Journal of Machine Learning Research 25 (2024) 1-55 Submitted 11/23; Published 12/24

Zeroth-order Stochastic Approximation Algorithms for DR-submodular Optimization

Yuefang Lian LIANYF@EMAILS.BJUT.EDU.CN
Institute of Operations Research and Information Engineering
Beijing University of Technology
Beijing, 100124, China

Xiao Wang* WXUCAS@OUTLOOK.COM
Pengcheng Laboratory
Shenzhen, 518066, China

Dachuan Xu XUDC@BJUT.EDU.CN
Institute of Operations Research and Information Engineering
Beijing University of Technology
Beijing, 100124, China

Zhongrui Zhao ZHONGRUI.ZHAO@MY.JCU.EDU.AU
College of Science and Engineering
James Cook University
Queensland, 4814, Australia

Editor: Silvia Villa

Abstract

In this paper, we study approximation algorithms for several classes of DR-submodular optimization problems, where DR is short for diminishing return. Following a newly introduced algorithm framework for zeroth-order stochastic approximation methods, we first propose algorithms **CG-ZOSA** and **RG-ZOSA** for smooth DR-submodular optimization based on the coordinate-wise gradient estimator and the randomized gradient estimator, respectively. Our theoretical analysis proves that **CG-ZOSA** can reach a solution whose expected objective value exceeds $(1-e^{-1}-\epsilon^2)\mathrm{OPT}-\epsilon$ after $\mathcal{O}(\epsilon^{-2})$ iterations and $\mathcal{O}(N^{2/3}d\epsilon^{-2})$ oracle calls, where d represents the problem dimension. On the other hand, **RG-ZOSA** improves the approximation ratio to $(1-e^{-1}-\epsilon^2/d)$ while maintaining the same overall oracle complexity. For non-smooth up-concave maximization problems, we propose a novel auxiliary function based on a smoothed objective function and introduce the **NZOSA** algorithm. This algorithm achieves an approximation ratio of $(1-e^{-1}-\epsilon\ln\epsilon^{-1}-\epsilon^2\ln\epsilon^{-1})$ with $\mathcal{O}(d\epsilon^{-2})$ iterations and $\mathcal{O}(N^{2/3}d^{3/2}\epsilon^{-3})$ oracle calls. We also extend **NZOSA** to handle a class of robust DR-submodular maximization problems. To validate the effectiveness of our proposed algorithms, we conduct experiments on both synthetic and real-world problems. The results demonstrate the superior performance and efficiency of our methods in solving DR-submodular optimization problems.

Keywords: DR-submodular optimization, stochastic optimization, zeroth-order gradient estimation, robust optimization, approximation algorithm

*. Corresponding author

1. Introduction

Submodular optimization has become increasingly popular in machine-learning due to its inherent property of diminishing returns. This powerful characteristic of submodular set-functions has led to applications in various domains, including sensor placement (Guestrin et al., 2005), dictionary learning (Das and Kempe, 2011), influence maximization (Kempe et al., 2003), data summarization (Lin and Bilmes, 2010, 2011), data subset selection (Mirzasoleiman et al., 2016; Wei et al., 2015), image summarization (Mirzasoleiman et al., 2018; Feldman et al., 2018; Tschiatschek et al., 2014) and optimal budget allocation (Soma et al., 2014). The continuous extension (e.g. multi-linear extension) of submodular set-functions has gained considerable attention due to its ability to enhance the effectiveness of continuous optimization methods in solving set submodular optimization problems (Lovász, 1983; Chekuri et al., 2014; Calinescu et al., 2011; Feldman et al., 2011). Moreover, many scenarios require maximizing continuous submodular functions, an extension of the concept of submodularity to continuous domains. These examples range from non-convex/non-concave quadratic programming (Bian et al., 2017a) to robust budget allocation (Staib and Jegelka, 2017), network security games (Wilder, 2018), and so on. More applications on continuous submodular optimizaiton are referred to (Bian et al., 2017b; Bach, 2019).

The recent surge in attention given to numerical algorithms for DR-submodular optimization is notable, especially for large-scale optimization problems. Several efficient algorithms have been proposed for stochastic DR-submodular maximization, for which stochastic oracles can be available through sampling from a distribution (Hassani et al., 2017; Mokhtari et al., 2020; Hassani et al., 2020; Zhang et al., 2020, 2022; Lian et al., 2024; Chen et al., 2020b; Pedramfar et al., 2023). In practical scenarios, however, it may be difficult to obtain information about the distribution function the random variable follows, with only historical data being accessible. Taking this limitation into account, we aim to address the finite-sum DR-submodular maximization problem:

$$\max_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) := \frac{1}{N}\sum_{t=1}^{N}\mathbf{f}(\mathbf{x},\xi_t). \tag{1}$$

Here, $\mathbf{f}(\cdot,\xi_t):\mathbb{R}^d\to\mathbb{R}, t\in\{1,\dots,N\}=:[N]$, are continuous, $f:\mathbb{R}^d\to\mathbb{R}$ is monotonically non-decreasing, non-negative and DR-submodular function (refer to Definition 1), and $\mathcal{X}\subseteq\mathbb{R}^d_+$ represents a compact and convex set. In order to characterize the worst-case performance of the objective function, our focus extends to a class of robust DR-submodular maximization problems:

$$\max_{\mathbf{x}\in\mathcal{X}} \frac{1}{N}\sum_{t=1}^{N}\left\{\mathbf{f}(\mathbf{x},\xi_t) := \min_{i\in[M]}\mathbf{f}_i(\mathbf{x},\xi_{i,t})\right\}, \tag{2}$$

where $\xi_t := \{\xi_{i,t}, i\in[M]\}$, $t\in[N]$, $\mathbf{f}_i(\cdot,\xi_{i,t}):\mathbb{R}^d\to\mathbb{R}$, $i\in[M], t\in[N]$ are monotonically non-decreasing, non-negative and DR-submodular functions, and $\mathcal{X}\subseteq\mathbb{R}^d_+$ represents a compact and convex set. The max-min problem is particularly advantageous for concave functions, as their concavity is preserved under point-wise minimization operations. However, it should be noted that the minimum of a set of DR-submodular functions is up-concave (refer to Definition 2), i.e., concave along any non-negative/non-positive direction, but not

necessarily DR-submodular. Additionally, when dealing with a group of smooth functions $\mathbf{f}_i(\mathbf{x}, \xi_{i,t})$, the minimum function $\mathbf{f}(\mathbf{x}, \xi_t)$ may become non-differentiable, thus losing smoothness. Therefore, up-concave maximization problem without smoothness has attracted our interest. In this work, our goal is to develop specialized approximation algorithms that utilize only stochastic function values provided by stochastic zeroth-order oracles to maximize smooth DR-submodular functions and non-smooth up-concave functions. When applying an approximation algorithm to solve the problem $\max_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x})$, the performance of the algorithm is evaluated based on the approximation ratio γ and complexity. We quantify the algorithm's performance by the following inequality:

$$\mathbb{E}\left[f(\mathbf{x}_{\text{output}})\right] \geq \gamma f(\mathbf{x}^*) - \epsilon,$$

where $\gamma \in (0, 1]$ is the approximation ratio, $\mathbf{x}^*$ represents the optimal point, $\mathbf{x}_{\text{output}}$ is the output returned by the algorithm, and ϵ denotes the accuracy of the algorithm. The accuracy ϵ is used to derive the iteration complexity and the oracle complexity of the algorithm. For simplicity, the oracle complexity of our approximation algorithm is in terms of total number of calls to stochastic zeroth-order oracles. Our objective is to design approximation algorithms that achieve a higher approximation ratio while maintaining lower complexity.

Derivative-free optimization algorithms have gained widespread attention due to their independence from the derivative information of the objective function (Larson et al., 2019; Conn et al., 2009), including trust-region methods (Menhorn et al., 2017; Chen et al., 2018b; Shashaani et al., 2018; Larson and Billups, 2016) and direct-search methods (Dzahini, 2022; Dzahini et al., 2023). These methods are particularly useful in scenarios where only function values are available due to their sound theoretical properties, but they can only find approximate stationary points. A specific category of algorithms, known as zeroth-order optimization algorithms, which typically rely on estimating gradients using available stochastic function values and sampling techniques, are tailored to optimize functions using only observed stochastic function values. Zeroth-order optimization algorithms, which are particularly valuable when gradient computations are prohibitively costly, have consequently garnered significant attention in the field of machine learning (Nesterov and Spokoiny, 2017; Ghadimi and Lan, 2013; Xu et al., 2023). Recently, zeroth-order approximation algorithms have found applications in various domains for DR-submodular optimization. For instance, they have been used to black-box adversarial attack models, where only black-box information is accessible when attacking submodular recommendation systems (Chen et al., 2020b). In addition, for the problem of D-optimal experience design, the gradient of the objective function requires inverting a potentially large matrix (Chen et al., 2018a), and the objective function of robust DR-submodular maximization (Wilder, 2018) is non-differentiable. However, designing stochastic zeroth-order approximation algorithms for the aforementioned DR-submodular optimization problems presents several challenges and issues that need to be addressed.

The zeroth-order version of continuous greedy algorithm and Frank-Wolfe algorithm have been designed for black-box DR-submodular maximization problem (Chen et al., 2020b; Pedramfar et al., 2023). However, the oracle complexity of the algorithm in (Chen et al., 2020b) is relatively high and strongly depends on the problem dimension d, and the performance of stochastic zeroth-order algorithms in (Pedramfar et al., 2023) seems not quite satisfying when applying to (1), as shown in Table 1. The large variance arising from

the errors of the approximate gradient may lead to divergence in the approximation analysis. Therefore, designing an efficient zeroth-order approximation algorithm for DR-submodular maximization with higher approximation ratio and lower oracle complexity is a challenging task. Moreover, the existing literature primarily focuses on smooth DR-submodular maximization problems, while the research on non-smooth up-concave maximization is still limited. Specifically, Lee et al. (2022) has introduced a sub-gradient method to deal with the non-differentiable up-concave function, but the approximation ratio still needs to be improved and there is no theoretical guarantee of zeroth-order algorithm for non-smooth DR-submodular maximization problem. Furthermore, note that the minimum of a set of smooth DR-submodular functions may be non-smooth, thus the challenge of solving a non-smooth up-concave maximization problem highlights the non-trivial nature of robust DR-submodular maximization. In fact, the specific structure of robust DR-submodular maximization poses additional challenges and opportunities for developments of zeroth-order approximation algorithms.

To address the aforementioned challenges, we have identified two main techniques to ensure desirable approximation ratio and oracle complexity in our analysis. The first one, inspired by the stochastic variance reduction technique in (Johnson and Zhang, 2013), is the double-loop zeroth-order stochastic approximation algorithm framework. Initially introduced for stochastic convex optimization (Johnson and Zhang, 2013), this technique has been extensively studied for general non-convex minimization problems and successfully improved the convergence rate in works such as (Reddi et al., 2016a,b; Allen-Zhu and Hazan, 2016). We will demonstrate that it can also reduce the variance in approximation analysis, allowing the zeroth-order algorithm to achieve the desired approximation guarantee more efficiently. The second technique we employ is leveraging an auxiliary function to enhance performance. Ranther than designing an approximate gradient for the original objective function, applying the stochastic gradient algorithm to a related auxiliary function has been shown to effectively improve the approximation ratio, as demonstrated in (Zhang et al., 2022). This technique was first introduced in (Alimonti, 1994; Khanna et al., 1998) as an alternative to classical local search methods like the greedy algorithm, and it has been employed to improve approximation algorithms for set function maximization problems by carefully selecting the auxiliary function (Filmus and Ward, 2014). More recently, the similar techniques in various forms have been used for continuous submodular maximization to boost the approximation ratio (Mitra et al., 2021; Zhang et al., 2022). However, the integral auxiliary function utilized in (Zhang et al., 2022) may cause divergence in the non-smooth case due to the introduced error. To tackle this issue, we introduce a finite-sum auxiliary function based on a smoothed function, which effectively mitigates the variance caused by the biased integral auxiliary function.

1.1 Contributions

Our main contributions are summarized as follows.

- For the smooth DR-submodular maximization problem, by incorporating the integral auxiliary function we propose two zeroth-order stochastic approximation algorithms, which fall into a generic framework for zeroth-order stochastic approximation methods (Algorithm 1). The first is the **CG-ZOSA** algorithm, which employs coordinate-wise

gradient estimation of an integral auxiliary function to design approximate gradients in Algorithm 1. It can find an approximate solution whose expected objective value exceeds $(1 - e^{-1} - \epsilon^2)\text{OPT} - \epsilon$ after $\mathcal{O}(\epsilon^{-2})$ iterations and $\mathcal{O}(N^{2/3}d\epsilon^{-2})$ oracle calls. The second is the **RG-ZOSA** algorithm, which uses randomized gradient estimation of an integral auxiliary function to design approximate gradients in Algorithm 1, improving the approximation ratio to $(1 - e^{-1} - \epsilon^2/d)$. Detailed comparison between our algorithms and existing works is presented in Table 1.

- For the non-smooth up-concave maximization problem, we propose a zeroth-order stochastic approximation algorithm by designing a finite-sum auxiliary function. This auxiliary function with finite-sum form allows for a trade-off between the approximation ratio and oracle complexity. Additionally, we propose the **NZOSA** algorithm, an enhanced version of **RG-ZOSA**, which achieves a $(1 - e^{-1} - \epsilon \ln \epsilon^{-1} - \epsilon^2 \ln \epsilon^{-1})$-approximation guarantee with $\mathcal{O}(d\epsilon^{-2})$ iterations and $\mathcal{O}(N^{2/3}d^{3/2}\epsilon^{-3})$ oracle calls under mild conditions. To the best of our knowledge, the study on zeroth-order algorithms for non-smooth up-concave maximization problems is novel in the literature. Detailed comparison between our algorithm and existing works is presented in Table 2.

- We extend **NZOSA** to handle a class of robust DR-submodular maximization problems and evaluate our proposed algorithms on synthetic and real-world tasks, including quadratic programming, multi-resolution data summarization and robust budget allocation problems. The numerical results demonstrate the effectiveness and efficiency of our algorithms.

1.2 Related Work

We next list research works that are closely related to our paper.

Stochastic submodular maximization. Submodular set-function maximization originates from combinatorial optimization (Nemhauser et al., 1978; Fisher et al., 1978). The application of continuous optimization techniques (Lovász, 1983; Chekuri et al., 2014) to solve submodular set-function optimization problems has attracted much attention. The appeal of such techniques lies in their ability to address a wider range of constraints associated with submodular set-function maximization. Consequently, the development of continuous greedy algorithms emerges as a solution for problems with multi-linear extensions, representing a special case of DR-submodular functions (Vondrák, 2008; Calinescu et al., 2011; Feldman et al., 2011). Subsequently, these algorithms are further analyzed in the context of general smooth and continuous DR-submodular functions (Bian et al., 2017a,b; Du, 2022).

To address scenarios where only the stochastic gradient information is available, researchers have developed stochastic first-order approximation algorithms (Mokhtari et al., 2020; Hassani et al., 2020; Zhang et al., 2020; Hassani et al., 2017; Zhang et al., 2022; Lian et al., 2024) for both monotone and non-monotone cases. One notable method is the stochastic continuous greedy algorithm (SCG), first proposed for monotone DR-submodular maximization by Mokhtari et al. (2020), with at least $(1 - 1/e)\text{OPT} - \epsilon$ approximation guarantee after $\mathcal{O}(\epsilon^{-3})$ stochastic gradient oracle calls. By incorporating variance reduction

Algorithm	Setting	Oracle	Utility	Iter. Comp.	Ora. Comp.
GA (Hassani et al., 2017)	stochastic	sto. grad.	$(1/2)\text{OPT}-\epsilon$	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-2})$
SCG (Mokhtari et al., 2020)	stochastic	sto. grad.	$(1-1/e)\text{OPT}-\epsilon$	$\mathcal{O}(\epsilon^{-3})$	$\mathcal{O}(\epsilon^{-3})$
BGA (Zhang et al., 2022)	stochastic	sto. grad.	$(1-1/e)\text{OPT}-\epsilon$	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-2})$
BCG (Chen et al., 2020b)	stochastic	sto. func.	$(1-1/e)\text{OPT}-\epsilon$	$\mathcal{O}(d\epsilon^{-3})$	$\mathcal{O}(d^3\epsilon^{-5})$
FW (Pedramfar et al., 2023)	stochastic	sto. func.	$(1/2)\text{OPT}-\epsilon$	$\tilde{\mathcal{O}}(d\epsilon^{-3})$	$\tilde{\mathcal{O}}(d\epsilon^{-5})$
CG-ZOSA (Theorem 12)	finite-sum	sto. func.	$(1-1/e-\epsilon^2)\text{OPT}-\epsilon$	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(N^{2/3}d\epsilon^{-2})$
RG-ZOSA (Theorem 17)	finite-sum	sto. func.	$(1-1/e-\epsilon^2/d)\text{OPT}-\epsilon$	$\mathcal{O}(d\epsilon^{-2})$	$\mathcal{O}(N^{2/3}d\epsilon^{-2})$

Table 1: Comparison of stochastic approximation algorithms for smooth monotone DR-submodular optimization. The terms "Iter. Comp." and "Ora. Comp." abbreviate iteration complexity and oracle complexity, respectively, indicating the total number of iterations and calls to zeroth-order oracles. "Sto. grad." and "Sto. func." are short for stochastic gradient and stochastic function value. The notation $\tilde{\mathcal{O}}$ is used to hide the logarithmic factor in the approximation error bound, and OPT represents the optimal objective value. It is worth noting that the works of stochastic algorithms presented here operate under the standard L-smoothness assumption and general convex set constraints, and the algorithm **RG-ZOSA** requires extra L_0-Lipschitz continuous assumption. Compared to existing zeroth-order approximation algorithms, our proposed algorithms significantly improve the complexities while achieving utility that approaches $(1-1/e)\text{OPT}$, when ϵ is sufficiently small. Additional relevant work that depends on supplementary assumptions (e.g. high-order smoothness) will be explored in Subsection 1.2.

Algorithm	Setting	Ora.	Utility	Iter. Comp.	Ora. Comp.
MP (Lee et al., 2022)	stochastic DR	sto. grad.	$(1/2)\text{OPT}-\epsilon$	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-2})$
NZOSA (Theorem 24)	finite-sum up-concave	sto. func.	$(1-e^{-1}-\epsilon\ln\epsilon^{-1}-\epsilon^2\ln\epsilon^{-1})\text{OPT}-\epsilon$	$\mathcal{O}(d\epsilon^{-2})$	$\mathcal{O}(N^{2/3}d^{3/2}\epsilon^{-3})$

Table 2: Comparison of stochastic approximation algorithms for non-smooth DR-submodular (up-concave) optimization. Note that the algorithm **NZOSA** can be applied to general non-smooth up-concave optimization, and can achieve a utility close to $(1-1/e)$ when ϵ becomes sufficiently small, though this comes at the cost of increased complexity.

techniques, the improved version SCG++ in (Hassani et al., 2020) achieves $(1-1/e)$ approximation ratio with $\mathcal{O}(\epsilon^{-2})$ stochastic gradient evaluations under high-order smoothness assumption. Recently Lian et al. (2024) further weakens the assumption conditions and obtain the same approximation guarantee by adopting stochastic variance reduction tech-

nique on SCG. Another efficient method for monotone DR-submodular maximization is the stochastic gradient method (Hassani et al., 2017; Zhang et al., 2022). In particular, the stochastic projected gradient ascent (GA) algorithm in (Hassani et al., 2017) reaches a 1/2-approximation to the global maximum after $\mathcal{O}(\varepsilon^{-2})$ iterations by identifying a stable point, and the boosting gradient ascent (BGA) algorithm proposed in (Zhang et al., 2022) yields a solution whose objective function achieves at least $(1-1/e)\mathrm{OPT}-\epsilon$ after $\mathcal{O}(\epsilon^{-2})$ iterations, which effectively improves the approximation ratio from 1/2 by designing an auxiliary function. For non-monotone DR-submodular maximization problems under down-closed convex set constraints, the SMCG++ algorithm introduced in (Hassani et al., 2020) achieves a $1/e$-approximation ratio after $\mathcal{O}(\epsilon^{-2})$ stochastic gradient evaluations. Furthermore, the SPIDER-CG algorithm presented in (Lian et al., 2024) provides a $\frac{1}{4}(1-\min_{x\in\mathcal{C}}\|x\|_\infty)\mathrm{OPT}-\epsilon$ guarantee with $\mathcal{O}(\epsilon^{-2})$ stochastic gradient oracles under general convex constraint sets.

Zeroth-order stochastic optimization. Zeroth-order optimization algorithms have become increasingly popular in scenarios where only function value information is available. A variety of algorithms have been proposed, such as those introduced by Ghadimi and Lan (2013); Duchi et al. (2015); Nesterov and Spokoiny (2017), which employ Gaussian smoothing techniques applicable to both convex and non-convex optimization problems. Accelerated zeroth-order algorithms based on variance reduction techniques have also been studied (Liu et al., 2018a; Ji et al., 2019). Additionally, research has also focused on zeroth-order algorithms for constrained optimization (Balasubramanian and Ghadimi, 2018; Huang et al., 2019; Chen et al., 2020a; Liu et al., 2018b) and non-smooth optimization (Huang et al., 2020; Ghadimi et al., 2016; Lin et al., 2022).

The study of zeroth-order stochastic approximation algorithms for DR-submodular maximization is yet limited. The black-box continuous greedy (BCG) algorithm (Chen et al., 2020b) achieves the tight $(1-e^{-1})\mathrm{OPT}-\epsilon$ approximation guarantee after $\mathcal{O}(d^3\epsilon^{-5})$ function evaluations in a stochastic setting. In (Pedramfar et al., 2023), a unified algorithm framework is presented for maximizing continuous DR-submodular functions, encompassing various settings and oracle access types. Among them, the stochastic Frank-Wolfe algorithm achieves a $(1/2)\mathrm{OPT}-\epsilon$ approximation guarantee under general convex constraint set $\mathcal{C}$ after $\mathcal{O}(d\epsilon^{-5})$ calls to stochastic zeroth-order oracles, and the stochastic continuous greedy algorithm achieves $(1-1/e)$-approximation for the case that the vector $\mathbf{0}$ belongs to the set $\mathcal{C}$. However, for more general non-smooth up-concave maximization, the study of zeroth-order approximation algorithms is still scarce.

Robust submodular maximization. Krause et al. (2008) studies the robust set submodular maximization problem $\max_{S\in 2^V}\min_{i\in[n]} f_i(S)$, where $f_i(S): 2^V \to \mathbb{R}$ for $i \in [n]$. Various extensions of robustness have been explored in works such as Chen et al. (2016); Staib et al. (2019); Adibi et al. (2022). For continuous robust submodular maximization, Wilder (2018) introduces a randomized smoothing technique to handle zero-sum games with submodular structure in network security games. Staib et al. (2019) proves that the objective becomes smooth under a high sample variance condition for distributionally robust submodular maximization defined by $\mathcal{X}^2$-divergence. Lee et al. (2022) systematically studies Hölder-smooth and non-smooth submodular problems, proposing a variant of the mirror-prox algorithm that achieves a $(1/2)\mathrm{OPT}-\epsilon$ guarantee for the non-smooth case. Additionally, the mirror-prox method is applied to robust submodular maximization to

obtain a solution whose function value is at least $(1/2)\mathrm{OPT}-\epsilon$. In our paper, we present the first zeroth-order approximation algorithm for robust DR-submodular maximization problem.

1.3 Application Examples

We next give several examples for DR-submodular optimization in real-world applications.

Quadratic programming. Non-convex quadratic programming has many applications in machine learning. Generally, it cannot be solved for a global optimal solution in polynomial time. However, the DR-submodular function introduces a non-convex structure that can be solved approximately. We consider to maximize a DR-submodular function f defined as

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x} + h^T\mathbf{x} + c,$$

subject to linear constraints. This type of problems arise in scheduling (Skutella, 2001; Bian et al., 2017b). The DR-submodular quadratic programming problem requires that all off-diagonal entries of $\mathbf{H}$ are non-positive, i.e., $\mathbf{H}_{i,j} \leq 0, \forall i, j$, and it is obviously smooth.

Multi-resolution data summarization. The multi-resolution summary problem (Bian et al., 2017b; Lee et al., 2022) with a non-differentiable utility function is regarded as a non-smooth up-concave maximization problem. Specifically, given a set $E = \{e_1, \ldots, e_k\}$, we assign each data e_i a non-negative weight $\mathbf{x}_i$ to measure its importance by which to determine whether the data is recommended. In general, the user will set a threshold τ to decide the set $S_\tau = \{e_i : \mathbf{x}_i \geq \tau\}$ which represents the level of details or resolution of summary. The weight of each point e_i is determined by a DR-submodular maximization problem where the utility function is given by

$$\sum_{i=1}^{k}\sum_{j=1}^{k}\phi(\mathbf{x}_j)s_{ij} - \sum_{i=1}^{k}\sum_{j=1}^{k}\mathbf{x}_i\mathbf{x}_j s_{ij}.$$

Here, $s_{ij} \geq 0$ is the similarity index between two items e_i and e_j, and $\phi(\cdot)$ is monotone and concave. It is obviously that utility function above is up-concave.

Optimal budget allocation. Optimal budget allocation is a classical submodular maximization problem (Bian et al., 2017b). We describe it as a bipartite graph $(S, T; W)$, where S and T are sets of advertising channels (e.g., TV and websites) and customers respectively, and $W \subseteq S \times T$ is an edge set. Each edge has a weight $\{p_{st}\}_{(s,t)\in W}$ which denotes the influence probability of channel s to customer t. The goal is to maximize the expected influence on customers by distributing the budgets, which concludes the time or space of advertisements (Soma et al., 2014; Hatano et al., 2015). More precisely, the total influence of customer t from all channels can be modeled as

$$I_t(\mathbf{x}) = 1 - \prod_{(s,t)\in W}(1-p_{st})^{\mathbf{x}_s},$$

where $\mathbf{x} \in \mathbb{R}_+^S$ denotes the budget assignment from S channels. The objective is to maximize $\sum_{t\in T} I_t(\mathbf{x})$, where $\mathbf{x}$ is constrained within some set $\mathcal{X}$. It is easy to prove that $I_t(\mathbf{x})$ is monotone and DR-submodular according to the second-order equivalent condition of

DR-submodularity (Bian et al., 2017b), i.e., $\frac{\partial^2 I_t(\mathbf{x})}{\partial \mathbf{x}_i \mathbf{x}_j} \leq 0, \forall i, j$. If there are N advertisers providing N schemes, let $\mathbf{x} = [\mathbf{x}^1, \ldots, \mathbf{x}^N]$, where $\mathbf{x}^i \in \mathbb{R}_+^S$. We aim to maximize the impact on the least affected customers with the influence of N advertisers, modeled as

$$\max_{\mathbf{x}} \sum_{i=1}^{N} \alpha_i \min_{t \in T} I_t(\mathbf{x}^i), \tag{3}$$

where $\alpha_i > 0$ denotes the weights. Obviously, (3) is a robust DR-submodular maximization problem.

1.4 Organization

The rest of our paper is organized as follows. Preliminaries, including relevant notations, definitions, zeroth-order gradient estimations and a generic zeroth-order stochastic approximation algorithm framework, are given in Section 2. In Section 3, we propose two specific zeroth-order stochastic approximation methods, **CG-ZOSA** and **RG-ZOSA**, and present their approximation analysis for smooth DR-submodular maximization problem. Section 4 introduces a novel finite-sum auxiliary function for smoothed functions, and then designs the **NZOSA** algorithm for non-smooth up-concave maximization. We further extend **NZOSA** to apply for a class of robust submodular optimization problems in Section 5. The efficiency of the proposed algorithms is validated by extensive experiments in Section 6. Finally, we present our conclusions in Section 7.

2. Preliminaries

2.1 Notations and Definitions

We denote basis vectors by $\mathbf{e}_i := (0, \ldots, 1, 0, \ldots, 0)^T \in \mathbb{R}^d$, where the i-th component is 1, and the radius and diameter of the constraint set $\mathcal{X}$ by $R := \max_{\mathbf{x} \in \mathcal{X}} \|\mathbf{x}\|$ and $D := \max_{\mathbf{x}, \mathbf{y} \in \mathcal{X}} \|\mathbf{x} - \mathbf{y}\|$ respectively, where $\|\cdot\|$ is ℓ_2-norm by default. For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, we denote $(\mathbf{x} \vee \mathbf{y})_i = \max\{\mathbf{x}_i, \mathbf{y}_i\}$ and $(\mathbf{x} \wedge \mathbf{y})_i = \min\{\mathbf{x}_i, \mathbf{y}_i\}$, and $\mathbf{x} \leq \mathbf{y}$ means $\mathbf{x}_i \leq \mathbf{y}_i, \forall i \in [d]$. Given $\mathbf{x} \in \mathbb{R}^d$ and $u > 0$, denote by $\mathbb{B}_d(\mathbf{x}, u) = \{\mathbf{y} \in \mathbb{R}^d | \|\mathbf{x} - \mathbf{y}\| \leq u\}$ the ℓ_2-norm ball of radius u centered at $\mathbf{x}$, and denote $\mathbb{S} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| = 1\}$ as the unit sphere. Given $\mathcal{D} \subseteq \mathbb{R}^d$, we say that $f : \mathcal{D} \to \mathbb{R}$ is L_0-Lipschitz continuous w.r.t. the norm $\|\cdot\|$ if $|f(\mathbf{x}) - f(\mathbf{y})| \leq L_0 \|\mathbf{x} - \mathbf{y}\|$ for any $\mathbf{x}, \mathbf{y} \in \mathcal{D}$, and that f is L-smooth if its gradient is L-Lipschitz continuous. A continuous function $f : \mathcal{D} \to \mathbb{R}$ is said to be submodular, if

$$f(\mathbf{x}) + f(\mathbf{y}) \geq f(\mathbf{x} \vee \mathbf{y}) + f(\mathbf{x} \wedge \mathbf{y})$$

holds for all $\mathbf{x}, \mathbf{y} \in \mathcal{D}$. If furthermore f is twice differentiable, the submodularity is equivalent to that all off-diagonal entries of its Hessian are non-positive, i.e., $\frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}_i \mathbf{x}_j} \leq 0, \forall \mathbf{x} \in \mathcal{D}$ and $i \neq j$.

Definition 1. (**DR-submodular function**) Given a convex set $\mathcal{D} \subseteq \mathbb{R}^d$, a continuous function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be DR-submodular over $\mathcal{D}$, if for any $\mathbf{x} \leq \mathbf{y} \in \mathcal{D}$ and $a \in \mathbb{R}_+$ such that $(a\mathbf{e}_i + \mathbf{x}) \in \mathcal{D}, (a\mathbf{e}_i + \mathbf{y}) \in \mathcal{D}$, $\forall i \in \{1, 2, \ldots, d\}$, the following diminishing return (DR) property holds:

$$f(a\mathbf{e}_i + \mathbf{x}) - f(\mathbf{x}) \geq f(a\mathbf{e}_i + \mathbf{y}) - f(\mathbf{y}).$$

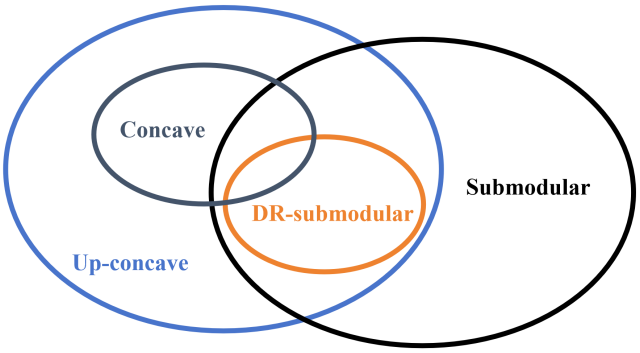


Figure 1: The relationship between concavity, up-concavity, DR-submodularity, and submodularity.

Note that Definition 1 is equivalent to $\nabla f(\mathbf{x}) \leq \nabla f(\mathbf{y}), \forall\, \mathbf{x} \geq \mathbf{y}$ when f is differentiable, and $\frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}_i \mathbf{x}_j} \leq 0, \forall \mathbf{x} \in \mathcal{D}$ and $\forall i, j$, when $f(\mathbf{x})$ is twice differentiable. If f is continuously differentiable, it holds that

$$\langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \geq f(\mathbf{x} \vee \mathbf{y}) + f(\mathbf{x} \wedge \mathbf{y}) - 2f(\mathbf{x}), \quad \forall\, \mathbf{x}, \mathbf{y} \in \mathcal{D}. \tag{4}$$

For more properties of DR-submodular functions, we refer interested readers to (Hassani et al., 2017; Bian et al., 2017a,b).

Definition 2. (**Up-concave function**) A function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be up-concave, if for any $\mathbf{x} \in \mathbb{R}^d$ and any non-negative direction $\mathbf{v} \geq \mathbf{0}$, $f(\mathbf{x} + \beta \mathbf{v})$ is concave w.r.t. $\beta \in \mathbb{R}$.

By Definition 2, if $f : \mathbb{R}^d \to \mathbb{R}$ is an up-concave function at any given point $\mathbf{x} \in \mathbb{R}^d$, there exists $g_{\mathbf{x}}$ such that

$$f(\mathbf{y}) \leq f(\mathbf{x}) + g_{\mathbf{x}}^T (\mathbf{y} - \mathbf{x}), \quad \forall \mathbf{y} \leq \mathbf{x} \text{ or } \mathbf{y} \geq \mathbf{x}. \tag{5}$$

Denote $\partial^{\uparrow} f(\mathbf{x})$ as the set of $g_{\mathbf{x}}$ that satisfies above inequality. If f is differentiable, $\partial^{\uparrow} f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}$, which also shows that f is concave along any non-negative/non-positive direction. Furthermore, if f is twice differentiable, it holds that

$$(\mathbf{y} - \mathbf{x})^T \nabla^2 f(\mathbf{x})(\mathbf{y} - \mathbf{x}) \leq 0, \quad \forall \mathbf{y} \leq \mathbf{x} \text{ or } \mathbf{y} \geq \mathbf{x}.$$

Moreover, it is obvious that a continuous concave function is up-concave. A DR-submodular function f must be up-concave according to the definition or the equivalent conditions of DR-submodular, but the converse is not true.

We present a description of the relationship among submodularity, DR-submodularity, concavity and up-concavity in Figure 1.

2.2 Zeroth-order Gradient Estimation

The focus of this paper is to design stochastic approximation algorithms based on zeroth-order gradient estimations and investigate their properties accordingly. There are two types

of zeroth-order gradient estimation that are widely used in the literature: the coordinate-wise gradient estimation and the randomized gradient estimation. Specifically, for $\mathbf{f}(\mathbf{x}, \xi_t)$, define

$$\textbf{(CoordGradEst)} \qquad \mathbf{g}_{\text{coord}}(\mathbf{x}, \xi_t) = \sum_{l=1}^{d} \frac{1}{2u} \left(\mathbf{f}(\mathbf{x} + u\mathbf{e}_l, \xi_t) - \mathbf{f}(\mathbf{x} - u\mathbf{e}_l, \xi_t)\right) \mathbf{e}_l,$$

$$\textbf{(RandGradEst)} \qquad \mathbf{g}_{\text{rand}}(\mathbf{x}, \nu, \xi_t) = \frac{d}{2u} \left(\mathbf{f}(\mathbf{x} + u\nu, \xi_t) - \mathbf{f}(\mathbf{x} - u\nu, \xi_t)\right) \nu,$$

where $u \in \mathbb{R}_+$ and $\nu \in \mathbb{R}^d$ is sampled following the uniform distribution on the unit sphere $\mathbb{S}$. Note that the randomized gradient estimator $\mathbf{g}_{\text{rand}}(\mathbf{x}, \nu, \xi_t)$ relies on the random variable ν, thus providing a stochastic approximation to $\nabla \mathbf{f}(\mathbf{x}, \xi_t)$. The coordinate-wise gradient estimator generates a deterministic approximation with the upper bound of the estimation error addressed in the lemma below. Proofs of lemmas in this subsection are presented in Appendix A.

Lemma 3. *Assume that $\mathbf{f}(\mathbf{x}, \xi_t)$ is L-smooth, then $\|\mathbf{g}_{\text{coord}}(\mathbf{x}, \xi_t) - \nabla \mathbf{f}(\mathbf{x}, \xi_t)\|^2 \leq L^2 d u^2$.*

To investigate the properties of the randomized gradient estimator $\mathbf{g}_{\text{rand}}(\mathbf{x}, \nu, \xi_t)$, we introduce a smoothing function for $\mathbf{f}(\mathbf{x}, \xi_t)$. Let μ be the uniform density w.r.t. Lebesgue measure on the $\mathbb{B}_d(\mathbf{0}, u)$, the smoothed function $\mathbf{f}_\mu : \mathbb{R}^d \times \Xi \to \mathbb{R}$ is defined as

$$\mathbf{f}_\mu(\mathbf{x}, \xi_t) := \int_{\mathbb{R}^d} \mathbf{f}(\mathbf{x} + \mathbf{y}, \xi_t) \mu(\mathbf{y}) d\mathbf{y} = \mathbb{E}_v \left[\mathbf{f}(\mathbf{x} + uv, \xi_t)\right], \tag{6}$$

where v is a random variable chosen uniformly from the d-dimensional unit ball $\mathbb{B}_d(\mathbf{0}, 1)$. Meanwhile, we define $f_\mu : \mathbb{R}^d \to \mathbb{R}$ as

$$f_\mu(\mathbf{x}) := \frac{1}{N} \sum_{t=1}^{N} \mathbf{f}_\mu(\mathbf{x}, \xi_t) = \mathbb{E}_v[f(\mathbf{x} + uv)]. \tag{7}$$

The following lemma summarizes the key properties of $f_\mu(\mathbf{x})$.

Lemma 4. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be monotonically non-decreasing, non-negative and DR-submodular (resp. up-concave). Then the following statements hold true.*

(i) $f_\mu(\mathbf{x})$ is monotonically non-decreasing, non-negative and DR-submodular (resp. up-concave).

(ii) If $f(\mathbf{x})$ is L-smooth, $f_\mu(\mathbf{x})$ is also L-smooth.

(iii) If $\mathbf{f}(\mathbf{x}, \xi_t)$ is L_0-Lipschitz continuous for any $t \in [N]$, $f_\mu(\mathbf{x})$ is differentiable and $\frac{L_0\sqrt{d}}{u}$-smooth, and moreover, $|f_\mu(\mathbf{x}) - f(\mathbf{x})| \leq L_0 u$.

It is noteworthy that the smoothing parameter of f_μ depends on u when f is merely L_0-Lipschitz continuous. This dependency is crucial for the algorithm design and theoretical analysis in the non-smooth case.

We now present the properties of the randomized gradient estimator as follows.

Lemma 5. *Suppose that $\mathbf{f}(\mathbf{x}, \xi_t)$ is L_0-Lipschitz continuous for any $t \in [N]$, then we have*

$$\mathbb{E}[\mathbf{g}_{\text{rand}}(\mathbf{x}, \nu, \xi_t)|\mathbf{x}] = \nabla \mathbf{f}_\mu(\mathbf{x}, \xi_t), \qquad \|\mathbf{g}_{\text{rand}}(\mathbf{x}, \nu, \xi_t) - \mathbf{g}_{\text{rand}}(\mathbf{y}, \nu, \xi_t)\|^2 \leq M_u^2 \|\mathbf{x} - \mathbf{y}\|^2, \quad (8)$$

where $M_u = \frac{dL_0}{u}$. Furthermore, it holds that

$$\mathbb{E}[\|\nabla f_\mu(\mathbf{x}, \xi_t) - \mathbf{g}_{\text{rand}}(\mathbf{x}, \nu, \xi_t)\|^2|\mathbf{x}] \leq 16\sqrt{2\pi} d L_0^2, \quad \forall t \in [N]. \quad (9)$$

Remark 6. *For the case that $\mathbf{f}(\mathbf{x}, \xi_t)$ is defined on $\mathcal{D} = \prod_{i=1}^d [0, a_i]$, the function may not be defined at the point $\mathbf{x} + u\nu$ or $\mathbf{x} + u\mathbf{e}_l$ if they fall outside of $\mathcal{D}$. To address this issue, we can either shrink the domain $\mathcal{D}$ to $\prod_{i=1}^d [u, a_i - u]$ with small $u > 0$ (Chen et al., 2020b), or construct a set $\mathcal{D}_u$ as described in (Pedramfar et al., 2023), ensuring that the above gradient estimators are well-defined. The purpose of these strategies is to ensure the meaningful gradient estimation computations within the feasible region. In our setting, for simplicity we define $\mathbf{f}(\mathbf{x}, \xi_\mathbf{t})$ in $\mathbb{R}^d$ for $t \in [N]$, meaning that the oracle can be queried for any point in $\mathbb{R}^d$.*

2.3 Zeroth-Order Stochastic Approximation Algorithm Framework

There are several stochastic gradient approximation methods in the literature for the DR-submodular maximization problem. The stochastic gradient ascent (GA) algorithm (Hassani et al., 2017) for the DR-submodular maximization problem in the expected form has been shown to achieve $1/2$ approximation ratio by randomly selecting an approximate gradient. To further improve the approximation ratio, Zhang et al. (2022) designed an approximate gradient of an auxiliary function F. Consequently, the approximation performance of the gradient ascent (GA) algorithm was improved from $1/2$ to $1 - 1/e$ with access to stochastic gradient information. It is important to note that both aforementioned methods follow a single-loop algorithm framework based on stochastic first-order information. These zeroth-order variants, which rely on a similar algorithm structure, may not enhance the approximation performance. For example, the zeroth-order gradient ascent algorithm (ZO-GA, presented in Algorithm 3), which directly replaces the stochastic gradient with a zeroth-order stochastic gradient estimation, can be shown to maintain an approximation ratio of $1/2$ after $\mathcal{O}(\frac{Nd}{\epsilon^2})$ oracle calls (see the approximation analysis in Appendix B). To achieve better approximation performance, the design of more efficient algorithm structures and a deeper analysis of these algorithms are desired.

The stochastic variance reduced gradient (SVRG) algorithm (Johnson and Zhang, 2013) is designed for minimizing a finite-sum of smooth functions, i.e., $f(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^N f_i(\mathbf{x})$. Within a double-loop framework, SVRG computes an approximate gradient at each inner iteration inside an epoch as

$$\mathbf{d}_k = \nabla f_i(\mathbf{x}_k) - \nabla f_i(\bar{\mathbf{x}}) + \nabla f(\bar{\mathbf{x}}),$$

where i is chosen uniformly and randomly from $[N]$, and the full gradient ∇f is computed at a snapshot point $\bar{\mathbf{x}}$ at the beginning of each inner loop. It is worthy to note that SVRG relies on gradient computations of component functions, whereas our problem setting only assumes access to zeroth-order oracles.

Here, we present a generic framework of zeroth-order stochastic approximation methods for DR-submodular optimization in Algorithm 1.

Algorithm 1 Zeroth-order Stochastic Approximation (ZOSA) Algorithm Framework

Input: Initial point $\mathbf{x}_m^0 = \mathbf{x}_0 \in \mathcal{X}$, positive integers m and S, step-sizes $\{\eta_j^{s+1}\}$.
Output: $\mathbf{x}_r$

1: **for** $s = 0, \dots S-1$ **do**
2: Set $\mathbf{x}_0^{s+1} = \mathbf{x}_m^s$
3: Compute zeroth-order approximate gradient $\mathbf{d}_0^{s+1}$ at $\mathbf{x}_0^{s+1}$
4: **for** $j = 0, \dots, m-1$ **do**
5: Compute zeroth-order approximate gradient $\mathbf{d}_j^{s+1}$ at $\mathbf{x}_j^{s+1}$
6: Compute $\mathbf{y}_{j+1}^{s+1} = \mathbf{x}_j^{s+1} + \eta_j^{s+1}\mathbf{d}_j^{s+1}$
7: Compute $\mathbf{x}_{j+1}^{s+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x} - \mathbf{y}_{j+1}^{s+1}\|^2$
8: **end for**
9: **end for**
10: Return $\mathbf{x}_r$ according to a certain type of probability distribution from $\mathbf{x}_m^S$ and $\mathbf{x}_j^{s+1}$, $j = 0, \dots, m-1; s = 0, \dots, S-1$.

We briefly describe the algorithm framework as follows. The input initializes the starting point, the number of iterations for the inner and outer loops and the step size at each iteration. Lines 2 and 3 set the starting point of each loop and compute the approximate gradient. Line 5 computes the zeroth-order approximate gradient at each inner iteration. Through Lines 6 and 7 we compute the next iterate, assuming availability of Euclidean projection onto the convex set $\mathcal{X}$. Line 10 returns a random vector according to a certain distribution.

To achieve desirable approximation guarantees, we attempt to leverage an auxiliary function and compute approximate gradients based on the proposed zeroth-order stochastic approximation (ZOSA) algorithm framework. In subsequent sections, we will propose zeroth-order stochastic approximation methods utilizing different gradient estimators and auxiliary function forms for both smooth and non-smooth cases, and provide the respective approximation analysis.

3. Zeroth-order Stochastic Approximation Methods for Smooth DR-Submodular Maximization

In this section we focus on smooth DR-submodular maximization in finite-sum form as described in (1), i.e.,

$$\max_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) := \frac{1}{N}\sum_{t=1}^{N} \mathbf{f}(\mathbf{x}, \xi_t), \tag{10}$$

where $\mathbf{f}(\cdot, \xi_t), t \in [N]$ are further required L-smooth.

Our purpose in this section is to design efficient zeroth-order gradient approximations that can be incorporated into Algorithm 1 for solving (10), and to present the corresponding approximation analysis. Specifically, we will adopt coordinate-wise gradient estimation and compute approximate gradients based on an integral auxiliary function in Subsection 3.1. Furthermore, to further improve the approximation performance, we will propose a

novel approach to approximate gradients based on randomized gradient estimation and the integral auxiliary function of a smoothed function in Subsection 3.2.

3.1 CG-ZOSA based on Coordinate-wise Gradient Estimation

Motivated by (Zhang et al., 2022), we assume that $\nabla \mathbf{f}(\theta \mathbf{x}, \xi_t)$ is Lebesgue integrable with respect to $\theta \in (0, 1]$. Define the auxiliary function $F : \mathbb{R}^d \to \mathbb{R}$ by

$$F(\mathbf{x}) := \frac{1}{N} \sum_{t=1}^{N} \mathbf{F}(\mathbf{x}, \xi_t), \quad \text{where} \quad \mathbf{F}(\mathbf{x}, \xi_t) := \int_0^1 \frac{e^{\theta-1}}{\theta} \mathbf{f}(\theta \mathbf{x}, \xi_t) d\theta, \quad \forall t \in [N]. \tag{11}$$

It is straightforward to obtain

$$F(\mathbf{x}) = \int_0^1 \frac{e^{\theta-1}}{\theta} f(\theta \mathbf{x}) d\theta,$$

which was first proposed in (Zhang et al., 2022) for weakly DR-submodular maximization as a representation of non-oblivious function. It is obvious that F is non-negative according to the non-negativeness of f, and the lemma below summarizes more properties of F. Before presenting the properties of F, we define the stationary point for a constrained maximization problem.

Definition 7. (**Stationary point**) *A vector $\mathbf{x} \in \mathcal{X}$ is called a stationary point of function $f : \mathbb{R}^d \to \mathbb{R}$ over the compact set $\mathcal{X} \subseteq \mathbb{R}^d$, if*

$$\max_{\mathbf{x} \in \mathcal{X}} \quad \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \leq 0.$$

It is worthy to note that properties indicated by Lemma 8 (ii) and (iii) relies on the DR-submodularity of f. The proof is given in Appendix C.

Lemma 8. *Let $\mathbf{x}^*$ be the optimal solution of (10) and $F(\mathbf{x})$ be defined by (11). The following statements hold true.*

(i) The gradient of $F(\mathbf{x})$ is given by $\nabla F(\mathbf{x}) = \int_0^1 e^{\theta-1} \nabla f(\theta \mathbf{x}) d\theta$, and $F(\mathbf{x})$ is $\frac{L}{e}$-smooth.

(ii) For any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, it holds that $\langle \mathbf{y} - \mathbf{x}, \nabla F(\mathbf{x}) \rangle \geq \left(1 - e^{-1}\right) f(\mathbf{y}) - f(\mathbf{x})$. Furthermore, $f(\mathbf{x}) \geq \left(1 - e^{-1}\right) f(\mathbf{x}^*)$, where $\mathbf{x}$ is a stationary point of $\max_{\mathbf{x} \in \mathcal{X}} F(\mathbf{x})$.

(iii) (Zhang et al., 2022, Theorem 2) For any $\mathbf{x} \in \mathcal{X}$, it holds that $F(\mathbf{x}) \leq (1 + \ln \tau)(f(\mathbf{x}) + \delta)$, where $\tau = \frac{LR^2}{\delta}$ and $\delta \in (0, LR^2]$.

Based on the integral auxiliary function defined above and the zeroth-order stochastic approximation algorithm framework described in Algorithm 1, we adopt coordinate-wise gradient estimation to compute an approximate gradient $\mathbf{d}_j^{s+1}$ for solving (10) as follows. Let random variable Θ follow the probability function

$$\mathrm{P}(\Theta \leq \theta) = \frac{\int_0^\theta e^{u-1} I_{[0,1]}(u) du}{\int_0^1 e^{u-1} du}, \tag{12}$$

where $I_{[0,1]}(u)$ is the indicator function on $[0,1]$. Let $\theta \in (0,1]$ be an i.i.d. sample from random variable Θ and $\mathcal{B}$ be a batch randomly selected from $\mathcal{N} := \{\xi_t\}_{t=1}^N$ with $|\mathcal{B}| = b$. Set

$$\mathbf{G}_{\text{coord}}(\mathbf{x}, \theta, \mathcal{B}) := (1 - e^{-1}) \cdot \frac{1}{b} \sum_{\xi_t \in \mathcal{B}} \mathbf{g}_{\text{coord}}(\theta \mathbf{x}, \xi_t), \tag{13}$$

where $\mathbf{g}_{\text{coord}}(\theta \mathbf{x}, \xi_t)$ is defined as (**RandGradEst**). Obviously, we have

$$\mathbb{E}_{\mathcal{B}}[\mathbf{G}_{\text{coord}}(\mathbf{x}, \theta, \mathcal{B})] = \mathbf{G}_{\text{coord}}(\mathbf{x}, \theta, \mathcal{N}) = \left(1 - e^{-1}\right) g_{\text{coord}}(\theta \mathbf{x}), \tag{14}$$

where

$$g_{\text{coord}}(\theta \mathbf{x}) := \frac{1}{N} \sum_{\xi_t \in \mathcal{N}} \mathbf{g}_{\text{coord}}(\theta \mathbf{x}, \xi_t).$$

It is worthy to note that $g_{\text{coord}}(\mathbf{x})$ is an estimate of $\nabla f(\mathbf{x})$. Moreover, by Lemma 3 and (14), we respectively have

$$\|g_{\text{coord}}(\mathbf{x}) - \nabla f(\mathbf{x})\|^2 \leq L^2 d u^2, \tag{15}$$

and

$$\begin{aligned}
\mathbb{E}[\mathbf{G}_{\text{coord}}(\mathbf{x}, \theta, \mathcal{B})|\mathbf{x}] = \mathbb{E}_{\theta \sim \Theta}[\mathbf{G}_{\text{coord}}(\mathbf{x}, \theta, \mathcal{N})|\mathbf{x}] &= \left(1 - e^{-1}\right) \mathbb{E}_{\theta \sim \Theta}[g_{\text{coord}}(\theta \mathbf{x})|\mathbf{x}] \\
&= \left(1 - e^{-1}\right) \int_{\theta=0}^{1} g_{\text{coord}}(\theta \mathbf{x}) d\mathrm{P}(\Theta \leq \theta) \\
&= \left(1 - e^{-1}\right) \int_0^1 \frac{e^{\theta - 1}}{\int_0^1 e^{u-1} du} g_{\text{coord}}(\theta \mathbf{x}) d\theta \\
&= \int_0^1 e^{\theta - 1} g_{\text{coord}}(\theta \mathbf{x}) d\theta =: \tilde{\nabla} F(\mathbf{x}).
\end{aligned} \tag{16}$$

Although, according to Lemma 8(i), $\mathbf{G}_{\text{coord}}(\mathbf{x}, \theta, \mathcal{B})$ is not an unbiased estimate of the exact gradient $\nabla F(\mathbf{x})$, it still provides a reliable approximation. We thus define the zeroth-order approximate gradient $\mathbf{d}_j^{s+1}$ at point $\mathbf{x}_j^{s+1}$, $j = 0, \ldots, m-1; s = 0, \ldots, S-1$, as

$$\mathbf{d}_j^{s+1} = \begin{cases} \mathbf{G}_{\text{coord}}(\mathbf{x}_0^{s+1}, \theta^{s+1}, \mathcal{N}), & j = 0 \\ \mathbf{G}_{\text{coord}}(\mathbf{x}_j^{s+1}, \theta^{s+1}, \mathcal{B}_j^{s+1}) - \mathbf{G}_{\text{coord}}(\mathbf{x}_0^{s+1}, \theta^{s+1}, \mathcal{B}_j^{s+1}) + \mathbf{G}_{\text{coord}}(\mathbf{x}_0^{s+1}, \theta^{s+1}, \mathcal{N}), & j > 0 \end{cases}, \tag{17}$$

where $\mathcal{B}_j^{s+1} \subseteq \mathcal{N}$ contains i.i.d. samples with $|\mathcal{B}_j^{s+1}| = b$, and $\theta^{s+1} \in (0,1]$ is an i.i.d. sample from the random variable Θ. For simplicity, we use **CG-ZOSA** to name the resulting algorithm that incorporates (17) to compute $\mathbf{d}_j^{s+1}$ in Algorithm 1.

The lemma below provides properties of the approximate gradient $\mathbf{d}_j^{s+1}$.

Lemma 9. *Let $\mathbf{d}_j^{s+1}, j = 0, \ldots, m-1, s = 0, \ldots S-1$ be computed through (17). Then it holds that*

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{d}_j^{s+1} | \mathbf{x}_j^{s+1}\right] = \tilde{\nabla} F(\mathbf{x}_j^{s+1}), \\
&\mathbb{E}\left[\|\nabla F(\mathbf{x}_j^{s+1}) - \mathbf{d}_j^{s+1}\|^2 | \mathbf{x}_j^{s+1}\right] \leq \frac{3\left(1 - e^{-1}\right)\left(1 - 2e^{-1}\right) L^2}{b} \|\mathbf{x}_j^{s+1} - \mathbf{x}_0^{s+1}\|^2 + Q,
\end{aligned}$$

where $Q = 6\left(1-e^{-1}\right)^2 L^2 du^2/b + 2\left(1-e^{-1}\right)^2 L^2 du^2 + 2L^2R^2$ *and* $\tilde{\nabla} F(\mathbf{x})$ *is defined as in* (16).

Proof Firstly, from $\mathbb{E}_{\mathcal{B}}\left[\mathbf{G}_{\text{coord}}(\mathbf{x},\theta,\mathcal{B})\right] = \mathbf{G}_{\text{coord}}(\mathbf{x},\theta,\mathcal{N})$ and (16), it is easy to obtain

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{d}_j^{s+1}|\mathbf{x}_j^{s+1}\right] \\
&= \mathbb{E}\left[\mathbf{G}_{\text{coord}}(\mathbf{x}_j^{s+1},\theta^{s+1},\mathcal{B}_j^{s+1}) - \mathbf{G}_{\text{coord}}(\mathbf{x}_0^{s+1},\theta^{s+1},\mathcal{B}_j^{s+1}) + \mathbf{G}_{\text{coord}}(\mathbf{x}_0^{s+1},\theta^{s+1},\mathcal{N})|\mathbf{x}_j^{s+1}\right] \\
&= \mathbb{E}\left[\mathbf{G}_{\text{coord}}(\mathbf{x}_j^{s+1},\theta^{s+1},\mathcal{B}_j^{s+1})|\mathbf{x}_j^{s+1}\right] = \tilde{\nabla} F(\mathbf{x}_j^{s+1}).
\end{aligned}$$

Secondly, from the L-smoothness of the function $\mathbf{f}(\mathbf{x},\xi_t)$ and Lemma 3, we obtain that for any $t \in [N]$,

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbf{g}_{\text{coord}}(\theta\mathbf{x},\xi_t) - \mathbf{g}_{\text{coord}}(\theta\mathbf{y},\xi_t)\|^2\right] \\
&\le 3\left(\mathbb{E}\left[\|\mathbf{g}_{\text{coord}}(\theta\mathbf{x},\xi_t) - \nabla\mathbf{f}(\theta\mathbf{x},\xi_t)\|^2 + \|\nabla\mathbf{f}(\theta\mathbf{x},\xi_t) - \nabla\mathbf{f}(\theta\mathbf{y},\xi_t)\|^2\right]\right. \\
&\quad \left. + \|\mathbf{g}_{\text{coord}}(\theta\mathbf{y},\xi_t) - \nabla\mathbf{f}(\theta\mathbf{y},\xi_t)\|^2\right]\right) \\
&\le 3\left(L^2\mathbb{E}_{\theta\sim\Theta}\left[\theta^2\|\mathbf{x}-\mathbf{y}\|^2\right] + 2L^2du^2\right) \\
&= 3\left(L^2\|\mathbf{x}-\mathbf{y}\|^2\int_0^1 \frac{e^{\theta-1}}{\int_0^1 e^{u-1}du}\theta^2 d\theta + 2L^2du^2\right) \\
&= 3\frac{\left(1-2e^{-1}\right)L^2}{\left(1-e^{-1}\right)}\|\mathbf{x}-\mathbf{y}\|^2 + 6L^2du^2.
\end{aligned} \tag{18}$$

Furthermore, it follows from the definition of $\mathbf{G}_{\text{coord}}(\mathbf{x},\theta,\mathcal{N})$ in (13) and the inequality $\|g_{\text{coord}}(\mathbf{x}) - \nabla f(\mathbf{x})\|^2 \le L^2du^2$ in (15) that

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbf{G}_{\text{coord}}(\mathbf{x},\theta,\mathcal{N}) - \nabla F(\mathbf{x})\|^2|\mathbf{x}\right] \\
&\le 2\mathbb{E}\left[\|\mathbf{G}_{\text{coord}}(\mathbf{x},\theta,\mathcal{N}) - \left(1-e^{-1}\right)\nabla f(\theta\mathbf{x})\|^2 + \|\left(1-e^{-1}\right)\nabla f(\theta\mathbf{x}) - \nabla F(\mathbf{x})\|^2|\mathbf{x}\right] \\
&= 2\mathbb{E}\left[\left(1-e^{-1}\right)^2\|g_{\text{coord}}(\theta\mathbf{x}) - \nabla f(\theta\mathbf{x})\|^2 + \|\int_0^1 e^{t-1}(\nabla f(\theta\mathbf{x}) - \nabla f(t\mathbf{x}))dt\|^2|\mathbf{x}\right] \\
&\le 2\left(1-e^{-1}\right)^2 L^2du^2 + 2\mathbb{E}_{\theta\sim\Theta}\left[\left\|\int_0^1 e^{t-1}|\theta-t|L\|\mathbf{x}\|dt\right\|^2|\mathbf{x}\right] \\
&\le 2\left(1-e^{-1}\right)^2 L^2du^2 + 2\mathbb{E}_{\theta\sim\Theta}\left[\int_0^1 e^{t-1}dt\int_0^1 e^{t-1}(\theta-t)^2L^2R^2dt|\mathbf{x}\right] \\
&= 2\left(1-e^{-1}\right)^2 L^2du^2 + 2\left(\int_0^1\int_0^1 e^{(\theta+t-2)}(\theta-t)^2dtd\theta\right)L^2R^2 \\
&\le 2\left(1-e^{-1}\right)^2 L^2du^2 + 2L^2R^2,
\end{aligned} \tag{19}$$

where the second inequality is due to the L-smoothness of f, the third one follows from the Cauchy–Schwarz inequality and the boundedness of $\mathbf{x}$ as well as

$$\left(\int a\cdot b dt\right)^2 \le \int a dt\cdot\int ab^2dt, \quad \forall a,b\ge 0,$$

the second equality is implied by $\int_0^1 e^{t-1}dt = 1 - \frac{1}{e}$ and (12), i.e.,

$$\begin{aligned}
&\mathbb{E}_{\theta\sim\Theta}\left[\int_0^1 e^{t-1}dt\int_0^1 e^{t-1}(\theta-t)^2L^2R^2dt|\mathbf{x}\right]\\
&=\int_{\theta=0}^1\frac{e^{\theta-1}}{1-\frac{1}{e}}\left(\int_{t=0}^1 e^{t-1}dt\int_{t=0}^1 e^{t-1}(\theta-t)^2dt\right)d\theta\cdot L^2R^2\\
&=\int_{\theta=0}^1 e^{\theta-1}\left(\int_{t=0}^1 e^{t-1}(\theta-t)^2dt\right)d\theta\cdot L^2R^2\\
&=\int_{\theta=0}^1\int_{t=0}^1 e^{\theta+t-2}(\theta-t)^2dtd\theta\cdot L^2R^2,
\end{aligned}$$

and the last inequality is derived from the fact that $\int_{\theta=0}^1\int_{t=0}^1 e^{(\theta+t-2)}(\theta-t)^2dtd\theta \leq 1$. Therefore, combining (18) and (19) we obtain

$$\begin{aligned}
&\mathbb{E}\left[\|\nabla F(\mathbf{x}_j^{s+1})-\mathbf{d}_j^{s+1}\|^2|\mathbf{x}_j^{s+1}\right]\\
&=\mathbb{E}\left[\|\nabla F(\mathbf{x}_j^{s+1})-\Big(\mathbf{G}_{\text{coord}}(\mathbf{x}_j^{s+1},\theta^{s+1},\mathcal{B}_j^{s+1})-\mathbf{G}_{\text{coord}}(\mathbf{x}_0^{s+1},\theta^{s+1},\mathcal{B}_j^{s+1})\right.\\
&\quad\left.+\mathbf{G}_{\text{coord}}(\mathbf{x}_0^{s+1},\theta^{s+1},\mathcal{N})\Big)\|^2|\mathbf{x}_j^{s+1}\right]\\
&=\mathbb{E}_{\theta^{s+1}}\left[\mathbb{E}_{\mathcal{B}_j^{s+1}}\left[\|\nabla F(\mathbf{x}_j^{s+1})-\mathbf{G}_{\text{coord}}(\mathbf{x}_j^{s+1},\theta^{s+1},\mathcal{N})+\mathbf{G}_{\text{coord}}(\mathbf{x}_j^{s+1},\theta^{s+1},\mathcal{N})\right.\right.\\
&\quad\left.\left.-\mathbf{G}_{\text{coord}}(\mathbf{x}_j^{s+1},\theta^{s+1},\mathcal{B}_j^{s+1})+\mathbf{G}_{\text{coord}}(\mathbf{x}_0^{s+1},\theta^{s+1},\mathcal{B}_j^{s+1})-\mathbf{G}_{\text{coord}}(\mathbf{x}_0^{s+1},\theta^{s+1},\mathcal{N})\|^2|\mathbf{x}_j^{s+1}\right]|\mathbf{x}_j^{s+1}\right]\\
&=\mathbb{E}_{\theta^{s+1}}\left[\mathbb{E}_{\mathcal{B}_j^{s+1}}\left[\|\mathbf{G}_{\text{coord}}(\mathbf{x}_j^{s+1},\theta^{s+1},\mathcal{B}_j^{s+1})-\mathbf{G}_{\text{coord}}(\mathbf{x}_0^{s+1},\theta^{s+1},\mathcal{B}_j^{s+1})+\mathbf{G}_{\text{coord}}(\mathbf{x}_0^{s+1},\theta^{s+1},\mathcal{N})\right.\right.\\
&\quad\left.\left.-\mathbf{G}_{\text{coord}}(\mathbf{x}_j^{s+1},\theta^{s+1},\mathcal{N})\|^2|\mathbf{x}_j^{s+1}\right]|\mathbf{x}_j^{s+1}\right]+\mathbb{E}[\|\nabla F(\mathbf{x}_j^{s+1})-\mathbf{G}_{\text{coord}}(\mathbf{x}_j^{s+1},\theta^{s+1},\mathcal{N})\|^2|\mathbf{x}_j^{s+1}]\\
&=\mathbb{E}\left[\|\mathbf{G}_{\text{coord}}(\mathbf{x}_j^{s+1},\theta^{s+1},\mathcal{B}_j^{s+1})-\mathbf{G}_{\text{coord}}(\mathbf{x}_0^{s+1},\theta^{s+1},\mathcal{B}_j^{s+1})+\mathbf{G}_{\text{coord}}(\mathbf{x}_0^{s+1},\theta^{s+1},\mathcal{N})\right.\\
&\quad\left.-\mathbf{G}_{\text{coord}}(\mathbf{x}_j^{s+1},\theta^{s+1},\mathcal{N})\|^2|\mathbf{x}_j^{s+1}\right]+\mathbb{E}[\|\nabla F(\mathbf{x}_j^{s+1})-\mathbf{G}_{\text{coord}}(\mathbf{x}_j^{s+1},\theta^{s+1},\mathcal{N})\|^2|\mathbf{x}_j^{s+1}]\\
&\leq\mathbb{E}\left[\|\mathbf{G}_{\text{coord}}(\mathbf{x}_j^{s+1},\theta^{s+1},\mathcal{B}_j^{s+1})-\mathbf{G}_{\text{coord}}(\mathbf{x}_0^{s+1},\theta^{s+1},\mathcal{B}_j^{s+1})\|^2|\mathbf{x}_j^{s+1}\right]\\
&\quad+\mathbb{E}[\|\nabla F(\mathbf{x}_j^{s+1})-\mathbf{G}_{\text{coord}}(\mathbf{x}_j^{s+1},\theta^{s+1},\mathcal{N})\|^2|\mathbf{x}_j^{s+1}]\\
&\leq\frac{\left(1-e^{-1}\right)^2}{b}\mathbb{E}\left[\|\mathbf{g}_{\text{coord}}(\theta^{s+1}\mathbf{x}_j^{s+1},\xi_t)-\mathbf{g}_{\text{coord}}(\theta^{s+1}\mathbf{x}_0^{s+1},\xi_t)\|^2|\mathbf{x}_j^{s+1}\right]\\
&\quad+\mathbb{E}\left[\|\nabla F(\mathbf{x}_j^{s+1})-\mathbf{G}_{\text{coord}}(\mathbf{x}_j^{s+1},\theta^{s+1},\mathcal{N})\|^2|\mathbf{x}_j^{s+1}\right]\\
&\leq\frac{3L^2\left(1-e^{-1}\right)\left(1-2e^{-1}\right)}{b}\|\mathbf{x}_j^{s+1}-\mathbf{x}_0^{s+1}\|^2+\frac{6\left(1-e^{-1}\right)^2L^2du^2}{b}\\
&\quad+2\left(1-e^{-1}\right)^2L^2du^2+2L^2R^2.
\end{aligned}$$

where the third equality is due to

$$\mathbb{E}_{\mathcal{B}_j^{s+1}}\left[\mathbf{G}_{\text{coord}}(\mathbf{x}_j^{s+1},\theta^{s+1},\mathcal{B}_j^{s+1})-\mathbf{G}_{\text{coord}}(\mathbf{x}_0^{s+1},\theta^{s+1},\mathcal{B}_j^{s+1})+\mathbf{G}_{\text{coord}}(\mathbf{x}_0^{s+1},\theta^{s+1},\mathcal{N})\right.$$

$$-\mathbf{G}_{\text{coord}}(\mathbf{x}_j^{s+1}, \theta^{s+1}, \mathcal{N})\Big] = 0,$$

the first inequality arises from the facts that $\mathbb{E}_{\mathcal{B}}\left[\mathbf{G}_{\text{coord}}(\mathbf{x}, \theta, \mathcal{B})\right] = \mathbf{G}_{\text{coord}}(\mathbf{x}, \theta, \mathcal{N})$ and $\mathbb{E}[\|\xi - \mathbb{E}[\xi]\|^2] \leq \mathbb{E}[\|\xi\|^2]$ for the random variable ξ, and the second inequality follows from the definition of $\mathbf{G}_{\text{coord}}(\mathbf{x}, \theta, \mathcal{B})$ as given in (13). The proof is completed. ■

The next lemma characterizes the relation between the auxiliary function values at two consecutive iteration points.

Lemma 10. *Let* $\mathbf{x}_j^{s+1}, j = 0, \ldots, m-1,\ s = 0, \ldots S-1$ *be generated by* **CG-ZOSA**. Then it holds that for any $\mathbf{y} \in \mathcal{X}$,

$$\begin{aligned} F(\mathbf{x}_{j+1}^{s+1}) \geq F(\mathbf{x}_j^{s+1}) + \frac{1}{2\eta_j^{s+1}}\left(\|\mathbf{x}_{j+1}^{s+1} - \mathbf{y}\|^2 - \|\mathbf{x}_j^{s+1} - \mathbf{y}\|^2\right) + \langle \mathbf{d}_j^{s+1}, \mathbf{y} - \mathbf{x}_j^{s+1}\rangle \\ - \eta_j^{s+1}\|\nabla F(\mathbf{x}_j^{s+1}) - \mathbf{d}_j^{s+1}\|^2 - \left(\frac{L}{2e} - \frac{1}{4\eta_j^{s+1}}\right)\|\mathbf{x}_{j+1}^{s+1} - \mathbf{x}_j^{s+1}\|^2. \end{aligned} \tag{20}$$

Proof It follows from the $\frac{L}{e}$-smoothness of $F(\mathbf{x})$ that for any $\mathbf{y} \in \mathcal{X}$,

$$\begin{aligned} F(\mathbf{x}_{j+1}^{s+1}) &\geq F(\mathbf{x}_j^{s+1}) + \langle \nabla F(\mathbf{x}_j^{s+1}), \mathbf{x}_{j+1}^{s+1} - \mathbf{x}_j^{s+1}\rangle - \frac{L}{2e}\|\mathbf{x}_{j+1}^{s+1} - \mathbf{x}_j^{s+1}\|^2 \\ &= F(\mathbf{x}_j^{s+1}) + \langle \mathbf{d}_j^{s+1}, \mathbf{x}_{j+1}^{s+1} - \mathbf{x}_j^{s+1}\rangle + \langle \nabla F(\mathbf{x}_j^{s+1}) - \mathbf{d}_j^{s+1}, \mathbf{x}_{j+1}^{s+1} - \mathbf{x}_j^{s+1}\rangle \\ &\quad - \frac{L}{2e}\|\mathbf{x}_{j+1}^{s+1} - \mathbf{x}_j^{s+1}\|^2 \\ &= F(\mathbf{x}_j^{s+1}) + \langle \mathbf{d}_j^{s+1}, \mathbf{x}_{j+1}^{s+1} - \mathbf{y}\rangle + \langle \mathbf{d}_j^{s+1}, \mathbf{y} - \mathbf{x}_j^{s+1}\rangle \\ &\quad + \langle \nabla F(\mathbf{x}_j^{s+1}) - \mathbf{d}_j^{s+1}, \mathbf{x}_{j+1}^{s+1} - \mathbf{x}_j^{s+1}\rangle - \frac{L}{2e}\|\mathbf{x}_{j+1}^{s+1} - \mathbf{x}_j^{s+1}\|^2 \\ &= F(\mathbf{x}_j^{s+1}) + \frac{1}{\eta_j^{s+1}}\langle \mathbf{y}_{j+1}^{s+1} - \mathbf{x}_j^{s+1}, \mathbf{x}_{j+1}^{s+1} - \mathbf{y}\rangle + \langle \mathbf{d}_j^{s+1}, \mathbf{y} - \mathbf{x}_j^{s+1}\rangle \\ &\quad + \langle \nabla F(\mathbf{x}_j^{s+1}) - \mathbf{d}_j^{s+1}, \mathbf{x}_{j+1}^{s+1} - \mathbf{x}_j^{s+1}\rangle - \frac{L}{2e}\|\mathbf{x}_{j+1}^{s+1} - \mathbf{x}_j^{s+1}\|^2 \\ &\geq F(\mathbf{x}_j^{s+1}) + \frac{1}{2\eta_j^{s+1}}\left(\|\mathbf{x}_{j+1}^{s+1} - \mathbf{x}_j^{s+1}\|^2 - \|\mathbf{x}_j^{s+1} - \mathbf{y}\|^2 + \|\mathbf{x}_{j+1}^{s+1} - \mathbf{y}\|^2\right) \\ &\quad + \langle \mathbf{d}_j^{s+1}, \mathbf{y} - \mathbf{x}_j^{s+1}\rangle + \langle \nabla F(\mathbf{x}_j^{s+1}) - \mathbf{d}_j^{s+1}, \mathbf{x}_{j+1}^{s+1} - \mathbf{x}_j^{s+1}\rangle - \frac{L}{2e}\|\mathbf{x}_{j+1}^{s+1} - \mathbf{x}_j^{s+1}\|^2 \end{aligned} \tag{21}$$

The second inequality of (21) uses the property of the projection operator: $\langle \mathbf{y}_{j+1}^{s+1} - \mathbf{x}_{j+1}^{s+1}, \mathbf{x}_{j+1}^{s+1} - \mathbf{y}\rangle \geq 0$ and $2\langle a, b\rangle = \|a\|^2 + \|b\|^2 - \|a-b\|^2$, through which we obtain

$$\begin{aligned} \langle \mathbf{y}_{j+1}^{s+1} - \mathbf{x}_j^{s+1}, \mathbf{x}_{j+1}^{s+1} - \mathbf{y}\rangle &= \langle \mathbf{y}_{j+1}^{s+1} - \mathbf{x}_{j+1}^{s+1}, \mathbf{x}_{j+1}^{s+1} - \mathbf{y}\rangle + \langle \mathbf{x}_{j+1}^{s+1} - \mathbf{x}_j^{s+1}, \mathbf{x}_{j+1}^{s+1} - \mathbf{y}\rangle \\ &\geq \langle \mathbf{x}_{j+1}^{s+1} - \mathbf{x}_j^{s+1}, \mathbf{x}_{j+1}^{s+1} - \mathbf{y}\rangle \\ &= \frac{1}{2}\left(\|\mathbf{x}_{j+1}^{s+1} - \mathbf{x}_j^{s+1}\|^2 - \|\mathbf{y} - \mathbf{x}_j^{s+1}\|^2 + \|\mathbf{x}_{j+1}^{s+1} - \mathbf{y}\|^2\right). \end{aligned}$$

Hence, by plugging the inequality

$$\langle \nabla F(\mathbf{x}_j^{s+1}) - \mathbf{d}_j^{s+1}, \mathbf{x}_{j+1}^{s+1} - \mathbf{x}_j^{s+1} \rangle \geq -\eta_j^{s+1} \|\nabla F(\mathbf{x}_j^{s+1}) - \mathbf{d}_j^{s+1}\|^2 - \frac{1}{4\eta_j^{s+1}} \|\mathbf{x}_{j+1}^{s+1} - \mathbf{x}_j^{s+1}\|^2$$

into (21) we derive the conclusion. ■

For simplicity, we denote $L_F = 2\left(1 - e^{-1}\right) L$. Thus, it is obvious that

$$L_F \geq \max\left\{ \frac{L}{e}, \sqrt{3\left(1 - e^{-1}\right)\left(1 - 2e^{-1}\right)} L \right\}.$$

For a better presentation of the main theorem in this subsection, we introduce an important inequality presented in the following lemma.

Lemma 11. *Given the parameters* $a_m = 0$, $a_j = a_{j+1}\left(1 + \frac{1}{m}\right) + \frac{\eta_j^{s+1} L_F^2}{b}$ *and*

$$b = m^2 < N, \quad \eta_j^{s+1} = \frac{1}{c_j^{s+1} L_F},$$

where c_j^{s+1} *is non-decreasing and* $c_j^{s+1} \geq 4\sqrt{2}, j = 0, \ldots, m-1, s = 0, \ldots, S-1$, *we have*

$$\frac{L_F}{2} + a_{j+1}(1+m) \leq \frac{1}{4\eta_j^{s+1}}. \tag{22}$$

Proof By $a_j = a_{j+1}(1 + \frac{1}{m}) + \eta_j^{s+1} L_F^2 / b$, we obtain

$$\begin{aligned}
a_j &= \left(a_{j+2}(1 + \frac{1}{m}) + \frac{\eta_{j+1}^{s+1} L_F^2}{b} \right)(1 + \frac{1}{m}) + \frac{\eta_j^{s+1} L_F^2}{b} \\
&= a_{j+2}(1 + \frac{1}{m})^2 + \frac{L_F^2}{b}\left(\eta_{j+1}^{s+1}(1 + \frac{1}{m}) + \eta_j^{s+1} \right) \\
&= a_m(1 + \frac{1}{m})^{m-j} + \frac{L_F^2}{b}\left(\eta_{m-1}^{s+1}(1 + \frac{1}{m})^{m-j-1} + \ldots + \eta_{j+1}^{s+1}(1 + \frac{1}{m}) + \eta_j^{s+1} \right) \\
&\leq \frac{\eta_j^{s+1} L_F^2 m}{b}\left((1 + \frac{1}{m})^{m-j} - 1 \right) \\
&= \frac{m L_F}{b c_j^{s+1}}\left(\left(1 + \frac{1}{m}\right)^{m-j} - 1 \right) \\
&\leq \frac{m L_F}{b c_j^{s+1}}(e - 1) \\
&\leq \frac{2 m L_F}{b c_j^{s+1}},
\end{aligned}$$

where the first inequality comes from $a_m = 0$ and $\eta_{m-1}^{s+1} \leq \ldots \leq \eta_{j+1}^{s+1} \leq \eta_j^{s+1}$, the fourth equality is because of $\eta_j^{s+1} = \frac{1}{c_j^{s+1} L_F}$, the second inequality is due to $(1 + 1/m)^{m-j} \leq e$ and

the last one is from $e-1 \leq 2$. Combining $c_{j+1}^{s+1} \geq c_j^{s+1} \geq 4\sqrt{2}$ and $b = m^2$, we have

$$\begin{aligned}\frac{L_F}{2} + a_{j+1}(1+m) &\leq \frac{L_F}{2} + \frac{2mL_F}{bc_{j+1}^{s+1}}(1+m) \\ &\leq \frac{L_F}{2} + \frac{2mL_F}{bc_j^{s+1}}(1+m) \\ &\leq \frac{L_F}{2} + \frac{4m^2 L_F}{bc_j^{s+1}} \\ &\leq \frac{1}{4\eta_j^{s+1}},\end{aligned}$$

where the last inequality is because of $\frac{L_F}{2} \leq \frac{1}{8\sqrt{2}\eta_j^{s+1}}$ and then $\frac{4m^2 L_F}{bc_j^{s+1}} = 4L_F^2\eta_j^{s+1} \leq \frac{1}{8\eta_j^{s+1}}$. Hence, the inequality (22) holds true. ■

The auxiliary function and the DR-submodularity provides non-convex optimization with an approximation guarantee, ensuring that the objective value at any stationary point of the auxiliary function F exceeds at least $(1-1/e)\mathrm{OPT}$, as shown in Lemma 8. We will demonstrate the approximation performance, including the approximation ratio and complexity of the **CG-ZOSA** algorithm by introducing a class of Lyapunov functions.

Theorem 12. *Let the parameters for the computation of* $\mathbf{d}_j^{s+1}$ *and the step-size* η_j^{s+1} *in* **CG-ZOSA** satisfy

$$b = m^2 < N, \quad u = \frac{1}{\sqrt{Smd}}, \quad \eta_j^{s+1} = \frac{1}{c_j^{s+1} L_F},$$

where $c_j^{s+1} = 4\sqrt{2}\sqrt{s(m-1)+j+1}$, *for* $j = 0, \ldots, m-1, s = 0, \ldots, S-1$ *and* $\eta_{m-1}^0 = \eta_{m-1}^1$. *And return* $\mathbf{x}_r$ *according to that* $\mathrm{P}(\mathbf{x}_r = \mathbf{x}_j^{s+1}) = \frac{1}{Sm+1+\ln\tau}$ *for* $j = 0, \ldots, m-1; s = 0, \ldots, S-1$, *and* $\mathrm{P}(\mathbf{x}_r = \mathbf{x}_m^S) = \frac{1+\ln\tau}{Sm+1+\ln\tau}$. *Then we obtain*

$$\mathbb{E}\left[f(\mathbf{x}_r)\right] \geq \left(1 - e^{-1} - \mathcal{O}\left(\frac{1}{Sm}\right)\right) f(\mathbf{x}^*) - \mathcal{O}\left(\frac{1}{\sqrt{Sm}}\right), \tag{23}$$

where $\mathbf{x}^*$ *is the optimal solution of* (10).

Proof To prove the theorem, we introduce the following Lyapunov function

$$L_j^{s+1} := \mathbb{E}\left[F(\mathbf{x}_j^{s+1}) - a_j\|\mathbf{x}_j^{s+1} - \mathbf{x}_0^{s+1}\|^2\right], j = 0, \ldots, m-1, s = 0, \ldots, S-1,$$

where $a_m = 0$ and $a_j = a_{j+1}\left(1 + \frac{1}{m}\right) + \frac{\eta_j^{s+1} L_F^2}{b}$. We bound L_{j+1}^{s+1} as follows

$$\begin{aligned}&L_{j+1}^{s+1} \\ &= \mathbb{E}\left[F(\mathbf{x}_{j+1}^{s+1}) - a_{j+1}\|\mathbf{x}_{j+1}^{s+1} - \mathbf{x}_j^{s+1} + \mathbf{x}_j^{s+1} - \mathbf{x}_0^{s+1}\|^2\right]\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}\left[F(\mathbf{x}_{j+1}^{s+1}) - a_{j+1}\left(\|\mathbf{x}_{j+1}^{s+1} - \mathbf{x}_{j}^{s+1}\|^2 + \|\mathbf{x}_{j}^{s+1} - \mathbf{x}_{0}^{s+1}\|^2 + 2\langle \mathbf{x}_{j+1}^{s+1} - \mathbf{x}_{j}^{s+1}, \mathbf{x}_{j}^{s+1} - \mathbf{x}_{0}^{s+1}\rangle\right)\right] \\
&\geq \mathbb{E}\left[F(\mathbf{x}_{j+1}^{s+1}) - a_{j+1}(1+m)\|\mathbf{x}_{j+1}^{s+1} - \mathbf{x}_{j}^{s+1}\|^2 - a_{j+1}\left(1+\frac{1}{m}\right)\|\mathbf{x}_{j}^{s+1} - \mathbf{x}_{0}^{s+1}\|^2\right] \\
&\geq \mathbb{E}\left[F(\mathbf{x}_{j}^{s+1}) + \frac{1}{2\eta_j^{s+1}}\left(\|\mathbf{x}_{j+1}^{s+1} - \mathbf{y}\|^2 - \|\mathbf{x}_{j}^{s+1} - \mathbf{y}\|^2\right) + \langle \mathbf{d}_j^{s+1}, \mathbf{y} - \mathbf{x}_j^{s+1}\rangle\right. \\
&\quad -\eta_j^{s+1}\|\nabla F(\mathbf{x}_j^{s+1}) - \mathbf{d}_j^{s+1}\|^2 - \left(\frac{L_F}{2} - \frac{1}{4\eta_j^{s+1}}\right)\|\mathbf{x}_{j+1}^{s+1} - \mathbf{x}_j^{s+1}\|^2 \\
&\quad \left.-a_{j+1}(1+m)\|\mathbf{x}_{j+1}^{s+1} - \mathbf{x}_{j}^{s+1}\|^2 - a_{j+1}\left(1+\frac{1}{m}\right)\|\mathbf{x}_{j}^{s+1} - \mathbf{x}_{0}^{s+1}\|^2\right] \\
&\geq \mathbb{E}\left[F(\mathbf{x}_{j}^{s+1}) + \frac{1}{2\eta_j^{s+1}}\left(\|\mathbf{x}_{j+1}^{s+1} - \mathbf{y}\|^2 - \|\mathbf{x}_{j}^{s+1} - \mathbf{y}\|^2\right) + \langle \mathbf{d}_j^{s+1}, \mathbf{y} - \mathbf{x}_j^{s+1}\rangle\right. \\
&\quad -\frac{\eta_j^{s+1}L_F^2}{b}\|\mathbf{x}_j^{s+1} - \mathbf{x}_0^{s+1}\|^2 - \eta_j^{s+1}Q - \left(\frac{L_F}{2} - \frac{1}{4\eta_j^{s+1}}\right)\|\mathbf{x}_{j+1}^{s+1} - \mathbf{x}_j^{s+1}\|^2 \\
&\quad \left.-a_{j+1}(1+m)\|\mathbf{x}_{j+1}^{s+1} - \mathbf{x}_{j}^{s+1}\|^2 - a_{j+1}\left(1+\frac{1}{m}\right)\|\mathbf{x}_{j}^{s+1} - \mathbf{x}_{0}^{s+1}\|^2\right] \\
&= \mathbb{E}\left[F(\mathbf{x}_{j}^{s+1}) + \frac{1}{2\eta_j^{s+1}}\left(\|\mathbf{x}_{j+1}^{s+1} - \mathbf{y}\|^2 - \|\mathbf{x}_{j}^{s+1} - \mathbf{y}\|^2\right) + \langle \mathbf{d}_j^{s+1}, \mathbf{y} - \mathbf{x}_j^{s+1}\rangle - \eta_j^{s+1}Q\right. \\
&\quad - \left(\frac{L_F}{2} + a_{j+1}(1+m) - \frac{1}{4\eta_j^{s+1}}\right)\|\mathbf{x}_{j+1}^{s+1} - \mathbf{x}_j^{s+1}\|^2 \\
&\quad \left.- \left(a_{j+1}\left(1+\frac{1}{m}\right) + \frac{\eta_j^{s+1}L_F^2}{b}\right)\|\mathbf{x}_j^{s+1} - \mathbf{x}_0^{s+1}\|^2\right] \\
&= L_j^{s+1} + \frac{1}{2\eta_j^{s+1}}\mathbb{E}\left[\|\mathbf{x}_{j+1}^{s+1} - \mathbf{y}\|^2 - \|\mathbf{x}_{j}^{s+1} - \mathbf{y}\|^2\right] + \mathbb{E}\left[\langle \mathbf{d}_j^{s+1}, \mathbf{y} - \mathbf{x}_j^{s+1}\rangle\right] - \eta_j^{s+1}Q \\
&\quad - \left(\frac{L_F}{2} + a_{j+1}(1+m) - \frac{1}{4\eta_j^{s+1}}\right)\mathbb{E}\left[\|\mathbf{x}_{j+1}^{s+1} - \mathbf{x}_j^{s+1}\|^2\right] \\
&\geq L_j^{s+1} + \frac{1}{2\eta_j^{s+1}}\mathbb{E}\left[\|\mathbf{x}_{j+1}^{s+1} - \mathbf{y}\|^2 - \|\mathbf{x}_{j}^{s+1} - \mathbf{y}\|^2\right] + \mathbb{E}\left[\langle \mathbf{d}_j^{s+1}, \mathbf{y} - \mathbf{x}_j^{s+1}\rangle\right] - \eta_j^{s+1}Q, \qquad (24)
\end{aligned}
$$

where $Q = 6\left(1-e^{-1}\right)^2 L^2 du^2/b + 2\left(1-e^{-1}\right)^2 L^2 du^2 + 2L^2R^2$. The first inequality comes from the Cauchy-Schwarz and Young's inequality, i.e.,

$$
\begin{aligned}
2\langle \mathbf{x}_{j+1}^{s+1} - \mathbf{x}_j^{s+1}, \mathbf{x}_j^{s+1} - \mathbf{x}_0^{s+1}\rangle &\leq 2\|\mathbf{x}_{j+1}^{s+1} - \mathbf{x}_j^{s+1}\|\|\mathbf{x}_j^{s+1} - \mathbf{x}_0^{s+1}\| \\
&\leq m\|\mathbf{x}_{j+1}^{s+1} - \mathbf{x}_j^{s+1}\|^2 + \frac{1}{m}\|\mathbf{x}_j^{s+1} - \mathbf{x}_0^{s+1}\|^2,
\end{aligned}
$$

the second inequality follows from (20) in Lemma 10 and $L_F \geq \frac{L}{e}$, the third inequality is due to Lemma 9 and $L_F \geq \sqrt{3\left(1-e^{-1}\right)\left(1-2e^{-1}\right)}L$, the last equality is derived from the definition of Lyapunov function L_j^{s+1} and a_j, and the final inequality is indicated by (22) in Lemma 11.

Now, we set $\mathbf{y} = \mathbf{x}^*$ and telescope (24) over all the iterations in epoch $s+1$, obtaining

$$\begin{aligned}
L_m^{s+1} &\geq L_0^{s+1} + \sum_{j=0}^{m-1} \mathbb{E}\left[\frac{1}{2\eta_j^{s+1}}\left(\|\mathbf{x}_{j+1}^{s+1} - \mathbf{x}^*\|^2 - \|\mathbf{x}_j^{s+1} - \mathbf{x}^*\|^2\right)\right. \\
&\quad \left. + \mathbb{E}\left[\langle \mathbf{d}_j^{s+1}, \mathbf{x}^* - \mathbf{x}_j^{s+1}\rangle | \mathbf{x}_j^{s+1}\right] - \eta_j^{s+1} Q\right] \\
&= L_0^{s+1} + \mathbb{E}\left[\sum_{j=1}^{m-1}\left(\frac{1}{2\eta_{j-1}^{s+1}} - \frac{1}{2\eta_j^{s+1}}\right)\|\mathbf{x}_j^{s+1} - \mathbf{x}^*\|^2 + \frac{1}{2\eta_{m-1}^{s+1}}\|\mathbf{x}_m^{s+1} - \mathbf{x}^*\|^2\right. \\
&\quad \left. - \frac{1}{2\eta_0^{s+1}}\|\mathbf{x}_0^{s+1} - \mathbf{x}^*\|^2 + \sum_{j=0}^{m-1}\left(\mathbb{E}\left[\langle \mathbf{d}_j^{s+1}, \mathbf{x}^* - \mathbf{x}_j^{s+1}\rangle | \mathbf{x}_j^{s+1}\right] - \eta_j^{s+1} Q\right)\right] \\
&\geq L_0^{s+1} + \mathbb{E}\left[\left(\frac{1}{2\eta_0^{s+1}} - \frac{1}{2\eta_{m-1}^{s+1}}\right) D^2 + \frac{1}{2\eta_{m-1}^{s+1}}\|\mathbf{x}_m^{s+1} - \mathbf{x}^*\|^2 - \frac{1}{2\eta_0^{s+1}}\|\mathbf{x}_0^{s+1} - \mathbf{x}^*\|^2\right. \\
&\quad \left. + \sum_{j=0}^{m-1}\left(\mathbb{E}\left[\langle \mathbf{d}_j^{s+1}, \mathbf{x}^* - \mathbf{x}_j^{s+1}\rangle | \mathbf{x}_j^{s+1}\right] - \eta_j^{s+1} Q\right)\right], \qquad (25)
\end{aligned}$$

where the last inequality holds because $\eta_{j-1}^{s+1} \geq \eta_j^{s+1}$ and $\|\mathbf{x}_j^{s+1} - \mathbf{x}^*\| \leq D$. We now focus on the last term on the right-hand side of (25). By Lemma 9 and according to Lemma 8(ii) with $\mathbf{x} = \mathbf{x}_j^{s+1}$ and $\mathbf{y} = \mathbf{x}^*$, we obtain

$$\begin{aligned}
&\mathbb{E}[\langle \mathbf{d}_j^{s+1}, \mathbf{x}^* - \mathbf{x}_j^{s+1}\rangle | \mathbf{x}_j^{s+1}] \\
&= \langle \tilde{\nabla} F(\mathbf{x}_j^{s+1}), \mathbf{x}^* - \mathbf{x}_j^{s+1}\rangle \\
&= \langle \nabla F(\mathbf{x}_j^{s+1}), \mathbf{x}^* - \mathbf{x}_j^{s+1}\rangle + \langle \tilde{\nabla} F(\mathbf{x}_j^{s+1}) - \nabla F(\mathbf{x}_j^{s+1}), \mathbf{x}^* - \mathbf{x}_j^{s+1}\rangle \\
&\geq \left(1 - e^{-1}\right) f(\mathbf{x}^*) - f(\mathbf{x}_j^{s+1}) - \|\tilde{\nabla} F(\mathbf{x}_j^{s+1}) - \nabla F(\mathbf{x}_j^{s+1})\|\|\mathbf{x}^* - \mathbf{x}_j^{s+1}\| \\
&\geq \left(1 - e^{-1}\right) f(\mathbf{x}^*) - f(\mathbf{x}_j^{s+1}) - \left(1 - e^{-1}\right) L\sqrt{d}uD, \qquad (26)
\end{aligned}$$

where the last inequality holds because the diameter of constraint set $\mathcal{X}$ is D and $\|g_{\text{coord}}(\mathbf{x}) - \nabla f(\mathbf{x})\|^2 \leq L^2 d u^2$, as shown in (15), i.e.,

$$\begin{aligned}
\|\tilde{\nabla} F(\mathbf{x}) - \nabla F(\mathbf{x})\| \leq \int_0^1 e^{\theta - 1}\left(g_{\text{coord}}(\theta\mathbf{x}) - \nabla f(\theta\mathbf{x})\right)\|d\theta &\leq \int_0^1 \left(e^{\theta-1} L\sqrt{d}u\right) d\theta \\
&= \left(1 - e^{-1}\right) L\sqrt{d}u,
\end{aligned}$$

and $\|\mathbf{x}^* - \mathbf{x}_j^{s+1}\| \leq D$. Furthermore, we have $L_m^{s+1} = \mathbb{E}[F(\mathbf{x}_m^{s+1})]$ because of $a_m = 0$ and $L_0^{s+1} = \mathbb{E}[F(\mathbf{x}_0^{s+1})] = \mathbb{E}[F(\mathbf{x}_m^s)]$, which is due to the setting $\mathbf{x}_0^{s+1} = \mathbf{x}_m^s$. Combining $\eta_0^{s+1} = \eta_{m-1}^s$, we obtain the following inequality from (25) and (26):

$$\begin{aligned}
\mathbb{E}\left[F(\mathbf{x}_m^{s+1})\right] &\geq \mathbb{E}\left[F(\mathbf{x}_m^s)\right] + \left(\frac{1}{2\eta_{m-1}^s} - \frac{1}{2\eta_{m-1}^{s+1}}\right) D^2 \\
&\quad + \frac{1}{2\eta_{m-1}^{s+1}}\mathbb{E}\left[\|\mathbf{x}_m^{s+1} - \mathbf{x}^*\|^2\right] - \frac{1}{2\eta_{m-1}^s}\mathbb{E}\left[\|\mathbf{x}_m^s - \mathbf{x}^*\|^2\right]
\end{aligned}$$

$$+\sum_{j=0}^{m-1}\mathbb{E}\left[\left(1-e^{-1}\right)f(\mathbf{x}^*)-f(\mathbf{x}_j^{s+1})-\left(1-e^{-1}\right)L\sqrt{d}uD-\eta_j^{s+1}Q\right]. \quad (27)$$

Adding the inequality (27) across all the epochs (for $s\in\{0,\ldots,S-1\}$), we have

$$\begin{aligned}\mathbb{E}\left[F(\mathbf{x}_m^S)\right]&\geq\mathbb{E}\left[F(\mathbf{x}_m^0)\right]+\left(\frac{1}{2\eta_{m-1}^0}-\frac{1}{2\eta_{m-1}^S}\right)D^2+\frac{1}{2\eta_{m-1}^S}\mathbb{E}\left[\|\mathbf{x}_m^S-\mathbf{x}^*\|^2\right]\\&\quad-\frac{1}{2\eta_{m-1}^0}\mathbb{E}\left[\|\mathbf{x}_m^0-\mathbf{x}^*\|^2\right]+\sum_{s=0}^{S-1}\sum_{j=0}^{m-1}\mathbb{E}\left[\left(1-e^{-1}\right)f(\mathbf{x}^*)-f(\mathbf{x}_j^{s+1})\right]\\&\quad-\sum_{s=0}^{S-1}\sum_{j=0}^{m-1}\left(\left(1-e^{-1}\right)L\sqrt{d}uD+\eta_j^{s+1}Q\right)\\&\geq-\frac{c_{m-1}^SL_F}{2}D^2+\sum_{s=0}^{S-1}\sum_{j=0}^{m-1}\mathbb{E}\left[\left(1-e^{-1}\right)f(\mathbf{x}^*)-f(\mathbf{x}_j^{s+1})\right]\\&\quad-\sum_{s=0}^{S-1}\sum_{j=0}^{m-1}\left(\left(1-e^{-1}\right)L\sqrt{d}uD+\eta_j^{s+1}Q\right),\end{aligned}$$

where $\eta_{m-1}^0=\eta_{m-1}^1$, and the second inequality follows from the non-negativeness of $F(\mathbf{x})$, $\|\mathbf{x}_m^S-\mathbf{x}^*\|^2$, $\|\mathbf{x}_m^0-\mathbf{x}^*\|^2\leq D^2$ and $\eta_{m-1}^S=\frac{1}{c_{m-1}^SL_F}$. Rearranging the above inequality, we derive

$$\begin{aligned}&\mathbb{E}\left[F(\mathbf{x}_m^S)+\sum_{s=0}^{S-1}\sum_{j=0}^{m-1}f(\mathbf{x}_j^{s+1})\right]\\&\geq Sm\left(1-e^{-1}\right)f(\mathbf{x}^*)-\frac{c_{m-1}^SL_F}{2}D^2-\sum_{s=0}^{S-1}\sum_{j=0}^{m-1}\left(\left(1-e^{-1}\right)L\sqrt{d}uD+\eta_j^{s+1}Q\right). \quad (28)\end{aligned}$$

According to Lemma 8(iii), we have

$$\begin{aligned}&\mathbb{E}\left[\sum_{s=0}^{S-1}\sum_{j=0}^{m-1}f(\mathbf{x}_j^{s+1})+(1+\ln\tau)\left(f(\mathbf{x}_m^S)+\delta\right)\right]\\&\geq Sm\left(1-e^{-1}\right)f(\mathbf{x}^*)-\frac{c_{m-1}^SL_F}{2}D^2-\sum_{s=0}^{S-1}\sum_{j=0}^{m-1}\left(\left(1-e^{-1}\right)L\sqrt{d}uD+\eta_j^{s+1}Q\right),\end{aligned}$$

where $\tau=\frac{LR^2}{\delta}$, $\delta\in(0,LR^2]$. Dividing both sides by $(Sm+1+\ln\tau)$ and noting $u=1/\sqrt{dSm}$ as well as the generation of $\mathbf{x}_r$, we obtain

$$\mathbb{E}\left[f(\mathbf{x}_r)\right]=\mathbb{E}\left[\sum_{s=0}^{S-1}\sum_{j=0}^{m-1}\frac{1}{Sm+1+\ln\tau}f(\mathbf{x}_j^{s+1})+\frac{1+\ln\tau}{Sm+1+\ln\tau}f(\mathbf{x}_m^S)\right]$$

$$
\begin{aligned}
&\geq \frac{Sm}{Sm+1+\ln\tau}\left(1-e^{-1}\right)f(\mathbf{x}^*) \\
&\quad - \frac{\frac{c_{m-1}^S L_F}{2}D^2 + (1+\ln\tau)\delta + \sum_{s=0}^{S-1}\sum_{j=0}^{m-1}\left(\left(1-e^{-1}\right)L\sqrt{d}uD + \eta_j^{s+1}Q\right)}{Sm+1+\ln\tau} \\
&= \left(1-e^{-1}-\frac{\left(1-e^{-1}\right)\ln\tau}{Sm+1+\ln\tau}\right)f(\mathbf{x}^*) \\
&\quad - \mathcal{O}\left(\frac{c_{m-1}^S}{Sm}+\frac{1}{Sm}+\frac{\sum_{s=0}^{S-1}\sum_{j=0}^{m-1}\left(\sqrt{d}u+\frac{1}{c_j^{s+1}}(du^2+1)\right)}{Sm}\right), \\
&= \left(1-e^{-1}-\frac{\left(1-e^{-1}\right)\ln\tau}{Sm+1+\ln\tau}\right)f(\mathbf{x}^*) - \mathcal{O}\left(\frac{c_{m-1}^S}{Sm}+\frac{\sum_{s=0}^{S-1}\sum_{j=0}^{m-1}\frac{1}{c_j^{s+1}}}{Sm}\right),
\end{aligned}
$$

where $Q = 6\left(1-e^{-1}\right)^2 L^2 du^2/b + 2\left(1-e^{-1}\right)^2 L^2 du^2 + 2L^2R^2$ from Lemma 9 and $\eta_j^{s+1} = \frac{1}{c_j^{s+1}L_F}$. Since $c_j^{s+1} = 4\sqrt{2}\sqrt{s(m-1)+j+1}$, where $j = 0,\ldots,m-1$, it follows that $c_{m-1}^S = 4\sqrt{2}\sqrt{Sm-(S-1)} \leq 4\sqrt{2}\sqrt{Sm}$, and

$$
\begin{aligned}
\sum_{s=0}^{S-1}\sum_{j=0}^{m-1}\frac{1}{c_j^{s+1}} &= \sum_{s=0}^{S-1}\sum_{j=0}^{m-1}\frac{1}{4\sqrt{2}\sqrt{(s-1)m+j+1}} \\
&\leq \sum_{s=0}^{S-1}\sum_{j=0}^{m-1}\frac{1}{2\sqrt{2}\sqrt{sm+j+1}} \\
&\leq \frac{\sqrt{Sm}}{\sqrt{2}}.
\end{aligned}
$$

Therefore, (23) is proved. ■

Corollary 13. *Under the same setting as Theorem 12 with additional condition $b = N^{2/3}$ and given $\epsilon > 0$, the output $\mathbf{x}_r$ of* **CG-ZOSA** satisfies

$$\mathbb{E}\left[f(\mathbf{x}_r)\right] \geq \left(1-e^{-1}-\epsilon^2\right)f(\mathbf{x}^*) - \epsilon,$$

after $\mathcal{O}(\epsilon^{-2})$ total iterations and $\mathcal{O}(dN^{2/3}\epsilon^{-2})$ zeroth-order oracle evaluations.

Proof Note that the number of function value evaluations in the inner iterations and outer iterations are $Sm \times 4b \times d$ and $S \times 2N \times d$, respectively. Thus, by setting $Sm = \mathcal{O}(\epsilon^{-2})$, $b = N^\beta$ with $\beta > 0$, and $b = m^2$, the total number of function value evaluations is in order of

$$\mathcal{O}(\frac{dN^\beta}{\epsilon^2} + \frac{dN^{1-(\beta/2)}}{\epsilon^2}).$$

It is straightforward to verify that when $\beta = 2/3$, the total complexity is optimal, i.e., $\mathcal{O}(dN^{2/3}\epsilon^{-2})$. ■

3.2 RG-ZOSA based on Randomized Gradient Estimation

In this section, besides the L-smoothness required in (10) we also assume L_0-Lipschitz continuity of function $\mathbf{f}(\mathbf{x}, \xi_t)$, where $t \in [N]$. Under this additional condition, we aim to further improve the approximation ratio of the algorithm. It is obvious that the f is also L-smooth and L_0-Lipschitz continuous. We employ randomized gradient estimation to approximate the exact gradient of the objective function. It is worth noting that both the sampling process and the design of the step-size need to be adjusted to accommodate the randomization in gradient estimation. As a result, the theoretical analysis and approximation performance of the algorithm are expected to differ.

To utilize the randomized gradient estimation, we first propose a mini-batch zeroth-order gradient estimator which plays an important role in the subsequent algorithm design. We randomly select an index set $B \subseteq [N]$ and generate i.i.d. samples $\{\nu_l\}_{l\in B}$ from a uniform distribution on the unit sphere $\mathbb{S}$. Let $\tilde{\mathcal{B}} = \{(\nu_l, \xi_l)\}_{l\in B}$ and define the mini-batch zeroth-order randomized gradient estimator at $\mathbf{x}$ as follows:

$$\bar{\mathbf{g}}_{\rm rand}(\mathbf{x}, \tilde{\mathcal{B}}) := \frac{1}{|B|}\sum_{l\in B} \mathbf{g}_{\rm rand}(\mathbf{x}, \nu_l, \xi_l). \tag{29}$$

Combining Lemma 5 with the L_0-Lipschitz continuity of $f(\mathbf{x})$, we can easily derive the relevant properties of $\bar{\mathbf{g}}_{\rm rand}(\mathbf{x}, \tilde{\mathcal{B}})$, i.e.,

$$\mathbb{E}[\bar{\mathbf{g}}_{\rm rand}(\mathbf{x}, \tilde{\mathcal{B}})|\mathbf{x}] = \nabla f_\mu(\mathbf{x}), \quad \mathbb{E}\left[\|\nabla f_\mu(\mathbf{x}) - \bar{\mathbf{g}}_{\rm rand}(\mathbf{x}, \tilde{\mathcal{B}})\|^2|\mathbf{x}\right] \le \frac{16\sqrt{2\pi}dL_0^2}{|B|}. \tag{30}$$

Similarly with the design of coordinate-wise gradient estimation based on integral auxiliary function, we define the integral auxiliary function of the smoothed function $f_\mu(\mathbf{x})$ as

$$F_\mu(\mathbf{x}) := \int_0^1 \frac{e^{\theta-1}}{\theta} f_\mu(\theta\mathbf{x})d\theta, \tag{31}$$

which is non-negative due to the non-negativeness of f_μ as shown in Lemma 4(i). Easily, we have $\nabla F_\mu(\mathbf{x}) = \int_0^1 e^{\theta-1}\nabla f_\mu(\theta\mathbf{x})d\theta$. We then pick an i.i.d. sample θ from random variable Θ in (12) and $\tilde{\mathcal{B}} = \{\nu_l, \xi_l\}_{l\in B}$ with $B \subseteq [N]$, and set

$$\mathbf{G}_{\rm rand}(\mathbf{x}, \theta, \tilde{\mathcal{B}}) := \left(1 - e^{-1}\right)\bar{\mathbf{g}}_{\rm rand}(\theta\mathbf{x}, \tilde{\mathcal{B}}), \tag{32}$$

where $\bar{\mathbf{g}}_{\rm rand}(\theta\mathbf{x}, \tilde{\mathcal{B}})$ is defined as (29). It follows from the expression of $\nabla F_\mu(\mathbf{x})$ as indicated by (31) that

$$\begin{aligned}
\mathbb{E}\left[\mathbf{G}_{\rm rand}(\mathbf{x}, \theta, \tilde{\mathcal{B}})|\mathbf{x}\right] &= \mathbb{E}_{\theta\sim\Theta}\left[\left(1 - e^{-1}\right)\mathbb{E}[\bar{\mathbf{g}}_{\rm rand}(\theta\mathbf{x}, \tilde{\mathcal{B}})|\mathbf{x}, \theta]\right] \\
&= \left(1 - e^{-1}\right)\mathbb{E}_{\theta\sim\Theta}\left[\nabla f_\mu(\theta\mathbf{x})\right] \\
&= \left(1 - e^{-1}\right)\int_0^1 \frac{e^{\theta-1}}{\int_0^1 e^{u-1}du}\nabla f_\mu(\theta\mathbf{x})d\theta \\
&= \int_0^1 e^{\theta-1}\nabla f_\mu(\theta\mathbf{x})d\theta = \nabla F_\mu(\mathbf{x}),
\end{aligned} \tag{33}$$

where the third equality is due to the distribution as given in $\mathbb{P}(\Theta \leq \theta)$ in (12). Thus, $\mathbf{G}_{\text{rand}}(\mathbf{x}, \theta, \tilde{\mathcal{B}})$ is an unbiased estimate to $\nabla F_\mu(\mathbf{x})$.

We now come to determine $\mathbf{d}_j^{s+1}$. We first randomly select an index set $B_j^{s+1} \subseteq [N]$ with the size $|B_j^{s+1}| = b$ and generate i.i.d. samples $\nu_l \in \mathbb{R}^d$ for $l \in B_j^{s+1}$ from a uniform distribution on the unit sphere $\mathbb{S}$. Denote $\tilde{\mathcal{B}}_j^{s+1} := \{(\nu_l, \xi_l)\}_{l \in B_j^{s+1}}$ and $\tilde{\mathcal{N}} := \{(\nu_l, \xi_l)\}_{l \in [N]}$ when $B_j^{s+1} = [N]$. Then compute the zeroth-order approximate gradient $\mathbf{d}_j^{s+1}$ through

$$\mathbf{d}_j^{s+1} = \begin{cases} \mathbf{G}_{\text{rand}}(\mathbf{x}_0^{s+1}, \theta^{s+1}, \tilde{\mathcal{N}}), & j = 0 \\ \mathbf{G}_{\text{rand}}(\mathbf{x}_j^{s+1}, \theta^{s+1}, \tilde{\mathcal{B}}_j^{s+1}) - \mathbf{G}_{\text{rand}}(\mathbf{x}_0^{s+1}, \theta^{s+1}, \tilde{\mathcal{B}}_j^{s+1}) + \mathbf{G}_{\text{rand}}(\mathbf{x}_0^{s+1}, \theta^{s+1}, \tilde{\mathcal{N}}), & j > 0 \end{cases}, \tag{34}$$

where $\theta^{s+1} \in (0, 1]$ is an i.i.d. sample from the random variable Θ. For convenience, we name Algorithm 1 with the design of (34) as **RG-ZOSA**.

The next lemma demonstrates that the approximate gradient $\mathbf{d}_j^{s+1}$ in (34) is an unbiased estimate of $\nabla F_\mu(\mathbf{x})$, and it provides an upper bound on the variance of $\mathbf{d}_j^{s+1}$.

Lemma 14. *For any $j = 0, \ldots, m-1$ and $s = 0, \ldots S-1$ in* **RG-ZOSA**, *it holds that $\mathbb{E}[\mathbf{d}_j^{s+1} | \mathbf{x}_j^{s+1}] = \nabla F_\mu(\mathbf{x}_j^{s+1})$ and*

$$\mathbb{E}\left[\|\nabla F_\mu(\mathbf{x}_j^{s+1}) - \mathbf{d}_j^{s+1}\|^2 | \mathbf{x}_j^{s+1}\right] \leq \frac{2\left(1 - e^{-1}\right)\left(1 - 2e^{-1}\right) M_u^2}{b} \|\mathbf{x}_j^{s+1} - \mathbf{x}_0^{s+1}\|^2 + \frac{16\sqrt{2\pi}\left(1 - e^{-1}\right)^2 L_0^2 d}{N} + L^2 R^2, \tag{35}$$

where $M_u = \frac{dL_0}{u}$.

Proof According to the unbiasedness of $\mathbb{E}[\mathbf{G}_{\text{rand}}(\mathbf{x}, \theta, \tilde{\mathcal{B}}) | \mathbf{x}]$ as (33), it is obvious to obtain that for any $j = 0, \ldots, m-1$ and $s = 0, \ldots S-1$,

$$\begin{aligned} &\mathbb{E}\left[\mathbf{d}_j^{s+1} | \mathbf{x}_j^{s+1}\right] \\ &= \mathbb{E}\left[\mathbf{G}_{\text{rand}}(\mathbf{x}_j^{s+1}, \theta^{s+1}, \tilde{\mathcal{B}}_j^{s+1}) - \mathbf{G}_{\text{rand}}(\mathbf{x}_0^{s+1}, \theta^{s+1}, \tilde{\mathcal{B}}_j^{s+1}) + \mathbf{G}_{\text{rand}}(\mathbf{x}_0^{s+1}, \theta^{s+1}, \tilde{\mathcal{N}}) | \mathbf{x}_j^{s+1}\right] \\ &= \mathbb{E}\left[\mathbf{G}_{\text{rand}}(\mathbf{x}_j^{s+1}, \theta^{s+1}, \tilde{\mathcal{B}}_j^{s+1}) | \mathbf{x}_j^{s+1}\right] = \nabla F_\mu(\mathbf{x}_j^{s+1}). \end{aligned}$$

To bound the variance $\mathbb{E}[\|\nabla F_\mu(\mathbf{x}_j^{s+1}) - \mathbf{d}_j^{s+1}\|^2 | \mathbf{x}_j^{s+1}]$, we first notice from expression of $\nabla F_\mu(\mathbf{x})$ that

$$\begin{aligned} &\mathbb{E}\left[\|\nabla F_\mu(\mathbf{x}) - \mathbf{G}_{\text{rand}}(\mathbf{x}, \theta, \tilde{\mathcal{N}})\|^2 | \mathbf{x}\right] \\ &= \mathbb{E}\left[\|\mathbf{G}_{\text{rand}}(\mathbf{x}, \theta, \tilde{\mathcal{N}}) - \left(1 - e^{-1}\right) \nabla f_\mu(\theta \mathbf{x})\|^2 + \|\left(1 - e^{-1}\right) \nabla f_\mu(\theta \mathbf{x}) - \nabla F_\mu(\mathbf{x})\|^2 | \mathbf{x}\right] \\ &= \mathbb{E}\left[\left(1 - e^{-1}\right)^2 \|\bar{\mathbf{g}}_{\text{rand}}(\theta \mathbf{x}, \tilde{\mathcal{N}}) - \nabla f_\mu(\theta \mathbf{x})\|^2 + \| \int_0^1 e^{t-1} \left(\nabla f_\mu(\theta \mathbf{x}) - \nabla f_\mu(t \mathbf{x})\right) dt\|^2 | \mathbf{x}\right] \\ &\leq \frac{16\sqrt{2\pi}\left(1 - e^{-1}\right)^2 d L_0^2}{N} + \left(\int_0^1 e^{2(t-1)} (t - \theta)^2 dt\right) L^2 R^2 \end{aligned}$$

$$\leq \frac{16\sqrt{2\pi}\left(1-e^{-1}\right)^2 dL_0^2}{N} + L^2R^2, \tag{36}$$

in which the first equality is derived by $\mathbb{E}_{\theta\sim\Theta}\left[\left(1-e^{-1}\right)\nabla f_\mu(\theta\mathbf{x})\right] = \nabla F_\mu(\mathbf{x})$ as shown in (33), the second equality is from the definitions of $\mathbf{G}_{\text{rand}}(\mathbf{x},\theta,\tilde{\mathcal{N}})$ in (32) and $\nabla F_\mu(\mathbf{x}) = \int_0^1 e^{t-1}\nabla f_\mu(t\mathbf{x})dt$, the first inequality is due to (30) with $|B| = N$ as well as the L-smoothness of f_μ from Lemma 4(ii). Then we obtain

$$\begin{aligned}
&\mathbb{E}\left[\|\nabla F_\mu(\mathbf{x}_j^{s+1}) - \mathbf{d}_j^{s+1}\|^2|\mathbf{x}_j^{s+1}\right] \\
&= \mathbb{E}\left[\|\nabla F_\mu(\mathbf{x}_j^{s+1}) - \Big(\mathbf{G}_{\text{rand}}(\mathbf{x}_j^{s+1},\theta^{s+1},\tilde{\mathcal{B}}_j^{s+1}) - \mathbf{G}_{\text{rand}}(\mathbf{x}_0^{s+1},\theta^{s+1},\tilde{\mathcal{B}}_j^{s+1})\right. \\
&\quad \left.+\mathbf{G}_{\text{rand}}(\mathbf{x}_0^{s+1},\theta^{s+1},\tilde{\mathcal{N}})\Big)\|^2|\mathbf{x}_j^{s+1}\right] \\
&= \mathbb{E}\left[\|-\mathbf{G}_{\text{rand}}(\mathbf{x}_j^{s+1},\theta^{s+1},\tilde{\mathcal{B}}_j^{s+1}) + \mathbf{G}_{\text{rand}}(\mathbf{x}_0^{s+1},\theta^{s+1},\tilde{\mathcal{B}}_j^{s+1}) + \nabla F_\mu(\mathbf{x}_j^{s+1}) - \nabla F_\mu(\mathbf{x}_0^{s+1})\|^2|\mathbf{x}_j^{s+1}\right] \\
&\quad +\mathbb{E}\left[\|\mathbf{G}_{\text{rand}}(\mathbf{x}_j^{s+1},\theta^{s+1},\tilde{\mathcal{N}}) - \mathbf{G}_{\text{rand}}(\mathbf{x}_0^{s+1},\theta^{s+1},\tilde{\mathcal{N}}) + \nabla F_\mu(\mathbf{x}_0^{s+1}) - \nabla F_\mu(\mathbf{x}_j^{s+1})\|^2|\mathbf{x}_j^{s+1}\right] \\
&\quad +\mathbb{E}\left[\|\nabla F_\mu(\mathbf{x}_j^{s+1}) - \mathbf{G}_{\text{rand}}(\mathbf{x}_j^{s+1},\theta^{s+1},\tilde{\mathcal{N}})\|^2|\mathbf{x}_j^{s+1}\right] \\
&\leq 2\mathbb{E}\left[\|\mathbf{G}_{\text{rand}}(\mathbf{x}_j^{s+1},\theta^{s+1},\tilde{\mathcal{B}}_j^{s+1}) - \mathbf{G}_{\text{rand}}(\mathbf{x}_0^{s+1},\theta^{s+1},\tilde{\mathcal{B}}_j^{s+1})\|^2|\mathbf{x}_j^{s+1}\right] \\
&\quad +\mathbb{E}\left[\|\nabla F_\mu(\mathbf{x}_j^{s+1}) - \mathbf{G}_{\text{rand}}(\mathbf{x}_j^{s+1},\theta^{s+1},\tilde{\mathcal{N}})\|^2|\mathbf{x}_j^{s+1}\right] \\
&\leq \frac{2\left(1-e^{-1}\right)^2}{b}\mathbb{E}\left[\|\mathbf{g}_{\text{rand}}(\theta^{s+1}\mathbf{x}_j^{s+1},\nu_l,\xi_l) - \mathbf{g}_{\text{rand}}(\theta^{s+1}\mathbf{x}_0^{s+1},\nu_l,\xi_l)\|^2|\mathbf{x}_j^{s+1}\right] \\
&\quad +\mathbb{E}\left[\|\nabla F_\mu(\mathbf{x}_j^{s+1}) - \mathbf{G}_{\text{rand}}(\mathbf{x}_j^{s+1},\theta^{s+1},\tilde{\mathcal{N}})\|^2|\mathbf{x}_j^{s+1}\right] \\
&\leq \frac{2\left(1-e^{-1}\right)^2 M_u^2}{b}\|\mathbf{x}_j^{s+1} - \mathbf{x}_0^{s+1}\|^2\int_0^1 \frac{e^{\theta^{s+1}-1}}{\int_0^1 e^{u-1}du}(\theta^{s+1})^2 d\theta^{s+1} \\
&\quad +\mathbb{E}\left[\|\nabla F_\mu(\mathbf{x}_j^{s+1}) - \mathbf{G}_{\text{rand}}(\mathbf{x}_j^{s+1},\theta^{s+1},\tilde{\mathcal{N}})\|^2|\mathbf{x}_j^{s+1}\right].
\end{aligned}$$

In above relations, the second equality comes from $\mathbb{E}[\mathbf{G}_{\text{rand}}(\mathbf{x},\theta,\tilde{\mathcal{B}})|\mathbf{x}] = \nabla F_\mu(\mathbf{x})$ in (33), the first inequality is due to the fact that $\mathbb{E}[\|\xi - \mathbb{E}[\xi]\|^2] \leq \mathbb{E}[\|\xi\|^2]$ for a random variable ξ and $|\tilde{\mathcal{B}}_j^{s+1}| \leq |\tilde{\mathcal{N}}|$, the second inequality comes from the definitions of $\mathbf{G}_{\text{rand}}(\mathbf{x},\theta,\tilde{\mathcal{B}})$ in (32) and $\bar{\mathbf{g}}_{\text{rand}}(\mathbf{x},\tilde{\mathcal{B}})$ in (29), the third one is from Lemma 5 with the assumption that $\mathbf{f}(\mathbf{x},\xi_t), t \in [N]$ are L_0-Lipschitz continuous, and the fact that θ^{s+1} is sampled from Θ following the probability function in (12).

Therefore, due to $\int_0^1 e^{u-1}du = 1-e^{-1}$ and $\int_0^1 e^{\theta-1}\theta^2 d\theta = 1-2e^{-1}$ in the inequality (36), we obtain inequality (35). ■

According to Lemma 4(ii) and Lemma 8(i), the auxiliary function $F_\mu(\mathbf{x})$, as defined in (31), is $\frac{L}{e}$-smooth due to the L-smoothness of $\mathbf{f}(\mathbf{x},\xi_t)$. Before proceeding, we present the following lemma, similar to Lemma 10, with the proof omitted.

Lemma 15. *For any $j = 0, \ldots, m-1$ and $s = 0, \ldots S-1$ in* **RG-ZOSA**, *it holds that for any $\mathbf{y} \in \mathcal{X}$,*

$$F_\mu(\mathbf{x}_{j+1}^{s+1}) \geq F_\mu(\mathbf{x}_j^{s+1}) + \frac{1}{2\eta_j^{s+1}}\left(\|\mathbf{x}_{j+1}^{s+1} - \mathbf{y}\|^2 - \|\mathbf{x}_j^{s+1} - \mathbf{y}\|^2\right) + \langle \mathbf{d}_j^{s+1}, \mathbf{y} - \mathbf{x}_j^{s+1}\rangle$$
$$- \eta_j^{s+1}\|\nabla F_\mu(\mathbf{x}_j^{s+1}) - \mathbf{d}_j^{s+1}\|^2 - \left(\frac{L}{2e} - \frac{1}{4\eta_j^{s+1}}\right)\|\mathbf{x}_{j+1}^{s+1} - \mathbf{x}_j^{s+1}\|^2. \tag{37}$$

Similarly with Lemma 11, we show a critical lemma as follows. For simplicity, we next denote

$$\hat{L} = \max\left\{\frac{L}{e}, \sqrt{2\left(1 - e^{-1}\right)\left(1 - 2e^{-1}\right)}M_u\right\}, \text{ where } M_u = \frac{dL_0}{u}.$$

Lemma 16. *Given the parameters $a_m = 0$, $a_j = a_{j+1}\left(1 + \frac{1}{m}\right) + \frac{\eta\hat{L}^2}{b}$ and*

$$b = m^2 < N, \quad \eta = \frac{1}{c\hat{L}},$$

where $c = 4\sqrt{2}$, we have

$$\frac{\hat{L}}{2} + a_{j+1}(1 + m) \leq \frac{1}{4\eta}.$$

Combined with Lemma 14, Lemma 15 and Lemma 16, we show the approximation performance for the **RG-ZOSA** algorithm as follows.

Theorem 17. *Let parameters for the computation of $\mathbf{d}_j^{s+1}$ and step-sizes η_j^{s+1} in* **RG-ZOSA** satisfy

$$b = m^2 \leq N, \quad u = \sqrt{\frac{d}{Sm}}, \quad \eta_j^{s+1} = \eta, \tag{38}$$

where $\eta = 1/c\hat{L}$ with $c = 4\sqrt{2}$ for $j = 0, \ldots, m-1$ and $s = 0, \ldots S-1$. Return $\mathbf{x}_r$ such that $\mathrm{P}(\mathbf{x}_r = \mathbf{x}_j^{s+1}) = \frac{1}{Sm+1+\ln\tau}$ for $j = 0, \ldots, m-1; s = 0, \ldots, S-1$, and $\mathrm{P}(\mathbf{x}_r = \mathbf{x}_m^S) = \frac{1+\ln\tau}{Sm+1+\ln\tau}$. Then it holds that

$$\mathbb{E}[f(\mathbf{x}_r)] \geq \left(1 - e^{-1} - \mathcal{O}\left(\frac{1}{Sm}\right)\right) f(\mathbf{x}^*) - O\left(\frac{\sqrt{d}}{\sqrt{Sm}}\right), \tag{39}$$

where $\mathbf{x}^*$ is the optimal solution of (10).

Proof We define the following Lyapunov function

$$L_{j+1}^{s+1} := \mathbb{E}\left[F_\mu(\mathbf{x}_j^{s+1}) - a_j\|\mathbf{x}_j^{s+1} - \mathbf{x}_0^{s+1}\|^2\right],$$

where the parameters $a_m = 0$ and $a_j = a_{j+1}\left(1 + \frac{1}{m}\right) + \frac{\eta\hat{L}^2}{b}$. Under the condition $c = 4\sqrt{2}$ and $b = m^2$, similar to (24), we bound L_{j+1}^{s+1} as follows

$$L_{j+1}^{s+1} \geq L_j^{s+1} + \frac{1}{2\eta}\mathbb{E}\left[\|\mathbf{x}_{j+1}^{s+1} - \mathbf{y}\|^2 - \|\mathbf{x}_j^{s+1} - \mathbf{y}\|^2\right] + \mathbb{E}\left[\langle \mathbf{d}_j^{s+1}, \mathbf{y} - \mathbf{x}_j^{s+1}\rangle\right]$$

$$- \left(\frac{16\sqrt{2\pi}\left(1-e^{-1}\right)^2 L_0^2 d}{N} + L^2 R^2 \right) \eta, \tag{40}$$

where we used Lemmas 14-16. Now, we set $\mathbf{y} = \mathbf{x}^*$, and telescope (40) over all the iteration in epoch s. Then from the unbiasedness of $\mathbf{d}_j^{s+1}$, it follows that

$$\begin{aligned} L_m^{s+1} \geq L_0^{s+1} + \frac{1}{2\eta}\mathbb{E}\left[\|\mathbf{x}_m^{s+1} - \mathbf{x}^*\|^2 - \|\mathbf{x}_0^{s+1} - \mathbf{x}^*\|^2\right] + \sum_{j=0}^{m-1} \mathbb{E}\left[\langle \nabla F_\mu(\mathbf{x}_j^{s+1}), \mathbf{x}^* - \mathbf{x}_j^{s+1}\rangle\right] \\ - \sum_{j=0}^{m-1} \left(\frac{16\sqrt{2\pi}\left(1-e^{-1}\right)^2 L_0^2 d}{N} + L^2 R^2 \right) \eta. \end{aligned} \tag{41}$$

Furthermore, we have $L_m^{s+1} = \mathbb{E}[F_\mu(\mathbf{x}_m^{s+1})]$ from $a_m = 0$ and the definition of $\mathbf{x}_0^{s+1}$, and $L_0^{s+1} = \mathbb{E}[F_\mu(\mathbf{x}_0^{s+1})] = \mathbb{E}[F_\mu(\mathbf{x}_m^s)]$, which is due to the setting $\mathbf{x}_0^{s+1} = \mathbf{x}_m^s$. Thus, following inequality holds from (41) and $\eta = 1/c\hat{L}$:

$$\begin{aligned} \mathbb{E}\left[F_\mu(\mathbf{x}_m^{s+1})\right] &\geq \mathbb{E}\left[F_\mu(\mathbf{x}_m^s)\right] + \frac{c\hat{L}}{2}\mathbb{E}\left[\|\mathbf{x}_m^{s+1} - \mathbf{x}^*\|^2 - \|\mathbf{x}_0^{s+1} - \mathbf{x}^*\|^2\right] \\ &\quad + \sum_{j=0}^{m-1} \mathbb{E}\left[\langle \nabla F_\mu(\mathbf{x}_j^{s+1}), \mathbf{x}^* - \mathbf{x}_j^{s+1}\rangle\right] \\ &\quad - \sum_{j=0}^{m-1} \frac{1}{c\hat{L}} \left(\frac{16\sqrt{2\pi}\left(1-e^{-1}\right)^2 L_0^2 d}{N} + L^2 R^2 \right) \\ &\geq \mathbb{E}\left[F_\mu(\mathbf{x}_m^s)\right] + \frac{c\hat{L}}{2}\mathbb{E}\left[\|\mathbf{x}_m^{s+1} - \mathbf{x}^*\|^2 - \|\mathbf{x}_0^{s+1} - \mathbf{x}^*\|^2\right] \\ &\quad + \sum_{j=0}^{m-1} \mathbb{E}\left[\left(1-e^{-1}\right) f(\mathbf{x}^*) - f(\mathbf{x}_j^{s+1}) - (2-e^{-1}) L_0 u\right] \\ &\quad - \sum_{j=0}^{m-1} \frac{1}{c\hat{L}} \left(\frac{16\sqrt{2\pi}\left(1-e^{-1}\right)^2 L_0^2 d}{N} + L^2 R^2 \right), \end{aligned} \tag{42}$$

where the last inequality is due to Lemma 8(ii) with F replaced by F_μ, $\mathbf{y} = \mathbf{x}^*$ and $\mathbf{x} = \mathbf{x}_j^{s+1}$ and Lemma 4(iii). Summing up above inequality over all the epochs (for $s \in \{0, \ldots, S-1\}$), we have

$$\begin{aligned} \mathbb{E}\left[F_\mu(\mathbf{x}_m^S)\right] &\geq \mathbb{E}\left[F_\mu(\mathbf{x}_m^0)\right] + \frac{c\hat{L}}{2}\mathbb{E}\left[\|\mathbf{x}_m^S - \mathbf{x}^*\|^2 - \|\mathbf{x}_0^1 - \mathbf{x}^*\|^2\right] \\ &\quad + \sum_{s=0}^{S-1}\sum_{j=0}^{m-1} \mathbb{E}\left[\left(1-e^{-1}\right) f(\mathbf{x}^*) - f(\mathbf{x}_j^{s+1}) - 2L_0 u\right] \\ &\quad - \sum_{s=0}^{S-1}\sum_{j=0}^{m-1} \frac{1}{c\hat{L}} \left(\frac{16\sqrt{2\pi}\left(1-e^{-1}\right)^2 L_0^2 d}{N} + L^2 R^2 \right) \\ &\geq -\frac{c\hat{L}}{2} D^2 + \sum_{s=0}^{S-1}\sum_{j=0}^{m-1} \mathbb{E}\left[\left(1-e^{-1}\right) f(\mathbf{x}^*) - f(\mathbf{x}_j^{s+1}) - 2L_0 u\right] \end{aligned}$$

$$-\sum_{s=0}^{S-1}\sum_{j=0}^{m-1}\frac{1}{c\hat{L}}\left(\frac{16\sqrt{2\pi}\left(1-e^{-1}\right)^2L_0^2d}{N}+L^2R^2\right),$$

where the second inequality is from the non-negativeness of $F_\mu(\mathbf{x})$ and $\mathbb{E}[\|\mathbf{x}_m^S-\mathbf{x}^*\|^2]$, and by considering the diameter of the constraint set $\mathcal{X}$. Furthermore, we rearrange the inequality, obtaining

$$\begin{aligned}&\mathbb{E}\left[F_\mu(\mathbf{x}_m^S)+\sum_{s=0}^{S-1}\sum_{j=0}^{m-1}f(\mathbf{x}_j^{s+1})\right]\\&\geq Sm\left(1-e^{-1}\right)f(\mathbf{x}^*)-\frac{c\hat{L}D^2}{2}-\sum_{s=0}^{S-1}\sum_{j=0}^{m-1}\frac{1}{c\hat{L}}\left(\frac{16\sqrt{2\pi}\left(1-e^{-1}\right)^2L_0^2d}{N}+L^2R^2+2L_0u\right).\end{aligned}\tag{43}$$

According to Lemma 8(iii) and Lemma 4(iii), we have

$$F_\mu(\mathbf{x})\leq(1+\ln\tau)(f_\mu(\mathbf{x})+\delta)\leq(1+\ln\tau)(f(\mathbf{x})+L_0u+\delta).$$

Thus, it indicates from (43) that

$$\begin{aligned}&\mathbb{E}\left[\sum_{s=0}^{S-1}\sum_{j=0}^{m-1}f(\mathbf{x}_j^{s+1})+(1+\ln\tau)\left(f(\mathbf{x}_m^S)+L_0u+\delta\right)\right]\\&\geq Sm\left(1-e^{-1}\right)f(\mathbf{x}^*)-\frac{c\hat{L}D^2}{2}-\sum_{s=0}^{S-1}\sum_{j=0}^{m-1}\frac{1}{c\hat{L}}\left(\frac{16\sqrt{2\pi}\left(1-e^{-1}\right)^2L_0^2d}{N}+L^2R^2+2L_0u\right).\end{aligned}$$

Dividing both sides by $(Sm+1+\ln\tau)$ and using the setting $u=\sqrt{\frac{d}{Sm}}$ yields

$$\begin{aligned}\mathbb{E}\left[f(\mathbf{x}_r)\right]&=\mathbb{E}\left[\sum_{s=0}^{S-1}\sum_{j=0}^{m-1}\frac{1}{Sm+1+\ln\tau}f(\mathbf{x}_j^{s+1})+\frac{1+\ln\tau}{Sm+1+\ln Z}f(\mathbf{x}_m^S)\right]\\&\geq\frac{Sm}{Sm+1+\ln\tau}\left(1-e^{-1}\right)f(\mathbf{x}^*)\\&\quad-\frac{\frac{c\hat{L}D^2}{2}+(2Sm+1+\ln\tau)L_0u+\frac{Sm}{c\hat{L}}\left(\frac{16\sqrt{2\pi}\left(1-e^{-1}\right)^2}{N}L_0^2d+L^2R^2\right)}{Sm+1+\ln\tau}\\&\geq\left(1-e^{-1}-\mathcal{O}\left(\frac{1}{Sm}\right)\right)f(\mathbf{x}^*)-\frac{\frac{cD^2}{2}(\frac{dL_0}{u}+L)+(1+\ln\tau)L_0u}{Sm+1+\ln\tau}\\&\quad-\frac{Sm\left[\left(2+\frac{16\sqrt{2\pi}(1-e^{-1})^2}{c\sqrt{2(1-e^{-1})(1-2e^{-1})N}}\right)L_0u+\frac{L^2R^2}{c\sqrt{2(1-e^{-1})(1-2e^{-1})}}\frac{u}{dL_0}\right]}{Sm+1+\ln\tau}\\&=\left(1-e^{-1}-\mathcal{O}\left(\frac{1}{Sm}\right)\right)f(\mathbf{x}^*)-\mathcal{O}\left(\frac{\sqrt{d}}{\sqrt{Sm}}\right),\end{aligned}\tag{44}$$

where $c = 4\sqrt{2}$ and the last inequality is from

$$\frac{dL_0}{u} + L \geq \hat{L} \geq \sqrt{2\left(1 - e^{-1}\right)\left(1 - 2e^{-1}\right)}\frac{dL_0}{u}.$$

Therefore, (39) can be derived. ■

Corollary 18. *Under the same conditions as Theorem 17 with additional setting $b = N^{2/3}$ and $\epsilon > 0$,* **RG-ZOSA** *outputs $\mathbf{x}_r$ satisfying that*

$$\mathbb{E}[f(\mathbf{x}_r)] \geq \left(1 - e^{-1} - \frac{\epsilon^2}{d}\right) f(\mathbf{x}^*) - \epsilon,$$

after $\mathcal{O}(d\epsilon^{-2})$ iterations and $\mathcal{O}(N^{2/3}d\epsilon^{-2})$ function evaluations.

Proof The proof is similar to Corollary 13, so we omit it here. ■

Remark 19. *We note that the L-smoothness of function f plays a crucial role in the approximation analysis of* **RG-ZOSA**. In the analysis of (36), the L-smoothness of f_μ helps provide the upper bound of variance for $\mathbf{d}_j^{s+1}$ as shown in Lemma 14, which further establishes the complexity of **RG-ZOSA** as given in Lemma 17. We would like to emphasize that without the L-smoothness property of f, if only assuming the L_0-Lipschitz continuity, we will not be able to obtain desirable approximation properties of **RG-ZOSA**. This can be explained by intuitive illustrations. Specifically, under mere L_0-Lipschitz continuity assumption of f and according to Lemma 4 (iii), f_μ is $\frac{L_0\sqrt{d}}{u}$-smooth, then in the conclusion and proof of Lemma 14 and Lemma 15, L should be replaced by $\frac{L_0\sqrt{d}}{u}$. Besides, $\hat{L}$, defined after Lemma 15, can be set as $\frac{L_0 d}{u}$ due to

$$\hat{L} = \frac{L_0 d}{u} \geq \max\left\{\frac{L}{e}, \sqrt{2(1 - e^{-1})(1 - 2e^{-1})}M_u\right\}, \text{ where } L = \frac{L_0\sqrt{d}}{u} \text{ and } M_u = \frac{L_0 d}{u}.$$

Consequently, in the proof of Theorem 17, the right hand side of (44) will contain the term

$$\begin{aligned}
&\frac{Sm}{Sm + 1 + \ln\tau} \cdot \left(\frac{16\sqrt{2\pi}\left(1 - e^{-1}\right)^2 L_0^2 d}{N \cdot c\hat{L}} + \frac{L^2 R^2}{c\hat{L}}\right) \\
&= \frac{Sm}{Sm + 1 + \ln\tau} \cdot \left(\frac{16\sqrt{2\pi}\left(1 - e^{-1}\right)^2 L_0}{N \cdot c} \cdot u + \frac{R^2 L_0}{c \cdot u}\right),
\end{aligned}$$

which cannot be controlled in a desired level no matter how to adjust the settings of S and m. Therefore, the smoothness of f is a key to analyse the properties of **RG-ZOSA**, which naturally poses an issue about how to tackle DR-submodular optimization problems without smoothness. We will address this issue in next section.

4. NZOSA: Zeroth-order Stochastic Approximation Method for Nonsmooth Up-Concave Maximization

In this section, we consider a class of nonsmooth up-concave maximization problems:

$$\max_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) := \frac{1}{N}\sum_{t=1}^{N}\mathbf{f}(\mathbf{x},\xi_t), \tag{45}$$

where $\mathbf{f}(\cdot,\xi_t):\mathbb{R}^d\to\mathbb{R}$, $t\in[N]$ are L_0-Lipschitz continuous, f is non-negative, monotonically non-decreasing and up-concave, and $\mathcal{X}\subseteq\mathbb{R}^d_+$ represents a compact and convex set.

We observe that the integral-form auxiliary functions introduced in the previous section require the parameter sampling in each outer loop, which produces additional errors of $\mathbf{d}_j^{s+1}$. As discussed in Remark 19, the absence of smoothness makes the previous analysis unsuitable. To design an efficient algorithm with desired properties for non-smooth up-concave submodular maximization problems, we give a finite-sum auxiliary function for the smoothed function that can be computed with certainty. A similar form of auxiliary function was proposed by (Mitra et al., 2021) for maximizing a combination of a continuous DR-submodular function and a concave function, but our finite-sum auxiliary function differs in that it does not depend on the accuracy parameter ϵ. Specifically, for (45), we introduce the auxiliary function $\hat{F}_\mu:\mathbb{R}^d\to\mathbb{R}_+$ as

$$\hat{F}_\mu(\mathbf{x}) = \frac{1}{N}\sum_{t=1}^{N}\hat{\mathbf{F}}_\mu(\mathbf{x},\xi_t),\quad \text{where}\quad \hat{\mathbf{F}}_\mu(\mathbf{x},\xi_t) := \frac{1}{Z}\sum_{z=1}^{Z}\frac{e^{\frac{z}{Z}-1}}{\frac{z}{Z}}\mathbf{f}_\mu\left(\frac{z}{Z}\mathbf{x},\xi_t\right), t\in[N] \tag{46}$$

with $Z\geq 3$ being an integer and $\mathbf{f}_\mu$ following the definition in (6). Obviously, by (7) we have

$$\hat{F}_\mu(\mathbf{x}) = \frac{1}{Z}\sum_{z=1}^{Z}\frac{e^{\frac{z}{Z}-1}}{\frac{z}{Z}}f_\mu\left(\frac{z}{Z}\mathbf{x}\right), \tag{47}$$

which is non-negative due to the non-negativeness of f_μ as shown in Lemma 4(i). The lemma below characterizes more properties of $\hat{F}_\mu$ that are essential for subsequent approximation analysis. Note that Lemma 20(iii) and (iv) relies on the up-concavity of f, and the proof is presented in Appendix D.

Lemma 20. *Let $\hat{F}_\mu$ be defined by (46). Then the following statements hold true.*

(i) The gradient of $\hat{F}_\mu(\mathbf{x})$ is given by

$$\nabla\hat{F}_\mu(\mathbf{x}) = \frac{1}{Z}\sum_{z=1}^{Z}e^{\frac{z}{Z}-1}\nabla f_\mu\left(\frac{z}{Z}\mathbf{x}\right),$$

and then $\hat{F}_\mu(\mathbf{x})$ is L_μ-smooth, where $L_\mu=\frac{\sqrt{d}L_0}{u}$.

(ii) For any $\mathbf{x}\in\mathcal{X}$, it holds that $\hat{F}_\mu(\mathbf{x})\leq(1+\ln Z)f_\mu(\mathbf{x})$.

(iii) For any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$*, it holds that*

$$\langle \mathbf{y} - \mathbf{x}, \nabla \hat{F}_\mu(\mathbf{x}) \rangle \geq \left(1 - e^{-1}\right) f(\mathbf{y}) - \left(1 + \frac{3 \ln Z}{Z}\right) f(\mathbf{x}) - 3L_0 u. \tag{48}$$

(iv) Let $\mathbf{x}^*$ *be an optimal solution of* (45) *and* $\mathbf{x}$ *be a stationary point of* $\max_{x \in \mathcal{X}} \hat{F}_\mu(\mathbf{x})$*, then*

$$f(\mathbf{x}) \geq (1 - e^{-1} - \frac{3 \ln Z}{Z}) f(\mathbf{x}^*) - 3L_0 u.$$

Similar to the sampling scheme of the mini-batch $\bar{\mathbf{g}}_{\text{rand}}(\mathbf{x}, \tilde{\mathcal{B}})$ in (29), we randomly pick $\tilde{\mathcal{B}} = \{\nu_l, \xi_l\}_{l \in B}$ where $B \subseteq [N]$ and z uniformly from $\{1, \ldots, Z\}$. Then set

$$\hat{\mathbf{G}}(\mathbf{x}, z, \tilde{\mathcal{B}}) := e^{\frac{z}{Z} - 1} \bar{\mathbf{g}}_{\text{rand}}(\frac{z}{Z}\mathbf{x}, \tilde{\mathcal{B}}) = e^{\frac{z}{Z} - 1} \frac{1}{|B|} \sum_{l \in B} \mathbf{g}_{\text{rand}}(\mathbf{x}, \nu_l, \xi_l). \tag{49}$$

We can prove that $\hat{\mathbf{G}}(\mathbf{x}, z, \tilde{\mathcal{B}})$ is an unbiased estimate of $\nabla \hat{F}_\mu(\mathbf{x})$. In fact, it follows from Lemma 20(i) that

$$\begin{aligned} \mathbb{E}\left[\hat{\mathbf{G}}(\mathbf{x}, z, \tilde{\mathcal{B}}) | \mathbf{x}\right] &= \mathbb{E}\left[e^{\frac{z}{Z} - 1} \bar{\mathbf{g}}_{\text{rand}}(\frac{z}{Z}\mathbf{x}, \tilde{\mathcal{B}}) | \mathbf{x}\right] \\ &= \frac{1}{Z} \sum_{z=1}^{Z} e^{\frac{z}{Z} - 1} \nabla f_\mu(\frac{z}{Z}\mathbf{x}) \\ &= \nabla \hat{F}_\mu(\mathbf{x}), \end{aligned} \tag{50}$$

where the second equality comes from $\mathbb{E}[\bar{\mathbf{g}}_{\text{rand}}(\frac{z}{Z}\mathbf{x}, \tilde{\mathcal{B}}) | \mathbf{x}] = \nabla f_\mu(\frac{z}{Z}\mathbf{x})$ as described in (30).

We randomly select an i.i.d. sample z^{s+1} from $\{1, 2, \ldots, Z\}$ at the s-th epoch, and select $\tilde{\mathcal{B}}_j^{s+1} = \{(\nu_l, \xi_l)\}_{l \in B_j^{s+1}}$ at the j-th inner iteration where $B_j^{s+1} \subseteq [N]$ with the size $|B_j^{s+1}| = b$. Then we generate i.i.d. samples $\nu_l \in \mathbb{R}^d, l \in B_j^{s+1}$ following a uniform distribution on the unit sphere $\mathbb{S}$ and denote $\tilde{\mathcal{N}} = \{(\nu_l, \xi_l)\}_{l \in [N]}$ when $B_j^{s+1} = [N]$. We now compute the zeroth-order approximate gradient $\mathbf{d}_j^{s+1}$ based on the finite-sum auxiliary function $\hat{F}_\mu$ through

$$\mathbf{d}_j^{s+1} = \begin{cases} \hat{G}(\mathbf{x}_0^{s+1}, \tilde{\mathcal{N}}), & j = 0, \\ \hat{\mathbf{G}}(\mathbf{x}_j^{s+1}, z^{s+1}, \tilde{\mathcal{B}}_j^{s+1}) - \hat{\mathbf{G}}(\mathbf{x}_0^{s+1}, z^{s+1}, \tilde{\mathcal{B}}_j^{s+1}) + \hat{G}(\mathbf{x}_0^{s+1}, \tilde{\mathcal{N}}), & j > 0. \end{cases} \tag{51}$$

Here,

$$\hat{G}(\mathbf{x}_0^{s+1}, \tilde{\mathcal{N}}) = \frac{1}{Z} \sum_{z=1}^{Z} e^{\frac{z}{Z} - 1} \bar{\mathbf{g}}_{\text{rand}}(\frac{z}{Z}\mathbf{x}_0^{s+1}, \tilde{\mathcal{N}}) = \frac{1}{Z} \sum_{z=1}^{Z} e^{\frac{z}{Z} - 1} \left(\frac{1}{N} \sum_{l \in [N]} \mathbf{g}_{\text{rand}}(\mathbf{x}, \nu_l, \xi_l)\right).$$

For simplicity, we call Algorithm 1 incorporating (51) as **NZOSA**.

In the lemma below, we demonstrate the unbiasedness and upper-bounded variance of the approximate gradient $\mathbf{d}_j^{s+1}$ defined in (51).

Lemma 21. *For any $j = 0, \ldots, m-1$ and $s = 0, \ldots S-1$ in* **NZOSA**, *it holds that* $\mathbb{E}\left[\mathbf{d}_j^{s+1}|\mathbf{x}_j^{s+1}\right] = \nabla \hat{F}_\mu(\mathbf{x}_j^{s+1})$ *and*

$$\mathbb{E}\left[\|\nabla \hat{F}_\mu(\mathbf{x}_j^{s+1}) - \mathbf{d}_j^{s+1}\|^2|\mathbf{x}_j^{s+1}\right] \leq \frac{2M_u^2}{b}\|\mathbf{x}_j^{s+1} - \mathbf{x}_0^{s+1}\|^2 + \frac{16\sqrt{2\pi}L_0^2 d}{ZN}, \tag{52}$$

where M_u is introduced in Lemma 5.

Proof Firstly, for any $j = 0, \ldots, m-1$ and $s = 0, \ldots, S-1$, we have that

$$\begin{aligned}
\mathbb{E}\left[\mathbf{d}_j^{s+1}|\mathbf{x}_j^{s+1}\right] &= \mathbb{E}\left[\hat{\mathbf{G}}(\mathbf{x}_j^{s+1}, z^{s+1}, \tilde{\mathcal{B}}_j^{s+1}) - \hat{\mathbf{G}}(\mathbf{x}_0^{s+1}, z^{s+1}, \tilde{\mathcal{B}}_j^{s+1}) + \hat{G}(\mathbf{x}_0^{s+1}, \tilde{\mathcal{N}})|\mathbf{x}_j^{s+1}\right] \\
&= \mathbb{E}\left[\hat{\mathbf{G}}(\mathbf{x}_j^{s+1}, z^{s+1}, \tilde{\mathcal{B}}_j^{s+1})|\mathbf{x}_j^{s+1}\right] \\
&= \nabla \hat{F}_\mu(\mathbf{x}_j^{s+1}),
\end{aligned}$$

where it uses the fact that $\mathbb{E}_{z^{s+1}, \tilde{\mathcal{B}}_j^{s+1}}[\hat{\mathbf{G}}(\mathbf{x}_0^{s+1}, z^{s+1}, \tilde{\mathcal{B}}_j^{s+1})] = \mathbb{E}[\hat{G}(\mathbf{x}_0^{s+1}, \tilde{\mathcal{N}})]$ and the final equality stems from the unbiasedness of $\hat{\mathbf{G}}(\mathbf{x}_j^{s+1}, z^{s+1}, \tilde{\mathcal{B}}_j^{s+1})$ as described in (50). Secondly, we provide an upper bound of the variance $\mathbb{E}[\|\nabla \hat{F}_\mu(\mathbf{x}_j^{s+1}) - \mathbf{d}_j^{s+1}\|^2|\mathbf{x}_j^{s+1}]$ as follows:

$$\begin{aligned}
&\mathbb{E}\left[\|\nabla \hat{F}_\mu(\mathbf{x}_j^{s+1}) - \mathbf{d}_j^{s+1}\|^2|\mathbf{x}_j^{s+1}\right] \\
&= \mathbb{E}\left[\|\nabla \hat{F}_\mu(\mathbf{x}_j^{s+1}) - \left(\hat{\mathbf{G}}(\mathbf{x}_j^{s+1}, z^{s+1}, \tilde{\mathcal{B}}_j^{s+1}) - \hat{\mathbf{G}}(\mathbf{x}_0^{s+1}, z^{s+1}, \tilde{\mathcal{B}}_j^{s+1}) + \hat{G}(\mathbf{x}_0^{s+1}, \tilde{\mathcal{N}})\right)\|^2|\mathbf{x}_j^{s+1}\right] \\
&= \mathbb{E}\left[\|-\hat{\mathbf{G}}(\mathbf{x}_j^{s+1}, z^{s+1}, \tilde{\mathcal{B}}_j^{s+1}) + \hat{\mathbf{G}}(\mathbf{x}_0^{s+1}, z^{s+1}, \tilde{\mathcal{B}}_j^{s+1}) + \nabla \hat{F}_\mu(\mathbf{x}_j^{s+1}) - \nabla \hat{F}_\mu(\mathbf{x}_0^{s+1})\|^2|\mathbf{x}_j^{s+1}\right] \\
&\quad + \mathbb{E}\left[\|\hat{G}(\mathbf{x}_j^{s+1}, \tilde{\mathcal{N}}) - \hat{G}(\mathbf{x}_0^{s+1}, \tilde{\mathcal{N}}) - \nabla \hat{F}_\mu(\mathbf{x}_j^{s+1}) + \nabla \hat{F}_\mu(\mathbf{x}_0^{s+1})\|^2|\mathbf{x}_j^{s+1}\right] \\
&\quad + \mathbb{E}\left[\|\nabla \hat{F}_\mu(\mathbf{x}_j^{s+1}) - \hat{G}(\mathbf{x}_j^{s+1}, \tilde{\mathcal{N}})\|^2|\mathbf{x}_j^{s+1}\right] \\
&\leq \mathbb{E}\left[\|\hat{\mathbf{G}}(\mathbf{x}_j^{s+1}, z^{s+1}, \tilde{\mathcal{B}}_j^{s+1}) - \hat{\mathbf{G}}(\mathbf{x}_0^{s+1}, z^{s+1}, \tilde{\mathcal{B}}_j^{s+1})\|^2|\mathbf{x}_j^{s+1}\right] + \mathbb{E}\left[\|\hat{G}(\mathbf{x}_j^{s+1}, \tilde{\mathcal{N}}) - \hat{G}(\mathbf{x}_0^{s+1}, \tilde{\mathcal{N}})\|^2|\mathbf{x}_j^{s+1}\right] \\
&\quad + \mathbb{E}\left[\|\nabla \hat{F}_\mu(\mathbf{x}_j^{s+1}) - \hat{G}(\mathbf{x}_j^{s+1}, \tilde{\mathcal{N}})\|^2|\mathbf{x}_j^{s+1}\right] \\
&\leq \frac{2}{b}\mathbb{E}\left[\|\mathbf{g}_{\text{rand}}(\frac{z^{s+1}}{Z}\mathbf{x}_j^{s+1}, \nu_l, \xi_l) - \mathbf{g}_{\text{rand}}(\frac{z^{s+1}}{Z}\mathbf{x}_0^{s+1}, \nu_l, \xi_l)\|^2|\mathbf{x}_j^{s+1}\right] \\
&\quad + \mathbb{E}\left[\|\nabla \hat{F}_\mu(\mathbf{x}_j^{s+1}) - \hat{G}(\mathbf{x}_j^{s+1}, \tilde{\mathcal{N}})\|^2|\mathbf{x}_j^{s+1}\right] \\
&\leq \frac{2}{b}M_u^2\|\frac{z^{s+1}}{Z}(\mathbf{x}_j^{s+1} - \mathbf{x}_0^{s+1})\|^2 + \mathbb{E}\left[\|\nabla \hat{F}_\mu(\mathbf{x}_j^{s+1}) - \hat{G}(\mathbf{x}_j^{s+1}, \tilde{\mathcal{N}})\|^2|\mathbf{x}_j^{s+1}\right] \\
&\leq \frac{2M_u^2}{b}\|\mathbf{x}_j^{s+1} - \mathbf{x}_0^{s+1}\|^2 + \mathbb{E}\left[\|\nabla \hat{F}_\mu(\mathbf{x}_j^{s+1}) - \hat{G}(\mathbf{x}_j^{s+1}, \tilde{\mathcal{N}})\|^2|\mathbf{x}_j^{s+1}\right],
\end{aligned} \tag{53}$$

where the second equality arises from (50), the first inequality is due to $\mathbb{E}[\|\xi - \mathbb{E}[\xi]\|^2] \leq \mathbb{E}[\|\xi\|^2]$ for a random variable ξ, the second one is due to the definitions of $\hat{\mathbf{G}}(\mathbf{x}, z, \tilde{B})$ in (49) and $\hat{G}(\mathbf{x}, \tilde{\mathcal{N}})$ in (51), $e^{\frac{z}{Z}-1} \leq 1$ and $|\tilde{\mathcal{B}}_j^{s+1}| \leq |\tilde{\mathcal{N}}|$, the third inequality can be derived by Lemma 5, and the last one is due to $z^{s+1} \leq Z$. For the second term on the right-hand side

of (53), note that by the unbiasedness and upper boundedness of variance for $\bar{g}_{\rm rand}(\mathbf{x}, \tilde{\mathcal{B}})$ in (30),

$$\begin{aligned}
&\mathbb{E}\left[\|\nabla \hat{F}_\mu(\mathbf{x}_j^{s+1}) - \hat{G}(\mathbf{x}_j^{s+1}, \tilde{\mathcal{N}})\|^2 | \mathbf{x}_j^{s+1}\right] \\
&= \mathbb{E}\left[\|\frac{1}{Z}\sum_{z=1}^{Z} e^{\frac{z}{Z}-1}\nabla f_\mu(\frac{z}{Z}\mathbf{x}_j^{s+1}) - \frac{1}{Z}\sum_{z=1}^{Z} e^{\frac{z}{Z}-1}\bar{\mathbf{g}}_{\rm rand}(\frac{z}{Z}\mathbf{x}_j^{s+1}, \tilde{\mathcal{N}})\|^2 | \mathbf{x}_j^{s+1}\right] \\
&= \frac{1}{Z^2}\sum_{z=1}^{Z}(e^{\frac{z}{Z}-1})^2 \mathbb{E}\left[\|\nabla f_\mu(\frac{z}{Z}\mathbf{x}_j^{s+1}) - \bar{\mathbf{g}}_{\rm rand}(\frac{z}{Z}\mathbf{x}_j^{s+1}, \tilde{\mathcal{N}})\|^2 | \mathbf{x}_j^{s+1}\right] \\
&\leq \frac{16\sqrt{2\pi}L_0^2 d}{ZN}. \qquad (54)
\end{aligned}$$

Therefore, the proof is completed by plugging (54) into (53). ■

Lemma 22. *For any $j = 0, \ldots, m-1$ and $s = 0, \ldots S-1$ in* **NZOSA**, *it holds that*

$$\begin{aligned}
\hat{F}_\mu(\mathbf{x}_{j+1}^{s+1}) \geq \hat{F}_\mu(\mathbf{x}_j^{s+1}) + \frac{1}{2\eta_j^{s+1}}\left(\|\mathbf{x}_{j+1}^{s+1} - \mathbf{y}\|^2 - \|\mathbf{x}_j^{s+1} - \mathbf{y}\|^2\right) + \langle \mathbf{d}_j^{s+1}, \mathbf{y} - \mathbf{x}_j^{s+1}\rangle \\
- \eta_j^{s+1}\|\nabla \hat{F}_\mu(\mathbf{x}_j^{s+1}) - \mathbf{d}_j^{s+1}\|^2 - \left(\frac{L_\mu}{2} - \frac{1}{4\eta_j^{s+1}}\right)\|\mathbf{x}_{j+1}^{s+1} - \mathbf{x}_j^{s+1}\|^2 \qquad (55)
\end{aligned}$$

for any $\mathbf{y} \in \mathcal{X}$, where $L_\mu = \frac{\sqrt{d}L_0}{u}$.

Proof According to the L_μ-smoothness of $\hat{F}_\mu$ as shown in Lemma 20 (i), we can follow a similar analysis of Lemma 10 to obtain the conclusion. The details are omitted here. ■

Similarly with Lemma 11 and Lemma 16, we show an auxiliary lemma as follows. We denote $\bar{M}_u = \frac{\sqrt{2}dL_0}{u} \geq \frac{\sqrt{d}L_0}{u} = L_\mu$.

Lemma 23. *Given the parameters $a_m = 0$, $a_j = a_{j+1}\left(1 + \frac{1}{m}\right) + \frac{\eta \bar{M}_u^2}{b}$ and*

$$b = m^2 < N, \quad \eta = \frac{1}{c\bar{M}_u},$$

where $c = 4\sqrt{2}$, we have

$$\frac{\bar{M}_u}{2} + a_{j+1}(1+m) \leq \frac{1}{4\eta}.$$

Combined with Lemmas 21-23, we show the approximation performance for the algorithm **NZOSA** as follows.

Theorem 24. *Let the parameters for the computation of $\mathbf{d}_j^{s+1}$ and the step-size η_j^{s+1} in* **NZOSA** satisfy

$$b = m^2 < N, \quad u = \sqrt{\frac{d}{Sm}}, \quad \eta_j^{s+1} = \eta, \quad j = 0, \ldots, m-1; s = 0, \ldots S-1,$$

where $\eta = 1/c\bar{M}_u$ *with* $c = 4\sqrt{2}$ *and* $\bar{M}_u = \frac{\sqrt{2}dL_0}{u}$. *Return* $\mathbf{x}_r$ *according to that* $\mathrm{P}(\mathbf{x}_r = \mathbf{x}_j^{s+1}) = \frac{1}{Sm+1+\ln Z}$ *for* $j = 0, \ldots, m-1; s = 0, \ldots, S-1$, *and* $\mathrm{P}(\mathbf{x}_r = \mathbf{x}_m^S) = \frac{1+\ln Z}{Sm+1+\ln Z}$. *Then it holds that*

$$\mathbb{E}\left[f(\mathbf{x}_r)\right] \geq \left(1 - e^{-1} - \frac{3\ln Z}{Z} - \frac{\ln Z}{Sm + \ln Z}\right) f(\mathbf{x}^*) - \mathcal{O}\left(\frac{\sqrt{d}}{\sqrt{Sm}}\right),$$

where $\mathbf{x}^*$ *is the optimal solution of* (45).

Proof To prove the conclusion we introduce the following Lyapunov function

$$L_{j+1}^{s+1} := \mathbb{E}\left[\hat{F}_\mu(\mathbf{x}_j^{s+1}) - a_j\|\mathbf{x}_j^{s+1} - \mathbf{x}_0^{s+1}\|^2\right],$$

where the parameter $a_m = 0$ and $a_j = a_{j+1}\left(1 + \frac{1}{m}\right) + \frac{\eta\bar{M}_u^2}{b}$. In analogy to (24), with the settings $c = 4\sqrt{2}$ and $b = m^2$, we derive that

$$L_{j+1}^{s+1} \geq L_j^{s+1} + \frac{1}{2\eta}\mathbb{E}\left[\|\mathbf{x}_{j+1}^{s+1} - \mathbf{y}\|^2 - \|\mathbf{x}_j^{s+1} - \mathbf{y}\|^2\right] + \mathbb{E}\left[\langle \mathbf{d}_j^{s+1}, \mathbf{y} - \mathbf{x}_j^{s+1}\rangle\right] - \frac{16\sqrt{2\pi}L_0^2 d}{ZN}\eta, \tag{56}$$

where we use Lemmas 21-23. Furthermore, by summing (56) with $\mathbf{y} = \mathbf{x}^*$ and $\eta = 1/c\bar{M}_u$ over all the iterations in s-th epoch, we obtain from $\mathbb{E}[\mathbf{d}_j^{s+1}|\mathbf{x}_j^{s+1}] = \nabla\hat{F}_\mu(\mathbf{x}_j^{s+1})$ that

$$\begin{aligned} L_m^{s+1} \geq L_0^{s+1} &+ \frac{c\bar{M}_u}{2}\mathbb{E}\left[\|\mathbf{x}_m^{s+1} - \mathbf{x}^*\|^2 - \|\mathbf{x}_0^{s+1} - \mathbf{x}^*\|^2\right] \\ &+ \sum_{j=0}^{m-1}\mathbb{E}\left[\langle\nabla\hat{F}_\mu(\mathbf{x}_j^{s+1}), \mathbf{x}^* - \mathbf{x}_j^{s+1}\rangle\right] - \sum_{j=0}^{m-1}\frac{16\sqrt{2\pi}}{cZN\bar{M}_u}L_0^2 d. \end{aligned} \tag{57}$$

Furthermore, we have $L_m^{s+1} = \mathbb{E}[\hat{F}_\mu(\mathbf{x}_m^{s+1})]$ from $a_m = 0$ and the definition of $\mathbf{x}_0^{s+1}$, and $L_0^{s+1} = \mathbb{E}[\hat{F}_\mu(\mathbf{x}_0^{s+1})] = \mathbb{E}[\hat{F}_\mu(\mathbf{x}_m^s)]$, which follows from the setting $\mathbf{x}_0^{s+1} = \mathbf{x}_m^s$. Thus, (57) indicates from Lemma 20(iii) with $\mathbf{y} = \mathbf{x}^*$ and $\mathbf{x} = \mathbf{x}_j^{s+1}$ that

$$\begin{aligned} \mathbb{E}\left[\hat{F}_\mu(\mathbf{x}_m^{s+1})\right] \geq \mathbb{E}\left[\hat{F}_\mu(\mathbf{x}_m^s)\right] &+ \frac{c\bar{M}_u}{2}\mathbb{E}\left[\|\mathbf{x}_m^{s+1} - \mathbf{x}^*\|^2 - \|\mathbf{x}_0^{s+1} - \mathbf{x}^*\|^2\right] \\ &+ \sum_{j=0}^{m-1}\mathbb{E}\left[\left(1 - e^{-1}\right)f(\mathbf{x}^*) - (1 + \frac{3\ln Z}{Z})f(\mathbf{x}_j^{s+1}) - 3L_0 u\right] - \sum_{j=0}^{m-1}\frac{16\sqrt{2\pi}}{cZN\bar{M}_u}L_0^2 d. \end{aligned} \tag{58}$$

Adding the inequality (58) across all the epochs (for $s \in \{0, \ldots, S-1\}$), we obtain

$$\begin{aligned} &\mathbb{E}\left[\hat{F}_\mu(\mathbf{x}_m^S)\right] \\ &\geq \mathbb{E}\left[\hat{F}_\mu(\mathbf{x}_m^0)\right] + \frac{c\bar{M}_u}{2}\mathbb{E}\left[\|\mathbf{x}_m^S - \mathbf{x}^*\|^2 - \|\mathbf{x}_0^1 - \mathbf{x}^*\|^2\right] \\ &\quad + \sum_{s=0}^{S-1}\sum_{j=0}^{m-1}\mathbb{E}\left[\left(1 - e^{-1}\right)f(\mathbf{x}^*) - (1 + \frac{3\ln Z}{Z})f(\mathbf{x}_j^{s+1}) - 3L_0 u\right] - \sum_{s=0}^{S-1}\sum_{j=0}^{m-1}\frac{16\sqrt{2\pi}}{cZN\bar{M}_u}L_0^2 d \end{aligned}$$

$$
\begin{aligned}
&\geq -\frac{c\bar{M}_u}{2}\mathbb{E}\left[\|\mathbf{x}_0^1-\mathbf{x}^*\|^2\right]+\sum_{s=0}^{S-1}\sum_{j=0}^{m-1}\mathbb{E}\left[\left(1-e^{-1}\right)f(\mathbf{x}^*)-(1+\frac{3\ln Z}{Z})f(\mathbf{x}_j^{s+1})-3L_0u\right]\\
&\quad-\sum_{s=0}^{S-1}\sum_{j=0}^{m-1}\frac{16\sqrt{2\pi}}{cZN\bar{M}_u}L_0^2d\\
&\geq -\frac{c\bar{M}_u}{2}D^2+\sum_{s=0}^{S-1}\sum_{j=0}^{m-1}\mathbb{E}\left[\left(1-e^{-1}-\frac{3\ln Z}{Z}\right)f(\mathbf{x}^*)-f(\mathbf{x}_j^{s+1})-3L_0u\right]\\
&\quad-\sum_{s=0}^{S-1}\sum_{j=0}^{m-1}\frac{16\sqrt{2\pi}}{cZN\bar{M}_u}L_0^2d, \qquad (59)
\end{aligned}
$$

where the second inequality arises from the non-negativity of $\hat{F}_\mu(\mathbf{x})$ and $\|\mathbf{x}_m^S-\mathbf{x}^*\|^2$. Furthermore, we rearrange the inequality, utilizing with $f(\mathbf{x}^*)\geq f(\mathbf{x}_j^{s+1})$, yielding

$$
\begin{aligned}
&\mathbb{E}\left[\hat{F}_\mu(\mathbf{x}_m^S)+\sum_{s=0}^{S-1}\sum_{j=0}^{m-1}f(\mathbf{x}_j^{s+1})\right]\\
&\geq Sm\left(1-e^{-1}-\frac{3\ln Z}{Z}\right)f(\mathbf{x}^*)-\frac{c\bar{M}_uD^2}{2}-\sum_{s=0}^{S-1}\sum_{j=0}^{m-1}\left(3L_0u+\frac{16\sqrt{2\pi}}{cZN\bar{M}_u}L_0^2d\right). \qquad (60)
\end{aligned}
$$

Recalling Lemma 20(ii) and Lemma 4(iii), we have

$$\hat{F}_\mu(\mathbf{x})\leq(1+\ln Z)f_\mu(\mathbf{x})\leq(1+\ln Z)(f(\mathbf{x})+L_0u).$$

Combining this inequality with (60) leads to

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{s=0}^{S-1}\sum_{j=0}^{m-1}f(\mathbf{x}_j^{s+1})+(1+\ln Z)(f(\mathbf{x}_m^S)+L_0u)\right]\\
&\geq Sm\left(1-e^{-1}-\frac{3\ln Z}{Z}\right)f(\mathbf{x}^*)-\frac{c\bar{M}_uD^2}{2}-\sum_{s=0}^{S-1}\sum_{j=0}^{m-1}\left(3L_0u+\frac{16\sqrt{2\pi}}{cZN\bar{M}_u}L_0^2d\right).
\end{aligned}
$$

Then through dividing both sides of the above inequality by $(Sm+1+\ln Z)$, we obtain

$$
\begin{aligned}
\mathbb{E}\left[f(\mathbf{x}_r)\right]&=\mathbb{E}\left[\sum_{s=0}^{S-1}\sum_{j=0}^{m-1}\frac{1}{Sm+1+\ln Z}f(\mathbf{x}_j^{s+1})+\frac{1+\ln Z}{Sm+1+\ln Z}f(\mathbf{x}_m^S)\right]\\
&\geq\frac{Sm}{Sm+1+\ln Z}\left(1-e^{-1}-\frac{3\ln Z}{Z}\right)f(\mathbf{x}^*)\\
&\quad-\frac{\frac{c\bar{M}_uD^2}{2}+(3Sm+1+\ln Z)L_0u+\frac{16\sqrt{2\pi}}{cZN\bar{M}_u}L_0^2d}{Sm+1+\ln Z}\\
&\geq\left((1-e^{-1}-\frac{3\ln Z}{Z})\left(1-\frac{1+\ln Z}{Sm+1+\ln Z}\right)\right)f(\mathbf{x}^*)
\end{aligned}
$$

$$
\begin{aligned}
&- \frac{\frac{cD^2L_0d}{\sqrt{2}u} + (1+\ln Z)L_0u + Sm\left(3+\frac{16\sqrt{\pi}}{cZN}\right)L_0u}{Sm+1+\ln Z} \\
&\geq \left(1-e^{-1}-\frac{3\ln Z}{Z}-\frac{\left(1-e^{-1}\right)(1+\ln Z)}{Sm+1+\ln Z}\right)f(\mathbf{x}^*) - \mathcal{O}\left((Sm)^{-1/2}d^{1/2}\right),
\end{aligned}
$$

where $\bar{M}_u = \frac{\sqrt{2}dL_0}{u}$, $u = \sqrt{\frac{d}{Sm}}$ and $c = 4\sqrt{2}$. Therefore, we derive the conclusion. ■

It is worth noting that the approximation ratio $(1 - e^{-1} - \frac{3\ln Z}{Z} - \frac{\ln Z}{Sm+\ln Z})$ in Theorem 24 depends on the parameter Z. Clearly, the value of Z offers a trade-off between the approximation ratio and complexity order.

Corollary 25. *Under the same conditions as Theorem 24 with additional setting that* $b = N^{2/3}$, $Z = \lceil\sqrt{Sm}\rceil$ *and* $\epsilon > 0$, *the* **NZOSA** algorithm achieves

$$
\mathbb{E}\left[f(\mathbf{x}_r)\right] \geq \left(1 - e^{-1} - \epsilon\ln\epsilon^{-1} - \epsilon^2\ln\epsilon^{-1}\right)f(\mathbf{x}^*) - \epsilon, \tag{61}
$$

after $\mathcal{O}(d\epsilon^{-2})$ iterations and $\mathcal{O}(N^{2/3}d^{3/2}\epsilon^{-3})$ function evaluations.

Proof To achieve ϵ-accuracy as shown in (61), we let $\sqrt{d}/\sqrt{Sm} = \mathcal{O}(\epsilon)$ in Theorem 24. It is clear that $Sm = \mathcal{O}(d/\epsilon^2)$, and thus $Z = \mathcal{O}(\sqrt{d}/\epsilon)$. With $b = N^{2/3}$, we have

$$
m = \sqrt{b} = N^{1/3}, \quad S = \frac{d/\epsilon^2}{m} = \mathcal{O}\left(\frac{d}{N^{1/3}\epsilon^2}\right).
$$

Thus, the total number of zeroth-order oracle calls in inner iterations and outer iterations is

$$
4b \times Sm = \mathcal{O}\left(\frac{dN^{2/3}}{\epsilon^2}\right), \quad Z \times 2SN = \mathcal{O}\left(\frac{d^{3/2}N^{2/3}}{\epsilon^3}\right),
$$

respectively. Therefore, we obtain the final oracle complexity. The proof is completed. ■

5. Application to Robust DR-Submodular Maximization

In this section, we focus on the robust DR-submodular maximization problem (2). We further assume that $\mathbf{f}_i(\cdot, \xi_{i,t}) : \mathbb{R}^d \to \mathbb{R}$, $i \in [M]$ are L_0-Lipschitz continuous and L-smooth DR-submodular functions. By Lemma 28 in Appendix E, we demonstrate that the objective function of (2), i.e., $\mathbf{f}(\cdot, \xi_t) := \min_{i\in[M]} \mathbf{f}_i(\cdot, \xi_{i,t})$, is non-negative, monotonically non-decreasing, up-concave and L_0-Lipschitz continuous. Then the problem (2) belongs to the class of non-smooth up-concave maximization problems, allowing for the application of **NZOSA** to solve it. However, to better adapt to the structure of (2), we propose a specific approach for computing the mini-batch zeroth-order gradient estimator in Algorithm 2. In Algorithm 2, to compute the randomized gradient estimator $\mathbf{g}_{\text{rand}}(\mathbf{x}, \nu_t, \xi_t)$, two optimal indices for the stochastic function at vectors $(\mathbf{x} + u\nu)$ and $(\mathbf{x} - u\nu)$ are chosen respectively. With zeroth-order oracles queried, the approximate gradient is defined as

$$
\mathbf{g}_{\text{rand}}(\mathbf{x}, \nu_l, \xi_l) = \frac{d}{2u}\left(\min_{i\in[M]} \mathbf{f}(\mathbf{x} + u\nu_l, \xi_{i,l}) - \min_{i\in[M]} \mathbf{f}(\mathbf{x} - u\nu_l, \xi_{i,l})\right)\nu_l
$$

Algorithm 2 Mini-batch Zeroth-order Gradient Estimator for Problem (2)

Input: Point $\mathbf{x} \in \mathbb{R}^n$, sample radius $u \in \mathbb{R}_+$, mini-batch $\tilde{\mathcal{B}} = \{\nu_l, \xi_l\}_{l \in B}$, where $B \subseteq [N]$, and $\mathbf{f}_i(\mathbf{x}, \xi_{i,t}), i \in [M], t \in [N]$.

1: **for** $l \in B$ **do**

2: Compute

$$\mathbf{g}_{\text{rand}}(\mathbf{x}, \nu_l, \xi_l) = \frac{d}{2u}(\mathbf{f}_{i_+}(\mathbf{x} + u\nu_l, \xi_{i_+,l}) - \mathbf{f}_{i_-}(\mathbf{x} - u\nu_l, \xi_{i_-,l}))\nu_l,$$

where $i_+ \in \arg\min_{i \in [M]} \mathbf{f}_i(\mathbf{x} + u\nu_l, \xi_{i,l})$ and $i_- \in \arg\min_{i \in [M]} \mathbf{f}_i(\mathbf{x} - u\nu_l, \xi_{i,l})$.

3: **end for**

4: Return $\bar{\mathbf{g}}_{\text{rand}}(\mathbf{x}, \tilde{\mathcal{B}}) = \frac{1}{|B|} \sum_{l \in B} \mathbf{g}_{\text{rand}}(\mathbf{x}, \nu_l, \xi_l)$

$$= \frac{d}{2u} \left(\mathbf{f}(\mathbf{x} + u\nu_l, \xi_l) - \mathbf{f}(\mathbf{x} - u\nu_l, \xi_l)\right) \nu_l.$$

Algorithm 2 returns an averaged stochastic approximation $\bar{\mathbf{g}}_{\text{rand}}(\mathbf{x}, \tilde{\mathcal{B}})$. It indicates from the L_0-Lipschitz continuity of $\mathbf{f}(\mathbf{x}, \xi_t)$ and Lemma 5 that

$$\mathbb{E}_{\tilde{\mathcal{B}}}[\bar{\mathbf{g}}_{\text{rand}}(\mathbf{x}, \tilde{\mathcal{B}})] = \frac{1}{N} \sum_{t=1}^{N} \nabla \mathbf{f}_\mu(\mathbf{x}, \xi_t),$$

where $\mathbf{f}_\mu(\mathbf{x}, \xi_t)$ is the smoothed function of $\mathbf{f}(\mathbf{x}, \xi_t)$.

For solving (2), we apply the **NZOSA** algorithm, which incorporates Algorithm 2 to compute zeroth-order gradient estimates. The main theoretical results are presented in the following theorem.

Theorem 26. *Given the computation of* $\mathbf{d}_j^{s+1}$ *in* (51)*, let the parameters in Algorithm 2 and the step-size* η_j^{s+1} *in* **NZOSA** *satisfy*

$$u = \sqrt{\frac{d}{Sm}}, \quad b = N^{2/3}, \quad m = N^{1/3}, \quad Z = \lceil \sqrt{Sm} \rceil, \quad \eta_j^{s+1} = \frac{1}{4\sqrt{2}\bar{M}_u},$$

where $\bar{M}_u = L_0 d/u$*. Return* $\mathbf{x}_r$ *according to that* $\mathrm{P}(\mathbf{x}_r = \mathbf{x}_j^{s+1}) = \frac{1}{Sm+1+\ln Z}$ *for* $j = 0, \ldots, m-1; s = 0, \ldots, S-1$*, and* $\mathrm{P}(\mathbf{x}_r = \mathbf{x}_m^S) = \frac{1+\ln Z}{Sm+1+\ln Z}$*. Then for a given* $\epsilon > 0$*, it achieves that*

$$\frac{1}{N} \sum_{t=1}^{N} \mathbb{E}\left[\mathbf{f}(\mathbf{x}_r, \xi_t)\right] \geq \left(1 - e^{-1} - \epsilon \ln \epsilon^{-1} - \epsilon^2 \ln \epsilon^{-1}\right) \mathrm{OPT} - \epsilon \tag{62}$$

after $\mathcal{O}(MN^{2/3}d^{3/2}\epsilon^{-3})$ *function value evaluations and* $\mathcal{O}(d\epsilon^{-2})$ *iterations, where* OPT *denotes the optimal value of problem* (2)*.*

Proof Note that each gradient computation requires $2M$ zeroth-order oracles, as shown in Algorithm 2. Similar to Corollary 25, to achieve (62) the total number of calls to the zeroth-order oracle in inner iterations and in outer iterations is

$$Sm \times 4Mb = \mathcal{O}\left(\frac{N^{2/3}Md}{\epsilon^2}\right) \quad \text{and} \quad Z \times 2SNM = \mathcal{O}\left(\frac{\sqrt{d}}{\epsilon} \times \frac{N^{2/3}dM}{\epsilon^2}\right),$$

respectively. Therefore, the total oracle complexity is in order $\mathcal{O}(MN^{2/3}d^{3/2}\epsilon^{-3})$. ■

Remark 27. *Compared to the mirror-prox* **(MP)** algorithm proposed in (Lee et al., 2022) which is a first-order method that uses approximations to sub-gradients, our proposed zeroth-order approximation algorithm improves the approximation ratio from $(1/2-\epsilon)$ to $(1-e^{-1}-\epsilon\ln\epsilon^{-1}-\epsilon^2\ln\epsilon^{-1})$ for robust DR-submodular maximization with only zeroth-order information available. But on the other hand, as shown in Table 2, the iteration and oracle complexity of **NZOSA** is higher than **MP**. We note that the reason for the high complexity is not due to zeroth-order estimation errors because replacing the zeroth-order gradient estimation with a stochastic gradient does not affect the final complexity. Specifically, assume $\mathbb{E}[\|\nabla f_\mu(x)-g_{\text{rand}}\|^2|x]\leq C$, where g_{rand} is the gradient estimation and C is a constant. Similar to (54), the error bound in Lemma 21 can still be described as

$$\mathbb{E}\left[\|\nabla\hat{F}_\mu(\mathbf{x}_j^{s+1})-\mathbf{d}_j^{s+1}\|^2|\mathbf{x}_j^{s+1}\right]\leq\frac{2M_u^2}{b}\|\mathbf{x}_j^{s+1}-\mathbf{x}_0^{s+1}\|^2+\frac{C}{Z}.$$

Thus, the total complexity will keep the same in this case. We further emphasize that the nature of the relatively higher complexity for **NZOSA** is caused by the smoothing technique and the algorithmic framework. The smoothing parameter of the smoothed function makes an impact on the iteration complexity through the first term above, while the finite-sum structure of the auxiliary function and the requirement to compute a full gradient with Z components in each outer iteration contributes to the oracle complexity. Notably, although the complexity is relatively higher, **NZOSA** owns a better approximation ratio. To compare approximation performances of the related zeroth-order algorithms, we replace the sub-gradient in **MP** with randomized gradient estimation leading to **ZO-MP**. There is an inherent trade-off in a fair comparison of these algorithms, and we will demonstrate their performances in next section through multi-resolution data summarization and robust budget allocation problem.

6. Numerical Experiments

In this section, we report numerical results evaluating the performance of our algorithms under various settings. We implement all the codes of competitor algorithms and our proposed algorithms in this. All experiments were implemented in PyCharm 2024.1 x64 using Python 3.10.9. We compare our proposed algorithms (**CG-ZOSA** and **RG-ZOSA)** with the baseline algorithms for the smooth case as follows:

- **BCG**: Black-box continuous greedy algorithm, a zeroth-order algorithm proposed in (Chen et al., 2020b) for solving monotone DR-submodualr maximization where only the function value is available.

- **ZO-GA**: Zeroth-order gradient ascent algorithm (Algorithm 3 in Appendix B), obtained by replacing the stochastic gradient in algorithm **GA** (Hassani et al., 2017) with the randomized gradient estimator as described in (29). We choose the iterate step-size $\eta_k=\frac{1}{\sqrt{k+1}}$ in the experiments.

- **FW**: Generalized DR-submodular Frank-Wolfe algorithm, which uses the black box gradient estimate (BBGE) proposed in (Pedramfar et al., 2023). We choose the parameter $\rho_n = 2/(n+3)^{2/3}$ and the step-size $\epsilon = \log(N)/N$, where n is the iteration index and N is the total number of iterations.

- **ZO-SPGD**: Zeroth-order stochastic projected gradient descent, proposed in (Liu et al., 2018b) for constrained non-convex optimization where only objective function values are available. We choose the parameter $\eta = 0.1$.

- **Free-FW**: Stochastic gradient free Frank-Wolfe algorithm, proposed in (Sahu et al., 2019) for general non-convex optimization. We choose the parameter $\rho_t = \frac{4}{(1+d)^{1/3}(t+8)^{2/3}}$ and the step-size $\gamma_t = 1/T^{3/4}$, where t is the iteration index, T is the total number of iterations and d denotes the problem dimension.

Although the theoretical guarantee requires a constant step-size of **RG-ZOSA** in our analysis, in order to achieve better algorithm performance we choose varying step-sizes in the implementation of Algorithm **RG-ZOSA**. It is noteworthy that the first three algorithms are tailored for non-convex optimization with DR-submodular structure and have quantified approximation guarantees of the optimal value. In contrast, the last two general stochastic algorithms are proposed to pursue the stationary point for general non-convex optimization. It is expected that the specific approximation methods for DR-submodular functions will perform better compared to the general non-convex algorithms. Furthermore, we compare **NZOSA** with the following algorithm for solving a non-smooth up-concave optimization problem and a robust submodular maximization problem.

- **ZO-MP**: Zeroth-order mirror-prox algorithm, obtained by replacing the stochastic sub-gradient oracle with stochastic zeroth-order oracle in the algorithm proposed in (Lee et al., 2022), with step-size $\eta = \frac{1}{2\sqrt{K}}$.

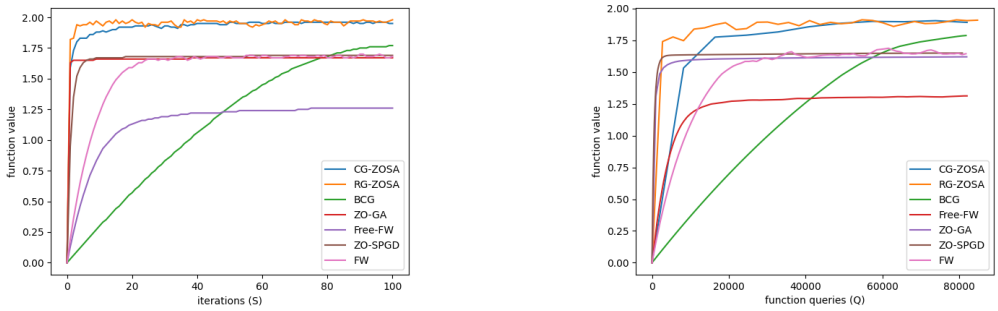
To report the numerical performances of algorithms, we present two types of figures: (objective) function value vs. iterations and (objective) function value vs. (oracle) queries. Each experiment is repeated 10 times, and the average value of the data is calculated for comparison.

Quadratic programming. In this setting, we apply the baselines and our algorithms to maximize the quadratic objective

$$f(\mathbf{x}) = \frac{1}{N}\sum_{t=1}^{N}\left\{\mathbf{f}(\mathbf{x},\xi_t) := \frac{1}{2}\mathbf{x}^T\mathbf{H}_t\mathbf{x} + h_t^T\mathbf{x}\right\},$$

where the symmetric matrix $\mathbf{H}_t \in \mathbb{R}^{d\times d}$ is randomly generated with entries $(\mathbf{H}_t)_{ij}$ uniformly distributed in $[-1,0]$, and the constraint set $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^d | \mathbf{A}\mathbf{x} \leq b, \mathbf{0} \leq \mathbf{x} \leq b, \mathbf{A} \in \mathbb{R}_+^{m\times d}, b \in \mathbb{R}_+^m\}$, with entries $\mathbf{A}_{ij}$ uniformly distributed in $[0,1]$. In our experiments, we set $N = 500$, $d = 3, m = 2, b = u = \mathbf{1}$ and $h_t = -\mathbf{H}_t^T u$, which guarantees the monotonicity and non-negativity of $\mathbf{f}(\mathbf{x},\xi_t)$.

Figure 2 presents a comparison of the numerical performances of the algorithms of interest. It is evident that our proposed algorithms, **CG-ZOSA** and **RG-ZOSA**, outperform



(a) Function value vs. iterations

(b) Function value vs. queries

Figure 2: Comparison of different zeroth-order algorithms for quadratic programming

all other algorithms in terms of reachable function value, projection-free algorithms (**FW**, **Free-FW** and **BCG**). Firstly, the maximum achievable function values of **ZO-GA** and **ZO-SPGD** are lower than those of our proposed algorithms, even though they require only a few additional oracle queries. In addition, the maximum function value achieved by **ZO-GA** is consistent with the theoretic guarantee described in Subsection 2.3. Secondly, our algorithms achieve higher function values within the same number of iterations and oracle queries compared to **BCG**. Notably, while **BCG** gradually approaches a similar function value as ours with increasing iteration numbers, it is significantly slower than our algorithms, as shown on Figure 2 (b). This observation validates, to a considerable extent, the theoretical analysis presented in Table 1, where **BCG** possess same (or similar) approximation ratios as ours but much higher complexities for small ϵ. Lastly, the algorithms for general non-convex optimization show poorer performance than the approximation algorithms with the same type for DR-submodular optimization, such as **Free-FW** vs. **FW**.

Multi-resolution data summarization. We consider the multi-resolution data summarization model which maximizes the utility function

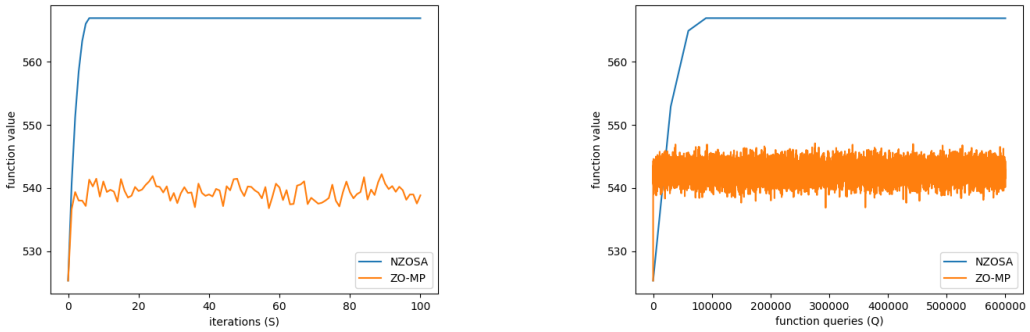
$$f(\mathbf{x}) = \frac{1}{N}\sum_{t=1}^{N}\left\{ f(\mathbf{x},\xi_t) := \sum_{i=1}^{d}\sum_{j=1}^{d}\phi(\mathbf{x}_j)s_{ij}^t - \sum_{i=1}^{d}\sum_{j=1}^{d}\mathbf{x}_i\mathbf{x}_j s_{ij}^t \right\}$$

over the set $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^d : 0 \leq \mathbf{x}_i \leq 1/2, i \in [1, \lceil d/5 \rceil]; 0 \leq \mathbf{x}_i \leq 1, i \in [\lceil d/5 \rceil + 1, d]; \sum_{i=1}^{d} \mathbf{x}_i \leq \frac{d}{3}\}$, where $\phi(\mathbf{x}_i)$ is given as the following piece-wise linear function defined as

$$\phi(\mathbf{x}_i) = \begin{cases} -3(\frac{1}{2})^{\mathbf{x}_i} + 4 & \text{if } \mathbf{x}_i \in [0, \frac{1}{2}], \\ -2(\frac{1}{2})^{\mathbf{x}_i} + 4 - (\frac{1}{2})^{\frac{1}{2}} & \text{if } \mathbf{x}_i \in [\frac{1}{2}, \frac{3}{4}], \\ -(\frac{1}{2})^{\mathbf{x}_i} + 4 - (\frac{1}{2})^{\frac{1}{2}} - (\frac{1}{2})^{\frac{3}{4}} & \text{if } \mathbf{x}_i \in [\frac{3}{4}, 1]. \end{cases}$$

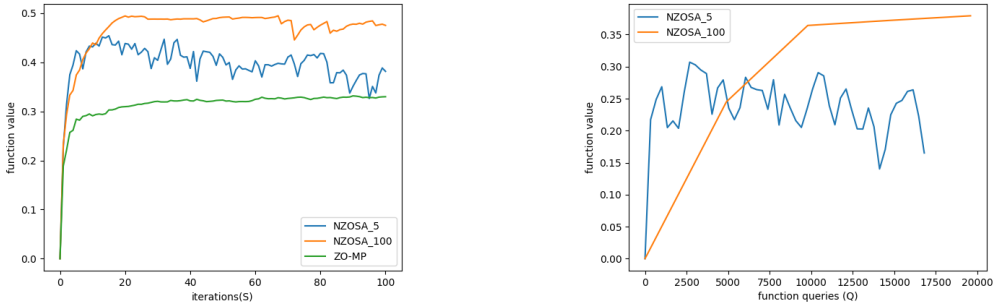
It is obvious that the function $\phi(\mathbf{x}_i)$ is concave. Therefore, the function $f(\mathbf{x})$ is up-concave. In our experiment, we sample N groups of similarity indices, $s_{ij}^t, t \in [N]$, from a uniform distribution on the interval $[0, 1]$, and we set $N = 1000$ and $d = 20$ in our experiments.

As shown in Figure 3, in the early stages of the algorithm, when the number of oracle queries is relatively small, **ZO-MP** achieves a higher function value. However, as the



(a) Function value vs. iterations (b) Function value vs. queries

Figure 3: Performance of our proposed algorithm for multi-resolution data summarization



(a) Function values vs. iterations (b) Function values vs. queries

Figure 4: Performance of our proposed algorithm for the robust budget allocation problem

algorithm progresses and the number of oracle queries becomes large, **NZOSA** reaches higher function values.

Robust budget allocation. In this experiment, we consider the robust optimal budget allocation problem (3) with $\alpha_i = 1/N$, and we set the constraint set as $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^d | 0 \leq \mathbf{x}_i \leq c_i, \sum_{i=1}^d \mathbf{x}_i \leq (\sum_{i=1}^d c_i)/3\}$. In our experiment, we set c_i is randomly chosen from $[2, 10]$, $N = 200$, $|S| = 20$, $|T| = 24$ (kon, 2017) and randomly sample $\{p_{st}\}_{(s,t)\in W}$ following a uniform distribution on $[0, 1]$. Moreover, we present algorithms' performances with different settings of the parameter $Z \in \{5, 100\}$ in Figure 4.

Observing Figure 4, we note that our proposed algorithm outperforms **ZO-MP**. Furthermore, as Z increases, the algorithm's approximation performance also improves, with a greater number of function values to be accessed.

7. Conclusion

In this paper, we studied stochastic approximation methods for DR-submodular optimization based on zeroth-order gradient estimations. Falling into a generic algorithm framework, specific algorithms tailored for three types of DR-submodular optimization problems were proposed. For the smooth DR-submodular maximization problem, we proposed two algorithms, which compute stochastic approximate gradients of an integral auxiliary function based on coordinate-wise gradient estimator and randomized gradient estimator, respectively. We established the approximation guarantees of both algorithms with iteration and oracle complexities being analyzed. We then presented a zeroth-order stochastic approximation method for non-smooth up-concave maximization based on a finite-sum auxiliary function. We then extended the algorithm to solve a class of robust DR-submodular maximization problems. Finally, numerical experiments were conducted to validate the effectiveness and efficiency of proposed algorithms.

Acknowledgements

This work was partially supported by National Natural Science Foundation of China (Nos. 12131003 and 12271278) and the Major Key Project of PCL (No. PCL2022A05) and the Talent Program of Guangdong Province (No. 2021QN02X160). Part of the work was done by the first author during her academic visit in Pengcheng Laboratory. We would like to thank the action editor and the anonymous referees for their helpful comments and suggestions on an earlier version of this paper.

Appendix A. Proofs of lemmas in Section 2

In this section, we present the technical proofs for the main lemmas stated in Section 2.

A.1 Proof of Lemma 3

Proof According to the definition of $\mathbf{g}_{\text{coord}}(\mathbf{x}, \xi_t)$ and mean value theorem, for any given $u > 0$, and there exists $\alpha_l \in (0, 1), l \in \{1, \ldots, d\}$, we have

$$\begin{aligned}
\|\mathbf{g}_{\text{coord}}(\mathbf{x}, \xi_t) - \nabla \mathbf{f}(\mathbf{x}, \xi_t)\|^2 &= \left\| \frac{1}{2u} \sum_{l=1}^{d} 2u \mathbf{e}_l \mathbf{e}_l^T \nabla \mathbf{f} \left((\mathbf{x} - u\mathbf{e}_l) + 2\alpha_l u \mathbf{e}_l, \xi_t \right) - \nabla \mathbf{f}(\mathbf{x}, \xi_t) \right\|^2 \\
&= \left\| \sum_{l=1}^{d} \left(\mathbf{e}_l \mathbf{e}_l^T (\nabla \mathbf{f}(\mathbf{x} + (2\alpha_l - 1)u\mathbf{e}_l, \xi_t) - \nabla \mathbf{f}(\mathbf{x}, \xi_t)) \right) \right\|^2 \\
&\leq \sum_{l=1}^{d} \|\nabla \mathbf{f}(\mathbf{x} + (2\alpha_l - 1)u\mathbf{e}_l, \xi_t) - \nabla \mathbf{f}(\mathbf{x}, \xi_t)\|^2 \\
&\leq L^2 \sum_{l=1}^{d} \|(2\alpha_l - 1)u\mathbf{e}_l\|^2 \leq L^2 d u^2,
\end{aligned}$$

where the first inequality follows from the definition of basis vector $\mathbf{e}_l$, and the second inequality is due to L-smoothness of function $\nabla \mathbf{f}(\mathbf{x}, \xi)$. ■

A.2 Proof of Lemma 4

Proof (i) It is straightforward to obtain from the monotonicity and the nonnegativity of $f(\mathbf{x})$ that $f_\mu(\mathbf{x})$ is monotone and non-negative. We now prove that $f_\mu(\mathbf{x})$ is up-concave. Since $f(\mathbf{x})$ is up-concave, for any $\mathbf{x} \in \mathbb{R}^d$ and a non-negative direction $\mathbf{v} \geq \mathbf{0}$, $f(\mathbf{x} + t\mathbf{v})$ is concave with respect with to t, i.e.,

$$f(\mathbf{x} + (\lambda t_1 + (1-\lambda)t_2)\mathbf{v}) \geq \lambda f(\mathbf{x} + t_1\mathbf{v}) + (1-\lambda) f(\mathbf{x} + t_2\mathbf{v}).$$

Then it holds that

$$\begin{aligned} f_\mu(\mathbf{x} + (\lambda t_1 + (1-\lambda)t_2)\mathbf{v}) &= \mathbb{E}_v\left[f\left(\mathbf{x} + uv + (\lambda t_1 + (1-\lambda)t_2)\,\mathbf{v}\right)\right] \\ &\geq \mathbb{E}_v[\lambda f(\mathbf{x} + uv + t_1\mathbf{v}) + (1-\lambda) f(\mathbf{x} + uv + t_2\mathbf{v})] \\ &= \lambda f_\mu(\mathbf{x} + t_1\mathbf{v}) + (1-\lambda) f_\mu(\mathbf{x} + t_2\mathbf{v}), \end{aligned}$$

which shows that $f_\mu(\mathbf{x} + t\mathbf{v})$ is concave with respect to t, and thus $f_\mu(\mathbf{x})$ is up-concave by Definition 2. Similarly, the function $f_\mu(\mathbf{x})$ is DR-submodular if the function $f(\mathbf{x})$ is DR-submodular.

(ii) If the function is L-smooth then we obtain that $f_\mu(\mathbf{x})$ is L_μ-smooth with $L_\mu \leq L$. Thus, the function $f_\mu(\mathbf{x})$ is L-smooth.

(iii) The function $f(\mathbf{x})$ is almost everywhere differentiable because its L_0-Lipschitz continuity, and then the differentiability of $f_\mu(\mathbf{x})$ can be derived following (Bertsekas, 1973). Furthermore, we have $\nabla f_\mu(\mathbf{x}) = \nabla \mathbb{E}_v[f(\mathbf{x} + uv)] = \mathbb{E}_v[g(\mathbf{x} + uv)]$, where $g(\mathbf{x} + uv) \in \partial^\uparrow f(\mathbf{x} + uv), v \in \mathbb{B}_d(\mathbf{0}, 1)$ and $u \in \mathbb{R}_+$. We can notice that

$$\begin{aligned} \|\nabla f_\mu(\mathbf{x}) - \nabla f_\mu(\mathbf{y})\|_2 &\leq \|\mathbb{E}_v[g(\mathbf{x} + uv)] - \mathbb{E}_v[g(\mathbf{y} + uv)]\|_\infty \\ &\leq \frac{2}{(2u)^d} L_0 (2u)^{d-1} \|\mathbf{x} - \mathbf{y}\|_1 \\ &\leq \frac{L_0\sqrt{d}}{u} \|\mathbf{x} - \mathbf{y}\|_2, \end{aligned}$$

where the second inequality is from (Duchi et al., 2012, Lemmas 11 and 12). Moreover, by the definition of $f_\mu(\mathbf{x})$ and L_0-Lipschitz continuous property of $f(\mathbf{x})$, we have

$$|f_\mu(\mathbf{x}) - f(\mathbf{x})| = |\mathbb{E}_v[f(\mathbf{x} + uv)] - f(\mathbf{x})| \leq L_0 u \mathbb{E}_v[\|v\|_2] \leq L_0 u,$$

where $v \in \mathbb{B}_d(\mathbf{0}, 1)$ and $\|v\|_2 \leq 1$. The proof is completed. ■

A.3 Proof of Lemma 5

Proof According to the definition of $\mathbf{g}_{\text{rand}}(\mathbf{x}, \nu, \xi_t)$ with fixed ξ_t, we observe that

$$\mathbb{E}\left[\mathbf{g}_{\text{rand}}(\mathbf{x}, \nu, \xi_t)|\mathbf{x}\right] = \mathbb{E}\left[\frac{d}{2u}\left(\mathbf{f}(\mathbf{x} + u\nu, \xi_t) - \mathbf{f}(\mathbf{x} - u\nu, \xi_t)\right)\nu|\mathbf{x}\right]$$

$$= \frac{1}{2}\mathbb{E}\left[\frac{d}{u}\mathbf{f}(\mathbf{x}+u\nu,\xi_t)\nu + \frac{d}{u}\mathbf{f}(\mathbf{x}+u(-\nu),\xi_t)(-\nu)|\mathbf{x}\right]$$
$$= \nabla \mathbf{f}_\mu(\mathbf{x},\xi_t).$$

We use the Young's inequality and the L_0-Lipschitz continuity of $\mathbf{f}(\mathbf{x},\xi_t)$, obtaining

$$\begin{aligned}
&\|\mathbf{g}_{\text{rand}}(\mathbf{x},\nu,\xi_t) - \mathbf{g}_{\text{rand}}(\mathbf{y},\nu,\xi_t)\|^2 \\
&\leq \frac{d^2}{2u^2}\left(\|\mathbf{f}(\mathbf{x}+u\nu,\xi_t) - \mathbf{f}(\mathbf{y}+u\nu,\xi_t)\|^2 + \|\mathbf{f}(\mathbf{x}-u\nu,\xi_t) - \mathbf{f}(\mathbf{y}-u\nu,\xi_t)\|^2\right) \\
&\leq \frac{d^2L_0^2}{u^2}\|\mathbf{x}-\mathbf{y}\|^2.
\end{aligned}$$

It remains to show that $\mathbb{E}\|\nabla f_\mu(\mathbf{x},\xi_t) - \mathbf{g}_{\text{rand}}(\mathbf{x},\nu,\xi_t)\|^2|\mathbf{x}] \leq 16\sqrt{2\pi}dL_0^2$. Using the inequality

$$\mathbb{E}\left[\|\mathbf{g}_{\text{rand}}(\mathbf{x},\nu,\xi_t)\|^2|\mathbf{x}\right] \leq 16\sqrt{2\pi}dL_0^2,$$

which has been proved in (Lin et al., 2022, Lemma D.1), we have

$$\begin{aligned}
\mathbb{E}\left[\|\nabla f_\mu(\mathbf{x},\xi_t) - \mathbf{g}_{\text{rand}}(\mathbf{x},\nu,\xi_t)\|^2|\mathbf{x}\right] &\leq \mathbb{E}\left[\|\mathbf{g}_{\text{rand}}(\mathbf{x},\nu,\xi_t)\|^2|\mathbf{x}\right] \\
&\leq 16\sqrt{2\pi}dL_0^2,
\end{aligned}$$

where the first inequality resulted from that $\mathbb{E}[\|\mathbb{E}[\xi] - \xi\|^2] \leq \mathbb{E}[\|\xi\|^2]$. The proof is completed. ■

Appendix B. Approximation analysis of ZO-GA algorithm

The first-order stochastic gradient ascent (GA) algorithm is studied for stochastic DR-maximization problem under convex constraint sets in (Hassani et al., 2017). This algorithm has been proved to reach 1/2 approximate ratio with $\mathcal{O}(1/\epsilon^2)$ stochastic gradient evaluations. In this section, we provide a zeroth-order variant of GA, named as ZO-GA, for smooth DR-submodular optimization (1). We adopt the randomized gradient estimation as described in Section 2.2 and present the algorithm framework as follows.

Algorithm 3 Zeroth-order Gradient Ascent (**ZO-GA**) Algorithm

Input: Initial point $\mathbf{x}_1 \in \mathcal{X}$, $K \in \mathbb{N}_+$ and step-sizes $\{\eta_k\}, k \in [K]$, radius u and batch size B.

Output: $\mathbf{x}_r$

1: **for** $k = 1, \dots K$ **do**
2: $\mathbf{y}_{k+1} = \mathbf{x}_k + \eta_k g_k$, where g_k is computed by the mini-batch randomized gradient estimation $\bar{\mathbf{g}}_{\text{rand}}(\mathbf{x}_k, \tilde{\mathcal{B}})$ in (29)
3: $\mathbf{x}_{k+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x} - \mathbf{y}_{k+1}\|_2$
4: **end for**

We now discuss the approximation behavior of ZO-GA for solving (10), where $f(\mathbf{x},\xi_t), t \in [N]$ are L-smooth and L_0-Lipschitz continuous, and f is monotone and DR-submodular. We

set parameters in ZO-GA satisfying

$$\eta_k = \frac{1}{L + \frac{8\sqrt{d}L_0}{R}\sqrt{k}}, \quad u = \frac{1}{\sqrt{K}}, \quad B = 1.$$

We analyze the approximation performance through two key points combining with existing results in (Hassani et al., 2017). Firstly, in ZO-GA the zeroth-order approximate gradient g_k is an unbiased estimation of the smoothed function f_μ with variance bounded by $\sigma^2 := 16\sqrt{2\pi}dL_0^2$ by Lemma 5. Then, following Theorem 4.3 in Hassani et al. (2017), we have that

$$\mathbb{E}[f_\mu(\mathbf{x}_r)] \geq \frac{\mathrm{OPT}}{2} - \left(\frac{R^2 L + \mathrm{OPT}}{2K} + \frac{R\sigma}{\sqrt{K}}\right),$$

where OPT denotes the optimal value of problem (1). Furthermore, by lemma 4 (iii) the error between the smoothed function f_μ and the function f is bounded by $L_0 u$, which is in order $\mathcal{O}(1/\sqrt{K})$. As a result, we obtain

$$\mathbb{E}[f(\mathbf{x}_r)] \geq \frac{\mathrm{OPT}}{2} - \mathcal{O}\left(\frac{\sqrt{d}}{\sqrt{K}}\right).$$

Hence, ZO-GA achieves a 1/2 approximate ratio after $\mathcal{O}(d/\epsilon^2)$ zeroth-order oracle calls.

Appendix C. Proof of lemma 8

Proof Firstly, it follows from the definition of function $F(\mathbf{x})$ as given in (11) that

$$\nabla F(\mathbf{x}) = \int_0^1 e^{\theta-1}\nabla f(\theta\mathbf{x})d\theta,$$

which indicates

$$\begin{aligned}\|\nabla F(\mathbf{x}) - \nabla F(\mathbf{y})\| &\leq \int_0^1 e^{\theta-1}\|\nabla \mathbf{f}(\theta\mathbf{x}) - \nabla f(\theta\mathbf{y})\|d\theta \\ &\leq L\|\mathbf{x}-\mathbf{y}\|\int_0^1 e^{\theta-1}\theta d\theta = \frac{L}{e}\|\mathbf{x}-\mathbf{y}\|.\end{aligned}$$

Therefore, $F(\mathbf{x})$ is $\frac{L}{e}$-smooth. Secondly, for any $\mathbf{x}, \mathbf{y} \in \mathcal{X} \subseteq \mathbb{R}_+^d$, we have

$$\begin{aligned}\langle \mathbf{y}-\mathbf{x}, \nabla F(\mathbf{x})\rangle &= \left\langle \mathbf{y}-\mathbf{x}, \int_0^1 e^{\theta-1}\nabla f(\theta\mathbf{x})d\theta \right\rangle \\ &= \int_0^1 e^{\theta-1}\left\langle \mathbf{y}, \nabla f(\theta\mathbf{x})\right\rangle d\theta - \int_0^1 e^{\theta-1}\left\langle \mathbf{x}, \nabla f(\theta\mathbf{x})\right\rangle d\theta \\ &\geq \int_0^1 e^{\theta-1}\langle \mathbf{y}\vee(\theta\mathbf{x}) - \theta\mathbf{x}, \nabla f(\theta\mathbf{x})\rangle d\theta - \int_0^1 e^{\theta-1}df(\theta\mathbf{x}) \\ &\geq \int_0^1 e^{\theta-1}\left(f(\mathbf{y}) - f(\theta\mathbf{x})\right)d\theta - e^{\theta-1}f(\theta\mathbf{x})|_{\theta=0}^{\theta=1} + \int_0^1 e^{\theta-1}f(\theta\mathbf{x})d\theta\end{aligned}$$

$$
\begin{aligned}
&= \left(\int_0^1 e^{\theta-1} d\theta\right) f(\mathbf{y}) - e^{\theta-1} f(\theta\mathbf{x})\Big|_0^1 \\
&\geq \left(1 - e^{-1}\right) f(\mathbf{y}) - f(\mathbf{x}),
\end{aligned}
$$

where the first inequality is because of $\theta\mathbf{x} + \mathbf{y} \geq \mathbf{y} \vee (\theta\mathbf{x})$ for $\theta \in (0, 1]$ and $\mathbf{x}, \mathbf{y} \geq \mathbf{0}$, and the monotonically non-decreasing property of f, the second inequality is due to (4) indicated by the DR-submodularity of f and the monotonically non-decreasing property of f, as follows,

$$
\langle \mathbf{y} \vee (\theta\mathbf{x}) - \theta\mathbf{x}, \nabla f(\theta\mathbf{x}) \rangle \geq f(\mathbf{y} \vee (\theta\mathbf{x})) - f(\theta\mathbf{x}) \geq f(\mathbf{y}) - f(\theta\mathbf{x}),
$$

and the last inequality is due to $f(\mathbf{0}) \geq 0$. Furthermore, if $\mathbf{x}$ is a stationary point for maximizing $F(\mathbf{x})$ over the set $\mathcal{X}$, we have

$$
0 \geq \max_{\mathbf{x} \in \mathcal{X}} \langle \mathbf{y} - \mathbf{x}, \nabla F(\mathbf{x}) \rangle \geq \left(1 - e^{-1}\right) f(\mathbf{y}) - f(\mathbf{x}), \forall \mathbf{y} \in \mathcal{X}.
$$

Therefore, $f(\mathbf{x}) \geq \left(1 - e^{-1}\right) f(\mathbf{x}^*)$, which completes the proof. ■

Appendix D. Proof of lemma 20

Proof (i) It is straightforward to obtain the gradient form of $\hat{F}_\mu$ from (47). Then it follows from the $\frac{\sqrt{d}L_0}{u}$-smoothness of f_μ (see Lemma 4(iii)) that

$$
\begin{aligned}
\|\nabla \hat{F}_\mu(\mathbf{x}) - \nabla \hat{F}_\mu(\mathbf{y})\| &= \|\frac{1}{Z} \sum_{z=1}^{Z} e^{\frac{z}{Z}-1} \left(\nabla f_\mu(\frac{z}{Z}\mathbf{x}) - \nabla f_\mu(\frac{z}{Z}\mathbf{y})\right) \| \\
&\leq \frac{1}{Z} \sum_{z=1}^{Z} e^{\frac{z}{Z}-1} \frac{\sqrt{d}L_0}{u} \frac{z}{Z} \|\mathbf{x} - \mathbf{y}\| \leq \frac{\sqrt{d}L_0}{u} \|\mathbf{x} - \mathbf{y}\|,
\end{aligned}
$$

where it utilizes Jensen's inequality and the fact that $z \leq Z$.

(ii) By the definition of $\hat{F}_\mu$ along with the non-negativeness and monotonicity properties of f_μ as stated in Lemma 4 (i), we obtain

$$
\hat{F}_\mu(\mathbf{x}) = \frac{1}{Z} \sum_{z=1}^{Z} \frac{e^{\frac{z}{Z}-1}}{\frac{z}{Z}} f_\mu(\frac{z}{Z}\mathbf{x}) \leq \frac{1}{Z} \sum_{z=1}^{Z} \frac{1}{\frac{z}{Z}} f_\mu(\mathbf{x}) = f_\mu(\mathbf{x}) \sum_{z=1}^{Z} \frac{1}{z} \leq f_\mu(\mathbf{x})(1 + \ln Z),
$$

where the last inequality comes from $\sum_{z=1}^{Z} \frac{1}{z} \leq 1 + \int_{z=1}^{Z} \frac{1}{z} = 1 + \ln Z$.

(iii) From the expression of $\nabla \hat{F}_\mu(\mathbf{x})$, we obtain

$$
\begin{aligned}
\langle \mathbf{y} - \mathbf{x}, \nabla \hat{F}_\mu(\mathbf{x}) \rangle &= \left\langle \mathbf{y} - \mathbf{x}, \frac{1}{Z} \sum_{z=1}^{Z} e^{\frac{z}{Z}-1} \nabla f_\mu(\frac{z}{Z}\mathbf{x}) \right\rangle \\
&\geq \frac{1}{Z} \sum_{z=1}^{Z} e^{\frac{z}{Z}-1} \left\langle \mathbf{y} \vee (\frac{z}{Z}\mathbf{x}) - \frac{z}{Z}\mathbf{x}, \nabla f_\mu(\frac{z}{Z}\mathbf{x}) \right\rangle - \frac{1}{Z} \sum_{z=1}^{Z} e^{\frac{z}{Z}-1} \left\langle \mathbf{x}, \nabla f_\mu(\frac{z}{Z}\mathbf{x}) \right\rangle
\end{aligned}
$$

$$\begin{aligned}
&\geq \frac{1}{Z}\sum_{z=1}^{Z} e^{\frac{z}{Z}-1}\left(f_\mu(\mathbf{y}) - f_\mu(\frac{z}{Z}\mathbf{x})\right) - \frac{1}{Z}\sum_{z=1}^{Z} e^{\frac{z}{Z}-1}\left\langle \mathbf{x}, \nabla f_\mu(\frac{z}{Z}\mathbf{x})\right\rangle \\
&\geq \left(1-e^{-1}\right) f_\mu(\mathbf{y}) - \frac{1}{Z}\sum_{z=1}^{Z} e^{\frac{z}{Z}-1} f_\mu(\frac{z}{Z}\mathbf{x}) - \frac{1}{Z}\sum_{z=1}^{Z} e^{\frac{z}{Z}-1}\left\langle \mathbf{x}, \nabla f_\mu(\frac{z}{Z}\mathbf{x})\right\rangle \\
&\geq \left(1-e^{-1}\right) f_\mu(\mathbf{y}) - (1+3\frac{\ln Z}{Z}) f_\mu(\mathbf{x}),
\end{aligned}$$

where the first inequality stems from $\mathbf{y} + \frac{z}{Z}\mathbf{x} \geq \mathbf{y} \vee \frac{z}{Z}\mathbf{x}$ for $z \leq Z$ and $\mathbf{x}, \mathbf{y} \geq \mathbf{0}$, and $\nabla f_\mu(\mathbf{x}) \geq \mathbf{0}$, the second one comes from (5) and the monotonicity of $f_\mu(\mathbf{x})$ (see Lemma 4(i)), namely,

$$\left\langle \mathbf{y} \vee (\frac{z}{Z}\mathbf{x}) - \frac{z}{Z}\mathbf{x}, \nabla f_\mu(\frac{z}{Z}\mathbf{x})\right\rangle \geq f_\mu(\mathbf{y} \vee \frac{z}{Z}\mathbf{x}) - f_\mu(\frac{z}{Z}\mathbf{x}) \geq f_\mu(\mathbf{y}) - f_\mu(\frac{z}{Z}\mathbf{x}),$$

the third inequality is due to $\frac{1}{Z}\sum_{z=1}^{Z} e^{\frac{z}{Z}-1} \geq 1 - e^{-1}$, and the fourth one is from (Mitra et al., 2021, Lemma 3.16) with $\nabla G(\mathbf{x}) = \nabla F(\mathbf{x})/e, G(\mathbf{x}) = f_\mu(\mathbf{x})/e$, and $\epsilon^{-1} = Z(Z \geq 3)$ that

$$\frac{1}{Z}\sum_{z=1}^{Z} e^{\frac{z}{Z}-1}\left\langle \mathbf{x}, \nabla f_\mu(\frac{z}{Z}\mathbf{x})\right\rangle \leq \left(1+3\frac{\ln Z}{Z}\right) f_\mu(\mathbf{x}) - \frac{1}{Z}\sum_{z=1}^{Z} e^{\frac{z}{Z}-1} f_\mu(\frac{z}{Z}\mathbf{x}).$$

Furthermore, it follows from $|f_\mu(\mathbf{x}) - f(\mathbf{x})| \leq L_0 u$ (see Lemma 4 (iii)) that

$$\begin{aligned}
\langle \mathbf{y} - \mathbf{x}, \nabla \hat{F}_\mu(\mathbf{x})\rangle &\geq \left(1-e^{-1}\right) f_\mu(\mathbf{y}) - (1+3\frac{\ln Z}{Z}) f_\mu(\mathbf{x}) \\
&\geq \left(1-e^{-1}\right) f(\mathbf{y}) - (1+3\frac{\ln Z}{Z}) f(\mathbf{x}) - (2 - e^{-1} + \frac{3\ln Z}{Z}) L_0 u \\
&\geq \left(1-e^{-1}\right) f(\mathbf{y}) - (1+3\frac{\ln Z}{Z}) f(\mathbf{x}) - 3L_0 u,
\end{aligned}$$

where we use that $\frac{3\ln Z}{Z} - e^{-1} \leq 1$ for the last inequality.

(iv) With $\mathbf{x}$ being a stationary point for $\max_{\mathbf{x}\in\mathcal{X}} \hat{F}_\mu(\mathbf{x})$ and $\mathbf{y} = \mathbf{x}^*$ in (48), we have

$$0 \geq \max\langle \mathbf{x}^* - \mathbf{x}, \nabla \hat{F}_\mu(\mathbf{x})\rangle \geq \left(1-e^{-1}\right) f(\mathbf{x}^*) - (1+3\frac{\ln Z}{Z}) f(\mathbf{x}) - 3L_0 u.$$

This, combined with $f(\mathbf{x}^*) \geq f(\mathbf{x})$, leads to the conclusion. ■

Appendix E. The Lemma in Section 5

Lemma 28. *Assume that $\mathbf{f}_i(\cdot, \xi_i) : \mathbb{R}^n \to \mathbb{R}$, $i \in [M]$, are non-negative, monotonically non-decreasing, DR-submodular and L_0-Lipschitz continuous. Then $\mathbf{f}(\cdot, \xi) := \min_{i\in[M]} \mathbf{f}_i(\cdot, \xi_i)$ is non-negative, monotonically non-decreasing, up-concave and L_0-Lipschitz continuous.*

Proof It is straightforward to show that $\mathbf{f}(\mathbf{x}, \xi)$ is non-negative and monotonic, given the non-negativity and monotonicity of $\mathbf{f}_i(\mathbf{x}, \xi_i), i \in [M]$ and the definition of $\mathbf{f}(\mathbf{x}, \xi)$. Moreover,

each function $\mathbf{f}_i(\mathbf{x}, \xi_i)$ satisfying DR-submodularity is up-concave by the definition of up-concave function, which demonstrates that $\mathbf{f}_i(\mathbf{x} + \beta\mathbf{v}, \xi_i)$ is concave w.r.t $\beta \in \mathbb{R}$ for $\mathbf{v} \geq \mathbf{0}$, i.e.,

$$\mathbf{f}_i\left(\mathbf{x} + (\alpha\beta_1 + (1-\alpha)\beta_2)\,\mathbf{v}, \xi_i\right) \geq \alpha\mathbf{f}_i(\mathbf{x}+\beta_1\mathbf{v}, \xi_i)+(1-\alpha)\mathbf{f}_i(\mathbf{x}+\beta_2\mathbf{v}, \xi_i), \quad \forall\alpha \in [0,1]; \forall\beta_1, \beta_2 \in \mathbb{R}.$$

Then it indicates

$$\begin{aligned}
\mathbf{f}\left(\mathbf{x} + (\alpha\beta_1 + (1-\alpha)\beta_2)\,\mathbf{v}, \xi\right) &= \min_{i\in[M]} \mathbf{f}_i(\mathbf{x} + (\alpha\beta_1 + (1-\alpha)\beta_2)\,\mathbf{v}, \xi_i) \\
&\geq \min_{i\in[M]} \left(\alpha\mathbf{f}_i(\mathbf{x} + \beta_1\mathbf{v}, \xi_i) + (1-\alpha)\mathbf{f}_i(\mathbf{x} + \beta_2\mathbf{v}, \xi_i)\right) \\
&\geq \alpha\mathbf{f}(\mathbf{x} + \beta_1\mathbf{v}, \xi) + (1-\alpha)\mathbf{f}(\mathbf{x} + \beta_2\mathbf{v}, \xi),
\end{aligned}$$

where the first inequality is from the concavity of $\mathbf{f}_i(\mathbf{x}+\beta\mathbf{v}, \xi_i)$ w.r.t. $\beta \geq 0$, and the second one comes from $\min_i(a_i(\mathbf{x}) + b_i(\mathbf{y})) \geq \min_i a_i(\mathbf{x}) + \min_i b_i(\mathbf{y})$. We next prove that $\mathbf{f}(\mathbf{x}, \xi)$ is L_0-Lipschitz continuous. Since for all i,

$$\min_i \mathbf{f}_i(\mathbf{x}, \xi_i) \leq \mathbf{f}_i(\mathbf{x}, \xi_i) = \mathbf{f}_i(\mathbf{x}, \xi_i) - \mathbf{f}_i(\mathbf{y}, \xi_i) + \mathbf{f}_i(\mathbf{y}, \xi) \leq L_0\|\mathbf{x} - \mathbf{y}\| + \mathbf{f}_i(\mathbf{y}, \xi_i),$$

it provides that $\mathbf{f}(\mathbf{x}, \xi) \leq L_0\|\mathbf{x} - \mathbf{y}\| + \mathbf{f}(\mathbf{y}, \xi)$. Similarly, we obtain for all i,

$$\min_i \mathbf{f}_i(\mathbf{y}, \xi_i) \leq L_0\|\mathbf{x} - \mathbf{y}\| + \mathbf{f}_i(\mathbf{x}, \xi_i).$$

Hence, $|\mathbf{f}(\mathbf{x}, \xi) - \mathbf{f}(\mathbf{y}, \xi)| \leq L_0\|\mathbf{x} - \mathbf{y}\|$, which completes the proof. ■

References

Corporate leaderships network dataset – KONECT, 2017. URL `http://konect.cc/networks/brunson_corporate-leadership`.

Arman Adibi, Aryan Mokhtari, and Hamed Hassani. Minimax optimization: The case of convex-submodular. In *International Conference on Artificial Intelligence and Statistics*, pages 3556–3580. PMLR, 2022.

Paola Alimonti. New local search approximation techniques for maximum generalized satisfiability problems. In *Italian Conference on Algorithms and Complexity*, pages 40–53. Springer, 1994.

Zeyuan Allen-Zhu and Elad Hazan. Variance reduction for faster non-convex optimization. In *International Conference on Machine Learning*, pages 699–707. PMLR, 2016.

Francis Bach. Submodular functions: from discrete to continuous domains. *Mathematical Programming*, 175:419–459, 2019.

Krishnakumar Balasubramanian and Saeed Ghadimi. Zeroth-order (non)-convex stochastic optimization via conditional gradient and gradient updates. *Advances in Neural Information Processing Systems*, 31, 2018.

Dimitri P Bertsekas. Stochastic optimization problems with nondifferentiable cost functionals. *Journal of Optimization Theory and Applications*, 12(2):218–231, 1973.

An Bian, Kfir Levy, Andreas Krause, and Joachim M Buhmann. Continuous DR-submodular maximization: Structure and algorithms. *Advances in Neural Information Processing Systems*, 30, 2017a.

Andrew An Bian, Baharan Mirzasoleiman, Joachim Buhmann, and Andreas Krause. Guaranteed non-convex optimization: Submodular maximization over continuous domains. In *International Conference on Artificial Intelligence and Statistics*, pages 111–120. PMLR, 2017b.

Gruia Calinescu, Chandra Chekuri, Martin Pal, and Jan Vondrák. Maximizing a monotone submodular function subject to a matroid constraint. *SIAM Journal on Computing*, 40 (6):1740–1766, 2011.

Chandra Chekuri, Jan Vondrák, and Rico Zenklusen. Submodular function maximization via the multilinear relaxation and contention resolution schemes. *SIAM Journal on Computing*, 43(6):1831–1879, 2014.

Jinghui Chen, Dongruo Zhou, Jinfeng Yi, and Quanquan Gu. A frank-wolfe framework for efficient and effective adversarial attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3486–3494, 2020a.

Lin Chen, Hamed Hassani, and Amin Karbasi. Online continuous submodular maximization. In *International Conference on Artificial Intelligence and Statistics*, pages 1896–1905. PMLR, 2018a.

Lin Chen, Mingrui Zhang, Hamed Hassani, and Amin Karbasi. Black box submodular maximization: Discrete and continuous settings. In *International Conference on Artificial Intelligence and Statistics*, pages 1058–1070. PMLR, 2020b.

Ruobing Chen, Matt Menickelly, and Katya Scheinberg. Stochastic optimization using a trust-region method and random models. *Mathematical Programming*, 169:447–487, 2018b.

Wei Chen, Tian Lin, Zihan Tan, Mingfei Zhao, and Xuren Zhou. Robust influence maximization. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 795–804, 2016.

Andrew R Conn, Katya Scheinberg, and Luis N Vicente. *Introduction to derivative-free optimization*. SIAM, 2009.

Abhimanyu Das and David Kempe. Submodular meets spectral: greedy algorithms for subset selection, sparse approximation and dictionary selection. In *International Conference on Machine Learning*, pages 1057–1064, 2011.

Donglei Du. Lyapunov function approach for approximation algorithm design and analysis: with applications in submodular maximization. *arXiv preprint arXiv:2205.12442*, pages 1–30, 2022.

John C Duchi, Peter L Bartlett, and Martin J Wainwright. Randomized smoothing for stochastic optimization. *SIAM Journal on Optimization*, 22(2):674–701, 2012.

John C Duchi, Michael I Jordan, Martin J Wainwright, and Andre Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015.

Kwassi Joseph Dzahini. Expected complexity analysis of stochastic direct-search. *Computational Optimization and Applications*, 81:179–200, 2022.

Kwassi Joseph Dzahini, Michael Kokkolaras, and Sébastien Le Digabel. Constrained stochastic blackbox optimization using a progressive barrier and probabilistic estimates. *Mathematical Programming*, 198(1):675–732, 2023.

Moran Feldman, Joseph Naor, and Roy Schwartz. A unified continuous greedy algorithm for submodular maximization. In *2011 IEEE 52nd Annual Symposium on Foundations of Computer Science*, pages 570–579, 2011.

Moran Feldman, Amin Karbasi, and Ehsan Kazemi. Do less, get more: Streaming submodular maximization with subsampling. *Advances in Neural Information Processing Systems*, 31, 2018.

Yuval Filmus and Justin Ward. Monotone submodular maximization over a matroid via non-oblivious local search. *SIAM Journal on Computing*, 43(2):514–542, 2014.

Marshall L Fisher, George L Nemhauser, and Laurence A Wolsey. An analysis of approximations for maximizing submodular set functions-II. *Mathematical Programming Studies*, 8:73–87, 1978.

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

Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155(1-2):267–305, 2016.

Carlos Guestrin, Andreas Krause, and Ajit Paul Singh. Near-optimal sensor placements in gaussian processes. In *International Conference on Machine Learning*, pages 265–272, 2005.

Hamed Hassani, Mahdi Soltanolkotabi, and Amin Karbasi. Gradient methods for submodular maximization. *Advances in Neural Information Processing Systems*, 30, 2017.

Hamed Hassani, Amin Karbasi, Aryan Mokhtari, and Zebang Shen. Stochastic conditional gradient++: (non) convex minimization and continuous submodular maximization. *SIAM Journal on Optimization*, 30(4):3315–3344, 2020.

Daisuke Hatano, Takuro Fukunaga, Takanori Maehara, and Ken-ichi Kawarabayashi. Lagrangian decomposition algorithm for allocating marketing channels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Feihu Huang, Bin Gu, Zhouyuan Huo, Songcan Chen, and Heng Huang. Faster gradient-free proximal stochastic methods for nonconvex nonsmooth optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1503–1510, 2019.

Feihu Huang, Lue Tao, and Songcan Chen. Accelerated stochastic gradient-free and projection-free methods. In *International Conference on Machine Learning*, pages 4519–4530. PMLR, 2020.

Kaiyi Ji, Zhe Wang, Yi Zhou, and Yingbin Liang. Improved zeroth-order variance reduced algorithms and analysis for nonconvex optimization. In *International Conference on Machine Learning*, pages 3100–3109. PMLR, 2019.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in Neural Information Processing Systems*, 26, 2013.

David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 137–146, 2003.

Sanjeev Khanna, Rajeev Motwani, Madhu Sudan, and Umesh Vazirani. On syntactic versus computational views of approximability. *SIAM Journal on Computing*, 28(1):164–191, 1998.

Andreas Krause, H Brendan McMahan, Carlos Guestrin, and Anupam Gupta. Robust submodular observation selection. *Journal of Machine Learning Research*, 9(12), 2008.

Jeffrey Larson and Stephen C Billups. Stochastic derivative-free optimization using a trust region framework. *Computational Optimization and applications*, 64:619–645, 2016.

Jeffrey Larson, Matt Menickelly, and Stefan M Wild. Derivative-free optimization methods. *Acta Numerica*, 28:287–404, 2019.

Duksang Lee, Nam Ho-Nguyen, and Dabeen Lee. Non-smooth and holder-smooth submodular maximization. *arXiv preprint arXiv:2210.06061*, 2022.

Yuefang Lian, Donglei Du, Xiao Wang, Dachuan Xu, and Yang Zhou. Stochastic variance reduction for dr-submodular maximization. *Algorithmica*, 86(5):1335–1364, 2024.

Hui Lin and Jeff Bilmes. Multi-document summarization via budgeted maximization of submodular functions. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 912–920, 2010.

Hui Lin and Jeff Bilmes. A class of submodular functions for document summarization. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 510–520, 2011.

Tianyi Lin, Zeyu Zheng, and Michael I Jordan. Gradient-free methods for deterministic and stochastic nonsmooth nonconvex optimization. *University of California, Berkeley*, 2022.

Sijia Liu, Bhavya Kailkhura, Pin-Yu Chen, Paishun Ting, Shiyu Chang, and Lisa Amini. Zeroth-order stochastic variance reduction for nonconvex optimization. *Advances in Neural Information Processing Systems*, 31, 2018a.

Sijia Liu, Xingguo Li, Pin-Yu Chen, Jarvis Haupt, and Lisa Amini. Zeroth-order stochastic projected gradient descent for nonconvex optimization. In *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 1179–1183. IEEE, 2018b.

László Lovász. Submodular functions and convexity. *Mathematical Programming The State of the Art: Bonn 1982*, pages 235–257, 1983.

Friedrich Menhorn, Florian Augustin, H-J Bungartz, and Youssef M Marzouk. A trust-region method for derivative-free nonlinear constrained stochastic optimization. *arXiv preprint arXiv:1703.04156*, 2017.

Baharan Mirzasoleiman, Amin Karbasi, Rik Sarkar, and Andreas Krause. Distributed submodular maximization. *The Journal of Machine Learning Research*, 17(1):8330–8373, 2016.

Baharan Mirzasoleiman, Stefanie Jegelka, and Andreas Krause. Streaming non-monotone submodular maximization: Personalized video summarization on the fly. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Siddharth Mitra, Moran Feldman, and Amin Karbasi. Submodular+concave. *arXiv preprint arXiv:2106.04769*, 2021.

Aryan Mokhtari, Hamed Hassani, and Amin Karbasi. Stochastic conditional gradient methods: From convex minimization to submodular maximization. *The Journal of Machine Learning Research*, 21(1):4232–4280, 2020.

George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functions-I. *Mathematical Programming*, 14(1): 265–294, 1978.

Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17:527–566, 2017.

Mohammad Pedramfar, Christopher John Quinn, and Vaneet Aggarwal. A unified approach for maximizing continuous DR-submodular functions. *arXiv preprint arXiv:2305.16671*, 2023.

Sashank J. Reddi, Ahmed Hefny, Suvrit Sra, Barnabás Póczos, and Alex Smola. Stochastic variance reduction for nonconvex optimization. In *International Conference on Machine Learning*, pages 314–323. PMLR, 2016a.

Sashank J. Reddi, Suvrit Sra, Barnabas Poczos, and Alexander J Smola. Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization. *Advances in Neural Information Processing Systems*, 29, 2016b.

Anit Kumar Sahu, Manzil Zaheer, and Soummya Kar. Towards gradient free and projection free stochastic optimization. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3468–3477. PMLR, 2019.

Sara Shashaani, Fatemeh S Hashemi, and Raghu Pasupathy. Astro-df: A class of adaptive sampling trust-region algorithms for derivative-free stochastic optimization. *SIAM Journal on Optimization*, 28(4):3145–3176, 2018.

Martin Skutella. Convex quadratic and semidefinite programming relaxations in scheduling. *Journal of the ACM (JACM)*, 48(2):206–242, 2001.

Tasuku Soma, Naonori Kakimura, Kazuhiro Inaba, and Ken-ichi Kawarabayashi. Optimal budget allocation: Theoretical guarantee and efficient algorithm. In *International Conference on Machine Learning*, pages 351–359. PMLR, 2014.

Matthew Staib and Stefanie Jegelka. Robust budget allocation via continuous submodular functions. In *International Conference on Machine Learning*, pages 3230–3240. PMLR, 2017.

Matthew Staib, Bryan Wilder, and Stefanie Jegelka. Distributionally robust submodular maximization. In *International Conference on Artificial Intelligence and Statistics*, pages 506–516. PMLR, 2019.

Sebastian Tschiatschek, Rishabh K Iyer, Haochen Wei, and Jeff A Bilmes. Learning mixtures of submodular functions for image collection summarization. *Advances in neural information processing systems*, 27, 2014.

Jan Vondrák. Optimal approximation for the submodular welfare problem in the value oracle model. In *Proceedings of the fortieth Annual ACM Symposium on Theory of Computing*, pages 67–74, 2008.

Kai Wei, Rishabh Iyer, and Jeff Bilmes. Submodularity in data subset selection and active learning. In *International Conference on Machine Learning*, pages 1954–1963. PMLR, 2015.

Bryan Wilder. Equilibrium computation and robust optimization in zero sum games with submodular structure. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Zi Xu, Zi-Qi Wang, Jun-Lin Wang, and Yu-Hong Dai. Zeroth-order alternating gradient descent ascent algorithms for a class of nonconvex-nonconcave minimax problems. *Journal of Machine Learning Research*, 24(313):1–25, 2023.

Mingrui Zhang, Zebang Shen, Aryan Mokhtari, Hamed Hassani, and Amin Karbasi. One sample stochastic frank-wolfe. In *International Conference on Artificial Intelligence and Statistics*, pages 4012–4023. PMLR, 2020.

Qixin Zhang, Zengde Deng, Zaiyi Chen, Haoyuan Hu, and Yu Yang. Stochastic continuous submodular maximization: Boosting via non-oblivious function. In *International Conference on Machine Learning*, pages 26116–26134. PMLR, 2022.