

Article

Stochastic Zeroth-Order Multi-Gradient Algorithm for Multi-Objective Optimization

Zhihao Li ¹, Qingtao Wu ^{1,2} , Moli Zhang ^{1,*}, Lin Wang ^{1,2} , Youming Ge ¹ and Guoyong Wang ³ 

¹ School of Information Engineering, Henan University of Science and Technology, Luoyang 471023, China; 220320040393@stu.haust.edu.cn (Z.L.); wqt8921@haust.edu.cn (Q.W.); linwang@haust.edu.cn (L.W.); geym@haust.edu.cn (Y.G.)

² Intelligent System Science and Technology Innovation Center, Longmen Laboratory, Luoyang 471023, China

³ School of Computer and Information Engineering, Luoyang Institute of Science and Technology, Luoyang 471023, China; wgy@lit.edu.cn

* Correspondence: zhangml@haust.edu.cn

Abstract: Multi-objective optimization (MOO) has become an important method in machine learning, which involves solving multiple competing objective problems simultaneously. Nowadays, many MOO algorithms assume that gradient information is easily available and use this information to optimize functions. However, when encountering situations where gradients are not available, such as black-box functions or non-differentiable functions, these algorithms become ineffective. In this paper, we propose a zeroth-order MOO algorithm named SZMG (stochastic zeroth-order multi-gradient algorithm), which approximates the gradient of functions by finite difference methods. Meanwhile, to avoid conflicting gradients between functions and reduce stochastic multi-gradient direction bias caused by stochastic gradients, an SGD-type method is adopted to acquire weight parameters. Under the non-convex setting and mild assumptions, the convergence rate is established for the SZMG algorithm. Simulation results demonstrate the effectiveness of the SZMG algorithm.

Keywords: multi-objective optimization; zeroth-order optimization; stochastic optimization

MSC: 68T07



Received: 10 January 2025

Revised: 27 January 2025

Accepted: 5 February 2025

Published: 14 February 2025

Citation: Li, Z.; Wu, Q.; Zhang, M.; Wang, L.; Ge, Y.; Wang, G. Stochastic Zeroth-Order Multi-Gradient Algorithm for Multi-Objective Optimization. *Mathematics* **2025**, *13*, 627. <https://doi.org/10.3390/math13040627>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In recent years, multi-objective optimization (MOO) has gained much attention in machine learning applications, including autonomous driving [1], language model training [2], safe reinforcement learning [3], and adversarial attacks [4]. MOO aims to optimize multiple, potentially competing objective functions simultaneously. Unlike single-objective optimization (SOO), which aims to find the minimum value of a single function, MOO problems rarely have a solution that is optimal for all objective functions. Therefore, the goal of MOO is to find the *Pareto optimality*, which is the optimal trade-off solution between objective functions.

To solve the MOO problems, various MOO algorithms have been proposed. A class of MOO methods involves heuristic algorithms inspired by natural behaviors, reactions, and communication mechanisms [5]. For example, NSGA-II (non-dominated sorting genetic algorithm II) [6], influenced by the process of biological evolution, explores the solution space to find optimal or near-optimal Pareto solutions by simulating genetic operations such as selection, crossover, and mutation. MOPSO (multi-objective particle swarm optimization) [7] and MOGWO (multi-objective gray wolf optimizer) [8], inspired

by the survival and behavioral activities of organisms, simulate collaborative mechanisms within a population to randomly search for Pareto solutions. The AMOSA (archived multi-objective simulated annealing) method [9], inspired by the annealing process in meals, explores the entire solution space by accepting poorer solutions with a probability. However, these algorithms rarely provide theoretical convergence guarantees and generally yield a set of solutions with each run, which can be computationally expensive for machine learning problems (Table 1).

Table 1. The time and space complexity of a single iteration.

Algorithms	Time Complexity	Space Complex
ZO-AdaMU [10]	$\mathcal{O}(q_1 q_2 d)$	$\mathcal{O}(d)$
BES [11]	$\mathcal{O}(q_1 q_2 d)$	$\mathcal{O}(q_1 q_2 d)$
RSGF [12]	$\mathcal{O}(q_1 q_2 d)$	$\mathcal{O}(q_1 q_2 d)$
ASMG [13]	$\mathcal{O}(M q_1 q_2 d)$	$\mathcal{O}(M q_1 q_2 + q_1 q_2 d)$
SZMG (This paper)	$\mathcal{O}(K(M^2 d + M q_1 q_2 d))$	$\mathcal{O}(M d + M^2 + q_1 q_2 d)$

In contrast, gradient-based MOO algorithms have a first-order theoretical convergence guarantee and are more efficient in finding Pareto solutions through gradient information [14]. One simple method is to adopt the scalarization methods, which convert MOO problems into SOO problems by presetting a group of weights, and then solve MOO problems by using SOO methods [15,16]. But scalarization methods face challenges with conflicting functions, such as conflicting gradients, which can impede the algorithm's ability to find Pareto solutions [17]. Additionally, pre-selecting weights is also a significant challenge. Reference [18] proposed a dynamic weight method called the multi-gradient descent algorithm (MGDA), which determines a multi-gradient descent direction in each iteration by solving a least-squares problem. The direction can optimize all objective functions simultaneously while avoiding the need for pre-selected weights. After that, various MGDA variants have been proposed [19,20].

The aforementioned works mainly focus on deterministic cases. In practical applications, many real-world optimization problems involve unknown or uncertain data parameters, and processing all the data can be extremely costly. An alternative approach is to use a stochastic optimization method. Mercier et al. [21] first proposed the stochastic version of MGDA, called SMGDA, by combining the stochastic gradient descent algorithm with MGDA. However, the paper did not identify a critical property of SMGDA—namely, that the stochastic multi-gradient direction is biased. Liu and Vicente [22] improved the analysis of SMGDA based on previous work and provided convergence analysis under convex and strongly convex settings. Subsequently, refs. [23,24] conducted theoretical studies of SMGDA under convex and non-convex settings.

However, the above works assume that gradient information is easily available. In some cases, gradients are not available, or the cost of obtaining gradients can be prohibitive, such as in black-box functions, where the explicit form and gradient information of the objective functions are unknown, making previous gradient-based optimization algorithms ineffective. Recently, some methods have been proposed to address the problems. Chen et al. [25] proposed a Hessian-aware zeroth-order optimization method, but papers mainly focus on convex settings. In practical applications, non-convex functions can more accurately represent actual scenarios. Ye et al. [13] proposed a multi-objective evolutionary strategy algorithm by iteratively updating the Gaussian distribution to find a Pareto solution. However, the paper assumes that the covariance matrix of the Gaussian distribution is diagonal, making it more suitable for separable functions where the

parameters are independent. Additionally, in the theoretical analysis, it assumes that all functions are bounded, which is an uncommon assumption in standard SOO and MOO convergence analysis. In this paper, a stochastic zeroth-order multi-gradient algorithm (SZMG) is proposed, which adopts milder assumptions that require at least one function to be bounded. As a zeroth-order variant of the method presented in [24], SZMG uses zeroth-order techniques to estimate gradients to address the issue of unavailable function gradients. The theoretical analysis of the SZMG algorithm is established under non-convex settings and milder assumptions. Numerical experiments demonstrate that the SZMG algorithm effectively optimizes both separable and non-separable functions.

The contributions of this paper are as follows:

- This paper introduces a new gradient-free MOO algorithm called stochastic zeroth-order multi-gradient descent algorithm (SZMG), which applies the zeroth-order optimization method to multi-objective optimization problems.
- The convergence rate of $\mathcal{O}(\sqrt{d}/\sqrt{T})$ is established for the SZMG algorithm under non-convex settings and mild assumptions, where d is the dimensionality of the model parameters and T is the number of iterations.
- Numerical experiments demonstrate the effectiveness of the SZMG algorithm.

The notations used in this paper are summarized in Table 2.

Table 2. Some important notations used in the paper.

Notation	Introduction
z	Decision variable
w	Data random variable independence of z
u	A random vector following the Gaussian distribution
ν	Smoothness parameter
d	The dimension of parameter z
M	Total number of functions
$h_i(z)$	The i -th objective function
$h_i^\nu(z)$	The smoothed version of function $h_i(z)$
$\nabla h_i(z)$	Gradient of the i -th objective function
$\nabla h_i^\nu(z)$	Gradient of the function $h_i^\nu(z)$
$s_i(z, u, w)$	Stochastic gradient of the function $h_i^\nu(z)$
$\nabla H(z)$	A matrix containing $\nabla h_1(z), \dots, \nabla h_M(z)$ as columns
$\nabla H_\nu(z)$	A matrix containing $\nabla h_1^\nu(z), \dots, \nabla h_M^\nu(z)$ as columns
$S_\nu(z, u, w)$	A matrix containing $s_1(z, u, w), \dots, s_M(z, u, w)$ as columns
λ_t^*	$\lambda_t^* = \arg \min_{\lambda} \ \nabla H(z_t)\lambda\ ^2$
$\lambda_{t,\nu}^*$	$\lambda_{t,\nu}^* = \arg \min_{\lambda} \ \nabla H_\nu(z_t)\lambda\ ^2$
$\lambda_{t,k+1}$	$\lambda_{t,k+1} = \Pi_{\Delta^M} \left(\lambda_{t,k} - \beta_{t,k} S_\nu(z_t, u'_t, w'_t)^\top S_\nu(z_t, u''_t, w''_t) \lambda_{t,k} \right)$
η, β	Learning rates
c, l, L	A constant greater than 0

2. Related Work

In this section, the development of gradient-based MOO and gradient-free MOO methods are briefly reviewed.

2.1. Gradient-Based Deterministic MOO

Gradient-based MOO methods can be divided into two categories—one is the scalarization method, as seen in References [15,16,26], which converts MOO problems into SOO problems by pre-selecting weights. Indeed, pre-selecting weights is a challenging task and does not resolve the issue of conflicting gradients. Another category is the non-scalarization method, which extends the ideas of SOO methods by finding a direction to optimize all

objective functions simultaneously. Examples include the steepest descent method [27], the proximal method [28], Newton’s method [29], the trust-region method [30], and the conjugate gradient method [31]. Reference [18] proposes a dynamic weighting method called MGDA, which determines a descent direction in each iteration by solving a quadratic subproblem. After that, various MGDA variants have been studied, such as methods for obtaining Pareto distributions [19,20] and minimizing average loss [32]. However, the above studies focus on determination cases. In practical applications, stochastic and gradient-free cases remain challenges.

2.2. Gradient-Based Stochastic MOO

In recent years, the stochastic versions of MGDA have been studied. Reference [21] proposed SMGDA, but the paper did not identify an important property of SMGDA—that the stochastic multi-gradient direction is biased. This property was found in Reference [22], which provides a convergence analysis under strongly convex and convex settings by adopting a linearly increasing batch size. However, using a linearly increasing batch size is impractical. To tackle this challenge, various methods have been proposed to reduce stochastic multi-gradient bias, such as adopting momentum mechanisms [23,33], SGD-type methods [24], and double sampling [34,35]. However, these works all focus on situations where gradients are available. When encountering situations where the gradients of the function are not obtainable, these methods become ineffective.

2.3. Gradient-Free MOO

Many algorithms study black-box MOO problems, such as Bayesian optimization algorithms, as seen in [36,37]. These methods are better suited for handling low-dimensional expensive MOO problems. Another category is heuristic algorithms, such as [6,7,38]. These methods generally explore the entire Pareto solutions, which are computationally expensive and typically lack convergence guarantees. Recently, Reference [25] proposed a Hessian-aware zeroth-order optimization method, but the paper mainly focused on convex settings. Reference [13] proposed a multi-objective evolutionary strategy algorithm by iteratively updating the Gaussian distribution to find a Pareto solution. However, the paper assumes that the covariance matrix of the Gaussian distribution is diagonal, which is more suitable for separable functions. Additionally, achieving theoretical convergence guarantees requires that all function values be bounded, which is uncommon in standard SOO and MOO theoretical analysis. In this paper, we propose a zeroth-order multi-objective optimization algorithm (SZMG), which adopts a milder assumption. Additionally, the theoretical analysis of the SZMG algorithm is established under non-convex settings and milder assumptions. In the numerical experiments, the SZMG algorithm has demonstrated effectiveness in optimizing both separable and non-separable functions.

3. Preliminaries

In this section, the basic concept of MOO problems, the MGDA algorithm, and the zeroth-order gradient estimation methods are introduced.

3.1. Problem Setup

This paper addresses the following stochastic MOO problem:

$$\min_z H(z) = (\mathbb{E}_w[h_1(z, w)], \dots, \mathbb{E}_w[h_M(z, w)]), \quad (1)$$

where $h_i : \mathbb{R}^d \rightarrow \mathbb{R}$ is the i -th continuously differentiable non-convex function with $h_m(z) = \mathbb{E}_w[h_m(z, w)]$, z is a d -dimensional decision variable, w is a random variable, and M is the number of objective functions. In this setting, it is assumed that the gradients

of the functions are not available, and only function values can be queried. Generally, the goal of MOO is to seek *Pareto optimality*, which means no other solution dominates it.

Definition 1 (Pareto optimality). (a) For any two solutions, $z_1, z_2 \in \mathbb{R}^d$, solution z_1 is said to dominate solution z_2 , if $\forall i \in \{1, 2, \dots, M\}$, $h_i(z_1) \leq h_i(z_2)$ and $H(z_1) \neq H(z_2)$. (b) A solution z^* is called Pareto-optimal if no other solution dominates it.

For multi-objective non-convex functions, the goal is generally to find a Pareto stationary point, which is a necessary condition for the Pareto optimal point [23,24,33,39].

Definition 2 (Pareto stationary point). If point z is a Pareto stationary point, then

$$\text{range}(\nabla H(z)) \cap (-\mathbb{R}_{++}^M) = \emptyset,$$

where $\nabla H(z) := (\nabla h_1(z), \nabla h_2(z), \dots, \nabla h_m(z))$, and $\mathbb{R}_{++}^M$ is the positive orthant cone.

3.2. Multi-Gradient Descent Algorithm

The multi-gradient descent algorithm is a first-order gradient algorithm that aims to avoid conflicting gradients between functions. In each iteration, the algorithm first computes the gradients of the function, then obtains the descent direction $\nabla H(z)\lambda^*(z)$, where weights $\lambda^*(z) = (\lambda_1^*(z), \dots, \lambda_M^*(z))^T$ are acquired by solving the following sub-problem:

$$\begin{aligned} \lambda^*(z) &\in \arg \min_{\lambda} \|\nabla H(z)\lambda\|^2 \\ \text{s.t. } \lambda &\in \Delta^m := \left\{ \lambda \in \mathbb{R}^m \mid \mathbf{1}^T \lambda = 1, \lambda \geq 0 \right\}. \end{aligned} \quad (2)$$

Then the parameter z is updated by the following equation:

$$z_{t+1} = z_t - \eta_t d(z_t), \quad (3)$$

where $d(z_t) = \nabla H(z_t)\lambda^*(z_t)$ is the t -th iteration descent direction and η_t is t -th iteration learning rate.

The standard MGDA is given in Algorithm 1, as follows:

Algorithm 1 Multi-gradient descent algorithm (MGDA)

- 1: **Input:** initial model parameter z_0 , learning rate η .
 - 2: **for** $t = 0, 1, \dots, T - 1$ **do**
 - 3: **for** objective function $i = 1, 2, \dots, M$ **do**
 - 4: Calculate gradients $\nabla h_i(z_t)$.
 - 5: **end for**
 - 6: Compute weights $\lambda^*(z_t)$ following (2).
 - 7: Update model parameters z_{t+1} following (3).
 - 8: **end for**
-

3.3. Zeroth-Order Optimization

ZOO is a method for approximating the gradient of a function by querying function values. Its fundamental idea is to use finite difference approximations of the directional derivatives of functions $h(z)$ along certain directions. In this paper, the Gaussian gradient estimator is utilized to approximate the gradients. For any function, $h_i(z)$, the Gaussian smoothed function of $h_i(z)$ is as follows:

$$h_i^v(z) = \mathbb{E}_u [h_i(z + \nu u)], \quad (4)$$

where u follows the standard Gaussian distribution, ν is a smoothness parameter. The gradient of $h_i^\nu(z)$ can be expressed as follows:

$$\nabla h_i^\nu(z) = \mathbb{E}_u \left[\frac{h_i(z + \nu u)}{\nu} u \right] \quad (5)$$

$$= \mathbb{E}_u \left[\frac{h_i(z + \nu u) - h_i(z)}{\nu} u \right], \quad (6)$$

where Equation (5) is the single-point estimation and Equation (6) is the two-point estimation. In general, Equation (5) has a greater gradient variance than Equation (6) [40,41]. In this paper, two-point estimation is adopted to approximate the gradient.

4. Stochastic Zeroth-Order Multi-Gradient Algorithm

To address the challenge of unavailable gradients, this paper applies the ZOO technique to MOO problems. The multi-objective problem (1) is converted into a Gaussian smoothing multi-objective problem, as follows:

$$\min_z H_\nu(z) = (h_1^\nu(z), \dots, h_M^\nu(z)), \quad (7)$$

where $h_i^\nu(z) = \mathbb{E}[h_i(z + \nu u, w)]$. Following Equation (6), the stochastic gradient of the function $h_i^\nu(z)$ is defined as follows:

$$s_i(z, u, w) = \frac{1}{q_1 q_2} \sum_{m=1}^{q_1} \sum_{n=1}^{q_2} \frac{h_i(z + \nu u_m, w_n) - h_i(z, w_n)}{\nu} u_m. \quad (8)$$

In this paper, it is assumed that $\mathbb{E}[h_i(z, w)] = h_i(z)$, which indicates that the stochastic gradient $s_i(z, u, w)$ is an unbiased gradient estimation of $\nabla h_i^\nu(z)$ due to the following:

$$\begin{aligned} \mathbb{E}[s_i(z, u, w)] &= \frac{1}{q_1 q_2} \mathbb{E}_u \left[\mathbb{E}_w \left[\sum_{m=1}^{q_1} \sum_{n=1}^{q_2} \frac{h_i(z + \nu u_m, w_n) - h_i(z, w_n)}{\nu} u_m | u \right] \right] \\ &= \frac{1}{q_1 q_2} \mathbb{E}_u \left[\sum_{m=1}^{q_1} \sum_{n=1}^{q_2} \frac{h_i(z + \nu u_m) - h_i(z)}{\nu} u_m \right] \\ &= \nabla h_i^\nu(z). \end{aligned} \quad (9)$$

According to Equation (9), an unbiased stochastic gradient of function $h_i^\nu(z)$ can be obtained by querying the function values. To avoid potential conflicts in gradient direction, such as $\langle s_i(z, u, w), s_j(z, u, w) \rangle < 0$, the MGDA-type method is adopted to determine the weight values by solving the following problem:

$$\begin{aligned} \min_{\lambda} & \|\nabla H_\nu(z) \lambda\|^2 \\ \text{s.t.} & \lambda \in \Delta^M := \left\{ \lambda \in \mathbb{R}^M \mid \mathbf{1}^\top \lambda = 1, \lambda \geq 0 \right\}, \end{aligned} \quad (10)$$

where $\nabla H_\nu(z) := (\nabla h_1^\nu(z), \dots, \nabla h_M^\nu(z))$.

According to References [22,33], replacing deterministic gradients $\nabla H_\nu(z)$ with stochastic gradients $S_\nu(z, u, w)$ in Equation (10) and directly obtaining the weight values will introduce a bias $\mathbb{E}[S_\nu(z, u, w) \lambda_\nu^*(z, u, w)] \neq \nabla H_\nu(z) \lambda_\nu^*(z)$. To reduce this bias, an SGD-type method from [24] is adopted to obtain the optimal weight value λ_ν^* .

The method for updating the weight λ_t is as follows:

$$\lambda_{t,k+1} = \Pi_{\Delta^M} \left(\lambda_{t,k} - \beta_{t,k} S_\nu(z_t, u'_t, w'_t)^\top S_\nu(z_t, u''_t, w''_t) \lambda_{t,k} \right), \quad (11)$$

where u'_t, w'_t, u''_t, w''_t are independent samples, $\beta_{t,k}$ is the learning rate for the weight λ_t at the k -th iteration, and Π_{Δ^M} represents a projection onto Δ^M . The gradient $S_v(z_t, u'_t, w'_t)^\top S_v(z_t, u''_t, w''_t) \lambda_{t,k}$ is unbiased for $\|\nabla H_v(z) \lambda\|^2$ because

$$\mathbb{E} \left[S_v(z_t, u'_t, w'_t)^\top S_v(z_t, u''_t, w''_t) \lambda_{t,k} \right] = \nabla H_v(z_t)^\top \nabla H_v(z_t) \lambda_{t,k}. \quad (12)$$

After obtaining the weights, the decision variable z_t will be updated by the following equation:

$$z_{t+1} = z_t - \eta_t S_v(z_t, u_t, w_t) \lambda_{t,K}, \quad (13)$$

where η_t is the learning rate and K is the total number of iterations for the weight.

The proposed SZMG algorithm is given in Algorithm 2.

Algorithm 2 Stochastic zeroth-order multi-gradient algorithm (SZMG)

```

1: Input: initial model parameter  $z_0$ , weights  $\lambda_0$ , learning rates  $\{\eta_t\}_{t=0}^T, \{\beta_t\}_{t=0}^T$ .
2: for  $t = 0, 1, \dots, T - 1$  do
3:   Set weights  $\lambda_{t,0} = \lambda_{t-1,K}$ .
4:   for  $k = 0, 1, \dots, K - 1$  do
5:     Calculate zeroth-order gradient  $S_v(z_t, u'_t, w'_t), S_v(z_t, u''_t, w''_t)$ .
6:     Update weights  $\lambda_{t,k+1}$  following (11).
7:   end for
8:   Calculate zeroth-order gradient  $S_v(z_t, u_t, w_t)$ .
9:   Update model parameters  $z_{t+1}$  following (13).
10: end for

```

5. Convergence Analysis

In this section, some standard assumptions are introduced first. Then, the error bound between gradient $S_v(z_t, u_t, w_t) \lambda_{t,K}$ and gradient $\nabla H(z_t) \lambda^*(z_t)$ is analyzed. Finally, the convergence rate of the SZMG algorithm is established. All the proofs are put in Appendix B.

Assumption 1. For any $z \in \mathbb{R}^d$, the estimation of $h_i(z)$ is unbiased such that $\mathbb{E}_w[h_i(z, w)] = h_i(z)$.

This assumption ensures that Equation (9) holds, making $s_i(z, u, w)$ an unbiased estimate of the gradients $\nabla h_i^v(z)$. This is also a standard assumption in stochastic zeroth-order optimization [12,40,42].

Assumption 2. All functions $h_1(z, w), \dots, h_M(z, w)$ are l -Lipschitz continuous for any w and $h_1(z), \dots, h_M(z)$ are L -smoothness, i.e.,

$$\|\nabla h_i(z_1, w)\| \leq l,$$

$$h_i(z_1) \leq h_i(z_2) + \langle \nabla h_i(z_2), z_1 - z_2 \rangle + \frac{L}{2} \|z_1 - z_2\|^2,$$

for any $z_1, z_2 \in \mathbb{R}^d$.

Applying Jensen's inequality, we have the following:

$$\|\nabla h_i(z)\|^2 \leq \mathbb{E} \left[\|\nabla h_i(z, w)\|^2 \right] \leq l^2, \quad (14)$$

which implies that functions $h_1(z), \dots, h_M(z)$ are l -Lipschitz continuous.

This is a standard assumption in MOO problems [23,24]. Here, the assumption that $h_i(z, w)$ is l -Lipschitz continuous is to ensure that gradient $s_i(z, u, w)$ is bounded for all $i \in \{1, \dots, M\}$, which is necessary to control the drift of the optimal weight λ , thereby establishing convergence guarantees in the non-convex setting. This assumption is used for zeroth-order stochastic optimization problems in [12,43]. The L -smoothness property of the function is a standard assumption used to establish the convergence of non-convex functions [44].

Assumption 3. For some $\lambda \in \Delta^M$, there exists a constant C_H such that $H_v(z_0)\lambda - \min_z H_v(z)\lambda \leq C_H$.

This assumption implies that there exists at least an objective function that is bounded at the initial point, such as $h_i^v(z_0) - \min_z h_i^v(z) \leq C_H$, which implies that the distance between the initial point and the Pareto optimal solution is finite. From this, it can be inferred that the algorithm can converge to the Pareto solution within a reasonable amount of time. Compared to [13], which requires the function values bounded for all objective functions at any point z , this is a milder assumption.

These assumptions are typically reasonable in the context of black-box adversarial attacks on an image classification neural network model. For example, in [45], the sparse adversarial attack is transformed into a multi-objective black-box optimization problem. In this context, the aforementioned assumptions hold.

The error upper bound between $S_v(z_t, u_t, w_t)\lambda_{t,K}$ and $\nabla H(z_t)\lambda^*(z_t)$ is established as follows:

Proposition 1. Suppose Assumptions 1 and 2 are satisfied. If we set $v = \Theta(\frac{1}{\sqrt{Td}})$, $K = T^2$, $\sqrt{d} \leq q_1 q_2 < d$, and $\beta_{t,k} = \frac{c}{\sqrt{k}}$, where $c > 0$ is a constant, we have the following:

$$\begin{aligned} & \mathbb{E} \left[\left\| \mathbb{E}[S_v(z_t, u_t, w_t)\lambda_{t,K}] - \nabla H(z_t)\lambda_t^* \right\|^2 | z_t \right] \\ & \leq \frac{c(\log(K) + 2)}{\sqrt{K}} \left(\frac{Mv^2}{2q_1 q_2} L^2 (d+6)^3 + \frac{2d+9+q_1 q_2}{q_1 q_2} Ml^2 \right)^2 \\ & \quad + (\log(K) + 2) \left(\frac{1}{c\sqrt{K}} + \sqrt{2}vMLl(d+3)^{3/2} \right) + \frac{Mv^2}{4} L^2 (d+3)^3 \\ & = \mathcal{O} \left(\frac{\log(K)\sqrt{d}}{K^{1/4}} \right). \end{aligned} \quad (15)$$

Proposition 1 shows that the error bound will converge to 0 as $K \rightarrow \infty$. The following theorem shows the SZMG can converge to a Pareto stationary point under a non-convex setting and mild assumptions:

Theorem 1. Suppose Assumptions 1, 2, and 3 are satisfied. If we set $\eta = \Theta(T^{-\frac{1}{2}})$, $v = \Theta(\frac{1}{\sqrt{Td}})$, $K = \Theta(T^2)$, $\sqrt{d} \leq q_1 q_2 < d$, and $\beta_{t,k} = \frac{c}{\sqrt{k}}$, where $c > 0$ is a constant, we have the following:

$$\begin{aligned}
& \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \left[\|\nabla H(z_t) \lambda_t^*\|^2 \right] \\
& \leq \frac{2C_H}{T\eta} + 4l \sqrt{\frac{1 + \log(T)}{cT}} + \frac{Mv^2}{2} L^2 (d+3)^3 \\
& \quad + \frac{Mv^2}{2q_1q_2} L^2 (d+6)^3 \left(2l \sqrt{\frac{2c + 2c \log(T)}{T}} + L\eta \right) \\
& \quad + \frac{2d+9+q_1q_2}{q_1q_2} Ml^2 \left(2l \sqrt{\frac{2c + 2c \log(T)}{T}} + L\eta \right) \\
& = \mathcal{O} \left(\frac{d \sqrt{\log(T)}}{q_1q_2 \sqrt{T}} \right). \tag{16}
\end{aligned}$$

Theorem 1 shows that the proposed SZMG algorithm converges to a Pareto stationary point with a convergence rate of $\mathcal{O}(\frac{d \sqrt{\log(T)}}{q_1q_2 \sqrt{T}})$. By choosing the parameter $q_1q_2 = \sqrt{d}$ and ignoring the term $\sqrt{\log(T)}$, which grows very slowly, the convergence rate of the SZMG algorithm reaches $\mathcal{O}(\sqrt{d}/\sqrt{T})$, which indicates that the SZMG algorithm has a convergence rate that matches the one in the zeroth-order stochastic single-objective non-convex setting [12]. To achieve an ϵ -optimal solution, the query complexity for each objective function is bounded by $\mathcal{O}(d^{9/2}/\epsilon^6)$.

To reduce function query complexity, the error between $S_v(z_t, u_t, w_t) \lambda_{t,K}$ and $\nabla H(z_t) \lambda^*(z_t)$ is not considered. The following theorem shows that the query complexity for converging to an ϵ -optimal solution is bounded by $\mathcal{O}(d^{3/2}/\epsilon^2)$.

Theorem 2. Suppose Assumptions 1–3 are satisfied. If we set $\eta = \Theta(T^{-1/2})$, $\beta = \Theta(T^{-1/2}d^{-1/2})$, $v = \Theta(T^{-1/2}d^{-1})$, $\sqrt{d} \leq q_1q_2 < d$, and $K = 1$, we have the following:

$$\begin{aligned}
& \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \left[\|\nabla H(z_t) \lambda_t^*\|^2 \right] \\
& \leq \frac{2C_H}{T\eta} + \beta \left(\frac{Mv^2}{2q_1q_2} L^2 (d+6)^3 + \frac{2d+9+q_1q_2}{q_1q_2} Ml^2 \right)^2 + \frac{2}{T\beta} \\
& \quad + \eta L \left(\frac{Mv^2}{2q_1q_2} L^2 (d+6)^3 + \frac{2d+9+q_1q_2}{q_1q_2} Ml^2 \right) + \frac{Mv^2}{2} L^2 (d+3)^3 \\
& = \mathcal{O} \left(\frac{d^{3/2}}{T^{1/2} (q_1q_2)^2} \right). \tag{17}
\end{aligned}$$

If we set the parameter $q_1q_2 = \sqrt{d}$, Theorem 2 shows that the SZMG algorithm converges to a Pareto stationary point with a convergence rate of $\mathcal{O}(\sqrt{d}/\sqrt{T})$. To achieve an ϵ -optimal solution, the query complexity for each function is bounded by $\mathcal{O}(d^{3/2}/\epsilon^2)$.

6. Experiments

In this section, several synthesis experiments are organized to verify the performance of the proposed SZMG algorithm.

6.1. Toy Example

To demonstrate that the proposed SZMG algorithm achieves better performance in stochastic multi-objective black-box settings, the first synthesis experiment is a

two-dimensional problem, which is used in [34]. The objective functions are defined as follows:

$$h_1(z) = f_1(z)g_1(z) + f_2(z)c_1(z) \quad (18)$$

$$h_2(z) = f_1(z)g_2(z) + f_2(z)c_2(z), \quad (19)$$

where the functions $f_1(z)$, $f_2(z)$, $g_1(z)$, $g_2(z)$, $c_1(z)$, and $c_2(z)$ are defined by the following equations:

$$\begin{aligned} f_1(z) &= \max(\tanh(0.5z_2), 0) \\ f_2(z) &= \max(\tanh(-0.5z_2), 0) \\ c_1(z) &= \left((-z_1 + 3.5)^2 + 0.1(-z_2 - 1)^2 \right) / 10 - 20 \\ &\quad - 2z_1w_1 - 5.5z_2w_2 \\ c_2(z) &= \left((-z_1 - 3.5)^2 + 0.1(-z_2 - 1)^2 \right) / 10 - 20 \\ &\quad + 2z_1w_1 - 5.5z_2w_2 \\ g_1(z) &= \log(\max(|0.5(-z_1 - 7) - \tanh(-z_2)|, \\ &\quad 0.000005)) + 6 \\ g_2(z) &= \log(\max(|0.5(-z_1 + 3) - \tanh(-z_2) + 2|, \\ &\quad 0.000005)) + 6, \end{aligned}$$

where $z = [z_1, z_2]^\top \in \mathbb{R}^2$ is the model parameter, and $w = [w_1, w_2]^\top \in \mathbb{R}^2$ represents stochastic data, which follow the standard Gaussian distribution, as follows:

The following three initial points are selected:

$$z_0 \in \{(-8.5, 7.5), (-10, -3), (9, 9)\}$$

The optimization trajectories of ZO-AdaMU (zeroth-order adapt momentum and uncertainty) [10], BES (Bernoulli smoothing) [11], RSGF (randomized stochastic gradient-free) [12], ASMG (adaptive stochastic multi-objective gradient) [13], MGDA [18], and SZMG are plotted in Figure 1. RSGF uses the Gaussian gradient estimator and BES adopts the variant of the Bernoulli distribution to estimate the gradient. ZO-AdaMU adopts the Gaussian gradient estimator by adapting perturbation with momentum and uncertainty. These three algorithms primarily focus on single-objective problems. In multi-objective scenarios, multiple objectives are aggregated into a single objective by applying equal weights to each individual objective. ASMG is an adaptive multi-objective gradient algorithm based on evolutionary strategies. MGDA is a first-order gradient algorithm. All algorithms adopt the Adam optimizer and the number of iterations is $T = 70,000$. Other hyperparameters are summarized in Table 3.

In Figure 1, the results show that algorithms SZMG and ASMG can effectively optimize stochastic multi-objective black-box functions while addressing the conflicting gradient problem. The three initial points have all converged to the Pareto front. In contrast, the ZO-AdaMU, BES, and RSGF methods all become trapped in a valley at initial points $(-8, 5, 7.5)$, $(9, 9)$, and fail to reach the Pareto front due to the dominant gradient of a specific objective.

Figure 2 examines the impact of different numbers of inner iterations K on the error between gradients $S_v(z_t, u_t, w_t)\lambda_{t,K}$, and $\nabla H(z_t)\lambda^*(z_t)$. The gradient deviation for each trajectory is measured using the average norm $\frac{1}{T} \sum_{t=1}^T \|S_v(z_t, u_t, w_t)\lambda_{t,K} - \nabla H(z_t)\lambda^*(z_t)\|$.

From the results, it can be seen that as K increases, the gradient deviation becomes smaller. This result validates Proposition 1. In Figure 1, the optimization trajectories of the SZMG algorithm are presented for different numbers of inner iterations K .

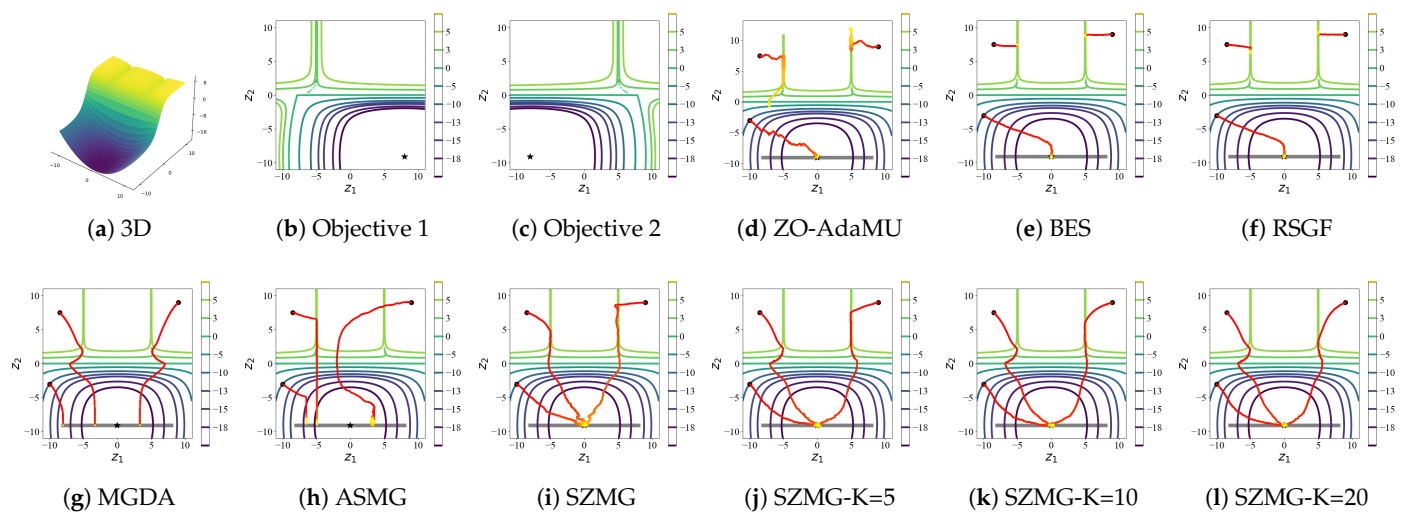


Figure 1. (a–c) show the 3D images of two objective functions, and the contour maps of objective 1 and objective 2, where the black $\star$ marks the minimum point and the background solid contours display the landscape. (d–l) show the optimization trajectories of different algorithms, where black $\bullet$ marks three different initial points, the gray bar is the Pareto front, and the black $\star$ denotes the point on the Pareto front corresponding to equal weights for each objective. All trajectories are shown fading from red (start) to yellow (end).

Table 3. Summary of hyperparameters.

	ν	q_1	q_2	η	K	γ	β
ZO-AdaMU	0.00001	8	8	0.002	-	-	-
BES	0.00001	8	8	0.002	-	-	-
RSGF	0.00001	8	8	0.002	-	-	-
ASMG	-	8	8	0.01	-	$\frac{1}{t+1}$	-
MGDA	-	-	full	0.002	-	-	-
SZMG	0.00001	8	8	0.002	1	-	$\frac{0.01}{\sqrt{k}}$

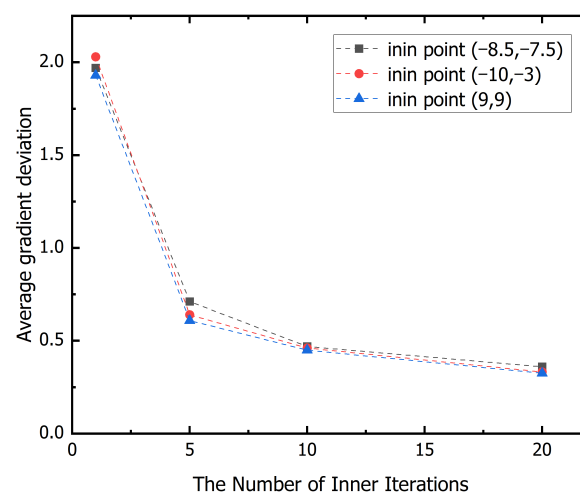


Figure 2. The average norm of gradient deviation with respect to the number of inner iterations K at the three initial points.

6.2. Separable and Non-Separable Functions

To validate the performance of the proposed SZMG algorithm under separable and non-separable functions, SZMG is compared to other algorithms using synthetic functions in 500 and 1000 dimensions.

For separable functions, we have the following:

$$H(z) = (h_1(z), h_2(z)) = \left(\sum_{i=1}^d 10^{\frac{2(i-1)}{d-1}} z_i^2, 10d - \sum_{i=1}^d 10 \cos(2\pi 10^{\frac{(i-1)}{d-1}} z_i) + \sum_{i=1}^d (10^{\frac{(i-1)}{d-1}} z_i)^2 \right). \quad (20)$$

For non-separable functions, we have the following:

$$H(x) = H(Rz) = (h_1(x), h_2(x)), \quad (21)$$

where $h_1(x) = \sum_{i=1}^d x_i^2 - 10 \cos(2\pi x_i) + 10$, $h_2(x) = \frac{1}{d} \sum_{i=1}^{d-1} 100(x_i^2 - x_{i+1})^2 + (x_i - 4)^2$, and R is an orthogonal matrix generated by applying the Gram-Schmidt procedure to a random matrix with a standard normal distribution.

The SZMG algorithm will be compared with ZO-AdaMU [10], BES [11], RSGF [12], and ASMG [13]. For ZO-AdaMU, BES, RSGF, and SZMG methods, the learning rate is selected from $\{0.01, 0.005, 0.001, 0.0005, 0.0001, 0.00005, 0.00001\}$, and the smoothing parameters ν is set to 0.00001. In the SZMG algorithm, the weight learning rate β is set to 0.005, and the number of inner iterations K is set to 1. The ASMG algorithm selects the learning rate from $\{5, 1, 0.5, 0.1, 0.05\}$. The total number of iterations T and the sample batch size q_1 are set to 5000 and 50, respectively, for all algorithms.

In Figures 3 and 4, the convergence results of five algorithms on the separable function under different dimensions are presented. Specifically, the function value with respect to the number of iterations is presented. The results demonstrate that the proposed SZMG algorithm achieves better performance. The ASMG algorithm exhibits faster convergence on the separable function. ZO-AdaMU, BES, and RSGF methods converge to a suboptimal Pareto solution due to conflicting gradients. In Figures 5 and 6, the convergence results of five algorithms on the non-separable functions under different dimensions are presented. The results can intuitively show that the SZMG algorithm outperforms ASMG in terms of convergence performance and results. These results show that the SZMG algorithm performs well on both separable and non-separable functions, demonstrating the effectiveness of the SZMG algorithm.

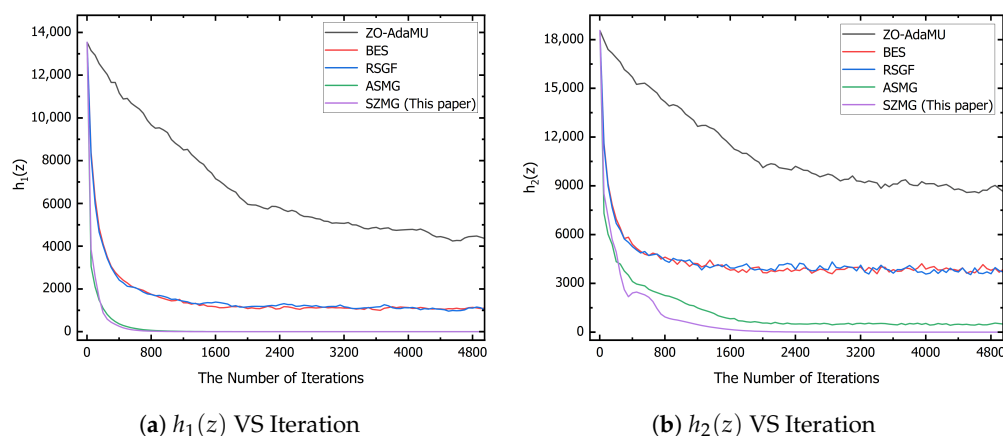


Figure 3. The two objective function values with respect to the number of iterations T on the separable function with 500 dimensions.

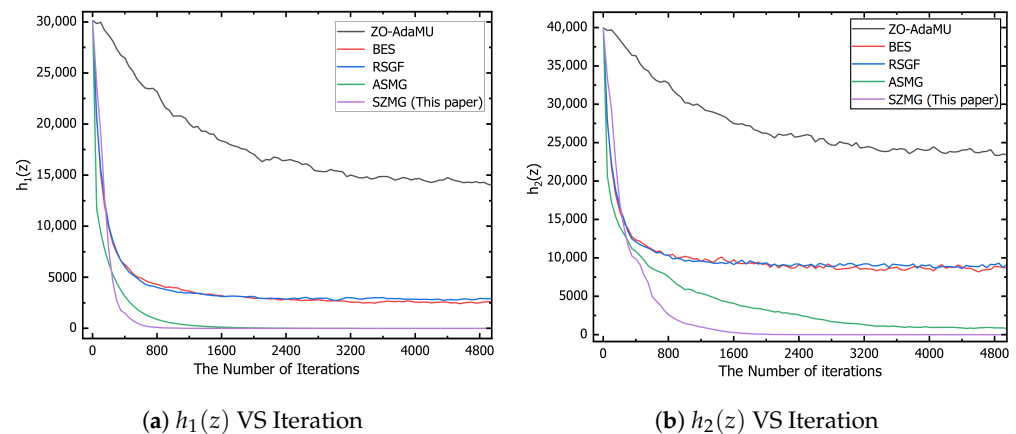


Figure 4. The two objective function values with respect to the number of iterations T on the separable function with 1000 dimensions.

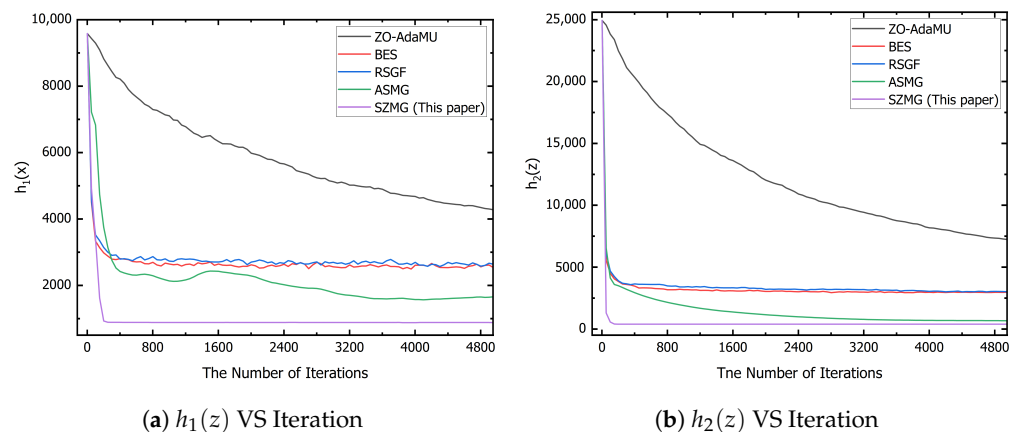


Figure 5. The two objective function values with respect to the number of iterations T on the non-separable function with 500 dimensions.

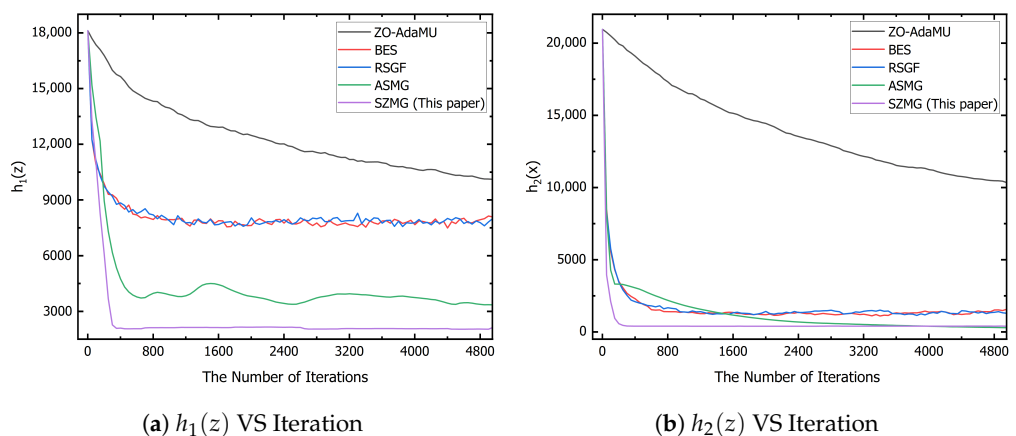


Figure 6. The two objective function values with respect to the number of iterations T on the non-separable function with 1000 dimensions.

6.2.1. Runtime

In this subsection, the runtimes of different algorithms on separable functions are compared to evaluate their computational complexity. The runtime of an algorithm is the average time per iteration over 5000 iterations. The experimental results were obtained using a PC equipped with a 3.60-GHz Intel Core i7-7700 CPU (Acer, Luoyang, China).

Figure 7 shows the average time per iteration for the methods ZO-AdaMU, BES, RSGF, ASMG, and SZMG across different dimensions of separable functions. From the results, the BES and RSGF methods have the shortest runtime per iteration because they do not need to compute weights. ZO-AdaMU takes the longest time due to requiring multiple uncertain perturbations internally. SZMG requires an additional gradient estimation and matrix operation to obtain the weights, making its runtime longer than RSGF. Compared to ASMG, SZMG has a shorter runtime, which indicates that the SZMG has better efficiency.

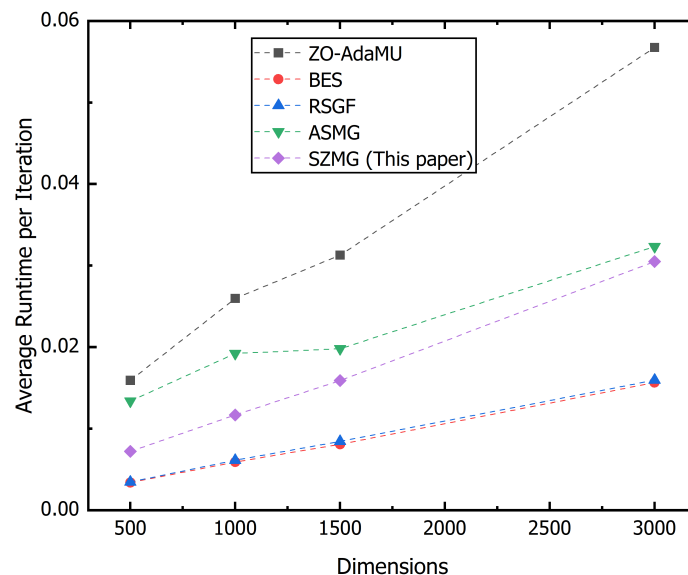


Figure 7. The average runtime per iteration with respect to the dimensions, d , on separable functions.

6.2.2. Scalability

In this subsection, the scalability of SZMG is studied on separable and non-separable functions at 1500 and 3000 dimensions.

Figures 8 and 9 show the convergence results of algorithms ZO-AdaMU, BES, RSGF, ASMG, and SZMG on separable functions at 1500 and 3000 dimensions. These results indicate that the SZMG algorithm has better performance and can converge to the Pareto solutions. Figures 10 and 11 present the convergence results on non-separable functions at 1500 and 3000 dimensions. These results show that SZMG delivers better performance, which implies that gradient-based optimization methods are better suited for handling high-dimensional problems; moreover, using dynamic weights based on SGD effectively addresses conflicting gradients. These results demonstrate the SZMG algorithm's effectiveness.

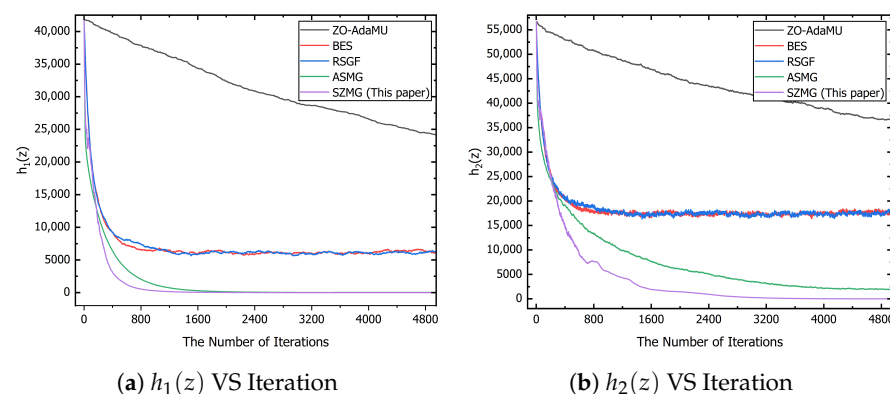


Figure 8. The two objective function values with respect to the number of iterations T on the separable function with 1500 dimensions.

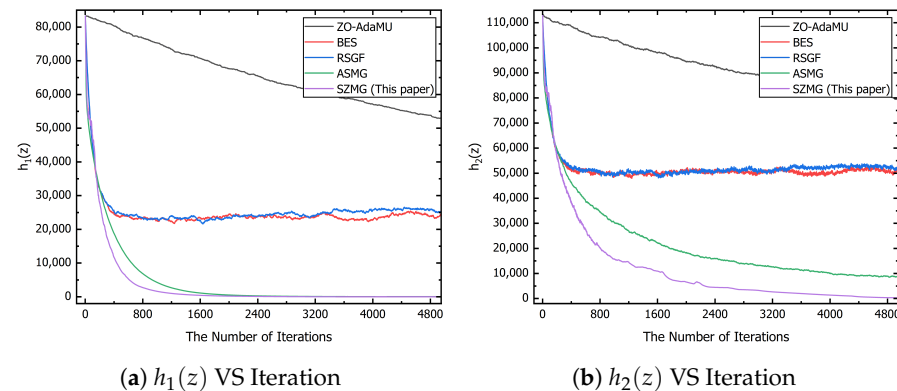


Figure 9. The two objective function values with respect to the number of iterations T on the separable function with 3000 dimensions.

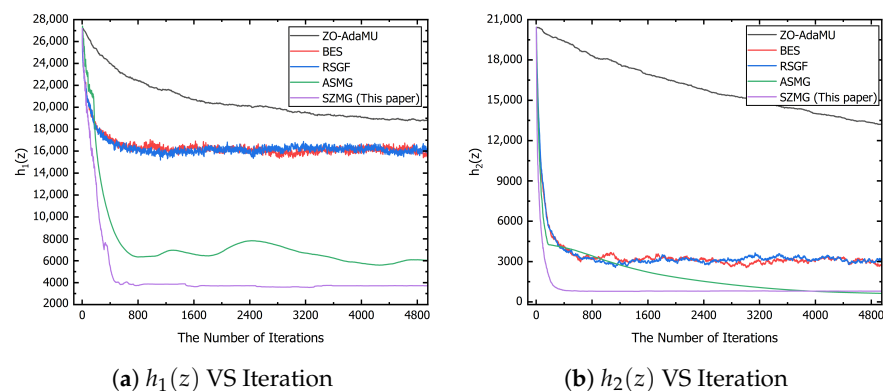


Figure 10. The two objective function values with respect to the number of iterations T on the non-separable function with 1500 dimensions.

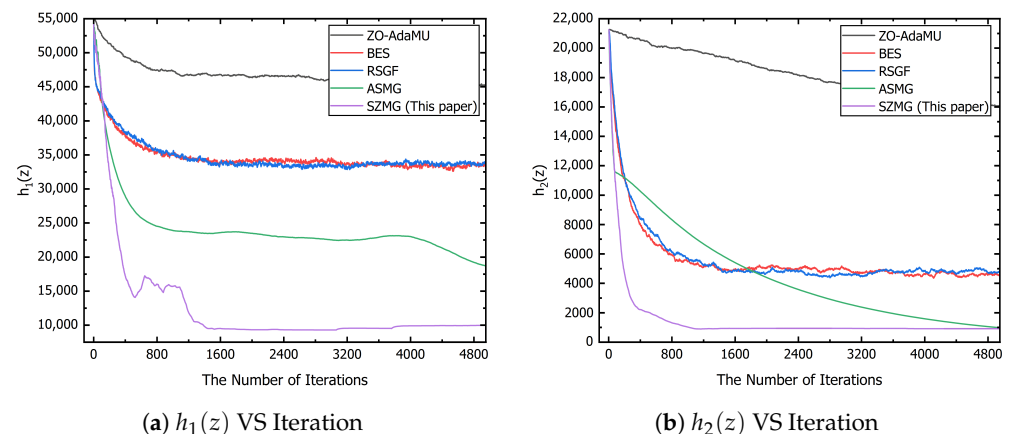


Figure 11. The two objective function values with respect to the number of iterations T on the non-separable function with 3000 dimensions.

6.2.3. Sensitiveness to Batch Size

The sample batch size is a key parameter in the proposed SZMG algorithm. A larger batch size can reduce the variance of zeroth-order gradients, leading to a more accurate gradient direction, but it also results in higher time and function query overheads. In this subsection, we study the impact of different batch sizes on the convergence of the SZMG algorithm for non-separable functions.

Figure 12 shows the impact of batch sizes 10, 30, 50, and 70 on the SAMG algorithm for the 1000-dimensional non-separable function. From the results, it can be observed that as

the batch size increases, the convergence trajectory of the SZMG algorithm becomes more stable and the convergence improves. Figure 13 shows the average runtime per iteration for different batch sizes. The runtime increases linearly with the batch size. Therefore, careful selection of an appropriate batch size is critical for balancing time consumption and stability for the SZMG algorithm.

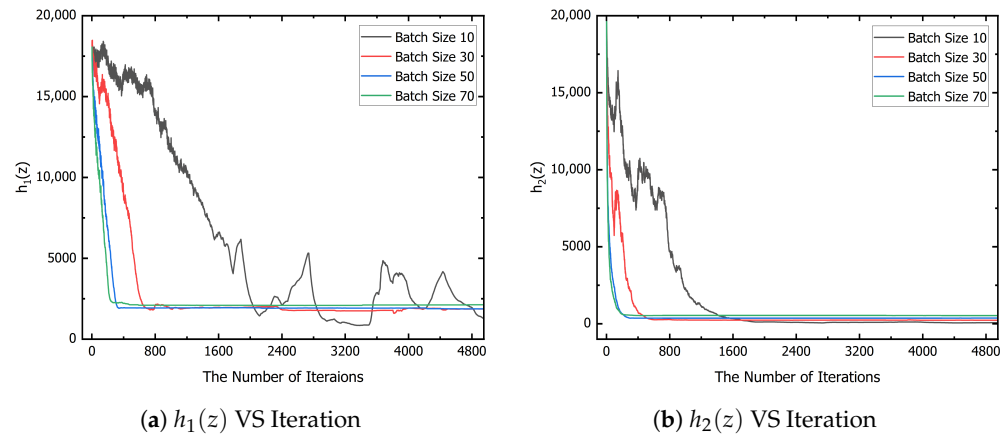


Figure 12. The two objective function values with respect to the number of iterations T on the non-separable function with 1000 dimensions.

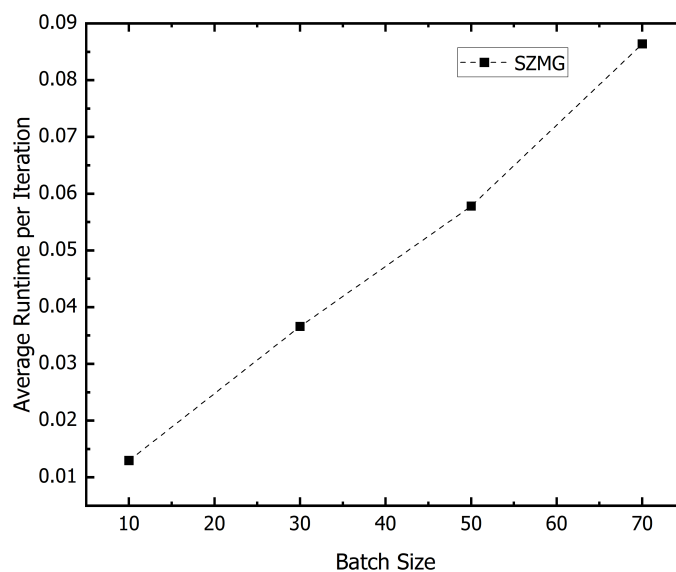


Figure 13. The average runtime per iteration with respect to the batch size on non-separable functions.

7. Discussion

In the experiments section, several numerical experiments were conducted to validate the practical performance of the proposed SZMG algorithm. The results show that SZMG demonstrates promising performance on smooth black-box functions. As a zeroth-order variant of first-order multi-objective optimization algorithms, SZMG exhibits a convergence rate comparable to first-order methods and can serve as a substitute for these algorithms in handling black-box multi-objective optimization problems or scenarios with expensive gradients. Theoretically, we have demonstrated that the convergence rate of SZMG is $\mathcal{O}(d)$ worse than the first-order multi-objective optimization methods.

Despite its advantages, SZMG has potential limitations such as the curse of dimensionality. As the dimensionality of the problem increases, the number of required function evaluations also rises significantly, leading to a sharp increase in computational costs.

Additionally, the theoretical convergence guarantees of SZMG are based on smooth functions, and its performance may degrade when applied to non-smooth functions. These limitations are also important research directions for zeroth-order multi-objective optimization algorithms.

In real-world applications, the SZMG algorithm has a wide range of application scenarios, such as black-box adversarial attacks. In single-objective optimization, zeroth-order algorithms are widely used in black-box attacks and have achieved excellent results [46]. Recently, Reference [45] explored sparse adversarial attacks as multi-objective problems, employing heuristic algorithms to generate adversarial examples. There is significant potential for applying the SZMG algorithm in this field.

8. Conclusions

In this paper, a gradient-free stochastic multi-objective optimization algorithm is proposed for addressing black-box multi-objective optimization problems. The algorithm applies zeroth-order optimization techniques to multi-objective optimization and adopts a dynamic weighting method to avoid conflicting gradients. Theoretically, this paper provides a convergence guarantee for the proposed algorithm under non-convex settings and mild assumptions. Moreover, synthetic experiments demonstrate the effectiveness of the proposed algorithm. In the future, it will be valuable to explore methods for zeroth-order algorithms to handle high-dimensional problems and introduce variance reduction techniques into zeroth-order optimization to reduce gradient variance and achieve better convergence performance.

Author Contributions: Conceptualization, Z.L., Q.W., M.Z., L.W., Y.G. and G.W.; methodology, Z.L. and Q.W.; software, Z.L. and Q.W.; validation, Z.L., Q.W., M.Z., L.W., Y.G. and G.W.; formal analysis, Z.L. and Q.W.; investigation, Z.L., Q.W. and M.Z.; resources, Q.W. and M.Z.; data curation, Z.L., Q.W. and M.Z.; writing—original draft preparation, Z.L.; writing—review and editing, Z.L., Q.W. and M.Z.; visualization, Z.L., Q.W. and M.Z.; supervision, Q.W.; project administration, Q.W.; funding acquisition, Q.W. and M.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (NSFC) under grant no. 62176113, and in part by the Key Technologies R & D Program of Henan Province under grant no. 242102211024, and in part by the Longmen Laboratory Frontier Exploration Project of Henan Province under grant no. LMQYTSKT035.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Extra Lemmas

In this section, some lemmas are presented for the subsequent proof.

Lemma A1 ([47]). *If functions $h_1(z), \dots, h_M(z)$ are l -Lipschitz continuous and exhibit L -smoothness, then $h_1^v(z), \dots, h_M^v(z)$ are also l -Lipschitz continuous and L -smoothness.*

Lemma A2. *If Assumptions 1 and 2 are satisfied, it holds that for all $z \in \mathbb{R}^d$*

$$\|\nabla H_v(z)\|^2 \leq Ml^2, \quad (\text{A1})$$

$$\|\nabla H_v(z) - \nabla H(z)\|^2 \leq \frac{Mv^2}{4}L^2(d+3)^3, \quad (\text{A2})$$

$$\mathbb{E}\left[\|S_v(z, u, w) - \nabla H_v(z)\|^2\right] \leq \frac{Mv^2}{2q_1q_2}L^2(d+6)^3 + \frac{(2d+9)}{q_1q_2}Ml^2. \quad (\text{A3})$$

Proof. Based on Lemma A1, we have the following:

$$\|\nabla h_i^v(z)\| \leq l. \quad (\text{A4})$$

Therefore, the squared norm of the gradient of $H_v(z)$ satisfies the following:

$$\begin{aligned} \|\nabla H_v(z)\|^2 &= \|(\nabla h_1^v(z), \dots, \nabla h_M^v(z))\|^2 \\ &\leq M \max_i \|\nabla h_i^v(z)\|^2 \\ &\leq Ml^2. \end{aligned} \quad (\text{A5})$$

Following Lemma 3 and Theorem 4 in [47], we have the following:

$$\|\nabla h_i^v(z) - \nabla h_i(z)\| \leq \frac{v}{2}L(d+3)^{3/2}, \quad (\text{A6})$$

$$\mathbb{E}\left[\left\|\frac{h_i(z+vu, w) - h_i(z, w)}{v}u\right\|^2\right] \leq \frac{v^2}{2}L^2(d+6)^3 + 2(d+4)l^2. \quad (\text{A7})$$

Therefore, we have the following:

$$\begin{aligned} \|\nabla H_v(z) - \nabla H(z)\|^2 &\leq M \max_i \|\nabla h_i^v(z) - \nabla h_i(z)\|^2 \\ &\leq \frac{Mv^2}{4}L^2(d+3)^3, \end{aligned} \quad (\text{A8})$$

$$\begin{aligned} &\mathbb{E}\left[\|s_i(z, u, w) - \nabla h_i^v(z)\|^2\right] \\ &= \frac{1}{q_1^2q_2^2}\mathbb{E}\left[\left\|\sum_{m=1}^{q_1}\sum_{n=1}^{q_2}\left[\frac{h_i(z+vu_m, w_n) - h_i(z, w_n)}{v}u_m\right] - \sum_{m=1}^{q_1}\sum_{n=1}^{q_2}[\nabla h_i^v(z)]\right\|^2\right] \\ &\leq \frac{1}{q_1q_2}\max_{m,n}\mathbb{E}\left[\left\|\frac{h_i(z+vu_m, w_n) - h_i(z, w_n)}{v}u_m - \nabla h_i^v(z)\right\|^2\right] \\ &\leq \frac{1}{q_1q_2}\max_{m,n}\mathbb{E}\left[\left\|\frac{h_i(z+vu_m, w_n) - h_i(z, w_n)}{v}u_m\right\|^2 + \|\nabla h_i^v(z)\|^2\right] \\ &\leq \frac{v^2L^2}{2q_1q_2}(d+6)^3 + \frac{(2d+9)}{q_1q_2}l^2, \end{aligned} \quad (\text{A9})$$

$$\begin{aligned} \mathbb{E}\left[\|S_v(z, u, w) - \nabla H_v(z)\|^2\right] &\leq M \max_i \mathbb{E}\left[\|s_i(z, u, w) - \nabla h_i^v(z)\|^2\right] \\ &\leq \frac{Mv^2}{2q_1q_2}L^2(d+6)^3 + \frac{(2d+9)}{q_1q_2}Ml^2. \end{aligned} \quad (\text{A10})$$

Therefore, the proof is completed. $\square$

Lemma A3. Suppose Assumptions 1 and 2 are satisfied, the following holds:

$$\mathbb{E} \left[\left\| S_v(z, u', w')^\top S_v(z, u'', w'') \right\|^2 \right] \leq \left(\frac{Mv^2}{2q_1q_2} L^2(d+6)^3 + \frac{2d+9+q_1q_2}{q_1q_2} Ml^2 \right)^2. \quad (\text{A11})$$

Proof.

$$\begin{aligned} & \mathbb{E} \left[\left\| S_v(z, u', w')^\top S_v(z, u'', w'') \right\|^2 \right] \\ & \leq \mathbb{E} \left[\left\| S_v(z, u', w') \right\|^2 \left\| S_v(z, u'', w'') \right\|^2 \right] \\ & = \mathbb{E} \left[\left\| S_v(z, u', w') - \nabla H_v(z) + \nabla H_v(z) \right\|^2 \left\| S_v(z, u'', w'') - \nabla H_v(z) + \nabla H_v(z) \right\|^2 \right] \\ & = \mathbb{E} \left[\left\| S_v(z, u', w') - \nabla H_v(z) \right\|^2 + \left\| \nabla H_v(z) \right\|^2 \right] \mathbb{E} \left[\left\| S_v(z, u'', w'') - \nabla H_v(z) \right\|^2 \right. \\ & \quad \left. + \left\| \nabla H_v(z) \right\|^2 \right] \\ & \leq \left(\frac{Mv^2}{2q_1q_2} L^2(d+6)^3 + \frac{2d+9+q_1q_2}{q_1q_2} Ml^2 \right)^2, \end{aligned} \quad (\text{A12})$$

where the first inequality is due to the Cauchy–Schwarz inequality, and the last inequality follows from Lemma A2. $\square$

Lemma A4. Suppose Assumptions 1 and 2 are satisfied, the following holds:

$$\begin{aligned} & -\mathbb{E}[\langle \nabla H_v(z_t) \lambda, \nabla H(z_t) \lambda_t \rangle] \\ & \leq \frac{1}{2\beta_t} \mathbb{E} \left[\left\| \lambda_t - \lambda \right\|^2 - \left\| \lambda_{t+1} - \lambda \right\|^2 \right] + \frac{\beta_t}{2} \mathbb{E} \left[\left\| S_v(z_t, u'_t, w'_t)^\top S_v(z_t, u''_t, w''_t) \right\|^2 \right] \\ & \quad - \mathbb{E} \left[\left\| \nabla H_v(z_t) \lambda_t \right\|^2 \right]. \end{aligned} \quad (\text{A13})$$

Proof. Following the updated form of weighting λ , we have the following:

$$\begin{aligned} & \mathbb{E} \left[\left\| \lambda_{t+1} - \lambda \right\|^2 \right] \\ & = \mathbb{E} \left[\left\| \Pi_{\Delta^M} \left(\lambda_t - \beta_t \left[S_v(z_t, u'_t, w'_t)^\top S_v(z_t, u''_t, w''_t) \lambda_t \right] \right) - \Pi_{\Delta^M}(\lambda) \right\|^2 \right] \\ & \leq \mathbb{E} \left[\left\| \lambda_t - \beta_t \left[S_v(z_t, u'_t, w'_t)^\top S_v(z_t, u''_t, w''_t) \lambda_t \right] - \lambda \right\|^2 \right] \\ & = \mathbb{E} \left[\left\| \lambda_t - \lambda \right\|^2 \right] + \beta_t^2 \mathbb{E} \left[\left\| S_v(z_t, u'_t, w'_t)^\top S_v(z_t, u''_t, w''_t) \lambda_t \right\|^2 \right] \\ & \quad - 2\beta_t \mathbb{E} \left[\left\langle \lambda_t - \lambda, \nabla H_v(z_t)^\top \nabla H_v(z_t) \lambda_t \right\rangle \right] \\ & \leq \mathbb{E} \left[\left\| \lambda_t - \lambda \right\|^2 \right] + \beta_t^2 \mathbb{E} \left[\left\| S_v(z_t, u'_t, w'_t)^\top S_v(z_t, u''_t, w''_t) \right\|^2 \left\| \lambda_t \right\|^2 \right] \\ & \quad - 2\beta_t \mathbb{E} \left[\left\langle \lambda_t - \lambda, \nabla H_v(z_t)^\top \nabla H_v(z_t) \right\rangle \right] \\ & \leq \mathbb{E} \left[\left\| \lambda_t - \lambda \right\|^2 \right] + \beta_t^2 \mathbb{E} \left[\left\| S_v(z_t, u'_t, w'_t)^\top S_v(z_t, u''_t, w''_t) \right\|^2 \right] \\ & \quad - 2\beta_t \mathbb{E} \left[\left\langle \lambda_t - \lambda, \nabla H_v(z_t)^\top \nabla H_v(z_t) \right\rangle \right], \end{aligned} \quad (\text{A14})$$

where the first inequality is due to the non-expansive property of projection onto the convex set, the second inequality results from the Cauchy–Schwarz inequality, and the last inequality follows $\left\| \lambda_t \right\|^2 \leq 1$.

For the last term on the right side, we have the following:

$$\begin{aligned}
 & -2\beta_t \mathbb{E} \left[\left\langle \lambda_t - \lambda, \nabla H_v(z_t)^\top \nabla H_v(z_t) \lambda_t \right\rangle \right] \\
 & = -2\beta_t \mathbb{E} \left[\langle \nabla H_v(z_t)(\lambda_t - \lambda), \nabla H_v(z_t) \lambda_t \rangle \right] \\
 & = -2\beta_t \mathbb{E} \left[\|\nabla H_v(z_t) \lambda_t\|^2 \right] + 2\beta_t \mathbb{E} \left[\langle \nabla H_v(z_t) \lambda, \nabla H_v(z_t) \lambda_t \rangle \right]. \quad (\text{A15})
 \end{aligned}$$

Plugging Equation (A15) into Equation (A14), and rearranging, the following holds:

$$\begin{aligned}
 & -\mathbb{E} \left[\langle \nabla H_v(z_t) \lambda, \nabla H_v(z_t) \lambda_t \rangle \right] \\
 & \leq \frac{1}{2\beta_t} \mathbb{E} \left[\|\lambda_t - \lambda\|^2 - \|\lambda_{t+1} - \lambda\|^2 \right] + \frac{\beta_t}{2} \mathbb{E} \left[\left\| S_v(z_t, u'_t, w'_t)^\top S_v(z_t, u''_t, w''_t) \right\|^2 \right] \\
 & \quad - \mathbb{E} \left[\|\nabla H_v(z_t) \lambda_t\|^2 \right]. \quad (\text{A16})
 \end{aligned}$$

Hence, the result of this lemma is obtained. $\square$

Lemma A5. Suppose Assumptions 1 and 2 are satisfied and $f(\lambda) = \|\nabla H(z_t) \lambda\|^2$ is a convex function. Setting $\lambda_t^* = \arg \min_{\lambda} \|\nabla H(z_t) \lambda\|^2$, learning rate $\beta_{t,k} = \frac{c}{\sqrt{k}}$, where $c \geq 0$ is a constant. Then for $K \geq 1$, the following holds:

$$\begin{aligned}
 f(\lambda_{t,K}) - f(\lambda_t^*) & \leq \frac{\log(K) + 2}{\sqrt{K}} \left(\frac{1}{c} + c \left(\frac{Mv^2}{2q_1q_2} L^2(d+6)^3 + \frac{2d+9+q_1q_2}{q_1q_2} Ml^2 \right)^2 \right) \\
 & \quad + \sqrt{2}vMLl(d+3)^{3/2}(\log(K) + 2). \quad (\text{A17})
 \end{aligned}$$

Proof. Following the updated form of weighting λ , we have the following:

$$\begin{aligned}
 & \mathbb{E} \left[\|\lambda_{t,k+1} - \lambda\|^2 \right] \\
 & = \mathbb{E} \left[\left\| \Pi_{\Delta^M}(\lambda_{t,k} - \beta_{t,k} S_v(z_t, u'_t, w'_t)^\top S_v(z_t, u''_t, w''_t) \lambda_{t,k}) - \Pi_{\Delta^M}(\lambda) \right\|^2 \right] \\
 & \leq \mathbb{E} \left[\left\| \lambda_{t,k} - \beta_{t,k} S_v(z_t, u'_t, w'_t)^\top S_v(z_t, u''_t, w''_t) \lambda_{t,k} - \lambda \right\|^2 \right] \\
 & = \mathbb{E} \left[\|\lambda_{t,k} - \lambda\|^2 \right] + \beta_{t,k}^2 \mathbb{E} \left[\left\| S_v(z_t, u'_t, w'_t)^\top S_v(z_t, u''_t, w''_t) \lambda_{t,k} \right\|^2 \right] \\
 & \quad - 2\beta_{t,k} \mathbb{E} \left[\left\langle \lambda_{t,k} - \lambda, \nabla H_v(z_t)^\top \nabla H_v(z_t) \lambda_{t,k} \right\rangle \right], \quad (\text{A18})
 \end{aligned}$$

where the inequality is due to the non-expansive property of projection onto the convex set.

Rearranging the above inequality, the following holds:

$$\begin{aligned}
 & \mathbb{E} \left[\left\langle \lambda_{t,k} - \lambda, \nabla H(z_t)^\top \nabla H(z_t) \lambda_{t,k} \right\rangle \right] \\
 & \leq \mathbb{E} \left[\left\langle \lambda_{t,k} - \lambda, \nabla H(z_t)^\top \nabla H(z_t) \lambda_{t,k} - \nabla H_v(z_t)^\top \nabla H_v(z_t) \lambda_{t,k} \right\rangle \right] \\
 & \quad + \frac{1}{2\beta_{t,k}} \mathbb{E} \left[\|\lambda_{t,k} - \lambda\|^2 - \|\lambda_{t,k+1} - \lambda\|^2 \right] \\
 & \quad + \frac{\beta_{t,k}}{2} \mathbb{E} \left[\left\| S_v(z_t, u'_t, w'_t)^\top S_v(z_t, u''_t, w''_t) \lambda_{t,k} \right\|^2 \right] \\
 & \leq \mathbb{E} \left[\|\lambda_{t,k} - \lambda\| \left\| \nabla H(z_t)^\top \nabla H(z_t) - \nabla H_v(z_t)^\top \nabla H_v(z_t) \right\| \|\lambda_{t,k}\| \right] \\
 & \quad + \frac{1}{2\beta_{t,k}} \mathbb{E} \left[\|\lambda_{t,k} - \lambda\|^2 - \|\lambda_{t,k+1} - \lambda\|^2 \right] \\
 & \quad + \frac{\beta_{t,k}}{2} \mathbb{E} \left[\left\| S_v(z_t, u'_t, w'_t)^\top S_v(z_t, u''_t, w''_t) \right\|^2 \|\lambda_{t,k}\|^2 \right] \\
 & \leq \underbrace{\sqrt{2} \mathbb{E} \left[\left\| \nabla H(z_t)^\top \nabla H(z_t) - \nabla H_v(z_t)^\top \nabla H_v(z_t) \right\| \right]}_{I_1} \\
 & \quad + \frac{1}{2\beta_{t,k}} \mathbb{E} \left[\|\lambda_{t,k} - \lambda\|^2 - \|\lambda_{t,k+1} - \lambda\|^2 \right] \\
 & \quad + \underbrace{\frac{\beta_{t,k}}{2} \mathbb{E} \left[\left\| S_v(z_t, u'_t, w'_t)^\top S_v(z_t, u''_t, w''_t) \right\|^2 \right]}_{I_2}, \tag{A19}
 \end{aligned}$$

where the second inequality is due to the Cauchy-Schwartz inequality and the last inequality follows from the $\|\lambda_{t,k} - \lambda\| \leq \sqrt{2}$ and $\|\lambda_{t,k}\|^2 \leq 1$.

Following the property of a convex function, $f(\lambda_1) \geq f(\lambda_2) + \langle \nabla f(\lambda_2), \lambda_1 - \lambda_2 \rangle$, the following holds:

$$\mathbb{E}[f(\lambda_{t,k}) - f(\lambda)] \leq \frac{1}{2\beta_{t,k}} \mathbb{E} \left[\|\lambda_{t,k} - \lambda\|^2 - \|\lambda_{t,k+1} - \lambda\|^2 \right] + \frac{\beta_{t,k}}{2} I_2 + I_1. \tag{A20}$$

Letting r be an element in $\{1, \dots, K-1\}$ and summing over $k = K-r, \dots, K$, the following holds:

$$\begin{aligned}
 \sum_{k=K-r}^K \mathbb{E}[f(\lambda_{t,k}) - f(\lambda)] & \leq \frac{1}{2\beta_{t,K-r}} \mathbb{E} \left[\|\lambda_{t,K-r} - \lambda\|^2 \right] + \sum_{k=K-r}^K I_1 + \frac{I_2}{2} \sum_{k=K-r}^K \beta_{t,k} \\
 & \quad + \sum_{k=K-r+1}^K \mathbb{E} \left[\frac{\|\lambda_{t,k} - \lambda\|^2}{2} \left(\frac{1}{\beta_{t,k}} - \frac{1}{\beta_{t,k-1}} \right) \right]. \tag{A21}
 \end{aligned}$$

Setting $\beta_{t,k} = \frac{c}{\sqrt{k}}$, $\lambda = \lambda_{t,K-r}$, we have the following:

$$\begin{aligned}
 \sum_{k=K-r}^K \mathbb{E}[f(\lambda_{t,k}) - f(\lambda_{K-r})] & \leq \sum_{k=K-r}^K I_1 + \frac{I_2 c}{2} \sum_{k=K-r}^K \frac{1}{\sqrt{k}} + \frac{1}{c} (\sqrt{K} - \sqrt{K-r}) \\
 & \leq I_1(r+1) + \left(\frac{1}{c} + I_2 c \right) (\sqrt{K} - \sqrt{K-r-1}) \\
 & = I_1(r+1) + \left(\frac{1}{c} + I_2 c \right) \frac{r+1}{\sqrt{K} + \sqrt{K-r-1}} \\
 & \leq \left(\frac{1}{c} + I_2 c \right) \frac{r+1}{\sqrt{K}} + I_1(r+1), \tag{A22}
 \end{aligned}$$

where the first inequality is due to $\|\lambda_{t,k} - \lambda_{t,K-r}\|^2 \leq 2$, the second inequality follows $\sum_{k=K-r}^K \frac{1}{\sqrt{k}} \leq 2(\sqrt{K} - \sqrt{K-r-1})$.

Letting $L_r = \frac{1}{r+1} \sum_{k=K-r}^K \mathbb{E}[f(\lambda_{t,k})]$, the following holds:

$$-\mathbb{E}[f(\lambda_{t,K-r})] \leq -L_r + \left(\frac{1}{c} + I_2 c\right) \frac{1}{\sqrt{K}} + I_1. \quad (\text{A23})$$

Following the definition of L_r , we have the following:

$$\begin{aligned} rL_{r-1} &= (r+1)L_r - \mathbb{E}[f(\lambda_{t,K-r})] \\ &\leq (r+1)L_r - L_r + \left(\frac{1}{c} + I_2 c\right) \frac{1}{\sqrt{K}} + I_1, \end{aligned} \quad (\text{A24})$$

and dividing by r , the following holds:

$$L_{r-1} \leq L_r + \left(\frac{1}{c} + I_2 c\right) \frac{1}{r\sqrt{K}} + \frac{I_1}{r}. \quad (\text{A25})$$

Summing over $r = 1, \dots, K-1$, the following holds:

$$\begin{aligned} L_0 = \mathbb{E}[f(\lambda_{t,K})] &\leq L_{K-1} + \left(\left(\frac{1}{c} + I_2 c\right) \frac{1}{\sqrt{K}} + I_1\right) \sum_{r=1}^{K-1} \frac{1}{r} \\ &\leq L_{K-1} + \left(\left(\frac{1}{c} + I_2 c\right) \frac{1}{\sqrt{K}} + I_1\right) (\log(K) + 1), \end{aligned} \quad (\text{A26})$$

where the last inequality is due to $\sum_{r=1}^{K-1} \frac{1}{r} \leq (1 + \log(K))$.

Now, using Equation (A21) with $r = K-1$, $\beta_{t,k} = \frac{c}{\sqrt{k}}$ and $\lambda = \lambda_t^*$, the following holds:

$$\begin{aligned} &\frac{1}{K} \sum_{k=1}^K \mathbb{E}[f(\lambda_{t,k}) - f(\lambda_t^*)] \\ &\leq \frac{1}{2K\beta_{t,1}} \mathbb{E}[\|\lambda_{t,1} - \lambda_t^*\|^2] + \sum_{k=2}^K \mathbb{E}\left[\frac{\|\lambda_{t,k} - \lambda_t^*\|^2}{2K} \left(\frac{1}{\beta_{t,k}} - \frac{1}{\beta_{t,k-1}}\right)\right] + I_1 + \frac{I_2}{2K} \sum_{k=1}^K \beta_{t,k} \\ &\leq \frac{1}{c\sqrt{K}} + \frac{cI_2}{\sqrt{K}} + I_1. \end{aligned} \quad (\text{A27})$$

It is easy to know the following:

$$\begin{aligned} L_{K-1} - f(\lambda_t^*) &= \frac{1}{K} \sum_{k=1}^K \mathbb{E}[f(\lambda_{t,k}) - f(\lambda_t^*)] \\ &\leq \frac{1}{c\sqrt{K}} + \frac{cI_2}{\sqrt{K}} + I_1. \end{aligned} \quad (\text{A28})$$

Plugging Equation (A28) into Equation (A26), the following holds:

$$f(\lambda_{t,K}) - f(\lambda_t^*) \leq \left(\left(\frac{1}{c} + I_2 c\right) \frac{1}{\sqrt{K}} + I_1\right) (\log(K) + 2). \quad (\text{A29})$$

For the term I_1 , we have the following:

$$\begin{aligned}
 & \sqrt{2}\mathbb{E}\left[\left\|\nabla H_v(z_t)^\top \nabla H_v(z_t) - \nabla H(z_t)^\top \nabla H(z_t)\right\|\right] \\
 &= \sqrt{2}\mathbb{E}\left[\left\|\nabla H_v(z_t)^\top \nabla H_v(z_t) - \nabla H_v(z_t)^\top \nabla H(z_t) + \nabla H_v(z_t)^\top \nabla H(z_t) \right. \right. \\
 &\quad \left. \left. - \nabla H(z_t)^\top \nabla H(z_t)\right\|\right] \\
 &\leq \sqrt{2}\left(\mathbb{E}\left[\left\|\nabla H_v(z_t)^\top \nabla H_v(z_t) - \nabla H_v(z_t)^\top \nabla H(z_t)\right\|\right] \right. \\
 &\quad \left. + \mathbb{E}\left[\left\|\nabla H_v(z_t)^\top \nabla H(z_t) - \nabla H(z_t)^\top \nabla H(z_t)\right\|\right]\right) \\
 &\leq \sqrt{2}\mathbb{E}[\|\nabla H_v(z_t) - \nabla H(z_t)\|(\|\nabla H_v(z_t)\| + \|\nabla H(z_t)\|)] \\
 &\leq \sqrt{2}\nu M L(d+3)^{3/2}, \tag{A30}
 \end{aligned}$$

where the first inequality is due to the triangle inequality, the second inequality follows the Cauchy–Schwartz inequality, and the last inequality is due to Lemma A2.

Following Lemma A3, for term I_2 , the following holds:

$$\mathbb{E}\left[\left\|S_v(z_t, u'_t, w'_t)^\top S_v(z_t, u''_t, w''_t)\right\|^2\right] \leq \left(\frac{M\nu^2}{2q_1q_2}L^2(d+6)^3 + \frac{2d+9+q_1q_2}{q_1q_2}Ml^2\right)^2. \tag{A31}$$

Plugging Equation (A30), Equation (A31) into (A29), the following holds:

$$\begin{aligned}
 f(\lambda_{t,K}) - f(\lambda_t^*) &\leq \frac{\log(K) + 2}{\sqrt{K}} \left(\frac{1}{c} + c \left(\frac{M\nu^2}{2q_1q_2}L^2(d+6)^3 + \frac{2d+9+q_1q_2}{q_1q_2}Ml^2 \right)^2 \right) \\
 &\quad + \sqrt{2}\nu M L(d+3)^{3/2}(\log(K) + 2). \tag{A32}
 \end{aligned}$$

Thus, this lemma is proved. $\square$

Lemma A6. Suppose Assumptions 1 and 2 are satisfied. For any $\lambda \in \triangle^M$, the following holds:

$$\|\nabla H(z_t)\lambda\|^2 \leq 2\|\nabla H_v(z_t)\lambda\|^2 + \frac{M\nu^2}{2}L^2(d+3)^3. \tag{A33}$$

Proof.

$$\begin{aligned}
 \|\nabla H(z_t)\lambda - \nabla H_v(z_t)\lambda\|^2 &\leq \|\nabla H(z_t) - \nabla H_v(z_t)\|^2 \|\lambda\|^2 \\
 &\leq \|\nabla H(z_t) - \nabla H_v(z_t)\|^2 \\
 &\leq \frac{M\nu^2}{4}L^2(d+3)^3. \tag{A34}
 \end{aligned}$$

where the first inequality follows the Cauchy–Schwartz inequality, the second inequality is due to $\|\lambda\|^2 \leq 1$, and the last inequality is due to Lemma A2.

For the lower bound of the left equation, we have the following:

$$\begin{aligned}
 & \|\nabla H(z_t)\lambda - \nabla H_v(z_t)\lambda\|^2 \\
 &= \|\nabla H(z_t)\lambda\|^2 + \|\nabla H_v(z_t)\lambda\|^2 - 2\langle \nabla H(z_t)\lambda, \nabla H_v(z_t)\lambda \rangle \\
 &\geq \frac{1}{2}\|\nabla H(z_t)\lambda\|^2 - \|\nabla H_v(z_t)\lambda\|^2, \tag{A35}
 \end{aligned}$$

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

Plugging Equation (A35) into Equation (A34), the following holds:

$$\|\nabla H(z_t)\lambda\|^2 \leq 2\|\nabla H_v(z_t)\lambda\|^2 + \frac{Mv^2}{2}L^2(d+3)^3. \quad (\text{A36})$$

Therefore, the result of this lemma is proved completely. $\square$

Appendix B. Proof of the Results in Section 5

Proof of Proposition 1. Based on the definitions of $S_v(z_t, u_t, w_t)$, we have the following:

$$\begin{aligned} & \mathbb{E} \left[\|\mathbb{E}[S_v(z_t, u_t, w_t)\lambda_{t,K}] - \nabla H(z_t)\lambda_t^*\|^2 | z_t \right] \\ &= \mathbb{E} \left[\|\nabla H_v(z_t)\lambda_{t,K} - \nabla H(z_t)\lambda_t^*\|^2 \right] \\ &= \mathbb{E} \left[\|\nabla H_v(z_t)\lambda_{t,K} - \nabla H(z_t)\lambda_{t,K} + \nabla H(z_t)\lambda_{t,K} - \nabla H(z_t)\lambda_t^*\|^2 \right] \\ &\leq 2 \underbrace{\mathbb{E} \left[\|\nabla H_v(z_t)\lambda_{t,K} - \nabla H(z_t)\lambda_{t,K}\|^2 \right]}_{I_1} + 2 \underbrace{\mathbb{E} \left[\|\nabla H(z_t)\lambda_{t,K} - \nabla H(z_t)\lambda_t^*\|^2 \right]}_{I_2}, \end{aligned} \quad (\text{A37})$$

where the inequality is due to Young's inequality.

For the term I_1 , we have the following:

$$\begin{aligned} \mathbb{E} \left[\|\nabla H_v(z_t)\lambda_{t,K} - \nabla H(z_t)\lambda_{t,K}\|^2 \right] &\leq \mathbb{E} \left[\|\nabla H_v(z_t) - \nabla H(z_t)\|^2 \|\lambda_{t,K}\|^2 \right] \\ &\leq \mathbb{E} \left[\|\nabla H_v(z_t) - \nabla H(z_t)\|^2 \right] \\ &\leq \frac{Mv^2}{4}L^2(d+3)^3, \end{aligned} \quad (\text{A38})$$

where the first inequality follows the Cauchy–Schwartz inequality, the second inequality is due to $\|\lambda_{t,K}\|^2 \leq 1$, and the last inequality follows Lemma A2.

For the term I_2 , we have the following:

$$\begin{aligned} & \mathbb{E} \left[\|\nabla H(z_t)\lambda_{t,K} - \nabla H(z_t)\lambda_t^*\|^2 \right] \\ &= \mathbb{E} \left[\|\nabla H(z_t)\lambda_{t,K}\|^2 + \|\nabla H(z_t)\lambda_t^*\|^2 - 2\langle \nabla H(z_t)\lambda_{t,K}, \nabla H(z_t)\lambda_t^* \rangle \right] \\ &\leq \mathbb{E} \left[\|\nabla H(z_t)\lambda_{t,K}\|^2 - \|\nabla H(z_t)\lambda_t^*\|^2 \right], \end{aligned} \quad (\text{A39})$$

where the inequality is due to the optimality condition that

$$\langle \nabla H(z_t)\lambda_{t,K}, \nabla H(z_t)\lambda_t^* \rangle \geq \|\nabla H(z_t)\lambda_t^*\|^2. \quad (\text{A40})$$

Following Lemma A5, the following holds:

$$\begin{aligned} & \mathbb{E} \left[\|\nabla H(z_t)\lambda_{t,K}\|^2 - \|\nabla H(z_t)\lambda_t^*\|^2 \right] \\ &\leq \frac{\log(K) + 2}{\sqrt{K}} \left(\frac{1}{c} + c \left(\frac{Mv^2}{2q_1q_2}L^2(d+6)^3 + \frac{2d+9+q_1q_2}{q_1q_2}Ml^2 \right)^2 \right) \\ &\quad + \sqrt{2}vMLl(d+3)^{3/2}(\log(K) + 2). \end{aligned} \quad (\text{A41})$$

Plugging I_1, I_2 into Equation (A37), the following holds:

$$\begin{aligned} & \mathbb{E} \left[\left\| \mathbb{E}[S_v(z_t, u_t, w_t) \lambda_{t,K}] - \nabla H(z_t) \lambda_t^* \right\|^2 | z_t \right] \\ & \leq \frac{Mv^2}{4} L^2 (d+3)^3 + \frac{\log(K) + 2}{\sqrt{K}} \left(\frac{1}{c} + c \left(\frac{Mv^2}{2q_1q_2} L^2 (d+6)^3 + \frac{2d+9+q_1q_2}{q_1q_2} ML^2 \right)^2 \right) \\ & \quad + \sqrt{2} v M L l (d+3)^{3/2} (\log(K) + 2). \end{aligned} \quad (\text{A42})$$

Therefore, Proposition 1 is proved completely. $\square$

Proof of Theorem 1. Following Lemma A1, it is known that $h_i^v(z)$ is L -smooth

$$\begin{aligned} h_i^v(z_{t+1}) & \leq h_i^v(z_t) + \langle \nabla h_i^v(z_t), z_{t+1} - z_t \rangle + \frac{L}{2} \|z_{t+1} - z_t\|^2 \\ & = h_i^v(z_t) + \eta_t \langle \nabla h_i^v(z_t), -S_v(z_t, u_t, w_t) \lambda_{t,K} \rangle + \frac{L\eta_t^2}{2} \|S_v(z_t, u_t, w_t) \lambda_{t,K}\|^2 \\ & \leq h_i^v(z_t) + \eta_t \langle \nabla h_i^v(z_t), -S_v(z_t, u_t, w_t) \lambda_{t,K} \rangle + \frac{L\eta_t^2}{2} \|S_v(z_t, u_t, w_t)\|^2 \|\lambda_{t,K}\|^2 \\ & \leq h_i^v(z_t) + \eta_t \langle \nabla h_i^v(z_t), -S_v(z_t, u_t, w_t) \lambda_{t,K} \rangle + \frac{L\eta_t^2}{2} \|S_v(z_t, u_t, w_t)\|^2, \end{aligned} \quad (\text{A43})$$

where the second inequality follows the Cauchy–Schwarz inequality, and the last inequality is due to $\|\lambda_{t,K}\|^2 \leq 1$.

When taking the expectation with respect to the condition z_t , the following holds:

$$\begin{aligned} \mathbb{E}[h_i^v(z_{t+1})] & \leq \mathbb{E}[h_i^v(z_t)] + \eta_t \mathbb{E}[\langle \nabla h_i^v(z_t), -\nabla H_v(z_t) \lambda_{t,K} \rangle] + \frac{L\eta_t^2}{2} \mathbb{E}[\|S_v(z_t, u_t, w_t)\|^2] \\ & = \mathbb{E}[h_i^v(z_t)] + \eta_t \mathbb{E}[\langle \nabla h_i^v(z_t), -\nabla H_v(z_t) \lambda_{t,\nu}^* \rangle] + \frac{L\eta_t^2}{2} \mathbb{E}[\|S_v(z_t, u_t, w_t)\|^2] \\ & \quad + \eta_t \mathbb{E}[\langle \nabla h_i^v(z_t), -\nabla H_v(z_t) \lambda_{t,K} + \nabla H_v(z_t) \lambda_{t,\nu}^* \rangle] \\ & \leq \mathbb{E}[h_i^v(z_t)] - \eta_t \mathbb{E}[\|\nabla H_v(z_t) \lambda_{t,\nu}^*\|^2] + \frac{L\eta_t^2}{2} \mathbb{E}[\|S_v(z_t, u_t, w_t)\|^2] \\ & \quad + \eta_t \mathbb{E}[\langle \nabla h_i^v(z_t), -\nabla H_v(z_t) \lambda_{t,K} + \nabla H_v(z_t) \lambda_{t,\nu}^* \rangle] \\ & \leq \mathbb{E}[h_i^v(z_t)] - \eta_t \mathbb{E}[\|\nabla H_v(z_t) \lambda_{t,\nu}^*\|^2] + \frac{L\eta_t^2}{2} \mathbb{E}[\|S_v(z_t, u_t, w_t)\|^2] \\ & \quad + \eta_t \mathbb{E}[\|\nabla h_i^v(z_t)\| \|\nabla H_v(z_t) \lambda_{t,K} - \nabla H_v(z_t) \lambda_{t,\nu}^*\|] \\ & \leq \mathbb{E}[h_i^v(z_t)] - \eta_t \mathbb{E}[\|\nabla H_v(z_t) \lambda_{t,\nu}^*\|^2] + \frac{L\eta_t^2}{2} \mathbb{E}[\|S_v(z_t, u_t, w_t)\|^2] \\ & \quad + \underbrace{\eta_t l \mathbb{E}[\|\nabla H_v(z_t) \lambda_{t,K} - \nabla H_v(z_t) \lambda_{t,\nu}^*\|]}_{I_1}, \end{aligned} \quad (\text{A44})$$

where the second inequality follows the optimality condition given by $\langle \nabla H_v(z_t) \lambda, \nabla H_v(z_t) \lambda_{t,\nu}^* \rangle \geq \|\nabla H_v(z_t) \lambda_{t,\nu}^*\|^2$ for any $\lambda \in \Delta^M$, as stated in [18]. The third inequality is due to the Cauchy–Schwarz inequality, and the last inequality follows from the l -Lipschitz continuous of function $h_i^v(z)$.

For the term I_1 , following Jensen's inequality, the following holds:

$$\begin{aligned} \mathbb{E}[\|\nabla H_v(z_t) \lambda_{t,K} - \nabla H_v(z_t) \lambda_{t,\nu}^*\|] & \leq \sqrt{\mathbb{E}[\|\nabla H_v(z_t) \lambda_{t,K} - \nabla H_v(z_t) \lambda_{t,\nu}^*\|^2]} \\ & \leq \sqrt{\mathbb{E}[\|\nabla H_v(z_t) \lambda_{t,K}\|^2 - \|\nabla H_v(z_t) \lambda_{t,\nu}^*\|^2]} \end{aligned} \quad (\text{A45})$$

where the last inequality is due to the optimality condition $\langle \nabla H_v(z_t)\lambda, \nabla H_v(z_t)\lambda_{t,v}^* \rangle \geq \|\nabla H_v(z_t)\lambda_{t,v}^*\|^2$.

Due to the fact that $f(\lambda) = \|\nabla H_v(z_t)\lambda\|^2$ is a convex function and $\sup_{\lambda, \lambda' \in \Delta^M} \|\lambda - \lambda'\|^2 \leq 2$, we set $\lambda_{t,v}^* = \arg \min_{\lambda} \|\nabla H_v(z_t)\lambda\|^2$, and the learning rate of weighting λ is $\beta_{t,k} = \frac{c}{\sqrt{k}}$ where c is a constant. Following Theorem 2 in [48] and Lemma A3, the following holds:

$$\begin{aligned} & \mathbb{E} \left[\|\nabla H_v(z_t)\lambda_{t,K}\|^2 - \|\nabla H_v(z_t)\lambda_{t,v}^*\|^2 \right] \\ & \leq \underbrace{\left(\frac{2}{c} + c \left(\frac{Mv^2}{2q_1q_2} L^2(d+6)^3 + \frac{2d+9+q_1q_2}{q_1q_2} Ml^2 \right)^2 \right)}_V \frac{2 + \log(K)}{\sqrt{K}}. \end{aligned} \quad (\text{A46})$$

Plugging Equations (A46) and (A45) into Equation (A44) and rearranging Equation (A44), the following holds:

$$\begin{aligned} & \mathbb{E} \left[\|\nabla H_v(z_t)\lambda_{t,v}^*\|^2 \right] \\ & \leq \frac{\mathbb{E}[h_i^v(z_t) - h_i^v(z_{t+1})]}{\eta_t} + l\sqrt{V} + \frac{L\eta_t}{2} \mathbb{E} \left[\|S_v(z_t, u_t, w_t)\|^2 \right] \\ & = \frac{\mathbb{E}[h_i^v(z_t) - h_i^v(z_{t+1})]}{\eta_t} + l\sqrt{V} + \frac{L\eta_t}{2} \mathbb{E} \left[\|S_v(z_t, u_t, w_t) - \nabla H_v(z_t) + \nabla H_v(z_t)\|^2 \right] \\ & = \frac{\mathbb{E}[h_i^v(z_t) - h_i^v(z_{t+1})]}{\eta_t} + l\sqrt{V} + \frac{L\eta_t}{2} \mathbb{E} \left[\|S_v(z_t, u_t, w_t) - \nabla H_v(z_t)\|^2 + \|\nabla H_v(z_t)\|^2 \right] \\ & \leq \frac{\mathbb{E}[h_i^v(z_t) - h_i^v(z_{t+1})]}{\eta_t} + l\sqrt{V} + \frac{L\eta_t}{2} \left(\frac{Mv^2}{2q_1q_2} L^2(d+6)^3 + \frac{2d+9+q_1q_2}{q_1q_2} Ml^2 \right), \end{aligned} \quad (\text{A47})$$

where the last inequality is due to Lemma A2.

Following Lemma A6, the following holds:

$$\begin{aligned} \mathbb{E} \left[\|\nabla H(z_t)\lambda_{t,v}^*\|^2 \right] & \leq \frac{2\mathbb{E}[h_i^v(z_t) - h_i^v(z_{t+1})]}{\eta_t} + 2l\sqrt{V} + \frac{Mv^2}{2} L^2(d+3)^3 \\ & \quad + L\eta_t \left(\frac{Mv^2}{2q_1q_2} L^2(d+6)^3 + \frac{2d+9+q_1q_2}{q_1q_2} Ml^2 \right). \end{aligned} \quad (\text{A48})$$

Due to $\|\nabla H(z_t)\lambda_t^*\|^2 \leq \|\nabla H(z_t)\lambda_{t,v}^*\|^2$, we have the following:

$$\begin{aligned} \mathbb{E} \left[\|\nabla H(z_t)\lambda_t^*\|^2 \right] & \leq \frac{2\mathbb{E}[h_i^v(z_t) - h_i^v(z_{t+1})]}{\eta_t} + 2l\sqrt{V} + \frac{Mv^2}{2} L^2(d+3)^3 \\ & \quad + L\eta_t \left(\frac{Mv^2}{2q_1q_2} L^2(d+6)^3 + \frac{2d+9+q_1q_2}{q_1q_2} Ml^2 \right). \end{aligned} \quad (\text{A49})$$

Setting $\eta_t = \eta$ and summing over $t \in \{0, \dots, T-1\}$, the following holds:

$$\begin{aligned}
 & \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} [\|\nabla H(z_t) \lambda_t^*\|^2] \\
 & \leq \frac{2\mathbb{E}[h_i^v(z_0) - h_i^v(z_T)]}{T\eta} + L\eta \left(\frac{Mv^2}{2q_1q_2} L^2(d+6)^3 + \frac{2d+9+q_1q_2}{q_1q_2} Ml^2 \right) + \frac{Mv^2}{2} L^2(d+3)^3 \\
 & \quad + 2l \sqrt{\left(\frac{2}{c} + c \left(\frac{Mv^2}{2q_1q_2} L^2(d+6)^3 + \frac{2d+9+q_1q_2}{q_1q_2} Ml^2 \right)^2 \right) \frac{2+\log(K)}{\sqrt{K}}} \\
 & \leq \frac{2C_H}{T\eta} + L\eta \left(\frac{Mv^2}{2q_1q_2} L^2(d+6)^3 + \frac{2d+9+q_1q_2}{q_1q_2} Ml^2 \right) + \frac{Mv^2}{2} L^2(d+3)^3 \\
 & \quad + 4l \sqrt{\frac{1+\log(T)}{cT}} + 2l \sqrt{\frac{2c+2c\log(T)}{T}} \left(\frac{Mv^2}{2q_1q_2} L^2(d+6)^3 + \frac{2d+9+q_1q_2}{q_1q_2} Ml^2 \right). \tag{A50}
 \end{aligned}$$

Therefore, Theorem 1 is proved completely. $\square$

Proof of Theorem 2. Following the smoothness of $h_i^v(z_t)$, the following holds:

$$h_i^v(z_{t+1}) \leq h_i^v(z_t) + \eta_t \langle \nabla h_i^v(z_t), -S_v(z_t, u_t, w_t) \lambda_t \rangle + \frac{L}{2} \eta_t^2 \|S_v(z_t, u_t, w_t) \lambda_t\|^2. \tag{A51}$$

For any $\lambda \in \Delta^M$, we have the following:

$$H_v(z_{t+1}) \lambda \leq H_v(z_t) \lambda + \eta_t \langle \nabla H_v(z_t) \lambda, -S_v(z_t, u_t, w_t) \lambda_t \rangle + \frac{L}{2} \eta_t^2 \|S_v(z_t, u_t, w_t)\|^2. \tag{A52}$$

When taking the expectation with respect to the condition z_t , the following holds:

$$\begin{aligned}
 \mathbb{E}[H_v(z_{t+1}) \lambda] & \leq \mathbb{E}[H_v(z_t) \lambda] + \eta_t \mathbb{E}[\langle \nabla H_v(z_t) \lambda, -\nabla H_v(z_t) \lambda_t \rangle] + \frac{L}{2} \eta_t^2 \mathbb{E}[\|S_v(z_t, u_t, w_t)\|^2] \\
 & = \mathbb{E}[H_v(z_t) \lambda] + \eta_t \mathbb{E}[\langle \nabla H_v(z_t) \lambda, -\nabla H_v(z_t) \lambda_t \rangle] \\
 & \quad + \frac{L}{2} \eta_t^2 \mathbb{E}[\|S_v(z_t, u_t, w_t) - \nabla H_v(z_t) + \nabla H_v(z_t)\|^2] \\
 & = \mathbb{E}[H_v(z_t) \lambda] + \eta_t \mathbb{E}[\langle \nabla H_v(z_t) \lambda, -\nabla H_v(z_t) \lambda_t \rangle] \\
 & \quad + \frac{L}{2} \eta_t^2 \mathbb{E}[\|S_v(z_t, u_t, w_t) - \nabla H_v(z_t)\|^2 + \|\nabla H_v(z_t)\|^2] \\
 & \leq \mathbb{E}[H_v(z_t) \lambda] + \eta_t \mathbb{E}[\langle \nabla H_v(z_t) \lambda, -\nabla H_v(z_t) \lambda_t \rangle] \\
 & \quad + \frac{L}{2} \eta_t^2 \left(\frac{Mv^2}{2q_1q_2} L^2(d+6)^3 + \frac{2d+9+q_1q_2}{q_1q_2} Ml^2 \right), \tag{A53}
 \end{aligned}$$

where the last inequality follows Lemma A2.

For the second term on the right side, following Lemma A4, the following holds:

$$\begin{aligned}
 & -\eta_t \mathbb{E}[\langle \nabla H_v(z_t) \lambda, \nabla H(z_t) \lambda_t \rangle] \\
 & \leq \frac{\eta_t}{2\beta_t} \mathbb{E}[\|\lambda_t - \lambda\|^2 - \|\lambda_{t+1} - \lambda\|^2] + \frac{\eta_t \beta_t}{2} \mathbb{E}[\|S_v(z_t, u'_t, w'_t)^\top S_v(z_t, u''_t, w''_t)\|^2] \\
 & \quad - \eta_t \mathbb{E}[\|\nabla H_v(z_t) \lambda_t\|^2]. \tag{A54}
 \end{aligned}$$

Plugging Equation (A54) into Equation (A53), and rearranging, the following holds:

$$\begin{aligned}
 \mathbb{E} \left[\|\nabla H_v(z_t) \lambda_t\|^2 \right] &\leq \frac{1}{\eta_t} \mathbb{E}[H_v(z_t) \lambda - H_v(z_{t+1}) \lambda] + \frac{1}{2\beta_t} \mathbb{E} \left[\|\lambda_t - \lambda\|^2 - \|\lambda_{t+1} - \lambda\|^2 \right] \\
 &\quad + \frac{\beta_t}{2} \mathbb{E} \left[\left\| S_v(z_t, u'_t, w'_t)^\top S_v(z_t, u''_t, w''_t) \right\|^2 \right] \\
 &\quad + \frac{L}{2} \eta_t \left(\frac{Mv^2}{2q_1q_2} L^2(d+6)^3 + \frac{2d+9+q_1q_2}{q_1q_2} Ml^2 \right) \\
 &\leq \frac{1}{\eta_t} \mathbb{E}[H_v(z_t) \lambda - H_v(z_{t+1}) \lambda] + \frac{1}{2\beta_t} \mathbb{E} \left[\|\lambda_t - \lambda\|^2 - \|\lambda_{t+1} - \lambda\|^2 \right] \\
 &\quad + \frac{\beta_t}{2} \left(\frac{Mv^2}{2q_1q_2} L^2(d+6)^3 + \frac{2d+9+q_1q_2}{q_1q_2} Ml^2 \right)^2 \\
 &\quad + \frac{L}{2} \eta_t \left(\frac{Mv^2}{2q_1q_2} L^2(d+6)^3 + \frac{2d+9+q_1q_2}{q_1q_2} Ml^2 \right), \tag{A55}
 \end{aligned}$$

where the last inequality is due to Lemma A3.

Following Lemma A4, the following holds:

$$\begin{aligned}
 \mathbb{E} \left[\|\nabla H(z_t) \lambda_t\|^2 \right] &\leq \frac{2}{\eta_t} \mathbb{E}[H_v(z_t) \lambda - H_v(z_{t+1}) \lambda] + \frac{1}{\beta_t} \mathbb{E} \left[\|\lambda_t - \lambda\|^2 - \|\lambda_{t+1} - \lambda\|^2 \right] \\
 &\quad + \frac{Mv^2}{2} L^2(d+3)^3 + \beta_t \left(\frac{Mv^2}{2q_1q_2} L^2(d+6)^3 + \frac{2d+9+q_1q_2}{q_1q_2} Ml^2 \right)^2 \\
 &\quad + \eta_t L \left(\frac{Mv^2}{2q_1q_2} L^2(d+6)^3 + \frac{2d+9+q_1q_2}{q_1q_2} Ml^2 \right). \tag{A56}
 \end{aligned}$$

Due to $\|\nabla H(z_t) \lambda_t^*\|^2 \leq \|\nabla H(z_t) \lambda_t\|^2$, we have the following:

$$\begin{aligned}
 \mathbb{E} \left[\|\nabla H(z_t) \lambda_t^*\|^2 \right] &\leq \frac{2}{\eta_t} \mathbb{E}[H_v(z_t) \lambda - H_v(z_{t+1}) \lambda] + \frac{1}{\beta_t} \mathbb{E} \left[\|\lambda_t - \lambda\|^2 - \|\lambda_{t+1} - \lambda\|^2 \right] \\
 &\quad + \frac{Mv^2}{2} L^2(d+3)^3 + \beta_t \left(\frac{Mv^2}{2q_1q_2} L^2(d+6)^3 + \frac{2d+9+q_1q_2}{q_1q_2} Ml^2 \right)^2 \\
 &\quad + \eta_t L \left(\frac{Mv^2}{2q_1q_2} L^2(d+6)^3 + \frac{2d+9+q_1q_2}{q_1q_2} Ml^2 \right). \tag{A57}
 \end{aligned}$$

Setting $\eta_t = \eta$, $\beta_t = \beta$ and summing over $t \in \{0, \dots, T-1\}$, the following holds:

$$\begin{aligned}
 &\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \left[\|\nabla H(z_t) \lambda_t^*\|^2 \right] \\
 &\leq \frac{2(H_v(z_0) \lambda - H_v(z_T) \lambda)}{T\eta} + \frac{\|\lambda_0 - \lambda\|^2 - \|\lambda_T - \lambda\|^2}{T\beta} + \frac{Mv^2}{2} L^2(d+3)^3 \\
 &\quad + \beta \left(\frac{Mv^2}{2q_1q_2} L^2(d+6)^3 + \frac{2d+9+q_1q_2}{q_1q_2} Ml^2 \right)^2 \\
 &\quad + \eta L \left(\frac{Mv^2}{2q_1q_2} L^2(d+6)^3 + \frac{2d+9+q_1q_2}{q_1q_2} Ml^2 \right) \\
 &\leq \frac{2C_H}{T\eta} + \frac{2}{T\beta} + \frac{Mv^2}{2} L^2(d+3)^3 + \beta \left(\frac{Mv^2}{2q_1q_2} L^2(d+6)^3 + \frac{2d+9+q_1q_2}{q_1q_2} Ml^2 \right)^2 \\
 &\quad + \eta L \left(\frac{Mv^2}{2q_1q_2} L^2(d+6)^3 + \frac{2d+9+q_1q_2}{q_1q_2} Ml^2 \right), \tag{A58}
 \end{aligned}$$

where the last inequality is due to Assumption 3 and $\sup_{\lambda, \lambda' \in \Delta^M} \|\lambda - \lambda'\|^2 \leq 2$. Thus, Theorem 2 is proved completely. $\square$

References

1. Lu, W.; Zhou, Y.; Wan, G.; Hou, S.; Song, S. L3-Net: Towards Learning Based LiDAR Localization for Autonomous Driving. In Proceedings of the 31th IEEE Conference on Computer Vision and Pattern Recognition, CVPR, Long Beach, CA, USA, 16–20 June 2019.
2. Fernandes, P.; Ghorbani, B.; Garcia, X.; Freitag, M.; Firat, O. Scaling Laws for Multilingual Neural Machine Translation. In Proceedings of the 40th International Conference on Machine Learning, ICML, Honolulu, HI, USA, 23–29 July 2023.
3. Satija, H.; Thomas, P.S.; Pineau, J.; Laroché, R. Multi-Objective SPIBB: Seldonian Offline Policy Improvement with Safety Constraints in Finite MDPs. In Proceedings of the 35th Advances in Neural Information Processing Systems, NeurIPS, Virtual, 6–14 December 2021.
4. Zhang, L.; Liu, X.; Mahmood, K.; Ding, C.; Guan, H. Dynamic Gradient Balancing for Enhanced Adversarial Attacks on Multi-Task Models. *CoRR* **2023**. [[CrossRef](#)]
5. Cui, Y.; Geng, Z.; Zhu, Q.; Han, Y. Review: Multi-objective optimization methods and application in energy saving. *Energy* **2017**, *125*, 681–704. [[CrossRef](#)]
6. Deb, K.; Agrawal, S.; Pratap, A.; Meyarivan, T. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans. Evol. Comput.* **2002**, *6*, 182–197. [[CrossRef](#)]
7. Coello, C.A.C.; Pulido, G.T.; Lechuga, M.S. Handling Multiple Objectives With Particle Swarm Optimization. *IEEE Trans. Evol. Comput.* **2004**, *8*, 256–279. [[CrossRef](#)]
8. Mirjalili, S.; Saremi, S.; Mirjalili, S.M.; dos Santos Coelho, L. Multi-objective grey wolf optimizer: A novel algorithm for multi-criterion optimization. *Expert Syst. Appl.* **2016**, *47*, 106–119. [[CrossRef](#)]
9. Bandyopadhyay, S.; Saha, S.; Maulik, U.; Deb, K. A Simulated Annealing-Based Multiobjective Optimization Algorithm: AMOSA. *IEEE Trans. Evol. Comput.* **2008**, *12*, 269–283. [[CrossRef](#)]
10. Jiang, S.; Chen, Q.; Pan, Y.; Xiang, Y.; Lin, Y.; Wu, X.; Liu, C.; Song, X. ZO-AdaMU Optimizer: Adapting Perturbation by the Momentum and Uncertainty in Zeroth-Order Optimization. In Proceedings of the 38th AAAI Conference on Artificial Intelligence, AAAI, Vancouver, BC, Canada, 20–27 February 2024. [[CrossRef](#)]
11. Gao, K.; Sener, O. Generalizing Gaussian Smoothing for Random Search. In Proceedings of the 39th International Conference on Machine Learning, ICML, Baltimore, MD, USA, 17–23 July 2022.
12. Ghadimi, S.; Lan, G. Stochastic First- and Zeroth-Order Methods for Nonconvex Stochastic Programming. *SIAM J. Optim.* **2013**, *23*, 2341–2368. [[CrossRef](#)]
13. Ye, F.; Lyu, Y.; Wang, X.; Zhang, Y.; Tsang, I.W. Adaptive Stochastic Gradient Algorithm for Black-box Multi-Objective Learning. In Proceedings of the 12th International Conference on Learning Representations, ICLR, Vienna, Austria, 7–11 May 2024.
14. Zerbinati, A.; Desideri, J.A.; Duvigneau, R. *Comparison Between MGDA and PAES for Multi-Objective Optimization*; Research Report RR-7667; INRIA: Le Chesnay-Rocquencourt, France, 2011.
15. Gass, S.; Saaty, T. The computational algorithm for the parametric objective function. *Nav. Res. Logist.* **1955**, *2*, 39–45. [[CrossRef](#)]
16. Haimes, Y. On a bicriterion formulation of the problems of integrated system identification and system optimization. *IEEE Trans. Syst. Man Cybern. Syst.* **1971**, *SMC-1*, 296–297. [[CrossRef](#)]
17. Chen, Z.; Ngiam, J.; Huang, Y.; Luong, T.; Kretschmar, H.; Chai, Y.; Anguelov, D. Just Pick a Sign: Optimizing Deep Multitask Models with Gradient Sign Dropout. In Proceedings of the 34th Advances in Neural Information Processing Systems, NeurIPS, Virtual, 6–12 December 2020.
18. Désidéri, J.A. Multiple-gradient descent algorithm (MGDA) for multiobjective optimization. *C. R. Math.* **2012**, *350*, 313–318. [[CrossRef](#)]
19. Lin, X.; Zhen, H.; Li, Z.; Zhang, Q.; Kwong, S. Pareto Multi-Task Learning. In Proceedings of the 33th Advances in Neural Information Processing Systems, NeurIPS, Vancouver, BC, Canada, 8–14 December 2019.
20. Mahapatra, D.; Rajan, V. Multi-Task Learning with User Preferences: Gradient Descent with Controlled Ascent in Pareto Optimization. In Proceedings of the 37th International Conference on Machine Learning, ICML, Virtual Event, 13–18 July 2020.
21. Mercier, Q.; Poirion, F.; Désidéri, J. A stochastic multiple gradient descent algorithm. *Eur. J. Oper. Res.* **2018**, *271*, 808–817. [[CrossRef](#)]
22. Liu, S.; Vicente, L.N. The stochastic multi-gradient algorithm for multi-objective optimization and its application to supervised machine learning. *Ann. Oper. Res.* **2024**, *339*, 1119–1148. [[CrossRef](#)]
23. Zhou, S.; Zhang, W.; Jiang, J.; Zhong, W.; Gu, J.; Zhu, W. On the Convergence of Stochastic Multi-Objective Gradient Manipulation and Beyond. In Proceedings of the 36th Advances in Neural Information Processing Systems, NeurIPS, New Orleans, LA, USA, 28 November–9 December 2022.

24. Xiao, P.; Ban, H.; Ji, K. Direction-oriented Multi-objective Learning: Simple and Provable Stochastic Algorithms. In Proceedings of the 37th Advances in Neural Information Processing Systems, NeurIPS, New Orleans, LA, USA, 10–16 December 2023.
25. Chen, R.; Li, Y.; Chai, T. Multi-Objective Derivative-Free Optimization Based on Hessian-Aware Gaussian Smoothing Method. In Proceedings of the 18th IEEE International Conference on Control & Automation, ICCA, Reykjavík, Iceland, 18–21 June 2024. [\[CrossRef\]](#)
26. Tammer, C.; Winkler, K. A new scalarization approach and applications in multicriteria d.c. optimization. *J. Nonlinear Convex Anal.* **2003**, *4*, 365–380.
27. Fliege, J.; Svaiter, B.F. Steepest descent methods for multicriteria optimization. *Math. Methods Oper. Res.* **2000**, *51*, 479–494. [\[CrossRef\]](#)
28. Bonnel, H.; Iusem, A.N.; Svaiter, B.F. Proximal Methods in Vector Optimization. *SIAM J. Optim.* **2005**, *15*, 953–970. [\[CrossRef\]](#)
29. Fliege, J.; Drummond, L.M.G.; Svaiter, B.F. Newton’s Method for Multiobjective Optimization. *SIAM J. Optim.* **2009**, *20*, 602–626. [\[CrossRef\]](#)
30. Carrizo, G.A.; Lotito, P.A.; Maciel, M.C. Trust region globalization strategy for the nonconvex unconstrained multiobjective optimization problem. *Math. Program.* **2016**, *159*, 339–369. [\[CrossRef\]](#)
31. Pérez, L.R.L.; da Fonseca Prudente, L. Nonlinear Conjugate Gradient Methods for Vector Optimization. *SIAM J. Optim.* **2018**, *28*, 2690–2720. [\[CrossRef\]](#)
32. Liu, L.; Li, Y.; Kuang, Z.; Xue, J.; Chen, Y.; Yang, W.; Liao, Q.; Zhang, W. Towards Impartial Multi-task Learning. In Proceedings of the 9th International Conference on Learning Representations, ICLR, Virtual Event, 3–7 May 2021.
33. Fernando, H.D.; Shen, H.; Liu, M.; Chaudhury, S.; Murugesan, K.; Chen, T. Mitigating Gradient Bias in Multi-objective Learning: A Provably Convergent Approach. In Proceedings of the 11th International Conference on Learning Representations, ICLR, Kigali, Rwanda, 1–5 May 2023.
34. Chen, L.; Fernando, H.D.; Ying, Y.; Chen, T. Three-Way Trade-Off in Multi-Objective Learning: Optimization, Generalization and Conflict-Avoidance. *J. Mach. Learn. Res.* **2024**, *25*, 193:1–193:53.
35. Fernando, H.D.; Chen, L.; Lu, S.; Chen, P.; Liu, M.; Chaudhury, S.; Murugesan, K.; Liu, G.; Wang, M.; Chen, T. Variance Reduction Can Improve Trade-Off in Multi-Objective Learning. In Proceedings of the 49th IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, Seoul, Republic of Korea, 14–19 April 2024. [\[CrossRef\]](#)
36. Laumanns, M.; Ocenasek, J. Bayesian Optimization Algorithms for Multi-objective Optimization. In Proceedings of the 7th Parallel Problem Solving from Nature, PPSN, Granada, Spain, 7–11 September 2002. [\[CrossRef\]](#)
37. Belakaria, S.; Deshwal, A.; Jayakodi, N.K.; Doppa, J.R. Uncertainty-Aware Search Framework for Multi-Objective Bayesian Optimization. In Proceedings of the 34th AAAI Conference on Artificial Intelligence, AAAI, New York, NY, USA, 7–12 February 2020. [\[CrossRef\]](#)
38. Zhang, Q.; Li, H. MOEA/D: A Multiobjective Evolutionary Algorithm Based on Decomposition. *IEEE Trans. Evol. Comput.* **2007**, *11*, 712–731. [\[CrossRef\]](#)
39. Fliege, J.; Vaz, A.I.F.; Vicente, L.N. Complexity of gradient descent for multiobjective optimization. *Optim. Methods Softw.* **2019**, *34*, 949–959. [\[CrossRef\]](#)
40. Liu, S.; Chen, P.; Kaikhura, B.; Zhang, G.; III, A.O.H.; Varshney, P.K. A Primer on Zeroth-Order Optimization in Signal Processing and Machine Learning: Principals, Recent Advances, and Applications. *IEEE Signal Process. Mag.* **2020**, *37*, 43–54. [\[CrossRef\]](#)
41. Berahas, A.S.; Cao, L.; Choromanski, K.; Scheinberg, K. A Theoretical and Empirical Comparison of Gradient Approximations in Derivative-Free Optimization. *Found. Comput. Math.* **2022**, *22*, 507–560. [\[CrossRef\]](#)
42. Balasubramanian, K.; Ghadimi, S. Zeroth-Order Nonconvex Stochastic Optimization: Handling Constraints, High Dimensionality, and Saddle Points. *Found. Comput. Math.* **2022**, *22*, 35–76. [\[CrossRef\]](#)
43. Chen, X.; Liu, S.; Xu, K.; Li, X.; Lin, X.; Hong, M.; Cox, D.D. ZO-AdaMM: Zeroth-Order Adaptive Momentum Method for Black-Box Optimization. In Proceedings of the 33th Advances in Neural Information Processing Systems, NeurIPS, Vancouver, BC, Canada, 8–14 December 2019.
44. Bottou, L.; Curtis, F.E.; Nocedal, J. Optimization Methods for Large-Scale Machine Learning. *SIAM Rev.* **2018**, *60*, 223–311. [\[CrossRef\]](#)
45. Williams, P.N.; Li, K. Black-Box Sparse Adversarial Attack via Multi-Objective Optimisation CVPR Proceedings. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, 17–24 June 2023.
46. Papernot, N.; McDaniel, P.D.; Goodfellow, I.J.; Jha, S.; Celik, Z.B.; Swami, A. Practical Black-Box Attacks against Machine Learning. In Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, AsiaCCS 2017, Abu Dhabi, United Arab Emirates, 2–6 April 2017.

-
47. Nesterov, Y.E.; Spokoiny, V.G. Random Gradient-Free Minimization of Convex Functions. *Found. Comput. Math.* **2017**, *17*, 527–566. [[CrossRef](#)]
 48. Shamir, O.; Zhang, T. Stochastic Gradient Descent for Non-smooth Optimization: Convergence Results and Optimal Averaging Schemes. In Proceedings of the 30th International Conference on Machine Learning, ICML, Atlanta, GA, USA, 16–21 June 2013.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.