NEW INSIGHT OF VARIANCE REDUCE IN ZERO-ORDER HARD-THRESHOLDING: MITIGATING GRADIENT ERROR AND EXPANSIVITY CONTRADICTIONS

Xinzhe Yuan[1], **William de Vazelhes**[2], **Bin Gu**[2,3*], **Huan Xiong**[1,2*]
[1] IASM, Harbin Institute of Technology
[2]Mohamed bin Zayed University of Artificial Intelligence
[3]School of Artificial Intelligence, Jilin University
`22s012015@stu.hit.edu.cn`
`{william.vazelhes, bin.gu}@mbzuai.ac.ae`
`huan.xiong.math@gmail.com`

ABSTRACT

Hard-thresholding is an important type of algorithm in machine learning that is used to solve ℓ_0 constrained optimization problems. However, the true gradient of the objective function can be difficult to access in certain scenarios, which normally can be approximated by zeroth-order (ZO) methods. The SZOHT algorithm is the only algorithm tackling ℓ_0 sparsity constraints with ZO gradients so far. Unfortunately, SZOHT has a notable limitation on the number of random directions due to the inherent conflict between the deviation of ZO gradients and the expansivity of the hard-thresholding operator. This paper approaches this problem by considering the role of variance and provides a new insight into variance reduction: mitigating the unique conflicts between ZO gradients and hard-thresholding. Under this perspective, we propose a generalized variance reduced ZO hard-thresholding algorithm as well as the generalized convergence analysis under standard assumptions. The theoretical results demonstrate the new algorithm eliminates the restrictions on the number of random directions, leading to improved convergence rates and broader applicability compared with SZOHT. Finally, we illustrate the utility of our method on a ridge regression problem as well as black-box adversarial attacks.

1 INTRODUCTION

ℓ_0 constrained optimization is a fundamental method in large-scale machine learning, particularly in high-dimensional problems. This approach is widely favored for achieving sparse learning. It offers numerous advantages, notably enhancing efficiency by reducing memory usage, computational demands, and environmental impact. Additionally, this constraint plays a crucial role in combatting overfitting and facilitating precise statistical estimation (Negahban et al., 2012; Raskutti et al., 2011; Bühlmann and Van De Geer, 2011; Yuan and Li, 2021). In this study, we focus on the following problem:

$$\min_{\theta \in \mathbb{R}^d} \mathcal{F}(\theta) = \frac{1}{n} \sum_{i=1}^{n} f_i(\theta), \quad s.t. \ \|\theta\|_0 \leq k, \tag{1}$$

Here, $\mathcal{F}(\theta)$ is the (regularized) empirical risk. $\|\theta\|_0$ represents the number of non-zero directions. d is the dimension of θ. Unfortunately, due to the ℓ_0 constraint, (1) becomes an NP-hard problem, rendering traditional methods unsuitable for its analysis.

Therefore, we consider using the hard-threshold iterative algorithm (Raskutti et al., 2011; Jain et al., 2014; Nguyen et al., 2017b; Yuan et al., 2017), which is a widely used technique for obtaining approximate solutions to NP-hard's ℓ_0 constrained optimization problems. Specifically, this technique alternates between the gradient step and the application of the hard threshold operator $\mathcal{H}_k(\theta)$. Operator

---

[*] Corresponding authors.

$\mathcal{H}_k(\theta)$ retains the top k elements of θ while setting all other directions to zero. The advantage of hard-thresholding over its convex relaxations is that it can achieve similar precision without the need for computationally intensive adjustments, such as tuning ℓ_1 penalties or constraints. Hard-thresholding was first used for its full gradient form(Jain et al., 2014). Nguyen (Nguyen et al., 2017b) developed a stochastic gradient descent SGD version of hard thresholding known as StoIHT. Nevertheless, StoIHT's convergence condition is overly stringent for practical applications(Li et al., 2016). To address this issue, (Zhou et al., 2018), (Shen and Li, 2017) and (Li et al., 2016) implemented variance reduction techniques to improve the performance of StoIHT in real-world problem-solving.

However, this type of StoIHT is still not suitable for many problems. For example, in certain graphical modeling tasks (Blumensath and Davies, 2009), obtaining the gradient is computationally hard. Even worse, in some settings, the gradient is inaccessible by nature, for instance in bandit problems (Shamir, 2017), black-box adversarial attacks(Tu et al., 2019; Chen et al., 2017; 2019), or reinforcement learning (Salimans et al., 2017; Mania et al., 2018; Choromanski et al., 2020). To address these challenges, zeroth-order (ZO) optimization methods have been developed(Nesterov and Spokoiny, 2017). These methods commonly replace the inaccessible gradient with its finite difference approximation which can be calculated by simply using the function evaluations. Subsequently, ZO methods have been adapted to handle convex constraint sets, rendering them suitable for solving the ℓ_1 convex relaxation of the problem (1)(Liu et al., 2018; Balasubramanian and Ghadimi, 2018). However, it's essential to highlight that in the context of sparse optimization, ℓ_1 regularization or constraints can introduce substantial estimation bias and result in inferior statistical properties when compared to ℓ_0 regularization and constraints(Fan and Li, 2001; Zhang, 2010).

To tackle this issue, a recent development introduced the Stochastic Zeroth-Order Hard-Thresholding algorithm (SZOHT)(de Vazelhes et al., 2022), specifically designed for ℓ_0 sparsity constraints and gradient-free optimization. Unfortunately, as the only available algorithm in zeroth-order hard-thresholding so far, SZOHT has notable limitations due to the inherent conflict between the deviation of ZO estimators and the expansivity of the hard-thresholding. This limitation makes the algorithm difficult to use in practice, and a natural question is proposed: Could we have a simple ZO hard-thresholding algorithm whose convergence does not rely on the number of q (the number of random directions used to estimate the gradient, further defined in Section 2)?

In this paper, we provide a positive response to this question. Our approach centers on the role of variance in addressing this problem. We firmly believe that variance reduction can offer a dual benefit. It not only holds the potential to accelerate convergence speed but, more importantly, it can effectively mitigate the unique conflicts associated with zero-order hard-thresholding. From this perspective, SZOHT is characterized by its limitation in restricting the sampling of zero-order gradients, essentially representing an incomplete approach to variance reduction. This incompleteness leads to strict conditions for SZOHT. In contrast, we have developed better algorithms by using historical gradients to reduce variance thoroughly. We then provide the convergence and complexity analysis for the generalized variance reduce algorithm under the standard assumptions of sparse learning, which are restricted strong smoothness (RSS), and restricted strong convexity (RSC) (Nguyen et al., 2017b; Shen and Li, 2017) to retain generality. These algorithms eliminate the restrictions on zero-order gradient steps, leading to improved convergence rates and broader applicability. Crucial to our analysis is to provide how variance reduction mitigates contradictions on the parameters q and k. Finally, we demonstrate the effectiveness of our method by applying it to both ridge regression problems and black-box adversarial attacks. Our results highlight that our method can achieve competitive performance when compared to state-of-the-art methods for zeroth-order algorithms designed to enforce sparsity.

The majority of our work can be summarized in three parts:

1. New Perspective on Resolving Conflicts Between Zeroth-Order Methods and Hard-Thresholding. Our paper acknowledges the necessity of mitigating this contradiction, emphasizing the demand for a more flexible and resilient approach. By employing the perspective of variance to analyze this issue, our paper presents a more practical and effective solution.

2. Variance Reduction: Another key innovation presented in the paper is the introduction of variance reduction. This concept provides a unique solution to ℓ_0-constrained zeroth-order optimization. By employing data-driven techniques to reduce variance, the paper not only

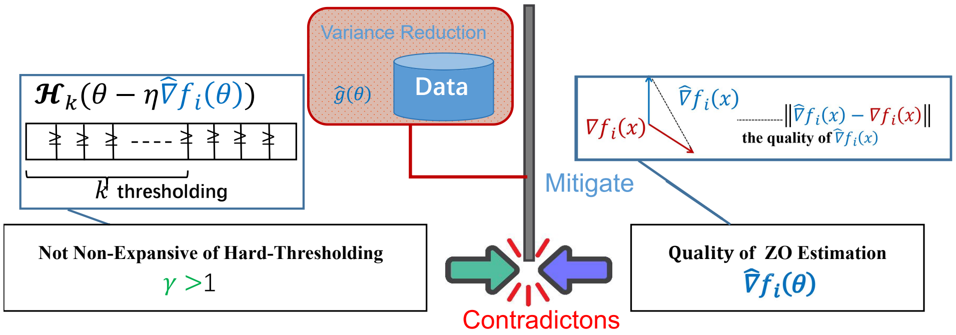

Variance Reduction
$\hat{g}(\theta)$
Data
$\mathcal{H}_k(\theta - \eta\hat{\nabla} f_i(\theta))$
k thresholding
$\hat{\nabla} f_i(x)$
$\nabla f_i(x)$
$\|\hat{\nabla} f_i(x) - \nabla f_i(x)\|$
the quality of $\hat{\nabla} f_i(x)$
Mitigate
Not Non-Expansive of Hard-Thresholding
$\gamma > 1$
Quality of ZO Estimation
$\hat{\nabla} f_i(\theta)$
Contradictons

Figure 1: Motivation of our algorithm.

enhances the algorithm's convergence but also expands its utility across a wider range of scenarios.

3. General Analysis: The introduction of a general analysis framework is another contribution to the paper. This framework systematically evaluates the performance and behavior of varying variance reduced algorithms under ℓ_0-constraint and ZO gradient.

2 PRELIMINARIES

Throughout this paper, we use $\|\theta\|$ to denote the Euclidean norm for a vector, $\|\theta\|_\infty$ to denote the maximum absolute component of that vector, and $\|\theta\|_0$ to denote the ℓ_0 norm (which is not a proper norm). The following two assumptions are widely adopted (Li et al., 2016; Nguyen et al., 2017b) and are needed in this paper.

Assumption 1 (Restricted strong convexity (RSC) (Li et al., 2016; Nguyen et al., 2017b))**.** *A differentiable function $\mathcal{F}$ is restricted ρ_s^--strongly convex at sparsity s if there exists a generic constant $\rho_s^- > 0$ such that for any θ, $\theta' \in \mathbb{R}^d$ with $\|\theta - \theta'\|_0 \leq s$, we have:*

$$\mathcal{F}(\theta) - \mathcal{F}(\theta') - \langle \nabla \mathcal{F}(\theta'), \theta - \theta' \rangle \geq \frac{\rho_s^-}{2}\|\theta - \theta'\|_2^2. \tag{2}$$

Assumption 2 (Restricted strong smoothness (RSS) (Li et al., 2016; Nguyen et al., 2017b))**.** *For any $i \in [n]$, a differentiable function f_i is restricted ρ_s^+-strongly smooth at sparsity level s if there exists a generic constant $\rho_s^+ > 0$ such that for any θ, $\theta' \in \mathbb{R}^d$ with $\|\theta - \theta'\|_0 \leq s$, we have*

$$\|\nabla f_i(\theta) - \nabla f_i(\theta')\| \leq \rho_s^+ \|\theta - \theta'\|.$$

We assume that the objective function $\mathcal{F}(\theta)$ satisfies the RSC condition and that each component function $f_i(\theta)_{i=1}^n$ satisfies the RSS condition. We also define the restricted condition number as $\kappa_s = \rho_s^+/\rho_s^-$. This assumption ensures that the objective function behaves like a strongly convex and smooth function over a sparse domain, even when it is non-convex.

2.1 ZO ESTIMATE

Then, we give our zeroth-order gradient estimator below adopted by (de Vazelhes et al., 2022):

$$\hat{\nabla} f(\theta) = \frac{d}{q\mu}\sum_{i=1}^{q}(f(\theta + \mu \boldsymbol{u}_i) - f(\theta))\boldsymbol{u}_i, \tag{3}$$

where each random direction $\boldsymbol{u}_i$ is a unit vector sampled uniformly from the set $\{\boldsymbol{u} \in \mathbb{R}^d : \|\boldsymbol{u}\|_0 \leq s_2, \|\boldsymbol{u}\| = 1\}$, q is the number of random unit vectors, and $\mu > 0$ is a constant called the *smoothing radius* (typically taken as small as possible, but no too small to avoid numerical errors). To obtain these vectors, we can first sample a random set of coordinates S of size s_2 from $[d]$. Following, we sample a random vector $\boldsymbol{u}$ supported on S, in other words, uniformly sampled from the set $\{u \in \mathbb{R}^d : \boldsymbol{u}_{[d]-S} = \mathbf{0}, \|\boldsymbol{u}\| = 1\}$. Especially, if $s_2 = d$, the general estimator is the usual vanilla estimator with uniform smoothing on the sphere (Gao et al., 2018). Additionally, for convenience, we define $\mathcal{I}^* = supp(\theta^*)$ as the support of θ^*. Let $\theta^{(r)}$ be a sparse vector with $\|\theta^{(r)}\|_0 \leq k$ and support $\mathcal{I}^{(r)} = supp(\theta^{(r)})$. Define, with $\mathcal{H}_{2k}(\cdot)$ the hard-thresholding operator of sparsity $2k$:

$$\widetilde{\mathcal{I}} = supp(\mathcal{H}_{2k}(\hat{\nabla}\mathcal{F}(\theta^*)) \cup supp(\theta^*).$$

and let $\mathcal{I} = \mathcal{I}^{(r)} + \mathcal{I}^{(r+1)} + \widetilde{\mathcal{I}}$. $\varepsilon_\mu = {\rho_s^+}^2 sd$, $\varepsilon_{\mathcal{I}} = \frac{2d}{q(s_2+2)}\left(\frac{(s-1)(s_2-1)}{d-1} + 3\right) + 2$, $\varepsilon_{\mathcal{I}^c} = \frac{2d}{q(s_2+2)}\left(\frac{s(s_2-1)}{-1}\right)$, $\varepsilon_{abs} = \frac{2d{\rho_s^+}^2 ss_2}{q}\left(\frac{(s-1)(s_2-1)}{d-1} + 1\right) + {\rho_s^+}^2 sd.$

2.2 REVISIT OF SZOHT

Based on this assumption and ZO estimation, de Vazelhes proposed the SZOHT algorithm (de Vazelhes et al., 2022). The iteration relationship of this algorithm is:

$$\theta^{(r+1)} = \mathcal{H}_k(\theta^{(r)} - \eta\hat{\nabla}\mathcal{F}(\theta^{(r)}))$$

where $\mathcal{H}_k(\cdot)$ is the hard-thresholding operator and $\hat{\nabla}\mathcal{F}(\theta^{(r)})$ is ZO gradient estimate defined by (3), η represents learning rate. SZOHT can address some ZO ℓ_0-constrained problems under specific conditions. However, it's important to note that the hard-thresholding operator, unlike the projection onto the ℓ_1 ball, lacks non-expansiveness. Consequently, it has the potential to divert the algorithm's iteration away from the desired solution. To deal with this challenge, SZOHT imposes stringent limitations on both k (hard-thresholding coefficients) and q. That is,

$$\left(1 - \frac{{\rho_s^-}^2}{(4\varepsilon_{\mathcal{I}}+1){\rho_s^+}^2}\right)\frac{k^*(4\varepsilon_{\mathcal{I}}+1)^2{\rho_s^+}^4}{{\rho_s^-}^4} \leq k \leq \frac{d-k^*}{2}$$

and

- **if** $s_2 > 1$: $q \geq \frac{16d(s_2-1)k^*\kappa^2}{(s_2+2)(d-1)}\left[18\kappa^2 - 1 + 2\sqrt{9\kappa^2(9\kappa^2-1) + \frac{1}{2} - \frac{1}{2k^*} + \frac{3}{2}\frac{d-1}{k^*(s_2-1)}}\right]$
- **if** $s_2 = 1$: $q \geq \frac{8\kappa^2 d}{\sqrt{\frac{d}{k^*}+1}}$

Evidently, these conditions are exceedingly stringent and may not be suitable for numerous real-world problems. Therefore, we urgently need an algorithm with fewer constraints.

3 GENERAL ANALYSIS WITH VARIANCE

In this section, we will analyze the random ZO hard-thresholding algorithm from the perspective of variance and provide a positive response to the above questions. These algorithms can be described using the following general iterative expression:

$$\theta^{(r+1)} = \mathcal{H}_k(\theta^{(r)} - \eta\hat{g}^{(r)}(\theta^{(r)})), \tag{4}$$

where $\hat{g}^{(r)}(\theta^{(r)})$ is the generalized gradient estimate (applicable to all ZO hard-thresholding algorithms). Let $\alpha = 1 + \frac{2\sqrt{k^*}}{\sqrt{k-k^*}}$. Then, we have:

Theorem 1. *Assume that each f_i is $(\rho_{s'}^+, s')-$RSS and that $\mathcal{F}$ is $(\rho_s^-, s)-$RSC. For any stochastic ZO hard-thresholding algorithm capable of expressing its iterative relationships as described in (4), we can establish the following:*

$$\begin{aligned}\mathbb{E}||\theta^{(r+1)} - \theta^*||_2^2 \leq (1+\eta^2{\rho_s^-}^2)\alpha\mathbb{E}||\theta^{(r)} - \theta^*||_2^2 + \eta^2\alpha\mathbb{E}||\hat{g}_{\mathcal{I}}^{(r)}(\theta^{(r)})||_2^2 - 2\eta\alpha\left[\mathcal{F}(\theta^{(r)}) - \mathcal{F}(\theta^*)\right] \\ + \alpha\frac{n^2\varepsilon_\mu\mu^2}{{\rho_s^-}^2}\end{aligned} \tag{5}$$

Remark 1. *Differing from the approach in (Yuan et al., 2017; Nguyen et al., 2017b; de Vazelhes et al., 2022), where the convergence inequality is segregated into linear convergence terms* *(represented as* $(1 + \eta^2 \rho^- s^2) \alpha \mathbb{E}||\theta^{(r)} - \theta^*||_2^2$ *in (5) and error terms (represented as* $\alpha \frac{n^2 \varepsilon \mu \mu^2}{\rho_s^{-2}} - 2\eta\alpha \left[\mathcal{F}(\theta^{(r)}) - \mathcal{F}(\theta^*)\right]$ *in (5), we have introduced the gradient squared term* $\eta^2 \alpha \mathbb{E}||\hat{g}^{(r)}(\theta^{(r)})||_2^2$ *to elucidate the role of variance better. We can transform (5) into the form of (Yuan et al., 2017; Nguyen et al., 2017b; de Vazelhes et al., 2022) by establishing an upper bound for the gradient squared term, which is often feasible for specific algorithms.*

Conflict analysis through variance. It is worth noting that among these three components, only the gradient squared term $\eta^2 \alpha \mathbb{E}\|\hat{g}^{(r)}(\theta^{(r)})\|_2^2$ encompasses both the hard-thresholding parameter (included by α) and the ZO gradient parameter (included by $\|\hat{g}^{(r)}(\theta^{(r)})\|_2^2$). In essence, this means that the conflict between expansivity and zeroth-order error can be fully encapsulated through the gradient squared term. More importantly, when our attention is directed towards the gradient squared term, we discover that in cases where the gradient estimation is unbiased, we obtain $\mathbb{E}\|\hat{g}^{(r)}(\theta^{(r)})\|_2^2 = \mathrm{Var}\|\hat{g}^{(r)}(\theta^{(r)})\|_2 + \left\|\nabla \mathcal{F}(\theta^{(r)})\right\|^2$, which means that $\mathbb{E}\|\hat{g}^{(r)}(\theta^{(r)})\|_2^2$ only related to the variance of gradient estimation. *This indicates that the conflict between the expansionary of hard-thresholding and ZO error is actually between hard-thresholding and the variance of gradient estimation.* In SZOHT, we have $\hat{g}^{(r)}(\theta^{(r)}) = \hat{\nabla}\mathcal{F}(\theta^{(r)})$. Then, the gradient squared term becomes $\eta^2 \alpha \mathbb{E}||\hat{\nabla}\mathcal{F}(\theta^{(r)})||_2^2$. In this scenario, to guarantee algorithm convergence, it becomes essential to ensure that the gradient squared term remains within a reasonable upper bound. Due to the fact that α is already required to satisfy certain conditions (which are generated by linear convergence terms and error terms), therefore, the sampling method for ZO gradients must be restricted, which leads to a reduction in the variance. However, due to the technique of sampling used to reduce the variance, the limitation on the number q of random directions is introduced into SZOHT.

Improvement plan. A natural idea is to use a more comprehensive variance reduction approach instead of only using sampling technique to reduce $\mathbb{E}||\hat{g}^{(r)}(\theta^{(r)})||_2^2$, which could effectively alleviate the conflict between ZO estimation and hard-thresholding, ultimately enabling the design of algorithms with fewer constraints, broader applicability, and enhanced convergence speed. Based on this perspective, we have developed a generalized variance reduction ZO hard-thresholding algorithm that leverages historical gradients. We will provide a detailed explanation of this algorithm in the next section.

4 *p*M-SZHT ALGORITHM FRAMEWORK

This section mainly presents the *p*M-SZHT algorithm framework along with its convergence analysis. This framework encompasses the majority of unbiased stochastic variance-reduction ZO hard-thresholding methods, providing a generalized result.Subsequently, we introduce the VR-SZHT algorithm, a special case under this framework. Additionally, we extend our discussion by introducing SARAH-ZHT (please note that the gradient estimate in this algorithm is biased) and providing its convergence analysis in the appendix.

4.1 *p*M-SZHT

We now present our generalized algorithm to solve the target problem (1), which we name pM-SZHT (p Memorization Stochastic Zeroth-Order Hard-Thresholding). Each iteration of our algorithm is composed of two steps: (i) the gradient estimation step, and (ii) the hard thresholding step, where the gradient estimation step includes the variance reduce estimation and zeroth-order estimation. We give the full formal description of our algorithm in Algorithm (1).

In the gradient estimation step, we are utilizing the p-Memorization framework, which was originally proposed by Hofmann (Hofmann et al., 2015) to analyze the sequential stochastic gradient algorithm for convex and smooth optimization problems. It's worth noting that our gradient estimation can be seen as its zeroth-order variant (the zeroth-order estimation is shown in Section 2.2). Here, we select in each iteration a random index set $J \subseteq [n]$ of memory locations to update according to:

$$\forall j \in [n] : \; \hat{a}_j^+ := \begin{cases} \hat{\nabla} f_j(\theta), & \text{if } j \in J \\ \hat{a}_j, & \text{otherwise} \end{cases}$$

such that any j has the same probability of p/n being updated[1], where p is the number of directions updated each time (see (Hofmann et al., 2015)). The value of p set J, $\forall j$, $\sum_{J \ni j} \boldsymbol{P}\{J\} - \frac{p}{n}$. Its probability is determined by some specific algorithm. For example, if $\boldsymbol{P}\{J\} = 1/\binom{n}{p}$ if $|J| = p$, and $\boldsymbol{P}\{J\} = 0$ otherwise, we obtain the p-SAGA-ZHT algorithm. If $\boldsymbol{P}\{\emptyset\} = 1 - \frac{p}{n}$ and $\boldsymbol{P}\{[1:n]\} = \frac{p}{n}$, we obtain a variant of the VR-SZHT algorithm from Section 4.2. Those are the ZO hard-thresholding versions of the algorithms mentioned in Hofmann et al. (2015); Gu et al. (2020).

In the hard thresholding step, we only keep the k largest (in magnitude) components of the current iterate $\theta^{(r)}$. This ensures that all our iterates (including the last one) are k-sparse. This hard-thresholding operator has been studied for instance in (Shen and Li, 2017), and possesses several interesting properties. Firstly, it can be seen as a projection on the ℓ_0 ball. Second, importantly, it is not non-expansive, contrary to other operators like the soft-thresholding operator (Shen and Li, 2017).

Algorithm 1 Stochastic variance reduced zeroth-order Hard-Thresholding with p-Memorization (pM-SZHT)

```
Input: Learning rate $\eta$, maximum number of iterations $T$, initial point $\theta^{(0)}$, number of random
    directions $q$, and number of coordinates to keep at each iteration $k$.
Output: $\theta^{(r)}$.
for $r = 1, \dots, T$ do
    Update $\hat{a}^{(r-1)}$
    Randomly sample $i_r \in \{1, 2, \dots, n\}$
    $\hat{g}^{(r-1)}(\theta^{(r-1)}) = \hat{\nabla} f_{i_r}(\theta^{(r-1)}) - \hat{a}_{i_r}^{(r-1)} + \frac{1}{n}\sum_{j=1}^{n} \hat{a}_j^{(r-1)}$
    $\theta^{(r)} = \mathcal{H}_k(\theta^{(r-1)} - \eta g^{(r-1)}(\theta^{(r-1)}))$
end for
```

Convergence Analysis: We provide the convergence analysis of pM-SZHT, using the assumptions from Section 2, and demonstrate the correctness of the conclusions made in Section 3 by assessing whether the algorithm converges independently of q.

Theorem 2. *Suppose $\mathcal{F}(\theta)$ satisfies the RSC condition and that the functions $\{f_i(\theta)\}_{i=1}^{n}$satisfy the RSS condition with $s = 2k + k^*$. For Algorithm 1, suppose that we run SZOHT with random supports of size s_2, q random directions, a learning rate of η, and k coordinates kept at each iteration. We have:*

$$[\mathbb{E}\mathcal{F}(\theta^{(r+1)}) - \mathcal{F}(\theta^*)] \leq \gamma[\mathbb{E}\mathcal{F}(\theta^{(r)}) - \mathcal{F}(\theta^*)] + 2L_\mu + L_r \tag{6}$$

here $L_\mu = \alpha\frac{n^2\varepsilon_\mu\mu^2}{\rho_s^{-2}} + 6\alpha\varepsilon_{abs}\mu^2 + 6\eta^2\alpha A_r$, $L_r = \sqrt{s}||\nabla\mathcal{F}(\theta^*)||_\infty\mathbb{E}||\theta^{(r)} - \theta^*||_2 + \eta^2(3\alpha((4\varepsilon_{\mathcal{I}}s + 2) + \varepsilon_{\mathcal{I}^c}(d-k))\mathbb{E}\|\nabla f_{i_t}(\theta^*)\|_\infty^2)$, $\gamma = \left(\frac{2\beta}{\rho_s^-} + 48\eta^2\alpha\rho_s^+\varepsilon_{\mathcal{I}} - 2\eta\alpha + 1 - \frac{p}{n}\right)$.

Remark 2. *(System error). This format of result is similar to the ones in (Yuan et al., 2017; Nguyen et al., 2017b; de Vazelhes et al., 2022), the right of (25) contains a linear convergence term $\gamma[\mathbb{E}\mathcal{F}(\theta^{(r)}) - \mathcal{F}(\theta^*)]$, and system error $2L_\mu + L_r$. We note that if $\mathcal{F}$ has a k^* -sparse unconstrained minimizer, which could happen in sparse reconstruction, or with overparameterized deep networks, then we would have $||\nabla\mathcal{F}(\theta^*)||_\infty = 0$ and $||\nabla f_{i_r}(\theta^*)||_\infty^2 = 0$, and hence that part of the system error $L_r$would vanish. In addition, we also have another system error L_r, which depends on the smoothing radius μ, due to the error from the ZO estimate and the iterative method of $\hat{a}$.*

From this theorem, we know that if the algorithm converges, η needs to lie in some specific interval.

Corollary 1. *If*

$$\eta' - \frac{\sqrt{\Delta}}{2(48\varepsilon_{\mathcal{I}}\alpha\rho_s^+ + \rho_s^-)} \leq \eta \leq \max\left\{\eta' + \frac{\sqrt{\Delta}}{2(48\varepsilon_{\mathcal{I}}\alpha\rho_s^+ + \rho_s^-)}, \frac{1}{48\varepsilon_{\mathcal{I}}\rho_s^+}\right\} \tag{7}$$

algorithm 1 converges. Here $\eta' = \frac{\alpha}{48\varepsilon_{\mathcal{I}}\alpha\rho_s^+ + \rho_s^-}$, $\Delta = 4\alpha^2 - 4(48\varepsilon_{\mathcal{I}}\alpha\rho_s^+ + \rho_s^-)(1 - \frac{p}{n} + \frac{2}{\rho_s^-})$.

---

[1] Originally, p-Memorization is called q-Memorization. We change it to p to avoid conflicting with random directions in zeroth order

Remark 3. *(Independence of q) When $k > k^*$, for any $q > 0$ the necessary condition $\Delta > 0$ for (7) holds. We emphasize here that variance reduction can only make q unable to determine whether to converge, but q can still affect the convergence speed. In other words, variance reduction can mitigate the conflict, but cannot resolve it.*

4.2 VR-SZHT

To offer a specific analysis, we introduce the VR-SZHT algorithm, which is the adaptation of the original SVRG method Johnson and Zhang (2013) to our ZO hard-thresholding setting. In addition to the previously mentioned convergence analysis, we will also provide a complexity analysis to demonstrate the advantages of this algorithm, which extend beyond existing algorithm.

Algorithm 2 Stochastic variance reduced zeroth-order Hard-Thresholding (VR-SZHT)

```
Input: Learning rate $\eta$, maximum number of iterations $T$, initial point $\theta^0$, SVRG update frequency
    $m$, number of random directions $q$, and number of coordinates to keep at each iteration $k$.
Output: $\theta^T$.
for $r = 1, \dots, T$ do
    $\theta^{(0)} = \theta^{r-1}$;
    $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \hat{\nabla} f_i(\theta^{(0)})$;
    for $t = 0, 1, \dots, m-1$ do
        Randomly sample $i_t \in \{1, 2, \dots, n\}$;
        Compute ZO estimate $\hat{\nabla} f_{i_t}(\theta^{(r)})$, $\hat{\nabla} f_{i_t}(\theta^{(0)})$;
        $\overline{\theta}^{(r+1)} = \theta^{(r)} - \eta(\hat{\nabla} f_{i_t}(\theta^{(r)}) - \hat{\nabla} f_{i_t}(\theta^{(0)}) + \hat{\mu}))$;
        $\theta^{(r+1)} = \mathcal{H}_k(\overline{\theta}^{(r+1)})$;
    end for
    $\theta^r = \theta^{(r+1)}$, random $t' \in [m-1]$;
end for
```

Theorem 3. *Suppose $\mathcal{F}(\theta)$ satisfies the RSC condition and that the functions $\{f_i(\theta)\}_{i=1}^n$ satisfy the RSS condition with $s = 2k + k^*$.When $\eta = \frac{\alpha\rho_S^-}{2(48\varepsilon_{\mathcal{I}}\alpha\rho_s^-\rho_s^+ + {\rho_s^-}^2)}$, we have:*

$$\delta\left[\mathcal{F}(\widetilde{\theta}^{(r)}) - \mathcal{F}(\theta^*)\right] \le \gamma'\mathbb{E}[\mathcal{F}(\widetilde{\theta}^{(r-1)}) - \mathcal{F}(\theta^*)] + L'_\mu + L. \tag{8}$$

Here $\beta = (1 + \eta^2{\rho_s^-}^2)\alpha$, $\delta = \frac{\beta^m - 1}{\beta - 1}(2\eta - 48\varepsilon_{\mathcal{I}}\eta^2\rho_s^+)\alpha$, $\gamma' = (\frac{2\beta^m}{\rho_s^-} + \frac{48\eta^2\rho_s^+\varepsilon_{\mathcal{I}}\alpha(\beta^m - 1)}{\beta - 1})$, $L'_\mu = \frac{2\beta^m}{\rho_s^-}\sqrt{s}\|\nabla\mathcal{F}(\theta^*)\|_\infty\mathbb{E}\|\widetilde{\theta}^{(r-1)} - \theta^*\|_2 + 6\eta^2\frac{\beta^m - 1}{\beta - 1}\alpha((4\varepsilon_{\mathcal{I}}s + 2) + \varepsilon_{\mathcal{I}^c}(d - k))\mathbb{E}||\nabla f_{i_t}(\theta^*)||_\infty^2 + 3||\nabla_{\mathcal{I}}\mathcal{F}(\theta^*)||_2^2)$, and $L' = \frac{\beta^m - 1}{\beta - 1}\alpha(72\eta^2\varepsilon_{abs}\mu^2 + \frac{n^2\varepsilon_\mu\mu^2}{{\rho_s^-}^2})$.

This theorem is similar to Theorem 2. And it is worth noting that q is also independent in VR-SZHT, and can be found in the appendix due to space limitations.

Corollary 2. *The ZO query complexity of the algorithm is $\mathcal{O}\left([n + \frac{\kappa^3}{\kappa^2 + 1}]\log\left(\frac{1}{\varepsilon}\right)\right)$. And the hard-thresholding query complexity is $\mathcal{O}\left(\log(\frac{1}{\varepsilon})\right)$.*

When comparing VR-SZHT with SZOHT, where the ZO query complexity of SZOHT is $\mathcal{O}\left((k + \frac{d}{s_2})\kappa^2\log\left(\frac{1}{\varepsilon}\right)\right)$ and the hard-thresholding query complexity is $\mathcal{O}\left(\kappa^2\log(\frac{1}{\varepsilon})\right)$, it becomes evident that the hard-thresholding query complexity of VR-SZHT is significantly reduced. Furthermore, as k becomes large, the ZO complexity is also reduced.

5 EXPERIMENTS

We now compare the performance of VR-SZHT, SAGA-SZHT, and SARAH-SZHT (an adaptation of the SARAH variance reduction method (Nguyen et al., 2017a) to our ZO hard-thresholding setting, for which we provide the convergence analysis in Appendix 6) with that of the following algorithms, in terms of IZO (iterative zeroth-order oracle, i.e. number of calls to f_i) and NHT (number of hard-thresholding operations):

- SZOHT (de Vazelhes et al., 2022): a vanilla stochastic ZO hard-thresholding algorithm.
- FGZOHT: the full gradient version of SZOHT.

Ridge Regression We first consider the following ridge regression problem, where malfunctions f_i are defined as follows: $f_i(\theta) = (x_i^\top \theta - y_i)^2 + \frac{\lambda}{2}\|\theta\|_2^2$, where λ is some regularization parameter. We generate each x_i randomly from a unit norm ball in $\mathbb{R}^d$, and a true random model θ^* from a normal distribution $\mathcal{N}(0, I_{d\times d})$. Each y_i is defined as $y_i = x_i^\top \theta^*$. We set the constants of the problem as such: $n = 10, d = 5, \lambda = 0.5$. Before training, we preprocess each column by subtracting its mean and dividing it by its empirical standard deviation. We run each algorithm with $k = 3, q = 200, \mu = 10^{-4}, s_2 = d = 5$, and for the variance reduced algorithms, we choose $m = 10$. For all algorithms, the learning rate η is found through grid-search in $\{0.005, 0.01, 0.05, 0.1, 0.5\}$: we choose the learning rates giving the lowest function value (averaged over several runs) at the end of training. We stop each algorithm once its number of IZO reaches 80,000. All curves are averaged over 3 runs, and we plot their mean and standard deviation in Figure 2. As we can observe, SZOHT converges to higher function values than other algorithms: this illustrates the advantage of the variance reduction techniques, which can allow to attain smaller function values than plain SZOHT, but at a cheaper cost than FGZOHT.

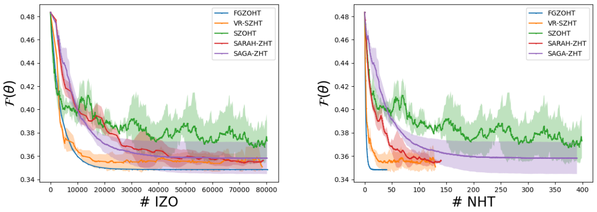


Figure 2: #IZO and #NHT on the ridge regression task.

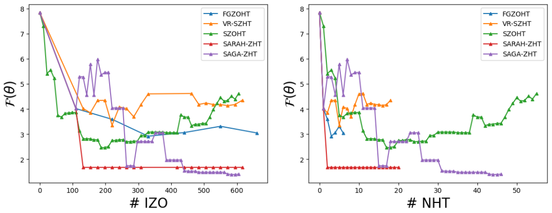


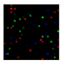
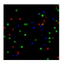
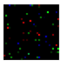
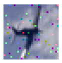
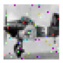
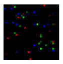
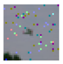
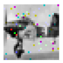
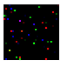
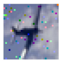
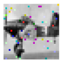
Figure 3: #IZO and #NHT on the few pixels adv. attacks (CIFAR-10), for the original class 'airplane'.

Few Pixels Universal Adversarial Attacks Finally, we consider a few-pixel universal adversarial attacks problem. Let some classifier be trained on a dataset of images. We assume that it can only be accessed as a black box, i.e. it only returns the log probabilities of each estimated class, given an input image. This is a typical real-life scenario, where for instance the model can only be accessed through a provider's API. We seek to find a single perturbation $\theta \in \mathbb{R}^d$, to apply to several images at once, (we denote those images by x_i, $i = \{1, \ldots, n\}$, and their true label as y_i) to make the predicted class for those images different than their true class. Further discussion on universal perturbations can be found in (Dezfooli et al., 2017). In addition, we seek an adversarial perturbation that is sparse, to preserve as much as possible the original image. As is usual in black-box adversarial attacks, we maximize the following Carlini-Wagner loss (Carlini and Wagner, 2017; Chen et al., 2017), which encourages the prediction from the model to be different from the true class:

$$f_i(\theta) = \max\{F_{y_i}(\text{clip}(x_i + \theta)) - \max_{j \neq y_i} F_j(\text{clip}(x_i + \theta)), 0\},$$

where x_i is the original i-th image (rescaled to have values in $[-0.5, 0.5]$), of true class y_i, clip denotes the clipping operation into $[-0.5, 0.5]$, θ is the universal perturbation that we seek to optimize, and

Table 1: Comparison of universal adversarial attacks on $n = 10$ images from the CIFAR-10 test-set, from the 'airplane' class. For each algorithm, the leftmost image is the sparse adversarial perturbation applied to each image in the row. ('auto' stands for 'automobile', and 'plane' for 'airplane')

Image ID	3	27	44	90	97	98	111	116	125	153
Original										
FGZOHT 	 plane	 plane	 plane	 **ship**	 **deer**	 plane	 plane	 plane	 **ship**	 **truck**
SZOHT 	 plane	 plane	 plane	 plane	 **deer**	 plane	 **bird**	 **bird**	 **ship**	 **truck**
VR-SZHT 	 plane	 plane	 **auto**	 plane	 **ship**	 plane	 plane	 plane	 **ship**	 **truck**
SAGA-SZHT 	 plane	 **frog**	 plane	 plane	 **deer**	 plane	 plane	 plane	 **ship**	 **ship**
SARAH-SZHT 	 plane	 plane	 **auto**	 **ship**	 **ship**	 plane	 **bird**	 **plane**	 **frog**	 **truck**

each function F_k outputs the log-probability of image x_i being of class k as predicted by the model, for $k \in \{1, .., K\}$, with K the number of classes (similarly to (Chen et al., 2017; Liu et al., 2018; Huang et al., 2019)). Similarly to Liu et al. (2018) (Appendix A.11), we evaluate the algorithm on a dataset of $n = 10$ images from the test-set of the CIFAR-10 dataset(Krizhevsky et al., 2009), of dimensionality $32 \times 32 \times 3 = 3,072$, from the same class 'airplane', which we display in Table 1. We take as model F a fixed neural network, already trained on the train-set of CIFAR-10, obtained from the supplementary material of (de Vazelhes et al., 2022). We set $k = 60$, $\mu = 0.001$, $q = 10$, $s_2 = d = 3,072$, and the number of inner iterations of the variance reduced algorithms to $m = 10$. We check at each iteration the number of IZO, and we stop training if it exceeds 600. Finally, for each algorithm, we grid-search the learning rate η in $\{0.001, 0.005, 0.01, 0.05\}$. The best learning rates (giving the curve which obtained the smallest minimum function value), are respectively: FGZOHT: 0.05, SZOHT: 0.005, VR-SZHT: 0.01, SAGA-SZHT: 0.05, SARAH-SZHT: 0.05. Our experiments are conducted on a workstation of 128 CPU cores. The training curves are presented in Figure 3: SAGA-SZHT obtains the lowest function value at the end of the training, followed by SARAH-SZHT. In terms of attack success rate, SARAH-SZHT presents the highest success rate, as it has successfully attacked 7/10 images. We provide further results, on 3 more classes ('ship', 'bird', and 'dog') in the appendix, which demonstrate even further the advantage of variance reduction methods in our setting.

6 CONCLUSION

In this paper, we introduce a novel approach to address sparse zero-order optimization problems and leverage it to enhance existing algorithms. We perform a comprehensive convergence analysis of the generalized variance reduction algorithm, showcasing how variance reduction can effectively mitigate the limitations inherent in existing algorithms. To substantiate our claims, we validate our algorithm through experiments involving ridge regression and adversarial attacks.

REFERENCES

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.

Thomas Blumensath and Mike E Davies. Iterative hard thresholding for compressed sensing. *Applied and computational harmonic analysis*, 27(3):265–274, 2009.

Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017.

PinYu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and ChoJui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pages 15–26, 2017.

Xiangyi Chen, Sijia Liu, Kaidi Xu, Xingguo Li, Xue Lin, Mingyi Hong, and David Cox. Zo-adamm: Zeroth-order adaptive momentum method for black-box optimization. *Advances in neural information processing systems*, 32, 2019.

Krzysztof Choromanski, Aldo Pacchiano, Jack Parker-Holder, Yunhao Tang, Deepali Jain, Yuxiang Yang, Atil Iscen, Jasmine Hsu, and Vikas Sindhwani. Provably robust blackbox optimization for reinforcement learning. In *Conference on Robot Learning*, pages 683–696. PMLR, 2020.

William de Vazelhes, Hualin Zhang, Huimin Wu, Xiaotong Yuan, and Bin Gu. Zeroth-order hard-thresholding: Gradient error vs. expansivity. *Advances in Neural Information Processing Systems*, 35:22589–22601, 2022.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in neural information processing systems*, 27, 2014.

SM Moosavi Dezfooli, F Alhussein, F Omar, F Pascal, and S Stefano. Analysis of universal adversarial perturbations. *arXiv preprint arXiv:1705.09554*, 2017.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

X. Gao, B. Jiang, and S. Zhang. On the information-adaptive variants of the admm: An iteration complexity perspective. *Journal of Scientific Computing*, 76(1):327–363, 2018.

Bin Gu, Wenhan Xian, Zhouyuan Huo, Cheng Deng, and Heng Huang. A unified q-memorization framework for asynchronous stochastic optimization. *The Journal of Machine Learning Research*, 21(1):7761–7813, 2020.

T. Hofmann, A. Lucchi, S. Lacoste-Julien, and B. Mcwilliams. Variance reduced stochastic gradient descent with neighbors. *Mathematics*, 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.

Prateek Jain, Ambuj Tewari, and Purushottam Kar. On iterative hard thresholding methods for high-dimensional m-estimation. *Advances in neural information processing systems*, 27, 2014.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26, 2013.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Xingguo Li, Raman Arora, Han Liu, Jarvis Haupt, and Tuo Zhao. Nonconvex sparse learning via stochastic optimization with progressive variance reduction. *arXiv preprint arXiv:1605.02711*, 2016.

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

Horia Mania, Aurelia Guy, and Benjamin Recht. Simple random search of static linear policies is competitive for reinforcement learning. *Advances in Neural Information Processing Systems*, 31, 2018.

Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. 2012.

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

Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *International conference on machine learning*, pages 2613–2621. PMLR, 2017a.

Nam Nguyen, Deanna Needell, and Tina Woolf. Linear convergence of stochastic iterative greedy algorithms with sparse constraints. *IEEE Transactions on Information Theory*, 63(11):6869–6895, 2017b.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE transactions on information theory*, 57(10):6976–6994, 2011.

Tim Salimans, Jonathan Ho, Xi Chen, Szymon Sidor, and Ilya Sutskever. Evolution strategies as a scalable alternative to reinforcement learning. *arXiv preprint arXiv:1703.03864*, 2017.

Ohad Shamir. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *The Journal of Machine Learning Research*, 18(1):1703–1713, 2017.

Jie Shen and Ping Li. A tight bound of hard thresholding. *The Journal of Machine Learning Research*, 18(1):7650–7691, 2017.

ChunChen Tu, Paishun Ting, PinYu Chen, Sijia Liu, Huan Zhang, Jinfeng Yi, ChoJui Hsieh, and ShinMing Cheng. Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, number 01, pages 742–749, 2019.

Joaquin Vanschoren, Jan N Van Rijn, Bernd Bischl, and Luis Torgo. Openml: networked science in machine learning. *ACM SIGKDD Explorations Newsletter*, 15(2):49–60, 2014.

Xiaotong Yuan and Ping Li. Stability and risk bounds of iterative hard thresholding. In *International Conference on Artificial Intelligence and Statistics*, pages 1702–1710. PMLR, 2021.

XiaoTong Yuan, Ping Li, and Tong Zhang. Gradient hard thresholding pursuit. *J. Mach. Learn. Res.*, 18(1):6027–6069, 2017.

CunHui Zhang. Nearly unbiased variable selection under minimax concave penalty. 2010.

Pan Zhou, Xiaotong Yuan, and Jiashi Feng. Efficient stochastic gradient hard thresholding. *Advances in Neural Information Processing Systems*, 31, 2018.

CONTENTS OF APPENDICES

1 RECALL ON VARIOUS VARIANCE REDUCTIONS ALGORITHMS

In this section, we provide a short overview of our introduced variance-reduced algorithms, as well as the corresponding references in the first-order setting, with the corresponding space cost for an epoch and time cost for an epoch, similar to the Table 2 in Gu et al. (2020) for improved clarity and to make our paper self-contained. The computational complexity of each algorithm is indeed similar to its first-order counterpart, when treating the sampling of the random directions at each iteration as a fixed cost.

Table 2: Exposition of the variance reduction algorithms in our paper. [1]: p-SAGA-SZHT, and SAGA-SZHT (i.e. p-SAGA-SZHT, with $p = 1$) are both instantiations of pM-SZHT, as we mention in Section 4.1. [2]: The pM-SZHT framework in Section 4.1 can also be specialized into a variant of SVRG, as described in Section 4.1. However, since the original version of SVRG Johnson and Zhang (2013) can often perform better than such variant in practice, we still have provided an independent analysis for the original SVRG algorithm Johnson and Zhang (2013) (adapted to our ZO hard-thresholding setting), in Section 4.2. [3]:p-SAGA is originally called q-SAGA in Hofmann et al. (2015), but we use p instead to avoid confusion with the number of random directions q. [4]: by *Original algorithm*, we refer to the corresponding algorithm which was first introduced in the first-order setting.

Algorithm	**Sec.**	**Original algorithm** [4]	**Space cost for an epoch**	**Time cost for an epoch**
SAGA-SZHT[1]	4.1	SAGA	$O(dn)$	$O(d)$
p-SAGA-SZHT[1]	4.1	p-SAGA[3]	$O(dn)$	$O(pd)$
VR-SZHT[2]	4.2	SVRG	$O(d)$	$O(dn)$
SARAH-SZHT	6	SARAH	$O(d)$	$O(dn)$

Below we also sum-up the corresponding references of the original (i.e. first-order) versions of each of the variance reduction algorithms above:

Table 3: Corresponding references for the first-order original versions of the variance reductions algorithms. [1] HT: hard-thresholding version of SVRG.

Algorithm	**References**
SAGA	Defazio et al. (2014); Hofmann et al. (2015); Gu et al. (2020)
p-SAGA	Hofmann et al. (2015); Gu et al. (2020)
SVRG	Johnson and Zhang (2013); Hofmann et al. (2015); Gu et al. (2020), HT[1]: Li et al. (2016)
SARAH	Nguyen et al. (2017a)

2 ADDITION FOR SECTION 2

Lemma 1. *(Proof in (Li et al., 2016, Lemma 3.3),) For $k > k^*$ and $\theta \in \mathbb{R}^d$, we have:*

$$\|\mathcal{H}_k(\theta) - \theta^*\|_2^2 \leq \left(1 + \frac{2\sqrt{k^*}}{\sqrt{k - k^*}}\right) \|\theta - \theta^*\|_2^2. \tag{9}$$

Remark 4. *In fact, through assumption 2, we can deduce another version of RSS conditions ((**?**, Lemma 1.2.3)), that is:*

$$f_i(\theta) - f_i(\theta') - \langle \nabla f_i(\theta'), \theta - \theta' \rangle \leq \frac{\rho_s^+}{2} \|\theta - \theta'\|_2^2. \tag{10}$$

Proof. For all θ, $\theta' \in \mathbb{R}^n$ we have

$$\begin{aligned} f(\theta') &= f(\theta) + \int_0^1 \langle f'(\theta + \tau(\theta' - \theta)), \theta' - \theta \rangle \, d\tau \\ &= f(\theta) + \langle f'(\theta), \theta' - \theta \rangle + \int_0^1 \langle f'(\theta + \tau(\theta' - \theta)) - f'(\theta), \theta' - \theta \rangle \, d\tau \end{aligned}$$

Therefore

$$
\begin{aligned}
|f(\theta') - f(\theta) - \langle f'(\theta), \theta' - \theta \rangle| &= \left| \int_0^1 \langle f'(\theta + \tau(\theta' - \theta)) - f'(\theta), \theta' - \theta \rangle \, d\tau \right| \\
&\leq \int_0^1 |\langle f'(\theta + \tau(\theta' - \theta)) - f'(\theta), \theta' - \theta \rangle| d\tau \\
&\leq \int_0^1 \|f'(\theta + \tau(\theta' - \theta)) - f'(\theta)\| \cdot \|\theta' - \theta\| d\tau \\
&\leq \int_0^1 \tau \rho_s^+ \|\theta' - \theta\|^2 d\tau = \frac{\rho_s^+}{2} \|\theta' - \theta\|^2.
\end{aligned}
$$

□

Lemma 2. *(Proof in de Vazelhes et al. (2022)) Let us consider any support $\mathcal{I} \subset [d]$ of size $s(|\mathcal{I}| = s)$. For the ZO gradient estimator in (3), with q random directions, and random supports of size s_2, and assuming that each f_i is (L_{s_2}, s_2)-RSS, with $\hat{\nabla}_{\mathcal{I}} f(x) := \mathcal{H}_k(\nabla f(\theta))$ on $\mathcal{I}$, we have:*

- $\|\mathbb{E}\hat{\nabla}_{\mathcal{I}} f_i(x) - \nabla_{\mathcal{I}} f_i(x)\|^2 \leq \varepsilon_\mu \mu^2$.
- $\mathbb{E}\|\hat{\nabla}_{\mathcal{I}} f_i(\theta)\|^2 \leq \varepsilon_{\mathcal{I}} \|\nabla_{\mathcal{I}} f_i(\theta)\|^2 + \varepsilon_{\mathcal{I}^c} \|\nabla_{\mathcal{I}^c} f(\theta)\|^2 + \varepsilon_{abs} \mu^2$.
- $\mathbb{E}\|\hat{\nabla}_{\mathcal{I}} f_i(\theta) - \nabla f_i(x)\|^2 \leq 2(\varepsilon_{\mathcal{I}} + 1)\|\nabla_{\mathcal{I}} f_i(\theta)\|^2 + 2\varepsilon_{\mathcal{I}^c} \|\nabla_{\mathcal{I}^c} f_i(\theta)\|^2 + 2\varepsilon_{abs} \mu^2$, *where* $\varepsilon_\mu = {\rho_s^+}^2 sd$, $\varepsilon_{\mathcal{I}} = \frac{2d}{q(s_2+2)}(\frac{(s-1)(s_2-1)}{d-1} + 3) + 2$, $\varepsilon_{\mathcal{I}^c} = \frac{2d}{q(s_2+2)}\left(\frac{s(s_2-1)}{-1}\right)$, $\varepsilon_{abs} = \frac{2d{\rho_s^+}^2 s s_2}{q}\left(\frac{(s-1)(s_2-1)}{d-1} + 1\right) + {\rho_s^+}^2 sd$.

3 PROOF OF SECTION 3

Here we prove the following inequality in Section 3:

$$
\begin{aligned}
\mathbb{E}||\theta^{(r)} - \eta \hat{g}_{\mathcal{I}}^{(r)}(\theta^{(r)}) - \theta^*||_2^2 \leq (1 + \eta^2 {\rho_s^-}^2)\mathbb{E}||\theta^{(r)} - \theta^*||_2^2 + \eta^2 \mathbb{E}||\hat{g}_{\mathcal{I}}^{(r)}(\theta^{(r)})||_2^2 - 2\eta \left[\mathcal{F}(\theta^{(r)}) - \mathcal{F}(\theta^*)\right] \\
+ \frac{n^2 \varepsilon_\mu \mu^2}{{\rho_s^-}^2}.
\end{aligned}
$$

Proof. We denote $v = \theta^{(r)} - \eta g_{\mathcal{I}}^{(r)}(\theta^{(r)})$ and $\mathcal{I} = \mathcal{I}^* \cup \mathcal{I}^{(r)} \cup \mathcal{I}^{(r+1)}$, where $\mathcal{I}^* = supp(\theta^*)$, $\mathcal{I}^{(r)} = supp(\theta^{(r)})$ and $\mathcal{I}^{(r+1)} = supp(\theta^{(r+1)})$

$$
\begin{aligned}
\mathbb{E}||v - \theta^*||_2^2 &= \mathbb{E}||\theta^{(r)} - \theta^*||^2 + \eta^2 \mathbb{E}||\hat{g}_{\mathcal{I}}^{(r)}(\theta^{(r)})||_2^2 - 2\eta \left\langle \theta^{(r)} - \theta^*, \mathbb{E}\hat{g}_{\mathcal{I}}^{(r)}(\theta^{(r)}) \right\rangle \\
&= \mathbb{E}||\theta^{(r)} - \theta^*||_2^2 + \eta^2 \mathbb{E}||\hat{g}_{\mathcal{I}}^{(r)}(\theta^{(r)})||_2^2 - 2\eta \left\langle \theta^{(r)} - \theta^*, \mathbb{E}\hat{\nabla}_{\mathcal{I}}\mathcal{F}(\theta^{(r)}) \right\rangle \\
&= \mathbb{E}||\theta^{(r)} - \theta^*||_2^2 + \eta^2 \mathbb{E}||\hat{g}_{\mathcal{I}}^{(r)}(\theta^{(r)})||_2^2 \\
&- 2\eta\mathbb{E} \left\langle \theta^{(r)} - \theta^*, \hat{\nabla}_{\mathcal{I}}\mathcal{F}(\theta^{(r)}) - \nabla_{\mathcal{I}}\mathcal{F}(\theta^{(r)}) \right\rangle - 2\eta\mathbb{E} \left\langle \theta^{(r)} - \theta^*, \nabla_{\mathcal{I}}\mathcal{F}(\theta^{(r)}) \right\rangle \\
&\leq \mathbb{E}||\theta^{(r)} - \theta^*||_2^2 + \eta^2 \mathbb{E}||\hat{g}_{\mathcal{I}}^{(r)}(\theta^{(r)})||_2^2 - 2\eta\mathbb{E} \left\langle \theta^{(r)} - \theta^*, \hat{\nabla}_{\mathcal{I}}\mathcal{F}(\theta^{(r)}) - \nabla_{\mathcal{I}}\mathcal{F}(\theta^{(r)}) \right\rangle \\
&- 2\eta \left[\mathcal{F}(\theta^{(r)}) - \mathcal{F}(\theta^*)\right] \\
&= \mathbb{E}||\theta^{(r)} - \theta^*||_2^2 + \eta^2 \mathbb{E}||\hat{g}_{\mathcal{I}}^{(r)}(\theta^{(r)})||_2^2 \\
&- 2\eta\mathbb{E} \left\langle \sqrt{\eta}\rho_s^-(\theta^{(r)} - \theta^*), \frac{1}{\sqrt{\eta}\rho_s^-}(\hat{\nabla}_{\mathcal{I}}\mathcal{F}(\theta^{(r)}) - \nabla_{\mathcal{I}}\mathcal{F}(\theta^{(r)})) \right\rangle - 2\eta \left[\mathcal{F}(\theta^{(r)}) - \mathcal{F}(\theta^*)\right] \\
&\leq (1 + \eta^2 {\rho_s^-}^2)\mathbb{E}||\theta^{(r)} - \theta^*||_2^2 + \eta^2 \mathbb{E}||\hat{g}_{\mathcal{I}}^{(r)}(\theta^{(r)})||_2^2 - 2\eta \left[\mathcal{F}(\theta^{(r)}) - \mathcal{F}(\theta^*)\right] \\
&+ \frac{1}{\rho_s^-}^2 \mathbb{E}||(\hat{\nabla}_{\mathcal{I}}\mathcal{F}(\theta^{(r)}) - \nabla_{\mathcal{I}}\mathcal{F}(\theta^{(r)}))||_2^2.
\end{aligned}
\tag{11}
$$

For $\mathbb{E}||(\hat{\nabla}_{\mathcal{I}}\mathcal{F}(\theta^{(r)}) - \nabla_{\mathcal{I}}\mathcal{F}(\theta^{(r)}))||_2^2$, we have:

$$
\begin{aligned}
\mathbb{E}||(\hat{\nabla}_{\mathcal{I}}\mathcal{F}(\theta^{(r)}) - \nabla_{\mathcal{I}}\mathcal{F}(\theta^{(r)}))||_2^2 &= ||\mathbb{E}_{\boldsymbol{u}}(\hat{\nabla}_{\mathcal{I}} \sum_{i=1}^n f_i(\theta^{(r)}) - \nabla_{\mathcal{I}} \sum_{i=1}^n f_i(\theta^{(r)}))||_2^2 \\
&\leq n \sum_{i=1}^n ||\mathbb{E}_{\boldsymbol{u}}(\hat{\nabla}_{\mathcal{I}} f_i(\theta^{(r)}) - \nabla_{\mathcal{I}} f_i(\theta^{(r)}))||_2^2 \\
&\leq n^2 \varepsilon_\mu \mu^2.
\end{aligned}
\tag{12}
$$

By constraining $\mathbb{E}||(\hat{\nabla}_{\mathcal{I}}\mathcal{F}(\theta^{(r)}) - \nabla_{\mathcal{I}}\mathcal{F}(\theta^{(r)}))||_2^2$, we can turn (11) into:

$$
\begin{aligned}
\mathbb{E}||\theta^{(r)} - \eta\hat{g}_{\mathcal{I}}^{(r)}(\theta^{(r)}) - \theta^*||_2^2 \leq (1 + \eta^2 {\rho_s^-}^2)\mathbb{E}||\theta^{(r)} - \theta^*||_2^2 + \eta^2 \mathbb{E}||\hat{g}_{\mathcal{I}}^{(r)}(\theta^{(r)})||_2^2 - 2\eta \left[\mathcal{F}(\theta^{(r)}) - \mathcal{F}(\theta^*)\right] \\
+ \frac{n^2 \varepsilon_\mu \mu^2}{{\rho_s^-}^2}.
\end{aligned}
\tag{13}
$$

□

4 PROOF OF pM-SZHT

4.1 PROOF OF THEOREM 1

Before providing the proof of Theorem 1, we need the following lemma:

Lemma 3. *Suppose that the functions $f_i(x)$ satisfies the RSS condition with $s = 2k + k^*$, For qM-SZHT, we can get:*

$$
\begin{aligned}
\mathbb{E}||\hat{g}^{(r)}(\theta^{(r)})||^2 \leq \frac{24\rho_s^+}{n} \sum_{u=1}^{r-1} (1 - \frac{q}{n})^{r-u-1} [\mathcal{F}(\theta^{(u)}) - \mathcal{F}(\theta^*)] + 24\rho_s^+ (1 - \frac{q}{n})^r [\mathcal{F}(\theta^{(0)}) - \mathcal{F}(\theta^*)] \\
+ 48k||F(\theta^r) - F(\theta^*)||^2 + 3((4\varepsilon_{\mathcal{I}} s + 2) + \varepsilon_{\mathcal{I}^c}(d - k))\mathbb{E}\|\nabla f_{i_r}(\theta^*)\|_\infty^2 + 6\varepsilon_{abs}\mu^2 + 6A_r.
\end{aligned}
\tag{14}
$$

Proof. we first get the upper bound of $\mathbb{E}||\hat{g}^{(r)}(\theta^{(r)})||^2$

$$\begin{aligned}
\mathbb{E}||\hat{g}^{(r)}(\theta^{(r)})||^2 &= \mathbb{E}||\hat{\nabla} f_{i_r}^{(r)}(\theta^{(r)}) - \hat{a}_{i_r}^{(r)} + \frac{1}{n}\sum_{j=1}^{n} \hat{a}_j^{(r)}||^2 \\
&= \mathbb{E}||\hat{\nabla} f_{i_r}^{(r)}(\theta^{(r)}) - \nabla f_{i_r}^{(r)}(\theta^*) - \hat{a}_{i_r}^{(r)} + \nabla f_{i_r}^{(r)}(\theta^*) + \frac{1}{n}\sum_{j=1}^{n} \hat{a}_j^{(r)} - \nabla F(\theta^*) + \nabla F(\theta^*)||^2 \\
&\le 3\mathbb{E}||\hat{a}_{i_r}^{(r)} - \nabla f_{i_r}^{(r)}(\theta^*) - \left(\frac{1}{n}\sum_{j=1}^{n} \hat{a}_j^{(r)} - \nabla F(\theta^*)\right)||^2 + 3\mathbb{E}||\hat{\nabla} f_{i_r}^{(r)}(\theta^{(r)}) - \nabla f_{i_r}^{(r)}(\theta^*)||^2 + 3||\nabla F(\theta^*)||^2 \\
&\le 3\mathbb{E}||\hat{a}_{i_r}^{(r)} - \nabla f_{i_r}^{(r)}(\theta^*)||^2 + 3\mathbb{E}||\hat{\nabla} f_{i_r}^{(r)}(\theta^{(r)}) - \nabla f_{i_r}^{(r)}(\theta^*)||^2 + 3||\nabla F(\theta^*)||^2 \\
&\le 6\mathbb{E}||a_{i_r}^{(r)} - \nabla f_{i_r}^{(r)}(\theta^*)||^2 + 6\mathbb{E}||\hat{a}_{i_r}^{(r)} - a_{i_r}^{(r)}||^2 + 3\mathbb{E}||\hat{\nabla} f_{i_r}^{(r)}(\theta^{(r)}) - \nabla f_{i_r}^{(r)}(\theta^*)||^2 + 3||\nabla F(\theta^*)||^2.
\end{aligned} \tag{15}$$

From Lemma 3 we get

$$\mathbb{E}\|\hat{\nabla}_{\mathcal{I}} f_i(\theta) - \nabla f_i(x)\|^2 \le 2(\varepsilon_{\mathcal{I}} + 1)\|\nabla_{\mathcal{I}} f_i(\theta)\|^2 + 2\varepsilon_{\mathcal{I}^c}\|\nabla_{\mathcal{I}^c} f_i(\theta)\|^2 + 2\varepsilon_{abs}\mu^2.$$

Then we get:

$$\begin{aligned}
\mathbb{E}||\hat{a}_{i_r}^{(r)} - a_{i_r}^{(r)}||^2 \le \frac{1}{n}\sum_{u=1}^{r-1}(1-\frac{q}{n})^{r-u-1} 2((\varepsilon_{\mathcal{I}} + 1)||\nabla_{\mathcal{I}} f_i(\theta^u)||^2 + \varepsilon_{\mathcal{I}^c}||\nabla_{\mathcal{I}^c} f_i(\theta^u)||^2 \\
+ \varepsilon_{abs}\mu^2) + (1-\frac{q}{n})^{r-1} 2((\varepsilon_{\mathcal{I}} + 1)||\nabla_{\mathcal{I}} f_i(\theta^0)||^2 + \varepsilon_{\mathcal{I}^c}||\nabla_{\mathcal{I}^c} f_i(\theta^0)||^2 + \varepsilon_{abs}\mu^2) = A_{r'}.
\end{aligned} \tag{16}$$

For $\mathbb{E}\|\hat{\nabla}_{\mathcal{I}} f_{i_r}(\theta) - \nabla_{\mathcal{I}} f_{i_r}(\theta^*)\|_2^2$,and Gu et al. (2020) we have:

$$\begin{aligned}
\mathbb{E}\|\hat{\nabla}_{\mathcal{I}} f_{i_r}(\theta) - \nabla_{\mathcal{I}} f_{i_r}(\theta^*)\|_2^2 \le 4\varepsilon_{\mathcal{I}}\mathbb{E}\|\nabla f_{i_r}(\theta^t) - \nabla f_{i_r}(\theta^*)\|^2 \\
+ ((4\varepsilon_{\mathcal{I}} s + 2) + \varepsilon_{\mathcal{I}^c}(d-k))\mathbb{E}\|\nabla f_{i_r}(\theta^*)\|_\infty^2 + 2\varepsilon_{abs}\mu^2.
\end{aligned} \tag{17}$$

From Gu et al. (2020), we know:

$$\begin{aligned}
6||a^{(r)} - \nabla f(\theta^*)||^2 + 12\varepsilon_{\mathcal{I}}\mathbb{E}||\nabla f^{(r)}(\theta^{(r)}) - \nabla f^{(r)}(\theta^*)||^2 \le \frac{24\rho_s^+}{n}\sum_{u=1}^{t-1}(1-\frac{q}{n})^{t-u-1}[\mathcal{F}(\theta^{(u)}) - \mathcal{F}(\theta^*)] \\
+ 24\rho_s^+(1-\frac{q}{n})^t[\mathcal{F}(\theta^{(0)}) - \mathcal{F}(\theta^*)] + 48\varepsilon_{\mathcal{I}}\mathbb{E}[\mathcal{F}(\theta^{(r)}) - \mathcal{F}(\theta^*)].
\end{aligned} \tag{18}$$

Taking (16),(18) into (15),

$$\begin{aligned}
\mathbb{E}||\hat{g}^{(r)}(\theta^{(r)})||^2 \le \frac{24\rho_s^+}{n}\sum_{u=1}^{t-1}(1-\frac{q}{n})^{t-u-1}[\mathcal{F}(\theta^{(u)}) - \mathcal{F}(\theta^*)] + 24\rho_s^+(1-\frac{q}{n})^t[\mathcal{F}(\theta^{(0)}) - \mathcal{F}(\theta^*)] \\
+ 48k\varepsilon_{\mathcal{I}}\mathbb{E}[\mathcal{F}(\theta^{(r)}) - \mathcal{F}(\theta^*)] + 3((4\varepsilon_{\mathcal{I}} s + 2) + \varepsilon_{\mathcal{I}^c}(d-k))\mathbb{E}\|\nabla f_{i_r}(\theta^*)\|_\infty^2 + 6\varepsilon_{abs}\mu^2 + 6A_t.
\end{aligned}$$

Here we get the conclusion. □

Then we can prove **Theorem 1**:

Proof. We denote $v = \theta^{(r)} - \eta g_{\mathcal{I}}^{(r)}(\theta^{(r)})$ and $\mathcal{I} = \mathcal{I}^* \cup \mathcal{I}^{(r)} \cup \mathcal{I}^{(r+1)}$, where $\mathcal{I}^* = supp(\theta^*)$, $\mathcal{I}^{(r)} = supp(\theta^{(r)})$ and $\mathcal{I}^{(r+1)} = supp(\theta^{(r+1)})$

$$\begin{aligned}
\mathbb{E}||v - \theta^*||_2^2 &= \mathbb{E}||\theta^{(r)} - \theta^*||^2 + \eta^2 \mathbb{E}||\hat{g}_{\mathcal{I}}^{(r)}(\theta^{(r)})||_2^2 - 2\eta \left\langle \theta^{(r)} - \theta^*, \mathbb{E}\hat{g}_{\mathcal{I}}^{(r)}(\theta^{(r)}) \right\rangle \\
&= \mathbb{E}||\theta^{(r)} - \theta^*||_2^2 + \eta^2 \mathbb{E}||\hat{g}_{\mathcal{I}}^{(r)}(\theta^{(r)})||_2^2 - 2\eta \left\langle \theta^{(r)} - \theta^*, \mathbb{E}\hat{\nabla}_{\mathcal{I}}\mathcal{F}(\theta^{(r)}) \right\rangle \\
&= \mathbb{E}||\theta^{(r)} - \theta^*||_2^2 + \eta^2 \mathbb{E}||\hat{g}_{\mathcal{I}}^{(r)}(\theta^{(r)})||_2^2 \\
&- 2\eta\mathbb{E} \left\langle \theta^{(r)} - \theta^*, \hat{\nabla}_{\mathcal{I}}\mathcal{F}(\theta^{(r)}) - \nabla_{\mathcal{I}}\mathcal{F}(\theta^{(r)}) \right\rangle - 2\eta\mathbb{E} \left\langle \theta^{(r)} - \theta^*, \nabla_{\mathcal{I}}\mathcal{F}(\theta^{(r)}) \right\rangle \\
&\leq \mathbb{E}||\theta^{(r)} - \theta^*||_2^2 + \eta^2 \mathbb{E}||\hat{g}_{\mathcal{I}}^{(r)}(\theta^{(r)})||_2^2 - 2\eta\mathbb{E} \left\langle \theta^{(r)} - \theta^*, \hat{\nabla}_{\mathcal{I}}\mathcal{F}(\theta^{(r)}) - \nabla_{\mathcal{I}}\mathcal{F}(\theta^{(r)}) \right\rangle \\
&- 2\eta \left[\mathcal{F}(\theta^{(r)}) - \mathcal{F}(\theta^*) \right] \\
&= \mathbb{E}||\theta^{(r)} - \theta^*||_2^2 + \eta^2 \mathbb{E}||\hat{g}_{\mathcal{I}}^{(r)}(\theta^{(r)})||_2^2 \\
&- 2\eta\mathbb{E} \left\langle \sqrt{\eta}\rho_s^-(\theta^{(r)} - \theta^*), \frac{1}{\sqrt{\eta}\rho_s^-}(\hat{\nabla}_{\mathcal{I}}\mathcal{F}(\theta^{(r)}) - \nabla_{\mathcal{I}}\mathcal{F}(\theta^{(r)})) \right\rangle - 2\eta \left[\mathcal{F}(\theta^{(r)}) - \mathcal{F}(\theta^*) \right] \\
&\leq (1 + \eta^2 {\rho_s^-}^2)\mathbb{E}||\theta^{(r)} - \theta^*||_2^2 + \eta^2 \mathbb{E}||\hat{g}_{\mathcal{I}}^{(r)}(\theta^{(r)})||_2^2 - 2\eta \left[\mathcal{F}(\theta^{(r)}) - \mathcal{F}(\theta^*) \right] \\
&+ {\frac{1}{\rho_s^-}}^2 \mathbb{E}||(\hat{\nabla}_{\mathcal{I}}\mathcal{F}(\theta^{(r)}) - \nabla_{\mathcal{I}}\mathcal{F}(\theta^{(r)}))||_2^2.
\end{aligned} \tag{19}$$

The first inequality follows from N-dimensional mean inequality and the second inequality follows from Assumption 2. For $\mathbb{E}||(\hat{\nabla}_{\mathcal{I}}\mathcal{F}(\theta^{(r)}) - \nabla_{\mathcal{I}}\mathcal{F}(\theta^{(r)}))||_2^2$, we have:

$$\begin{aligned}
\mathbb{E}||(\hat{\nabla}_{\mathcal{I}}\mathcal{F}(\theta^{(r)}) - \nabla_{\mathcal{I}}\mathcal{F}(\theta^{(r)}))||_2^2 &= ||\mathbb{E}_{\boldsymbol{u}}(\hat{\nabla}_{\mathcal{I}}\sum_{i=1}^n f_i(\theta^{(r)}) - \nabla_{\mathcal{I}}\sum_{i=1}^n f_i(\theta^{(r)}))||_2^2 \\
&\leq n\sum_{i=1}^n ||\mathbb{E}_{\boldsymbol{u}}(\hat{\nabla}_{\mathcal{I}} f_i(\theta^{(r)}) - \nabla_{\mathcal{I}} f_i(\theta^{(r)}))||_2^2 \\
&\leq n^2\varepsilon_\mu\mu^2.
\end{aligned} \tag{20}$$

By constraining $\mathbb{E}||(\hat{\nabla}_{\mathcal{I}}\mathcal{F}(\theta^{(r)}) - \nabla_{\mathcal{I}}\mathcal{F}(\theta^{(r)}))||_2^2$, we can turn (19) into:

$$\begin{aligned}
\mathbb{E}||\theta^{(r)} - \eta\hat{g}_{\mathcal{I}}^{(r)}(\theta^{(r)}) - \theta^*||_2^2 &\leq (1 + \eta^2 {\rho_s^-}^2)\mathbb{E}||\theta^{(r)} - \theta^*||_2^2 + \eta^2\mathbb{E}||\hat{g}_{\mathcal{I}}^{(r)}(\theta^{(r)})||_2^2 \\
&- 2\eta \left[\mathcal{F}(\theta^{(r)}) - \mathcal{F}(\theta^*) \right] + \frac{n^2\varepsilon_\mu\mu^2}{{\rho_s^-}^2}.
\end{aligned} \tag{21}$$

Let $\alpha = 1 + \frac{2\sqrt{k^*}}{\sqrt{k - k^*}}$. Using Lemma 1, we have:

$$\begin{aligned}
\mathbb{E}||\theta^{(r+1)} - \theta^*||_2^2 &\leq (1 + \eta^2 {\rho_s^-}^2)\alpha\mathbb{E}||\theta^{(r)} - \theta^*||_2^2 + \alpha\frac{n^2\varepsilon_\mu\mu^2}{{\rho_s^-}^2} - 2\eta\alpha \left[\mathcal{F}(\theta^{(r)}) - \mathcal{F}(\theta^*) \right] \\
&+ \eta^2(\alpha\frac{24\rho_s^+}{n}\sum_{u=1}^{t-1}(1 - \frac{q}{n})^{t-u-1}[\mathcal{F}(\theta^{(u)}) - \mathcal{F}(\theta^*)] \\
&+ 24\alpha\rho_s^+(1 - \frac{q}{n})^t[\mathcal{F}(\theta^{(0)}) - \mathcal{F}(\theta^*)] + 48\alpha k\varepsilon_{\mathcal{I}}\mathbb{E}[\mathcal{F}(\theta^{(r)}) - \mathcal{F}(\theta^*)] \\
&+ 3\alpha((4\varepsilon_{\mathcal{I}}s + 2) + \varepsilon_{\mathcal{I}^c}(d - k))\mathbb{E}\|\nabla f_{i_r}(\theta^*)\|_\infty^2 + 6\alpha\varepsilon_{abs}\mu^2 + 6\alpha A).
\end{aligned} \tag{22}$$

Let $L_t = \alpha\frac{n^2\varepsilon_\mu\mu^2}{\rho_s^{-2}} + \eta^2(3\alpha((4\varepsilon_{\mathcal{I}}s+2)+\varepsilon_{\mathcal{I}^c}(d-k))\mathbb{E}\|\nabla f_{i_r}(\theta^*)\|_\infty^2 + 6\alpha\varepsilon_{abs}\mu^2 + 6\alpha A_t)$, $\beta = (1+\eta^2\rho_s^{-2})\alpha$, then:

$$
\begin{aligned}
\mathbb{E}||\theta^{(r+1)} - \theta^*||_2^2 \leq (1+\eta^2\rho_s^{-2})\alpha\mathbb{E}||\theta^{(r)} - \theta^*||_2^2 + L - 2\eta\alpha\left[\mathcal{F}(\theta^{(r)}) - \mathcal{F}(\theta^*)\right] \\
+ \eta^2(\alpha\frac{24\rho_s^+}{n}\sum_{u=1}^{t-1}(1-\frac{q}{n})^{t-u-1}[\mathcal{F}(\theta^{(u)}) - \mathcal{F}(\theta^*)] + 24\alpha\rho_s^+(1-\frac{q}{n})^t[\mathcal{F}(\theta^{(0)}) - \mathcal{F}(\theta^*)] \\
+ 48\alpha\rho_s^+\varepsilon_{\mathcal{I}}\mathbb{E}[\mathcal{F}(\theta^{(r)}) - \mathcal{F}(\theta^*)].
\end{aligned}
\tag{23}
$$

Here, we use RSC and RSS condition, we have:

$$
\begin{aligned}
\mathcal{F}(\theta^{(r+1)}) - \mathcal{F}(\theta^*) &\leq (\frac{2\beta}{\rho_s^-} + 48\eta^2\alpha k\varepsilon_{\mathcal{I}} - 2\eta\alpha)\left[\mathcal{F}(\theta^{(r)}) - \mathcal{F}(\theta^*)\right] + L \\
&+ \eta^2(\alpha\frac{24\rho_s^+}{n}\sum_{u=1}^{r-1}(1-\frac{q}{n})^{r-u-1}[\mathcal{F}(\theta^{(u)}) - \mathcal{F}(\theta^*)] + 24\alpha\rho_s^+(1-\frac{q}{n})^t[\mathcal{F}(\theta^{(0)}) - \mathcal{F}(\theta^*)] \\
&- \frac{2}{\rho_s^+}\mathbb{E}\left\langle\nabla\mathcal{F}(\theta^*), \widetilde{\theta}^* - \theta^{(r+1)}\right\rangle + \frac{2\beta}{\rho_s^-}\mathbb{E}\left\langle\nabla\mathcal{F}(\theta^*), \widetilde{\theta}^* - \theta^{(r)}\right\rangle \\
&\leq (\frac{2\beta}{\rho_s^-} + 48\eta^2\alpha k\varepsilon_{\mathcal{I}} - 2\eta\alpha)\left[\mathcal{F}(\theta^{(r)}) - \mathcal{F}(\theta^*)\right] + L_t + \sqrt{s}||\nabla\mathcal{F}(\theta^*)||_\infty\mathbb{E}||\theta^{(r)} - \theta^*||_2 \\
&+ \eta^2(\alpha\frac{24\rho_s^+}{n}\sum_{u=1}^{t-1}(1-\frac{q}{n})^{t-u-1}\left[\mathcal{F}(\theta^{(u)}) - \mathcal{F}(\theta^*)\right] + 24\alpha\rho_s^+(1-\frac{q}{n})^t\left[\mathcal{F}(\theta^{(0)}) - \mathcal{F}(\theta^*)\right].
\end{aligned}
\tag{24}
$$

From Gu et al. (2020)[Lemma11], we have:

$$
\begin{aligned}
[\mathbb{E}\mathcal{F}(\theta^{(r+1)}) - \mathcal{F}(\theta^*)] &\leq (\left(\frac{2\beta}{\rho_s^-} + 48\eta^2\alpha\rho_s^+\varepsilon_{\mathcal{I}} - 2\eta\alpha + 1 - \frac{p}{n}\right))[\mathbb{E}\mathcal{F}(\theta^{(r)}) - \mathcal{F}(\theta^*)] \\
&+ 2(\alpha\frac{n^2\varepsilon_\mu\mu^2}{\rho_s^{-2}} + 6\alpha\varepsilon_{abs}\mu^2 + 6\eta^2\alpha A_r) + (\sqrt{s}||\nabla\mathcal{F}(\theta^*)||_\infty\mathbb{E}||\theta^{(r)} - \theta^*||_2 \\
&+ \eta^2(3\alpha((4\varepsilon_{\mathcal{I}}s+2)+\varepsilon_{\mathcal{I}^c}(d-k))\mathbb{E}\|\nabla f_{i_r}(\theta^*)\|_\infty^2))
\end{aligned}
\tag{25}
$$

□

5 VR-SZHT

5.1 ALGORITHM

Since the p-M algorithm is difficult to provide specific parameter analysis, we present a special case of the p-M algorithm, named VR-SZHT, for analysis. In the stochastic variance reduced gradient part, different from the stochastic gradient, we use $\hat{\nabla} f_{i_r}(\theta^{(r)}) - \hat{\nabla} f_{i_r}(\theta^{(0)}) + \hat{\mu}$ to replace $\hat{\nabla} f_i(\theta^{(r)})$. Through this iteration, we reduce the variance and stabilize the algorithm.

5.2 CONVERGENCE

In this section, we will provide a convergence analysis for VR-SZHT. Recall that the unknown sparse vector of interest is denoted as $\theta^* \in \mathbb{R}^d$, where $\|\theta^*\|_0 \geq k^*$, and the hard-thresholding operator $\mathcal{H}_k : \mathbb{R}^d \to \mathbb{R}^d$ keeps the largest k entries (in magnitude) and the rest is set to zero. For ease of notation, we use $\mathbb{E}(\cdot) = \mathbb{E}_{\boldsymbol{u},i_r}(\cdot)$, and we denote the full gradient and the stochastic variance reduced

Algorithm 3 Stochastic variance reduced zeroth-order Hard-Thresholding (VR-SZHT)

```
Input: Learning rate $\eta$, maximum number of iterations $T$, initial point $\theta^0$, SVRG update frequency $m$, number of random directions $q$, and number of coordinates to keep at each iteration $k$.
Output: $\theta^T$.
  for $r = 1, \dots, T$ do
    $\theta^{(0)} = \theta^{r-1}$;
    $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} \hat{\nabla} f_i(\theta^{(0)})$;
    for $t = 0, 1, \dots, m-1$ do
      Randomly sample $i_r \in \{1, 2, \dots, n\}$;
      Compute ZO estimate $\hat{\nabla} f_{i_r}(\theta^{(r)})$, $\hat{\nabla} f_{i_r}(\theta^{(0)})$;
      $\overline{\theta}^{(r+1)} = \theta^{(r)} - \eta(\hat{\nabla} f_{i_r}(\theta^{(r)}) - \hat{\nabla} f_{i_r}(\theta^{(0)}) + \hat{\mu})$;
      $\theta^{(r+1)} = \mathcal{H}_k(\overline{\theta}^{(r+1)})$;
    end for
    $\theta^r = \theta^{(t')}$, random $t' \in [m-1]$
  end for
```

gradient by:

$$
\begin{aligned}
\mu(\widetilde{\theta}) &= \nabla \mathcal{F}(\widetilde{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(\widetilde{\theta}), g^{(r)}(\theta^{(r)}) \\
&= \nabla f_{i_r}(\theta^t) - \nabla f_{i_r}(\theta^{(0)}) + \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(\theta^{(0)}); \\
\hat{\mu}(\widetilde{\theta}) &= \hat{\nabla} \mathcal{F}(\widetilde{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \hat{\nabla} f_i(\widetilde{\theta}), \hat{g}^{(r)}(\theta^{(r)}) \\
&= \hat{\nabla} f_{i_r}(\theta^t) - \hat{\nabla} f_{i_r}(\theta^{(0)}) + \frac{1}{n} \sum_{i=1}^{n} \hat{\nabla} f_i(\theta^{(0)}); \\
\hat{g}^{(r)}(\theta^{(r)}) &= \hat{\nabla} f_{i_r}(\theta^{(r)}) - \mathbb{E} \hat{\nabla} f_{i_r}(\widetilde{\theta}) + \hat{\mu}(\widetilde{\theta}).
\end{aligned} \tag{26}
$$

Theorem 4. *Suppose $\mathcal{F}(\theta)$ satisfies the RSC condition and that the functions $\{f_i(\theta)\}_{i=1}^{n}$satisfy the RSS condition with $s = 2k + k^*$. Let $\mathcal{I}^* = supp(\theta^*)$ denote the support of θ^*. Let $\theta^{(r)}$ be a sparse vector with $\|\theta^{(r)}\|_0 \le k$ and support $\mathcal{I}^{(r)} = supp(\theta^{(r)})$.*

$$
\widetilde{\mathcal{I}} = supp(\mathcal{H}_{2k}(\hat{\nabla} \mathcal{F}(\theta^*)) \cup supp(\theta^*).
$$

Then, we have:

$$
\begin{aligned}
&\frac{\beta^m - 1}{\beta - 1}(2\eta - 48\varepsilon_{\mathcal{I}} \eta^2 \rho_s^+)\alpha \left[\mathcal{F}(\widetilde{\theta}^{(r)}) - \mathcal{F}(\theta^*)\right] \le \\
&(\frac{2\beta^m}{\rho_s^-} + \frac{48\eta^2 \rho_s^+ \varepsilon_{\mathcal{I}} \alpha(\beta^m - 1)}{\beta - 1}) \mathbb{E}[\mathcal{F}(\widetilde{\theta}^{(r-1)}) - \mathcal{F}(\theta^*)] \\
&+ \frac{2\beta^m}{\rho_s^-} \sqrt{s} \|\nabla \mathcal{F}(\theta^*)\|_\infty \mathbb{E}\|\widetilde{\theta}^{(r-1)} - \theta^*\|_2 + \frac{\beta^m - 1}{\beta - 1} \alpha L.
\end{aligned} \tag{27}
$$

Before presenting the proof of convergence, it is necessary to examine the boundary of the (ZO) SVRG gradient $\hat{g}_{\mathcal{I}}^{(r)}(\theta^{(r)})$. For this purpose, we first derive the following three lemmas.

Lemma 4. *Under the condition of Theorem 4, for any $\mathcal{I} \supseteq (\mathcal{I}^* \cup \mathcal{I}^{(r)})$, we have $\mathbb{E}[\hat{g}^{(r)}(\theta^{(r)})] = \mathbb{E}_{\boldsymbol{u}} \hat{\nabla} \mathcal{F}(\theta^{(r)})$ and*

$$
\begin{aligned}
\mathbb{E}\|\hat{g}_{\mathcal{I}}^{(r)}(\theta^{(r)})\|_2^2 \le\ & 12\varepsilon_{\mathcal{I}} \mathbb{E}(\|\nabla f_{i_r}(\theta^t) - \nabla f_{i_r}(\theta^*)\|^2 \\
&+ \|\nabla f_{i_r}(\theta^0) - \nabla f_{i_r}(\theta^*)\|^2) \\
&+ 6((4\varepsilon_{\mathcal{I}} s + 2) + \varepsilon_{\mathcal{I}^c}(d - k)) \mathbb{E}\|\nabla f_{i_r}(\theta^*)\|_\infty^2 \\
&+ 12\varepsilon_{abs} \mu^2 + 3\|\nabla_{\mathcal{I}} \mathcal{F}(\theta^*)\|_2^2.
\end{aligned} \tag{28}
$$

Proof. It is straightforward that the stochastic variance reduced gradient satisfies:

$$\mathbb{E}\hat{g}^{(r)}(\theta^{(r)}) = \mathbb{E}\hat{\nabla} f_{i_r}(\theta^{(r)}) - \mathbb{E}\hat{\nabla} f_{i_r}(\widetilde{\theta}) + \hat{\mu}(\widetilde{\theta}) = \mathbb{E}_{\boldsymbol{u}}\hat{\nabla}\mathcal{F}(\theta^{(r)})$$

Thus $\hat{g}^{(r)}(\theta^{(r)})$ is a unbiased estimator of $\hat{\nabla}\mathcal{F}(\theta^{(r)})$. The form of an inapplicable zero order gradient is the same. As a result, the first claim is verified.

For the second claim, we have:

$$\begin{aligned}
\mathbb{E}\|g_{\mathcal{I}}^{(r)}(\theta^{(r)})\|_2^2 =& \mathbb{E}\hat{\nabla}\mathcal{F}(\theta^{(r)}) \\
\overset{①}{\leq}& 3\mathbb{E}\|\hat{\nabla}_{\mathcal{I}} f_{i_r}(\theta^{(r)}) - \nabla_{\mathcal{I}} f_{i_r}(\theta^*)\|_2^2 + 3\|\nabla_{\mathcal{I}}\mathcal{F}(\theta^*)\|_2^2 \\
&+ 3\mathbb{E}\|[\hat{\nabla}_{\mathcal{I}} f_{i_r}(\theta^{(0)}) - \nabla_{\mathcal{I}} f_{i_r}(\theta^*)] - \hat{\nabla}_{\mathcal{I}}\mathcal{F}(\theta^{(0)}) + \nabla_{\mathcal{I}}\mathcal{F}(\theta^*)\|_2^2 \\
=& 3\mathbb{E}\|\hat{\nabla}_{\mathcal{I}} f_{i_r}(\theta^{(r)}) - \nabla_{\mathcal{I}} f_{i_r}(\theta^*)\|_2^2 + 3\|\nabla_{\mathcal{I}}\mathcal{F}(\theta^*)\|_2^2 \\
&+ 3\mathbb{E}_{\boldsymbol{u}}[\mathbb{E}_{i_r}\|[\hat{\nabla}_{\mathcal{I}} f_{i_r}(\theta^{(0)}) - \nabla_{\mathcal{I}} f_{i_r}(\theta^*)] - \hat{\nabla}_{\mathcal{I}}\mathcal{F}(\theta^{(0)}) + \nabla_{\mathcal{I}}\mathcal{F}(\theta^*)\|_2^2] \\
\overset{②}{\leq}& 3\mathbb{E}\|\hat{\nabla}_{\mathcal{I}} f_{i_r}(\theta^{(r)}) - \nabla_{\mathcal{I}} f_{i_r}(\theta^*)\|_2^2 + 3\|\nabla_{\mathcal{I}}\mathcal{F}(\theta^*)\|_2^2 \\
&+ 3\mathbb{E}_{\boldsymbol{u}}[\mathbb{E}_{i_r}\|\hat{\nabla}_{\mathcal{I}} f_{i_r}(\theta^{(0)}) - \nabla_{\mathcal{I}} f_{i_r}(\theta^*)\|_2^2] \\
=& 3\mathbb{E}\|\hat{\nabla}_{\mathcal{I}} f_{i_r}(\theta^{(r)}) - \nabla_{\mathcal{I}} f_{i_r}(\theta^*)\|_2^2 + 3\|\nabla_{\mathcal{I}}\mathcal{F}(\theta^*)\|_2^2 \\
&+ 3\mathbb{E}\|\hat{\nabla}_{\mathcal{I}} f_{i_r}(\theta^{(0)}) - \nabla_{\mathcal{I}} f_{i_r}(\theta^*)\|_2^2
\end{aligned} \tag{29}$$

The inequality ① follows from the power mean inequality$\|a+b+c\|_2^2 \leq 3\|a\|_2^2+3\|b\|_2^2+3\|c\|_2^2$, and ② is follows from $\mathbb{E}\|x - \mathbb{E}x\|_2^2 \leq \mathbb{E}\|x\|_2^2$. Now we focus on $\mathbb{E}\|\hat{\nabla}_{\mathcal{I}} f_{i_r}(\theta) - \nabla_{\mathcal{I}} f_{i_r}(\theta^*)\|_2^2$. Actually, the boundary of $\mathbb{E}\|\hat{\nabla}_{\mathcal{I}} f_{i_r}(\theta) - \nabla_{\mathcal{I}} f_{i_r}(\theta^*)\|_2^2$ is available in Lemma 3, that is

$$\begin{aligned}
\mathbb{E}\|\hat{\nabla}_{\mathcal{I}} f_{i_r}(\theta) - \nabla_{\mathcal{I}} f_{i_r}(\theta^*)\|_2^2 \leq& 4\varepsilon_{\mathcal{I}}\mathbb{E}\|\nabla f_{i_r}(\theta^t) - \nabla f_{i_r}(\theta^*)\|^2 \\
&+ ((4\varepsilon_{\mathcal{I}} s + 2) + \varepsilon_{\mathcal{I}^c}(d-k))\mathbb{E}\|\nabla f_{i_r}(\theta^*)\|_\infty^2 + 2\varepsilon_{abs}\mu^2
\end{aligned} \tag{30}$$

Taking (30) into (29):

$$\begin{aligned}
\mathbb{E}\|g_{\mathcal{I}}^{(r)}(\theta^{(r)})\|_2^2 \leq& 12\varepsilon_{\mathcal{I}}\mathbb{E}(\|\nabla f_{i_r}(\theta^t) - \nabla f_{i_r}(\theta^*)\|^2 + \|\nabla f_{i_r}(\theta^0) - \nabla f_{i_r}(\theta^*)\|^2) \\
&+ 6((4\varepsilon_{\mathcal{I}} s + 2) + \varepsilon_{\mathcal{I}^c}(d-k))\mathbb{E}\|\nabla f_{i_r}(\theta^*)\|_\infty^2 + 12\varepsilon_{abs}\mu^2 + 3\|\nabla_{\mathcal{I}}\mathcal{F}(\theta^*)\|_2^2
\end{aligned}$$

□

The unbiasedness of SVRG implies that $\mathbb{E}[\hat{g}^{(r)}(\theta^{(r)})] = \mathbb{E}_{\boldsymbol{u}}\hat{\nabla}\mathcal{F}(\theta^{(r)})$ is obvious. As for inequality (49), we first use the mean inequality to divide it into three parts, and then prove it using Lemma 1. Now we provide the proof for Theorem 3:

Proof. We denote $v = \theta^{(r)} - \eta g_{\mathcal{I}}^{(r)}(\theta^{(r)})$ and $\mathcal{I} = \mathcal{I}^* \cup \mathcal{I}^{(r)} \cup \mathcal{I}^{(r+1)}$, where $\mathcal{I}^* = supp(\theta^*)$, $\mathcal{I}^{(r)} = supp(\theta^{(r)})$ and $\mathcal{I}^{(r+1)} = supp(\theta^{(r+1)})$

$$\begin{aligned}
\mathbb{E}||v-\theta^*||_2^2 &= \mathbb{E}||\theta^{(r)}-\theta^*||^2 + \eta^2\mathbb{E}||\hat{g}_{\mathcal{I}}^{(r)}(\theta^{(r)})||_2^2 - 2\eta\left\langle \theta^{(r)}-\theta^*, \mathbb{E}\hat{g}_{\mathcal{I}}^{(r)}(\theta^{(r)})\right\rangle \\
&= \mathbb{E}||\theta^{(r)}-\theta^*||_2^2 + \eta^2\mathbb{E}||\hat{g}_{\mathcal{I}}^{(r)}(\theta^{(r)})||_2^2 - 2\eta\left\langle \theta^{(r)}-\theta^*, \mathbb{E}\hat{\nabla}_{\mathcal{I}}\mathcal{F}(\theta^{(r)})\right\rangle \\
&= \mathbb{E}||\theta^{(r)}-\theta^*||_2^2 + \eta^2\mathbb{E}||\hat{g}_{\mathcal{I}}^{(r)}(\theta^{(r)})||_2^2 \\
&- 2\eta\mathbb{E}\left\langle \theta^{(r)}-\theta^*, \hat{\nabla}_{\mathcal{I}}\mathcal{F}(\theta^{(r)}) - \nabla_{\mathcal{I}}\mathcal{F}(\theta^{(r)})\right\rangle - 2\eta\mathbb{E}\left\langle \theta^{(r)}-\theta^*, \nabla_{\mathcal{I}}\mathcal{F}(\theta^{(r)})\right\rangle \\
&\leq \mathbb{E}||\theta^{(r)}-\theta^*||_2^2 + \eta^2\mathbb{E}||\hat{g}_{\mathcal{I}}^{(r)}(\theta^{(r)})||_2^2 - 2\eta\mathbb{E}\left\langle \theta^{(r)}-\theta^*, \hat{\nabla}_{\mathcal{I}}\mathcal{F}(\theta^{(r)}) - \nabla_{\mathcal{I}}\mathcal{F}(\theta^{(r)})\right\rangle \\
&- 2\eta\left[\mathcal{F}(\theta^{(r)}) - \mathcal{F}(\theta^*)\right] \\
&= \mathbb{E}||\theta^{(r)}-\theta^*||_2^2 + \eta^2\mathbb{E}||\hat{g}_{\mathcal{I}}^{(r)}(\theta^{(r)})||_2^2 - 2\eta\mathbb{E}\left\langle \sqrt{\eta}\rho_s^-(\theta^{(r)}-\theta^*), \frac{1}{\sqrt{\eta}\rho_s^-}(\hat{\nabla}_{\mathcal{I}}\mathcal{F}(\theta^{(r)}) - \nabla_{\mathcal{I}}\mathcal{F}(\theta^{(r)}))\right\rangle \\
&- 2\eta\left[\mathcal{F}(\theta^{(r)}) - \mathcal{F}(\theta^*)\right] \\
&\leq (1+\eta^2{\rho_s^-}^2)\mathbb{E}||\theta^{(r)}-\theta^*||_2^2 + \eta^2\mathbb{E}||\hat{g}_{\mathcal{I}}^{(r)}(\theta^{(r)})||_2^2 - 2\eta\left[\mathcal{F}(\theta^{(r)}) - \mathcal{F}(\theta^*)\right] \\
&+ {\frac{1}{\rho_s^-}}^2\mathbb{E}||(\hat{\nabla}_{\mathcal{I}}\mathcal{F}(\theta^{(r)}) - \nabla_{\mathcal{I}}\mathcal{F}(\theta^{(r)}))||_2^2
\end{aligned} \tag{31}$$

For $\mathbb{E}||(\hat{\nabla}_{\mathcal{I}}\mathcal{F}(\theta^{(r)}) - \nabla_{\mathcal{I}}\mathcal{F}(\theta^{(r)}))||_2^2$, we have:

$$\begin{aligned}
\mathbb{E}||(\hat{\nabla}_{\mathcal{I}}\mathcal{F}(\theta^{(r)}) - \nabla_{\mathcal{I}}\mathcal{F}(\theta^{(r)}))||_2^2 &= ||\mathbb{E}_{\boldsymbol{u}}(\hat{\nabla}_{\mathcal{I}}\sum_{i=1}^n f_i(\theta^{(r)}) - \nabla_{\mathcal{I}}\sum_{i=1}^n f_i(\theta^{(r)}))||_2^2 \\
&\leq n\sum_{i=1}^n ||\mathbb{E}_{\boldsymbol{u}}(\hat{\nabla}_{\mathcal{I}}f_i(\theta^{(r)}) - \nabla_{\mathcal{I}}f_i(\theta^{(r)}))||_2^2 \\
&\leq n^2\varepsilon_\mu\mu^2
\end{aligned} \tag{32}$$

The first inequality follows from N-dimensional mean inequality and the second inequality follows from Assumption 2. By constraining $\mathbb{E}||(\hat{\nabla}_{\mathcal{I}}\mathcal{F}(\theta^{(r)}) - \nabla_{\mathcal{I}}\mathcal{F}(\theta^{(r)}))||_2^2$, we can turn (46) into:

$$\begin{aligned}
\mathbb{E}||\theta^{(r)} - \eta\hat{g}_{\mathcal{I}}^{(r)}(\theta^{(r)}) - \theta^*||_2^2 &\leq (1+\eta^2{\rho_s^-}^2)\mathbb{E}||\theta^{(r)}-\theta^*||_2^2 + \eta^2\mathbb{E}||\hat{g}_{\mathcal{I}}^{(r)}(\theta^{(r)})||_2^2 - 2\eta\left[\mathcal{F}(\theta^{(r)}) - \mathcal{F}(\theta^*)\right] \\
&+ \frac{n^2\varepsilon_\mu\mu^2}{{\rho_s^-}^2}
\end{aligned} \tag{33}$$

Let $\alpha = 1 + \frac{2\sqrt{k^*}}{\sqrt{k-k^*}}$. Using Lemma 2, we have:

$$\begin{aligned}
\mathbb{E}||\theta^{(r+1)}-\theta^*||_2^2 &\leq (1+\eta^2{\rho_s^-}^2)\alpha\mathbb{E}||\theta^{(r)}-\theta^*||_2^2 + \eta^2\alpha\mathbb{E}||\hat{g}_{\mathcal{I}}^{(r)}(\theta^{(r)})||_2^2 - 2\eta\alpha\left[\mathcal{F}(\theta^{(r)}) - \mathcal{F}(\theta^*)\right] \\
&+ \alpha\frac{n^2\varepsilon_\mu\mu^2}{{\rho_s^-}^2} \\
&\leq (1+\eta^2{\rho_s^-}^2)\alpha\mathbb{E}||\theta^{(r)}-\theta^*||_2^2 + \alpha\frac{n^2\varepsilon_\mu\mu^2}{{\rho_s^-}^2} - 2\eta\alpha\left[\mathcal{F}(\theta^{(r)}) - \mathcal{F}(\theta^*)\right] \\
&+ \eta^2\alpha(12\varepsilon_{\mathcal{I}}\mathbb{E}(||\nabla f_{i_r}(\boldsymbol{x}^t) - \nabla f_{i_r}(\theta^*)||^2 + ||\nabla f_{i_r}(\theta^0) - \nabla f_{i_r}(\theta^*)||^2) \\
&+ 6((4\varepsilon_{\mathcal{I}}s+2) + \varepsilon_{\mathcal{I}^c}(d-k))\mathbb{E}||\nabla f_{i_r}(\theta^*)||_\infty^2 + 12\varepsilon_{abs}\mu^2 + 3||\nabla_{\mathcal{I}}\mathcal{F}(\theta^*)||_2^2)
\end{aligned} \tag{34}$$

For $||\nabla f_{i_r}(\boldsymbol{\theta}^t) - \nabla f_{i_r}(\theta^*)||^2$, we can easily get

$$\mathbb{E}||\nabla f_{i_r}(\boldsymbol{\theta}^t) - \nabla f_{i_r}(\theta^*)||^2 = \frac{1}{n}||\nabla f_{i_r}(\boldsymbol{\theta}^t) - \nabla f_{i_r}(\theta^*)||^2 \leq 4\rho_s^+[\mathcal{F}(\theta) - \mathcal{F}(\theta^*)]$$

by RSS condition. As a reasult, we have:

$$\begin{aligned}\mathbb{E}||\theta^{(r+1)} - \theta^*||_2^2 \leq & (1+\eta^2\rho_s^{-2})\alpha\mathbb{E}||\theta^{(r)} - \theta^*||_2^2 + \alpha\frac{n^2\varepsilon_\mu\mu^2}{\rho_s^{-2}} - \alpha(2\eta - 48\varepsilon_\mathcal{I}\eta^2\rho_s^+)\left[\mathcal{F}(\theta^{(r)}) - \mathcal{F}(\theta^*)\right] \\ & + 48\eta^2\rho_s^+\alpha\varepsilon_\mathcal{I}\mathbb{E}[\mathcal{F}(\theta^0) - \mathcal{F}(\theta^*)] \\ & + 6\eta^2\alpha((4\varepsilon_\mathcal{I}s+2) + \varepsilon_{\mathcal{I}^c}(d-k))\mathbb{E}||\nabla f_{i_r}(\theta^*)||_\infty^2 + 12\varepsilon_{abs}\mu^2 + 3||\nabla_\mathcal{I}\mathcal{F}(\theta^*)||_2^2)\end{aligned} \tag{35}$$

Let $L = 6\eta^2((4\varepsilon_\mathcal{I}s+2) + \varepsilon_{\mathcal{I}^c}(d-k))\mathbb{E}||\nabla f_{i_r}(\theta^*)||_\infty^2 + 12\varepsilon_{abs}\mu^2 + 3||\nabla_\mathcal{I}\mathcal{F}(\theta^*)||_2^2) + \frac{n^2\varepsilon_\mu\mu^2}{\rho_s^{-2}}$, $\beta = (1+\eta^2\rho_s^{-2})\alpha$, then:

$$\begin{aligned}\mathbb{E}||\theta^{(r+1)} - \theta^*||_2^2 + \alpha(2\eta - 48\varepsilon_\mathcal{I}\eta^2\rho_s^+)\left[\mathcal{F}(\theta^{(r)}) - \mathcal{F}(\theta^*)\right] \leq & \beta\mathbb{E}||\theta^{(r)} - \theta^*||_2^2 \\ & + 48\eta^2\rho_s^+\varepsilon_\mathcal{I}\alpha\mathbb{E}[\mathcal{F}(\theta^0) - \mathcal{F}(\theta^*)] + \alpha L\end{aligned} \tag{36}$$

By summing (51) over $t = 0, \dots, m-1$, we have:

$$\begin{aligned}\mathbb{E}||\theta^{(m)} - \theta^*||_2^2 + \frac{\beta^m - 1}{\beta - 1}(2\eta - 48\varepsilon_\mathcal{I}\eta^2\rho_s^+)\alpha\left[\mathcal{F}(\widetilde{\theta}^{(r)}) - \mathcal{F}(\theta^*)\right] \leq & \beta^m\mathbb{E}||\theta^{(r)} - \theta^*||_2^2 \\ & + 48\eta^2\rho_s^+\varepsilon_\mathcal{I}\alpha\frac{\beta^m - 1}{\beta - 1}\mathbb{E}[\mathcal{F}(\widetilde{\theta}^{(r-1)}) - \mathcal{F}(\theta^*)] + \frac{\beta^m - 1}{\beta - 1}\alpha\end{aligned} \tag{37}$$

Through RSC condition and the definition of $\widetilde{I}$, it further follows from (52) that:

$$\begin{aligned}\frac{\beta^m - 1}{\beta - 1}(2\eta - 48\varepsilon_\mathcal{I}\eta^2\rho_s^+)\alpha\left[\mathcal{F}(\widetilde{\theta}^{(r)}) - \mathcal{F}(\theta^*)\right] \leq & \left(\frac{2\beta^m}{\rho_s^-} + \frac{48\eta^2\rho_s^+\varepsilon_\mathcal{I}\alpha(\beta^m - 1)}{\beta - 1}\right)\mathbb{E}[\mathcal{F}(\widetilde{\theta}^{(r-1)}) - \mathcal{F}(\theta^*)] \\ & + \frac{2\beta^m}{\rho_s^-}\mathbb{E}\left\langle\nabla\mathcal{F}(\theta^*), \widetilde{\theta}^{(r-1)} - \theta^*\right\rangle + \frac{\beta^m - 1}{\beta - 1}\alpha L \\ \leq & \left(\frac{2\beta^m}{\rho_s^-} + \frac{48\eta^2\rho_s^+\varepsilon_\mathcal{I}\alpha(\beta^m - 1)}{\beta - 1}\right)\mathbb{E}[\mathcal{F}(\widetilde{\theta}^{(r-1)}) - \mathcal{F}(\theta^*)] \\ & + \frac{2\beta^m}{\rho_s^-}\sqrt{s}||\nabla\mathcal{F}(\theta^*)||_\infty\mathbb{E}||\widetilde{\theta}^{(r-1)} - \theta^*||_2 + \frac{\beta^m - 1}{\beta - 1}\alpha L\end{aligned} \tag{38}$$

Here $\varepsilon_\mu = \rho_s^{+2}sd$, $\varepsilon_\mathcal{I} = \frac{2d}{q(s_2+2)}(\frac{(s-1)(s_2-1)}{d-1} + 3) + 2$, $\varepsilon_{\mathcal{I}^c} = \frac{2d}{q(s_2+2)}\left(\frac{s(s_2-1)}{-1}\right)$, $\varepsilon_{abs} = \frac{2d\rho_s^{+2}ss_2}{q}\left(\frac{(s-1)(s_2-1)}{d-1} + 1\right) + \rho_s^{+2}sd$ □

5.3 RELATIONSHIP BETWEEN PARAMETERS

Upon completing the proof, we can conclude that the convergence of the algorithm is contingent on the coefficient of $\left[\mathcal{F}(\widetilde{\theta}^{(r)}) - \mathcal{F}(\theta^*)\right]$, $[\mathcal{F}(\widetilde{\theta}^{(r-1)}) - \mathcal{F}(\theta^*)]$, which is determined by the values of m, η, $\varepsilon_\mathcal{I}$ and α. In the subsequent section, we will investigate the relationship among these parameters. Recall (53), we have

$$\begin{aligned}\frac{\beta^m - 1}{\beta - 1}(2\eta - 48\varepsilon_\mathcal{I}\eta^2\rho_s^+)\alpha\left[\mathcal{F}(\widetilde{\theta}^{(r)}) - \mathcal{F}(\theta^*)\right] \leq & (\frac{2\beta^m}{\rho_s^-} + \frac{48\eta^2\rho_s^+\varepsilon_\mathcal{I}\alpha(\beta^m - 1)}{\beta - 1})\mathbb{E}[\mathcal{F}(\widetilde{\theta}^{(r-1)}) - \mathcal{F}(\theta^*)] \\ & + \frac{2\beta^m}{\rho_s^-}\sqrt{s}||\nabla\mathcal{F}(\theta^*)||_\infty\mathbb{E}||\widetilde{\theta}^{(r-1)} - \theta^*||_2 + \frac{\beta^m - 1}{\beta - 1}\alpha L\end{aligned}$$

Actually, if the algorithm converges, we have:

$$\frac{\beta - 1}{\alpha(\beta^m - 1)(2\eta - 48\varepsilon_\mathcal{I}\eta^2\rho_s^+)}(\frac{2\beta^m}{\rho_s^-} + \frac{48\eta^2\rho_s^+\varepsilon_\mathcal{I}\alpha(\beta^m - 1)}{\beta - 1}) \leq 1,$$

which is:

$$\frac{24\varepsilon_{\mathcal{I}}\eta\rho_s^+}{(1-24\varepsilon_{\mathcal{I}}\eta\rho_s^+)} + \frac{(\beta-1)\beta^m}{\eta\alpha\rho_s^-(\beta^m-1)(1-24\varepsilon_{\mathcal{I}}\eta\rho_s^+)} \leq 1. \tag{39}$$

(39) provides us with a more detailed understanding of the relationship between the zeroth-order gradient estimator, the stochastic variance reduced gradient estimator, and the hard-thresholding operator when the algorithm converges. Based on this equation, we can derive two corollaries as follows:

Corollary 3. *When the learning rate $\eta \in [\eta_{\min}, \min\{\eta_{\max}, \frac{1}{48\varepsilon_{\mathcal{I}}\rho_s^+}\}]$, there always exists an m such that the algorithm converges.Here*

$$\eta_{\max} = \frac{\alpha\rho_s^- + \sqrt{(\alpha\rho_s^-)^2 - 4(48\varepsilon_{\mathcal{I}}\alpha\rho_s^-\rho_s^+ + {\rho_s^-}^2)(\alpha-1)}}{2(48\varepsilon_{\mathcal{I}}\alpha\rho_s^-\rho_s^+ + {\rho_s^-}^2)};$$

$$\eta_{\min} = \frac{\alpha\rho_s^- - \sqrt{(\alpha\rho_s^-)^2 - 4(48\varepsilon_{\mathcal{I}}\alpha\rho_s^-\rho_s^+ + {\rho_s^-}^2)(\alpha-1)}}{2(48\varepsilon_{\mathcal{I}}\alpha\rho_s^-\rho_s^+ + {\rho_s^-}^2)}.$$

Remark 5. *This corollary reveals that there is a range of values for η. When $\eta < \frac{1}{48\varepsilon_{\mathcal{I}}\rho_s^+}$, we have*

$$\beta^{-m} \leq \frac{(1-48\varepsilon_{\mathcal{I}}\eta\rho_s^+)\eta\alpha\rho_s^- - (\beta-1)}{(1-48\varepsilon_{\mathcal{I}}\eta\rho_s^+)\eta\alpha\rho_s^-}. \tag{40}$$

This means that if

$$(1-48\varepsilon_{\mathcal{I}}\eta\rho_s^+)\eta\alpha\rho_s^- - (\beta-1) > 0, \tag{41}$$

holds, there is always a suitable value of m that makes the algorithm converge. We recall $\beta = (1+\eta^2{\rho_s^-}^2)\alpha$ and (40) becomes:

$$(48\varepsilon_{\mathcal{I}}\alpha\rho_s^-\rho_s^+ + {\rho_s^-}^2)\eta^2 - \alpha\rho_s^-\eta + \alpha - 1 < 0. \tag{42}$$

If we solve (42) , we will obtain $\eta_{\min} < \eta < \eta_{\max}$. To achieve optimal convergence, we suggest setting $\eta = \frac{\alpha\rho_S^-}{2(48\varepsilon_{\mathcal{I}}\alpha\rho_s^-\rho_s^+ + {\rho_s^-}^2)}$. We will revisit this recommendation later on in Remark 6.

Remark 6. *The value of k in equation (42) is subject to a limitation. For the algorithm to converge, we require that $\alpha^2 > 4(48\varepsilon_{\mathcal{I}}\alpha\kappa_s + 1)(\alpha-1)$, which is a polynomial of k. Therefore, we can obtain the boundary of k by solving this inequality. Note that SARAH-SZHT can be proven in the same way as VR-SZHT, so it will not be repeated in this paper.*

Proof. let $\eta \leq \frac{C_3}{\varepsilon_{\mathcal{I}}\rho_s^+} < \frac{1}{72\varepsilon_{\mathcal{I}}\rho_s^+} < \frac{C_3}{144\rho_s^+}$, then

$$\frac{24\varepsilon_{\mathcal{I}}\eta\rho_s^+}{1-24\varepsilon_{\mathcal{I}}\eta\rho_s^+} < \frac{24C_3}{1-24C_3} < \frac{1}{2}$$

If $k \geq C_1\kappa_s^2k^*$, $\eta \geq \frac{C_2}{\rho_s^+}$ with $C_2 < C_3 < C_3\varepsilon_{\mathcal{I}}$, and $A = \frac{2}{\sqrt{C_1 - 1}} + \frac{C_3^2}{4\kappa_s} + \frac{C_3^2}{2\sqrt{C_1 - 1}\kappa_s^2}$. Then we have $\alpha \leq 1 + \frac{2}{\sqrt{C_1 - 1}\kappa_s}$, $\beta \leq 1 + \frac{A}{\kappa_s}$ and

$$\begin{aligned}\frac{\beta^m(\beta-1)}{\eta\rho_s^-\alpha(1-24\varepsilon_{\mathcal{I}}\eta\rho_s^+)(\beta^m-1)} &\leq \frac{\frac{A}{\kappa_s}}{\frac{2C_2}{3\kappa_s}(1-(1+\frac{A}{\kappa_s})^{-m})} \\ &= \frac{3A}{2C_2(1-(1+\frac{A}{\kappa_s})^{-m})}\end{aligned}$$

It is guaranteed $\frac{\beta^m(\beta-1)}{\eta\rho_s^-\alpha(1-24\varepsilon_{\mathcal{I}}\eta\rho_s^+)(\beta^m-1)} < \frac{1}{2}$ if we have

$$m > \log_{1+\frac{A}{\kappa_s}}\frac{C_2}{C_2-3A} = \frac{\log\frac{C_2}{C_2-3A}}{\log(1+\frac{A}{\kappa_s})}$$

Using the fact that $ln(1+x) > \frac{x}{2}$

$$\frac{\log \frac{C_2}{C_2-3A}}{\log(1+\frac{A}{\kappa_s})} \leq \log \frac{C_2}{C_2-3A}\left(\frac{2\kappa_s}{A}\right)$$

Then $\frac{\beta^m(\beta-1)}{\eta\rho_s^- \alpha(1-24\varepsilon_{\mathcal{I}}\eta\rho_s^+)(\beta^m-1)} < \frac{1}{2}$ holds if m satisfies

$$\begin{aligned} m &\geq \log\left(\frac{C_2}{C_2-3A}\right)\left(\frac{2\kappa_s}{A}\right) \\ &= \log\left(\frac{C_2}{C_2-3A}\right)\left(\frac{2\kappa_s^3}{\frac{2}{\sqrt{C_1-1}}\kappa_s^2 + \frac{C_3^2}{4}\kappa_s + \frac{C_3^2}{2\sqrt{C_1-1}}}\right) \\ &= \mathcal{O}\left(\frac{\kappa^3}{\kappa^2+1}\right) \end{aligned}$$

if we want to contorl the error below ε, we need $\mathcal{O}\left(\log(\frac{1}{\varepsilon})\right)$ outer iterations. And for each outer iterations, we need to calculate a full (ZO) gradient and m stochastic variance reduced (ZO) gradients. So the query complexity of the algorithm is $[n + \frac{\kappa^3}{\kappa^2+1}]\log\left(\frac{1}{\varepsilon}\right)$. When κ_s is large, the complexity will tend to $[n+\kappa]\log\left(\frac{1}{\varepsilon}\right)$. When κ_s is small, the complexity will tend to $[n+\kappa^3]\log\left(\frac{1}{\varepsilon}\right)$. Thus VR-SZHT yields a significant improvement over SZOHT. ☐

Corollary 4. *The query complexity of the algorithm is* $\mathcal{O}([n + \frac{\kappa^3}{\kappa^2+1}]\log\left(\frac{1}{\varepsilon}\right))$.

If we want the error to be below ε, we need $\mathcal{O}\left(\log(\frac{1}{\varepsilon})\right)$ outer iteration. For each outer iteration, we need to calculate a full (ZO) gradient and m stochastic variance reduced (ZO) gradients. So the query complexity of the algorithm is $\mathcal{O}([n + \frac{\kappa^3}{\kappa^2+1}]\log\left(\frac{1}{\varepsilon}\right))$. When κ_s is large, the complexity will tend to $\mathcal{O}([n+\kappa]\log\left(\frac{1}{\varepsilon}\right))$. When κ_s is small, the complexity will tend to $\mathcal{O}([n+\kappa^3]\log\left(\frac{1}{\varepsilon}\right))$. Thus, VR-SZHT yields a significant improvement over SZOHT. See the appendix for the specific proof.

6 SARAH-SZHT

Algorithm: With the development of variance reduction theory, an iterative form of gradient descent algorithm has emerged, such as SARAH Nguyen et al. (2017a), etc. Such algorithms are similar to SVRG, in that they utilize both inner and outer loops. Unfortunately, the p-Memorization framework is not applicable to such algorithms. Therefore, in this section, we will attempt to apply the iterative form of variance descent algorithm to gradient simulation.

Algorithm 4 Stochastic variance reduced zeroth-order Hard-Thresholding with SARAH (SARAH-SZHT)

```
Input: Learning rate $\eta$, maximum number of iterations $T$, initial point $\theta^0$, number of random
    directions $q$, maximum error $\varepsilon$, and number of coordinates to keep at each iteration $k$.
Output: $\theta^T$.
for $r = 1, \ldots, T$ do
    $\theta^0 = \widetilde{\theta}^{(r-1)}$
    $g^{(0)} = \frac{1}{n}\sum_{i=1}^{n} \hat{\nabla} f_i(\theta_0)$
    $\theta^{(1)} = \theta^{(0)} - \eta g^{(0)}$
    for $t = 1, ..., m-1$ do
        Sample $i_t$ uniformly at random from $[n]$
        $\hat{g}^{(t)} = \hat{\nabla} f_{i_t}(\theta^{(t)}) - \hat{\nabla} f_{i_t}(\theta^{(t-1)}) + \hat{g}^{t-1}$
        $\theta^{(t+1)} = \theta^{(t)} - \eta g^t$
    end for
    Set $\widetilde{\theta}^{(r)} = \theta^{(d)}$ with $d$ chosen uniformly at random from $\{0, 1, \ldots, m\}$.
end for
```

Convergence Analysis: In this section, we will provide a convergence analysis for SARAH-SZHT. using the assumptions from section 2, and discussing the relationship between parameters, providing a positive answer to the question from the section. Following the analysis in section 3, we first obtain the boundary of $\mathbb{E}\|g_{\mathcal{I}}^{(t)}(\theta^{(t)})\|_2^2$.

Lemma 5. *Suppose that the Assumption 1 and 3, for $\eta < 2/\rho_s^+$ we can get:*

$$\begin{aligned}\mathbb{E}\|g_{\mathcal{I}}^{(t)}(\theta^{(t)})\|_2^2 \le{}& 12\varepsilon_{\mathcal{I}}\mathbb{E}(\|\nabla f_{i_t}(\theta^{(t)}) - \nabla f_{i_t}(\theta^*)\|^2 + \|\nabla f_{i_t}(\theta^{(t-1)}) - \nabla f_{i_t}(\theta^*)\|^2)\\ &+ 6((4\varepsilon_{\mathcal{I}}s + 2) + \varepsilon_{\mathcal{I}^c}(d-k))\mathbb{E}\|\nabla f_{i_t}(\theta^*)\|_\infty^2 + 12\varepsilon_{abs}\mu^2 + 3\|\nabla_{\mathcal{I}}\mathcal{F}(\theta^*)\|_2^2.\end{aligned} \tag{43}$$

Therefore, we derive the following conclusion.

Theorem 5. *Suppose the same assumption as Theorem 1. For Algorithm 4, we have:*

$$\begin{aligned}\gamma'\mathbb{E}\left[\mathcal{F}(\widetilde{\theta}^{(r)}) - \mathcal{F}(\theta^*)\right] \le{}& \left(\frac{2\beta^{m-1}}{\rho_s^-} + 48\eta^2\rho_s^+\varepsilon_{\mathcal{I}}\alpha\beta^{m-2}\right)\mathbb{E}[\mathcal{F}(\widetilde{\theta}^{(r-1)}) - \mathcal{F}(\theta^*)]\\ &+ \frac{2\beta^{m-1}}{\rho_s^-}\sqrt{s}\|\nabla\mathcal{F}(\theta^*)\|_\infty\mathbb{E}\|\widetilde{\theta}^{(r-1)} - \theta^*\|_2 + \frac{\beta^{m-1}-1}{\beta-1}\alpha L,\end{aligned}$$

here $\gamma' = \left(\frac{\beta^{m-1}-1}{\beta-1}2\eta - \left(\frac{\beta^{m-1}-1}{\beta-1} + \frac{\beta^{m-2}-1}{\beta-1}\right)48\varepsilon_{\mathcal{I}}\eta^2\rho_s^+)\alpha\right)$.

Based on the coefficient of $\mathbb{E}\left[\mathcal{F}(\widetilde{\theta}^{(r)}) - \mathcal{F}(\theta^*)\right]$ and $\mathbb{E}\left[\mathcal{F}(\widetilde{\theta}^{(r-1)}) - \mathcal{F}(\theta^*)\right]$, we derive the boundary of η when the algorithm converges.

Corollary 5. *If*

$$\frac{\alpha - \sqrt{\alpha^2 - 4(48\varepsilon_{\mathcal{I}}\alpha\rho_s^+ + \alpha\rho_s^-)(\alpha-1)}}{2(48\varepsilon_{\mathcal{I}}\alpha\rho_s^+ + \alpha\rho_s^-)} \le \eta \le \frac{\alpha + \sqrt{\alpha^2 - 4(48\varepsilon_{\mathcal{I}}\alpha\rho_s^+ + \alpha\rho_s^-)(\alpha-1)}}{2(48\varepsilon_{\mathcal{I}}\alpha\rho_s^+ + \alpha\rho_s^-)},$$

there always exists an m such that the algorithm converges.

Remark 7. *Like pM-SZHT, from corollary 5, we know that q does not play a decisive role in convergence. This indicates that it is feasible to use variance reduction algorithms instead of restricting q to make the algorithm converge.*

Remark 8. *Unfortunately, in the SARAH-SZHT algorithm, the gradient estimation $\hat{g}$ will accumulate errors during the inner loop process. This will cause the algorithm to not converge when m is large.*

Corollary 6. *The query complexity of the algorithm is $\mathcal{O}([n + \frac{\kappa^3}{\kappa^2+1}]\log(\frac{1}{\varepsilon}))$.*

6.1 PROOF OF LEMMA 2

Proof. It is straightforward that the stochastic variance reduced gradient satisfies:

$$\mathbb{E}\hat{g}^{(r)}(\theta^{(r)}) = \mathbb{E}\hat{\nabla}f_{i_r}(\theta^{(r)}) - \mathbb{E}\hat{\nabla}f_{i_r}(\theta^{(r-1)}) + \hat{g}^{(r-1)}(\theta^{(r-1)}) = \mathbb{E}\hat{\nabla}f_{i_r}(\theta^{(r)}).$$

Thus $\hat{g}^{(r)}(\theta^{(r)})$ is a unbiased estimator of $\hat{\nabla}\mathcal{F}(\theta^{(r)})$. The form of inapplicable zero order gradient is the same. As a result, the first claim is verified.

For the second claim, we have:

$$\begin{aligned}\mathbb{E}\|g_{\mathcal{I}}^{(r)}(\theta^{(r)})\|_2^2 \overset{①}{\le}{}& 3\mathbb{E}\|\hat{\nabla}_{\mathcal{I}}f_{i_r}(\theta^{(r)}) - \nabla_{\mathcal{I}}f_{i_r}(\theta^*)\|_2^2 + 3\|\nabla_{\mathcal{I}}\mathcal{F}(\theta^*)\|_2^2\\ &+ 3\mathbb{E}\|[\hat{\nabla}_{\mathcal{I}}f_{i_r}(\theta^{(r-1)}) - \nabla_{\mathcal{I}}f_{i_r}(\theta^*)] - \hat{g}_{\mathcal{I}}^{(r-1)}(\theta^{(r-1)}) + \nabla_{\mathcal{I}}\mathcal{F}(\theta^*)\|_2^2\\ ={}& 3\mathbb{E}\|\hat{\nabla}_{\mathcal{I}}f_{i_r}(\theta^{(r)}) - \nabla_{\mathcal{I}}f_{i_r}(\theta^*)\|_2^2 + 3\|\nabla_{\mathcal{I}}\mathcal{F}(\theta^*)\|_2^2\\ &+ 3\mathbb{E}_{\boldsymbol{u}}[\mathbb{E}_{i_r}\|[\hat{\nabla}_{\mathcal{I}}f_{i_r}(\theta^{(r-1)}) - \nabla_{\mathcal{I}}f_{i_r}(\theta^*)] - \hat{\nabla}_{\mathcal{I}}\mathcal{F}(\theta^{(r-1)}) + \nabla_{\mathcal{I}}\mathcal{F}(\theta^*)\|_2^2]\\ \overset{②}{\le}{}& 3\mathbb{E}\|\hat{\nabla}_{\mathcal{I}}f_{i_r}(\theta^{(r)}) - \nabla_{\mathcal{I}}f_{i_r}(\theta^*)\|_2^2 + 3\|\nabla_{\mathcal{I}}\mathcal{F}(\theta^*)\|_2^2\\ &+ 3\mathbb{E}_{\boldsymbol{u}}[\mathbb{E}_{i_r}\|\hat{\nabla}_{\mathcal{I}}f_{i_r}(\theta^{(r-1)}) - \nabla_{\mathcal{I}}f_{i_r}(\theta^*)\|_2^2]\\ ={}& 3\mathbb{E}\|\hat{\nabla}_{\mathcal{I}}f_{i_r}(\theta^{(r)}) - \nabla_{\mathcal{I}}f_{i_r}(\theta^*)\|_2^2 + 3\|\nabla_{\mathcal{I}}\mathcal{F}(\theta^*)\|_2^2\\ &+ 3\mathbb{E}\|\hat{\nabla}_{\mathcal{I}}f_{i_r}(\theta^{(r-1)}) - \nabla_{\mathcal{I}}f_{i_r}(\theta^*)\|_2^2.\end{aligned} \tag{44}$$

The inequality ① follows from the power mean inequality $\|a+b+c\|_2^2 \leq 3\|a\|_2^2+3\|b\|_2^2+3\|c\|_2^2$, and ② is follows from $\mathbb{E}\|x-\mathbb{E}x\|_2^2 \leq \mathbb{E}\|x\|_2^2$. Now we focus on $\mathbb{E}\|\hat{\nabla}_{\mathcal{I}} f_{i_r}(\theta) - \nabla_{\mathcal{I}} f_{i_r}(\theta^*)\|_2^2$. Actually, the boundary of $\mathbb{E}\|\hat{\nabla}_{\mathcal{I}} f_{i_r}(\theta) - \nabla_{\mathcal{I}} f_{i_r}(\theta^*)\|_2^2$ is available in Lemma 2, that is

$$\begin{aligned} \mathbb{E}\|\hat{\nabla}_{\mathcal{I}} f_{i_r}(\theta) - \nabla_{\mathcal{I}} f_{i_r}(\theta^*)\|_2^2 \leq{} & 4\varepsilon_{\mathcal{I}} \mathbb{E}\|\nabla f_{i_r}(\theta^t) - \nabla f_{i_r}(\theta^*)\|^2 \\ & + ((4\varepsilon_{\mathcal{I}} s + 2) + \varepsilon_{\mathcal{I}^c}(d-k))\mathbb{E}\|\nabla f_{i_r}(\theta^*)\|_\infty^2 + 2\varepsilon_{abs}\mu^2. \end{aligned} \tag{45}$$

Taking (45) into (44):

$$\begin{aligned} \mathbb{E}\|g_{\mathcal{I}}^{(r)}(\theta^{(r)})\|_2^2 \leq{} & 12\varepsilon_{\mathcal{I}} \mathbb{E}(\|\nabla f_{i_r}(\theta^{(r)}) - \nabla f_{i_r}(\theta^*)\|^2 + \|\nabla f_{i_r}(\theta^{(r-1)}) - \nabla f_{i_r}(\theta^*)\|^2) \\ & + 6((4\varepsilon_{\mathcal{I}} s + 2) + \varepsilon_{\mathcal{I}^c}(d-k))\mathbb{E}\|\nabla f_{i_r}(\theta^*)\|_\infty^2 + 12\varepsilon_{abs}\mu^2 + 3\|\nabla_{\mathcal{I}} \mathcal{F}(\theta^*)\|_2^2. \end{aligned}$$

□

6.2 PROOF OF THEOREM 2

Proof. We denote $v = \theta^{(r)} - \eta g_{\mathcal{I}}^{(r)}(\theta^{(r)})$ and $\mathcal{I} = \mathcal{I}^* \cup \mathcal{I}^{(r)} \cup \mathcal{I}^{(r+1)}$, where $\mathcal{I}^* = supp(\theta^*)$, $\mathcal{I}^{(r)} = supp(\theta^{(r)})$ and $\mathcal{I}^{(r+1)} = supp(\theta^{(r+1)})$

$$\begin{aligned} \mathbb{E}||v-\theta^*||_2^2 &= \mathbb{E}||\theta^{(r)} - \theta^*||^2 + \eta^2 \mathbb{E}||\hat{g}_{\mathcal{I}}^{(r)}(\theta^{(r)})||_2^2 - 2\eta \left\langle \theta^{(r)} - \theta^*, \mathbb{E}\hat{g}_{\mathcal{I}}^{(r)}(\theta^{(r)}) \right\rangle \\ &= \mathbb{E}||\theta^{(r)} - \theta^*||_2^2 + \eta^2 \mathbb{E}||\hat{g}_{\mathcal{I}}^{(r)}(\theta^{(r)})||_2^2 - 2\eta \left\langle \theta^{(r)} - \theta^*, \mathbb{E}\hat{\nabla}_{\mathcal{I}} \mathcal{F}(\theta^{(r)}) \right\rangle \\ &= \mathbb{E}||\theta^{(r)} - \theta^*||_2^2 + \eta^2 \mathbb{E}||\hat{g}_{\mathcal{I}}^{(r)}(\theta^{(r)})||_2^2 \\ &\quad - 2\eta \mathbb{E}\left\langle \theta^{(r)} - \theta^*, \hat{\nabla}_{\mathcal{I}} \mathcal{F}(\theta^{(r)}) - \nabla_{\mathcal{I}} \mathcal{F}(\theta^{(r)}) \right\rangle - 2\eta \mathbb{E}\left\langle \theta^{(r)} - \theta^*, \nabla_{\mathcal{I}} \mathcal{F}(\theta^{(r)}) \right\rangle \\ &\leq \mathbb{E}||\theta^{(r)} - \theta^*||_2^2 + \eta^2 \mathbb{E}||\hat{g}_{\mathcal{I}}^{(r)}(\theta^{(r)})||_2^2 - 2\eta \mathbb{E}\left\langle \theta^{(r)} - \theta^*, \hat{\nabla}_{\mathcal{I}} \mathcal{F}(\theta^{(r)}) - \nabla_{\mathcal{I}} \mathcal{F}(\theta^{(r)}) \right\rangle \\ &\quad - 2\eta \left[\mathcal{F}(\theta^{(r)}) - \mathcal{F}(\theta^*)\right] \\ &= \mathbb{E}||\theta^{(r)} - \theta^*||_2^2 + \eta^2 \mathbb{E}||\hat{g}_{\mathcal{I}}^{(r)}(\theta^{(r)})||_2^2 - 2\eta \mathbb{E}\left\langle \sqrt{\eta}\rho_s^-(\theta^{(r)} - \theta^*), \frac{1}{\sqrt{\eta}\rho_s^-}(\hat{\nabla}_{\mathcal{I}} \mathcal{F}(\theta^{(r)}) - \nabla_{\mathcal{I}} \mathcal{F}(\theta^{(r)})) \right\rangle \\ &\quad - 2\eta \left[\mathcal{F}(\theta^{(r)}) - \mathcal{F}(\theta^*)\right] \\ &\leq (1 + \eta^2 {\rho_s^-}^2)\mathbb{E}||\theta^{(r)} - \theta^*||_2^2 + \eta^2 \mathbb{E}||\hat{g}_{\mathcal{I}}^{(r)}(\theta^{(r)})||_2^2 - 2\eta \left[\mathcal{F}(\theta^{(r)}) - \mathcal{F}(\theta^*)\right] \\ &\quad + \frac{1}{\rho_s^-}^2 \mathbb{E}||(\hat{\nabla}_{\mathcal{I}} \mathcal{F}(\theta^{(r)}) - \nabla_{\mathcal{I}} \mathcal{F}(\theta^{(r)}))||_2^2. \end{aligned} \tag{46}$$

For $\mathbb{E}||(\hat{\nabla}_{\mathcal{I}} \mathcal{F}(\theta^{(r)}) - \nabla_{\mathcal{I}} \mathcal{F}(\theta^{(r)}))||_2^2$, we have:

$$\begin{aligned} \mathbb{E}||(\hat{\nabla}_{\mathcal{I}} \mathcal{F}(\theta^{(r)}) - \nabla_{\mathcal{I}} \mathcal{F}(\theta^{(r)}))||_2^2 &= ||\mathbb{E}_{\boldsymbol{u}}(\hat{\nabla}_{\mathcal{I}} \sum_{i=1}^n f_i(\theta^{(r)}) - \nabla_{\mathcal{I}} \sum_{i=1}^n f_i(\theta^{(r)}))||_2^2 \\ &\leq n \sum_{i=1}^n ||\mathbb{E}_{\boldsymbol{u}}(\hat{\nabla}_{\mathcal{I}} f_i(\theta^{(r)}) - \nabla_{\mathcal{I}} f_i(\theta^{(r)}))||_2^2 \\ &\leq n^2 \varepsilon_\mu \mu^2. \end{aligned} \tag{47}$$

The first inequality follows from N-dimensional mean inequality and the second inequality follows from Assumption 2. By constraining $\mathbb{E}||(\hat{\nabla}_{\mathcal{I}} \mathcal{F}(\theta^{(r)}) - \nabla_{\mathcal{I}} \mathcal{F}(\theta^{(r)}))||_2^2$, we can turn (46) into:

$$\begin{aligned} \mathbb{E}||\theta^{(r)} - \eta \hat{g}_{\mathcal{I}}^{(r)}(\theta^{(r)}) - \theta^*||_2^2 \leq{} & (1 + \eta^2 {\rho_s^-}^2)\mathbb{E}||\theta^{(r)} - \theta^*||_2^2 + \eta^2 \mathbb{E}||\hat{g}_{\mathcal{I}}^{(r)}(\theta^{(r)})||_2^2 - 2\eta \left[\mathcal{F}(\theta^{(r)}) - \mathcal{F}(\theta^*)\right] \\ & + \frac{n^2 \varepsilon_\mu \mu^2}{{\rho_s^-}^2}. \end{aligned} \tag{48}$$

Let $\alpha = 1 + \frac{2\sqrt{k^*}}{\sqrt{k-k^*}}$. Using Lemma 2, we have:

$$\begin{aligned}
\mathbb{E}||\theta^{(r+1)} - \theta^*||_2^2 &\leq (1 + \eta^2 {\rho_s^-}^2)\alpha\mathbb{E}||\theta^{(r)} - \theta^*||_2^2 + \eta^2\alpha\mathbb{E}||\hat{g}_{\mathcal{I}}^{(r)}(\theta^{(r)})||_2^2 - 2\eta\alpha\left[\mathcal{F}(\theta^{(r)}) - \mathcal{F}(\theta^*)\right] \\
&\quad + \alpha\frac{n^2\varepsilon_\mu\mu^2}{{\rho_s^-}^2} \\
&\leq (1 + \eta^2 {\rho_s^-}^2)\alpha\mathbb{E}||\theta^{(r)} - \theta^*||_2^2 + \alpha\frac{n^2\varepsilon_\mu\mu^2}{{\rho_s^-}^2} - 2\eta\alpha\left[\mathcal{F}(\theta^{(r)}) - \mathcal{F}(\theta^*)\right] \\
&\quad + \eta^2\alpha(12\varepsilon_{\mathcal{I}}\mathbb{E}(||\nabla f_{i_r}(\boldsymbol{x}^t) - \nabla f_{i_r}(\theta^*)||^2 + ||\nabla f_{i_r}(\theta^{(r-1)}) - \nabla f_{i_r}(\theta^*)||^2) \\
&\quad + 6((4\varepsilon_{\mathcal{I}}s + 2) + \varepsilon_{\mathcal{I}^c}(d-k))\mathbb{E}||\nabla f_{i_r}(\theta^*)||_\infty^2 + 12\varepsilon_{abs}\mu^2 + 3||\nabla_{\mathcal{I}}\mathcal{F}(\theta^*)||_2^2).
\end{aligned} \tag{49}$$

For $||\nabla f_{i_r}(\boldsymbol{\theta}^t) - \nabla f_{i_r}(\theta^*)||^2$, we can easily get

$$\mathbb{E}||\nabla f_{i_r}(\boldsymbol{\theta}^t) - \nabla f_{i_r}(\theta^*)||^2 = \frac{1}{n}||\nabla f_{i_r}(\boldsymbol{\theta}^t) - \nabla f_{i_r}(\theta^*)||^2 \leq 4\rho_s^+[\mathcal{F}(\theta) - \mathcal{F}(\theta^*)]$$

by RSS condition. As a reasult, we have:

$$\begin{aligned}
\mathbb{E}||\theta^{(r+1)} - \theta^*||_2^2 &\leq (1 + \eta^2 {\rho_s^-}^2)\alpha\mathbb{E}||\theta^{(r)} - \theta^*||_2^2 + \alpha\frac{n^2\varepsilon_\mu\mu^2}{{\rho_s^-}^2} - \alpha(2\eta - 48\varepsilon_{\mathcal{I}}\eta^2\rho_s^+)\left[\mathcal{F}(\theta^{(r)}) - \mathcal{F}(\theta^*)\right] \\
&\quad + 48\eta^2\rho_s^+\alpha\varepsilon_{\mathcal{I}}\mathbb{E}[\mathcal{F}(\theta^{(r-1)}) - \mathcal{F}(\theta^*)] \\
&\quad + 6\eta^2\alpha((4\varepsilon_{\mathcal{I}}s + 2) + \varepsilon_{\mathcal{I}^c}(d-k))\mathbb{E}||\nabla f_{i_r}(\theta^*)||_\infty^2 + 12\varepsilon_{abs}\mu^2 + 3||\nabla_{\mathcal{I}}\mathcal{F}(\theta^*)||_2^2).
\end{aligned} \tag{50}$$

Let $L = 6\eta^2((4\varepsilon_{\mathcal{I}}s + 2) + \varepsilon_{\mathcal{I}^c}(d-k))\mathbb{E}||\nabla f_{i_r}(\theta^*)||_\infty^2 + 12\varepsilon_{abs}\mu^2 + 3||\nabla_{\mathcal{I}}\mathcal{F}(\theta^*)||_2^2) + \frac{n^2\varepsilon_\mu\mu^2}{{\rho_s^-}^2}$, $\beta = (1 + \eta^2 {\rho_s^-}^2)\alpha$, then:

$$\begin{aligned}
\mathbb{E}||\theta^{(r+1)} - \theta^*||_2^2 + \alpha(2\eta - 48\varepsilon_{\mathcal{I}}\eta^2\rho_s^+)\mathbb{E}\left[\mathcal{F}(\theta^{(r)}) - \mathcal{F}(\theta^*)\right] &\leq \beta\mathbb{E}||\theta^{(r)} - \theta^*||_2^2 \\
&\quad + 48\eta^2\rho_s^+\varepsilon_{\mathcal{I}}\alpha\mathbb{E}[\mathcal{F}(\theta^{(r-1)}) - \mathcal{F}(\theta^*)] + \alpha L.
\end{aligned} \tag{51}$$

By summing (51) over $t = 1, \ldots, m-1$, Let $G = \mathbb{E}||\theta^{(m)} - \theta^*||_2^2 + \frac{\beta^{m-1}-1}{\beta-1}(2\eta - 48\varepsilon_{\mathcal{I}}\eta^2\rho_s^+)\alpha\left[\mathcal{F}(\widetilde{\theta}^{(r)}) - \mathcal{F}(\theta^*)\right]$we have:

$$\begin{aligned}
G &\leq \beta^{m-1}\mathbb{E}||\theta^{(r)} - \theta^*||_2^2 + 48\eta^2\rho_s^+\varepsilon_{\mathcal{I}}\alpha\frac{\beta^{m-2}-1}{\beta-1}\mathbb{E}[\mathcal{F}(\widetilde{\theta}^{(r)}) - \mathcal{F}(\theta^*)] \\
&\quad + 48\eta^2\rho_s^+\varepsilon_{\mathcal{I}}\alpha\beta^{m-2}\mathbb{E}[\mathcal{F}(\widetilde{\theta}^{(r-1)}) - \mathcal{F}(\theta^*)] + \frac{\beta^{m-1}-1}{\beta-1}\alpha L.
\end{aligned} \tag{52}$$

That is

$$\begin{aligned}
&\mathbb{E}||\theta^{(m)} - \theta^*||_2^2 + (\frac{\beta^{m-1}-1}{\beta-1}2\eta - (\frac{\beta^{m-1}-1}{\beta-1} + \frac{\beta^{m-2}-1}{\beta-1})48\varepsilon_{\mathcal{I}}\eta^2\rho_s^+)\alpha\left[\mathcal{F}(\widetilde{\theta}^{(r)}) - \mathcal{F}(\theta^*)\right] \\
&\leq \beta^{m-1}\mathbb{E}||\theta^{(r)} - \theta^*||_2^2 + 48\eta^2\rho_s^+\varepsilon_{\mathcal{I}}\alpha\beta^{m-2}\mathbb{E}[\mathcal{F}(\widetilde{\theta}^{(r-1)}) - \mathcal{F}(\theta^*)] + \frac{\beta^{m-1}-1}{\beta-1}\alpha L.
\end{aligned}$$

Through RSC condition and the definition of $\widetilde{I}$, it further follows from (52) that:

$$\begin{aligned}
&\mathbb{E}||\theta^{(m)} - \theta^*||_2^2 + (\frac{\beta^{m-1}-1}{\beta-1}2\eta - (\frac{\beta^{m-1}-1}{\beta-1} + \frac{\beta^{m-2}-1}{\beta-1})48\varepsilon_{\mathcal{I}}\eta^2\rho_s^+)\alpha\left[\mathcal{F}(\widetilde{\theta}^{(r)}) - \mathcal{F}(\theta^*)\right] \\
&\leq \left(\frac{2\beta^{m-1}}{\rho_s^-} + 48\eta^2\rho_s^+\varepsilon_{\mathcal{I}}\alpha\beta^{m-2}\right)\mathbb{E}[\mathcal{F}(\widetilde{\theta}^{(r-1)}) - \mathcal{F}(\theta^*)] + \frac{2\beta^{m-1}}{\rho_s^-}\mathbb{E}\left\langle\nabla\mathcal{F}(\theta^*), \widetilde{\theta}^* - \theta^{(r-1)}\right\rangle + \frac{\beta^{m-1}-1}{\beta-1}\alpha L \\
&\leq \left(\frac{2\beta^{m-1}}{\rho_s^-} + 48\eta^2\rho_s^+\varepsilon_{\mathcal{I}}\alpha\beta^{m-2}\right)\mathbb{E}[\mathcal{F}(\widetilde{\theta}^{(r-1)}) - \mathcal{F}(\theta^*)] + \frac{2\beta^{m-1}}{\rho_s^-}\sqrt{s}||\nabla\mathcal{F}(\theta^*)||_\infty\mathbb{E}||\widetilde{\theta}^{(r-1)} - \theta^*||_2 + \frac{\beta^{m-1}-1}{\beta-1}\alpha L.
\end{aligned} \tag{53}$$

Here $\varepsilon_\mu = \rho_s^{+2} sd$, $\varepsilon_{\mathcal{I}} = \frac{2d}{q(s_2+2)}(\frac{(s-1)(s_2-1)}{d-1} + 3) + 2$, $\varepsilon_{\mathcal{I}^c} = \frac{2d}{q(s_2+2)}\left(\frac{s(s_2-1)}{-1}\right)$, $\varepsilon_{abs} = \frac{2d\rho_s^{+2} ss_2}{q}\left(\frac{(s-1)(s_2-1)}{d-1} + 1\right) + \rho_s^{+2} sd$ □

6.3 PROOF OF COROLLARY 2

If the algorithm converges, we have:

$$\frac{2\beta^{m-1} - 1}{\alpha\rho_s^-} + 48\varepsilon_{\mathcal{I}}\eta^2\rho_s^+\beta^{m-2} \leq \frac{\beta^{m-1} - 1}{\beta - 1}2\eta - \beta^{m-2}48\varepsilon_{\mathcal{I}\eta^2\rho_s^+}.$$

which is:

$$\frac{\beta - 1}{\alpha\rho_s^-\eta} + \frac{48\varepsilon_{\mathcal{I}}\eta^2\rho_s^+(\beta - 1)}{\beta} \leq 1 - \frac{1}{\beta^{m-1}}. \tag{54}$$

If there is a $m \in \mathbb{N}^+$ that holds (54), then there must be:

$$\frac{\beta - 1}{\alpha\rho_s^-\eta} + \frac{48\varepsilon_{\mathcal{I}}\eta^2\rho_s^+(\beta - 1)}{\beta} \leq 1.$$

Recall $\beta = \alpha(1 + \eta^2\rho_s^-)^2$, we have:

$$(48\varepsilon_{\mathcal{I}}\alpha\rho_s^+ + \alpha\rho_s^-)\eta^2 - \alpha\eta + (\alpha - 1) \leq 0.$$

This is a quadratic inequality about η, and a conclusion can be obtained through discriminant.

7 EXTRA EXPERIMENTS

7.1 RIDGE REGRESSION

In this section we provide additional curves for the first problem described in section 5, which we recall here for sake of completeness. We consider a ridge regression problem, where each function f_i is defined as follows:

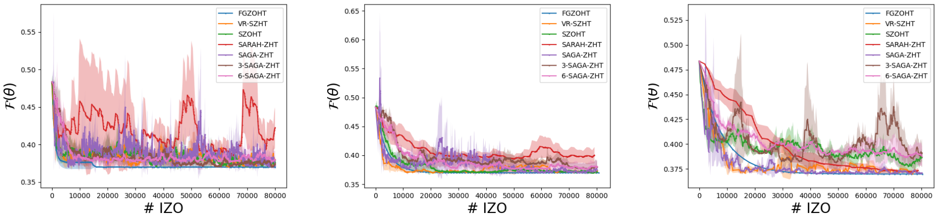
$$f_i(\theta) = (x_i^\top\theta - y_i)^2 + \frac{\lambda}{2}\|\theta\|_2^2,$$

where λ is some regularization parameter.

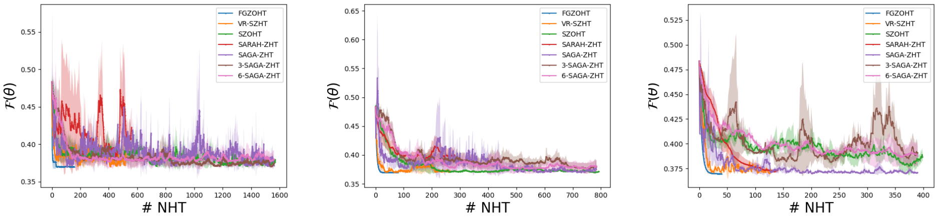
7.2 SYNTHETIC EXPERIMENT

Experimental Setting First, as in the main paper, we consider a synthetic dataset: we generate each x_i randomly from a unit norm ball in $\mathbb{R}^d$, and a true random model θ^* from a normal distribution $\mathcal{N}(0, I_{d\times d})$. Each y_i is defined as $y_i = x_i^\top\theta^*$. We set the constants of the problem as such: $n = 10, d = 5, \lambda = 0.5$. Before training, we pre-process each column by subtracting its mean and dividing it by its empirical standard deviation. We run each algorithm with $\mu = 10^{-4}, s_2 = d = 5$, and for the variance reduced algorithms, we choose $m = 10$. For all algorithms, the learning rate η is found through grid-search in $\{0.005, 0.01, 0.05, 0.1, 0.5\}$, and we keep the η giving the lowest function value (averaged over 3 runs) at the end of training. We stop each algorithm once its number of IZO reaches 80,000. We plot in Figures 4, 5, and 6 the mean and standard deviation of the curves for a value of $k = 2$, 3, and 4 respectively.

Results and Discussion We can observe the several phenomena on the Figures 4, 5, and 6. First, we can observe that for larger k, the algorithms converge to lower function values (which is natural because optimization is then over a larger set), but also, the algorithms are more stable (for example, SARAH-SZHT converges more easily with $k = 4$ than with $k = 2$), which is due to the hard-thresholding operator being more non-expansive. Then, although a larger number of random directions q may slow down the query complexity (IZO), we observe that it can also stabilize some algorithms that would otherwise be less unstable, such as SARAH-SZHT (which converges better for $q = 200$ than for $q = 50$).

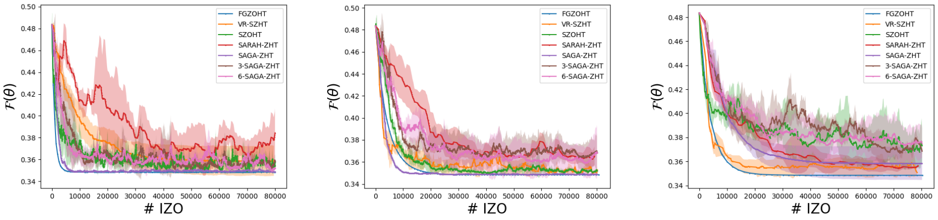


(a) q=50 (b) q=100 (c) q=200

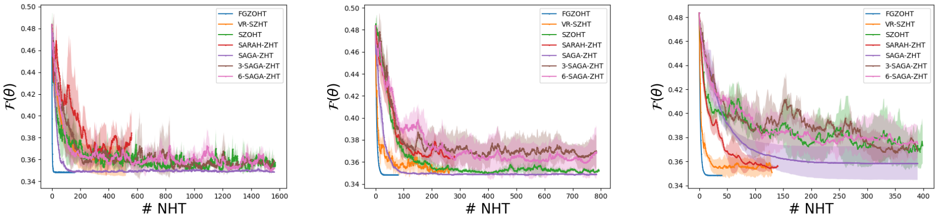


(d) q=50 (e) q=100 (f) q=200

Figure 4: #IZO (up) and #NHT (down) on the ridge regression task, synthetic example (k=2).

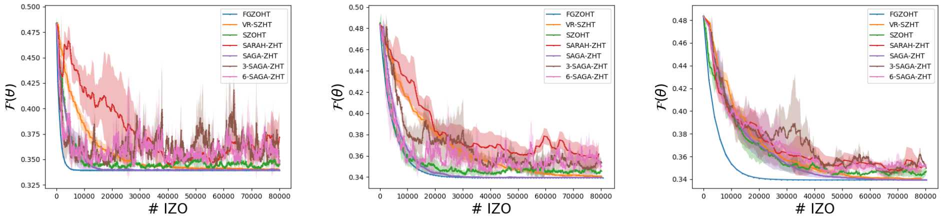


(a) q=50 (b) q=100 (c) q=200

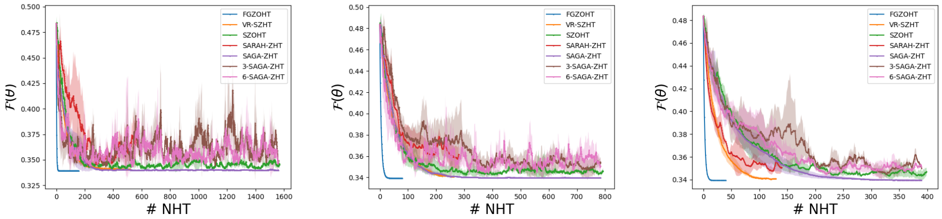


(d) q=50 (e) q=100 (f) q=200

Figure 5: #IZO (up) and #NHT (down) on the ridge regression task, synthetic example (k=3).



(a) q=50 (b) q=100 (c) q=200



(d) q=50 (e) q=100 (f) q=200

Figure 6: #IZO (up) and #NHT (down) on the ridge regression task, synthetic example (k=4).

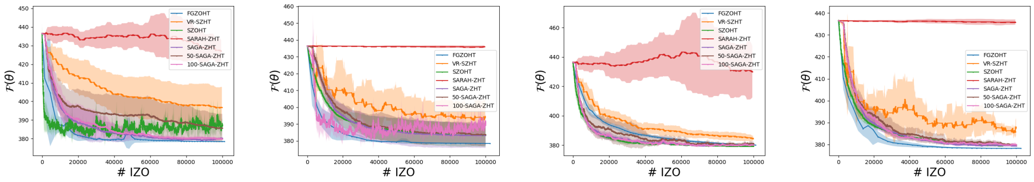
7.3 REAL-LIFE DATASETS

Experimental Setting Second, we now compare the above algorithms on the following open source real-life datasets (obtained from OpenML Vanschoren et al. (2014)), of which a summary is presented in Table 4. We take $\lambda = 0.5$. Similarly as above, before training, we pre-process each column by subtracting its mean and dividing it by its empirical standard deviation. We run each algorithm with $\mu = 10^{-4}$, $s_2 = d$ (where d depends on the dataset), and for the variance reduced algorithms with an inner and outer loop (VR-SZHT and SARAH-SZHT), we choose $m = \lfloor \frac{1}{n}2 \rfloor$. For all algorithms, the learning rate η is found through grid-search in $\{10^{-i},\ i \in \{1, ..., 7\}\}$, and we choose the one giving the lowest function value (averaged over 5 runs) at the end of training. We stop each algorithm once its number of IZO reaches 100,000. We plot the optimization curves (averaged over the 5 runs) for several values of q and k, to study their impact on the convergence.

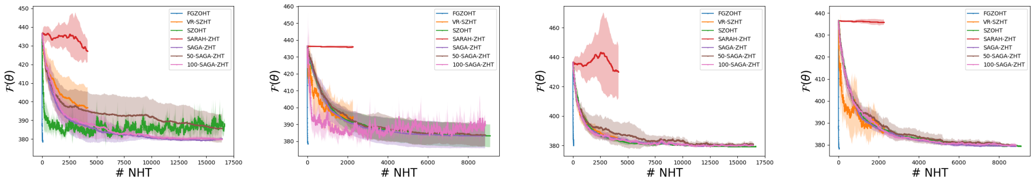
Table 4: Datasets used in the comparison. *Reference:* Dua and Graff (2017), *Source:*Vanschoren et al. (2014), downloaded with `scikit-learn` Pedregosa et al. (2011).

DATASET	d	n
BODYFAT$^{(2)}$	14	252
AUTO-PRICE$^{(3)}$	15	159

Results and Discussion We present our results in Figures 7 and 8. Those results are consistent with preliminary results on the synthetic dataset from Section 7.1, namely, that overall, although taking a larger q may worsen the IZO complexity, it can help some algorithms to converge more smoothly, by reducing the error of the zeroth-order estimator. Additionally, we can observe that taking a larger k often helps to achieve smoother convergence. Finally, consistently across experiments, we observe that SARAH-SZHT has difficulties converging: this seems to indicate that SARAH-SZHT may be highly impacted by the errors introduced by the zeroth-order estimator. SARAH-SZHT could potentially be improved by a more careful choice of the number of inner iterations, and/or by running SARAH+, which is an adaptive version of SARAH Nguyen et al. (2017b), which we leave for future work.

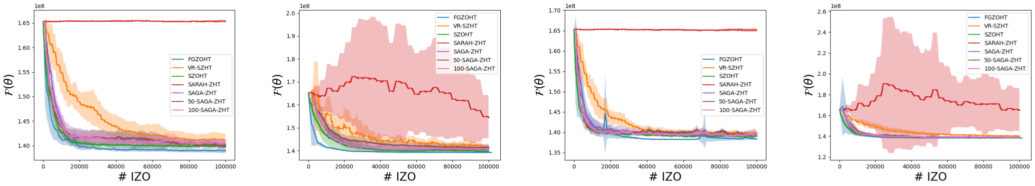


(a) q=5, k=5 (b) q=10, k=5 (c) q=5, k=10 (d) q=10, k=10

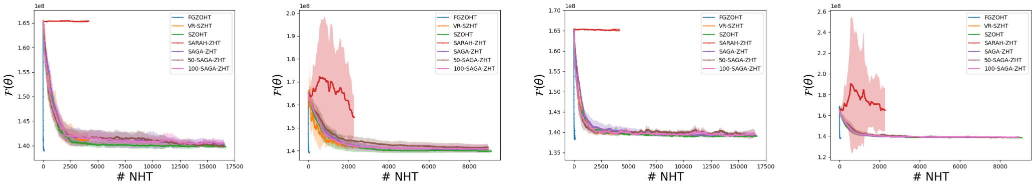


(e) q=5, k=5 (f) q=10, k=5 (g) q=5, k=10 (h) q=10, k=10

Figure 7: #IZO (up) and #NHT (down) on the ridge regression task, `bodyfat` dataset.



(a) q=5, k=5 (b) q=10, k=5 (c) q=5, k=10 (d) q=10, k=10



(e) q=5, k=5 (f) q=10, k=5 (g) q=5, k=10 (h) q=10, k=10

Figure 8: #IZO (up) and #NHT (down) on the ridge regression task, `autoprice` dataset.

7.4 ADDITIONAL RESULTS FOR UNIVERSAL ADVERSARIAL ATTACKS

In this section , we provide additional results for the universal adversarial attacks setting from our Experiments Section, for the 3 additional classes: 'ship', 'bird', and 'dog'. As we can observe in Figures 10 , 9, and 11 below respectively, in most of such cases, there is a variance-reduced algorithm which can achieve better performance than the vanilla zeroth-order hard-thresholding algorithms, (for instance, SARAH-ZHT in Figure 9, and SAGA-ZHT in Figure 10) which demonstrates the applicability of such algorithms. Correspondingly, this can also be verified by observing the misclassification success in Table 6, 5 and 7: even if a smaller value for the cost does not necessarily imply a strictly higher attack success rate, still, overall, more successful universal attacks also have a higher success rate of attack.

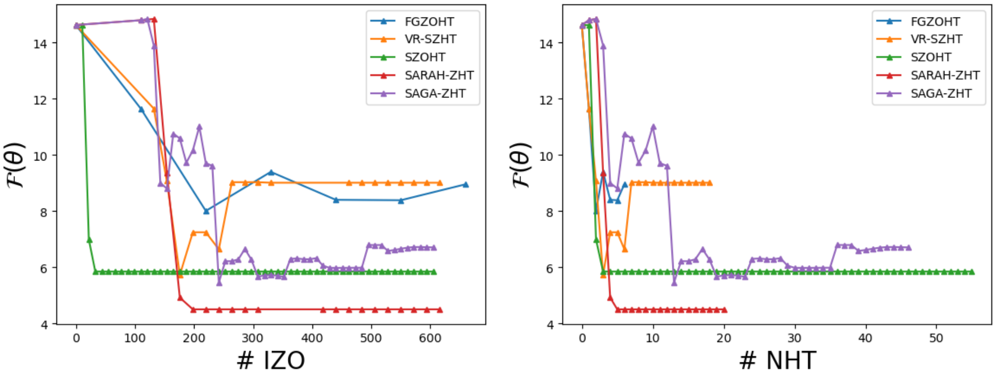
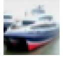
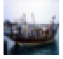
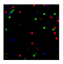
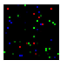
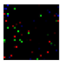
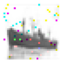
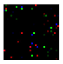
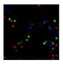


Figure 9: #IZO and #NHT on the few pixels adversarial attacks task (CIFAR-10), for the original class 'ship'.

Table 5: Comparison of universal adversarial attacks on $n = 10$ images from the CIFAR-10 test-set, from the 'ship' class. For each algorithm, the leftmost image is the sparse adversarial perturbation applied to each image in the row. ('auto' stands for 'automobile', and 'plane' for 'airplane')

Image ID	1	15	18	51	54	55	72	73	79	80
Original										
FGZOHT 	 ship	 **frog**	 ship	 ship	 ship	 ship	 ship	 ship	 ship	 ship
SZOHT 	 ship	 **frog**	 ship	 **deer**	 ship	 **deer**	 ship	 ship	 ship	 **plane**
VR-SZHT 	 ship	 **frog**	 **truck**	 **plane**	 ship	 ship	 ship	 ship	 ship	 **auto**
SAGA-SZHT 	 ship	 **frog**	 **truck**	 ship	 **plane**	 **deer**	 ship	 ship	 ship	 ship
SARAH-SZHT 	 ship	 **frog**	 **auto**	 **auto**	 ship	 ship	 ship	 ship	 ship	 ship

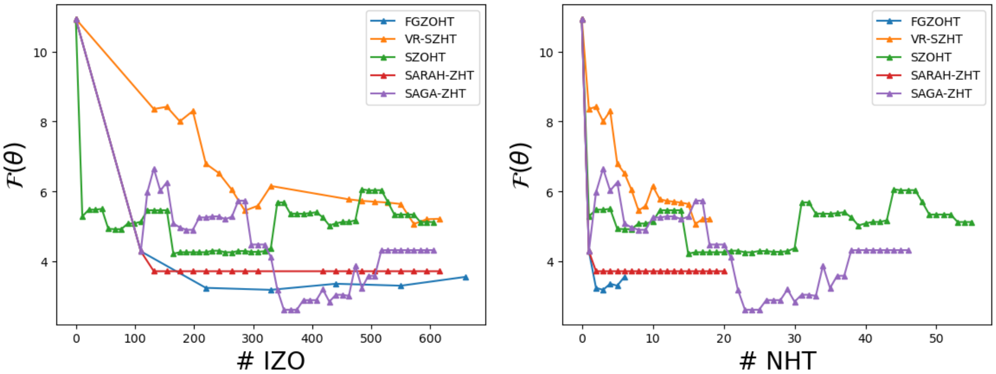
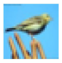
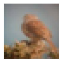
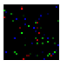
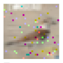
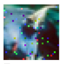
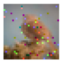
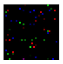
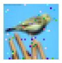
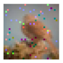
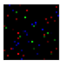
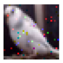
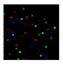
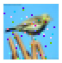
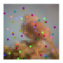
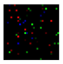
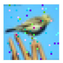
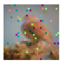
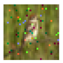


Figure 10: #IZO and #NHT on the few pixels adversarial attacks task (CIFAR-10), for the original class 'bird'.

Table 6: Comparison of universal adversarial attacks on $n = 10$ images from the CIFAR-10 test-set, from the 'bird' class. For each algorithm, the leftmost image is the sparse adversarial perturbation applied to each image in the row.

Image ID	65	67	70	75	86	113	123	129	138	149
Original										
FGZOHT 	 **frog**	 **bird**	 **deer**	 **dog**	 **deer**	 bird	 bird	 **deer**	 **frog**	 bird
SZOHT 	 **frog**	 bird	 bird	 bird	 bird	 bird	 bird	 **ship**	 bird	 bird
VR-SZHT 	 **frog**	 bird	 bird	 bird	 bird	 **frog**	 bird	 **ship**	 bird	 bird
SAGA-SZHT 	 **frog**	 bird	 **deer**	 **dog**	 bird	 bird	 **cat**	 **deer**	 **frog**	 bird
SARAH-SZHT 	 **frog**	 bird	 **frog**	 bird	 bird	 **frog**	 bird	 bird	 **frog**	 bird

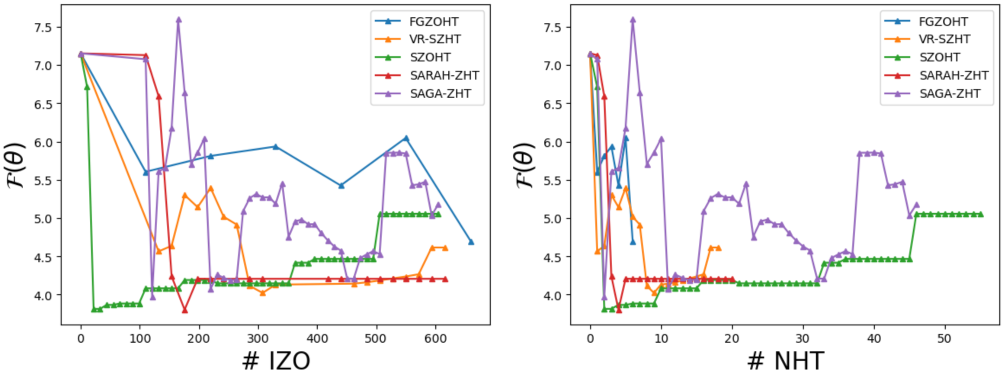
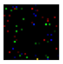
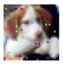
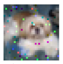
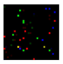
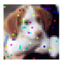
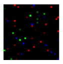
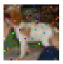
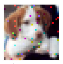
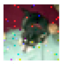
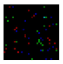
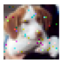
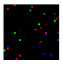
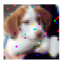
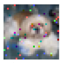
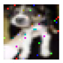


Figure 11: #IZO and #NHT on the few pixels adversarial attacks task (CIFAR-10), for the original class ’dog’.

Table 7: Comparison of universal adversarial attacks on $n = 10$ images from the CIFAR-10 test-set, from the ’dog’ class. For each algorithm, the leftmost image is the sparse adversarial perturbation applied to each image in the row.

Image ID	12	16	31	33	39	42	101	128	141	148
Original										
FGZOHT 	 **bird**	 dog	 **deer**	 dog	 dog	 **cat**	 **cat**	 **cat**	 dog	 dog
SZOHT 	 **frog**	 dog	 **bird**	 dog	 dog	 **cat**	 dog	 **cat**	 dog	 dog
VR-SZHT 	 dog	 dog	 **bird**	 dog	 dog	 **horse**	 dog	 dog	 dog	 dog
SAGA-SZHT 	 dog	 dog	 dog	 dog	 dog	 **horse**	 dog	 **cat**	 dog	 dog
SARAH-SZHT 	 **deer**	 dog	 dog	 **frog**	 dog	 **cat**	 **frog**	 dog	 dog	 dog