Contemporary Mathematics
https://ojs.wiserpub.com/index.php/CM/

Research Article

A Theoretical and Experimental Exploration in Permutation Randomization on Nonsmooth Nonconvex Optimization

Arif Hassan Zidan[1], Afrar Jahin[1], Yu Bao[2], Tianming Liu[3], Wei Zhang[1*]

[1] School of Computer and Cyber Sciences, Augusta University, Augusta, GA, 30912, United States
[2] Department of Graduate Psychology, James Madison University, Harrisonburg, VA, 22807, United States
[3] School of Computing, University of Georgia, Athens, GA, 30912, United States
E-mail: wzhang2@augusta.edu

Received: 31 August 2025; **Revised:** 21 October 2025; **Accepted:** 5 November 2025

Abstract: While gradient-based optimizers that incorporate randomization often demonstrate superior performance on complex optimizations, the theoretical foundations of this advantage remain underexplored. A central question arises: What role does randomization play in dimension-free, nonsmooth, nonconvex optimization? To address this gap, we examine both the theoretical and empirical impact of permutation randomization within gradient-based optimization frameworks, using it as a representative case to investigate broader implications. From a theoretical perspective, our analysis reveals that permutation randomization disrupts the shrinkage behavior characteristic of gradient-based optimizers, allowing for continued progress toward the global optimum with sufficient iterations. Moreover, we prove that permutation randomization preserves the convergence rate of the underlying optimizer. Empirically, we conduct extensive numerical experiments comparing permutation-randomized optimizers with three baseline methods. These experiments span tasks such as training deep neural networks with stacked architectures and optimizing noisy objective functions. The results not only support our theoretical findings but also demonstrate the practical benefits of permutation randomization. In summary, this work provides both rigorous theoretical justification and compelling empirical evidence for the effectiveness of permutation randomization, establishing a foundation for extending such analyses to broader implications of randomized strategies.

***Keywords*:** permutation randomization, gradient descent, nonsmooth nonconvex optimization, machine learning, Deep Neural Networks (DNNS)

MSC: 05A05, 90C30, 60C05

1. Introduction

Optimization problems in machine learning are typically addressed using numerical optimizers that rely on either full gradients, computed over the entire dataset, or stochastic gradients, calculated from individual data points or mini-batches. Full-gradient methods deliver theoretical guarantees of eventual convergence, while stochastic gradient methods provide advantages in terms of convergence speed and computational efficiency [1–3]. In parallel, a growing body of research has explored hybrid optimization strategies that combine full and stochastic gradients with additional techniques such as

Copyright ©2026 Wei Zhang, et al.
DOI: https://doi.org/10.37256/cm.7220268464
This is an open-access article distributed under a CC BY license
(Creative Commons Attribution 4.0 International License)
https://creativecommons.org/licenses/by/4.0/

noise injection, dynamic batch sizing, and advanced sampling methods, and an interactive approach for multi-objective optimization. These approaches have been shown to achieve favorable convergence rates and improved accuracy in practice [4–11].

With the rapid advancement of optimization theory over the past several decades, significant research efforts have focused on nonsmooth nonconvex optimization problems. Notable approaches include gradient sampling methods [12–15], bundle methods [16–18], and subgradient methods [19–21]. In particular, Burke et al. developed several gradient-based optimizers for nonsmooth nonconvex objective functions by sampling gradients at a set of randomly generated nearby points [12, 13]. These sampled gradients are then used to construct a local search direction that approximates the ε-steepest descent direction, where ε denotes the sampling radius [15]. Further contributions by Daniilidis and Drusvyatskiy examined the limitations of traditional subgradient methods in nonsmooth nonconvex settings. They demonstrated that the standard subgradient method may fail to converge to any Clarke stationary point, even when applied to Lipschitz-continuous functions [20]. Moreover, recent work has introduced a randomized variant of Goldstein's subgradient method, establishing a dimension-independent complexity bound for finding a (δ, ε)-Goldstein stationary point. This work builds on foundational concepts from Clarke's generalized gradients [3, 22–25].

Despite significant progress with state-of-the-art optimizers, efficiently approximating the global optimum of nonsmooth nonconvex objective functions remains a major challenge [3]. Even advanced methods such as Adaptive Moment Estimation (ADAM) [26] and Stochastic Variance Reduced Gradient (SVRG) [27–29] often struggle with these complex landscapes, particularly in high-dimensional settings. A notable example arises in the context of Deep Neural Networks (DNNs), whose objective functions are typically nonsmooth and nonconvex, and only Lipschitz continuous (see Definition 1) [3]. This is largely due to the depth of DNNs and the use of non-differentiable activation functions, such as Rectified Linear Units (ReLUs) [3, 30, 31].

Importantly, although it has been demonstrated that deterministic, dimension-free, first-order optimizers cannot outperform randomized methods in smooth or convex regimes [3, 25, 32–34], the complexity of nonsmooth nonconvex problems suggests that randomization could play a critical role in improving optimization outcomes. Therefore, developing a theoretical understanding of why and how randomization enhances the performance of gradient-based optimizers in these challenging settings is both timely and necessary [3]. In this work, we aim to contribute to this understanding by analyzing the efficacy of permutation randomization in nonsmooth nonconvex optimization and validating our theoretical findings through comprehensive numerical experiments.

Our Contributions: This work presents comprehensive theoretical and experimental studies to uncover the advancements of permutation randomization, providing a partial answer to the fundamental question raised in [3]. Our key contributions are summarized as follows:

1. Limitation of Gradient-Based Optimizers without Randomization (Theorem 1): Considering a gradient-based optimizer as an operator (see Definition 2) [35, 36], we show that such optimizers usually converge to a fixed point or a stationary point, e.g., a Clarke or (δ, ε)-Goldstein stationary point. Crucially, without randomization, the optimizer cannot guarantee to approximate a closed cube of the global optimum (Definitions 5 and 6).

2. Advancement in Permutation Randomization (Theorem 2): We theoretically establish that incorporating permutation randomization enables collectively covering the entire domain of the objective function over sufficiently large iterations. This property makes it feasible to approximate a closed cube (Definition 6) that contains the global optimum.

3. Permutation Randomization Preserves Convergence Rate (Theorem 3): A natural concern is whether permutation randomization adversely affects the convergence rate of the underlying optimizer. We demonstrate that it does not: the norm of the permutation operator (Definition 3) is 1, implying that it preserves the convergence rate of the incorporated optimizer. Moreover, we provide a theoretical upper bound for this convergence rate in Theorem 3.

4. Experimental Validation: Various numerical experiments demonstrate that integrating permutation randomization into ADAM [26] consistently improves accuracy compared to three peer optimizers. Additionally, in noisy optimization tasks, ADAM with permutation randomization outperforms all baselines, further validating our theoretical insights.

Full proofs and extended analysis can be found in Appendices A and B.

2. Preliminaries

This section introduces the fundamental optimization problem investigated in this work, along with the key definitions and assumptions that underpin the subsequent analyses. We consider the following minimization problem:

$$\min_{x\subseteq\mathbb{R}^D} f(x). \tag{1}$$

We assume that the real objective function $f:\mathbb{R}^D\to\mathbb{R}$ is Lipschitz continuous (please refer to Definition 1) over a closed set I and $x=\{x_1,\ x_2,\ \cdots,\ x_D\}\subseteq I$. In addition, we define a closed subset $I'\subseteq I$, within which f is convex /smooth, while it is nonsmooth and nonconvex on the remaining $I\backslash I'$. Notably, to support the theoretical analyses, we first introduce a set of mathematical symbols, summarized in Table 1, that will be used throughout this work.

Table 1. The definitions of mathematical symbols

Symbols	Definitions
$\mathscr{G}$	The gradient operator
$\mathscr{G}^t$	The gradient operator in t iterations
$\mathscr{R}$	The randomization operator
$\mathscr{R}^t$	The randomization operator in t iterations
I	The domain of an objective function
I_0	An initial set
T	The maximum iteration
$B(x,\ \delta)$	A closed cube with center x and radius $\delta>0$
D	Dimensionality of the objective function variable

Furthermore, the following assumptions and definitions serve as essential prerequisites for the theoretical analyses of permutation randomization in nonsmooth nonconvex optimization problems.

Assumption 1 For a set of points $\{x_1,\ x_2,\ \cdots,\ x_D\}\subseteq I$, representing all dimensions of the objective function variable, we assume the existence of a closed cube that covers each point. Specifically, for each coordinate x_i, there exists a closed cube $B(x_i,\ \delta_i),\ 1\leq i\leq D,\ \delta_i>0$, such that $x_1\in B(x_1,\ \delta_1),\ x_2\in B(x_2,\ \delta_2),\ \cdots,\ x_D\in B(x_D,\ \delta_D)$. Notably, x_i is the center of closed cube $B(x_i,\ \delta_i)$ and δ_i denotes the radius.

Notably, Assumption 1 is novel in gradient-based optimization. To explain the effectiveness of randomization using fundamental concepts and theorems from set theory and functional analysis, we adopt the assumption that a single point should be contained within a closed cube.

Definition 1 (Lipschitz Continuous Function) For any $x,\ y\subseteq\mathbb{R}^D$, $\|f(x)-f(y)\|\leq L\|x-y\|$, $L\in\mathbb{R}$ and $L<\infty$, we call f Lipschitz continuity and can rewrite as $f\in Lip(I)$.

Definition 2 (Gradient-based Optimization Operator) To conveniently employ Real Analysis and Functional Analysis for theoretical analyses, we denote a gradient-based optimizer without randomization as an operator: $\mathscr{G}:\mathbb{R}\to\mathbb{R}^D$. $\mathscr{G}$ can represent any gradient-based optimizer without randomization, such as Gradient Descent (GD) [37], Nesterov accelerated gradient descent [38], and ADAM [26]. For instance, to optimize the objective function f using GD, if $x\in\mathbb{R}$, we have $\mathscr{G}\cdot f(x)=x-\alpha\cdot\nabla f(x)$, and α denotes the step size of GD [37].

Definition 3 (Permutation Randomization Operator) We define permutation randomization as an operator: $\mathscr{R}:\mathbb{R}^D\to\mathbb{R}^D$, $D\in\mathbb{N}$, $D>1$. Specifically, when applied to a vector in three dimensions, permutation randomization reorders all components. For example: $\mathscr{R}\cdot\{x_1,\ x_2,\ x_3\}\to\{x_3,\ x_1,\ x_2\}$.

Definition 4 (An Initial Set) We initialize a closed interval $I_0 = \{x_{1,\,0},\, x_{2,\,0},\, \cdots,\, x_{D,\,0}\} \subseteq I$ for the gradient-based optimization operator $\mathscr{G}$, and begin optimizing the objective function f. For example, let $D = 3$ and $I_0 = \{10,\, 20,\, 60\}$ be the initial set for the gradient-based optimizer.

Definition 5 (The Iterative Format of Gradient-based Optimization and Permutation Randomization Operator) Suppose t is the current iteration, we denote the iterative format of the gradient-based optimization and permutation randomization operator as $\mathscr{G}^t$ and $\mathscr{R}^t$, respectively. Naturally, we can represent the output of a gradient-based optimizer in iteration t as a closed interval $\mathscr{G}^t \cdot f(I_0) \to \{x_{1,\,t},\, x_{2,\,t},\, \cdots,\, x_{D,\,t}\}$. Similarly, the output of randomization applied to a set of points represents $\mathscr{R}^t \cdot I_0 \to \{\hat{x}_{1,\,t},\, \hat{x}_{2,\,t},\, \cdots,\, \hat{x}_{D,\,t}\}$. For example, $\mathscr{G}^2 \cdot f(I_0) = \mathscr{G} \cdot \mathscr{G} \cdot f(I_0) = \{x_{1,\,2},\, x_{2,\,2},\, \cdots,\, x_{D,\,2}\}$ and $\mathscr{R}^2 \cdot \mathscr{G}^2 \cdot f(I_0) = \mathscr{R} \cdot \mathscr{R} \cdot \mathscr{G}^2 \cdot f(I_0) \cdot \{x_{1,\,2},\, x_{2,\,2},\, \cdots,\, x_{D,\,2}\} = \{\hat{x}_{1,\,2},\, \hat{x}_{2,\,2},\, \cdots,\, \hat{x}_{D,\,2}\}$.

Definition 6 (The Closed Cubes of Global Optimum) The closed cubes of the global optimum are represented as a series of closed cubes as $B(x_{1,\,gbest},\, \delta_1),\, B(x_{2,\,gbest},\, \delta_2),\, \cdots,\, B(x_{D,\,gbest},\, \delta_D) \subseteq I$.

Definition 7 (A Class of Closed Cubes Cover The Gradient) According to Definition 5 and Assumption 1, any gradient-based optimizer incorporating permutation randomization $\mathscr{R}$ applied on $f(I_0)$ in iterations t is denoted as $\mathscr{R}^t \cdot \mathscr{G}^t \cdot f(I_0) = \{\hat{x}_{1,\,t},\, \hat{x}_{2,\,t},\, \cdots,\, \hat{x}_{D,\,t}\}$. Naturally, a set of closed cubes that cover these points can be denoted as $B(\hat{x}_{1,\,t},\, \delta_{1,\,t}),\, B(\hat{x}_{2,\,t},\, \delta_{2,\,t}),\, \cdots,\, B(\hat{x}_{D,\,t},\, \delta_{D,\,t})$.

3. The limitation of conventional gradient-based optimizers

In this section, we theoretically demonstrate that conventional gradient-based optimizers (i.e., those without randomization) have limited capability to explore the global optimum once the number of iterations becomes sufficiently large.

In particular, Lemma 1 establishes a rigorous prerequisite for Theorem 1. In addition, Theorem 1 demonstrates that any gradient-based optimizer without randomization can approximate a closed cube containing the global optimum within a maximum of T iterations. The result shows that, in nonsmooth nonconvex settings, such optimizers progressively narrow their search region and ultimately become trapped at the nearest stationary point-for example, a $(\delta,\, \varepsilon)$-Goldstein stationary point [22–25], after a sufficiently large number of iterations t.

Intuitively, this showcases that the gradient-based optimizers tend to "*lock in*" too quickly in the late stage: they focus on a small local region of the search space and lose the ability to escape toward the true global solution.

Lemma 1 (Contraction Property of Gradient-based Optimizer without Randomization) According to Definitions 2, 3, 4, and 5, a gradient-based optimizer without randomization $\mathscr{G} : \mathbb{R} \to \mathbb{R}^D$ within iterations t denoted as $\mathscr{G} \cdot f(I_0) = \{x_{1,\,t},\, x_{2,\,t},\, \cdots,\, x_{D,\,t}\},\ \left\|\mathscr{G}^{t+1}\right\| \leq \left\|\mathscr{G}^t\right\|,\, \forall t \in [1,\, T]$ holds.

Theorem 1 (Limitation of Gradient-based Optimizer without Randomization) According to Definitions 2, 3, 4, 5, and 6, suppose current iteration as t, maximum iteration as T, and initialized interval as I_0, if a gradient-based optimizer without randomization $\mathscr{G}$ can approximate the closed cubes of global optimum $B(x_{i,\,gbest},\, \delta_i),\, 1 \leq i \leq D,\, \delta_i > 0$, for all closed cubes of global optimum $B(x_{1,\,gbest},\, \delta_1),\, B(x_{2,\,gbest},\, \delta_2),\, \cdots,\, B(x_{D,\,gbest},\, \delta_D) \subseteq \mathscr{G}^t \cdot f(I_0)$ must hold.

The detailed proof of Lemma 1 and Theorem 1 can be viewed in Appendix A, Supplementary Material.

4. Theoretical analyses on permutation randomization

In this section, we theoretically analyze the advantages of permutation randomization. Unlike the gradient-based optimizer without randomization, the gradient-based optimizer incorporating permutation randomization guarantees to approximate the closed cube of global optimum, only considering that the maximum iteration T is sufficiently large.

Figure 1 illustrates the contrasting outcomes of applying a gradient-based optimizer with and without coordinate randomization. Figure 1a demonstrates that a conventional gradient-based optimizer often becomes trapped in multiple local optima, primarily due to the vanishing gradient phenomenon near these points. In contrast, Figure 1b depicts the behavior of a gradient-based optimizer enhanced with coordinate randomization. One of the key advantages of

permutation randomization is its capability to disrupt the contraction or shrinkage property of conventional gradient-based optimizers by reordering all coordinates. This reordering provides an opportunity to escape local optima and potentially cover the entire domain of the objective function, thereby approaching the global optimum. In addition, the following Lemmas 2 and 3 are fundamental to our theoretical analyses employing real analysis [39] and functional analysis [35]. These Lemmas are representative theorems in set theory [39].

Lemma 2 (Heine-Borel Theorem) [39] Assume Γ is a close and bounded set and $\{g_i\}_{i=1}^{K} = G$ is a closed set. Then $\bigcup_{i=1}^{K} g_i \supseteq \Gamma$, and $\overline{\overline{G}} = \aleph_0$.

Lemma 3 (Vitali Covering Theorem) [39] Assume $\{B_i\}_{i=1}^{n}$ are closed sets and $\forall B_i \cap B_j = \emptyset$, $i \neq j$, $\Gamma \subseteq \mathbb{R}$, and $m^\star(\Gamma) < \infty$, if $m^\star(\Gamma \backslash \bigcup_{i=1}^{n} B_i) < \varepsilon$, $\forall \varepsilon > 0$, holds, $\{B_i\}_{i=1}^{n}$ defines a Vitali Covering of Γ. And $m^\star(\cdot)$ denotes a Lebesgue outer measure.

Importantly, we present Theorem 2 as follows to demonstrate that an optimizer incorporating permutation randomization guarantees that all output intervals such as $\mathscr{R}^t \cdot \mathscr{G}^t \cdot f(I_0) \to [\hat{x}_t, \hat{y}_t]$ (please refer to Definition 7) can cover a closed cube of the global optimum.

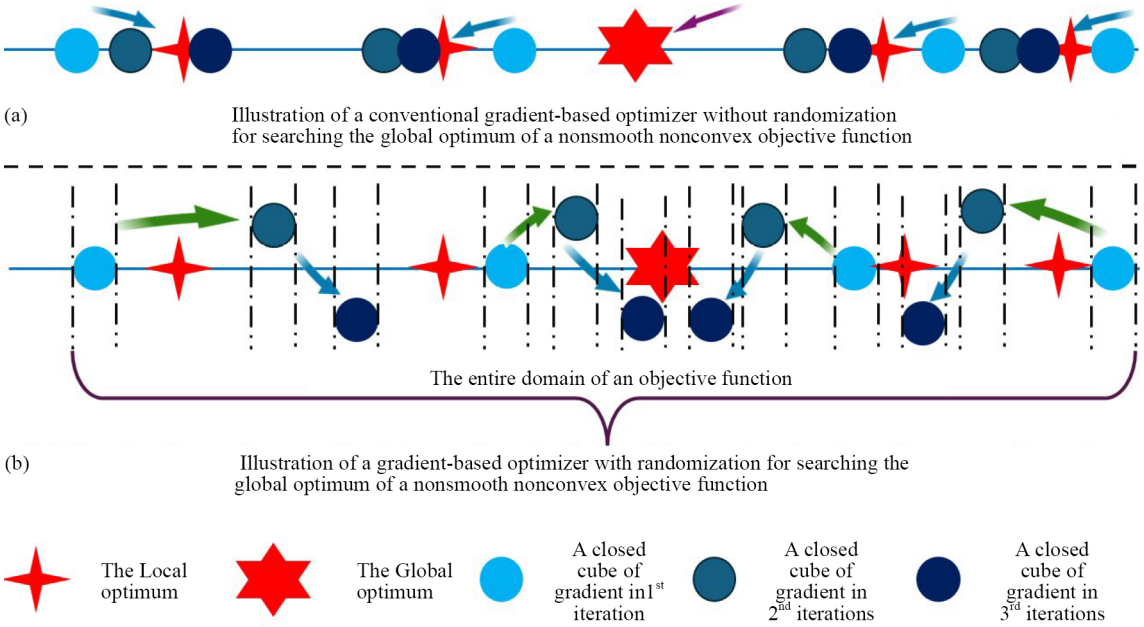


Figure 1. An illustration of a gradient-based optimizer incorporating permutation randomization can cover a closed cube of global optimum when the maximum iteration T is sufficiently large

Theorem 2 (The Advantage of Gradient-based Optimizer Incorporating Permutation Randomization) According to Definitions 1, 2, 3, 4, and 6, given a Lipschitz continuous but nonsmooth nonconvex objective function f (please refer to Definition 1), an optimizer without randomization $\mathscr{G}$ (please refer to Definition 2), a randomization operator $\mathscr{R}$ (please refer to Definition 3), an initialized set $I_0 = \{x_{1,0}, x_{2,0}, \cdots, x_{D,0}\}$ (please refer to Definition 4), the close cubes of the global optimum are denoted as $B(x_{1,gbest}, \delta_1)$, $B(x_{2,gbest}, \delta_2)$, $\cdots$, $B(x_{D,gbest}, \delta_D)$ (please refer to Definition 6), suppose the maximum iteration is T, the following (1), (2), and (3) hold:

(1) According to Definition 7, $\mathscr{R}^t \cdot \mathscr{G}^t \cdot f(I_0) = \{\hat{x}_{1,t}, \hat{x}_{2,t}, \cdots, \hat{x}_{D,t}\} \subseteq I$ holds. According to Assumption 1 and D, T are sufficiently large, there are $B(\hat{x}_{k,i}, \delta_{k,i}) \cap B(\hat{x}_{k,j}, \delta_{k,j}) \neq \emptyset, 1 < i, j < T$, inferring from Lemma 2:

$$\forall B(x_{i,gbest}, \delta_i) \subseteq I \subseteq \bigcup_{t=1}^{T} \bigcup_{i=1}^{D} B(\hat{x}_{i,t}, \delta_{i,t}),\ 1 \leq i \leq D,\ 1 \leq t \leq T \tag{2}$$

(2) According to Definition 7, $\mathscr{R}^t \cdot \mathscr{G}^t \cdot f(I_0) = \{\hat{x}_{1,\,t},\, \hat{x}_{2,\,t},\, \cdots,\, \hat{x}_{D,\,t}\} \subseteq I$ holds. According to Assumption 1, and D, T are sufficiently large, there are $B(\hat{x}_{k,\,i},\, \delta_{k,\,i}) \cap B(\hat{x}_{k,\,j},\, \delta_{k,\,j}) = \emptyset,\ 1 < i,\ j < T$, inferring from Lemma 3:

$$\begin{aligned} &\forall B(x_{i,gbest},\, \delta_i) \subseteq I \subseteq \bigcup_{t=1}^{T} \bigcup_{i=1}^{D} B(\hat{x}_{i,\,t},\, \delta_{i,\,t}) \cup E,\ 1 \le i \le D,\ 1 \le t \le T \\ &\overline{\overline{E}} = \aleph_0. \end{aligned} \tag{3}$$

(3) Otherwise, if some closed cubes are overlapped, and others are not, it needs to consider (1) and (2) comprehensively. In brief, the randomized intervals generated by any optimizer can cover the closed cubes of the global optimum.

In Theorem 2, it concludes that an set of closed cubes of the global optimum $B(x_{i,\,gbest},\, \delta_i),\ 1 \le i \le D,\ \delta_i > 0$ (please refer to Definition 6) can be covered in the union of a series closed cubes, such as $B(\hat{x}_{i,\,t},\, \delta_{i,\,t}),\ 1 \le i \le D,\ 1 \le t \le T$, generated via the gradient-based optimizer incorporating randomization, given sufficiently large iterations. For instance, optimize an objective function f with an initialized interval $I_0 = \{10,\ 11,\ 20\}$ and the global best optimum $I_{gbest} = \{19,\ 16,\ 15\}$. Suppose we employ a gradient-based optimizer $\mathscr{G}$, due to contraction property of gradient-based optimizer, we have $\mathscr{G} \cdot f(I_0) = \{12,\ 13,\ 14\}$, $\mathscr{G}^2 \cdot f(I_0) = \{14,\ 15,\ 16\}$, and $\mathscr{G}^3 \cdot f(I_0) = \{15,\ 16,\ 19\}$. If we assume the global optimum is $\{19,\ 16,\ 15\}$, the last randomization can facilitate an accurate approximation to the global optimum, such as $\mathscr{R} \cdot \{15,\ 16,\ 19\} \to \{19,\ 16,\ 15\}$. Meanwhile, we can derive the global optimum $\{19,\ 16,\ 15\} \subseteq \{12,\ 13,\ 14\} \cup \{14,\ 15,\ 16\} \cup \{15,\ 16,\ 19\}$.

Furthermore, as D becomes sufficiently large (e.g., $D \to \infty$), Eqs. (2) and (3) incorporate an increasing number of gradient cubes to effectively cover the entire domain of the objective function.

Moreover, we provide Lemma 4 and Theorem 3 to analyze the potential influence of permutation randomization on the convergence rate. In detail, supposing the permutation randomization as an operator (please refer to Definition 3, we can prove that the norm of $\mathscr{R}$ is equal to 1 and discuss the upper bound of the gradient-based optimizer incorporating randomization.

Lemma 4 (Norm of Permutation Randomization Operator is Equal to 1) Suppose the permutation randomization as an operator $\mathscr{R} : \mathbb{R}^D \to \mathbb{R}^D$, $\|\mathscr{R}\| = 1$ holds, if $D < \infty$.

It is not difficult to prove Lemma 4. According to Definition 3, we can infer:

$$\|\mathscr{R} \cdot \{x_1,\, x_2,\, \cdots,\, x_D\}\| = \|\{\hat{x}_1,\, \hat{x}_2,\, \cdots,\, \hat{x}_D\}\|. \tag{4}$$

According to Definition 5, Lemma 4, and finite-dimensionality space, for any $X,\ Y \in I$, we have:

$$\left\|\mathscr{R}^t \cdot \mathscr{G}^t \cdot (f(X) - f(Y))\right\| \le \left\|\mathscr{R}^t\right\| \cdot \left\|\mathscr{G}^t \cdot (f(X) - f(Y))\right\| = \left\|\mathscr{G}^t \cdot (f(X) - f(Y))\right\|. \tag{5}$$

Let X be $I_1 = \mathscr{G} \cdot f(I_0)$ and Y be I_0, inferring from Eq. (5), we have:

$$\left\|\mathscr{R}^t \cdot \mathscr{G}^t \cdot (f(I_1) - f(I_0))\right\| \le \left\|\mathscr{G}^{t+1} \cdot f(I_0) - \mathscr{G}^t \cdot f(I_0))\right\|. \tag{6}$$

The upper bound on the convergence rate of gradient-based optimizer [40] incorporating randomization can reach $\left\|\mathscr{G}^{t+1} \cdot f(I_0) - \mathscr{G}^t \cdot f(I_0))\right\|$; thus, we revealed the upper bound of the convergence rate of a gradient-based optimizer incorporating permutation randomization should be equivalent to the convergence rate of original gradient-based optimizer

and proved that permutation randomization could maintain the convergence rate of incorporated optimizer and shown in Theorem 3.

Theorem 3 (Permutation Randomization Preserves the Convergence Rate of the Incorporated Gradient-Based Optimizer) According to Definition 7, the convergence rate of a gradient-based optimizer incorporating randomization should be smaller than or equal to the convergence rate of the incorporated gradient-based optimizer $\mathscr{G}$ without randomization.

Suppose that incorporating randomization with a smaller norm (e.g., $\|\mathscr{R}\| < 1$), according to Eq. (6), might decrease the convergence rate of the original optimizer. For instance, Burke's work [15] demonstrates that randomization with a continuously reduced norm should benefit local searches but influence convergence. Furthermore, the following Theorem discusses the upper bound on the convergence rate of a gradient-based optimizer incorporating permutation randomization to approximate a closed cube of the global optimum.

Theorem 4 (Upper Bound on Convergence Rate of Randomized Gradient-based Optimizer Approximates a Closed Cube of Global Optimum) According to Definitions 5, 6, and Theorem 3, considering the objective function $\psi : \mathbb{R}^1 \to \mathbb{R}^1$, if maximum convergence rate is less than $2 \cdot \left|\psi(B(x_{gbest},\ \delta))\right|$, $\{x,\ x_{gbest} \in I \subseteq \mathbb{R}^D,\ \psi(x) \leq \psi(B(x_{gbest},\ \delta))\} = \bigcup_{k=1}^{T} \bigcup_{m=1}^{T} \bigcap_{t=m}^{T} \left\{ x \in I \subseteq \mathbb{R}^D : max[\mathscr{R}^t \cdot \mathscr{G}^t \cdot \psi(x)] + \frac{1}{k} \right\}$ holds.

Our theoretical studies indicate the efficacy of permutation randomization on nonsmooth nonconvex optimization. The following experimental studies can validate theoretical analyses and further demonstrate the advantages of permutation randomization.

5. Experimental results

In this section, we empirically investigate the performance of permutation randomization with a gradient-based optimizer. Specifically, we incorporate the permutation randomization and a representative gradient-based optimizer ADAM [26] as randomized ADAM (see pseudo-code in Algorithm 1, Appendix C, Supplementary Material) to validate reconstruction loss, such as $\|InputMatrix - WeightMatrix \times FeatureMatrix\|_F^2 / \|InputMatrix\|_F^2$, via multiple numerical experiments on unsupervised deep nonlinear learning methods such as Deep Nonlinear Matrix Factorization (DNMF) [41–44] and Deep Belief Network (DBN) [45], and noisy problems. Meanwhile, ADMM [46], ADAM [26], and SVRG [27, 29, 28] are included as peer optimizers to validate permutation randomization. In particular, DNMF is comparable to DNNs, which are widely employed in biomedical signal processing and medical image analysis. For example, DNMF provides an important methodological framework for detecting functional connectivity derived from functional Magnetic Resonance Imaging (fMRI) at multiple spatial scales [47–49], thereby enabling the discovery of complex brain functional organization.

Importantly, permutation randomization is only applied when the difference between the current and previous gradient reaches the given threshold. The threshold measuring the difference between the lastious and current gradient is 1.0×10^{-2}. We also expect to investigate that permutation randomization enables continuous update of the gradient during the late stage [22, 24, 50].

Furthermore, we introduce the objective functions of DNMF, presented below:

$$\min_{Z_i \in \mathbb{R}^{m \times n}} \bigcup_{i=1}^{k} \|Z_i\|_1 \tag{7}$$

$$s.t. \quad (\prod_{i=1}^{k} X_i) \cdot \mathscr{N}_k(Y_k) + Z_k = S. \tag{8}$$

Optimizing the objective function of DNMF is more challenging than most regression problems [43, 44] since more variables (e.g., three variables) are involved in the optimization. In detail, X_i represents the weight/mixing matrix at the i^{th} layer, Y_k denotes the feature matrix at the k^{th} layer, Z_k represents the noise/background components at the k^{th} layer, and S defines the input data matrix. In addition, $\mathcal{N}_k$ represents the activation function in k^{th} layer. Notably, ReLU is determined as an activation function [31]. The reconstruction loss is defined as $\frac{\left\|(\prod_{i=1}^{k} X_i)\cdot Y_k + Z_k - S\right\|_F^2}{\|S\|_F^2}$ [43, 44].

In addition, randomized ADAM and three other peer optimizers are validated on public biomedical data on *OpenfMRI* [51]. We employed a data augmentation technique [52] to expand the original dataset from 29 subjects to a total of 100 subjects. In this experimental study, only the maximum iterations are manually determined as two hundred iterations, and other parameters are tuned as default settings in [46, 53, 54]. Meanwhile, the experimental studies are validated on the CPU cluster (Intel Xeon Gold 6246R) and NVIDIA CUDA cores (NVIDIA DGX A100 P3687).

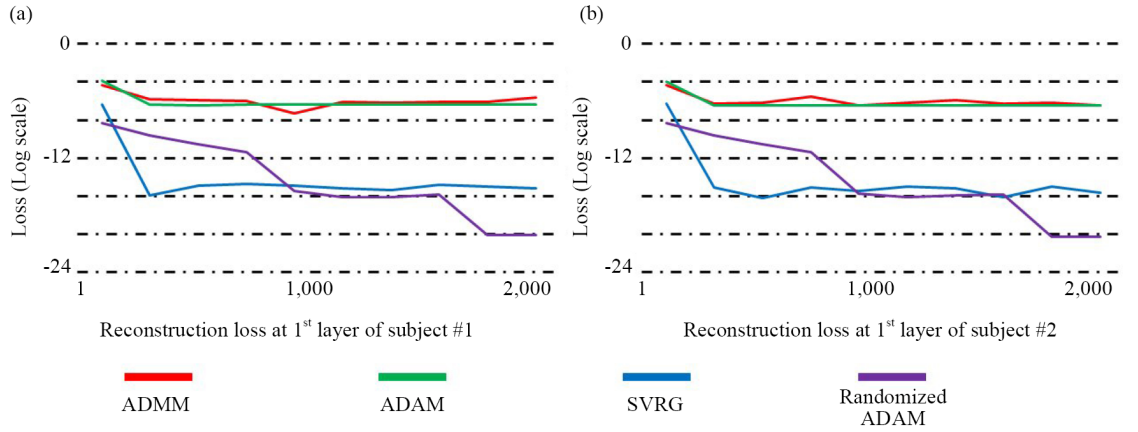


Figure 2. The reconstruction loss comparison of randomized ADAM and the other three peer optimizers within two thousand iterations of randomly selected two subjects at the first layer

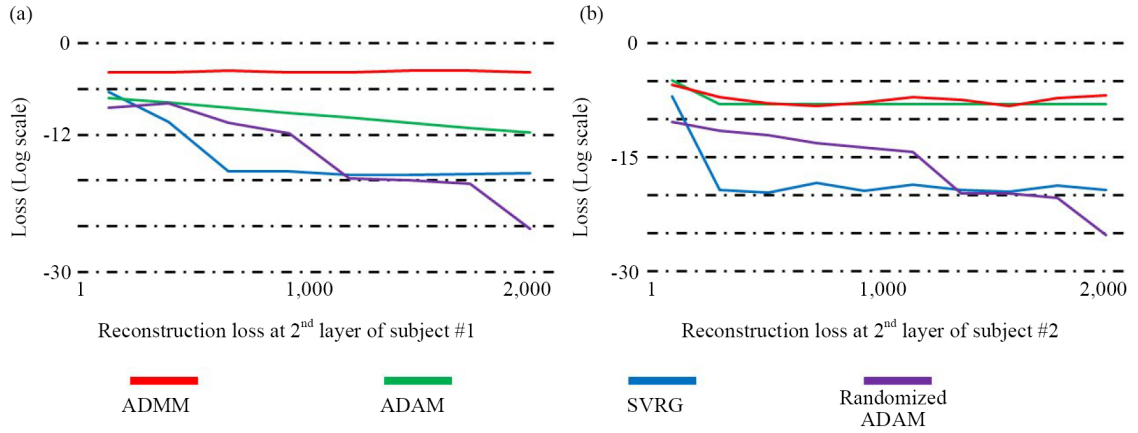


Figure 3. The reconstruction loss comparison of randomized ADAM and the other three peer optimizers within two thousand iterations of randomly selected two subjects at the second layer

Moreover, the Figures (from Figures 2-4 present all experimental results within 2,000 iterations. All figures provide the reconstruction loss of the subjects' signal matrices. Overall, randomized ADAM and three other peer optimizers (ADMM [46], ADAM [26], and SVRG [27]) are applied to optimize the objective function of DNMF (please refer to Eqs. (7) and (8). Notably, with an automatic estimation technique provided in [42, 41], all subjects' signal matrices can

be consistently decomposed into two layers. Notably, in Figures 2 and 3, we provide randomly selected four subjects' reconstruction accuracy curves of the first and second layers.

According to Figures 2-4, ADAM incorporating permutation randomization can achieve the highest reconstruction accuracy. Overall, SVRG and randomized ADAM perform better than ADMM and ADAM. Notably, the gradient of randomized ADAM can continuously update even in the late stage, after 1,600 iterations in Figure 2a and b. Specifically, when the number of iterations is smaller than 1,000, SVRG can achieve a higher accuracy than randomized ADAM. In contrast, if the number of iterations is larger than 1,000, randomized ADAM achieves the highest reconstruction accuracy. In addition, to evaluate the performance of ADAM incorporating permutation randomization on all subjects, we also provide the averaged reconstruction loss in Figure 3 that can further compare the performance of randomized ADAM with other peer algorithms.

Meanwhile, we aggregate all subjects as group-wise data [51] and employ a 3-layer DBN to further validate the performance of randomized ADAM and the other two peer optimizers. We also introduce a noisy optimization problem to investigate the performance of permutation randomization on nonsmooth nonconvex problems [34, 55, 56].

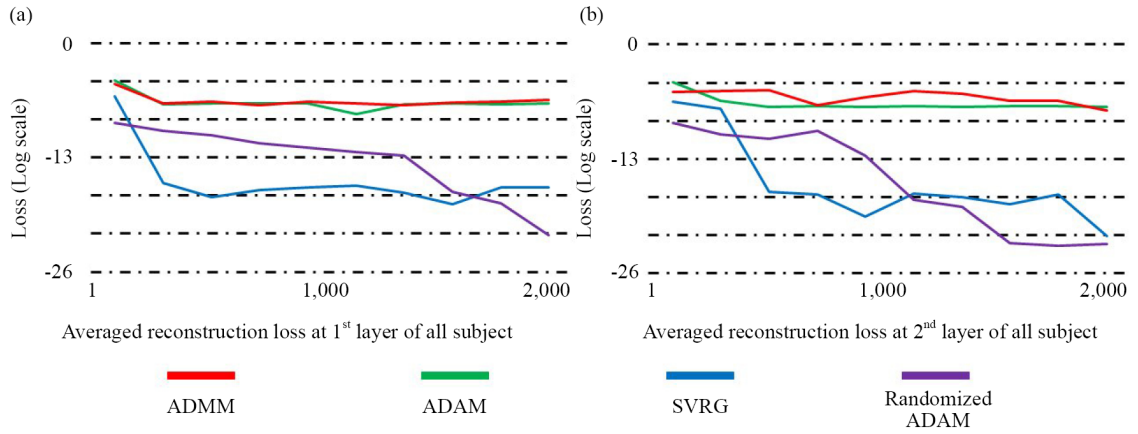


Figure 4. The averaged reconstruction loss comparison of randomized ADAM and the other three peer optimizers within two thousand iterations across all subjects at first and second layers, respectively

Consistent with earlier numerical experiments, the randomized gradient-based optimizer is less prone to becoming trapped in stationary points in both DNNs and noisy DNMF (introducing random noise into the gradient at every 10 iterations), as shown in Figures 6 and 7 in Appendix C of the Supplementary Materials. For additional evaluation, we applied each method using default parameter settings [26, 53, 56, 57] to a logistic regression task [58] using the publicly available breast cancer dataset [59], visualized in Figure 8 in Appendix C of the Supplementary Material.

6. Statistical analyses

In this section, we present a quantitative analysis of the previously reported experimental results. Since all gradient-based optimizers examined in our empirical study are iterative algorithms, the reconstruction loss between adjacent iterations is not statistically independent (see Figures 2-4). This lack of independence limits the applicability of traditional statistical tests such as t-tests and confidence intervals for comparing iterative reconstruction performance [60]. To address this limitation, we adopt Intra-class Correlation Coefficients (ICCs)-a widely used descriptive statistical technique for assessing consistency in grouped quantitative measurements [61, 62]. Figure 5c and d report the ICCs for randomized ADAM and three peer optimizers. Notably, randomized ADAM and SVRG exhibit the highest levels of consistency. For example, at the second layer of DNMF, ICCs for randomized ADAM exceed 0.80 at the first layer and surpass 0.90 at the second layer, indicating strong internal reliability. These results highlight the performance of permutation randomization

as comparable to SVRG, which uses averaged gradients for updates. In addition to consistency, we evaluated time consumption across all methods. As shown in Figures 5a and b, the runtime of permutation randomization is higher than ADAM and ADMM, but substantially lower than SVRG, striking a promising balance between performance and efficiency. The mean and standard deviation of these results are summarized in Table 2.

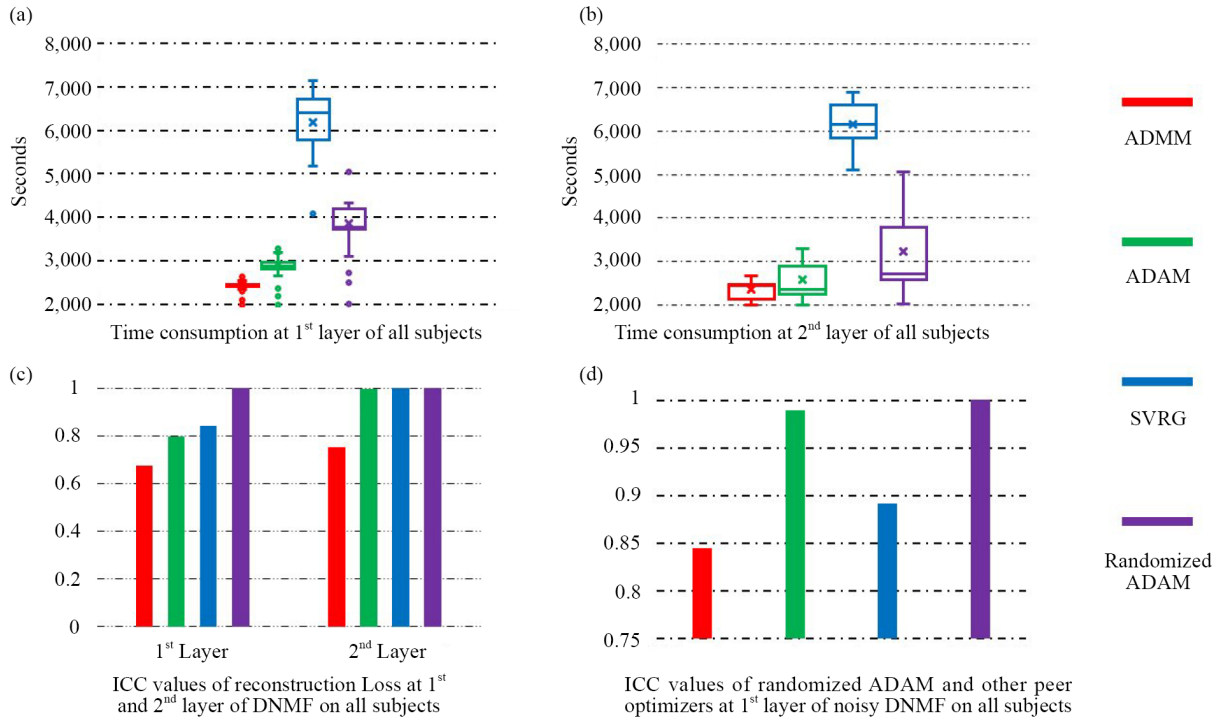


Figure 5. An illustration of time-consumption and consistency comparisons of randomized ADAM and other peer optimizers. The box plots in (a) and (b) represent the time consumption of four algorithms using all subjects; (c) provides the ICC values to demonstrate an overall consistency

Table 2. Time consumption comparison of all optimizers in seconds

Time consumption at 1st layer of DNMF	Time consumption at 2nd layer of DNMF
ADMM 451.58 ± 93.86	ADMM 287.42 ± 78.15
ADAM 901.65 ± 196.96	ADAM 579.98 ± 53.41
SVRG 4,538.50 ± 305.49	SVRG 4,441.60 ± 547.12
Randomized ADAM 1,562.50 ± 185.31	Randomized ADAM 1,228.70 ± 195.09

7. Conclusion

In this work, we conducted both theoretical and empirical analyses of permutation randomization in nonsmooth nonconvex optimization. Our results showcased that integrating permutation randomization into gradient-based optimizers broadens the search dynamics, enabling the algorithm to explore diverse closed cubes that contain the global optimum rather than shrinking prematurely toward suboptimal regions. Importantly, we prove that this strategy preserves the original optimizer's convergence rate, ensuring that no efficiency is lost.

On the experimental side, we validated these insights through extensive numerical studies, including the training of deep neural networks with stacked architectures and the optimization of noisy objective functions. The consistent performance gains highlight the practical value of permutation randomization in real-world scenarios.

In summary, these findings deliver both a rigorous theoretical foundation and compelling empirical evidence for the utility of randomization in gradient-based methods. We envision this work as a starting point for broader investigations into alternative randomization techniques, their role in overcoming fundamental challenges of nonsmooth and nonconvex optimization, and their potential to advance applications in machine learning and beyond.

Acknowledgement

We gratefully acknowledge the AU-GSU Seed Award Program (AUGSU00016) for funding the experiments of this study.

Conflict of interest

The authors declare no competing financial interest.

References

[1] Hazan E, Agarwal A, Kale S. Logarithmic regret algorithms for online convex optimization. *Machine Learning*. 2007; 69(2): 169-192. Available from: https://doi.org/10.1007/s10994-007-5016-8.

[2] Nemirovski A, Juditsky A, Lan G, Shapiro A. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*. 2009; 19(4): 1574-1609. Available from: https://doi.org/10.1137/070704277.

[3] Jordan MI, Kornowski G, Lin T, Shamir O, Zampetakis M. Deterministic nonsmooth nonconvex optimization. In: *Proceedings of The Thirty Sixth Annual Conference on Learning Theory*. Bangalore, India: PMLR; 2023. p.4570-4597.

[4] Zhang Y, Wainwright MJ, Duchi JC. Communication-efficient algorithms for statistical optimization. *Journal of Machine Learning Research*. 2013; 14(68): 3321-3363.

[5] Shalev-Shwartz S, Zhang T. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*. 2013; 14(16): 567-599.

[6] Johnson R, Zhang T. Accelerating stochastic gradient descent using predictive variance reduction. In: *Proceedings of the 27th International Conference on Neural Information Processing Systems*. USA: Curran Associates, Inc.; 2013. p.315-323.

[7] Defazio A, Bach F, Lacoste-Julien S. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In: *Proceedings of the 28th International Conference on Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press; 2014. p.1646-1654.

[8] Arjevani Y, Shamir O. Communication complexity of distributed convex learning and optimization. In: *Proceedings of the 29th International Conference on Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press; 2015. p.1756-1764.

[9] Lin H, Mairal J, Harchaoui Z. A universal catalyst for first-order optimization. In: *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press; 2015. p.3384-3392.

[10] Allen-Zhu Z. The first direct acceleration of stochastic gradient methods. *Journal of Machine Learning Research*. 2017; 18(1): 8194-8244.

[11] El Sayed MA, Farahat FA, Elsisy MA. A novel interactive approach for solving uncertain bi-level multi-objective supply chain model. *Computers & Industrial Engineering*. 2022; 169: 108225. Available from: https://doi.org/10.1016/j.cie.2022.108225.

[12] Burke JV, Lewis AS, Overton ML. Approximating subdifferentials by random sampling of gradients. *Mathematics of Operations Research*. 2002; 27(3): 567-584. Available from: https://doi.org/10.1287/moor.27.3.567.317.

[13] Burke JV, Lewis AS, Overton ML. A robust gradient sampling algorithm for nonsmooth, nonconvex optimization. *SIAM Journal on Optimization*. 2005; 15(3): 751-779. Available from: https://doi.org/10.1137/030601296.

[14] Kiwiel KC. Convergence of the gradient sampling algorithm for nonsmooth nonconvex optimization. *SIAM Journal on Optimization*. 2007; 18(2): 379-388. Available from: https://doi.org/10.1137/050639124.

[15] Burke JV, Curtis FE, Lewis AS, Overton ML, Simões LM. Gradient sampling methods for nonsmooth optimization. In: Bagirov A, Gaudioso M, Karmitsa N, Mäkelä M, Taheri S. (eds.) *Numerical Nonsmooth Optimization: State of the Art Algorithms*. Cham: Springer; 2020. p.201-225.

[16] Kiwiel KC. Restricted step and Levenberg-Marquardt techniques in proximal bundle methods for nonconvex nondifferentiable optimization. *SIAM Journal on Optimization*. 1996; 6(1): 227-249. Available from: https://doi.org/10.1137/0806013.

[17] Atenas F, Sagastizábal C, Silva PJ, Solodov M. A unified analysis of descent sequences in weakly convex optimization, including convergence rates for bundle methods. *SIAM Journal on Optimization*. 2023; 33(1): 89-115. Available from: https://doi.org/10.1137/21M1465445.

[18] Gaudioso M, Taheri S, Bagirov AM, Karmitsa N. Bundle enrichment method for nonsmooth difference of convex programming problems. *Algorithms*. 2023; 16(8): 394. Available from: https://doi.org/10.3390/a16080394.

[19] Benaïm M, Hofbauer J, Sorin S. Stochastic approximations and differential inclusions. *SIAM Journal on Control and Optimization*. 2005; 44(1): 328-348. Available from: https://doi.org/10.1137/S0363012904439301.

[20] Daniilidis A, Drusvyatskiy D. Pathological subgradient dynamics. *SIAM Journal on Optimization*. 2020; 30(2): 1327-1338. Available from: https://doi.org/10.1137/19M1288430.

[21] Bolte J, Pauwels E. Conservative set valued fields, automatic differentiation, stochastic gradient methods and deep learning. *Mathematical Programming*. 2020; 188(1): 19-51. Available from: https://doi.org/10.1007/s10107-020-01501-5.

[22] Clarke FH. Necessary conditions for nonsmooth variational problems. In: *Optimal Control Theory and Its Applications*. Berlin, Heidelberg: Springer; 1974. p.70-91.

[23] Clarke FH. Generalized gradients and applications. *Transactions of the American Mathematical Society*. 1975; 205: 247-262.

[24] Clarke FH. Generalized gradients of Lipschitz functionals. *Advances in Mathematics*. 1981; 40(1): 52-67. Available from: https://doi.org/10.1016/0001-8708(81)90022-5.

[25] Clarke FH. *Optimization and Nonsmooth Analysis*. Philadelphia, PA: Society for Industrial and Applied Mathematics; 1990.

[26] Kingma DP, Ba J. Adam: A method for stochastic optimization. *arXiv:14126980*. 2014. Available from: https://arxiv.org/abs/1412.6980.

[27] Lian X, Wang M, Liu J. Finite-sum composition optimization via variance reduced gradient descent. In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. Fort Lauderdale, FL, USA: PMLR; 2017. p.1159-1167.

[28] Nan T, Gao Y, Kroer C. Convergence of extragradient SVRG for variational inequalities: Error bounds and increasing iterate averaging. *arXiv:230601796*. 2023. Available from: https://arxiv.org/abs/2306.01796.

[29] Dubois-Taine B, Vaswani S, Babanezhad R, Schmidt M, Lacoste-Julien S. SVRG meets AdaGrad: Painless variance reduction. *Machine Learning*. 2021; 111(12): 4359-4409. Available from: https://doi.org/10.1007/s10994-022-06265-x.

[30] LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015; 521(7553): 436-444. Available from: https://doi.org/10.1038/nature14539.

[31] Nair V, Hinton GE. Rectified linear units improve restricted Boltzmann machines. In: *Proceedings of the 27th International Conference on Machine Learning*. Haifa, Israel: Omnipress; 2010. p.807-814.

[32] Nesterov Y. Smooth minimization of non-smooth functions. *Mathematical Programming*. 2005; 103(1): 127-152. Available from: https://doi.org/10.1007/s10107-004-0552-5.

[33] Nesterov Y. *Lectures on Convex Optimization*. 2nd ed. Cham: Springer; 2018.

[34] Arjevani Y, Carmon Y, Duchi JC, Foster DJ, Srebro N, Woodworth B. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*. 2022; 199(1): 165-214. Available from: https://doi.org/10.1007/s10107-022-01822-7.

[35] Rudin W. *Functional Analysis*. 2nd ed. New York: McGraw-Hill; 1973.

[36] Yosida K. *Functional Analysis*. 6th ed. Berlin, Heidelberg: Springer Science & Business Media; 2012.

[37] Von Oswald J, Niklasson E, Randazzo E, Sacramento J, Mordvintsev A, Zhmoginov A, et al. Transformers learn in-context by gradient descent. In: *International Conference on Machine Learning*. Honolulu, Hawaii, USA: PMLR; 2023. p.35151-35174.

[38] Lin J, Song C, He K, Wang L, Hopcroft JE. Nesterov accelerated gradient and scale invariance for adversarial attacks. *arXiv:190806281*. 2019. Available from: https://arxiv.org/abs/1908.06281.

[39] Royden HL. *Real Analysis*. 2nd ed. New York: Macmillan; 1968.

[40] Wang L, Shen B. On the parallelization upper bound for asynchronous stochastic gradient descent in non-convex optimization. *Journal of Optimization Theory and Applications*. 2023; 196(3): 900-935. Available from: https://doi.org/10.1007/s10957-022-02141-9.

[41] Wen Z, Yin W, Zhang Y. Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm. *Mathematical Programming Computation*. 2012; 4(4): 333-361. Available from: https://doi.org/10.1007/s12532-012-0044-1.

[42] Shen Y, Wen Z, Zhang Y. Augmented Lagrangian alternating direction method for matrix separation based on low-rank factorization. *Optimization Methods and Software*. 2014; 29(2): 239-263. Available from: https://doi.org/10.1080/10556788.2012.700713.

[43] Trigeorgis G, Bousmalis K, Zafeiriou S, Schuller BW. A deep semi-NMF model for learning hidden representations. In: *Proceedings of the 31st International Conference on Machine Learning*. Beijing, China: PMLR; 2014. p.1692-1700.

[44] Trigeorgis G, Bousmalis K, Zafeiriou S, Schuller BW. A deep matrix factorization method for learning attribute representations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2016; 39(3): 417-429. Available from: https://doi.org/10.1109/TPAMI.2016.2554555.

[45] Hinton GE. Deep belief networks. *Scholarpedia*. 2009; 4(5): 5947. Available from: https://doi.org/10.4249/scholarpedia.5947.

[46] Nishihara R, Lessard L, Recht B, Packard A, Jordan MI. A general analysis of the convergence of ADMM. *arXiv:150202009*. 2015. Available from: https://arxiv.org/abs/1502.02009.

[47] Xie H, Han X, Xiao G, Xu H, Zhang Y, Zhang G, et al. Multifocal fluorescence video-rate imaging of centimetre-wide arbitrarily shaped brain surfaces at micrometric resolution. *Nature Biomedical Engineering*. 2024; 8(6): 740-753. Available from: https://doi.org/10.1038/s41551-023-01155-6.

[48] Zhang W, Cohen AD, McCrea M, Mukherjee P, Wang YP. Deep linear matrix approximate reconstruction with integrated BOLD signal denoising reveals reproducible hierarchical brain connectivity networks from multiband multi-echo fMRI. *Frontiers in Neuroscience*. 2025; 19: 1577029. Available from: https://doi.org/10.3389/fnins.2025.1577029.

[49] Zidan AH, Jahin A, Bao Y, Zhang W. Semi-nonlinear deep efficient reconstruction for unveiling linear and nonlinear spatial features of the human brain. *Electronics*. 2025; 14(3): 554. Available from: https://doi.org/10.3390/electronics14030554.

[50] Goldstein AA. Optimization of Lipschitz continuous functions. *Mathematical Programming*. 1977; 13(1): 14-22. Available from: https://doi.org/10.1007/BF01584320.

[51] Cohen AD, Wang YP. *Multiband multi-echo BOLD fMRI*. 2018. Available from: https://openfmri.org/dataset/ds000258/ [Accessed 17th July 2018].

[52] Pei S, Wang C, Cao S, Lv Z. Data augmentation for fMRI-based functional connectivity and its application to cross-site ADHD classification. *IEEE Transactions on Instrumentation and Measurement*. 2022; 72: 1-15. Available from: https://doi.org/10.1109/TIM.2022.3232670.

[53] Lei L, Jordan MI. Less than a single pass: Stochastically controlled stochastic gradient. In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. Fort Lauderdale, Florida, USA: Proceedings of Machine Learning Research; 2017. p.148-156.

[54] Cutkosky A, Orabona F. Momentum-based variance reduction in nonconvex SGD. In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Vancouver, Canada: Curran Associates, Inc.; 2019. p.15236-15245.

[55] Forneron JJ. Noisy, non-smooth, non-convex estimation of moment condition models. *arXiv:230107196*. 2023. Available from: https://doi.org/10.48550/arXiv.2301.07196.

[56] Lei Y. Stability and generalization of stochastic optimization with nonconvex and nonsmooth problems. In: *Proceedings of The Thirty Sixth Annual Conference on Learning Theory*. Bangalore, India: PMLR; 2023. p.191-227.

[57] Allen-Zhu Z, Yuan Y. Improved SVRG for non-strongly-convex or sum-of-non-convex objectives. In: *Proceedings of The 33rd International Conference on Machine Learning*. New York: PMLR; 2016. p.1080-1089.

[58] Schober P, Vetter TR. Logistic regression in medical research. *Anesthesia & Analgesia*. 2021; 132(2): 365-366. Available from: https://doi.org/10.1213/ANE.0000000000005247.

[59] Shut M. *Breast cancer data*. 2023. Available from: https://www.kaggle.com/datasets/marshuu/breast-cancer [Accessed 1st January 2023].

[60] Field A. *Discovering Statistics Using IBM SPSS Statistics*. 4th ed. London: SAGE Publications; 2013.

[61] Bartko JJ. On various intraclass correlation reliability coefficients. *Psychological Bulletin*. 1976; 83(5): 762-765. Available from: https://doi.org/10.1037/0033-2909.83.5.762.

[62] Bujang MA, Baharum N. A simplified guide to determination of sample size requirements for estimating the value of intraclass correlation coefficient: A review. *Archives of Orofacial Science*. 2017; 12(1): 1-11.

Appendix A

Lemma 1 (Contraction Property of Gradient-based Optimizer without Randomization) According to Definitions 2, 3, 4, 5, and Assumption 1, a gradient-based optimizer without randomization $\mathscr{G} : \mathbb{R} \to \mathbb{R}^D$ within t iterations denoted as $\mathscr{G}^t \cdot f(I_0), \forall t \in [1,\ T], \mathscr{G} \cdot f(I_t) = I_{t+1}, I_{t+1} \subseteq I_t$.

Proof. Proof by contradiction. Assume $\mathscr{G} \cdot f(I_t) = I_{t+1} \supseteq I_t$ and initialized closed interval as I_0, we can infer:

$$I_T \supseteq \cdots \supseteq I_t \supseteq \cdots \supseteq I_0. \tag{9}$$

Let $T \to \infty$, we have $I_T \to (-\infty,\ +\infty)$ which implies:

$$||\mathscr{G}^t \cdot f(I_0)|| \to \infty. \tag{10}$$

Based on the concept of a contraction operator, we infer that $\mathscr{G}$ is not a contraction and therefore cannot guarantee convergence. This supports our theoretical result that a gradient-based optimizer must exhibit contraction behavior, by inducing the shrinkage of input closed sets or intervals, to ensure convergence properties.

Theorem 1 (Limitation of Gradient-based Optimizer without Randomization) According to Definitions 2, 3, 4, 5, and 6, suppose current iteration as t, maximum iteration as T, and initialized interval as I_0 if a gradient-based optimizer without randomization $\mathscr{G}$ can approximate the closed cubes of global optimum $B(x_{i,\ gbest}, \delta_i),\ \delta_i > 0,\ 1 \leq i \leq D$, for any $t \in [1,\ T]$, $B(x_{i,\ gbest},\ \delta_i) \subseteq I_t = \mathscr{G}^t \cdot f(I_0)$ must hold.

Proof. Inferring from Lemma 1, we have:

$$\mathscr{G}^T \cdot f(I_0) \subseteq \cdots \subseteq \mathscr{G}^t \cdot f(I_0) \subseteq \cdots \subseteq \mathscr{G}^1 \cdot f(I_0). \tag{11}$$

Let $t \to \infty$, consider $\mathscr{G}$ can approximate a closed cube of global optimum $B(x_{gbest},\ \delta),\ \delta > 0$. And inferring from Cantor Intersection Theorem, we have:

$$\forall i \in [1,\ D],\ B(x_{i,\ gbest},\ \delta_i) \subseteq \bigcap_{t=1}^{T} \mathscr{G}^t \cdot f(I_0). \tag{12}$$

Therefore, we conclude that for a conventional gradient-based optimizer to successfully approach the global optimum, the intervals generated in each iteration, denoted as $I_t,\ 1 \leq t \leq T$, must include the global optimum.

Appendix B

Theorem 2 (The Advantage of Gradient-based Optimizer Incorporating Permutation Randomization) According to Definitions 1, 2, 3, 4, and 6, given a Lipschitz continuous but nonsmooth nonconvex objective function f (please refer to Definition 1), an optimizer without randomization $\mathscr{G}$ (please refer to Definition 2), a randomization operator $\mathscr{R}$ (please refer to Definition 3), an initialized internal $I_0 = \{x_{1,\,0},\, x_{2,\,0},\, \cdots,\, x_{D,\,0}\}$ (please refer to Definition 4), the close cubes of the global optimum are denoted as $B(x_{1,\,gbest},\, \delta_1),\, B(x_{2,\,gbest},\, \delta_2),\, \cdots,\, B(x_{D,\,gbest},\, \delta_D)$ (please refer to Definition 6, suppose the maximum iteration is T, the following (1), (2), and (3) hold:

(1) According to Definition 7, $\mathscr{R}^t \cdot \mathscr{G}^t \cdot f(I_0) = \{\hat{x}_{1,\,t},\, \hat{x}_{2,\,t},\, \cdots,\, \hat{x}_{D,\,t}\} \subseteq I$ holds. According to Assumption 1 and D, T are sufficiently large, there are $B(\hat{x}_{i,\,t},\, \delta_{i,\,t}) \cap B(\hat{x}_{i,\,j},\, \delta_{i,\,j}) \neq \emptyset,\, 1 \leq t,\, j \leq T,\, 1 \leq i \leq D$, inferring from Lemma 2:

$$\forall B(x_{i,\,gbest},\, \delta_i) \subseteq I \subseteq \bigcup_{t=1}^{T} \bigcup_{i=1}^{D} B(\hat{x}_{i,\,t}, \delta_{i,\,t}),\, 1 \leq t \leq T,\, 1 \leq i \leq D. \tag{13}$$

(2) According to Definition 7, $\mathscr{R}^t \cdot \mathscr{G}^t \cdot f(I_0) = \{\hat{x}_{1,\,t},\, \hat{x}_{2,\,t},\, \cdots,\, \hat{x}_{D,\,t}\} \subseteq I$ holds. According to Assumption 1, and D, T are sufficiently large, there are $B(\hat{x}_{i,\,t},\, \delta_{i,\,t}) \cap B(\hat{x}_{i,\,j},\, \delta_{i,\,j}) = \emptyset,\, 1 \leq t \leq T,\, 1 \leq i \leq D$, inferring from Lemma 3:

$$\begin{gathered} \forall B(x_{i,\,gbest},\, \delta_i) \subseteq I \subseteq \bigcup_{t=1}^{T} \bigcup_{i=1}^{D} B(\hat{x}_{i,\,t},\, \delta_{i,\,t}) \cup E,\, 1 \leq t \leq T,\, 1 \leq i \leq D \\ \overline{\overline{E}} = \aleph_0. \end{gathered} \tag{14}$$

(3) Otherwise, if some closed cubes are overlapped, and others are not, it needs to consider (1) and (2) comprehensively. In brief, the randomized intervals generated by any optimizer can cover the closed cubes of the global optimum.

Proof. According to Definition 2, considering a gradient-based optimizer:

$$\mathscr{G} \cdot I_0 \subseteq I. \tag{15}$$

According to Definition 3, considering a gradient-based optimizer with permutation randomization, we have:

$$\mathscr{R} \cdot \mathscr{G} \cdot I_0 \subseteq I. \tag{16}$$

According to Definition 5, We can derive the following equation:

$$\mathscr{G}^t \cdot I_0 = \{x_{1,\,t},\, x_{2,\,t},\, \cdots,\, x_{D,\,t}\} \subseteq I. \tag{17}$$

In addition, for any $t \in \mathbb{N}$, we have:

$$\mathscr{R}^t \cdot \mathscr{G}^t \cdot I_0 = \{\hat{x}_{1,\,t},\, \hat{x}_{2,\,t},\, \cdots,\, x_{D,\,t}\} \subseteq I. \tag{18}$$

On the one hand, if we consider the overlap through all generated closed cubes via $\mathscr{R}^t \cdot \mathscr{G}^t \cdot I_0$:

$$\begin{aligned} &\forall j,\ k \in [1,\ T],\ j \neq k,\ B(\hat{x}_{i,\ j},\ \delta_{i,\ j}) \subseteq I,\ B(\hat{x}_{i,\ k},\ \delta_{i,\ k} \subseteq I,\ 1 \leq j,\ k \leq T,\ 1 \leq i \leq D \\ &B(\hat{x}_{i,\ j},\ \delta_{i,\ j}) \cap B(\hat{x}_{i,\ k},\ \delta_{i,\ k}) \neq \varnothing. \end{aligned} \tag{19}$$

Suppose T and D are sufficiently large, inferring from Lemma 2, according to Definition 6, if we denote the closed cubes of global optimum as $B(x_{1,\ gbest},\ \delta_1),\ B(x_{2,\ gbest},\ \delta_2),\ \cdots,\ B(x_{D,\ gbest})$, we have:

$$I \subseteq \bigcup_{t=1}^{T} \bigcup_{i=1}^{D} B(\hat{x}_{i,\ t},\ \delta_{i,\ t}). \tag{20}$$

We can also derive:

$$\forall i \in [1,\ D],\ \bigcup_{t=1}^{T} \bigcup_{i=1}^{D} B(\hat{x}_{i,\ t},\ \delta_{i,\ t}) \supseteq I \supset B(x_{i,\ gbest},\ \delta_i). \tag{21}$$

On the other hand, if we consider no overlap through all generated intervals via $\mathscr{R}^t \cdot \mathscr{G}^t \cdot I$:

$$\begin{aligned} &\forall j,\ k \in [1,\ T],\ j \neq k,\ B(\hat{x}_{i,\ j},\ \delta_{i,\ j}) \subseteq I,\ B(\hat{x}_{i,\ k},\ \delta_{i,\ k}) \subseteq I \\ &B(\hat{x}_{i,\ j},\ \delta_{i,\ j}) \cap B(\hat{x}_{i,\ k},\ \delta_{i,\ k}) \neq \varnothing. \end{aligned} \tag{22}$$

According to Lemma 3, suppose T is sufficiently large, we have:

$$\begin{aligned} &m^* \left(I \backslash \bigcup_{t=1}^{T} \bigcup_{i=1}^{D} B(\hat{x}_{i,\ t},\ \delta_{i,\ t}) \right) < \varepsilon \\ &E = I \backslash \bigcup_{t=1}^{T} \bigcup_{i=1}^{D} B(\hat{x}_{i,\ t},\ \delta_{i,\ t}) \\ &\left(\bigcup_{t=1}^{T} \bigcup_{i=1}^{D} B(\hat{x}_{i,\ t},\ \delta_{i,\ t}) \right) \bigcup E \supseteq I. \end{aligned} \tag{23}$$

It also indicates:

$$\begin{aligned} &^{\forall} i \in [1,\ D],\ \left(\bigcup_{t=1}^{T}\bigcup_{i=1}^{D} B(\hat{x}_{i,\ t},\ \delta_{i,\ t})\right)\bigcup E \supseteq I \supseteq B(x_{i,\ gbest},\ \delta_i) \\ &\overline{\overline{E}} = \aleph_0. \end{aligned} \tag{24}$$

Otherwise, it is obvious that $\{B(x_{1,\ gbest},\ \delta_1),\ B(x_{2,\ gbest},\ \delta_2),\ \cdots,\ B(x_{D,\ gbest},\ \delta_D)\}$ can be included in a series of overlapped and non-overlapped closed intervals by comprehensively using Eqs. (21) and (24).

Lemma 4 (Norm of Permutation Randomization Operator is Equal to 1) Suppose the permutation randomization as an operator $\mathscr{R} : \mathbb{R}^D \to \mathbb{R}^D$, $\|\mathscr{R}\| = 1$ holds, if $D < \infty$.

Proof. Considering $\mathscr{R}$ applying on finite-dimensional space:

$$\mathscr{R} \cdot \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{bmatrix} = \begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \\ \vdots \\ \hat{x}_D \end{bmatrix}. \tag{25}$$

Inferring from Eq. (25), we have:

$$\hat{x}_1 = x_i,\ \hat{x}_2 = x_j,\ \cdots,\ \hat{x}_D = x_k,\ i,\ j,\ k \in [1,\ D]. \tag{26}$$

Eq. (14), we have:

$$||\{x_1,\ x_2,\ \cdots,\ x_D\}|| = ||\{\hat{x}_1,\ \hat{x}_2,\ \cdots,\ \hat{x}_D\}||. \tag{27}$$

According to the concept of operator norm [35], we can derive the following:

$$||\mathscr{R}|| = sup\frac{\mathscr{R} \cdot ||\{x_1,\ x_2,\ \cdots,\ x_D\}||}{||\{x_1,\ x_2,\ \cdots,\ x_D\}||} = sup\frac{||\{\hat{x}_1,\ \hat{x}_2,\ \cdots,\ \hat{x}_D\}||}{||\{x_1,\ x_2,\ \cdots,\ x_D\}||} = 1. \tag{28}$$

Theorem 3 (Permutation Randomization Preserves the Convergence Rate of the Incorporated Gradient-Based Optimizer) According to Assumption 1 and Definition 7, the convergence rate of a gradient-based optimizer incorporating randomization should be smaller than or equal to the convergence rate of the originally incorporated gradient-based optimizer $\mathscr{G}$ without randomization.

Proof. Inferring from the concept of contraction operator, we have:

$$\begin{aligned} &||\mathscr{G} \cdot (f(X) - f(Y))|| \leq c||\mathscr{G} \cdot (f(X) - f(Y))|| \\ &0 < c < 1. \end{aligned} \tag{29}$$

We can rewrite the left side of Eq. (28) as:

$$||\mathscr{G} \cdot (f(I_{t+1}) - f(I_t))||. \tag{30}$$

Then, we have:

$$||\mathscr{G} \cdot (f(I_{t+1}) - f(I_t))|| \leq c \cdot ||(f(I_{t+1}) - f(I_t))||. \tag{31}$$

Considering the incorporation of optimizer and randomization as $\mathscr{R} \cdot \mathscr{G} \cdot f(x)$, we have

$$||\mathscr{R} \cdot \mathscr{G} \cdot (f(I_{t+1}) - f(I_t))|| \leq ||\mathscr{R}|| \cdot ||\mathscr{G} \cdot (f(I_{t+1}) - f(I_t))||. \tag{32}$$

Inferring from Lemma 4, it is obvious that we have:

$$||\mathscr{R}|| \cdot ||\mathscr{G} \cdot (f(I_{t+1}) - f(I_t))|| = ||\mathscr{G} \cdot (f(I_{t+1}) - f(I_t))|| \leq c \cdot ||f(I_{t+1}) - f(I_t)||. \tag{33}$$

Eq. (33) implies permutation randomization $\mathscr{R}$ can maintain the convergence rate of the original gradient-based optimizer $\mathscr{G}$.

Theorem 4 (Upper Bound on Convergence Rate of Randomized Gradient-based Optimizer Approximates an Closed Cube of Global Optimum) According to Definitions 5, 6, and Theorem 3, considering the objective function $\psi : \mathbb{R}^1 \to \mathbb{R}^1$, if maximum convergence rate is less than $2 \cdot \left| f(B(x_{gbest},\ \delta)) \right|$, $\{x,\ x_{gbest} \in I \subseteq \mathbb{R}^D,\ f(x) \leq \psi(B(x_{gbest},\ \delta))\} = \bigcup_{k=1}^{T} \bigcup_{m=1}^{T} \bigcap_{t=m}^{T} \left\{ x \in I \subseteq \mathbb{R}^D : max[\mathscr{R}^t \cdot \mathscr{G}^t \cdot \psi(x)] + \frac{1}{k} \right\}$ holds.

Proof. Let $E_{t,\ k} = \left[x \in I : max_{x \in I} \mathscr{R}^t \cdot \mathscr{G}^t \cdot \psi(x) \leq \psi(B(x_{gbest},\ \delta)) + \frac{1}{k} \right]$.

On the one hand, we need to prove $\{x,\ x_{gbest} \in \mathbb{R}^1 : \psi(x) \leq \psi(B(x_{gbest},\ \delta))\} \subseteq \bigcup_{k=1}^{T} \bigcup_{m=1}^{T} \bigcap_{t=m}^{T} \left\{ x \in \mathbb{R}^1 : max_{x \in I} \mathscr{R}^t \cdot \mathscr{G}^t \cdot f(x) \leq f(B(x_{gbest},\ \delta)) + \frac{1}{k} \right\}$.

Suppose $x_{gbest} \in \mathbb{R}^1$ and $x_0 \in \{x,\ x_{gbest} \in \mathbb{R}^1 : f(x) \leq f(B(x_{gbest},\ \delta))\}$, according to *Cauchy Theorem*, we have $lim_{t \to \infty} \psi_t(x_0) = \psi(x_0) \leq \psi(B(x_{gbest},\ \delta))$ and derive:

$$\forall k_0 \in \mathbb{N},\ \exists t_0 \geq t_0,\ \psi_t(x_0) \leq \psi(B(x_{gbest},\ \delta)) + \frac{1}{k_0}. \tag{34}$$

Eq. (22) indicates that $x_0 \in E_{t,\ k_0} (t \geq t_0)$. It also demonstrates x_0 belongs to the lower limit of $\{E_{t,\ k_0}\}$.
According to the Definition of the lower limit, we have:

$$\forall x_0 \in I,\ x_0 \in \bigcup_{m=1}^{T} \bigcap_{t=m}^{T} E_{t,\ k_0}. \tag{35}$$

Thus, we proved $\{x,\ x_{gbest} \in I \subseteq \mathbb{R}^D : f(x) \leq f(B(x_{gbest},\ \delta))\} \subseteq \bigcup_{k=1}^{T} \bigcup_{m=1}^{T} \bigcap_{t=m}^{T} \left\{ x \in I \subseteq \mathbb{R}^D : max_{x \in I} \mathscr{R}^t \cdot \mathscr{G}^t \cdot f(x) \leq f(B(x_{gbest},\ \delta)) + \frac{1}{k} \right\}$.

On the other hand, we need to prove $\{x,\ x_{gbest} \in I \subseteq \mathbb{R}^D : f(x) \leq f(B(x_{gbest},\ \delta))\} \supseteq \bigcup_{k=1}^{T} \bigcup_{m=1}^{T} \bigcap_{t=m}^{T} \left\{ x \in I \subseteq \mathbb{R}^D : max_{x \in I} \mathscr{R}^t \cdot \mathscr{G}^t \cdot f(x) \leq f(B(x_{gbest},\ \delta)) + \frac{1}{k} \right\}$.

Suppose $x_0 \in \bigcup_{k=1}^{T} \bigcup_{m=1}^{T} \bigcap_{t=m}^{T} \left\{ x \in I \subseteq \mathbb{R}^D : max_{x \in \mathbb{R}^1} \mathscr{R}^t \cdot \mathscr{G}^t \cdot f(x) \leq f(B(x_{gbest},\ \delta)) + \frac{1}{k} \right\}$, obviously, we have $x_0 \in \bigcup_{m=1}^{T} \bigcap_{t=m}^{T} \left\{ x \in I \subseteq \mathbb{R}^D : max_{x \in \mathbb{R}^1} \mathscr{R}^t \cdot \mathscr{G}^t \cdot f(x) \leq \psi(B(x_{gbest},\ \delta)) + \frac{1}{k} \right\}$.

Importantly, it implies that $x_0 \in \bigcup_{m=1}^{T} \bigcap_{t=m}^{T} R_{t,\ K_0}$ which is the lower limit of $\{E_{t,\ k_0}\}$.

Thus, $\exists t_0,\ x_0 \in E_{t,\ k_0},\ t \geq t_0$, we can derive:

$$\psi_t(x_0) \leq \psi(B(x_{gbest},\ \delta)) + \frac{1}{k_0} (t \geq t_0). \tag{36}$$

Let t and k_0 are sufficiently large, we have $f(x_0) \leq f(B(x_{gbest},\ \delta))$.

So far, we proved $\{x,\ x_{gbest} \in \mathbb{R}^1 : \psi(x) \leq \psi(B(x_{gbest},\ \delta))\} \supseteq \bigcup_{k=1}^{T} \bigcup_{m=1}^{T} \bigcap_{t=m}^{T} \left\{ x \in \mathbb{R}^1 : max_{x \in I} \mathscr{R}^t \cdot \mathscr{G}^t \cdot \psi(x) \leq \psi(B(x_{gbest},\ \delta)) + \frac{1}{k} \right\}$.

Moreover, according to Theorem 3 and Eq. (24), we have:

$$|max_{x \in I} \{\mathscr{R}^t \cdot \mathscr{G}^t \cdot \psi(x)\} - max_{x \in \mathbb{R}^1} \{\mathscr{R}^{t+1} \cdot \mathscr{G}^{t+1} \cdot \psi(x)\}| \leq 2 \cdot |f(x_{gbest},\ \delta)|. \tag{37}$$

Therefore, we can finally conclude that the upper bound of the convergence rate of a gradient-based optimizer in each dimensionality incorporating permutation randomization is:

$$\frac{1}{\sqrt{T}} \leq 2 \cdot |f(x_{i,\ gbest},\ \delta)|,\ 1 \leq i \leq D. \tag{38}$$

Appendix C. Additional experiments

Figure 6 presents a comparison between randomized ADAM and two other leading optimizers, ADAM and SVRG, in optimizing a three-layer DBN. Notably, randomized ADAM achieves the highest reconstruction accuracy among the methods compared.

Additionally, Figure 7 shows the average reconstruction loss for randomized ADAM and two other peer optimizers over 2,000 iterations in the optimization of noisy DNMF. Specifically, this model introduces noise into the gradient g_t at every 10 iterations to enhance the complexity of the optimization process.

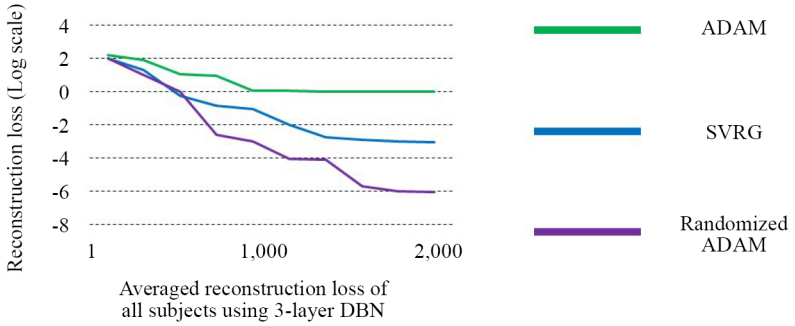


Figure 6. An illustration of reconstruction loss comparisons of randomized ADAM and other peer optimizers on optimizing a 3-layer DBN

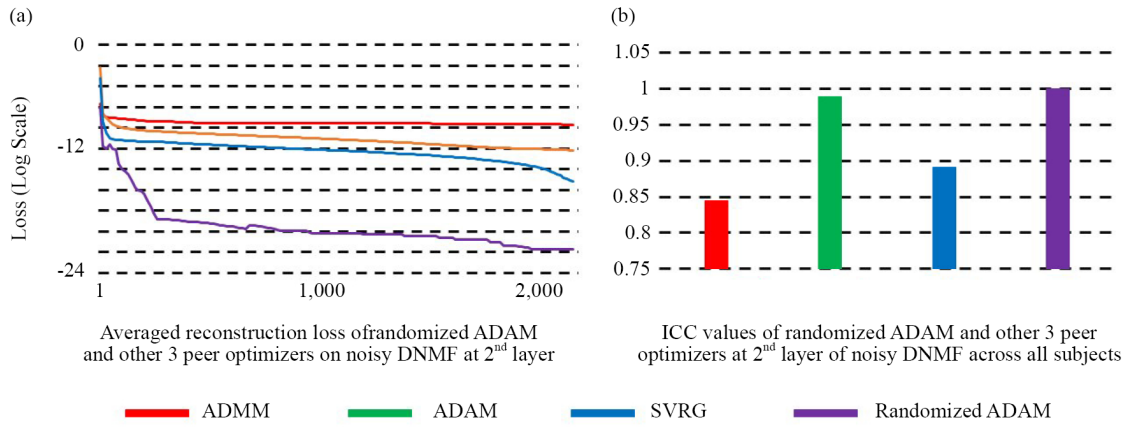
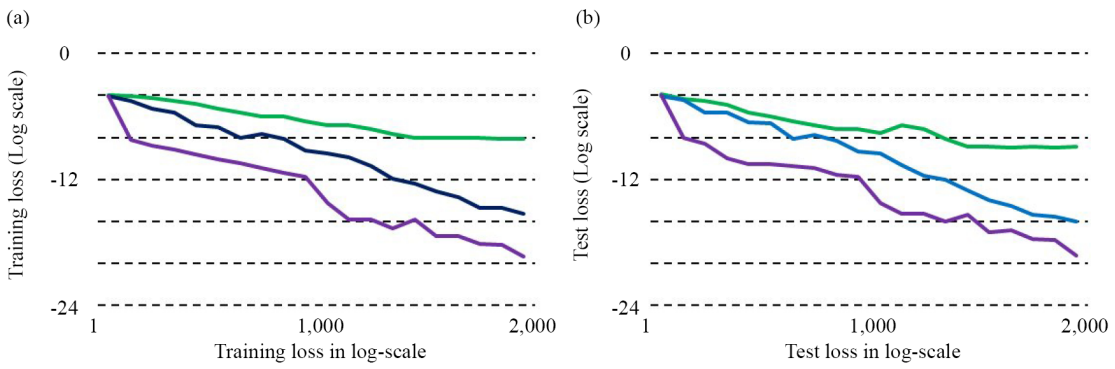


Figure 7. An illustration of reconstruction loss comparisons of randomized ADAM and other peer optimizers



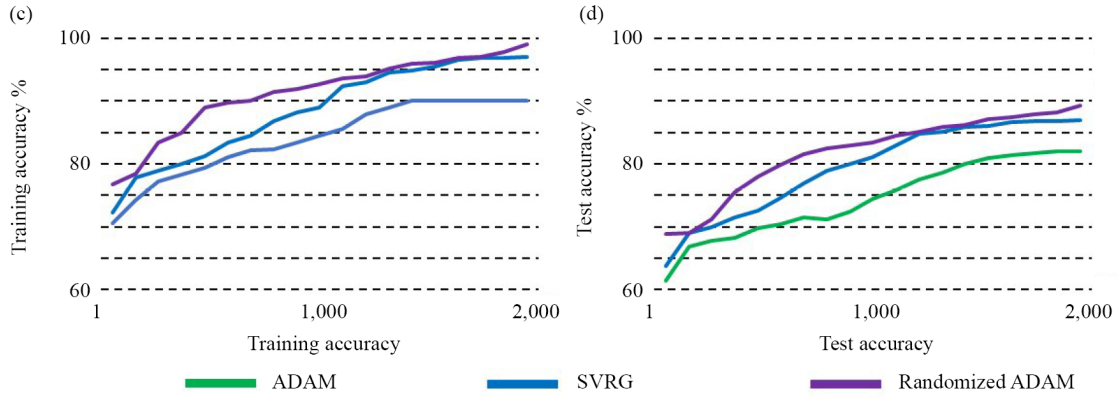


Figure 8. An illustration of reconstruction loss comparisons of randomized ADAM and other peer optimizers on solving the logistic regression problem

Finally, we provide the pseudo-code of randomized ADAM below:

Algorithm 1. Randomized ADAM: Stochastic ADAM

Input: α denotes step size; β_1 and β_2 are exponential decay;
θ_0 is initial parameter vector; m_0 is initialized 1st moment vector;
v_0 is initialized 2nd moment vector;
t is current iteration; ε is initialized as a threshold; T is the maximum iteration.
while $t < T$
$g_t \leftarrow \nabla_x f(x_t)$
$M_t \leftarrow \beta_1 M_{t-1} + (1-\beta_1) g_t$
$V_t \leftarrow \beta_2 V_{t-1} + (1-\beta_1) g_t$
$\hat{M}_t \leftarrow \dfrac{M_t}{1-\beta_1^t}$
$\hat{V}_t \leftarrow \dfrac{V_t}{1-\beta_2^t}$
$x_t \leftarrow x_{t-1} - \alpha \cdot \dfrac{\hat{M}_t}{\sqrt{\hat{V}_t} + \varepsilon}$
if $\|g_t - g_{t-1}\| < \varepsilon$
$\hat{x}_t \leftarrow \mathscr{R}(x_t)$
$x_t \leftarrow \hat{x}_t$
End if
$t \leftarrow t+1$
End while.