CLIPVR-MGDA: OPTIMAL STOCHASTIC ALGORITHM FOR GENERALIZED-SMOOTH MULTI-OBJECTIVE OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-objective optimization (MOO) has become increasingly important in modern machine learning applications. While many stochastic balanced descent algorithms (MGDA) have been developed for MOO, they typically suffer from suboptimal sample complexity for finding a Pareto stationary point. This has motivated MOCO+, a variance-reduced algorithm for MOO. However, its convergence guarantee relies on restrictive assumptions and applies only to smooth objective functions that are uncommon in applications with complex models. We develop *ClipVR-MGDA*, an optimal stochastic algorithm for nonconvex MOO under a broad class of generalized-smooth objectives. *ClipVR-MGDA* integrates a matrix-valued SPIDER estimator with MGDA-type updates to construct variance-reduced multi-gradient estimates across all objectives. To enable effective multi-gradient variance reduction under generalized-smooth objectives, we further introduce carefully designed gradient-clipping schemes for both the parameter update and the preference vector update. Under standard assumptions, we prove that *ClipVR-MGDA* converges to an ϵ-Pareto stationary point with the optimal $\mathcal{O}(\epsilon^{-3})$ sample complexity. Experiments on multi-task learning benchmarks demonstrate improved performance over existing stochastic MOO baselines.

1 INTRODUCTION

Multi-objective optimization (MOO) provides a mathematical framework for learning problems in which multiple objectives need to be optimized simultaneously. It has been widely used in multi-task learning applications (Sener and Koltun, 2018), where different tasks or performance criteria may compete with each other. A standard formulation of MOO is written as

$$\min_{x\in\mathcal{X}} F(x) := \big(f^1(x), \ldots, f^K(x)\big),$$

where K is the total number of objectives, and each $f^k : \mathcal{X} \to \mathbb{R}$ denotes one objective. The minimum in this problem is understood with respect to the partial order induced by objective-wise comparison. In the stochastic setting, each objective is usually written as $f^k(x) := \mathbb{E}_\xi[f^k(x;\xi)]$ for a given variable x, where ξ denotes a data sample. Since different objectives may not admit a single common minimizer, MOO usually adopts *Pareto optimality* and *Pareto stationarity* as the solution concepts. Compared with the classical notion of stationarity for scalar optimization, these concepts better reflect the trade-offs among different objectives.

A key challenge in MOO algorithm design is handling the conflicts among objective gradients, which may point in incompatible directions, so descending along one objective may improve one task while harming other tasks. This motivates constructing update directions that appropriately balance progress across objectives. The Multiple-Gradient Descent Algorithm (MGDA) (Désidéri, 2012) addresses this issue by seeking a conflict-aware (CA) descent direction through solving a subproblem that maximizes the minimal improvement among objectives. Following this principle, several related methods, such as CAGrad (Liu et al., 2021a) and PCGrad (Yu et al., 2020), have been proposed to provide more systematic treatments of gradient conflict in MOO. This line of research has also been extended to the stochastic setting, yielding algorithms such as SMG (Liu and Vicente, 2021), MOCO (Fernando et al., 2023) and MODO (Chen et al., 2023a).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Despite their simplicity, the existing stochastic MOO algorithms are designed for smooth objectives. This condition is often violated in modern ML models, including deep networks and foundation models (Zhang et al., 2020; Li et al., 2023). In addition, these algorithms suffer from suboptimal sample complexity due to the variance of stochastic gradients across multiple objectives. These issues have recently motivated variance-reduced methods for stochastic MOO. For instance, the MOCO+ algorithm (Fernando et al., 2024) adapts the STORM variance-reduction scheme (Cutkosky and Orabona, 2019) to reduce multi-objective gradient noise and achieves an improved sample complexity. However, its convergence guarantee is proved only under standard smooth objectives and relies on restrictive assumptions, such as uniformly bounded gradients and Lipschitz continuous objective functions, which further limit its practical performance in solving complex MOO problems. This motivates our goal in this work:

> *Can we develop an efficient stochastic algorithm for nonconvex MOO beyond smooth objectives, avoiding restrictive assumptions, while achieving optimal sample complexity?*

1.1 Our Contributions

We propose a new variance-reduction algorithm, named *ClipVR-MGDA*, for nonconvex MOO under generalized-smooth objective functions. Our algorithm integrates a matrix-valued, variance-reduced multi-gradient estimator based on the SPIDER scheme (Fang et al., 2018) with a double-clipping update rule, yielding the optimal sample complexity $\mathcal{O}(\epsilon^{-3})$ without requiring restrictive assumptions. Our main contributions are summarized as follows.

Mean-square generalized smoothness for MOO. We introduce a mean-square generalized-smoothness condition (see Assumption 3) to characterize objectives in stochastic nonconvex MOO. Under this condition, each objective admits a gradient-dependent mean-square smoothness parameter. This extends the classical mean-square smoothness condition commonly used in the analysis of variance reduction methods, and covers a broad class of modern machine learning models that exhibit generalized-smooth geometry.

Variance reduction for MOO without restrictive assumptions. *ClipVR-MGDA* provides the first variance reduction scheme for nonconvex MOO under generalized-smooth objectives. In particular, it leverages the SPIDER scheme to construct matrix-valued variance-reduced multi-gradient estimates across objectives. Compared with the existing variance-reduced MOO algorithm MOCO+ (which applies only to smooth objectives), our algorithm applies to the broader class of generalized-smooth MOO problems and avoids restrictive assumptions, including Lipschitz continuity of the objective functions and bounded gradients. Technically, we introduce a *multi-objective tracking sequence* to track the evolution of variance-reduced multi-gradient across objectives along an arbitrary simplex direction. This tracking sequence is compatible with the conflict-aware structure of MGDA-type updates and enables a tight characterization of optimization progress in MOO.

Double-clipping MGDA updates and optimal complexity. The variance-reduced multi-gradient estimator alone is insufficient to guarantee convergence. This is because the standard clipping used in SPIDER does not directly provide control of the estimator along arbitrary simplex directions, nor does it handle the additional preference-weight bias terms that arise when moving from single-objective optimization to generalized-smooth MOO. To address this issue, we propose a double-clipping scheme for the MGDA-type updates. For the model parameter update, we apply a customized clipping rule to control the update increment, which is required both to stabilize variance reduction and to keep iterates within the local neighborhood where the generalized-smooth descent lemma remains valid. For the preference vector update, we apply a separate clipping rule to control the update in the simplex and limit the accumulated bias induced by changing weights across iterations. Together, these two clipping mechanisms, combined with the variance-reduced multi-gradient estimator, yield an optimal sample complexity for our algorithm.

2 Related Works

Multi-objective optimization (MOO). From the multi-objective optimization (MOO) viewpoint, multi-task learning can be treated as the simultaneous optimization of several task losses. In this formulation, the main challenge is to construct an update direction that balances potentially conflicting task gradients. Gradient-balancing methods address this issue by adjusting both the model

update direction and the preference vector over objectives, so that different tasks can make balanced progress. A representative early work is Sener and Koltun (2018), which formulates multi-task learning (MTL) as a MOO problem and shows that MGDA (Désidéri, 2012) provides an effective framework for computing such common descent directions.

Existing gradient-balancing methods can be roughly organized into two lines. The first line focuses on practical conflict-mitigation rules. Methods such as CAGrad (Liu et al., 2021a), IMTL-Grad (Liu et al., 2021b), PCGrad (Yu et al., 2020), GradDrop (Chen et al., 2020), and RotoGrad (Javaloy and Valera, 2021) modify, project, or rebalance task gradients to reduce conflicts during training. These methods are often effective in practice, but they generally do not provide strong convergence guarantees in stochastic nonconvex settings.

The second line studies MGDA-type algorithms with provable convergence guarantees. Liu and Vicente (2021) establishes the first convergence result for stochastic MGDA, although the analysis requires growing batch sizes. Later works, including MoDo (Chen et al., 2023a) and SDMGrad (Xiao et al., 2023), remove this requirement by using double-sampling techniques to control the bias of gradient-related estimators while keeping the batch size fixed. MoCo (Fernando et al., 2023) and its variance-reduced extension (Fernando et al., 2024) further improve estimation quality by introducing auxiliary tracking variables. However, many of these provable methods still rely on bounded-gradient assumptions. A notable exception is Zhang et al. (2024), which proves convergence of double-sampling-based stochastic MGDA under generalized smoothness (Li et al., 2023), thereby relaxing the bounded-gradient requirement. More recently, Yang et al. (2026) further introduces a double-clipping technique to refine the convergence analysis.

Variance Reduction. Variance reduction has played a central role in stochastic optimization by reducing the variance of gradient estimators while preserving relatively low per-iteration cost. Early methods such as SAGA (Defazio et al., 2014) exploit stored gradient information in finite-sum problems, while recursive estimators such as SARAH (Nguyen et al., 2017) further improve estimator efficiency through biased but tightly controlled updates. More recent methods, including SPIDER (Fang et al., 2018) and STORM (Cutkosky and Orabona, 2019), show that recursive variance-reduction schemes can achieve near-optimal complexity guarantees for stochastic nonconvex optimization. These developments have made variance reduction a standard tool for designing theoretically grounded stochastic algorithms, and they also motivate its incorporation into gradient-balancing methods for multi-objective optimization.

Generalized Smoothness. The standard L-smoothness has been shown to be restrictive for modern non-convex setting. In particular, Zhang et al. (2020) observed that the local smoothness level may vary significantly along the optimization trajectory and can grow with the gradient norm. It motivated the (L_0, L_1)-smoothness condition, under which the Hessian norm is controlled by a linear function of $\|\nabla f(x)\|$. Subsequent works further studied optimization under unbounded or relaxed smoothness conditions, including generalized SignSGD (Crawshaw et al., 2022) and other adaptive methods (Crawshaw et al., 2022; Hübler et al., 2024). More general notions of generalized smoothness have also been proposed. Chen et al. (2023b) introduced a broader generalized-smooth framework that covers functions beyond the classical (L_0, L_1) setting, such as high-order polynomial and exponential-type objectives. Chen et al. (2023b) studied both convex and nonconvex optimization under generalized smoothness and established the equivalence between Hessian-based ℓ-smoothness and the local (r, ℓ)-smoothness formulation. These results provide useful descent lemmas for algorithms whose updates are controlled by clipping or normalization.

3 Preliminaries on Generalized-Smooth Functions

The standard L-smoothness assumption has been widely adopted in the analysis of gradient-based optimization methods. It posits that a differentiable function $f : \mathcal{X} \to \mathbb{R}$ has globally Lipschitz-continuous gradient, i.e., there exists $L > 0$ such that

$$\|\nabla f(x) - \nabla f(y)\| \le L\|x - y\|, \quad \forall x, y \in \mathcal{X}.$$

This condition provides a global control of gradient variation and underpins the classical descent lemma that is central to the convergence analysis of first-order methods. However, in many modern learning problems, particularly those involving deep neural networks and foundational models, their

smoothness constants may fail to be globally bounded. Empirically, the smoothness parameter in such models often appears to scale with the gradient norm (Zhang et al., 2020; Li et al., 2023). This observation motivates the following generalized smoothness condition (Li et al., 2023).

Definition 1 (*Generalized $\mathcal{L}$-smoothness, (Li et al., 2023)*)**.** *A real-valued twice-differentiable function $f : \mathcal{X} \to \mathbb{R}$ is called generalized $\mathcal{L}$-smooth if*

$$\|\nabla^2 f(x)\| \leq \mathcal{L}(\|\nabla f(x)\|)$$

almost everywhere in $\mathcal{X}$, where $\mathcal{L} : [0, +\infty) \to (0, +\infty)$ is a continuous non-decreasing function.

This definition recovers the standard L-smoothness when $\mathcal{L}$ is uniformly bounded. More generally, it allows the smoothness level to increase with the gradient norm $\|\nabla f(x)\|$, thereby capturing cases in which the Hessian spectrum cannot be controlled by a fixed global constant. A closely related definition is the following local generalized smoothness condition, which is often more convenient for establishing descent estimates along the iterates.

Definition 2 (*$(r, \mathcal{L})$-smoothness, (Li et al., 2023)*)**.** *Let $B(x, R)$ denote the Euclidean ball centered at x with radius R. A real-valued differentiable function $f : \mathcal{X} \to \mathbb{R}$ is called $(r, \mathcal{L})$-smooth if $B\big(x,\, r(\|\nabla f(x)\|)\big) \subseteq \mathcal{X}$ for any $x \in \mathcal{X}$, and for any $x_1, x_2 \in B\big(x,\, r(\|\nabla f(x)\|)\big)$,*

$$\|\nabla f(x_1) - \nabla f(x_2)\| \leq \mathcal{L}(\|\nabla f(x)\|)\, \|x_1 - x_2\|,$$

where $r, \mathcal{L} : [0, +\infty) \to (0, +\infty)$ are continuous functions, r is non-increasing, $\mathcal{L}$ is non-decreasing.

The above definition implies that, around each point x, the function behaves like a smooth function on a neighborhood whose radius depends on $\|\nabla f(x)\|$. Within this neighborhood, the effective smoothness constant is $\mathcal{L}(\|\nabla f(x)\|)$. The following proposition connects this local definition with the generalized $\mathcal{L}$-smoothness in Definition 1.

Proposition 1 (Equivalence of generalized smoothness, Proposition 3.2 in Li et al. (2023))**.** *Any $(r, \mathcal{L})$-smooth function is $\mathcal{L}$-smooth. Moreover, if f is $\mathcal{L}$-smooth and is differentiable and closed within its open domain $\mathcal{X}$, then f is (r, m)-smooth with $m(u) = \mathcal{L}(u + a)$, and $r(u) = \frac{a}{m(u)}$ for any $a > 0$.*

This equivalence allows us to switch flexibly between the Hessian-based generalized smoothness condition and the local generalized smoothness condition. In particular, as later we show, the $(r, m)-$smoothness is useful for verifying that each algorithmic update stays within a neighborhood in which a descent lemma applies.

4 Multi-Objective Optimization and Existing Algorithms

4.1 Problem Formulation

In contrast to single-objective optimization, which aims to minimize a single scalar objective function, multi-objective optimization (MOO) seeks to optimize multiple objective functions simultaneously. Specifically, we consider the following MOO problem

$$\text{(MOO):} \quad \min_{x \in \mathcal{X}} F(x) := (f^1(x), f^2(x), ..., f^K(x)),$$

where each $f^k : \mathcal{X} \to \mathbb{R}$ denotes the k-th objective function. Since $F(x)$ is vector-valued, the minimization above is understood with respect to the partial order induced by componentwise comparison, rather than the usual scalar order. Specifically, for two objective vectors $u, v \in \mathbb{R}^K$, we say that u *dominates* v if $u^k \leq v^k$ for all $k \in [K]$, with strict inequality for at least one objective. A feasible solution is *Pareto optimal* if no other feasible solution improves at least one objective without worsening another. This partial-order structure is a key distinction between MOO and scalar optimization. Because objectives may conflict, there may not exist a direction that decreases all objectives simultaneously. Accordingly, instead of the single-objective first-order condition $\|\nabla f(x)\| = 0$, MOO quantifies first-order stationarity via the following notion of Pareto stationary.

Definition 3. *A point $x \in \mathbb{R}^m$ is a Pareto stationary point if $\min_{w \in \Delta_K} \|\nabla F(x)w\| = 0$, where Δ_K is the probability simplex over $[K]$. Furthermore, a point x is an ϵ-accurate Pareto stationary point if*

$$\min_{w \in \Delta_K} \|\nabla F(x)w\| \leq \epsilon.$$

This definition is the multi-objective analogue of first-order stationarity in single-objective optimization. When the minimum value equals zero, there exists a convex combination of objective gradients that vanishes, indicating that no common first-order descent direction exists under the Pareto stationary criterion.

4.2 Existing Stochastic Algorithms for MOO

The multiple-gradient descent algorithm (MGDA) is a classical gradient-based method for MOO (Désidéri, 2012). It uses a gradient descent update of the form $x_{t+1} = x_t - \alpha_t d(x_t)$, where $d(x) = \nabla F(x) w^*(x)$ is the balanced gradient, and the preference vector w^* is obtained by solving the following subproblem.

$$w^*(x) \in \arg\min_{w} \|\nabla F(x) w\|^2, \ \text{ s.t. } w \in \Delta_K := \left\{ w \in \mathbb{R}^K \mid \mathbf{1}^\top w = 1, w \geq 0 \right\}.$$

Such balanced gradient addresses the issue that gradients of different objectives may conflict with one another by seeking an update direction that yields the best worst-case improvement across objectives. Many advanced MGDA variants have been proposed in the literature, e.g., (Liu et al., 2021a; Fernando et al., 2023; Chen et al., 2023a; Xiao et al., 2023; Liu and Vicente, 2021; Zhang et al., 2024; Liu et al., 2021a). In particular, Liu and Vicente (2021) establish the convergence guarantee for stochastic MGDA under smooth and strongly convex objectives using growing batch sizes. Zhang et al. (2024) further establish the convergence of (stochastic) MGDA for nonconvex objectives under generalized smoothness using independent stochastic gradient samples, and establish a sample complexity of $\mathcal{O}(\epsilon^{-4})$ to achieve an ϵ-Pareto stationary point. Please refer to Table 1 for a comprehensive comparison of the existing stochastic algorithms for nonconvex MOO.

Table 1: Comparison of related stochastic algorithms for nonconvex MOO.

Method	Smoothness[1]	Additional Assumption(s)[2]	Complexity[3]
SMG (Liu and Vicente, 2021)	(LS)	(OD)	N/A
CR-MOGM (Zhou et al., 2022)	(LS)	(BF) and (BG)	$\mathcal{O}(\epsilon^{-4})$
MoCo (Fernando et al., 2023)	(LS)	(BF) and (BG)	$\mathcal{O}(\epsilon^{-4})$
MoDo (Chen et al., 2023a)	(LS)	(BG)	$\mathcal{O}(\epsilon^{-4})$
SDMGrad (Xiao et al., 2023)	(LS)	(BG)	$\mathcal{O}(\epsilon^{-4})$
MOCO+ (Fernando et al., 2024)	(LS)	(BG)	$\mathcal{O}(\epsilon^{-3})$
Stochastic MGDA (Zhang et al., 2024)	(GS)	None	$\mathcal{O}(\epsilon^{-4})$
ClipVR-MGDA(ours)	**(msGS)**	**None**	$\mathcal{O}(\epsilon^{-3})$

[1] (LS) : L-smoothness; (GS) : generalized ℓ-smoothness; (msGS): mean-square generalized smoothness
[2] (OD) : optimal-direction in each iteration; (BF) : bounded function values; (BG) : bounded gradient.
[3] Complexity refers to the sample complexity for finding an ϵ-Pareto stationary point.

While simple and effective, existing stochastic MGDA methods for nonconvex MOO have suboptimal sample complexity. This motivates developing variance-reduced algorithms for MOO. Recently, Fernando et al. (2024) applied the STORM variance reduction scheme (Cutkosky and Orabona, 2019) to improve the sample complexity of MGDA in nonconvex MOO. Specifically, the proposed MOCO+ algorithm takes the following updates.

$$\begin{aligned}
x_{t+1} &= x_t - \alpha_t Y_t w_t \\
w_{t+1} &= \Pi_{\Delta^M} \left(w_t - \gamma_t Y_t^\top Y_t w_t \right) \\
y_{t+1}^k &= \Pi_L \left(\nabla f^k(x_{t+1}, \xi_{t+1}) + (1 - \beta_{t+1}) \left(y_t^k - \nabla f^k(x_t, \xi_{t+1}) \right) \right), \ \forall k \in [K],
\end{aligned}$$

where the multi-objective gradient estimates $\{y^k\}_{k=1}^K$ are constructed via the STORM scheme. Under smooth objectives, Fernando et al. (2024) proved that MOCO+ achieves the optimal sample complexity $\mathcal{O}(\epsilon^{-3})$. Nevertheless, their convergence analysis relies on additional restrictive assumptions, such as Lipschitz continuity of the objective functions and bounded gradients.

In this work, we further advance variance-reduced methods for nonconvex MOO under generalized-smooth objective functions, and obtain the optimal sample complexity without the aforementioned restrictive assumptions.

5 MGDA WITH CLIPPED MULTI-GRADIENT VARIANCE REDUCTION

In this section, we propose *ClipVR-MGDA*, a new algorithm aiming to establish convergence in terms of ϵ-Pareto stationary points for nonconvex MOO under the generalized-smooth geometry condition defined as Assumption 3. *ClipVR-MGDA* can also be viewed as a variance-reduced extension of stochastic MGDA under generalized smoothness (Zhang et al., 2024).

5.1 ALGORITHM DESIGN

Algorithm 1 ClipVR-MGDA

Initialize: $S_1, S_2, q, T, K, \beta, \rho, x_0, w_0$.
for $t = 0, ..., T-1$ **do**
  if $t \bmod q = 0$ **then**
    Compute $v_t^k = \nabla f_{S_1}^k(x_t)$, $\forall k \in [K]$
  else
    Compute $v_t^k = v_{t-1}^k + \nabla f_{S_2}^k(x_t) - \nabla f_{S_2}^k(x_{t-1})$, $\forall k \in [K]$
  end if
  Assemble $V_t = (v_t^1, ..., v_t^K)$
  $x_{t+1} = x_t - \alpha_t V_t w_t$
  $w_{t+1} = \Pi_{\Delta^K}\left(w_t - \beta(\gamma_t V_t^\top V_t + \rho I_K) w_t\right)$
end for

We present *ClipVR-MGDA* in Algorithm 1. Our algorithm design exhibits the following key features.

- **Variance-reduced multi-gradient estimate.** *ClipVR-MGDA* utilizes a matrix-valued SPIDER estimator V_t to approximate the multi-gradient in MOO. For each objective k, the corresponding gradient estimate v_t^k is updated via the SPIDER recursion Fang et al. (2018). Every q iterations, a refresh step with a large batch size $|S_1|$ is performed to align V_t with $\nabla F(x_t)$; in the intermediate iterations, a smaller batch size $|S_2|$ is used for incremental updates. This matrix-valued variance-reduced estimator tracks all objectives simultaneously (and hence any simplex-weighted combination), yields bounds over the simplex and is therefore central to our convergence analysis.
- **Double gradient-clipping.** In order to stabilize the variance reduction updates under generalized-smooth objectives, we introduce two gradient-clipping parameters, i.e., $\alpha_t = \min\{c_1/\|V_t w_t\|, c_2\}$ and $\gamma_t = \min\{c_3/\|V_t w_t\|, c_4\}$ for some constants $c_1, c_2, c_3, c_4 > 0$. For the model parameter update, α_t controls the update increment of x_t, which is required both to stabilize variance reduction and to keep iterates within the local neighborhood where the generalized-smooth descent lemma remains valid. For the preference vector update, γ_t stabilizes the update of w_t in the simplex. We note that the clipping on α_t alone is not sufficient to guarantee convergence, since the instability induced by the variance-reduced estimator and the potentially large local gradient variation can propagate to the preference vector update. Therefore, γ_t is introduced to mitigate this effect on the simplex-weight dynamics and to limit the accumulated MOO-specific bias induced by changing weights across iterations. Together, these two clipping mechanisms yield a more balanced and stable MOO algorithm.

5.2 TECHNICAL ASSUMPTIONS

To establish convergence, we impose the following assumptions:

Assumption 1. *Each objective function $f^k : \mathcal{X} \to \mathbb{R}$, $\forall k \in [K]$ is twice differentiable. There exists a constant $F_{init} > 0$, such that $f^k(x_0) - f^{k,*} \le F_{init}$, where $f^{k,*} := \inf_{x \in \mathcal{X}} f^k(x) > -\infty$.*

Assumption 2. *The following statements hold regarding the stochastic gradient $\nabla f^k(x; \xi)$.*

1. *Unbiased Estimator: $\mathbb{E}_\xi[\nabla f^k(x; \xi)] = \nabla f^k(x)$.*
2. *Bounded Variance: There exists some $\sigma \ge 0$ such that*

$$\mathbb{E}_\xi[\|\nabla f^k(x; \xi) - \nabla f^k(x)\|^2] \le \sigma^2, \quad \forall x \in \mathcal{X}, k \in [K].$$

Assumption 3 (Mean-square generalized smoothness)**.** *There exists a positive non-decreasing function $g : [0, \infty) \to (0, \infty)$ such that for all $k \in [K]$ and all $x, y \in \mathcal{X}$,*

$$\mathbb{E}_\xi \left\| \nabla f^k(x;\xi) - \nabla f^k(y;\xi) \right\|^2 \le g\big(\|\nabla f^k(x)\|\big)^2 \|x - y\|^2.$$

There exists a constant $a > 0$, and denote $\ell(u) := g(u + a)$, $r(u) := \frac{a}{\ell(u)}$ for all $u \ge 0$, such that ℓ is continuous non-decreasing, r is non-increasing. Moreover, $B\big(x,\ r\big(\|\nabla f^k(x)\|\big)\big) \subseteq \mathcal{X}$ for every $k \in [K]$ and every $x \in \mathcal{X}$. Finally, $\varphi(s) := s^2/(2\ell(2s))$ is non-decreasing and unbounded.

The above mean-square generalized smoothness condition extends the generalized smoothness condition in Definition 2 introduced by Li et al. (2023), which holds for many modern machine learning models. This mean-square version is specially designed for the stochastic setting and is aligned with the standard mean-square smoothness assumption.

Proposition 2. *Under Assumption 3, the following statements hold for all $k \in [K]$.*

1. For all $x, y \in \mathcal{X}$, we have

$$\mathbb{E}_\xi \left\| \nabla f^k(x;\xi) - \nabla f^k(y;\xi) \right\|^2 \le \ell\big(\|\nabla f^k(x)\|\big)^2 \|x - y\|^2,$$

and

$$\|\nabla f^k(x) - \nabla f^k(y)\| \le g\big(\|\nabla f^k(x)\|\big) \|x - y\| \le \ell(\|\nabla f^k(x)\|)\|x - y\|.$$

2. The function f^k is (r, ℓ)-smooth, and hence ℓ-smooth.

5.3 Convergence Analysis

We obtain the following convergence guarantee for ***ClipVR-MGDA***, showing in particular that it attains the optimal sample complexity $\mathcal{O}(\epsilon^{-3})$.

Theorem 1. *Let Assumptions 1, 2, and 3 hold. Given $0 < \theta, \epsilon < 1$, set hyperparameters $n_0 = \mathcal{O}(1)$, $\beta = \mathcal{O}(\epsilon^{-2})$, $\rho = \mathcal{O}(\epsilon^2)$, $\delta = \mathcal{O}(1)$, and $\eta_1 = \Theta(\epsilon)$. Choose batch sizes and block size $S_1 = \mathcal{O}(\epsilon^{-2})$, $S_2 = \mathcal{O}(\epsilon^{-1})$, and $q = \mathcal{O}(\epsilon^{-1})$. Set the clipping learning rates as $\gamma_t = \min\{\frac{\eta_1}{\|V_t w_t\|}, \frac{1}{2}\}$ and $\alpha_t = \delta\gamma_t = \min\{\frac{\eta_1}{n_0 \ell(M+1)\|V_t w_t\|}, \frac{1}{2n_0\ell(M+1)}\}$. Then, with stochastic gradient complexity $\mathcal{O}(\epsilon^{-3})$, the iterates generated by Algorithm 1 satisfy, with probability at least $1 - \theta$,*

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla F(x_t) w_t\| \le \epsilon.$$

The proof of the theorem can be found in Appendix C. Our analysis is based on a stopping-time argument. We first restrict the convergence analysis to the interval before the stopping time τ. With a suitable definition of τ, all local conditions required by generalized smoothness and variance reduction remain valid on this interval. The stopping time is used to localize the trajectory to the event on which the objective values, gradient norms, and SPIDER tracking errors remain controlled. After proving the desired descent estimate on this localized event, we choose the parameters so that the stopping event fails with probability at most θ. We then decompose the descent inequality into the following three key components.

$$\begin{aligned}(F(x_{t+1}) - F(x_t))w \le &-\frac{\alpha_t}{2}\big(1 - \alpha_t \ell(M+1)\big)\|V_t w_t\|^2 \\ &+ \frac{\alpha_t}{2}\|(V_t - \nabla F(x_t))w\|^2 - \alpha_t \langle V_t w - V_t w_t, V_t w_t\rangle,\end{aligned}$$

which includes a main descent term $\|V_t w_t\|$, an estimation error term $\|(V_t - \nabla F(x_t))w\|$, and a multi-objective preference bias term $\langle V_t w - V_t w_t, V_t w_t\rangle$. This decomposition mirrors the classical descent structure in (Fang et al., 2018; Zhang et al., 2024) while isolating the additional bias term introduced by the multi-objective and variance-reduced setting.

The main technical novelties in handling these terms are summarized as follows:

- **Multi-gradient variance reduction.** Leveraging the recursive structure of SPIDER, we show that for any simplex vector w, the variance-reduced multi-gradient estimation error satisfies $\mathbb{E}\|(V_t - \nabla F(x_t))w\|^2 = \mathcal{O}(\epsilon^2)$. This bound not only controls the estimation error term in the descent decomposition, but also supports the analysis of the clipped preference vector update, where residual terms involving $\|V_t^\top V_t w_t\|^2$ must be controlled via the matrix error $V_t - \nabla F(x_t)$. Importantly, it is also crucial for the final Pareto-stationarity argument, since the analysis must hold uniformly over arbitrary simplex directions rather than only along the algorithm's current preference weight w_t.

- **Double gradient-clipping.** The clipping parameter α_t (for the parameter update) is used to establish the estimation error bound $\mathbb{E}\|(V_t - \nabla F(x_t))w\|^2 = \mathcal{O}(\epsilon^2)$ and to control the increment $\|x_{t+1} - x_t\| \leq \mathcal{O}(\epsilon)$, which both stabilizes the variance reduction scheme and keeps the iterates within the local neighborhood where the generalized-smooth descent lemma applies. The other clipping parameter γ_t stabilizes the simplex-weight update. In particular, the preference bias term $-\alpha_t\langle V_t w - V_t w_t, V_t w_t\rangle$ is linked to a telescoping term of the form $\frac{\alpha_t}{\gamma_t}\big(\|w_t - w\|^2 - \|w_{t+1} - w\|^2\big)$. Accordingly, we set the ratio $\delta = \alpha_t/\gamma_t$ to be constant so that these terms telescope across iterations. Finally, enforcing the clipping condition $\gamma_t\|V_t w_t\| \leq \eta_1$ controls the residual terms induced by the projected preference update, preventing the matrix estimator V_t from inducing unstable changes in the simplex variable.

6 EXPERIMENTS

In this section, we evaluate *ClipVR-MGDA* from the perspectives of Pareto stationarity and training stability. We compare *ClipVR-MGDA* with several representative MOO baselines, including double-sampling MGDA (Zhang et al., 2024), MoDo (Chen et al., 2023a), SDMGrad (Xiao et al., 2023), MoCo (Fernando et al., 2023), and MOCO+ (Fernando et al., 2024). For all algorithms, we perform a broad hyperparameter search, further refine the most promising configurations, and report results using the best-performing setting. Each experiment is run with multiple random seeds. To ensure a fair comparison across methods with different batch size and update rule, we plot all curves against the cumulative number of gradient queries, rather than the iteration number. We refer to Section E in the appendix for the ablation study and full experimental and hyperparameter details.

Synthetic Dataset. We first test the proposed *ClipVR-MGDA* over a synthetic MOO dataset. We construct a controlled stochastic multi-objective quadratic problem, where the objectives are defined as a collection of different quadratic functions. The problem contains $K = 3$ objectives in $d = 20$ dimensions. All stochastic methods are run for 500 iterations and averaged over 5 random seeds. Since different algorithms require different numbers of stochastic-gradient evaluations per iteration, we report convergence as a function of cumulative stochastic-gradient queries; this directly accounts for the mini-batch sizes and oracle calls used by each update. The primal learning rates are on the order of 10^{-3}, typically around 5×10^{-3}. The results in Figure 1 show that, under the setting $\sigma = 512$, *ClipVR-MGDA* clearly outperforms the baselines in terms of the multi-gradient norm while also achieving the lowest or near-lowest objective loss. This suggests that the proposed clipping and variance-reduction improve both query efficiency and training stability in noisy stochastic MOO problems.

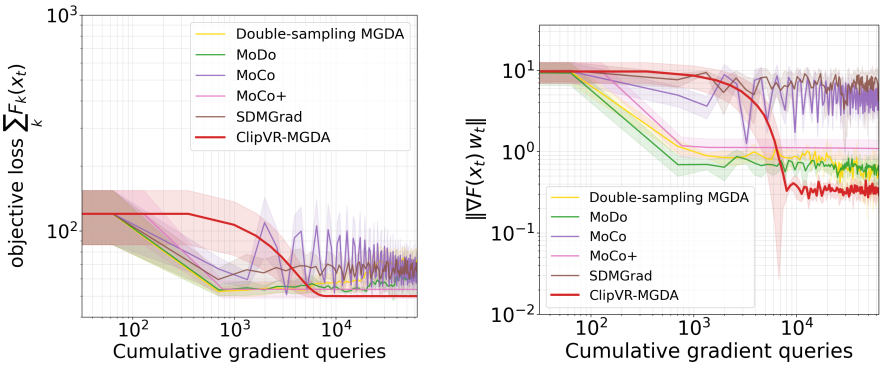

Figure 1: Loss/Multi-gradient vs. Cumulative Gradient Queries on Synthetic Dataset

Multi-MNIST. We conduct another experiment on the 2-digit Multi-MNIST benchmark (Sun, 2019), intending to compare *ClipVR-MGDA* with baselines in real-world data. Most of the baseline implementations can be found on LibMTL (Lin and Zhang, 2023). We adopt the neural network architecture summarized in Table 10, which consists of a shared encoder and two task-specific classification heads, for the left/right-digit classification task. Each task is trained with the cross-entropy loss using training batch size 128 and test batch size 256. For fairness, we fine-tune all methods so that their effective step sizes are at the 10^{-2} scale, and use the same optimizer setting without additional weight decay during training. The detail of hyper-parameter settings are summarized in Table 9. The results in Figure 2 show that *ClipVR-MGDA* is more query-efficient than the baselines, achieving the best test accuracy and test loss with the same cumulative gradient-query budget.

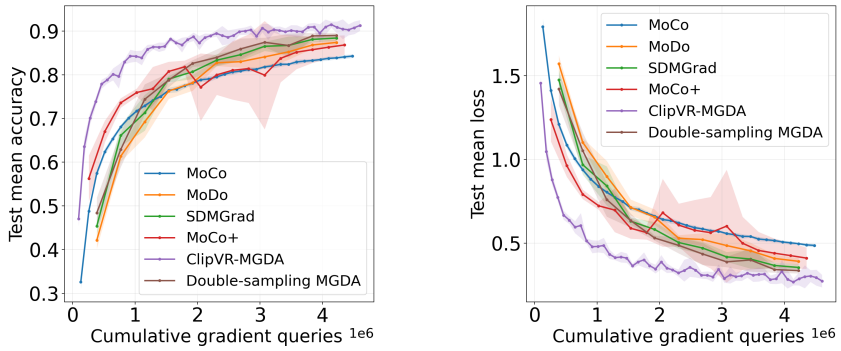

Figure 2: Test Mean Accuracy/Loss vs. Cumulative Gradient Queries on Multi-MNIST Dataset

7 Conclusion

In this work, we studied stochastic nonconvex MOO under generalized smoothness. We proposed ClipVR-MGDA, which combines matrix-valued SPIDER tracking with MGDA-style conflict-aware updates and a double-clipping mechanism for stabilizing both model and preference-vector updates. Under a mean-square generalized-smoothness condition, we established convergence to an ϵ-Pareto stationary point with optimal $O(\epsilon^{-3})$ sample complexity, without requiring bounded gradients or Lipschitz continuity of the objective functions. Experiments on synthetic and multi-task learning benchmarks demonstrate improved query efficiency and training stability over existing stochastic MOO baselines.

AI USE STATEMENT

We used generative AI tools for language editing, readability improvement, and assistance with literature search and discovery. Generative AI was not used to produce theoretical results or proofs, construct datasets, design experiments, or draw conclusions from experimental results. All AI-assisted revisions and literature-related outputs were manually reviewed by the authors, who take full responsibility for the final manuscript.

REFERENCES

Yossi Arjevani, Yair Carmon, John C. Duchi, Dylan J. Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization, 2022. URL https://arxiv.org/abs/1912.02365.

Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1):165–214, 2023.

Arthur Asuncion, David Newman, et al. Uci machine learning repository, 2007.

Dimitri P Bertsekas. Nonlinear programming. *Journal of the Operational Research Society*, 48(3): 334–334, 1997.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Lisha Chen, Heshan Fernando, Yiming Ying, and Tianyi Chen. Three-way trade-off in multi-objective learning: Optimization, generalization and conflict-avoidance. *Advances in Neural Information Processing Systems*, 36:70045–70093, 2023a.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International conference on machine learning*, pages 794–803. PMLR, 2018.

Zhao Chen, Jiquan Ngiam, Yanping Huang, Thang Luong, Henrik Kretzschmar, Yuning Chai, and Dragomir Anguelov. Just pick a sign: Optimizing deep multitask models with gradient sign dropout. *Advances in Neural Information Processing Systems*, 33:2039–2050, 2020.

Ziyi Chen, Yi Zhou, Yingbin Liang, and Zhaosong Lu. Generalized-smooth nonconvex optimization is as efficient as smooth nonconvex optimization. In *International Conference on Machine Learning*, 2023b.

Michael Crawshaw, Mingrui Liu, Francesco Orabona, Wei Zhang, and Zhenxun Zhuang. Robustness to unbounded smoothness of generalized SignSGD. In *Advances in Neural Information Processing Systems*, volume 35, pages 9955–9968, 2022.

Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex SGD. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, pages 15210–15219, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/b8002139cdde66b87638f7f91d169d96-Abstract.html.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems 27 (NeurIPS 2014)*, pages 1646–1654, 2014. URL https://proceedings.neurips.cc/paper/2014/hash/937964195d6fb3a55cd7cc578165f058-Abstract.html.

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

Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/1543843a4723ed2ab08e18053ae6dc5b-Paper.pdf.

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

Heshan Devaka Fernando, Han Shen, Miao Liu, Subhajit Chaudhury, Keerthiram Murugesan, and Tianyi Chen. Mitigating gradient bias in multi-objective learning: A provably convergent approach. In *The eleventh international conference on learning representations*, 2023.

Jenna C Fromer and Connor W Coley. Computer-aided multi-objective optimization in small molecule discovery. *Patterns*, 4(2), 2023.

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

Horst W Hamacher and K-H Küfer. Inverse radiation therapy planning—a multiple objective optimization approach. *Discrete Applied Mathematics*, 118(1-2):145–161, 2002.

Conor F Hayes, Roxana Rădulescu, Eugenio Bargiacchi, Johan Källström, Matthew Macfarlane, Mathieu Reymond, Timothy Verstraeten, Luisa M Zintgraf, Richard Dazeley, Fredrik Heintz, et al. A practical guide to multi-objective reinforcement learning and planning. *arXiv preprint arXiv:2103.09568*, 2021.

Florian Hübler, Ilyas Fatkhullin, and Niao He. Parameter-agnostic optimization under relaxed smoothness. In *International Conference on Artificial Intelligence and Statistics*, 2024.

Adrián Javaloy and Isabel Valera. Rotograd: Gradient homogenization in multitask learning. *arXiv preprint arXiv:2103.02631*, 2021.

Haochuan Li, Jian Qian, Yi Tian, Alexander Rakhlin, and Ali Jadbabaie. Convex and non-convex optimization under generalized smoothness, 2023. URL https://arxiv.org/abs/2306.01264.

Baijiong Lin and Yu Zhang. LibMTL: A Python library for multi-task learning. *Journal of Machine Learning Research*, 24(209):1–7, 2023.

Baijiong Lin, Weisen Jiang, Pengguang Chen, Feiyang Ye, Yu Zhang, Ying-Cong Chen, Shu Liu, Ivor W Tsang, and James T Kwok. Dual-balancing for multi-task learning. *Neural Networks*, page 108317, 2025.

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

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. *Advances in Neural Information Processing Systems*, 34:18878–18890, 2021a.

Liyang Liu, Yi Li, Zhanghui Kuang, Jing-Hao Xue, Yimin Chen, Wenming Yang, Qingmin Liao, and Wayne Zhang. Towards impartial multi-task learning. In *International conference on learning representations*, 2021b.

Suyun Liu and Luis Nunes Vicente. The stochastic multi-gradient algorithm for multi-objective optimization and its application to supervised machine learning, 2021. URL https://arxiv.org/abs/1907.04472.

David Love and Güzin Bayraksan. Phi-divergence constrained ambiguous stochastic programs for data-driven optimization. *Technical report, Department of Integrated Systems Engineering, The Ohio State University, Columbus, Ohio*, 2015.

I Maruşciac. On fritz john type optimality criterion in multi-objective optimization. *Mathematica-Revue d'analyse numérique et de théorie de l'approximation. L'analyse numérique et la théorie de l'approximation*, pages 109–114, 1982.

Lam M. Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *Proceedings of the 34th International Conference on Machine Learning (ICML 2017)*, volume 70 of *Proceedings of Machine Learning Research*, pages 2613–2621, 2017. URL https://proceedings.mlr.press/v70/nguyen17b.html.

Amirhossein Reisizadeh, Haochuan Li, Subhro Das, and Ali Jadbabaie. Variance-reduced clipping for non-convex optimization. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

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

Aman Sinha, Hongseok Namkoong, Riccardo Volpi, and John Duchi. Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017.

Shao-Hua Sun. Multi-digit mnist for few-shot learning, 2019. URL https://github.com/shaohua0116/MultiDigitMNIST.

Moritz Werling, Julius Ziegler, Sören Kammel, and Sebastian Thrun. Optimal trajectory generation for dynamic street scenarios in a frenet frame. In *2010 IEEE international conference on robotics and automation*, pages 987–993. IEEE, 2010.

Peiyao Xiao, Hao Ban, and Kaiyi Ji. Direction-oriented multi-objective learning: Simple and provable stochastic algorithms. *Advances in Neural Information Processing Systems*, 36:4509–4533, 2023.

Yufeng Yang, Fangning Zhuo, Ziyi Chen, Heng Huang, and Yi Zhou. Distributionally robust multi-objective optimization, 2026. URL https://arxiv.org/abs/2605.05660.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33: 5824–5836, 2020.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations*, 2020.

Qi Zhang, Peiyao Xiao, Shaofeng Zou, and Kaiyi Ji. Mgda converges under generalized smoothness, provably. *arXiv preprint arXiv:2405.19440*, 2024.

Yong Zheng and David (Xuejun) Wang. Multi-objective recommendations. In *KDD '21*, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383325.

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 S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 38103–38115. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/f91bd64a3620aad8e70a27ad9cb3ca57-Paper-Conference.pdf.

Appendix

A Matrix-Valued Variance-Reduced Multi-Gradient Estimate Tracking **15**

B Technical Lemmas for ClipVR-MGDA **16**

B.1 Decomposition of Multi-objective Descent Lemma 16

B.2 A Unified Bound for the Steps with w_t Update 18

B.3 Bias Bound via Variance-Reduced Estimation 19

C Proof of the Final Convergence Guarantee **22**

C.1 Sample Complexity 22

C.2 Stopping Time and Convergence 23

D Auxiliary Results **26**

D.1 Proof of Proposition 2 26

E Experiment Detail **27**

E.1 Ablation Study 27

E.2 Experiment on Multi-MNIST 29

E.3 Experiment on Synthetic Dataset 30

F Frequently Asked Questions **31**

A MATRIX-VALUED VARIANCE-REDUCED MULTI-GRADIENT ESTIMATE TRACKING

Following the principle of SPIDER (Fang et al., 2018), this section introduces the matrix-valued variance-reduced tracking framework used in our analysis. The goal is to track the multi-gradient matrix $\nabla F(x_t)$ and to control its estimation error along arbitrary simplex directions $w \in \Delta_K$.

Background. For the multi-objective case, consider matrix-valued deterministic sequence

$$\{Q(x_t)\}_{t=0}^{\infty} := \{(q^1(x_t), \dots, q^K(x_t))\}_{t=0}^{\infty},$$

where $q^k(x_t) \in \mathbb{R}^N$ for each $k \in [K]$. Naturally, for any $t \geq 1$, and any $s \in \{1, \dots, t\}$, one can rewrite such sequence as

$$Q(x_t) = Q(x_{t-s}) + \sum_{\nu=t-s+1}^{t} \big(Q(x_\nu) - Q(x_{\nu-1})\big).$$

The SPIDER Fang et al. (2018) would like to construct a stochastic tracking sequence for $Q(x_t)$ when only its unbiased stochastic estimator $\tilde{Q}(x_t; \xi_{t+1}) := \{(\tilde{q}^1(x_t; \xi_{t+1}), \dots, \tilde{q}^K(x_t; \xi_{t+1}))\}_{t=0}^{\infty}$ is available instead of the exact quantity $Q(x_t)$.

To match our algorithmic timeline, we let $\mathcal{F}_t$ information available *before* drawing the fresh sample used at iteration $t+1$. More precisely, $\mathcal{F}_t = \sigma(x_0, \xi_1, x_1, \xi_2, ..., x_t)$ is the natural filtration generated by all randomness up to time t, so that x_t, w_t, and V_t are all $\mathcal{F}_t$-measurable, while the fresh sample ξ_{t+1} used to construct the next estimator is independent of $\mathcal{F}_t$. For brevity, we write $\xi_{1:t+1} := (\xi_1, \dots, \xi_{t+1})$ and $x_{0:t} := (x_0, \dots, x_t)$.

Suppose that, for each $\nu \geq 1$, we have access to an unbiased stochastic estimator $\tilde{Q}(x_\nu; \xi_{\nu+1}) - \tilde{Q}(x_{\nu-1}; \xi_{\nu+1})$ of the increment $Q(x_\nu) - Q(x_{\nu-1})$, namely

$$\mathbb{E}\Big[\tilde{Q}(x_\nu; \xi_{\nu+1}) - \tilde{Q}(x_{\nu-1}; \xi_{\nu+1}) \mid \mathcal{F}_\nu\Big] = Q(x_\nu) - Q(x_{\nu-1}).$$

Then we define the corresponding tracking estimator

$$V_t := V_{t-s} + \sum_{\nu=t-s+1}^{t} \Big(\tilde{Q}(x_\nu; \xi_{\nu+1}) - \tilde{Q}(x_{\nu-1}; \xi_{\nu+1})\Big)$$

where $\tilde{Q}(x_0)$ is the initial estimator available at the beginning.

For our task, taking $Q(x_t) = \nabla F(x_t)$ and $\tilde{Q}(x_t; \xi_{t+1}) = \nabla F(x_t; \xi_{t+1}), \tilde{Q}(x_{t-1}; \xi_{t+1}) = \nabla F(x_{t-1}; \xi_{t+1})$, its recursive update can be obtained by taking $s = 1$, that is

$$V_t = V_{t-1} + (\nabla F(x_t; \xi_{t+1}) - \nabla F(x_{t-1}; \xi_{t+1})).$$

To further analyze the quality of this estimator, we typically consider two aspects: expectation and variance. The expectation characterizes its average accuracy and reveals whether the estimator exhibits systematic bias, namely, whether it deviates from the target quantity on average. The variance quantifies the magnitude of its stochastic fluctuations around the target quantity, and thus reflects the reliability of the estimator by describing how much random error is accumulated along the stochastic tracking process.

Moreover, both of them should be conditional version, based on our goal to tracking a changing sequence, instead of a fixed point.

We first consider the expectation. What is important here is the following propagation property:

$$\mathbb{E}[V_t - Q(x_t) \mid \mathcal{F}_{t-s+1}] = \mathbb{E}[V_{t-s} - Q(x_{t-s}) \mid \mathcal{F}_{t-s+1}] = V_{t-s} - Q(x_{t-s}),$$

where we use the tower property together with the conditional unbiasedness of each stochastic increment. In other words, along the sub-path from $t-s$ to t, the SPIDER does not introduce any additional conditional bias to expectation: the conditional bias of V_t as an estimator of $Q(x_t)$ is exactly the anchor error inherited from V_{t-s}.

Moreover, if V_{t-s} is an accurate estimator of $Q(x_{t-s})$, then V_t remains accurate for $Q(x_t)$ in expectation throughout the current block. In particular, if the restart estimator is exact, i.e., $V_{t-s} = Q(x_{t-s})$, then

$$\mathbb{E}[V_t - Q(x_t) \mid \mathcal{F}_{t-s+1}] = 0,$$

so V_t is conditionally unbiased for $Q(x_t)$ within the block. More generally, even when V_{t-s} is only a stochastic but unbiased restart estimator, no further bias is accumulated inside the block; the estimation error of V_t is then governed by the variance accumulated from the local stochastic increment terms.

To analyze its variance, it is enough to study its behavior on an arbitrary w, or at least fixed $w \in \Delta_K$, which is all we care about. This is because $V_t - Q(x_t)$ is a linear map, and a linear map is completely determined by its action on vectors. Hence, by characterizing $(V_t - Q(x_t))w$ for arbitrary $w \in \Delta_K$, we obtain a precise description of the estimator error in all directions of interest. We formalize this below. Similar to the Prop. 1 in Fang et al. (2018), we now give the bound as the second-moment for the bias caused by such estimation.

B Technical Lemmas for ClipVR-MGDA

Let $F > 0$ be some constant. We will specify the restriction on its value at the end of the stopping-time section. Let

$$M := \sup\{z \geq 0 \mid \varphi(z) \leq F\}.$$

In addition, let $H_1, J_1 > 0$ be some constants and let $\theta \in (0,1)$ be probabilities, all of which will be determined before they are used in the following proof.

For Theorem 1, we set

$$\eta_1 := \frac{\theta\epsilon}{4(J_1 + \sqrt{K})} \in \Theta(\epsilon), \qquad n_0 \in \{n_0 \in \mathcal{O}(1) \mid n_0 \geq 1\},$$
$$S_1 := \frac{2\sigma^2}{\eta_1^2}, \qquad S_2 := \frac{2\sigma}{\eta_1 n_0}, \qquad q := \frac{\sigma n_0}{\eta_1}, \qquad \delta := \frac{1}{n_0 \ell(M+1)},$$
$$\gamma_t := \min\left\{\frac{\eta_1}{\|V_t w_t\|}, \frac{1}{2}\right\}, \qquad \beta := \frac{1}{K(M^2+\eta_1^2)}, \qquad \rho := \min\left\{\sqrt{\frac{\eta_1^2}{\beta}}, \eta_1^2\right\},$$
$$\alpha_t := \delta\gamma_t = \min\left\{\frac{\eta_1}{n_0\ell(M+1)\|V_t w_t\|}, \frac{1}{2n_0\ell(M+1)}\right\} = \min\left\{\frac{\delta\eta_1}{\|V_t w_t\|}, \frac{\delta}{2}\right\}.$$

and the iteration complexity $T > 0$ will be given at the end of proof for both cases, respectively. We define the stopping times as

$$\tau_1 := \min\{t \mid \exists k \in [K],\ f^k(x_{t+1}) - f^{k,*} > F\} \wedge T,$$
$$\tau_2 := \min\{t \mid \exists k \in [K],\ \|v_t^k - \nabla f^k(x_t)\| \geq H_1\} \wedge T,$$
$$\tau := \min\{\tau_1, \tau_2\}.$$

B.1 Decomposition of Multi-objective Descent Lemma

To simplify the one-step analysis, we use an equivalent one-step shifted indexing convention for the w-update throughout this part. Before proving the main theorem, we first present several lemmas needed for the proof.

Lemma 1. *Under the assumptions of Theorem 1, suppose there exist a stopping time τ and a constant M such that $\|\nabla f^k(x_t)\| \le M$ for all $k \in [K]$ and $t < \tau$. Then, for any $w \in \Delta_K$, we have*

$$F(x_{t+1})w - F(x_t)w \le -\frac{\eta_1\|V_t w_t\| - 2\eta_1^2}{4n_0\ell(M+1)} + \underbrace{\frac{\alpha_t}{2}\|(V_t - \nabla F(x_t))w\|^2}_{B_1} - \underbrace{\alpha_t\langle V_t w - V_t w_t, V_t w_t\rangle}_{B_2}. \tag{1}$$

Here, B_1 denotes the variance-reduced estimate error, and B_2 denotes the MOO-specific bias from the mismatch between w and w_t.

Proof. By the definition of the stopping time τ and the choice $M := \sup\{z \ge 0 \mid \varphi(z) \le F\}$, we have $\|\nabla f^k(x_t)\| \le M$ for all $k \in [K]$ and $t < \tau$. By Proposition 2 and Lemma 3.3 in Li et al. (2023), we first verify that each update step $t < \tau$ remains inside the local neighborhood with radius $r(\|\nabla f^k(x_t)\|) = \frac{b}{\ell(\|\nabla f^k(x_t)\|+b)}$, $\forall k \in [K]$ for some $b > 0$. Indeed, taking $b = 1$,

$$\|x_{t+1} - x_t\| = \|\alpha_t V_t w_t\| \le \frac{\eta_1}{n_0\ell(M+1)\|V_t w_t\|}\|V_t w_t\| = \frac{\eta_1}{n_0\ell(M+1)} \le r(M+1) \le r(\|\nabla f^k(x_t)\|),$$

where we employ $1/n_0 \le 1$, $\eta_1 < 1$ and non-increasingness of r.

Hence, the descent lemma from Lemma 3.3 in Li et al. (2023) is applicable, and for each $k \in [K]$,

$$\begin{aligned} f^k(x_{t+1}) &\le f^k(x_t) + \langle \nabla f^k(x_t), x_{t+1} - x_t\rangle + \frac{\ell(M+1)}{2}\|x_{t+1} - x_t\|^2 \\ &\le f^k(x_t) - \alpha_t\langle \nabla f^k(x_t), V_t w_t\rangle + \frac{\alpha_t^2\ell(M+1)}{2}\|V_t w_t\|^2. \end{aligned}$$

Therefore, for any $w \in \Delta_K$, where Δ_K denotes the K-dimensional simplex,

$$F(x_{t+1})w - F(x_t)w \le -\alpha_t\langle \nabla F(x_t)w, V_t w_t\rangle + \frac{\alpha_t^2\ell(M+1)}{2}\|V_t w_t\|^2.$$

We next decompose the first term on the right-hand side into the main descent part and two bias/error terms. Specifically,

$$\begin{aligned} -\alpha_t\langle \nabla F(x_t)w, V_t w_t\rangle &= -\alpha_t\langle \nabla F(x_t)w - V_t w_t, V_t w_t\rangle - \alpha_t\|V_t w_t\|^2 \\ &= -\alpha_t\langle \nabla F(x_t)w - V_t w, V_t w_t\rangle - \alpha_t\langle V_t w - V_t w_t, V_t w_t\rangle - \alpha_t\|V_t w_t\|^2 \\ &\le \frac{\alpha_t}{2}\|(V_t - \nabla F(x_t))w\|^2 - \frac{\alpha_t}{2}\|V_t w_t\|^2 - \alpha_t\langle V_t w - V_t w_t, V_t w_t\rangle, \end{aligned}$$

where the last inequality is due to Young's inequality.

Combining the above two inequalities, we obtain

$$F(x_{t+1})w - F(x_t)w \le -\frac{\alpha_t}{2}\big(1 - \alpha_t\ell(M+1)\big)\|V_t w_t\|^2 + \frac{\alpha_t}{2}\|(V_t - \nabla F(x_t))w\|^2 - \alpha_t\langle V_t w - V_t w_t, V_t w_t\rangle. \tag{2}$$

Next, using $\min\{|x|, x^2/2\} \ge |x| - 2$ and $\alpha_t \le \frac{1}{2n_0\ell(M+1)} \le \frac{1}{2\ell(M+1)}$, we have

$$\begin{aligned} \frac{\alpha_t}{2}\big(1 - \ell(M+1)\alpha_t\big)\|V_t w_t\|^2 &\ge \frac{\alpha_t}{4}\|V_t w_t\|^2 \\ &= \frac{\eta_1^2}{4n_0\ell(M+1)}\min\left(\left\|\frac{V_t w_t}{\eta_1}\right\|, \frac{1}{2}\left\|\frac{V_t w_t}{\eta_1}\right\|^2\right) \\ &\ge \frac{\eta_1\|V_t w_t\| - 2\eta_1^2}{4n_0\ell(M+1)}. \end{aligned}$$

Therefore, following equation (2), we now have

$$\begin{aligned} F(x_{t+1})w - F(x_t)w &\le -\frac{\eta_1\|V_t w_t\| - 2\eta_1^2}{4n_0\ell(M+1)} + \frac{\alpha_t}{2}\|(V_t - \nabla F(x_t))w\|^2 - \alpha_t\langle V_t w - V_t w_t, V_t w_t\rangle \\ &= -\frac{\eta_1\|V_t w_t\|}{4n_0\ell(M+1)} + \frac{\eta_1^2}{2n_0\ell(M+1)} + \frac{\alpha_t}{2}\|(V_t - \nabla F(x_t))w\|^2 - \alpha_t\langle V_t w - V_t w_t, V_t w_t\rangle. \end{aligned}$$

□

B.2 A Unified Bound for the Steps with w_t Update

In this part, we derive a unified one-step estimate for the iterations where w_t is updated. This bound will be used to derive the descent estimate lemma needed for the summation argument.

Lemma 2. *Under the assumptions of Theorem 1, suppose the conditions of Lemma 1 hold. Then, for any reference vector $w \in \Delta_K$, we have the followings hold for $t < \tau$:*

$$F(x_{t+1})w - F(x_t)w \leq -\frac{\eta_1\|V_t w_t\|}{4n_0\ell(M+1)} + \frac{\alpha_t}{2}\|(V_t - \nabla F(x_t))w\|^2 + \frac{\delta}{2\beta}(\|w_t - w\|^2 - \|w_{t+1} - w\|^2) + \delta\rho + \delta\beta\rho^2 + 2\delta\beta\eta_1^2\|V_t - \nabla F(x_t)\|^2 + 3\delta\eta_1^2. \tag{3}$$

Equivalently,

$$\|V_t w_t\| \leq \frac{4}{\delta\eta_1}\big(F(x_t)w - F(x_{t+1})w\big) + \frac{2\alpha_t}{\delta\eta_1}\|(V_t - \nabla F(x_t))w\|^2 + \frac{2}{\beta\eta_1}\big(\|w_t - w\|^2 - \|w_{t+1} - w\|^2\big) + \frac{4}{\eta_1}\big(\rho + \beta\rho^2 + 3\eta_1^2\big) + 8\beta\eta_1\|V_t - \nabla F(x_t)\|^2. \tag{4}$$

Proof. To this end, for any reference vector $w \in \Delta$, we first estimate the $\{w_t\}$-update:

$$\begin{aligned}
\|w_{t+1} - w\|^2 &= \|\Pi_\Delta(w_t - \beta[\gamma_t V_t^\top V_t w_t + \rho w_t]) - w\|^2 \\
&\leq \|(w_t - \beta[\gamma_t V_t^\top V_t w_t + \rho w_t]) - w\|^2 \\
&= \|w_t - w\|^2 - 2\beta\langle w_t - w, (\gamma_t V_t^\top V_t + \rho I)w_t\rangle + \beta^2\|(\gamma_t V_t^\top V_t + \rho I)w_t\|^2 \\
&\leq \|w_t - w\|^2 - 2\beta\langle w_t - w, \gamma_t V_t^\top V_t w_t\rangle - 2\beta\langle w_t - w, \rho w_t\rangle + \beta^2\|(\gamma_t V_t^\top V_t + \rho I)w_t\|^2 \\
&\leq \|w_t - w\|^2 - 2\beta\langle w_t - w, \gamma_t V_t^\top V_t w_t\rangle - 2\beta\langle w_t - w, \rho w_t\rangle + 2\beta^2\gamma_t^2\|V_t^\top V_t w_t\|^2 + 2\beta^2\rho^2,
\end{aligned}$$

where in the last step we use $w_t \in \Delta$, and hence $\|w_t\| \leq \|w_t\|_1 = 1$.

Rearrangement of the above yields

$$\begin{aligned}
-\alpha_t\langle V_t w - V_t w_t, V_t w_t\rangle &= \frac{\alpha_t}{2\gamma_t\beta}\cdot 2\beta\langle w_t - w, \gamma_t V_t^\top V_t w_t\rangle \\
&\leq \frac{\alpha_t}{2\gamma_t\beta}\Big(\|w_t - w\|^2 - \|w_{t+1} - w\|^2 - 2\beta\langle w_t - w, \rho w_t\rangle + 2\beta^2\gamma_t^2\|V_t^\top V_t w_t\|^2 + 2\beta^2\rho^2\Big) \\
&\leq \frac{\alpha_t}{2\gamma_t\beta}\big(\|w_t - w\|^2 - \|w_{t+1} - w\|^2\big) + \frac{\alpha_t\rho}{\gamma_t} + \alpha_t\gamma_t\beta\|V_t^\top V_t w_t\|^2 + \frac{\alpha_t\beta\rho^2}{\gamma_t},
\end{aligned}$$

where we use $w_t, w \in \Delta_K$, and thus $\langle w_t - w, w_t\rangle = \|w_t\|^2 - \langle w, w_t\rangle \leq 1$.

Substituting

$$\begin{aligned}
\|V_t^\top V_t w_t\|^2 &= \|(V_t - \nabla F(x_t) + \nabla F(x_t))^\top V_t w_t\|^2 \\
&\leq 2\|(V_t - \nabla F(x_t))^\top V_t w_t\|^2 + 2\|\nabla F(x_t)^\top V_t w_t\|^2 \\
&\leq 2\|V_t - \nabla F(x_t)\|^2\|V_t w_t\|^2 + 2\|\nabla F(x_t)\|^2\|V_t w_t\|^2.
\end{aligned}$$

into the previous bound yields

$$\begin{aligned}
-\alpha_t\langle V_t w - V_t w_t, V_t w_t\rangle &\leq \frac{\alpha_t}{2\gamma_t\beta}\big(\|w_t - w\|^2 - \|w_{t+1} - w\|^2\big) + \frac{\alpha_t\rho}{\gamma_t} + \frac{\alpha_t\beta\rho^2}{\gamma_t} \\
&\quad + 2\alpha_t\gamma_t\beta\|V_t - \nabla F(x_t)\|^2\|V_t w_t\|^2 + 2\alpha_t\gamma_t\beta\|\nabla F(x_t)\|^2\|V_t w_t\|^2 \\
&= \frac{\delta}{2\beta}\big(\|w_t - w\|^2 - \|w_{t+1} - w\|^2\big) + \delta\rho + \delta\beta\rho^2 \\
&\quad + 2\delta\gamma_t^2\beta\|V_t - \nabla F(x_t)\|^2\|V_t w_t\|^2 + 2\delta\gamma_t^2\beta\|\nabla F(x_t)\|^2\|V_t w_t\|^2.
\end{aligned}$$

Combining the above estimate with the descent inequality for the x-update in equation (2), and applying $\gamma_t \leq \eta_1/\|V_t w_t\|$, $\beta \leq KM^2$, we obtain

$$F(x_{t+1})w - F(x_t)w \leq -\frac{\eta_1\|V_t w_t\|}{4n_0\ell(M+1)} + \frac{\alpha_t}{2}\|(V_t - \nabla F(x_t))w\|^2 + \frac{\delta}{2\beta}(\|w_t - w\|^2 - \|w_{t+1} - w\|^2) + \delta\rho + \delta\beta\rho^2$$

$$+ 2\delta\beta\eta_1^2\|V_t - \nabla F(x_t)\|^2 + (1+2)\delta\eta_1^2.$$

Moreover, after simple rearrangement, it is now

$$\begin{aligned}\frac{\delta\eta_1}{4}\|V_t w_t\| &= \frac{\eta_1\|V_t w_t\|}{4n_0\ell(M+1)} \\ &\le \big(F(x_t)w - F(x_{t+1})w\big) + \frac{\alpha_t}{2}\|(V_t - \nabla F(x_t))w\|^2 + \frac{\delta}{2\beta}\big(\|w_t - w\|^2 - \|w_{t+1} - w\|^2\big) + \delta\rho + \delta\beta\rho^2 \\ &\quad + 2\delta\beta\eta_1^2\|V_t - \nabla F(x_t)\|^2 + (1+2)\delta\eta_1^2,\end{aligned}$$

and equivalently,

$$\begin{aligned}\|V_t w_t\| \le{}& \frac{4}{\delta\eta_1}\big(F(x_t)w - F(x_{t+1})w\big) + \frac{2\alpha_t}{\delta\eta_1}\|(V_t - \nabla F(x_t))w\|^2 \\ &+ \frac{2}{\beta\eta_1}\big(\|w_t - w\|^2 - \|w_{t+1} - w\|^2\big) + \frac{4}{\eta_1}\big(\rho + \beta\rho^2 + 3\eta_1^2\big) + 8\beta\eta_1\|V_t - \nabla F(x_t)\|^2\end{aligned}$$

□

B.3 Bias Bound via Variance-Reduced Estimation

In this section, we analyze the proposed variance-reduced multi-gradient estimate, which extends SPIDER (Fang et al., 2018) to MOO setting. We follow the notation introduced in Appendix A.

Proposition 3 (Conditional second-moment decomposition over a sub-path)**.** *Fix any $w \in \Delta_K$. Suppose that for every $\nu \in \{t-s+1, \dots, t\}$,*

$$\mathbb{E}\Big[\tilde{Q}(x_\nu; \xi_{\nu+1}) - \tilde{Q}(x_{\nu-1}; \xi_{\nu+1}) \,\Big|\, \mathcal{F}_\nu\Big] = Q(x_\nu) - Q(x_{\nu-1}).$$

Define

$$\Delta_{Q,\nu} := \tilde{Q}(x_\nu; \xi_{\nu+1}) - \tilde{Q}(x_{\nu-1}; \xi_{\nu+1}) - \big(Q(x_\nu) - Q(x_{\nu-1})\big), \qquad \nu = t-s+1, \dots, t,$$

and naturally

$$\big(V_t - Q(x_t)\big)w = \big(V_{t-s} - Q(x_{t-s})\big)w + \sum_{\nu=t-s+1}^{t} \Delta_{Q,\nu} w.$$

Then

$$\mathbb{E}\Big[\big\|\big(V_t - Q(x_t)\big)w\big\|^2 \,\Big|\, \mathcal{F}_{t-s+1}\Big] = \big\|\big(V_{t-s} - Q(x_{t-s})\big)w\big\|^2 + \sum_{\nu=t-s+1}^{t} \mathbb{E}\big[\|\Delta_{Q,\nu} w\|^2 \,\big|\, \mathcal{F}_{t-s+1}\big]. \tag{5}$$

Proof. By the assumption and the $\mathcal{F}_\nu$-measurability of $x_{\nu-1}$ and x_ν, we have

$$\mathbb{E}[\Delta_{Q,\nu} \mid \mathcal{F}_\nu] = 0, \qquad \nu = t-s+1, \dots, t.$$

Hence, by the tower property and the inclusion $\mathcal{F}_{t-s+1} \subseteq \mathcal{F}_\nu$,

$$\mathbb{E}[\Delta_{Q,\nu} \mid \mathcal{F}_{t-s+1}] = \mathbb{E}[\mathbb{E}[\Delta_{Q,\nu} \mid \mathcal{F}_\nu] \,|\, \mathcal{F}_{t-s+1}] = 0.$$

Therefore, by $\big(V_t - Q(x_t)\big)w = \big(V_{t-s} - Q(x_{t-s})\big)w + \sum_{\nu=t-s+1}^{t} \Delta_{Q,\nu} w$, we obtain

$$\begin{aligned}\mathbb{E}\Big[\big\|\big(V_t - Q(x_t)\big)w\big\|^2 \,\Big|\, \mathcal{F}_{t-s+1}\Big] &= \mathbb{E}\left[\left\|\big(V_{t-s} - Q(x_{t-s})\big)w + \sum_{\nu=t-s+1}^{t} \Delta_{Q,\nu} w\right\|^2 \,\middle|\, \mathcal{F}_{t-s+1}\right] \\ &= \big\|\big(V_{t-s} - Q(x_{t-s})\big)w\big\|^2 + \sum_{\nu=t-s+1}^{t} \mathbb{E}\big[\|\Delta_{Q,\nu} w\|^2 \,\big|\, \mathcal{F}_{t-s+1}\big] \\ &\quad + 2\sum_{\nu=t-s+1}^{t} \mathbb{E}\big[\big\langle \big(V_{t-s} - Q(x_{t-s})\big)w,\, \Delta_{Q,\nu} w\big\rangle \,\big|\, \mathcal{F}_{t-s+1}\big] \\ &\quad + 2\sum_{t-s+1 \le i < j \le t} \mathbb{E}[\langle \Delta_{Q,i} w, \Delta_{Q,j} w\rangle \,|\, \mathcal{F}_{t-s+1}].\end{aligned}$$

All the cross terms vanish, because of

1. for each $\nu = t-s+1, \dots, t$, since $\big(V_{t-s} - Q(x_{t-s})\big)w$ is $\mathcal{F}_{t-s+1}$-measurable,

$$\mathbb{E}\big[\big\langle \big(V_{t-s} - Q(x_{t-s})\big)w,\, \Delta_{Q,\nu} w\big\rangle \,\big|\, \mathcal{F}_{t-s+1}\big] = \big\langle \big(V_{t-s} - Q(x_{t-s})\big)w,\, \mathbb{E}[\Delta_{Q,\nu} w \mid \mathcal{F}_{t-s+1}]\big\rangle = 0.$$

2. for $t-s+1 \le i < j \le t$, since $\Delta_{Q,i} w$ is $\mathcal{F}_j$-measurable,

$$\mathbb{E}[\langle \Delta_{Q,i} w, \Delta_{Q,j} w\rangle \,|\, \mathcal{F}_{t-s+1}] = \mathbb{E}[\langle \Delta_{Q,i} w,\, \mathbb{E}[\Delta_{Q,j} w \mid \mathcal{F}_j]\rangle \,|\, \mathcal{F}_{t-s+1}] = 0.$$

Substituting these into the above expansion yields

$$\mathbb{E}\Big[\big\|\big(V_t - Q(x_t)\big)w\big\|^2 \,\Big|\, \mathcal{F}_{t-s+1}\Big] = \big\|\big(V_{t-s} - Q(x_{t-s})\big)w\big\|^2 + \sum_{\nu=t-s+1}^{t} \mathbb{E}\big[\|\Delta_{Q,\nu} w\|^2 \,\big|\, \mathcal{F}_{t-s+1}\big],$$

which proves equation (5). □

Lemma 3. *Suppose that*

1) for every $\nu \in [t]$,

$$\mathbb{E}\Big[\tilde{Q}(x_\nu; \xi_{\nu+1}) - \tilde{Q}(x_{\nu-1}; \xi_{\nu+1}) \mid \mathcal{F}_\nu\Big] = Q(x_\nu) - Q(x_{\nu-1});$$

2) for $\tilde{Q}(x_t; \xi_{t+1}) := \{(\tilde{q}^1(x_t; \xi_{t+1}), \dots, \tilde{q}^K(x_t; \xi_{t+1}))\}_{t=0}^{\infty}$, there exists a constant L such that

$$\mathbb{E}_\xi \|\tilde{q}^k(x;\xi) - \tilde{q}^k(y;\xi)\|^2 \le L^2 \|x-y\|^2, \qquad \forall x, y \in \mathcal{X},\ \forall k \in [K];$$

3) there exists a constant r_1 such that

$$\|x_\nu - x_{\nu-1}\| \le r_1, \qquad \forall \nu \in [t].$$

Let S_* be a mini-batch and define mini-batch stochastic estimator

$$\tilde{Q}(x_t; S_*) := \frac{1}{|S_*|} \sum_{\xi \in S_*} \tilde{Q}(x_t; \xi).$$

Consider the recursive estimator

$$V_\nu := V_{\nu-1} + \tilde{Q}(x_\nu; S_*) - \tilde{Q}(x_{\nu-1}; S_*), \qquad \nu \in [t].$$

Then for any fixed $w \in \Delta_K$,

$$\mathbb{E}[\|(V_t - Q(x_t))w\|^2 \mid \mathcal{F}_t] \le \frac{L^2 r_1^2}{|S_*|} + \|(V_{t-1} - Q(x_{t-1}))w\|^2, \tag{6}$$

$$\mathbb{E}[\|(V_t - Q(x_t))w\|^2 \mid \mathcal{F}_0] \le \frac{tL^2 r_1^2}{|S_*|} + \mathbb{E}[\|(V_0 - Q(x_0))w\|^2 \mid \mathcal{F}_0]. \tag{7}$$

Proof. First note that

$$V_t - Q(x_t) = V_{t-1} - Q(x_{t-1}) + \Big(\tilde{Q}(x_t; S_*) - \tilde{Q}(x_{t-1}; S_*) - (Q(x_t) - Q(x_{t-1}))\Big).$$

Hence, applying Prop. 3 by taking $s = 1$, we obtain

$$\begin{aligned}\mathbb{E}[\|(V_t - Q(x_t))w\|^2 \mid \mathcal{F}_t] &= \|(V_{t-1} - Q(x_{t-1}))w\|^2 \\ &\quad + \mathbb{E}\Big[\|(\tilde{Q}(x_t; S_*) - \tilde{Q}(x_{t-1}; S_*) - (Q(x_t) - Q(x_{t-1})))w\|^2 \mid \mathcal{F}_t\Big].\end{aligned}$$

For the second term,

$$\begin{aligned}&\mathbb{E}\Big[\|(\tilde{Q}(x_t; S_*) - \tilde{Q}(x_{t-1}; S_*) - (Q(x_t) - Q(x_{t-1})))w\|^2 \mid \mathcal{F}_t\Big] \\ &\le \frac{1}{|S_*|} \mathbb{E}_\xi \|(\tilde{Q}(x_t; \xi) - \tilde{Q}(x_{t-1}; \xi))w\|^2\end{aligned}$$

$$\leq \frac{1}{|S_*|}\mathbb{E}_\xi\Big[\sum_{k\in[K]}\|\tilde{q}^k(x_t;\xi)-\tilde{q}^k(x_{t-1};\xi)\|^2 w^k\Big]$$
$$\leq \frac{L^2}{|S_*|}\|x_t-x_{t-1}\|^2 \leq \frac{L^2 r_1^2}{|S_*|}.$$

Therefore,

$$\mathbb{E}[\|(V_t-Q(x_t))w\|^2 \mid \mathcal{F}_t] \leq \frac{L^2 r_1^2}{|S_*|} + \|(V_{t-1}-Q(x_{t-1}))w\|^2.$$

Iterating this inequality and using the tower property gives

$$\mathbb{E}[\|(V_t-Q(x_t))w\|^2 \mid \mathcal{F}_0] \leq \frac{tL^2 r_1^2}{|S_*|} + \mathbb{E}[\|(V_0-Q(x_0))w\|^2 \mid \mathcal{F}_0].$$

□

Lemma 4. *Set the parameters as above. Fix any $\kappa \in \mathbb{N}$, let $t_0 := \kappa q$, and let $t \in \{t_0, ..., t_0+q-1\}$. In the Algorithm 1, under the condition of Lemma 2, for any $w \in \Delta_K$ and $t < \tau$, we have*

$$\mathbb{E}\big[\|(V_t-\nabla F(x_t))w\|^2 \,\big|\, \mathcal{F}_{t_0}\big] \leq \eta_1^2,$$

and thus

$$\mathbb{E}\left[\|V_t-\nabla F(x_t)\|_{op}^2 \mid \mathcal{F}_{t_0}\right] \leq \mathbb{E}\left[\|V_t-\nabla F(x_t)\|_F^2 \mid \mathcal{F}_{t_0}\right] \leq K\eta_1^2$$

Proof. By the restart update at $t = t_0$,

$$\begin{aligned}\mathbb{E}\big[\|(V_{t_0}-\nabla F(x_{t_0}))w\|^2 \,\big|\, \mathcal{F}_{t_0}\big] &= \mathbb{E}\big[\|(\nabla F_{S_1}(x_{t_0})-\nabla F(x_{t_0}))w\|^2 \,\big|\, \mathcal{F}_{t_0}\big] \\ &\leq \mathbb{E}\left[\sum_{k\in[K]}\|\nabla f_{S_1}^k(x_{t_0})-\nabla f^k(x_{t_0})\|^2 w^k \,\middle|\, \mathcal{F}_{t_0}\right] \\ &\leq \frac{\sigma^2}{S_1} = \frac{\eta_1^2}{2}.\end{aligned}$$

Here we used $w \in \Delta_K$ and the choice $S_1 = 2\sigma^2/\eta_1^2$.

Now fix any t. By the step-size bound, with $r_1 = \frac{\eta_1}{\ell(M+1)n_0}$ we have

$$\|x_t - x_{t-1}\| \leq r_1, \forall t.$$

By taking $V_t = V_t, Q(x_t) = \nabla F(x_t), \tilde{Q}(x_t;\xi_{t+1}) = \nabla F(x_t;\xi_{t+1})$, applying Lemma 3, starting from any re-start step t_0 and consider $t = t_0, ..., t_0+q-1$, it derives

$$\begin{aligned}\mathbb{E}[\|(V_t-\nabla F(x_t))w\|^2 \mid \mathcal{F}_{t_0}] &\leq \frac{(t-t_0)\ell(M+1)^2 r_1^2}{|S_2|} + \mathbb{E}[\|(V_{t_0}-\nabla F(x_{t_0}))w\|^2 \mid \mathcal{F}_{t_0}] \\ &\leq \ell(M+1)^2 \cdot q \cdot r_1^2 \cdot \frac{1}{S_2} + \mathbb{E}[\|(V_{t_0}-\nabla F(x_{t_0}))w\|^2 \mid \mathcal{F}_{t_0}] \\ &\leq \ell(M+1)^2 \cdot \frac{\sigma n_0}{\eta_1} \cdot \frac{\eta_1^2}{\ell(M+1)^2 n_0^2} \cdot \frac{\eta_1 n_0}{2\sigma} + \eta_1^2/2 \\ &= \eta_1^2.\end{aligned}$$

Thus, we have

$$\|V_t-\nabla F(x_t)\|_{op}^2 \leq \|V_t-\nabla F(x_t)\|_F^2 = \sum_{k\in[K]}\|(V_t-\nabla F(x_t))e_k\|^2,$$

where the first inequality holds in general, and e_k is the standard basic vector with k^{th} ordinate equal to 1, o.w. 0. We finish the proof by taking conditional expectation, which preserves the property. □

C Proof of the Final Convergence Guarantee

In this section, we complete the proof of Theorem 1 under the parameter choices and settings specified at the beginning of the Section B. The proof consists of three parts: deriving the sample-complexity bound, controlling the stopping-time event, and obtaining the final high-probability convergence result. Recall the Theorem 1:

Theorem 1. *Let Assumptions 1, 2, and 3 hold. Given $0 < \theta, \epsilon < 1$, set hyperparameters $n_0 = \mathcal{O}(1)$, $\beta = \mathcal{O}(\epsilon^{-2})$, $\rho = \mathcal{O}(\epsilon^2)$, $\delta = \mathcal{O}(1)$, and $\eta_1 = \Theta(\epsilon)$. Choose batch sizes and block size $S_1 = \mathcal{O}(\epsilon^{-2})$, $S_2 = \mathcal{O}(\epsilon^{-1})$, and $q = \mathcal{O}(\epsilon^{-1})$. Set the clipping learning rates as $\gamma_t = \min\{\frac{\eta_1}{\|V_t w_t\|}, \frac{1}{2}\}$ and $\alpha_t = \delta\gamma_t = \min\{\frac{\eta_1}{n_0\ell(M+1)\|V_t w_t\|}, \frac{1}{2n_0\ell(M+1)}\}$. Then, with stochastic gradient complexity $\mathcal{O}(\epsilon^{-3})$, the iterates generated by Algorithm 1 satisfy, with probability at least $1-\theta$,*

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F(x_t)w_t\| \le \epsilon.$$

C.1 Sample Complexity

From the equation (4), we have the basic inequality

$$\begin{aligned}\|V_t w_t\| \le \frac{4}{\delta\eta_1}\big(F(x_t)w - F(x_{t+1})w\big) + \frac{2\alpha_t}{\delta\eta_1}\|(V_t - \nabla F(x_t))w\|^2 \\ + \frac{2}{\beta\eta_1}\big(\|w_t - w\|^2 - \|w_{t+1} - w\|^2\big) + \frac{4}{\eta_1}\big(\rho + \beta\rho^2\big) + 8\beta\eta_1\|V_t - \nabla F(x_t)\|^2 + 12\eta_1.\end{aligned} \tag{8}$$

Therefore, taking telescopic sum for $t = 0, ...\tau - 1$, where $\tau \le T$, we obtain

$$\begin{aligned}\sum_{t=0}^{\tau-1}\|V_t w_t\| &\le (F(x_0) - F(x_\tau))w \cdot \frac{4}{\delta\eta_1} + \sum_{t=0}^{\tau-1}\frac{2\alpha_t}{\delta\eta_1}\|(V_t - \nabla F(x_t))w\|^2 \\ &\quad + (\frac{1}{\beta} + \rho\tau + \beta\rho^2\tau)\cdot\frac{4}{\eta_1} + \sum_{t=0}^{\tau-1}8\beta\eta_1\|V_t - \nabla F(x_t)\|^2 + 12\eta_1\tau \\ &\le (F(x_0) - F(x_\tau))w \cdot \frac{4}{\delta\eta_1} + \sum_{t=0}^{\tau-1}\frac{1}{\eta_1}\|(V_t - \nabla F(x_t))w\|^2 \\ &\quad + (\frac{1}{\beta} + \eta_1^2\tau + \eta_1^2\tau)\cdot\frac{4}{\eta_1} + \sum_{t=0}^{\tau-1}8\beta\eta_1\|V_t - \nabla F(x_t)\|^2 + 12\eta_1\tau,\end{aligned}$$

where the last inequality follows from $\alpha_t \le \frac{1}{2n_0\ell(M+1)} = \delta/2$ and the choice of ρ.

Multiplying both sides by $1/T$, we obtain

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{\tau-1}\|V_t w_t\| &\le (F(x_0) - F(x_\tau))w \cdot \frac{4}{\delta\eta_1 T} + \frac{1}{\eta_1 T}\sum_{t=0}^{\tau-1}\|(V_t - \nabla F(x_t))w\|^2 \\ &\quad + (\frac{1}{\beta T} + \eta_1^2 + \eta_1^2)\cdot\frac{4}{\eta_1} + \frac{1}{T}\sum_{t=0}^{\tau-1}8\beta\eta_1\|V_t - \nabla F(x_t)\|^2 + 12\eta_1 \\ &\le \eta_1 + \frac{1}{\eta_1 T}\sum_{t=0}^{\tau-1}\|(V_t - \nabla F(x_t))w\|^2 \\ &\quad + (\eta_1 + 4\eta_1 + 4\eta_1) + \frac{1}{T}\sum_{t=0}^{\tau-1}8\beta\eta_1\|V_t - \nabla F(x_t)\|^2 + 12\eta_1 \\ &= \frac{1}{\eta_1 T}\sum_{t=0}^{\tau-1}\|(V_t - \nabla F(x_t))w\|^2 + \frac{1}{T}\sum_{t=0}^{\tau-1}\beta\|V_t - \nabla F(x_t)\|^2 \cdot 8\eta_1 + 22\eta_1,\end{aligned}$$

where the second inequality uses the choice $\frac{1}{T} \leq \frac{\delta\eta_1^2}{4(F(x_0)-F(x_\tau))w}$, $\frac{1}{T} \leq \frac{\beta\eta_1^2}{4}$.

Taking total expectation on both sides and applying Lemma 4 on each block starting from κq, we obtain

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{\tau-1}\mathbb{E}\|V_t w_t\| &\leq \frac{1}{\eta_1 T}\sum_{t=0}^{\tau-1}\mathbb{E}\|(V_t-\nabla F(x_t))w\|^2 + \frac{1}{T}\sum_{t=0}^{\tau-1}\beta\mathbb{E}\|V_t-\nabla F(x_t)\|^2\cdot 8\eta_1 + 22\eta_1 \\
&= \frac{1}{\eta_1 T}\sum_{t=0}^{\tau-1}\mathbb{E}\left[\mathbb{E}\Big[\|(V_t-\nabla F(x_t))w\|^2 \mid \mathcal{F}_{\lfloor t/q\rfloor q}\Big]\right] \\
&\quad + \frac{1}{T}\sum_{t=0}^{\tau-1}\beta\mathbb{E}\left[\mathbb{E}\Big[\|V_t-\nabla F(x_t)\|^2 \mid \mathcal{F}_{\lfloor t/q\rfloor q}\Big]\right]\cdot 8\eta_1 + 22\eta_1 \\
&\leq \frac{\tau}{\eta_1 T}\cdot\eta_1^2 + \frac{\tau}{T}\beta\cdot K\eta_1^2\cdot 8\eta_1 + 22\eta_1 \\
&\leq \eta_1 + 8\eta_1 + 22\eta_1 \leq J_1\eta_1, \qquad (9)
\end{aligned}$$

for some $J_1 \in \{J_1 \in O(1) \mid 31 \leq J_1\}$. Finally, taking $T = \lfloor(\frac{4F_{init}}{\delta}+\frac{4}{\beta})\eta_1^{-2}\rfloor + 1$, the total number of oracle accesses can be upper bounded by using $\tau \leq T$ and the property $S_1 = qS_2$, that is

$$\begin{aligned}
\lceil T\cdot\frac{1}{q}\rceil S_1 + TS_2 &\leq 2TS_2 + S_1 \\
&\leq 2\Big((\frac{4F_{init}}{\delta}+\frac{4}{\beta})\eta_1^{-2}+1\Big)S_2 + S_1 \\
&\leq 2\Big((\frac{4F_{init}}{\delta}+\frac{4}{\beta})\eta_1^{-2}+1\Big)\frac{2\sigma}{n_0\eta_1} + \frac{2\sigma^2}{\eta_1^2} \\
&\leq \Big(8F_{init}\cdot n_0\ell(M+1)\cdot\eta_1^{-2} + 8\cdot K(M^2+\eta_1^2)\cdot\eta_1^{-2}+1\Big)\frac{2\sigma}{n_0\eta_1} + \frac{2\sigma^2}{\eta_1^2} \\
&= O(\eta_1^{-3}) = O(\epsilon^{-3}).
\end{aligned}$$

We remain the final iteration complexity result to the next section.

C.2 Stopping Time and Convergence

Recall that, for some constant $F > 0$, we define the stopping time as:

$$\begin{aligned}
\tau_1 &:= \min\{t \mid \exists k \in [K], f^k(x_{t+1}) - f^{k,*} > F\} \wedge T, \\
\tau_2 &:= \min\{t \mid \exists k \in [K], \|v_t^k - \nabla f^k(x_t)\| \geq H_1\} \wedge T, \\
\tau &:= \min\{\tau_1, \tau_2\}.
\end{aligned}$$

In this section, for a small probability $\theta \in (0,1)$, we prove the following two claims one by one:

- **Claim 1** : event $\{\tau < T\}$ is a tail event.
- **Claim 2** : event $\{\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F(x_t)w_t\| \leq \epsilon\}$ happens with high probability $1-\theta$.

C.2.1 Claim 1 : The Tail Event

We claim the event $\{\tau < T\}$ is with small probability. To begin with, taking the event decomposition gives us two events to consider:
$\{\tau < T\} = \{\tau = \tau_2, \tau < T\} \sqcup \{\tau = \tau_1, \tau \neq \tau_2, \tau < T\} = \{\tau = \tau_2, \tau < T\} \sqcup \{\tau = \tau_1, \tau < \tau_2, \tau < T\}$..

We first consider the likelihood of the first event. Based on union bound, taking $H_1^2 = \frac{8KT\eta_1^2}{\theta} = O(1)$:

$$\mathbb{P}(\{\tau = \tau_2 < T\}) \leq \sum_{k=1}^{K}\sum_{t=0}^{T-1}\mathbb{P}\Big(\|v_t^k - \nabla f^k(x_t)\| \geq H_1\Big) \leq KT\cdot\frac{\eta_1^2}{H_1^2} = \frac{\theta}{8},$$

where in the last inequality, we employ Markov inequality as $\mathbb{P}(X \geq a) = \mathbb{P}(X^2 \geq a^2) \leq \mathbb{E}[X^2]/a^2$ for $X \geq 0, a > 0$, and total expectation of Lemma 4 with $w = e^k$.

We now turn to the second event. Following the standard decomposition trick, we obtain

$$\{f^k(x_{t+1}) - f^{k,*} > F\} \subseteq \{f^k(x_{t+1}) - f^k(x_t) > F/2\} \cup \{f^k(x_t) - f^{k,*} > F/2\},$$

so the two sub-cases we need to consider for the event $\{\tau = \tau_1, \tau < \tau_2, \tau < T\}$ are

$$\mathcal{E}_1 := \Big\{\exists k \in [K] \mid f^k(x_{\tau+1}) - f^k(x_\tau) > F/2, \text{ and } \|v_t^j - \nabla f^j(x_t)\| < H_1, \forall j \in [K], t \in [\tau]\Big\},$$
$$\mathcal{E}_2 := \Big\{\exists k \in [K] \mid f^k(x_\tau) - f^{k,*} > F/2, \text{ and } \|v_t^j - \nabla f^j(x_t)\| < H_1, \forall j \in [K], t \in [\tau]\Big\}.$$

The first sub-case has probability 0. On the event $\{\tau = \tau_1, \tau < \tau_2, \tau < T\}$, the estimator error is controlled up to time τ, and the clipped update keeps $x_{\tau+1}$ inside the local neighborhood of x_τ. Hence, applying equation (3) at $t = \tau$ and dropping all non-positive terms gives, for any $w \in \Delta_K$,

$$\begin{aligned} F(x_{\tau+1})w - F(x_\tau)w &\le -\frac{\eta_1\|V_\tau w_\tau\|}{4n_0\ell(M+1)} + \frac{\alpha_\tau}{2}\|(V_\tau - \nabla F(x_\tau))w\|^2 + \frac{\delta}{2\beta}(\|w_\tau - w\|^2 - \|w_{\tau+1} - w\|^2) \\ &\quad + \delta\rho + \delta\beta\rho^2 + 2\delta\beta\eta_1^2\|V_\tau - \nabla F(x_\tau)\|^2 + (1+2)\delta\eta_1^2 \\ &\le \frac{\alpha_\tau}{2}\|(V_\tau - \nabla F(x_\tau))w\|^2 + \frac{\delta}{\beta} + \delta\rho + \delta\beta\rho^2 + 2\delta\beta\eta_1^2\|V_\tau - \nabla F(x_\tau)\|^2 + (2+1)\delta\eta_1^2. \end{aligned}$$

Hence, taking stand basis vectors $w = e^k$ for every single $k \in [K]$ generates

$$\begin{aligned} f^k(x_{\tau+1}) - f^k(x_\tau) &= F(x_{\tau+1})e^k - F(x_\tau)e^k \\ &\le \frac{\alpha_\tau}{2}\|(V_\tau - \nabla F(x_\tau))e^k\|^2 + \frac{\delta}{\beta} + \delta\rho + \delta\beta\rho^2 + 2\delta\beta\eta_1^2\|V_\tau - \nabla F(x_\tau)\|^2 + (2+1)\delta\eta_1^2 \\ &\le \underbrace{\frac{\delta}{4}\cdot H_1^2 + \frac{\delta}{\beta} + \delta\rho + \delta\beta\rho^2 + 2\delta\beta\eta_1^2 \cdot KH_1^2 + (2+1)\delta\eta_1}_{=:B_3 \in O(1+1+\eta_1^2+\eta_1^2+1+\eta_1)=O(1)} \\ &\le F/2, \end{aligned} \tag{10}$$

where in the second inequality, we utilize $\|V_\tau - \nabla F(x_\tau)\|^2 \le K\|v_\tau - \nabla f(x_\tau)\|^2 \le KH_1^2$ given by stopping time, and also set F satisfying $B_3 \le F/2$. It does imply the first sub-case should have probability zero, that is $\mathbb{P}(\mathcal{E}_1) = 0$.

For the second sub-case, we simply consider the cover of $\mathcal{E}_2$:

$$\mathcal{E}_2' := \{\exists k \in [K] \mid f^k(x_\tau) - f^{k,*} \ge F/2\} \supseteq \mathcal{E}_2.$$

Starting with the summation of equation (3), we obtain

$$\begin{aligned} F(x_\tau)w - F^*w \le F(x_0)w - F^*w &- \sum_{t=0}^{\tau-1}\frac{\eta_1\|V_t w_t\|}{4n_0\ell(M+1)} + \sum_{t=0}^{\tau-1}\frac{\alpha_t}{2}\|(V_t - \nabla F(x_t))w\|^2 \\ &+ \frac{\delta}{\beta} + \delta\rho\tau + \delta\beta\rho^2\tau + \sum_{t=0}^{\tau-1} 2\delta\beta\eta_1^2\|V_t - \nabla F(x_t)\|^2 + 3\delta\eta_1^2\tau. \end{aligned}$$

Taking total expectation, skipping the non-positive term, and applying Lemma 4, we derive

$$\begin{aligned} \mathbb{E}\Big[F(x_\tau)w - F^*w\Big] &\le F(x_0)w - F^*w - \mathbb{E}\Big[\sum_{t=0}^{\tau-1}\frac{\eta_1\|V_t w_t\|}{4n_0\ell(M+1)}\Big] + \sum_{t=0}^{\tau-1}\frac{\alpha_t}{2}\mathbb{E}\|(V_t - \nabla F(x_t))w\|^2 \\ &\quad + \frac{\delta}{\beta} + \delta\rho\tau + \delta\beta\rho^2\tau + \sum_{t=0}^{\tau-1} 2\delta\beta\eta_1^2\mathbb{E}\|V_t - \nabla F(x_t)\|^2 + 3\delta\eta_1^2\tau \\ &\le \underbrace{F_{init} + \frac{T\delta\eta_1^2}{4} + \frac{\delta}{\beta} + \delta\rho T + \delta\beta\rho^2 T + 2T\delta\beta\eta_1^2 \cdot K\eta_1^2 + 3\delta\eta_1^2 T}_{=:\tilde{B}_3 \in O(1+1+1+1+1+1+1)=O(1)} \\ &\quad - \mathbb{E}\Big[\sum_{t=0}^{\tau-1}\frac{\eta_1\|V_t w_t\|}{4n_0\ell(M+1)}\Big] \end{aligned}$$

$$\leq \theta F/(8K) - \mathbb{E}\Big[\sum_{t=0}^{\tau-1} \frac{\eta_1 \|V_t w_t\|}{4n_0 \ell(M+1)}\Big] \leq \theta F/(8K),$$

which holds for any $w \in \Delta_K$ because of convex combination, and we also further restrict our F to satisfy $\tilde{B}_3 \leq \theta F/(8K)$. Thus, for every single objective k, we

$$\mathbb{P}\Big(f^k(x_\tau) - f^{k,*} \geq F/2\Big) = \mathbb{P}\Big(F(x_\tau)e^k - F^* e^k \geq F/2\Big) \leq \frac{\mathbb{E}[F(x_\tau)w - F^* w]}{F/2} \leq \theta/(4K),$$

where e^k is the stand basis vector of k. Thus, by consider the basic property of measure $\mathbb{P}(\cup_{k\in[K]} A_k) \leq \sum_{k\in[K]} \mathbb{P}(A_k)$, we obtain

$$\mathbb{P}(\mathcal{E}_2) \leq \mathbb{P}(\mathcal{E}_2') = \mathbb{P}\Big(\exists k \in [K] \mid f^k(x_\tau) - f^{k,*} \geq F/2\Big) \leq \sum_{k\in[K]} \mathbb{P}\Big(f^k(x_\tau) - f^{k,*} \geq F/2\Big) \leq \theta/4.$$

Therefore, combining both cases above, it's straightforward to achieve

$$\begin{aligned}
\mathbb{P}(\tau < T) &= \mathbb{P}(\tau = \tau_2, \tau < T) + \mathbb{P}(\tau = \tau_1, \tau \neq \tau_2, \tau < T) \\
&\leq \mathbb{P}(\tau = \tau_2, \tau < T) + \Big(\mathbb{P}(\mathcal{E}_1) + \mathbb{P}(\mathcal{E}_2)\Big) \\
&\leq \theta/8 + 0 + \theta/4 \leq \theta/2,
\end{aligned}$$

which concludes that $\{\tau < T\}$ is a tail event by taking a sufficiently small θ as probability.

C.2.2 Claim 2 : The High probability of Convergence

In this section, we will prove our claim 2, as well the convergence result holds with high probability. For brevity, let's define

$$A := \left\{\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla F(x_t) w_t\| \leq \epsilon\right\}, \qquad B := \{\tau = T\}.$$

and from basic probability we have

$$\mathbb{P}(A) = 1 - \mathbb{P}(A^c) = 1 - \mathbb{P}(A^c \cap B) - \mathbb{P}(A^c \cap B^c) \geq 1 - \mathbb{P}(A^c \cap B) - \mathbb{P}(B^c).$$

By the stopping-time argument established previously,

$$\mathbb{P}(B^c) = \mathbb{P}(\tau < T) \leq \frac{\theta}{2}.$$

It remains to upper bound $\mathbb{P}(A^c \cap B)$. On the event $B = \{\tau = T\}$, we notice $\sum_{t=0}^{T-1} \|\nabla F(x_t) w_t\| = \sum_{t=0}^{\tau-1} \|\nabla F(x_t) w_t\|$ and the triangle inequality gives us $\frac{1}{T}\sum_{t=0}^{\tau-1} \|\nabla F(x_t) w_t\| \leq \frac{1}{T}\sum_{t=0}^{\tau-1} \|V_t w_t\| + \frac{1}{T}\sum_{t=0}^{\tau-1} \|(V_t - \nabla F(x_t)) w_t\|$, and then we obtain

$$\begin{aligned}
A^c \cap B &= \left\{\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla F(x_t) w_t\| > \epsilon\right\} \cap B \\
&\subseteq \left\{\frac{1}{T}\sum_{t=0}^{\tau-1} \|\nabla F(x_t) w_t\| > \epsilon\right\} \\
&\subseteq \left\{\frac{1}{T}\sum_{t=0}^{\tau-1} \|V_t w_t\| + \frac{1}{T}\sum_{t=0}^{\tau-1} \|(V_t - \nabla F(x_t)) w_t\| > \epsilon\right\} \\
&\subseteq \left\{\frac{1}{T}\sum_{t=0}^{\tau-1} \|V_t w_t\| \geq \epsilon/2\right\} \cup \left\{\frac{1}{T}\sum_{t=0}^{\tau-1} \|(V_t - \nabla F(x_t)) w_t\| > \epsilon/2\right\},
\end{aligned}$$

where in the last decomposition, we employ the standard threshold decomposition that $X + Y > \epsilon$ implies $X > \epsilon/2$ or $Y > \epsilon/2$. Therefore, it holds that

$$\mathbb{P}(A^c \cap B) \leq \mathbb{P}\left(\frac{1}{T}\sum_{t=0}^{\tau-1} \|V_t w_t\| > \frac{\epsilon}{2}\right) + \mathbb{P}\left(\frac{1}{T}\sum_{t=0}^{\tau-1} \|(V_t - \nabla F(x_t)) w_t\| > \frac{\epsilon}{2}\right).$$

To bound the first term, by Markov's inequality and the previously established bound in equation (9),

$$\mathbb{P}\left(\frac{1}{T}\sum_{t=0}^{\tau-1}\|V_t w_t\| > \frac{\epsilon}{2}\right) \le \frac{2}{\epsilon}\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{\tau-1}\|V_t w_t\|\right] \le \frac{2J_1\eta_1}{\epsilon}.$$

For the second term, again by Markov's inequality,

$$\mathbb{P}\left(\frac{1}{T}\sum_{t=0}^{\tau-1}\|(V_t - \nabla F(x_t))w_t\| > \frac{\epsilon}{2}\right) \le \frac{2}{\epsilon}\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{\tau-1}\|(V_t - \nabla F(x_t))w_t\|\right].$$

Notice $\tau \le T$, and $\|w_t\|_2 \le \|w_t\|_1 = 1$, $\forall\, w_t \in \Delta_K$, following Lemma 4, one should obtain

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{\tau-1}\|(V_t - \nabla F(x_t))w_t\|\right] &= \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\mathbf{1}_{\{t<\tau\}}\cdot\|(V_t - \nabla F(x_t))w_t\|\right] \\
&\le \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|(V_t - \nabla F(x_t))w_t\| \\
&\le \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|V_t - \nabla F(x_t)\|_{\mathrm{op}} \\
&\le \frac{1}{T}\sum_{t=0}^{T-1}\sqrt{\mathbb{E}\|V_t - \nabla F(x_t)\|_{\mathrm{op}}^2} \le \sqrt{K}\eta_1,
\end{aligned}$$

which directly implies

$$\mathbb{P}\left(\frac{1}{T}\sum_{t=0}^{\tau-1}\|(V_t - \nabla F(x_t))w_t\| > \frac{\epsilon}{2}\right) \le \frac{2\sqrt{K}\,\eta_1}{\epsilon}.$$

By taking η_1, J_1 as shown, combining the above estimates yields

$$\mathbb{P}(A^c \cap B) \le \frac{2J_1\eta_1}{\epsilon} + \frac{2\sqrt{K}\eta_1}{\epsilon} = \frac{2(J_1+\sqrt{K})\eta_1}{\epsilon} \le \theta/2,$$

and thus

$$\mathbb{P}(A) \ge 1 - \mathbb{P}(A^c \cap B) - \mathbb{P}(B^c) \ge 1 - \frac{\theta}{2} - \frac{\theta}{2} = 1 - \theta.$$

Consequently,

$$\mathbb{P}\left(\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F(x_t)w_t\| \le \epsilon\right) \ge 1 - \theta,$$

which completes the proof.

D Auxiliary Results

D.1 Proof of Proposition 2

Proof. For the first item, we take the non-decreasing of g as well $\ell(u) := g(u+a)$ for fixed non-negative constant a and any $u \ge 0$, which directly gives us

$$\mathbb{E}_\xi\|\nabla f^k(x;\xi) - \nabla f^k(y;\xi)\|^2 \le g(\|\nabla f^k(x)\| + a)^2\|x-y\|^2 = \ell(\|\nabla f^k(x)\|)^2\|x-y\|^2.$$

The second inequality can be derived by Jensen's inequality,

$$\begin{aligned}
\|\nabla f^k(x) - \nabla f^k(y)\|^2 &= \|\mathbb{E}_\xi(\nabla f^k(x;\xi) - \nabla f^k(y;\xi))\|^2 \\
&\le \mathbb{E}_\xi\|\nabla f^k(x;\xi) - \nabla f^k(y;\xi)\|^2 \\
&\le g(\|\nabla f^k(x)\|)^2\|x-y\|^2
\end{aligned}$$

$$\leq \ell(\|\nabla f^k(x)\|)^2\|x-y\|^2.$$

For the second item, fix any $x \in \mathcal{X}$ and denote $u := \|\nabla f^k(x)\|$. By Assumption 3, we already have

$$B\big(x,\ r\big(\|\nabla f^k(x)\|\big)\big) \subseteq \mathcal{X}.$$

It remains to verify the local gradient-difference bound. We first show that the gradient norm is controlled inside a small neighborhood. Let $x_1, x_2 \in B\big(x,\ r\big(\|\nabla f^k(x)\|\big)\big)$, thus by the first item,

$$\|\nabla f^k(x_1) - \nabla f^k(x)\| \leq g\big(\|\nabla f^k(x)\|\big)\|x_1 - x\| \leq \ell\big(\|\nabla f^k(x)\|\big)\, r\big(\|\nabla f^k(x)\|\big) = a,$$

and hence

$$\|\nabla f^k(x_1)\| \leq \|\nabla f^k(x)\| + \|\nabla f^k(x_1) - \nabla f^k(x)\| \leq \|\nabla f^k(x)\| + a = u + a.$$

Therefore, using again the first item and the monotonicity of g, we obtain

$$\begin{aligned}\|\nabla f^k(x_1) - \nabla f^k(x_2)\| &\leq g\big(\|\nabla f^k(x_1)\|\big)\|x_1 - x_2\| \\ &\leq g(u+a)\|x_1 - x_2\| \\ &= \ell(u)\|x_1 - x_2\| \\ &= \ell\big(\|\nabla f^k(x)\|\big)\|x_1 - x_2\|.\end{aligned}$$

Thus, f^k is (r, ℓ)-smooth. The ℓ-smoothness then follows from Proposition 3.2 in Li et al. (2023). □

E Experiment Detail

In this section, we provide ablation studies and experimental details for the experiments in the main text.

E.1 Ablation Study

Our theoretical analysis of the clipping mechanism shows that it significantly improves training stability in complicated settings while maintaining comparable training speed and final performance, as further demonstrated by the following experiment.

E.1.1 Step Size

For the ablation study, we test the following four SPIDER-MGDA variants with different step-size strategies (clipped/fixed), as summarized in Table 2.

Table 2: SPIDER-MGDA variants used in the ablation study.

Models	α_t	γ_t
Double clip (ClipVR-MGDA)	clip	clip
Primal clip only	clip	fixed
Dual clip only	fixed	clip
No clip	fixed	fixed

On the *Multi-MNIST* dataset, the double-clip variant exhibits the lowest cross-seed variability throughout training and the smoothest temporal training trajectories. It is substantially more stable than the no-clip and dual-clip-only variants, and slightly more stable than the primal-clip-only variant, while maintaining comparable final accuracy and loss. The detailed results are reported in Table 3. The first result column reports final performance, while the second and third quantify cross-seed and temporal stability, respectively; lower values indicate greater stability.

Table 3: Ablation results on Multi-MNIST. Lower values indicate greater stability for the cross-seed and temporal stability metrics.

Variant	Acc. %/Loss	Cross-seed stability Avg. trajectory SD (Acc. pp/Loss)	Temporal stability mean $\lvert\Delta^2\rvert$ (Acc. pp/Loss)
Double-clip	89.869 ± 1.176 / 0.3135 ± 0.0406	$\mathbf{1.263}/\mathbf{0.0351}$	$\mathbf{2.967 \pm 0.244}$ / $\mathbf{0.0797 \pm 0.0088}$
Primal-only	89.822 ± 0.481 / 0.3128 ± 0.0163	$1.329/0.0363$	3.018 ± 0.491 / 0.0820 ± 0.0151
Dual-only	89.201 ± 0.825 / 0.3320 ± 0.0241	$3.689/0.1066$	6.020 ± 1.919 / 0.1776 ± 0.0656
No-clip	89.849 ± 0.443 / 0.3119 ± 0.0162	$4.027/0.1995$	7.604 ± 3.488 / 0.3772 ± 0.3594

E.1.2 CLIP

Moreover, we conduct the same test on the Office-31 dataset with different fixed clip terms c, denoting $\gamma_t = \min\{\eta/\|d_t\|, c\}$. The results under different choices of c are reported in Table 4. As c increases, the accuracy and loss performance of the dual-only and no-clip variants deteriorate substantially, while the double-clip variant remains stable. The other metrics further confirm that double clipping improves both cross-seed stability and temporal stability throughout training. Note that the identical Double-clip results across different values of c follow directly from the algorithm, since these values yield the same effective updates. Moreover, it indicates that the double-clip method is less sensitive to the choice of c than the other variants, providing a wider range of practically effective hyperparameter values.

Table 4: Ablation results on Office-31 under different choices of c. Lower values indicate greater stability for the cross-seed and temporal stability metrics.

c	Variant	Acc. %/Loss	Cross-seed stability Avg. trajectory SD (Acc. pp/Loss)	Temporal stability mean $\lvert\Delta^2\rvert$ (Acc. pp/Loss)
0.10	Double-clip	$\mathbf{89.09 \pm 0.49}$ / 0.745 ± 0.005	$\mathbf{1.034}/\mathbf{0.0143}$	1.280 ± 0.301 / 0.0202 ± 0.0034
0.10	Primal-only	87.91 ± 0.69 / 0.816 ± 0.007	$1.104/0.0203$	$\mathbf{1.227 \pm 0.280}$ / $\mathbf{0.0176 \pm 0.0032}$
0.10	Dual-only	86.00 ± 1.39 / $\mathbf{0.666 \pm 0.034}$	$1.527/0.0395$	3.186 ± 0.388 / 0.0870 ± 0.0216
0.10	No-clip	83.21 ± 1.92 / 0.789 ± 0.077	$2.925/0.1019$	2.914 ± 0.987 / 0.0855 ± 0.0269
0.14	Double-clip	$\mathbf{89.09 \pm 0.49}$ / $\mathbf{0.745 \pm 0.005}$	$\mathbf{1.034}/\mathbf{0.0143}$	1.280 ± 0.301 / 0.0202 ± 0.0034
0.14	Primal-only	87.76 ± 0.46 / 0.825 ± 0.006	$1.048/0.0145$	$\mathbf{1.208 \pm 0.165}$ / $\mathbf{0.0177 \pm 0.0030}$
0.14	Dual-only	83.39 ± 1.73 / 0.751 ± 0.068	$2.892/0.0836$	4.942 ± 0.812 / 0.1406 ± 0.0245
0.14	No-clip	75.84 ± 3.41 / 1.018 ± 0.108	$4.126/0.1181$	6.071 ± 1.582 / 0.1468 ± 0.0263
0.20	Double-clip	$\mathbf{89.09 \pm 0.49}$ / $\mathbf{0.745 \pm 0.005}$	$1.034/\mathbf{0.0143}$	1.280 ± 0.301 / 0.0202 ± 0.0034
0.20	Primal-only	87.68 ± 0.52 / 0.829 ± 0.007	$\mathbf{0.968}/0.0145$	$\mathbf{1.206 \pm 0.306}$ / $\mathbf{0.0175 \pm 0.0030}$
0.20	Dual-only	72.42 ± 6.54 / 1.093 ± 0.201	$6.873/0.2327$	8.046 ± 2.472 / 0.2327 ± 0.0906
0.20	No-clip	64.65 ± 1.59 / 1.352 ± 0.058	$4.383/0.1502$	6.981 ± 2.928 / 0.2480 ± 0.1032
0.30	Double-clip	$\mathbf{89.09 \pm 0.49}$ / $\mathbf{0.745 \pm 0.005}$	$1.034/\mathbf{0.0143}$	$\mathbf{1.280 \pm 0.301}$ / 0.0202 ± 0.0034
0.30	Primal-only	87.44 ± 0.44 / 0.848 ± 0.008	$\mathbf{1.009}/0.0161$	1.284 ± 0.223 / $\mathbf{0.0180 \pm 0.0030}$
0.30	Dual-only	49.59 ± 6.46 / 1.904 ± 0.235	$6.338/0.3803$	9.928 ± 1.962 / 0.4898 ± 0.1905
0.30	No-clip	39.91 ± 6.81 / 2.279 ± 0.276	$3.231/0.2242$	6.647 ± 3.168 / 0.4403 ± 0.1942
0.60	Double-clip	$\mathbf{89.09 \pm 0.49}$ / $\mathbf{0.745 \pm 0.005}$	$\mathbf{1.034}/\mathbf{0.0143}$	1.280 ± 0.301 / 0.0202 ± 0.0034
0.60	Primal-only	87.39 ± 0.86 / 0.848 ± 0.010	$1.185/0.0161$	$\mathbf{1.216 \pm 0.153}$ / $\mathbf{0.0171 \pm 0.0031}$
0.60	Dual-only	25.88 ± 4.01 / 3.153 ± 0.398	$2.386/0.3147$	5.143 ± 1.293 / 0.6484 ± 0.3653
0.60	No-clip	26.63 ± 4.74 / 3.009 ± 0.232	$2.841/0.1904$	4.623 ± 1.726 / 0.4952 ± 0.2090

E.1.3 OTHER HYPER-PARAMETERS

We conduct sensitivity tests on different datasets. Again, we simplify the clipping strategy as

$$\alpha_t = \delta\gamma_t, \qquad \gamma_t = \min\{\eta/\|d_t\|, c\}.$$

Our method shows good robustness to the choices of δ and η on the Multi-MNIST dataset. We first vary δ while fixing η at its default value, and then vary η while fixing δ at its default value. The corresponding sensitivity results are reported in Table 5. As shown, the performance remains nearly unchanged over a broad range around the default values, while only overly small choices lead to

Table 5: Sensitivity of ClipVR-MGDA to δ and η on Multi-MNIST.

δ	Final Acc. (%)	η	Final Acc. (%)
0.25	89.84 ± 0.24	0.00375	90.41 ± 0.34
0.50	91.47 ± 0.53	0.00750	91.46 ± 0.28
1.00	$\mathbf{91.67 \pm 0.36}$	0.01500	91.67 ± 0.36
2.00	91.65 ± 0.32	0.03000	91.49 ± 0.93
4.00	91.46 ± 0.21	0.06000	$\mathbf{91.91 \pm 0.77}$

noticeable degradation. This indicates that δ and η meaningfully regulate the optimization dynamics but do not require precise tuning.

E.1.4 Ablation Configuration.

The detailed experimental settings for Multi-MNIST and Office-31 are summarized in Tables 6 and 7, respectively.

Table 6: Experimental details for Multi-MNIST.

Item	Setting
Dataset	Multi-MNIST (64,000 training and 20,000 test samples)
Model	HPS with a shared CNN encoder and two task-specific linear heads
Objectives	Left- and right-digit classification
Random seeds	5 independent runs
Metric aggregation	Uniform average of the two objective-level test metrics
Reported results	Mean $\pm$ SD across seeds
Training budget	50 epochs (25,000 updates and 4,600,448 gradient queries per run)
Ablation settings	Double-clip, primal-only, dual-only, and no-clip; $c = 0.015$
Cross-seed stability	Trajectory average of the pointwise SD across seeds
Temporal stability	mean $\lvert\Delta^2\rvert$ along each trajectory, reported as mean $\pm$ SD across seeds
Training	Train/test batch sizes $128/256$ and SGD learning rate 0.015
Controls	Same model, batch sizes, training budget, and non-varied hyperparameters

Table 7: Experimental details for Office-31.

Item	Setting
Dataset	Office-31: Amazon, DSLR, and Webcam
Model	HPS with a shared ImageNet-pretrained ResNet-18 and domain-specific linear heads
Random seeds	5 independent runs
Metric aggregation	Uniform average of the three domain-level test metrics
Reported results	Mean $\pm$ SD across seeds
Training budget	50 epochs (1,300 updates and 156,000 gradient queries per run)
Ablation settings	Double-clip, primal-only, dual-only, and no-clip; $c \in \{0.10, 0.14, 0.20, 0.30, 0.60\}$
Cross-seed stability	Trajectory average of the pointwise SD across seeds
Temporal stability	mean $\lvert\Delta^2\rvert$ along each trajectory, reported as mean $\pm$ SD across seeds
Training	Batch size 64 and task-head Adam learning rate 10^{-4}

E.2 Experiment on Multi-MNIST

We evaluate the proposed *ClipVR-MGDA* method (Algorithm 1) and the baselines on the Multi-MNIST dataset (Sun, 2019). The main experimental protocol has been described above; the re-

maining shared settings and algorithm-specific hyperparameters are summarized in Tables 8, 9 and 10.

Table 8: Shared experimental settings for the query-efficiency comparison.

Setting	Value
Dataset	Generated Multi-MNIST
Tasks	Left-digit and right-digit classification
Architecture	HPS hard-parameter-sharing network
Optimizer	SGD
Momentum	0
Weight decay	0
Training batch size	128
Test batch size	256
Generated image size	64×64
Digit-pair offset	5000
Number of seeds	5
Seeds	$\{0, 1, 2, 3, 4\}$
Training horizon	50 epochs
Balanced-gradient logging	Full-gradient logging on shared encoder parameters

Table 9: Algorithm-specific hyperparameters in the query-efficiency comparison.

Method	Update scale	Algorithm-specific parameters
MoCo	$\alpha = 0.01$	$\beta = 0.7$, $\beta_\sigma = 0.5$, $\gamma = 0.1$, $\gamma_\sigma = 0.5$, $\rho = 0$
MoDo	$\alpha = 0.01$	$\gamma = 3 \times 10^{-4}$, $\rho = 0.1$
SDMGrad	$\alpha = 0.01$	$\lambda = 1.0$, inner iterations $= 20$
MoCo+	$\approx 10^{-2}$	$k = 0.03$, $w = 0.001$, $\sigma = 0.1$, $\gamma = 0.003$, $c = 150$, $L = 5$
Double-sampling MGDA	$\alpha = 0.015$	$\beta = 0.1$, $\rho = 0$
ClipVR-MGDA (Ours)	$\approx 10^{-2}$	$q = 16$, $S_2 = 32$, $S_1 = 512$, $\eta_1 = 0.015$, $\delta = 1$, $\beta = 0.3$, $\rho = 10^{-4}$

Table 10: Neural network architecture used for Multi-MNIST experiments.

Component	Layer	Output dimension
Input	Generated Multi-MNIST image	$1 \times 64 \times 64$
Shared encoder	Conv2d$(1, 32, 3 \times 3)$ + BatchNorm + ReLU + MaxPool(2)	$32 \times 32 \times 32$
Shared encoder	Conv2d$(32, 64, 3 \times 3)$ + BatchNorm + ReLU + MaxPool(2)	$64 \times 16 \times 16$
Shared encoder	AdaptiveAvgPool(9×9) + Flatten	$64 \cdot 9 \cdot 9$
Shared encoder	Linear$(64 \cdot 9 \cdot 9, 128)$ + ReLU	128
Left-task head	Linear$(128, 10)$	10 logits
Right-task head	Linear$(128, 10)$	10 logits

E.3 Experiment on Synthetic Dataset

Experiment Design. We provide the details needed to reproduce the synthetic experiment in Section 6. Each objective is a quadratic function $f_k(x) = \frac{1}{2}(x - a_k)^\top H_k(x - a_k)$ with gradient $\nabla f_k(x) = H_k(x - a_k)$, for $k = 1, \ldots, K$. The centers a_k are fixed across all runs and placed at radius 5, while the Hessians H_k are diagonal with entries linearly spaced from 1 to 3. The stochastic oracle uses curvature noise. For each stochastic sample, the diagonal curvature is perturbed as $\widehat{H}_k = \mathrm{diag}\big(\max\{H_{k,j} + \sigma z_j,\, 0.1\}\big)_{j=1}^{d}$, $z_j \sim \mathrm{Unif}[-1, 1]$.

The initial point is sampled from a standard normal distribution using the corresponding random seed. Stochasticity is introduced through a noisy gradient oracle, whose noise level is controlled by σ; in the main high-noise experiment we use $\sigma = 512$. All reported curves are averaged over 5 random seeds.

Step-size protocol. We compare all methods as a function of cumulative stochastic-gradient queries, so that algorithms with different per-iteration oracle costs are normalized on the horizontal axis. The step-size and batch settings are listed in Table 11, where DS-MGDA denotes double-sampling stochastic MGDA. Specifically, for ClipVR-MGDA, given the estimator V_t and weight w_t, ClipVR-MGDA uses the hyper-parameters $\alpha_{\max} = 1$, $\alpha_{\text{clip}} = 0.05$, $\gamma_{\max} = 1$, $\gamma_{\text{clip}} = 0.01$, $\beta = 3 \times 10^{-4}$, $\rho = 0.05$ for clip step-size $\alpha_t = \min\{\alpha_{\max}, \alpha_{\text{clip}}/\|V_t w_t\|_2\}$ and $\gamma_t = \min\{\gamma_{\max}, \gamma_{\text{clip}}/\|V_t w_t\|_2\}$.

Table 11: Hyperparameters for the synthetic experiment.

Method	x-step	Aux. params	Batch/estimator
DS-MGDA	5×10^{-3}	$\beta = 10^{-6}$	3×21
MODO	5×10^{-3}	$\gamma = 10^{-6}$	2×32
MOCO	7.5×10^{-3}	$\beta = 5\times10^{-2}, \gamma = 10^{-6}$	64
MOCO+	$\eta = 10^{-2}$	$\gamma = 10^{-6}, c_\beta = 20, w = 100$	2×28
SDMGrad	5×10^{-3}	$\beta = 10^{-5}, S = 1$	21
ClipVR-MGDA(Ours)	clipped	$\beta = 3\times10^{-4}$	$S_1 = 352, S_2 = 32, q = 10$

F Frequently Asked Questions

This section collects several recurring questions regarding the theoretical and practical aspects of our method.

Q1. How does the number of objectives affect the convergence and sample complexity? The dependence on the number of objectives is explicit in our theoretical analysis. In particular, the stepsize contains a factor of order $\Theta\Big(\epsilon/(J_1 + \sqrt{K})\Big)$ with $J_1 = O(1)$, which leads to an overall stochastic-gradient complexity of $O\Big(K(J_1 + \sqrt{K})^3\epsilon^{-3}\Big) = O(K^{5/2}\epsilon^{-3})$. Therefore, increasing the number of objectives introduces a polynomial dependence on K, while the optimal $O(\epsilon^{-3})$ dependence on the target stationarity accuracy remains unchanged.

Q2. Why are two gradient estimators used for each objective? The two-estimator construction is primarily introduced for variance reduction. It follows the standard recursive-estimation principle underlying methods such as SPIDER (Fang et al., 2018) and SARAH (Nguyen et al., 2017): a periodically refreshed large-batch estimator controls accumulated estimation error, while a small-batch recursive estimator tracks gradient variation between refreshes. This structure is essential to obtain the desired stochastic sample complexity and is therefore mainly theoretically motivated, rather than being introduced as an independent empirical heuristic.

Q3. Are minibatches shared across different objectives? Does the analysis require independent stochastic gradients across objectives? In our implementation, the same minibatch is shared across all objectives at each iteration. Consequently, stochastic-gradient noise across different objectives can be correlated. Our analysis does not require independently sampled minibatches for different objectives, and is consistent with this shared-minibatch implementation.

Q4. How should the clipping parameters be selected in practice? The theoretically prescribed parameters are used to establish convergence and sample-complexity guarantees, but practitioners do not need to know quantities such as the target accuracy or stopping-time constants in advance. Consider the simplified primal clipping rule $\alpha_t = \min\{\delta\eta/\|d_t\|, \delta c\}$. When $\|d_t\|$ is large, the update behaves as normalized gradient descent with movement scale $\delta\eta$; when $\|d_t\|$ is small, it

behaves as a fixed-step method with effective learning rate δc. A practical choice is therefore to first select δc to match the learning rate of a standard fixed-step baseline, then choose η according to a desired normalized-update scale, and finally adjust c using the typical magnitude of $\eta/\|d_t\|$ observed in short preliminary runs. After changing c, δ can be adjusted correspondingly so that δc remains fixed.

Q5. Does the method require prior knowledge of the target accuracy ϵ in practice? No. The explicit dependence on ϵ appears in the theoretical parameterization used to establish convergence rates and oracle complexity. The practical implementation can instead use fixed or empirically calibrated clipping parameters, selected according to the procedure above, without requiring the target stationarity accuracy to be known beforehand.

Q6. In what sense is the $O(\epsilon^{-3})$ complexity optimal? The single-objective stochastic nonconvex optimization problem is a special case of our multi-objective generalized-smooth setting. Existing lower bounds (Arjevani et al., 2022) establish an $\Omega(\epsilon^{-3})$ stochastic first-order oracle complexity for this special case under the corresponding stationarity criterion. Therefore, no algorithm can generally improve the dependence on ϵ over the full problem class. Since our method achieves $O(\epsilon^{-3})$ dependence on ϵ, its accuracy dependence is optimal up to problem-dependent and objective-number factors.

Q7. Does the generalized-smoothness assumption cover models beyond globally smooth quadratic objectives? Yes. According to Zhang et al. (2020), the assumption contains the classical mean-square smoothness condition as a special case when the generalized smoothness function is constant. It therefore covers standard stochastic models such as least squares, logistic regression, generalized linear models, and softmax cross-entropy under standard moment conditions. More generally, it also allows the local smoothness level to vary with the gradient magnitude, which is consistent with empirical observations reported for modern neural-network training.

Q8. What new analytical difficulties arise when SPIDER-style variance reduction is extended to stochastic multi-objective optimization with adaptive preference weights? The extension is not a straightforward combination of SPIDER, MGDA, and gradient clipping. Existing SPIDER analyses typically track the stochastic gradient of a single objective. In stochastic MGDA, however, the preference vector changes adaptively over time, while Pareto-stationarity must be characterized relative to an arbitrary reference vector $w \in \Delta_K$. Consequently, controlling only the currently selected direction $V_t w_t$ is insufficient. Our analysis instead introduces a matrix-valued SPIDER estimator V_t that jointly tracks all objectives and establishes an arbitrary-direction error bound of the form

$$\mathbb{E}\Big[\|(V_t - \nabla F(x_t))w\|^2\Big] = O(\epsilon^2), \qquad \forall w \in \Delta_K.$$

Thus, a single tracking sequence can simultaneously support any simplex-weighted multi-gradient required by the MGDA and Pareto-stationarity analyses.

The interaction between this matrix-valued estimator and the evolving preference vector introduces additional terms that do not appear in the standard single-objective SPIDER analysis. In particular, the descent argument contains the MOO-specific mismatch term

$$-\alpha_t \langle V_t w - V_t w_t,\, V_t w_t \rangle ,$$

which compares an arbitrary reference direction with the adaptive direction used at iteration t. Using the projected preference-vector update, this term yields a potential-difference bound involving

$$\frac{\alpha_t}{2\gamma_t \beta} \left(\|w_t - w\|^2 - \|w_{t+1} - w\|^2 \right),$$

together with an additional residual depending jointly on the matrix estimator, the current multi-gradient, and the evolving simplex weight.

These terms create two difficulties. First, the usual MGDA potential no longer telescopes automatically because its coefficient depends on α_t/γ_t. Second, the preference update produces additional estimator-dependent residuals that must be controlled simultaneously with the SPIDER tracking

error. Our coordinated adaptive-step design is introduced specifically to resolve these issues. By enforcing

$$\alpha_t = \delta\gamma_t,$$

the potential coefficient becomes iteration-independent, restoring the telescoping argument. At the same time, the preference-step rule controls the residual terms, while the clipping rule keeps the iterates within the local region required by the generalized-smoothness descent analysis.

Therefore, the main technical novelty is not merely the use of variance reduction or clipping in a multi-objective problem, but the development of a matrix-valued arbitrary-direction tracking analysis and a coordinated adaptive-step mechanism that makes this tracking compatible with an evolving MGDA preference vector. This coupling is what allows the stochastic multi-objective analysis to close and yields the resulting $O(\epsilon^{-3})$ dependence under mean-square generalized smoothness.