Solving High-Dimensional Multi-Objective Optimization Problems with Low Effective Dimensions*

Hong Qian and **Yang Yu**
National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, 210023, China
Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing, 210023, China
{qianh,yuy}@lamda.nju.edu.cn

Abstract

Multi-objective (MO) optimization problems require simultaneously optimizing two or more objective functions. An MO algorithm needs to find solutions that reach different optimal balances of the objective functions, i.e., optimal Pareto front, therefore, high dimensionality of the solution space can hurt MO optimization much severer than single-objective optimization, which was little addressed in previous studies. This paper proposes a general, theoretically-grounded yet simple approach ReMO, which can scale current derivative-free MO algorithms to the high-dimensional non-convex MO functions with low effective dimensions, using random embedding. We prove the conditions under which an MO function has a low effective dimension, and for such functions, we prove that ReMO possesses the desirable properties of optimal Pareto front preservation, time complexity reduction, and rotation perturbation invariance. Experimental results indicate that ReMO is effective for optimizing the high-dimensional MO functions with low effective dimensions, and is even effective for the high-dimensional MO functions where all dimensions are effective but most only have a small and bounded effect on the function value.

Introduction

Solving sophisticated optimization problems plays an essential role in the development of artificial intelligence. In some real-world applications, we need to simultaneously optimize two or more objective functions instead of only one, which leads to the progress of multi-objective (MO) optimization.

Let $\boldsymbol{f}(\boldsymbol{x}) = \big(f_1(\boldsymbol{x}), \ldots, f_m(\boldsymbol{x})\big)$ denote a multi-objective function. In this paper, we assume that the optimization problems are deterministic, i.e., each call of $\boldsymbol{f}$ returns the same function value for the same solution $\boldsymbol{x}$. Furthermore, we focus on the derivative-free optimization. That is to say, $\boldsymbol{f}$ is regarded as a black-box function, and we can only perform the MO optimization based on the sampled solutions and their function values. Other information such as gradient is not used or even not available. Since the derivative-free optimization methods do not depend on gradient, they are suitable for a wide range of sophisticated real-world optimization problems, such as non-convex functions, non-differentiable functions, and discontinuous functions.

The MO optimization has achieved many remarkable applications such as in reinforcement learning (Moffaert and Nowé 2014), constrained optimization (Qian, Yu, and Zhou 2015a), and software engineering (Harman, Mansouri, and Zhang 2012; Minku and Yao 2013). Two reasons accounting for its successful applications. First, the MO algorithm can find the solutions that reach different optimal balances of the objectives (e.g., performance and cost), which can satisfy different demands from different users. Besides, it has been shown that MO optimization can do better than single-objective optimization in some machine learning tasks (Li et al. 2014; Qian, Yu, and Zhou 2015b; 2015c).

Driven by the demand from real-world applications, despite the hardness of MO optimization such as the objectives are often conflicted with each other, derivative-free MO optimization methods have obtained substantial achievements. Well-known MO optimization methods, such as improved version of the strength Pareto evolutionary algorithm (SPEA2) (Zitzler, Laumanns, and Thiele 2001), region-based selection in evolutionary multi-objective optimization (PESA-II) (Corne et al. 2001), non-dominated sorting genetic algorithm II (NSGA-II) (Deb et al. 2002), and multi-objective evolutionary algorithm based on decomposition (MOEA/D) (Zhang and Li 2007), non-dominated neighbor immune algorithm (NNIA) (Gong et al. 2008), etc., have been proposed and successfully applied in various applications.

Problem. Previous studies have shown that derivative-free MO methods are effective and efficient for the MO functions in low-dimensional solution space. However, MO optimization methods may lose their power for the high-dimensional MO functions, because the convergence rate is slow or the computational cost of each iteration is high in high-dimensional solution space. Furthermore, high dimensionality of the solution space can hurt MO optimization much severer than single-objective optimization, since an MO algorithm needs to find a set of solutions that reaches different optimal balances of the objectives. Therefore, scalability becomes one of the main bottlenecks of MO optimization, which restricts the further applications of it.

*This research was supported by the NSFC (61375061, 61333014, 61305067), Jiangsu Science Foundation (BK20160066), and Foundation for the Author of National Excellent Doctoral Dissertation of China (201451).

Copyright © 2017, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Related Work. Recently, there emerged studies focusing on addressing the issue of scaling derivative-free MO optimization algorithms to high-dimensional solution space, such as (Wang et al. 2015; Ma et al. 2016; Zhang et al. 2016). In (Wang et al. 2015), the algorithm is designed on the basis of analyzing the relationship between the decision variables and the objective functions. In (Ma et al. 2016), the relationship between the decision variables is analyzed in order to design the smart algorithm. In (Zhang et al. 2016), the problem of scalability is handled through decision variable clustering. Although the remarkable empirical performance of these algorithms when addressing high-dimension MO optimization problems, almost all of these algorithms lack the solid theoretical foundations. The questions of when and why these algorithms work need to be answered theoretically in order to guide the better design and application of high-dimensional MO algorithms. Compared with the theoretical progress for derivative-free high-dimensional single-objective optimization (Wang et al. 2013; Kaban, Bootkrajang, and Durrant 2013; Friesen and Domingos 2015; Kandasamy, Schneider, and Póczos 2015; Qian and Yu 2016), the theoretically-grounded MO methods are deficient and thus quite appealing. Furthermore, some existing approaches to scaling MO algorithms to high-dimensional MO problems are not general and only restricted to the special algorithms. It is desirable that more MO algorithms can possess the scalability.

On the other hand, it has been observed that, for a wide class of high-dimensional single-objective optimization problems, such as hyper-parameter optimization in machine learning tasks (Bergstra and Bengio 2012; Hutter, Hoos, and Leyton-Brown 2014), the objective function value is only affected by a few dimensions instead of all dimensions. And we call the dimensions which affect the function value as the effective dimensions. For the high-dimensional single-objective optimization problems with low effective dimensions, the random embedding technique which possesses the solid theoretical foundation has been proposed (Wang et al. 2013; 2016). Due to the desirable theoretical property of random embedding, it has been applied to cooperate with some state-of-the-art derivative-free single-objective optimization methods and shown to be effective. Successful cases including Bayesian optimization (Wang et al. 2013; 2016), simultaneous optimistic optimization (Qian and Yu 2016), and estimation of distribution algorithm (Sanyang and Kaban 2016). These successful examples inspire us that we can extend the random embedding technique to high-dimensional MO functions with low effective dimensions, and at the same time inherit the theoretical merits of random embedding. Besides, it would be desirable if the extension of random embedding is suitable for any MO algorithm rather than only some special algorithms.

Our Contributions. This paper proposes a general, theoretically-grounded yet simple approach ReMO, which can scale any derivative-free MO optimization algorithm to the high-dimensional non-convex MO optimization problems with low effective dimensions using random embedding. ReMO performs the optimization by employing arbitrary derivative-free MO optimization algorithm in low-dimensional solution space, where the function values of solutions are evaluated through embedding it into the original high-dimensional solution space. Theoretical and experimental results verify the effectiveness of ReMO for scalability. The contributions of this paper are:

- Disclosing a sufficient and necessary condition as well as a sufficient condition under which the high-dimensional MO functions have the effective dimensions.
- Proving that ReMO possesses three desirable theoretical properties: Pareto front preservation, time complexity reduction, and rotation perturbation invariance (i.e., robust to rotation perturbation).
- Showing that ReMO is effective to improve the scalability of current derivative-free MO optimization algorithms for the high-dimensional non-convex MO problems with low effective dimensions, and is even effective for the high-dimensional MO problems where all dimensions are effective but most only have a small and bounded effect on the function value.

The rest of the paper is organized as follows. Section 2 introduces the notations of MO optimization and reviews random embedding for the high-dimensional single-objective optimization. Section 3 extends random embedding to the high-dimensional MO optimization and presents the proposed general approach ReMO. Section 4 shows three desirable theoretical properties of ReMO, and Section 5 shows the experimental results. Section 6 concludes the paper.

Background

Multi-Objective Optimization

Multi-objective (MO) optimization simultaneously optimizes two or more objective functions as Definition 1. In this paper, we consider minimization problems.

Definition 1 (Multi-Objective Optimization)
Given m objective functions $f_1, \ldots, f_m$ defined on the solution space $\mathcal{X} \subseteq \mathbb{R}^D$, the minimum multi-objective optimization aims to find the solution $\boldsymbol{x}^* \in \mathcal{X}$ s.t.

$$\boldsymbol{x}^* = \operatorname*{argmin}_{\boldsymbol{x} \in \mathcal{X}} \boldsymbol{f}(\boldsymbol{x}) = \operatorname*{argmin}_{\boldsymbol{x} \in \mathcal{X}} \big(f_1(\boldsymbol{x}), \ldots, f_m(\boldsymbol{x})\big),$$

where $\boldsymbol{f}(\boldsymbol{x}) = \big(f_1(\boldsymbol{x}), \ldots, f_m(\boldsymbol{x})\big)$ is the objective function vector of the solution $\boldsymbol{x}$.

This paper only considers $\mathcal{X} = \mathbb{R}^D$, i.e., unconstrained MO optimization problems. In most cases, it is impossible that there exist a solution $\boldsymbol{x} \in \mathbb{R}^D$ such that $\boldsymbol{x}$ is optimal for all the objectives, since the objectives are often conflicted with each other and optimizing one objective alone will degrade the other objectives. Thus, MO optimization attempts to find out a set of solutions that reach different optimal balances of the objective functions according to some criteria. A widely-used criterion is the Pareto optimality that utilizes the *dominance relationship* between solutions as Definition 2. The solution set with Pareto optimality is called the Pareto set as Definition 3.

Definition 2 (Dominance Relationship)
Let $\boldsymbol{f} = (f_1, \dots, f_m)\colon \mathbb{R}^D \to \mathbb{R}^m$ be the objective function vector, where $\mathbb{R}^D$ is the solution space and $\mathbb{R}^m$ is the objective space. For two solutions $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^D$:

- *$\boldsymbol{x}_1$ weakly dominates $\boldsymbol{x}_2$ iff $f_i(\boldsymbol{x}_1) \le f_i(\boldsymbol{x}_2)$ for all $i \in \{1, \dots, m\}$. Denote the weak dominance relationship as $\boldsymbol{x}_1 \preceq_{\boldsymbol{f}} \boldsymbol{x}_2$.*
- *$\boldsymbol{x}_1$ dominates $\boldsymbol{x}_2$ iff $\boldsymbol{x}_1 \preceq_{\boldsymbol{f}} \boldsymbol{x}_2$ (i.e., $\boldsymbol{x}_1$ weakly dominates $\boldsymbol{x}_2$) and $f_i(\boldsymbol{x}_1) < f_i(\boldsymbol{x}_2)$ for some $i \in \{1, \dots, m\}$. Denote the dominance relationship as $\boldsymbol{x}_1 \prec_{\boldsymbol{f}} \boldsymbol{x}_2$.*

Definition 3 (Pareto Optimality)
Let $\boldsymbol{f} = (f_1, \dots, f_m)\colon \mathbb{R}^D \to \mathbb{R}^m$ be the objective function vector, where $\mathbb{R}^D$ is the solution space and $\mathbb{R}^m$ is the objective space. A solution $\boldsymbol{x}$ is called Pareto optimal if there exists no other solution in $\mathbb{R}^D$ which dominates $\boldsymbol{x}$. A solution set is called the Pareto set if it only contains the Pareto optimal solutions. The collection of objective function vectors of the Pareto set is called the Pareto front of the Pareto set.

Under the criterion of Pareto optimality, MO optimization aims at finding the largest Pareto set $\mathcal{PS}$ that is also called the *optimal Pareto set*, and the corresponding *optimal Pareto front* denoted as $\mathcal{PF}$. To be specific, for each member in $\mathcal{PF}$, MO optimization attempts to find at least one corresponding solution in $\mathcal{PS}$.

Random Embedding for Single-Objective Optimization

Before introducing our general approach to handling a large class of high-dimensional MO functions, in this section we will review the previously proposed random embedding technique for optimizing the high-dimensional single-objective functions with low effective dimensions, and also will present the desirable theoretical property of it.

It has been observed that, for a wide class of high-dimensional single-objective optimization problems, such as hyper-parameter optimization in machine learning (Bergstra and Bengio 2012; Hutter, Hoos, and Leyton-Brown 2014), the objective function value is only affected by a few effective dimensions. Here, we adopt the concept of effective dimension introduced in (Wang et al. 2013; 2016) as Definition 4. We call the effective dimension of single-objective function as S-effective dimension in order to distinguish it from the effective dimension of MO functions that will be formally defined in the next section.

Definition 4 (S-Effective Dimension)
A function $f\colon \mathbb{R}^D \to \mathbb{R}$ is said to have S-effective dimension d_e with $d_e \le D$, if

- *there exists a linear subspace $\mathcal{V} \subseteq \mathbb{R}^D$ with dimension $d_{\mathcal{V}}$ such that for all $\boldsymbol{x} \in \mathbb{R}^D$, we have $f(\boldsymbol{x}) = f(\boldsymbol{x}_e + \boldsymbol{x}_c) = f(\boldsymbol{x}_e)$, where $\boldsymbol{x}_e \in \mathcal{V} \subseteq \mathbb{R}^D$, $\boldsymbol{x}_c \in \mathcal{V}^\perp \subseteq \mathbb{R}^D$ and $\mathcal{V}^\perp$ is the orthogonal complement of $\mathcal{V}$.*
- *$d_e = \min_{\mathcal{V} \in \mathbb{V}} d_{\mathcal{V}}$, where $\mathbb{V}$ is the collection of all the subspaces $\mathcal{V}$ with the property described above.*

We call $\mathcal{V}$ the effective subspace of f and $\mathcal{V}^\perp$ the constant subspace of f.

Intuitively, Definition 4 means that the function value of $f(\boldsymbol{x})$ only varies along the effective subspace $\mathcal{V}$, and does not vary along the constant subspace $\mathcal{V}^\perp$. For the high-dimensional single-objective functions with low S-effective dimensions, Theorem 1 (Wang et al. 2013; 2016) below implies that the random embedding technique is effective. Let $\mathcal{N}(0, 1)$ denote the standard Gaussian distribution, i.e., mean = 0 and variance = 1. The proof of Theorem 1 can be found in (Wang et al. 2013; 2016).

Theorem 1
Given a function $f\colon \mathbb{R}^D \to \mathbb{R}$ with S-effective dimension d_e, and a random matrix $\boldsymbol{A} \in \mathbb{R}^{D \times d}$ with independent members sampled from $\mathcal{N}(0, 1)$ where $d \ge d_e$, then, with probability 1, for any $\boldsymbol{x} \in \mathbb{R}^D$ there exists $\boldsymbol{y} \in \mathbb{R}^d$ such that $f(\boldsymbol{x}) = f(\boldsymbol{A}\boldsymbol{y})$.

From Theorem 1, we know that, given a high-dimensional single-objective function with low S-effective dimension (i.e., $d_e \ll D$) and a random embedding matrix $\boldsymbol{A} \in \mathbb{R}^{D \times d}$, for any maximizer $\boldsymbol{x}^* \in \mathbb{R}^D$, there must exist $\boldsymbol{y}^* \in \mathbb{R}^d$ such that $f(\boldsymbol{A}\boldsymbol{y}^*) = f(\boldsymbol{x}^*)$ with probability 1. Namely, random embedding enables us to optimize the lower-dimensional function $g(\boldsymbol{y}) = f(\boldsymbol{A}\boldsymbol{y})$ in $\mathbb{R}^d$ instead of optimizing the original high-dimensional $f(\boldsymbol{x})$ in $\mathbb{R}^D$, while the function value is still evaluated in the original solution space.

Due to the desirable theoretical property of random embedding, this technique has been applied in some state-of-the-art derivative-free single-objective optimization methods for optimizing high-dimensional single-objective functions with low S-effective dimensions. The performance of them is remarkable, and successful examples include Bayesian optimization (Wang et al. 2013; 2016), simultaneous optimistic optimization (Qian and Yu 2016), and estimation of distribution algorithm (Sanyang and Kaban 2016).

Multi-Objective Optimization via Random Embedding

Inspired from the successful and remarkable cases of applying random embedding to handle the high-dimensional single-objective functions with low S-effective dimensions (Wang et al. 2013; 2016; Qian and Yu 2016; Sanyang and Kaban 2016), in this section, we extend the concept of effective dimension from single-objective functions to MO functions. Conditions under which $\boldsymbol{f}$ has the effective dimension are disclosed, which are useful to verify the existence of the effective dimension. For the high-dimensional MO functions with low effective dimensions, we extend random embedding for handling this function class in a more general way, and thus propose the approach of multi-objective optimization via random embedding (ReMO).

High-Dimensional Multi-Objective Functions with Low Effective Dimensions

The concept of effective dimension for MO functions is formally defined as Definition 5. We call the effective dimension of MO optimization problem as M-effective dimension in order to distinguish it from the effective dimension of single-objective optimization problem.

DEFINITION 5 (M-Effective Dimension)
Let $\boldsymbol{f} = (f_1, \ldots, f_m) \colon \mathbb{R}^D \to \mathbb{R}^m$ be the objective function vector, where $\mathbb{R}^D$ is the solution space and $\mathbb{R}^m$ is the objective space, then, $\boldsymbol{f}$ is said to have M-effective dimension ϑ_e with $\vartheta_e \leq D$, if

- *there exists a linear subspace $\mathcal{V} \subseteq \mathbb{R}^D$ with dimension $\vartheta_{\mathcal{V}}$ such that for all $\boldsymbol{x} \in \mathbb{R}^D$, we have $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x}_e + \boldsymbol{x}_c) = \boldsymbol{f}(\boldsymbol{x}_e)$, where $\boldsymbol{x}_e \in \mathcal{V} \subseteq \mathbb{R}^D$, $\boldsymbol{x}_c \in \mathcal{V}^{\perp} \subseteq \mathbb{R}^D$ and $\mathcal{V}^{\perp}$ is the orthogonal complement of $\mathcal{V}$.*
- *$\vartheta_e = \min_{\mathcal{V} \in \mathbb{V}} \vartheta_{\mathcal{V}}$, where $\mathbb{V}$ is the collection of all the subspaces $\mathcal{V}$ with the property described above.*

We call $\mathcal{V}$ the ***effective subspace*** *of $\boldsymbol{f}$ and $\mathcal{V}^{\perp}$ the* ***constant subspace*** *of $\boldsymbol{f}$.*

Similar to the case of high-dimensional single-objective functions, Definition 5 intuitively indicates that there exists as least one linear subspace $\mathcal{V} \subseteq \mathbb{R}^D$ called effective subspace along which $\boldsymbol{f}(\boldsymbol{x})$ varies, and its orthogonal complement $\mathcal{V}^{\perp}$ called constant subspace makes no effects on $\boldsymbol{f}(\boldsymbol{x})$. It is worthwhile to point out that the definition not only includes the cases of axis-aligned M-effective dimensions but also is not limited to this special cases. It can be verified directly that an effective subspace of $\boldsymbol{f}(\boldsymbol{x})$ is also an effective subspace of each $f_i(\boldsymbol{x})$. Therefore, if $\boldsymbol{f}(\boldsymbol{x})$ has M-effective dimension ϑ_e, then each $f_i(\boldsymbol{x})$ has S-effective dimension $d_e^{(i)} \leq \vartheta_e$ for $i = 1, \ldots, m$, where $d_e^{(i)}$ denotes the S-effective dimension of $f_i(\boldsymbol{x})$.

Given the definition of M-effective dimension, a natural question that we are interested in is how to verify the existence of it. To answer this question, theoretically, we first derive a sufficient and necessary condition under which $\boldsymbol{f}(\boldsymbol{x})$ has the M-effective dimension as Theorem 2. We denote the transpose of matrix $\boldsymbol{M}$ as $\boldsymbol{M}^{\top}$. If a matrix $\boldsymbol{M} \in \mathbb{R}^{D \times D}$ satisfying $\boldsymbol{M}\boldsymbol{M}^{\top} = \boldsymbol{M}^{\top}\boldsymbol{M} = \boldsymbol{I}$, then we call $\boldsymbol{M}$ the orthogonal matrix, where $\boldsymbol{I}$ is the identity matrix.

THEOREM 2
Given any $\boldsymbol{f} = (f_1, \ldots, f_m) \colon \mathbb{R}^D \to \mathbb{R}^m$ and an orthogonal matrix $\boldsymbol{M} \in \mathbb{R}^{D \times D}$, let $\boldsymbol{f}_{\boldsymbol{M}}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{M}\boldsymbol{x}) = (f_1(\boldsymbol{M}\boldsymbol{x}), \cdots, f_m(\boldsymbol{M}\boldsymbol{x}))$, then $\boldsymbol{f}$ has the M-effective dimension ϑ_e if and only if $\boldsymbol{f}_{\boldsymbol{M}}$ has the M-effective dimension ϑ_e.

The proof of Theorem 2 is shown in the appendix[1]. Theorem 2 indicates that $\boldsymbol{f}$ has the effective subspace if and only if the rotation of $\boldsymbol{f}$ has the effective subspace, and they share the same M-effective dimension. This observation provides the possibility that we may verify the existence of M-effective dimension easily via rotating $\boldsymbol{f}$ in a smart way.

In addition to Theorem 2, we also derive a sufficient condition under which $\boldsymbol{f}(\boldsymbol{x})$ has the M-effective dimension as Theorem 3.

THEOREM 3
Given any $\boldsymbol{f} = (f_1, \ldots, f_m)$, if each f_i has at least one effective subspace $\mathcal{V}_i \subseteq \mathbb{R}^D$ with dimension d_i such that $\mathcal{V}_i$ is orthogonal to $\mathcal{V}_j$ for any $i \neq j \in \{1, \ldots, m\}$, then $\boldsymbol{f}$ has the M-effective dimension $\vartheta_e \leq \sum_{i=1}^{m} d_i$.

Algorithm 1 Multi-Objective Optimization via Random Embedding (ReMO)

```
Input:
    MO function $\boldsymbol{f}(\boldsymbol{x}) = (f_1(\boldsymbol{x}), \ldots, f_m(\boldsymbol{x}))$;
    Derivative-free MO optimization algorithm $\mathcal{M}$;
    Number of function evaluation budget $n$;
    Upper bound of the M-effective dimension $\vartheta$ $(\geq \vartheta_e)$.
Procedure:
1: Generate a random matrix $\boldsymbol{A} \in \mathbb{R}^{D \times \vartheta}$ with $\boldsymbol{A}_{i,j} \sim \mathcal{N}(0, 1)$.
2: Apply $\mathcal{M}$ to optimize the low-dimensional MO function $\boldsymbol{g}(\boldsymbol{y}) = \boldsymbol{f}(\boldsymbol{A}\boldsymbol{y}) = (f_1(\boldsymbol{A}\boldsymbol{y}), \ldots, f_m(\boldsymbol{A}\boldsymbol{y}))$ with $n$ function evaluations, where $\boldsymbol{y} \in \mathbb{R}^{\vartheta}$.
3: Obtain the approximate optimal Pareto set $\mathcal{PS}'_{\boldsymbol{g}}$ of $\boldsymbol{g}$ as well as the approximate optimal Pareto front $\mathcal{PF}'_{\boldsymbol{g}}$ of $\boldsymbol{g}$ found by $\mathcal{M}$.
4: Let $\mathcal{PS}'_{\boldsymbol{f}} = \{\boldsymbol{A}\boldsymbol{y} \mid \boldsymbol{y} \in \mathcal{PS}'_{\boldsymbol{g}}\}$ and $\mathcal{PF}'_{\boldsymbol{f}} = \mathcal{PF}'_{\boldsymbol{g}}$.
5: return $\mathcal{PS}'_{\boldsymbol{f}}$ and $\mathcal{PF}'_{\boldsymbol{f}}$.
```

The proof of Theorem 3 is shown in the appendix. Theorem 3 inspires us that we may verify the existence of M-effective dimension of $\boldsymbol{f}$ via each f_i, which decomposes the problem into relatively easy problems. Let $\mathcal{S}_1 + \mathcal{S}_2$ denote the sum of linear vector subspaces $\mathcal{S}_1, \mathcal{S}_2 \subseteq \mathbb{R}^D$. Namely, $\mathcal{S}_1 + \mathcal{S}_2 = \{\boldsymbol{x}_1 + \boldsymbol{x}_2 \mid \boldsymbol{x}_1 \in \mathcal{S}_1, \boldsymbol{x}_2 \in \mathcal{S}_2\}$. It is easy to verify that the sum of linear subspaces $\mathcal{S}_1 + \mathcal{S}_2 \subseteq \mathbb{R}^D$ is also a linear subspace. The proof of Theorem 3 implies that $\sum_{i=1}^{m} \mathcal{V}_i \subseteq \mathbb{R}^D$ is an effective subspaces of $\boldsymbol{f}$, which means that we can construct the effective subspace of $\boldsymbol{f}$ via the effective subspace of each f_i.

The ReMO Approach

For the high-dimensional MO functions with low M-effective dimensions (i.e., $\vartheta_e \ll D$), we extend random embedding to optimize this function class in a more general way, and propose the multi-objective optimization via random embedding (ReMO) as depicted in Algorithm 1.

If $\boldsymbol{f}$ has the M-effective dimension, given an upper bound of the M-effective dimension $\vartheta \geq \vartheta_e$ (instead of knowing ϑ_e exactly), ReMO first generates a random embedding matrix $\boldsymbol{A} \in \mathbb{R}^{D \times \vartheta}$ with each member i.i.d. sampled from the standard Gaussian distribution $\mathcal{N}(0, 1)$ as line 1. Then, ReMO applies some MO optimization algorithm to optimize the lower-dimensional function $\boldsymbol{g}(\boldsymbol{y}) = \boldsymbol{f}(\boldsymbol{A}\boldsymbol{y}) = (f_1(\boldsymbol{A}\boldsymbol{y}), \ldots, f_m(\boldsymbol{A}\boldsymbol{y}))$ in $\mathbb{R}^{\vartheta}$ while the function value is still evaluated in the original solution space $\mathbb{R}^D$, and then gets the approximate optimal Pareto set $\mathcal{PS}'_{\boldsymbol{g}}$ and the approximate optimal Pareto front $\mathcal{PF}'_{\boldsymbol{g}}$ of $\boldsymbol{g}$ as line 2 to 3. It is worthwhile to point out that the derivative-free MO optimization algorithm $\mathcal{M}$ equipped in ReMO can be quite general and is not restricted to any special algorithm. That is to say, we can equip ReMO with any well-known derivative-free MO algorithm such as improved version of the strength Pareto evolutionary algorithm (SPEA2) (Zitzler, Laumanns,

[1] The appendix presenting all the proofs can be found in the homepage of the author `http://cs.nju.edu.cn/yuy`.

and Thiele 2001), region-based selection in evolutionary multi-objective optimization (PESA-II) (Corne et al. 2001), non-dominated sorting genetic algorithm II (NSGA-II) (Deb et al. 2002), multi-objective evolutionary algorithm based on decomposition (MOEA/D) (Zhang and Li 2007), non-dominated neighbor immune algorithm (NNIA) (Gong et al. 2008) and the like. At last, as line 4 to 5, ReMO constructs the approximate optimal Pareto set $\mathcal{PS}'_{\boldsymbol{f}}$ and the approximate optimal Pareto front $\mathcal{PF}'_{\boldsymbol{f}}$ of $\boldsymbol{f}$ from $\mathcal{PS}'_{\boldsymbol{g}}$ and $\mathcal{PF}'_{\boldsymbol{g}}$, and then returns them as the output. Obviously, $\mathcal{PS}'_{\boldsymbol{f}} \subseteq \{\boldsymbol{Ay} \mid \boldsymbol{y} \in \mathcal{PS}_{\boldsymbol{g}}\}$ and $\mathcal{PF}'_{\boldsymbol{f}} \subseteq \mathcal{PF}_{\boldsymbol{g}} \subseteq \mathcal{PF}_{\boldsymbol{f}}$. In the next section, we will show that $\{\boldsymbol{Ay} \mid \boldsymbol{y} \in \mathcal{PS}_{\boldsymbol{g}}\} \subseteq \mathcal{PS}_{\boldsymbol{f}}$ and $\mathcal{PF}_{\boldsymbol{g}} = \mathcal{PF}_{\boldsymbol{f}}$, therefore, $\mathcal{PS}'_{\boldsymbol{f}} \subseteq \mathcal{PS}_{\boldsymbol{f}}$ and $\mathcal{PF}'_{\boldsymbol{f}} \subseteq \mathcal{PF}_{\boldsymbol{f}}$.

ReMO is suitable for a general function class, we only need to get an upper bound of the M-effective dimension rather than knowing ϑ_e exactly, and any derivative-free MO algorithm can be cooperated with ReMO flexibly. All of these reflect that ReMO is a general approach to optimizing the high-dimensional MO functions with low M-effective dimensions. Besides, the implementation of ReMO is simple.

Theoretical Study

If we apply ReMO to optimize the high-dimensional $\boldsymbol{f}$ with low M-effective dimension ϑ_e, theoretically, we show that ReMO inherits the merits of random embedding and possesses the desirable theoretical properties of optimal Pareto front preservation, time complexity reduction, and rotation perturbation invariance.

Optimal Pareto Front Preservation

Let $\boldsymbol{g}(\boldsymbol{y}) = \boldsymbol{f}(\boldsymbol{Ay}) = (f_1(\boldsymbol{Ay}), \ldots, f_m(\boldsymbol{Ay}))$, where $\boldsymbol{y} \in \mathbb{R}^{\vartheta}$ with $\vartheta_e \leq \vartheta \leq D$. Theorem 4 proves that, with probability 1, the optimal Pareto front of $\boldsymbol{g}(\boldsymbol{y})$ is as same as that of $\boldsymbol{f}(\boldsymbol{x})$ and the optimal Pareto set of $\boldsymbol{f}(\boldsymbol{x})$ can be recovered from that of $\boldsymbol{g}(\boldsymbol{y})$, i.e., optimal Pareto front and optimal Pareto set preservation. Denote the optimal Pareto set of $\boldsymbol{f}, \boldsymbol{g}$ as $\mathcal{PS}_{\boldsymbol{f}} \subseteq \mathbb{R}^D, \mathcal{PS}_{\boldsymbol{g}} \subseteq \mathbb{R}^{\vartheta}$, and denote the optimal Pareto front of $\boldsymbol{f}, \boldsymbol{g}$ as $\mathcal{PF}_{\boldsymbol{f}}, \mathcal{PF}_{\boldsymbol{g}} \subseteq \mathbb{R}^m$, respectively.

THEOREM 4
Given any $\boldsymbol{f} = (f_1, \ldots, f_m) \colon \mathbb{R}^D \to \mathbb{R}^m$ with M-effective dimension ϑ_e, then, with probability 1, we have that $\{\boldsymbol{Ay} \mid \boldsymbol{y} \in \mathcal{PS}_{\boldsymbol{g}}\} \subseteq \mathcal{PS}_{\boldsymbol{f}}$ and $\mathcal{PF}_{\boldsymbol{g}} = \mathcal{PF}_{\boldsymbol{f}}$.

The proof of Theorem 4 is shown in the appendix. Theorem 4 implies that, in the procedure of ReMO, optimizing the lower-dimensional function $\boldsymbol{g}(\boldsymbol{y})$ in $\mathbb{R}^{\vartheta}$ instead of optimizing the original high-dimensional function $\boldsymbol{f}(\boldsymbol{x})$ in $\mathbb{R}^D$ will not miss any part of $\mathcal{PS}_{\boldsymbol{f}}$ and $\mathcal{PF}_{\boldsymbol{f}}$ if $\mathcal{M}$ can find $\mathcal{PS}_{\boldsymbol{g}}$ and $\mathcal{PF}_{\boldsymbol{g}}$ perfectly. Here, an MO optimization algorithm $\mathcal{M}$ can find the optimal Pareto set $\mathcal{PS}$ perfectly means that, for each member in the optimal Pareto front $\mathcal{PF}$, $\mathcal{M}$ can find at least one corresponding solution in $\mathcal{PS}$.

Time Complexity Reduction

Since ReMO only optimizes the lower-dimensional function $\boldsymbol{g}$ and can preserve the optimal Pareto front, the time complexity of ReMO is less than that of optimizing $\boldsymbol{f}$ directly, as Theorem 5.

THEOREM 5
Given any $\boldsymbol{f} = (f_1, \ldots, f_m) \colon \mathbb{R}^D \to \mathbb{R}^m$ with M-effective dimension ϑ_e, assume that $\mathcal{M}$ can find the $\mathcal{PS}_{\boldsymbol{f}}$ and $\mathcal{PF}_{\boldsymbol{f}}$ with time complexity $\mathcal{O}(\phi(D, m))$, then, with probability 1, ReMO equipped with the same algorithm $\mathcal{M}$ can find the $\mathcal{PS}_{\boldsymbol{f}}$ and $\mathcal{PF}_{\boldsymbol{f}}$ with time complexity $\mathcal{O}(\phi(\vartheta, m))$, where $\phi(d, m)$ is a monotone increasing function with respect to the dimension of solution space d.

The proof of Theorem 5 is shown in the appendix. From Theorem 5, we know that how much the time complexity can be reduced relies on the form of function ϕ, and thus depends on the specific function $\boldsymbol{f}$ and the specific algorithm $\mathcal{M}$.

Rotation Perturbation Invariance

At last, we theoretically show that ReMO is robust to rotation perturbation (i.e., rotation perturbation invariance) as Theorem 6. The similar result of the single-objective function optimization can be found in (Wang et al. 2016).

THEOREM 6
Given any $\boldsymbol{f} = (f_1, \ldots, f_m) \colon \mathbb{R}^D \to \mathbb{R}^m$ with M-effective dimension ϑ_e and an orthogonal matrix $\boldsymbol{M} \in \mathbb{R}^{D \times D}$, let $\boldsymbol{f_M}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{Mx}) = (f_1(\boldsymbol{Mx}), \cdots, f_m(\boldsymbol{Mx}))$ denote the rotation perturbation of $\boldsymbol{f}$, if we apply ReMO to optimize $\boldsymbol{f_M}$, ReMO can still find the $\mathcal{PS}_{\boldsymbol{f}}$ and $\mathcal{PF}_{\boldsymbol{f}}$ of $\boldsymbol{f}$ with probability 1.

The proof of Theorem 6 is shown in the appendix. Theorem 6 indicates that the rotation perturbation of $\boldsymbol{f}$ may not affect ReMO finding the $\mathcal{PS}_{\boldsymbol{f}}$ and $\mathcal{PF}_{\boldsymbol{f}}$. That is to say, ReMO is robust to rotation perturbation.

Experiments

On Functions with M-Effective Dimensions

We first verify the effectiveness of ReMO empirically on three high-dimensional bi-objective (i.e., $m = 2$) optimization testing functions ZDT1^0, ZDT2^0 and ZDT3^0 with low M-effective dimensions. They are constructed based on the first three bi-objective functions ZDT1, ZDT2 and ZDT3 introduced in (Zitzler, Deb, and Thiele 2000). The dimensions of ZDT1, ZDT2 and ZDT3 are all 30. ZDT1^0, ZDT2^0 and ZDT3^0 are constructed to meet the M-effective dimension assumption as follows: ZDT1, ZDT2 and ZDT3 are embedded into a D-dimensional solution space $\mathcal{X} \subseteq \mathbb{R}^D$ with $D \gg 30$. The embedding is done by firstly adding additional $D - 30$ dimensions, but the additional dimensions have no effects on the function value; secondly, the embedded functions are rotated via an orthogonal matrix. Thus, the dimensions of ZDT1^0, ZDT2^0 and ZDT3^0 are D while their M-effective dimensions are 30. For these bi-objective testing functions, ZDT1^0 has a convex optimal Pareto front, ZDT2^0 has a non-convex optimal Pareto front, and ZDT3^0 has a discontinuous optimal Pareto front. The second objective functions of ZDT1^0, ZDT2^0 and ZDT3^0 are all non-convex.

For the practical issues in experiments, we set the high-dimensional solution space $\mathcal{X} = [-1, 1]^D$ and the low-dimensional solution space $\mathcal{Y} = [-1, 1]^{\vartheta}$, instead of $\mathbb{R}^D$ and $\mathbb{R}^{\vartheta}$. To implement the MO algorithm in $\mathcal{Y}$, since there may

Table 1: Comparing the achieved hyper-volume indicators of the algorithms on 10000-dimensional multi-objective functions *with* low M-effective dimensions (mean ± standard derivation). In each row, an entry of Re-NSGA-II (or Re-MOEA/D) is bold if its mean value is better than NSGA-II (or MOEA/D); and an entry of Re-NSGA-II (or Re-MOEA/D) is marked with bullet if it is significantly better than NSGA-II (or MOEA/D) by t-test with 5% significance level.

Algorithm	*NSGA-II*	*Re-NSGA-II*	*MOEA/D*	*Re-MOEA/D*
ZDT1^0	0.4176±0.0099	**0.8633**±0.0398●	0.5935±0.0078	**0.6718**±0.0153●
ZDT2^0	0.2259±0.0355	**0.7076**±0.0333●	0.4313±0.0261	**0.6931**±0.0294●
ZDT3^0	0.4248±0.0191	**0.8289**±0.0209●	0.5868±0.0254	**0.6818**±0.0243●

Table 2: Comparing the achieved hyper-volume indicators of the algorithms on 10000-dimensional multi-objective functions *without* low M-effective dimensions (mean ± standard derivation). In each row, an entry of Re-NSGA-II (or Re-MOEA/D) is bold if its mean value is better than NSGA-II (or MOEA/D); and an entry of Re-NSGA-II (or Re-MOEA/D) is marked with bullet if it is significantly better than NSGA-II (or MOEA/D) by t-test with 5% significance level.

Algorithm	*NSGA-II*	*Re-NSGA-II*	*MOEA/D*	*Re-MOEA/D*
ZDT1$^\varepsilon$	0.2681±0.0138	**0.4308**±0.0220●	0.4063±0.0245	0.3961±0.0160
ZDT2$^\varepsilon$	0.0939±0.0140	**0.3208**±0.0237●	0.1934±0.0334	**0.2998**±0.0310●
ZDT3$^\varepsilon$	0.2799±0.0240	**0.4666**±0.0152●	0.3404±0.0189	**0.3804**±0.0230●

exist $\boldsymbol{y}' \in \mathcal{Y}$ s.t. $\boldsymbol{A}\boldsymbol{y}' \notin \mathcal{X}$ and thus $\boldsymbol{f}$ cannot be evaluated at point $\boldsymbol{A}\boldsymbol{y}'$. To address this problem, we use Euclidean projection, i.e., $\boldsymbol{A}\boldsymbol{y}'$ is projected to $\mathcal{X}$ when it is outside $\mathcal{X}$ by $P_{\mathcal{X}}(\boldsymbol{A}\boldsymbol{y}') = \mathrm{argmin}_{\boldsymbol{x}\in\mathcal{X}} \|\boldsymbol{x} - \boldsymbol{A}\boldsymbol{y}'\|_2$. We employ two well-known derivative-free MO optimization methods to minimize ZDT1^0, ZDT2^0 and ZDT3^0: non-dominated sorting genetic algorithm II (NSGA-II) (Deb et al. 2002), and multi-objective evolutionary algorithm based on decomposition (MOEA/D) (Zhang and Li 2007). The implementations of them are both by their authors. When applying the random embedding technique to these optimization methods, we denote them by the prefix "Re-". Thus, we have combinations including Re-NSGA-II and Re-MOEA/D. NSGA-II and MOEA/D are compared with Re-NSGA-II and Re-MOEA/D on these three testing functions with dimension $D = 10000(\gg \vartheta_e = 30)$. We set the function evaluation budget $n = 0.3D = 3000$ and the upper bound of M-effective dimension $\vartheta = 50 > \vartheta_e = 30$. To measure the performance of each algorithm, we adopt the hyper-volume indicator with reference point $(1, 4)$ that quantifies the closeness of the approximate optimal Pareto front each algorithm found to the true optimal Pareto front. The hyper-volume indicator ranges from 0 to 1, and the larger the better. Each algorithm is repeated 30 times independently. The hyper-volume indicator is reported in Table 1.

Table 1 shows that, for high-dimensional MO functions with low M-effective dimensions, ReMO is always significantly better than applying derivative-free MO methods in high-dimensional solution space directly on all the testing functions, whenever we choose NSGA-II or MOEA/D. And the advantage of ReMO is particularly obvious when comparing Re-NSGA-II with NSGA-II. This indicates that ReMO is effective for MO functions with M-effective dimensions, and its effectiveness is general.

Due to the restriction of space, the additional experimental results which show the approximate optimal Pareto front each algorithm found for ZDT1^0, ZDT2^0 and ZDT3^0 can be found in the appendix.

On Functions without M-Effective Dimensions

Furthermore, we also verify the effectiveness of ReMO empirically on three high-dimensional bi-objective optimization testing functions ZDT1$^\varepsilon$, ZDT2$^\varepsilon$ and ZDT3$^\varepsilon$ where all dimensions are effective but most only have a small and bounded effect (up to ε but not zero effect) on the function value. This MO function class, where the M-effective dimension assumption may no longer hold, is more general since the MO functions with M-effective dimensions (i.e., $\varepsilon = 0$) are special cases of this function class. To meet this requirement, ZDT1$^\varepsilon$, ZDT2$^\varepsilon$ and ZDT3$^\varepsilon$ are constructed on the basis of ZDT1, ZDT2 and ZDT3 as follows: ZDT1, ZDT2 and ZDT3 are embedded into a D-dimensional solution space $\mathcal{X}$ with $D = 10000$. The embedding is done by adding additional $D - 30$ dimensions $\sum_{i=31}^{D} x_i/D$. We set the function evaluation budget $n = 0.3D = 3000$ and set the reference point as $(1, 4)$ to calculate the hyper-volume indicator (the larger the better). Each algorithm is repeated 30 times independently. The achieved hyper-volume indicator is reported in Table 2.

Table 2 indicates that, except Re-MOEA/D on ZDT1$^\varepsilon$, ReMO is always significantly better than applying MO methods directly (especially when comparing Re-NSGA-II with NSGA-II). This implies that, for high-dimensional MO functions where all dimensions are effective but most only have a small and bounded impact, ReMO still works well and can also be applied.

Conclusion

This paper proposes a general, theoretically-grounded yet simple approach ReMO that can scale any derivative-free MO optimization algorithm to the high-dimensional non-convex MO functions with low M-effective dimensions via random embedding. Theoretically, we disclose the conditions under which the high-dimensional MO functions have the M-effective dimensions, and prove that ReMO possesses the desirable theoretical properties of Pareto front preservation, time complexity reduction, and rotation perturbation invariance (i.e., robust to rotation perturbation) for such kind of functions. Experimental results show that ReMO is effective to improve the scalability of current derivative-free MO optimization algorithms, and even may be effective for the high-dimensional MO functions without low M-effective dimensions. In the future, we will apply ReMO to more sophisticated real-world tasks.

References

Bergstra, J., and Bengio, Y. 2012. Random search for hyper-parameter optimization. *Journal of Machine Learning Research* 13:281–305.

Corne, D. W.; Jerram, N. R.; Knowles, J. D.; Oates, M. J.; and J, M. 2001. PESA-II: Region-based selection in evolutionary multiobjective optimization. In *Proceedings of the 3rd ACM Conference on Genetic and Evolutionary Computation*, 283–290.

Deb, K.; Agrawal, S.; Pratap, A.; and Meyarivan, T. 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation* 6(2):182–197.

Friesen, A. L., and Domingos, P. M. 2015. Recursive decomposition for nonconvex optimization. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, 253–259.

Gong, M.; Jiao, L.; Du, H.; and Bo, L. 2008. Multiobjective immune algorithm with nondominated neighbor-based selection. *Evolutionary Computation* 16(2):225–255.

Harman, M.; Mansouri, S. A.; and Zhang, Y. 2012. Search-based software engineering: Trends, techniques and applications. *ACM Computing Surveys* 45(1):11.

Hutter, F.; Hoos, H.; and Leyton-Brown, K. 2014. An efficient approach for assessing hyperparameter importance. In *Proceedings of the 31th International Conference on Machine Learning*, 754–762.

Kaban, A.; Bootkrajang, J.; and Durrant, R. J. 2013. Towards large scale continuous EDA: a random matrix theory perspective. In *Proceedings of the 15th Annual Conference on Genetic and Evolutionary Computation*, 383–390.

Kandasamy, K.; Schneider, J.; and Póczos, B. 2015. High dimensional bayesian optimisation and bandits via additive models. In *Proceedings of the 32nd International Conference on Machine Learning*, 295–304.

Li, L.; Yao, X.; Stolkin, R.; Gong, M.; and He, S. 2014. An evolutionary multiobjective approach to sparse reconstruction. *IEEE Transactions on Evolutionary Computation* 18(6):827–845.

Ma, X.; Liu, F.; Qi, Y.; Wang, X.; Li, L.; Jiao, L.; Yin, M.; and Gong, M. 2016. A multiobjective evolutionary algorithm based on decision variable analyses for multiobjective optimization problems with large-scale variables. *IEEE Transactions on Evolutionary Computation* 20(2):275–298.

Minku, L. L., and Yao, X. 2013. Software effort estimation as a multi-objective learning problem. *ACM Transactions on Software Engineering and Methodology* 22(4):35.

Moffaert, K. V., and Nowé, A. 2014. Multi-objective reinforcement learning using sets of pareto dominating policies. *Journal of Machine Learning Research* 15(1):3483–3512.

Qian, H., and Yu, Y. 2016. Scaling simultaneous optimistic optimization for high-dimensional non-convex functions with low effective dimensions. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, 2000–2006.

Qian, C.; Yu, Y.; and Zhou, Z.-H. 2015a. On constrained boolean pareto optimization. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, 389–395.

Qian, C.; Yu, Y.; and Zhou, Z.-H. 2015b. Pareto ensemble pruning. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, 2935–2941.

Qian, C.; Yu, Y.; and Zhou, Z.-H. 2015c. Subset selection by pareto optimization. In *Advances in Neural Information Processing Systems 28*, 1765–1773.

Sanyang, M. L., and Kaban, A. 2016. REMEDA: Random Embedding EDA for optimising functions with intrinsic dimension. In *Proceedings of the 14th International Conference on Parallel Problem Solving from Nature*.

Wang, Z.; Zoghi, M.; Hutter, F.; Matheson, D.; and Freitas, N. D. 2013. Bayesian optimization in high dimensions via random embeddings. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, 1778–1784.

Wang, H.; Jiao, L.; Shang, R.; He, S.; and Liu, F. 2015. A memetic optimization strategy based on dimension reduction in decision space. *Evolutionary Computation* 23(1):69–100.

Wang, Z.; Zoghi, M.; Hutter, F.; Matheson, D.; and Freitas, N. D. 2016. Bayesian optimization in a billion dimensions via random embeddings. *Journal of Artificial Intelligence Research* 55:361–387.

Zhang, Q., and Li, H. 2007. MOEA/D: A multiobjective evolutionary algorithm based on decomposition. *IEEE Transactions on Evolutionary Computation* 11(6):712–731.

Zhang, X.; Tian, Y.; Cheng, R.; and Jin, Y. 2016. A decision variable clustering-based evolutionary algorithm for large-scale many-objective optimization. *IEEE Transactions on Evolutionary Computation*.

Zitzler, E.; Deb, K.; and Thiele, L. 2000. Comparison of multiobjective evolutionary algorithms: Empirical results. *Evolutionary Computation* 8(2):173–195.

Zitzler, E.; Laumanns, M.; and Thiele, L. 2001. SPEA2: Improving the Strength Pareto Evolutionary Algorithm. Technical Report 103, Computer Engineering and Networks Laboratory (TIK), ETH Zurich, Zurich, Switzerland.

Appendix of "Solving High-Dimensional Multi-Objective Optimization Problems with Low Effective Dimensions"

Hong Qian and **Yang Yu**
National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, 210023, China
Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing, 210023, China
{qianh,yuy}@lamda.nju.edu.cn

In this appendix, we present the proofs of Theorem 1 to 6 in the paper in Section 1 to 6, respectively. Section 7 presents the additional experimental results of the paper.

Proof of Theorem 1

This section presents the proof of Theorem 1 in the paper. Let $\mathcal{N}(0,1)$ denote the standard Gaussian distribution, i.e., mean $= 0$ and variance $= 1$, and $\boldsymbol{M}^\top$ denote the transpose of matrix $\boldsymbol{M}$. The proof of Theorem 1 is shown below which can be found in (Wang et al. 2013; 2016).

THEOREM 1
Given a function $f\colon \mathbb{R}^D \to \mathbb{R}$ with S-effective dimension d_e, and a random matrix $\boldsymbol{A} \in \mathbb{R}^{D\times d}$ with independent members sampled from $\mathcal{N}(0,1)$ where $d \geq d_e$, then, with probability 1, for any $\boldsymbol{x} \in \mathbb{R}^D$ there exists $\boldsymbol{y} \in \mathbb{R}^d$ such that $f(\boldsymbol{x}) = f(\boldsymbol{A}\boldsymbol{y})$.

Proof. Since f has the effective dimension d_e, there exists an effective subspace $\mathcal{V} \subseteq \mathbb{R}^D$ with $\dim(\mathcal{V}) = d_e$. In addition, any $\boldsymbol{x} \in \mathbb{R}^D$ can be decomposed as $\boldsymbol{x} = \boldsymbol{x}_e + \boldsymbol{x}_c$, where $\boldsymbol{x}_e \in \mathcal{V}$, $\boldsymbol{x}_c \in \mathcal{V}^\perp$ and $\mathcal{V}^\perp$ is the orthogonal complement of $\mathcal{V}$. By the definition of effective subspace, we have $f(\boldsymbol{x}) = f(\boldsymbol{x}_e + \boldsymbol{x}_c) = f(\boldsymbol{x}_e)$. Therefore, it suffices to show that, with probability 1, for any $\boldsymbol{x}_e \in \mathcal{V}$ there exists $\boldsymbol{y} \in \mathbb{R}^d$ such that $f(\boldsymbol{x}_e) = f(\boldsymbol{A}\boldsymbol{y})$.

Let $\boldsymbol{\Phi} \in \mathbb{R}^{D\times d_e}$ be a matrix whose columns form a standard orthonormal basis of $\mathcal{V}$. Thus, for any $\boldsymbol{x}_e \in \mathcal{V}$, there exists $\boldsymbol{c} \in \mathbb{R}^{d_e}$ such that $\boldsymbol{x}_e = \boldsymbol{\Phi}\boldsymbol{c}$. Let us for now assume that $\boldsymbol{\Phi}^\top\boldsymbol{A}$ has rank d_e. If $\operatorname{rank}(\boldsymbol{\Phi}^\top\boldsymbol{A}) = d_e$, there must exist $\boldsymbol{y} \in \mathbb{R}^d$ such that $(\boldsymbol{\Phi}^\top\boldsymbol{A})\boldsymbol{y} = \boldsymbol{c}$, because $\operatorname{rank}(\boldsymbol{\Phi}^\top\boldsymbol{A}) = \operatorname{rank}([\boldsymbol{\Phi}^\top\boldsymbol{A}, \boldsymbol{c}])$. The orthonormal projection of $\boldsymbol{A}\boldsymbol{y}$ onto $\mathcal{V}$ is given by $\boldsymbol{\Phi}\boldsymbol{\Phi}^\top\boldsymbol{A}\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{c} = \boldsymbol{x}_e$. Thus, $\boldsymbol{A}\boldsymbol{y} = \boldsymbol{x}_e + \tilde{\boldsymbol{x}}$, where $\tilde{\boldsymbol{x}} \in \mathcal{V}^\perp$ since $\boldsymbol{x}_e$ is the orthonormal projection of $\boldsymbol{A}\boldsymbol{y}$ onto $\mathcal{V}$. Therefore, we have $f(\boldsymbol{A}\boldsymbol{y}) = f(\boldsymbol{x}_e + \tilde{\boldsymbol{x}}) = f(\boldsymbol{x}_e)$.

At last, it remains to verify that $\operatorname{rank}(\boldsymbol{\Phi}^\top\boldsymbol{A}) = d_e$ with probability 1. Let $\boldsymbol{A}_e \in \mathbb{R}^{D\times d_e}$ be a sub-matrix of $\boldsymbol{A}$ consisting of any d_e columns of $\boldsymbol{A}$, which are i.i.d. samples distributed according to $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, where $\boldsymbol{I}$ is the identity matrix. Let $\boldsymbol{a}_i$ denote any column of $\boldsymbol{A}_e$. By the standard orthonormal property of $\boldsymbol{\Phi}$, $\boldsymbol{\Phi}^\top\boldsymbol{a}_i$ are i.i.d. samples from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Phi}^\top\boldsymbol{\Phi}) = \mathcal{N}(\boldsymbol{0}_{d_e}, \boldsymbol{I}_{d_e\times d_e})$, and thus we have $\boldsymbol{\Phi}^\top\boldsymbol{A}_e$, when considered as an element of $\mathbb{R}^{d_e^2}$, is a sample from $\mathcal{N}(\boldsymbol{0}_{d_e^2}, \boldsymbol{I}_{d_e^2\times d_e^2})$. On the other hand, the set of singular matrices in $\mathbb{R}^{d_e^2}$ has Lebesgue measure zero, since it is the zero set of a polynomial (i.e., the determinant function) and polynomial functions are Lebesgue measurable. Furthermore, the Gaussian distribution is absolutely continuous with respect to the Lebesgue measure, so $\boldsymbol{\Phi}^\top\boldsymbol{A}_e$ is almost surely non-singular, which means that it has rank d_e, and thus the same holds for $\boldsymbol{\Phi}^\top\boldsymbol{A}$ whose columns contain the columns of $\boldsymbol{\Phi}^\top\boldsymbol{A}_e$. □

Proof of Theorem 2

This section presents the proof of Theorem 2 in the paper. A matrix $\boldsymbol{M} \in \mathbb{R}^{D\times D}$ is called the orthogonal matrix if and only if $\boldsymbol{M}\boldsymbol{M}^\top = \boldsymbol{M}^\top\boldsymbol{M} = \boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix.

THEOREM 2
Given any $\boldsymbol{f} = (f_1, \ldots, f_m)\colon \mathbb{R}^D \to \mathbb{R}^m$ and an orthogonal matrix $\boldsymbol{M} \in \mathbb{R}^{D\times D}$, let $\boldsymbol{f}_{\boldsymbol{M}}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{M}\boldsymbol{x}) = (f_1(\boldsymbol{M}\boldsymbol{x}), \cdots, f_m(\boldsymbol{M}\boldsymbol{x}))$, then $\boldsymbol{f}$ has the M-effective dimension ϑ_e if and only if $\boldsymbol{f}_{\boldsymbol{M}}$ has the M-effective dimension ϑ_e.

Proof. On the one hand, if $\boldsymbol{f}_{\boldsymbol{M}}$ has the M-effective dimension, let $\mathcal{V} \subseteq \mathbb{R}^D$ be any effective subspace of $\boldsymbol{f}_{\boldsymbol{M}}$ with dimension ϑ. Let $\{\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_\vartheta\}$ be a basis of $\mathcal{V}$, and $\{\boldsymbol{\alpha}_{\vartheta+1}, \ldots, \boldsymbol{\alpha}_D\}$ be a basis of $\mathcal{V}^\perp$ where $\mathcal{V}^\perp$ is the orthogonal complement of $\mathcal{V}$. Let $\mathcal{V}_{\boldsymbol{M}} = \{\boldsymbol{M}\boldsymbol{x} \mid \boldsymbol{x} \in \mathcal{V}\} \subseteq \mathbb{R}^D$. Since $\boldsymbol{M}$ is an orthogonal matrix, we have that $\boldsymbol{M}^{-1}$ always exists and $\boldsymbol{M}^{-1} = \boldsymbol{M}^\top$. We can verify directly that $\mathcal{V}_{\boldsymbol{M}}$ is also a linear subspace if $\mathcal{V}$ is a linear subspace, $\{\boldsymbol{M}\boldsymbol{\alpha}_1, \ldots, \boldsymbol{M}\boldsymbol{\alpha}_\vartheta\}$ is a basis of $\mathcal{V}_{\boldsymbol{M}}$ if $\{\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_\vartheta\}$ is a basis of $\mathcal{V}$ and $\boldsymbol{M}$ is an orthogonal matrix, and $\{\boldsymbol{M}\boldsymbol{\alpha}_{\vartheta+1}, \ldots, \boldsymbol{M}\boldsymbol{\alpha}_D\}$ is a basis of $\mathcal{V}_{\boldsymbol{M}}^\perp$ if $\{\boldsymbol{\alpha}_{\vartheta+1}, \ldots, \boldsymbol{\alpha}_D\}$ is a basis of $\mathcal{V}^\perp$ and $\boldsymbol{M}$ is an orthogonal matrix. Therefore, the dimension of $\mathcal{V}_{\boldsymbol{M}}$ is also ϑ. We now prove that $\mathcal{V}_{\boldsymbol{M}}$ is an effective subspace of $\boldsymbol{f}$. For any $\boldsymbol{x} \in \mathbb{R}^D$, since $\boldsymbol{x} = \sum_{i=1}^{\vartheta} a_i\boldsymbol{M}\boldsymbol{\alpha}_i + \sum_{j=\vartheta+1}^{D} a_j\boldsymbol{M}\boldsymbol{\alpha}_j$ where $\sum_{i=1}^{\vartheta} a_i\boldsymbol{M}\boldsymbol{\alpha}_i \in \mathcal{V}_{\boldsymbol{M}}$ and $\sum_{j=\vartheta+1}^{D} a_j\boldsymbol{M}\boldsymbol{\alpha}_j \in \mathcal{V}_{\boldsymbol{M}}^\perp$, we have that $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{f}(\sum_{i=1}^{\vartheta} a_i\boldsymbol{M}\boldsymbol{\alpha}_i + \sum_{j=\vartheta+1}^{D} a_j\boldsymbol{M}\boldsymbol{\alpha}_j) = \boldsymbol{f}_{\boldsymbol{M}}(\sum_{i=1}^{\vartheta} a_i\boldsymbol{\alpha}_i + \sum_{j=\vartheta+1}^{D} a_j\boldsymbol{\alpha}_j) = \boldsymbol{f}_{\boldsymbol{M}}(\sum_{i=1}^{\vartheta} a_i\boldsymbol{\alpha}_i) =$

Copyright © 2017, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

$\boldsymbol{f}_{\boldsymbol{M}}(\boldsymbol{M}^{-1}\sum_{i=1}^{\vartheta} a_i\boldsymbol{M}\boldsymbol{\alpha}_i) = \boldsymbol{f}(\sum_{i=1}^{\vartheta} a_i\boldsymbol{M}\boldsymbol{\alpha}_i)$, where the third equality is by that $\mathcal{V}$ is an effective subspace of $\boldsymbol{f}_{\boldsymbol{M}}$ and $\sum_{i=1}^{\vartheta} a_i\boldsymbol{\alpha}_i \in \mathcal{V}$, and the last equality is by that $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{f}_{\boldsymbol{M}}(\boldsymbol{M}^{-1}\boldsymbol{x})$. This shows that $\mathcal{V}_{\boldsymbol{M}}$ is an effective subspace of $\boldsymbol{f}$.

On the other hand, if $\boldsymbol{f}$ has the M-effective dimension, let $\mathcal{V} \subseteq \mathbb{R}^D$ be any effective subspace of $\boldsymbol{f}$ with dimension ϑ. Consider the linear subspace $\mathcal{V}_{\boldsymbol{M}^{-1}} = \{\boldsymbol{M}^{-1}\boldsymbol{x} \mid \boldsymbol{x} \in \mathcal{V}\} \subseteq \mathbb{R}^D$ with dimension ϑ. With the same arguments above, we can verify that $\mathcal{V}_{\boldsymbol{M}^{-1}}$ is an effective subspace of $\boldsymbol{f}_{\boldsymbol{M}}$, which proves the theorem. □

Proof of Theorem 3

This section presents the proof of Theorem 3 in the paper. Before presenting the proof of Theorem 3, we first show three lemmas below that will be used in the poof of Theorem 3. Let $\mathcal{S}_1 + \mathcal{S}_2$ denote the sum of linear vector subspaces $\mathcal{S}_1, \mathcal{S}_2 \subseteq \mathbb{R}^D$. Namely, $\mathcal{S}_1 + \mathcal{S}_2 = \{\boldsymbol{x}_1 + \boldsymbol{x}_2 \mid \boldsymbol{x}_1 \in \mathcal{S}_1, \boldsymbol{x}_2 \in \mathcal{S}_2\}$. It is easy to verify that the sum of linear subspaces $\mathcal{S}_1 + \mathcal{S}_2 \subseteq \mathbb{R}^D$ is also a linear subspace. Let $\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle$ denote the inner product of two vectors $\boldsymbol{x}_1, \boldsymbol{x}_2$.

LEMMA 1
Given two linear subspaces $\mathcal{S}_1, \mathcal{S}_2 \subseteq \mathbb{R}^D$, let $\mathcal{S}_1^{\perp}$, $\mathcal{S}_2^{\perp}$ and $(\mathcal{S}_1 + \mathcal{S}_2)^{\perp}$ denote the orthogonal complement subspaces of $\mathcal{S}_1, \mathcal{S}_2$ and $\mathcal{S}_1 + \mathcal{S}_2$ respectively, then $(\mathcal{S}_1 + \mathcal{S}_2)^{\perp} = \mathcal{S}_1^{\perp} \cap \mathcal{S}_2^{\perp}$.

Proof. On the one hand, given any $\boldsymbol{x} \in (\mathcal{S}_1 + \mathcal{S}_2)^{\perp}$, by the definition of $\mathcal{S}_1 + \mathcal{S}_2$, we have that $\mathcal{S}_1, \mathcal{S}_2 \subseteq \mathcal{S}_1 + \mathcal{S}_2$ and thus $\boldsymbol{x} \in \mathcal{S}_1^{\perp} \cap \mathcal{S}_2^{\perp}$, i.e., $(\mathcal{S}_1 + \mathcal{S}_2)^{\perp} \subseteq \mathcal{S}_1^{\perp} \cap \mathcal{S}_2^{\perp}$.

On the other hand, given any $\boldsymbol{x} \in \mathcal{S}_1^{\perp} \cap \mathcal{S}_2^{\perp}$, for all $\boldsymbol{x}_1 \in \mathcal{S}_1$ and $\boldsymbol{x}_2 \in \mathcal{S}_2$, we have that $\langle \boldsymbol{x}, \boldsymbol{x}_1 \rangle = \langle \boldsymbol{x}, \boldsymbol{x}_2 \rangle = 0$. Therefore, $\langle \boldsymbol{x}, \boldsymbol{x}_1 + \boldsymbol{x}_2 \rangle = \langle \boldsymbol{x}, \boldsymbol{x}_1 \rangle + \langle \boldsymbol{x}, \boldsymbol{x}_2 \rangle = 0$ for all $\boldsymbol{x}_1 \in \mathcal{S}_1$ and $\boldsymbol{x}_2 \in \mathcal{S}_2$. That is to say, $\boldsymbol{x} \in (\mathcal{S}_1 + \mathcal{S}_2)^{\perp}$ and thus $\mathcal{S}_1^{\perp} \cap \mathcal{S}_2^{\perp} \subseteq (\mathcal{S}_1 + \mathcal{S}_2)^{\perp}$. Therefore, we have that $(\mathcal{S}_1 + \mathcal{S}_2)^{\perp} = \mathcal{S}_1^{\perp} \cap \mathcal{S}_2^{\perp}$. □

LEMMA 2
Given a linear subspace $\mathcal{S}_1 \subseteq \mathbb{R}^D$, let $\mathcal{S}_1^{\perp}$ denote its orthogonal complement subspace, if there exists a linear subspace $\mathcal{S}_2 \subseteq \mathbb{R}^D$ such that $\mathcal{S}_1 + \mathcal{S}_2 = \mathbb{R}^D$ and $\mathcal{S}_2$ is orthogonal to $\mathcal{S}_1$, then $\mathcal{S}_2 = \mathcal{S}_1^{\perp}$.

Proof. Since $\mathcal{S}_1 + \mathcal{S}_2 = \mathbb{R}^D$ as well as $\mathcal{S}_2$ is orthogonal to $\mathcal{S}_1$, we have that $\mathcal{S}_2$ is also an orthogonal complement subspace of $\mathcal{S}_1$. To prove $\mathcal{S}_2 = \mathcal{S}_1^{\perp}$, it suffices to verify that the orthogonal complement subspace of $\mathcal{S}_1$ is unique. Assume that there exist two linear subspaces $\mathcal{S}_1', \mathcal{S}_1'' \subseteq \mathbb{R}^D$ such that $\mathcal{S}_1 + \mathcal{S}_1' = \mathbb{R}^D$, $\mathcal{S}_1 + \mathcal{S}_1'' = \mathbb{R}^D$, and $\mathcal{S}_1', \mathcal{S}_1''$ are both orthogonal to $\mathcal{S}_1$. Given any $\boldsymbol{x} \in \mathcal{S}_1'$, we have $\boldsymbol{x} = \boldsymbol{x}_1 + \boldsymbol{x}_2$, where $\boldsymbol{x}_1 \in \mathcal{S}_1$ and $\boldsymbol{x}_2 \in \mathcal{S}_1''$. Since $\mathcal{S}_1', \mathcal{S}_1''$ are both orthogonal to $\mathcal{S}_1$, we have

$$\begin{aligned}\langle \boldsymbol{x}, \boldsymbol{x}_1 \rangle &= \langle \boldsymbol{x}_1 + \boldsymbol{x}_2, \boldsymbol{x}_1 \rangle = \langle \boldsymbol{x}_1, \boldsymbol{x}_1 \rangle + \langle \boldsymbol{x}_2, \boldsymbol{x}_1 \rangle \\ &= \langle \boldsymbol{x}_1, \boldsymbol{x}_1 \rangle = 0.\end{aligned}$$

That is to say $\boldsymbol{x}_1 = \boldsymbol{0}$ and thus $\boldsymbol{x} \in \mathcal{S}_1''$, i.e., $\mathcal{S}_1' \subseteq \mathcal{S}_1''$. On the other hand, given any $\boldsymbol{x} \in \mathcal{S}_1''$, we have $\boldsymbol{x} = \boldsymbol{x}_1 + \boldsymbol{x}_2$, where $\boldsymbol{x}_1 \in \mathcal{S}_1$ and $\boldsymbol{x}_2 \in \mathcal{S}_1'$. Similarly, we can verify that $\boldsymbol{x}_1 = \boldsymbol{0}$ and thus $\boldsymbol{x} \in \mathcal{S}_1'$, i.e., $\mathcal{S}_1'' \subseteq \mathcal{S}_1'$. Therefore, $\mathcal{S}_1' = \mathcal{S}_1''$, which indicates that $\mathcal{S}_2 = \mathcal{S}_1^{\perp}$. □

LEMMA 3
Given two linear subspaces $\mathcal{S}_1, \mathcal{S}_2 \subseteq \mathbb{R}^D$ with dimensions dim$(\mathcal{S}_1)$ and dim$(\mathcal{S}_2)$, then dim$(\mathcal{S}_1 + \mathcal{S}_2) \leq$ dim$(\mathcal{S}_1)$ + dim$(\mathcal{S}_2)$.

Proof. Let $p = \dim(\mathcal{S}_1)$, $q = \dim(\mathcal{S}_1)$, $\{\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_p\}$ be a basis of $\mathcal{S}_1$ and $\{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_q\}$ be a basis of $\mathcal{S}_2$. Given any $\boldsymbol{x} \in \mathcal{S}_1 + \mathcal{S}_2$, we have that $\boldsymbol{x} = \boldsymbol{x}_1 + \boldsymbol{x}_2$, where $\boldsymbol{x}_1 = \sum_{i=1}^{p} a_i\boldsymbol{\alpha}_i \in \mathcal{S}_1$ and $\boldsymbol{x}_2 = \sum_{i=1}^{q} b_i\boldsymbol{\beta}_i \in \mathcal{S}_2$. Thus, $\boldsymbol{x} = \sum_{i=1}^{p} a_i\boldsymbol{\alpha}_i + \sum_{i=1}^{q} b_i\boldsymbol{\beta}_i$, i.e., $\boldsymbol{x}$ can be linearly represented by $\{\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_p, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_q\}$, which indicates that $\dim(\mathcal{S}_1 + \mathcal{S}_2) \leq p + q$. □

On the basis of Lemma 1, 2 and 3, we show the proof of Theorem 3 below.

THEOREM 3
Given any $\boldsymbol{f} = (f_1, \ldots, f_m)$, if each f_i has at least one effective subspace $\mathcal{V}_i \subseteq \mathbb{R}^D$ with dimension d_i such that $\mathcal{V}_i$ is orthogonal to $\mathcal{V}_j$ for any $i \neq j \in \{1, \ldots, m\}$, then $\boldsymbol{f}$ has the M-effective dimension $\vartheta_e \leq \sum_{i=1}^{m} d_i$.

Proof. If each f_i has the S-effective dimension, we have that there exist subspaces $\mathcal{V}_1, \ldots, \mathcal{V}_m \subseteq \mathbb{R}^D$ such that $\mathcal{V}_i$ is the effective subspace of f_i, where $i = 1, \ldots, m$. That is to say, for any $\boldsymbol{x} \in \mathbb{R}^D$ and each $i = 1, \ldots, m$, $f_i(\boldsymbol{x}) = f_i(\boldsymbol{x}_e^{(i)} + \boldsymbol{x}_c^{(i)}) = f_i(\boldsymbol{x}_e^{(i)})$, where $\boldsymbol{x}_e^{(i)} \in \mathcal{V}_i$ and $\boldsymbol{x}_c^{(i)} \in \mathcal{V}_i^{\perp} \subseteq \mathbb{R}^D$. Since the sum of linear subspaces is also a linear subspace, we consider the linear subspace $\mathcal{V} = \mathcal{V}_1 + \mathcal{V}_2 + \cdots + \mathcal{V}_m \subseteq \mathbb{R}^D$. Let $\mathcal{V}^{\perp}$ denote the orthogonal complement subspace of $\mathcal{V}$. We have that $\mathcal{V} + \mathcal{V}^{\perp} = \mathcal{V}_1 + \mathcal{V}_2 + \cdots + \mathcal{V}_m + \mathcal{V}^{\perp} = \mathbb{R}^D$. Since $\mathcal{V}_i$ is orthogonal to $\mathcal{V}_j$ for any $i \neq j \in \{1, \ldots, m\}$, given any $\boldsymbol{x} \in \mathbb{R}^D$, let $\boldsymbol{x}_e = \sum_{i=1}^{m} \boldsymbol{x}_e^{(i)} \in \mathcal{V}$, then $\boldsymbol{x} = \boldsymbol{x}_e + \boldsymbol{x}_c$, where $\boldsymbol{x}_c \in \mathcal{V}^{\perp}$. To prove the theorem, it is sufficient to prove that $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x}_e)$, i.e., $f_i(\boldsymbol{x}) = f_i(\boldsymbol{x}_e)$ for each $i = 1, \ldots, m$.

Without loss of generality, we only prove the case when $i = 1$, since the proof procedure for $i = 2, \ldots, m$ is as same as that for $i = 1$. For f_1 with effective subspace $\mathcal{V}_1$, on the one hand, we have that $f_1(\boldsymbol{x}) = f_1(\boldsymbol{x}_e^{(1)})$, where $\boldsymbol{x}_e^{(1)} \in \mathcal{V}_1$. On the other hand, since $\mathcal{V}^{\perp}$ is the orthogonal complement of $\mathcal{V} = \mathcal{V}_1 + \mathcal{V}_2 + \cdots + \mathcal{V}_m$, $\mathcal{V}^{\perp} = (\mathcal{V}_1 + \mathcal{V}_2 + \cdots + \mathcal{V}_m)^{\perp}$. By Lemma 1, we know that $\mathcal{V}_1$ is orthogonal to $\mathcal{V}^{\perp}$. Under the condition that $\mathcal{V}_1$ is orthogonal to $\mathcal{V}_j$ for any $j = 2, \ldots, m$, we have that $\mathcal{V}_1$ is orthogonal to $\mathcal{V}_2 + \cdots + \mathcal{V}_m + \mathcal{V}^{\perp}$. Since $\mathcal{V}_1 + \mathcal{V}_2 + \cdots + \mathcal{V}_m + \mathcal{V}^{\perp} = \mathbb{R}^D$, by Lemma 2, we know that $\mathcal{V}_1^{\perp} = \mathcal{V}_2 + \cdots + \mathcal{V}_m + \mathcal{V}^{\perp}$. Therefore, $\sum_{i=2}^{m} \boldsymbol{x}_e^{(i)} \in \mathcal{V}_1^{\perp}$ and $f_1(\boldsymbol{x}_e) = f_1(\boldsymbol{x}_e^{(1)} + \sum_{i=2}^{m} \boldsymbol{x}_e^{(i)}) = f_1(\boldsymbol{x}_e^{(1)})$. To sum up, we have that $f_1(\boldsymbol{x}) = f_1(\boldsymbol{x}_e)$ and the same conclusion holds when $i = 2, \ldots, m$.

Therefore, $\mathcal{V} = \mathcal{V}_1 + \mathcal{V}_2 + \cdots + \mathcal{V}_m \subseteq \mathbb{R}^D$ is one of the effective subspaces of $\boldsymbol{f}(\boldsymbol{x})$. By Lemma 3, we know that $\dim(\mathcal{V}) \leq \sum_{i=1}^{m} d_i$. This proves that $\boldsymbol{f}$ has the M-effective dimension $\vartheta_e \leq \dim(\mathcal{V}) \leq \sum_{i=1}^{m} d_i$. □

Proof of Theorem 4

This section presents the proof of Theorem 4 in the paper. Before presenting the proof of Theorem 4, we first show

Lemma 4 below that will be used in the poof of Theorem 4.

LEMMA 4

Given any $\boldsymbol{f} = (f_1, \ldots, f_m)\colon \mathbb{R}^D \to \mathbb{R}^m$ with M-effective dimension ϑ_e, and a random matrix $\boldsymbol{A} \in \mathbb{R}^{D\times\vartheta}$ with independent members sampled from $\mathcal{N}(0,1)$ where $\vartheta \geq \vartheta_e$, then, with probability 1, for any $\boldsymbol{x} \in \mathbb{R}^D$ there exists $\boldsymbol{y} \in \mathbb{R}^\vartheta$ such that $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{A}\boldsymbol{y})$.

Proof. If $\boldsymbol{f}$ has the M-effective dimension, then there exists an effective subspace $\mathcal{V} \subseteq \mathbb{R}^D$ of $\boldsymbol{f}$ with dimension ϑ_e. Since any $\boldsymbol{x} \in \mathbb{R}^D$ can be decomposed as $\boldsymbol{x} = \boldsymbol{x}_e + \boldsymbol{x}_c$, where $\boldsymbol{x}_e \in \mathcal{V}$, $\boldsymbol{x}_c \in \mathcal{V}^\perp$ and $\mathcal{V}^\perp$ is the orthogonal complement of $\mathcal{V}$, by the definition of M-effective subspace, we have that $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x}_e + \boldsymbol{x}_c) = \boldsymbol{f}(\boldsymbol{x}_e)$. Thus, it is sufficient to show that, with probability 1, for any $\boldsymbol{x}_e \in \mathcal{V}$ there exists $\boldsymbol{y} \in \mathbb{R}^d$ such that $\boldsymbol{f}(\boldsymbol{x}_e) = \boldsymbol{f}(\boldsymbol{A}\boldsymbol{y})$. Let $\boldsymbol{\Psi} \in \mathbb{R}^{D\times\vartheta_e}$ be a matrix whose columns form a standard orthonormal basis of $\mathcal{V}$. Thus, for any $\boldsymbol{x}_e \in \mathcal{V}$, there exists $\boldsymbol{c} \in \mathbb{R}^{\vartheta_e}$ such that $\boldsymbol{x}_e = \boldsymbol{\Psi}\boldsymbol{c}$. Let us for now assume that $\boldsymbol{\Psi}^\top\boldsymbol{A}$ has rank ϑ_e. If $\text{rank}(\boldsymbol{\Psi}^\top\boldsymbol{A}) = \vartheta_e$, there must exist $\boldsymbol{y} \in \mathbb{R}^\vartheta$ such that $(\boldsymbol{\Psi}^\top\boldsymbol{A})\boldsymbol{y} = \boldsymbol{c}$, because $\text{rank}(\boldsymbol{\Psi}^\top\boldsymbol{A}) = \text{rank}([\boldsymbol{\Psi}^\top\boldsymbol{A}, \boldsymbol{c}])$. The orthonormal projection of $\boldsymbol{A}\boldsymbol{y}$ onto $\mathcal{V}$ is given by $\boldsymbol{\Psi}\boldsymbol{\Psi}^\top\boldsymbol{A}\boldsymbol{y} = \boldsymbol{\Psi}\boldsymbol{c} = \boldsymbol{x}_e$. Thus, $\boldsymbol{A}\boldsymbol{y} = \boldsymbol{x}_e + \tilde{\boldsymbol{x}}$, where $\tilde{\boldsymbol{x}} \in \mathcal{V}^\perp$ since $\boldsymbol{x}_e$ is the orthonormal projection of $\boldsymbol{A}\boldsymbol{y}$ onto $\mathcal{V}$. Since $\mathcal{V}$ is the effective subspace of $\boldsymbol{f}$, we have $\boldsymbol{f}(\boldsymbol{A}\boldsymbol{y}) = \boldsymbol{f}(\boldsymbol{x}_e + \tilde{\boldsymbol{x}}) = \boldsymbol{f}(\boldsymbol{x}_e)$. At last, we can verify that $\text{rank}(\boldsymbol{\Psi}^\top\boldsymbol{A}) = \vartheta_e$ with probability 1 directly with the same arguments in Theorem 1. □

Let $\boldsymbol{g}(\boldsymbol{y}) = \boldsymbol{f}(\boldsymbol{A}\boldsymbol{y}) = (f_1(\boldsymbol{A}\boldsymbol{y}), \ldots, f_m(\boldsymbol{A}\boldsymbol{y}))$, where $\boldsymbol{y} \in \mathbb{R}^\vartheta$ with $\vartheta_e \leq \vartheta \leq D$. Based on Lemma 4, we prove that, with probability 1, the optimal Pareto front of $\boldsymbol{g}(\boldsymbol{y})$ is as same as that of $\boldsymbol{f}(\boldsymbol{x})$ and the optimal Pareto set of $\boldsymbol{f}(\boldsymbol{x})$ can be recovered from that of $\boldsymbol{g}(\boldsymbol{y})$, i.e., optimal Pareto front and optimal Pareto set preservation. Denote the optimal Pareto set of $\boldsymbol{f}, \boldsymbol{g}$ as $\mathcal{PS}_{\boldsymbol{f}} \subseteq \mathbb{R}^D, \mathcal{PS}_{\boldsymbol{g}} \subseteq \mathbb{R}^\vartheta$, and denote the optimal Pareto front of $\boldsymbol{f}, \boldsymbol{g}$ as $\mathcal{PF}_{\boldsymbol{f}}, \mathcal{PF}_{\boldsymbol{g}} \subseteq \mathbb{R}^m$, respectively.

THEOREM 4

Given any $\boldsymbol{f} = (f_1, \ldots, f_m)\colon \mathbb{R}^D \to \mathbb{R}^m$ with M-effective dimension ϑ_e, then, with probability 1, we have that $\{\boldsymbol{A}\boldsymbol{y} \mid \boldsymbol{y} \in \mathcal{PS}_{\boldsymbol{g}}\} \subseteq \mathcal{PS}_{\boldsymbol{f}}$ and $\mathcal{PF}_{\boldsymbol{g}} = \mathcal{PF}_{\boldsymbol{f}}$.

Proof. For any $\boldsymbol{A}\boldsymbol{y} \in \{\boldsymbol{A}\boldsymbol{y} \mid \boldsymbol{y} \in \mathcal{PS}_{\boldsymbol{g}}\}$, there exists no other solution in $\mathbb{R}^\vartheta$ which dominates $\boldsymbol{y}$. If $\boldsymbol{A}\boldsymbol{y} \notin \mathcal{PS}_{\boldsymbol{f}}$, then there must exist $\boldsymbol{x}' \in \mathbb{R}^D$ such that $\boldsymbol{x}' \prec_{\boldsymbol{f}} \boldsymbol{A}\boldsymbol{y}$. By Lemma 4, we know that, with probability 1, there exists $\boldsymbol{y}' \in \mathbb{R}^\vartheta$ such that $\boldsymbol{g}(\boldsymbol{y}') = \boldsymbol{f}(\boldsymbol{A}\boldsymbol{y}') = \boldsymbol{f}(\boldsymbol{x}')$. Since $\boldsymbol{g}(\boldsymbol{y}) = \boldsymbol{f}(\boldsymbol{A}\boldsymbol{y})$, we have that $\boldsymbol{y}' \prec_{\boldsymbol{g}} \boldsymbol{y}$, which make a contradiction. Therefore, $\boldsymbol{A}\boldsymbol{y} \in \mathcal{PS}_{\boldsymbol{f}}$ and $\{\boldsymbol{A}\boldsymbol{y} \mid \boldsymbol{y} \in \mathcal{PS}_{\boldsymbol{g}}\} \subseteq \mathcal{PS}_{\boldsymbol{f}}$ with probability 1.

On the one hand, since $\boldsymbol{g}(\boldsymbol{y}) = \boldsymbol{f}(\boldsymbol{A}\boldsymbol{y}) \in \mathbb{R}^m$ and $\{\boldsymbol{A}\boldsymbol{y} \mid \boldsymbol{y} \in \mathcal{PS}_{\boldsymbol{g}}\} \subseteq \mathcal{PS}_{\boldsymbol{f}}$ with probability 1, we have that $\mathcal{PF}_{\boldsymbol{g}} \subseteq \mathcal{PF}_{\boldsymbol{f}}$ with probability 1. On the other hand, given any $\boldsymbol{z} \in \mathcal{PF}_{\boldsymbol{f}}$, there exists $\boldsymbol{x} \in \mathcal{PS}_{\boldsymbol{f}}$ such that $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{z}$. By Lemma 4, we know that, with probability 1, there exists $\tilde{\boldsymbol{y}} \in \mathbb{R}^\vartheta$ such that $\boldsymbol{g}(\tilde{\boldsymbol{y}}) = \boldsymbol{f}(\boldsymbol{A}\tilde{\boldsymbol{y}}) = \boldsymbol{f}(\boldsymbol{x})$. Therefore, $\boldsymbol{g}(\tilde{\boldsymbol{y}}) = \boldsymbol{z}$ with probability 1. If $\tilde{\boldsymbol{y}} \notin \mathcal{PS}_{\boldsymbol{g}}$, there must exist $\hat{\boldsymbol{y}} \in \mathbb{R}^\vartheta$ such that $\hat{\boldsymbol{y}} \prec_{\boldsymbol{g}} \tilde{\boldsymbol{y}}$, and thus $\boldsymbol{g}(\hat{\boldsymbol{y}}) = \boldsymbol{f}(\boldsymbol{A}\hat{\boldsymbol{y}})$ is strictly better than $\boldsymbol{g}(\tilde{\boldsymbol{y}}) = \boldsymbol{z}$, which contradicts the condition of $\boldsymbol{z} \in \mathcal{PF}_{\boldsymbol{f}}$. Therefore, $\tilde{\boldsymbol{y}} \in \mathcal{PS}_{\boldsymbol{g}}$. Now, we get that $\boldsymbol{z} \in \mathcal{PF}_{\boldsymbol{g}}$ and $\mathcal{PF}_{\boldsymbol{f}} \subseteq \mathcal{PF}_{\boldsymbol{g}}$ with probability 1. To sum up, $\mathcal{PF}_{\boldsymbol{g}} = \mathcal{PF}_{\boldsymbol{f}}$ with probability 1. □

Proof of Theorem 5

This section presents the proof of Theorem 5 in the paper.

THEOREM 5

Given any $\boldsymbol{f} = (f_1, \ldots, f_m)\colon \mathbb{R}^D \to \mathbb{R}^m$ with M-effective dimension ϑ_e, assume that $\mathcal{M}$ can find the $\mathcal{PS}_{\boldsymbol{f}}$ and $\mathcal{PF}_{\boldsymbol{f}}$ with time complexity $\mathcal{O}(\phi(D, m))$, then, with probability 1, ReMO equipped with the same algorithm $\mathcal{M}$ can find the $\mathcal{PS}_{\boldsymbol{f}}$ and $\mathcal{PF}_{\boldsymbol{f}}$ with time complexity $\mathcal{O}(\phi(\vartheta, m))$, where $\phi(d, m)$ is a monotone increasing function with respect to the dimension of solution space d.

Proof. Since $\mathcal{M}$ can find the $\mathcal{PS}_{\boldsymbol{f}}$ and $\mathcal{PF}_{\boldsymbol{f}}$ of $\boldsymbol{f}$ with time complexity $\mathcal{O}(\phi(D, m))$, from the procedure of ReMO as depicted in Algorithm 1 in the paper, we have that $\mathcal{M}$ can find the $\mathcal{PS}_{\boldsymbol{g}}$ and $\mathcal{PF}_{\boldsymbol{g}}$ of $\boldsymbol{g}$ with time complexity $\mathcal{O}(\phi(\vartheta, m))$.

By Theorem 4, i.e., $\{\boldsymbol{A}\boldsymbol{y} \mid \boldsymbol{y} \in \mathcal{PS}_{\boldsymbol{g}}\} \subseteq \mathcal{PS}_{\boldsymbol{f}}$ and $\mathcal{PF}_{\boldsymbol{g}} = \mathcal{PF}_{\boldsymbol{f}}$ with probability 1, we can conclude that, with probability 1, ReMO equipped with $\mathcal{M}$ can find the $\mathcal{PS}_{\boldsymbol{f}}$ and $\mathcal{PF}_{\boldsymbol{f}}$ of $\boldsymbol{f}$ with time complexity $\mathcal{O}(\phi(\vartheta, m))$. □

Proof of Theorem 6

This section presents the proof of Theorem 6 in the paper. Before presenting the proof of Theorem 6, we first show Lemma 5 below that will be used in the poof of Theorem 6. A matrix $\boldsymbol{M} \in \mathbb{R}^{D\times D}$ is called the orthogonal matrix if and only if $\boldsymbol{M}\boldsymbol{M}^\top = \boldsymbol{M}^\top\boldsymbol{M} = \boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix.

LEMMA 5

Given an orthogonal matrix $\boldsymbol{M} \in \mathbb{R}^{D\times D}$, let $\boldsymbol{A} \in \mathbb{R}^{D\times\vartheta}$ be a random matrix with independent members sampled from $\mathcal{N}(0,1)$, then, the members of $\boldsymbol{M}^{-1}\boldsymbol{A}$ are also i.i.d. random variables sampled from $\mathcal{N}(0,1)$.

Proof. Since $\boldsymbol{M} \in \mathbb{R}^{D\times D}$ is an orthogonal matrix, namely, $\boldsymbol{M}\boldsymbol{M}^\top = \boldsymbol{M}^\top\boldsymbol{M} = \boldsymbol{I}$, we have that $\boldsymbol{M}^{-1} = \boldsymbol{M}^\top$. It is sufficient to prove that the members of $\boldsymbol{M}^\top\boldsymbol{A}$ are also i.i.d. random variables sampled from $\mathcal{N}(0,1)$.

Let $m_{i,j}$ and $a_{i,j}$ denote the members that lie in the i-th row and j-th column of $\boldsymbol{M}^\top$ and $\boldsymbol{A}$ respectively, then the member lying in the i-th row and j-th column of $\boldsymbol{M}^\top\boldsymbol{A}$ is $\sum_{k=1}^D m_{i,k}a_{k,j}$. Since $\boldsymbol{A}$ is a random matrix with independent members sampled from $\mathcal{N}(0,1)$, $\sum_{k=1}^D m_{i,k}a_{k,j}$ is also a random variable sampled from Gaussian distribution with expectation

$$\mathbb{E}[\sum_{k=1}^{D} m_{i,k}a_{k,j}] = \sum_{k=1}^{D} m_{i,k}\mathbb{E}[a_{k,j}] = 0,$$

and variance

$$\mathbb{V}[\sum_{k=1}^{D} m_{i,k}a_{k,j}] = \sum_{k=1}^{D} m_{i,k}^2\mathbb{V}[a_{k,j}] = \sum_{k=1}^{D} m_{i,k}^2 = 1,$$

where the last equality is by the property of orthogonal matrix. Consider another member $\sum_{k=1}^{D} m_{p,k}a_{k,q}$ that lies in the p-th row and q-th column of $\boldsymbol{M}^\top \boldsymbol{A}$, where $p \neq i$ or $q \neq j$. If $p = i$ and $q \neq j$, by the definition of $\boldsymbol{A}$, we know that $\sum_{k=1}^{D} m_{i,k}a_{k,j}$ and $\sum_{k=1}^{D} m_{p,k}a_{k,q}$ are independent. If $p \neq i$ and $q = j$, consider the covariance between $\sum_{k=1}^{D} m_{i,k}a_{k,j}$ and $\sum_{k=1}^{D} m_{p,k}a_{k,j}$

$$\begin{aligned}
&\mathtt{Cov}[\sum_{k=1}^{D} m_{i,k}a_{k,j}, \sum_{k=1}^{D} m_{p,k}a_{k,j}] \\
&= \mathbb{E}[(\sum_{k=1}^{D} m_{i,k}a_{k,j}) \cdot (\sum_{k=1}^{D} m_{p,k}a_{k,j})] \\
&\qquad\qquad - \mathbb{E}[\sum_{k=1}^{D} m_{i,k}a_{k,j}] \cdot \mathbb{E}[\sum_{k=1}^{D} m_{p,k}a_{k,j}] \\
&= \mathbb{E}[(\sum_{k=1}^{D} m_{i,k}a_{k,j})(\sum_{k=1}^{D} m_{p,k}a_{k,j})] \\
&= \mathbb{E}[\sum_{k=1}^{D} m_{i,k}m_{p,k}a_{k,j}^2] \\
&= \sum_{k=1}^{D} m_{i,k}m_{p,k} = 0,
\end{aligned}$$

where the third equality is by $\mathbb{E}[a_{k,j}a_{k',j}] = \mathbb{E}[a_{k,j}] \cdot \mathbb{E}[a_{k',j}] = 0$ for $k \neq k'$, the fourth equality is by $\mathbb{E}[a_{k,j}^2] = \mathbb{V}[a_{k,j}^2] + \mathbb{E}^2[a_{k,j}] = 1$, and the last equality is by the property of orthogonal matrix. Since $\sum_{k=1}^{D} m_{i,k}a_{k,j}$ and $\sum_{k=1}^{D} m_{p,k}a_{k,q}$ are both Gaussian random variable, $\sum_{k=1}^{D} m_{i,k}a_{k,j}$ and $\sum_{k=1}^{D} m_{p,k}a_{k,q}$ are independent, which proves the lemma. □

On the basis of Lemma 5, we show that ReMO possesses the property of rotation perturbation invariance.

THEOREM 6
Given any $\boldsymbol{f} = (f_1, \ldots, f_m)\colon \mathbb{R}^D \to \mathbb{R}^m$ with M-effective dimension ϑ_e and an orthogonal matrix $\boldsymbol{M} \in \mathbb{R}^{D\times D}$, let $\boldsymbol{f}_{\boldsymbol{M}}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{M}\boldsymbol{x}) = (f_1(\boldsymbol{M}\boldsymbol{x}), \cdots, f_m(\boldsymbol{M}\boldsymbol{x}))$ denote the rotation perturbation of $\boldsymbol{f}$, if we apply ReMO to optimize $\boldsymbol{f}_{\boldsymbol{M}}$, ReMO can still find the $\mathcal{PS}_{\boldsymbol{f}}$ and $\mathcal{PF}_{\boldsymbol{f}}$ of $\boldsymbol{f}$ with probability 1.

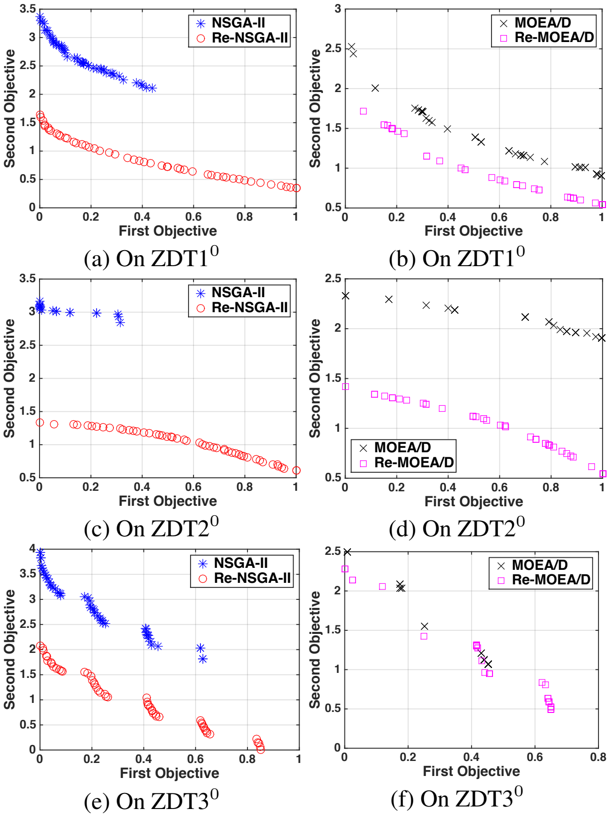
Proof. During the procedure of ReMO optimizing $\boldsymbol{f}_{\boldsymbol{M}}$, we use the random matrix $\boldsymbol{M}^{-1}\boldsymbol{A} \in \mathbb{R}^{D\times\vartheta}$ instead of $\boldsymbol{A}$. By Theorem 2, Theorem 4 and Lemma 5, we know that ReMO can find the $\mathcal{PS}_{\boldsymbol{f}_{\boldsymbol{M}}}$ and $\mathcal{PF}_{\boldsymbol{f}_{\boldsymbol{M}}}$ of $\boldsymbol{f}_{\boldsymbol{M}}$ with probability 1. Furthermore, noticing that $\boldsymbol{f}_{\boldsymbol{M}}(\boldsymbol{M}^{-1}\boldsymbol{A}\boldsymbol{y}) = \boldsymbol{f}(\boldsymbol{M}\boldsymbol{M}^{-1}\boldsymbol{A}\boldsymbol{y}) = \boldsymbol{f}(\boldsymbol{A}\boldsymbol{y})$, we have that $\mathcal{PS}_{\boldsymbol{f}_{\boldsymbol{M}}} = \mathcal{PS}_{\boldsymbol{f}}$ and $\mathcal{PF}_{\boldsymbol{f}_{\boldsymbol{M}}} = \mathcal{PF}_{\boldsymbol{f}}$, which proves the theorem. □

Additional Experimental Results

On Functions with M-Effective Dimensions

We verify the effectiveness of ReMO empirically on three high-dimensional bi-objective (i.e., $m = 2$) optimization testing functions ZDT1^0, ZDT2^0 and ZDT3^0 with low M-effective dimensions. They are constructed based on the first three bi-objective functions ZDT1, ZDT2 and ZDT3 introduced in (Zitzler, Deb, and Thiele 2000). The dimensions of ZDT1, ZDT2 and ZDT3 are all 30. ZDT1^0, ZDT2^0 and ZDT3^0 are constructed to meet the M-effective dimension assumption as follows: ZDT1, ZDT2 and ZDT3 are embedded into a D-dimensional solution space $\mathcal{X} \subseteq \mathbb{R}^D$ with $D \gg 30$. The embedding is done by firstly adding additional $D - 30$ dimensions, but the additional dimensions have no effects on the function value; secondly, the embedded functions are rotated via an orthogonal matrix. Thus, the dimensions of ZDT1^0, ZDT2^0 and ZDT3^0 are D while their M-effective dimensions are 30. For these bi-objective testing functions, ZDT1^0 has a convex optimal Pareto front, ZDT2^0 has a non-convex optimal Pareto front, and ZDT3^0 has a discontinuous optimal Pareto front. The second objective functions of ZDT1^0, ZDT2^0 and ZDT3^0 are all non-convex.

For the practical issues in experiments, we set the high-dimensional solution space $\mathcal{X} = [-1, 1]^D$ and the low-dimensional solution space $\mathcal{Y} = [-1, 1]^\vartheta$, instead of $\mathbb{R}^D$ and $\mathbb{R}^\vartheta$. To implement the MO algorithm in $\mathcal{Y}$, since there may exist $\boldsymbol{y}' \in \mathcal{Y}$ s.t. $\boldsymbol{A}\boldsymbol{y}' \notin \mathcal{X}$ and thus $\boldsymbol{f}$ cannot be evaluated at point $\boldsymbol{A}\boldsymbol{y}'$. To address this problem, we use Euclidean projection, i.e., $\boldsymbol{A}\boldsymbol{y}'$ is projected to $\mathcal{X}$ when it is outside $\mathcal{X}$ by $P_{\mathcal{X}}(\boldsymbol{A}\boldsymbol{y}') = \mathrm{argmin}_{\boldsymbol{x}\in\mathcal{X}} \|\boldsymbol{x} - \boldsymbol{A}\boldsymbol{y}'\|_2$. We employ two well-



(a) On ZDT1^0 (b) On ZDT1^0

(c) On ZDT2^0 (d) On ZDT2^0

(e) On ZDT3^0 (f) On ZDT3^0

Figure 1: Comparing the approximate optimal Pareto front each algorithm found with dimension $D = 10000$.

known derivative-free MO optimization methods to minimize ZDT1^0, ZDT2^0 and ZDT3^0: non-dominated sorting genetic algorithm II (NSGA-II) (Deb et al. 2002), and multi-objective evolutionary algorithm based on decomposition (MOEA/D) (Zhang and Li 2007). The implementations of them are both by their authors. When applying the random embedding technique to these optimization methods, we denote them by the prefix "Re-". Thus, we have combinations including Re-NSGA-II and Re-MOEA/D. NSGA-II and MOEA/D are compared with Re-NSGA-II and Re-MOEA/D on these three testing functions with dimension $D = 10000 (\gg \vartheta_e = 30)$. We set the function evaluation budget $n = 0.3D = 3000$ and the upper bound of M-effective dimension $\vartheta = 50 > \vartheta_e = 30$. The approximate optimal Pareto front each algorithm found for ZDT1^0, ZDT2^0 and ZDT3^0 is depicted in Figure 1.

Figure 1 shows that, for high-dimensional MO functions with low M-effective dimensions, the performance of directly applying derivative-free MO methods in high-dimensional solution space is unsatisfying. Meanwhile, ReMO that enables optimization to be implemented in the low-dimensional solution space has the better performance on all the testing functions whenever we choose Re-NSGA-II or Re-MOEA/D, which implies that ReMO is effective and its effectiveness is general. To be specific, Figure 1 reflects the effectiveness of ReMO from two aspects. First, ReMO can find the significantly higher-quality solutions on both objectives, especially in (a), (c), (d), (e), since some solutions found by MO are dominated by those of ReMO. Second, the distribution of approximate optimal Pareto front found by ReMO is more diverse and uniform, especially in (a), (c), (e), (f), which means that ReMO can find the better solutions that reach different optimal balances of the objective functions.

References

Deb, K.; Agrawal, S.; Pratap, A.; and Meyarivan, T. 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation* 6(2):182–197.

Wang, Z.; Zoghi, M.; Hutter, F.; Matheson, D.; and Freitas, N. D. 2013. Bayesian optimization in high dimensions via random embeddings. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, 1778–1784.

Wang, Z.; Zoghi, M.; Hutter, F.; Matheson, D.; and Freitas, N. D. 2016. Bayesian optimization in a billion dimensions via random embeddings. *Journal of Artificial Intelligence Research* 55:361–387.

Zhang, Q., and Li, H. 2007. MOEA/D: A multiobjective evolutionary algorithm based on decomposition. *IEEE Transactions on Evolutionary Computation* 11(6):712–731.

Zitzler, E.; Deb, K.; and Thiele, L. 2000. Comparison of multiobjective evolutionary algorithms: Empirical results. *Evolutionary Computation* 8(2):173–195.