On the Inherent Privacy of Zeroth-Order Projected Gradient Descent

Devansh Gupta **Meisam Razaviyayn** **Vatsal Sharan**
University of Southern California, Los Angeles

Abstract

Differentially private zeroth-order optimization methods have recently gained popularity in private fine tuning of machine learning models due to their reduced memory requirements. Current approaches for privatizing zeroth-order methods rely on adding Gaussian noise to the estimated zeroth-order gradients. However, since the search direction in the zeroth-order methods is inherently random, researchers including Tang et al. (2024) and Zhang et al. (2024a) have raised an important question: is the inherent noise in zeroth-order estimators sufficient to ensure the overall differential privacy of the algorithm? This work settles this question for a class of oracle-based optimization algorithms where the oracle returns zeroth-order gradient estimates. In particular, we show that for a fixed initialization, there exist strongly convex objective functions such that running (Projected) Zeroth-Order Gradient Descent (ZO-GD) is not differentially private. Furthermore, we show that even with random initialization and without revealing (initial and) intermediate iterates, the privacy loss in ZO-GD can grow superlinearly with the number of iterations when minimizing convex objective functions.

1 Introduction

The fine-tuning of pretrained large language models (LLMs) has demonstrated state-of-the-art performance across a range of downstream applications. However, two main challenges hinder the wide adoption of these models: the substantial memory requirements of gradient-based optimizers used for fine-tuning and the critical need to protect the privacy of domain-specific fine-tuning data. As fine-tuning LLMs grows increasingly memory-intensive, a range of strategies has emerged to address this issue. In particular, zeroth-order (ZO) optimization methods recently have gained traction due to their memory efficiency, as they do not require explicit gradient computations. Instead, the zeroth-order gradients can be computed using forward step only, significantly reducing memory use compared to gradient computation.

In a pioneering approach, Malladi et al. (2023) introduced a memory-efficient technique for fine-tuning LLMs using two-point Simultaneous Perturbation Stochastic Approximation (SPSA) estimators (Spall, 1992), enabling large model fine-tuning on memory-limited devices. Since then, zeroth-order methods have gained popularity in dealing with large machine learning models due to their memory efficiency and favorable upper bounds on gap from optimality under certain conditions on the Hessian of the objective function (Zhang et al., 2024c,a; Guo et al., 2024).

Another major concern in training LLMs is *privacy.* As large parameterized models are increasingly used in sensitive data applications, these models must protect sensitive information, especially given privacy regulations like the E.U. General Data Protection Regulation and the California Consumer Privacy Act. This requirement led to significant research into differential privacy (DP), a robust framework ensuring that machine learning models do not compromise the privacy of their contributors (Dwork et al., 2010a). As a result, there has been a growing focus on developing methods that fine-tune LLMs while adhering to differential privacy standards, leading to numerous theoretical advance-

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume TBD. Copyright 2024 by the author(s).

ments in private optimization (Dwork et al., 2010b; Chaudhuri and Hsu, 2011; Cohen et al., 2023; Gonen et al., 2019; Block et al., 2024; Chaudhuri et al., 2011; Bassily et al., 2014; Lowy et al., 2024; Arora et al., 2023; Feldman et al., 2020; Bassily et al., 2019; Gopi et al., 2022; Altschuler and Talwar, 2022; Carmon et al., 2023; Gopi et al., 2023) and practical applications in the industry (Abadi et al., 2016; Yousefpour et al., 2022; Erlingsson et al., 2014; Ding et al., 2017; Rogers et al., 2021). Nevertheless, most existing work in this area has focused on first-order optimization/training algorithms, highlighting the need to explore differentially private zeroth-order optimization techniques to combine memory efficiency with privacy protection.

Motivated by the memory efficiency and empirical success of ZO methods in fine-tuning LLMs, Tang et al. (2024) and Zhang et al. (2024a) introduced differentially private and memory-efficient algorithms based on ZO optimization techniques. Both noted that the inherent noise in zeroth-order gradient estimates might contribute to privacy protection. As a result, they highlighted that the inherent noise in the ZO estimators was not considered in their privacy analyses and posed the following open question:

Open Problem: *Is additive (Gaussian) noise necessary for ensuring privacy for ZO Stochastic Gradient Descent (SGD)?*

In this work, we address this question through the following key **contributions**:

1. We propose a class of oracles that generalizes multiple point zeroth-order estimators and show that any estimator in our class is not differentially private.

2. We show that for a generalized setting which subsumes the algorithms proposed in Tang et al. (2024) and Zhang et al. (2024a), the ZO method is not private on its own and the presence of additive noise is necessary to preserve privacy of the algorithm. This answers the open problem posed by Tang et al. (2024) and Zhang et al. (2024a).

3. We further show that, even with random initialization and without disclosing intermediate iterates, optimizing specific types of objectives using ZO-GD results in a superlinear increase in privacy loss as the number of iterations grows. This finding suggests that, despite random initialization and privacy amplification through iterations, the inherent randomness of zeroth-order methods is insufficient to guarantee meaningful privacy in practice.

1.1 Related Work and Existing Results

ZO Optimization. The idea of minimizing functions based on function evaluations has origins from control theory (Spall, 1992). Such an approach is useful when obtaining gradients is either impossible or too costly, making ZO methods favorable for minimizing non-smooth or even discontinuous functions. For example, Wibisono et al. (2012); Nesterov and Spokoiny (2015) gave upper bounds on the gap between the optimal solutions and returned solutions after a finite number of iterations in the case of convex functions. They relied on gradient oracles based on the finite difference method and show that these oracles estimate the gradient of the smoothed version of the function, as we discuss in Section 2. To understand the fundamental limit on the performance of such algorithms, Shamir (2013) showed the existance of convex functions with a lower bound on the optimality gap for any ZO algorithm that queries a single point per iteration. Duchi et al. (2015) extended this result to algorithms that query multiple points per iteration. Recently Malladi et al. (2023); Yue et al. (2023) provided nearly-dimension-independent upper bounds on the optimality gap for ZO algorithms, primarily depending on a measure termed as the effective dimension of the problem which relates to some notion of a local rank of the hessian of the function. However, in the worst case, this effective dimension is equal to the actual dimension of the problem, restoring the lower bounds obtained for their specific cases in Shamir (2013) and Duchi et al. (2015). However, from an application perspective, the primary reason ZO methods have gained traction is that many over-parameterized models, such as pretrained Large Language Models (LLMs), are shown to have a low effective dimension (Hu et al., 2022). Malladi et al. (2023); Zhang et al. (2024a) leverage this insight to justify the effectiveness of ZO algorithms in fine-tuning LLMs.

Private Convex Optimization. On the other hand, optimization is the foundation of modern large-scale machine learning, making it essential to understand private optimization to fully grasp privacy in machine learning. Along this step, Bassily et al. (2014) gave the first tight upper and lower bounds for minimizing convex and strongly convex functions. Their idea for minimizing smooth functions was making Stochastic Gradient Descent (SGD) differentially private by updating their parameters with a noisy version of the gradient estimate. This algorithm achieves optimal excess risk gap (upto logarithmic terms) in private empirical risk minimization (ERM) (Bassily et al., 2014) and stochastic convex optimization (SCO) for smooth objectives (Bassily et al., 2019). Moreover, a small (yet effective) modification to this algorithm achieves optimal rates in just one pass over the entire data (Feldman et al., 2020). There have been further works which have also solved the problem of differentially private convex optimization under non-smooth conditions (Bassily et al., 2019; Kulkarni et al., 2021; Feldman et al., 2020) and general norms (Asi et al., 2021; Gopi et al., 2023). Other than private SGD, Bassily et al. (2014) proposed an exponential-mechanism-based algorithm which achieved optimal risk bounds up to constant factors for ε-DP minimization. Gopi et al. (2022) further extended the exponential mechanism to (ε, δ) DP SCO and obtained upper bounds on the excess population loss. Recently, Carmon et al. (2023) proposed a new estimator specifically made for accelerated private optimization giving improvements on the query complexity of private SCO.

Private ZO optimization is an area which is relatively new and there have been recent works which give guarantees on private ZO optimization for smooth convex problems (Malladi et al., 2023), smooth non-convex problems (Zhang et al., 2024a; Tang et al., 2024) and non-smooth non-convex problems (Zhang et al., 2024b).

On the other hand, the work on inherent privacy in zeroth-order optimization is relatively sparse. Tang et al. (2024) provided an *empirical* evaluation of the privacy of Projected ZO-SGD using a privacy accounting technique based on membership-inference attacks on ML models (Shokri et al., 2017). They empirically showed that ZO SGD without any additive noise gave almost the same privacy as regular SGD with additive noise corresponding to $\varepsilon = 10$ and $\delta = 10^{-5}$, giving some *experimental evidence* to conjecture the possible presence of the inherent privacy in ZO optimization. However, we prove that there exist worst case objectives and datapoints where popular ZO optimization algorithms (including the algorithm studied in Tang et al. (2024)) are not private.

2 Problem Setting and Preliminaries

Notation. We use $\|\cdot\|_2$ for the Euclidean L_2 norm. We denote $\mathbb{P}[E]$ as the probability of any event E. $\mathbb{E}[X]$ denotes the expectation of any random variable X. For any distribution $\mathcal{D}$, $supp(\mathcal{D})$ represents the support of the distribution, $\mathcal{N}(0, \sigma^2 I_d)$ is the isotropic normal distribution with mean 0 and covariance $\sigma^2 I_d$. For any set S, $Unif(S)$ is the uniform distribution over the set S. $D\mathbb{B}^d = \{x \in \mathbb{R}^d : \|x\|_2 \leq D\}$. $\mathbf{0}_d$, $\mathbf{1}_d$, and $\mathbf{e}_i$ represent d dimensional all zero, all one, and the i^{th} standard basis vectors, respectively. For a vector $v \in \mathbb{R}^d$, $\{v\}_i$ represents the i^{th} coordinate of the vector; $[n]$ denotes the set of natural numbers less than or equal to n. $\Pi_D(x) = \arg\min_{u \in D} \|x - u\|_2$ is used to denote the projection of $x \in \mathbb{R}^d$ onto the set D. For any function f, $dom(f)$ represents the domain of the function f.

Problem Setting. We consider the problem of minimizing the empirical loss function with respect to the dataset $\mathcal{D} = \{d_1, d_2, ..., d_n\}$ given a closed convex set $\mathcal{C} \subseteq \mathbb{R}^d$

$$\min_{w \in \mathcal{C}} \mathcal{L}(w; \mathcal{D})$$

where $\mathcal{L}(w; \cdot) : \mathcal{C} \to \mathbb{R}$ is convex and L-Lipschitz. There can be additional assumptions on $\mathcal{L}(w; \cdot)$ such as Δ strong convexity. Recall that a function g is called L-Lipschitz if $\|g(x) - g(y)\|_2 \leq L \|x - y\|_2$ for all $x, y \in \mathcal{C}$ and it is called μ-Strongly Convex when for all $x, y \in \mathcal{C}$, $f(x) \geq f(y) + \langle z, y - x \rangle + \frac{\mu}{2} \|y - x\|_2^2$ where z is any subgradient of f. Next, we define the notion of Differential Privacy.

Definition 2.1 (Differential Privacy (Dwork et al., 2006))**.** *Two datasets of the same size are neighbouring if $|\mathcal{D} \Delta \mathcal{D}'| = 2$ where Δ represents the symmetric set difference. Let $\varepsilon \geq 0$, $\delta \in [0, 1)$. A randomized algorithm $\mathcal{A}$ is (ε, δ)-differentially private (DP) if for all pairs of neighbouring data sets $\mathcal{D}, \mathcal{D}'$, we have*

$$\mathbb{P}(\mathcal{A}(\mathcal{D}) \in O) \leq e^{\varepsilon} \mathbb{P}(\mathcal{A}(\mathcal{D}') \in O) + \delta, \qquad (1)$$

for any measurable set O.

ZO Stochastic Gradient Descent. The algorithmic framework that we will be operating under is the *projected stochastic ZO descent* as given in Algorithm 1. Algorithm 1 requires access to *zeroth order oracles* to query the update direction at each step. We define such oracles below.

Definition 2.2. *A zeroth order oracle $\mathcal{O}$ is an oracle which takes a function $f : \mathbb{R}^d \to \mathbb{R}$, a single* *point $w \in \mathbb{R}^d$ and returns a probability measure over a subset of $\mathbb{R}^d$ using only function evaluations on different points depending on w.*

Notably, ZO oracle may require multiple function evaluations, which may not necessarily depend on the point w. In the design of ZO algorithms, these oracles are typically implemented using well-established ZO estimators. Below, we provide examples of such estimators.

Example 2.3. *Here we list three popular ZO estimators: Simultaneous Perturbation Stochastic Approximation (SPSA) introduced by Spall (1992), Finite Difference (FD) introduced by Nesterov and Spokoiny (2015), and Single Point (SP) estimator introduced by Flaxman et al. (2005).*

1. *$SPSA_\xi(f, w)$: Sample $Z \sim \mathcal{N}(0, I_d)$ and return $\widehat{\nabla}_1 f_\xi(w) := \frac{f(w+\xi Z)-f(w-\xi Z)}{2\xi} Z$.*

2. *$FD_\xi(f, w)$: Sample $Z \sim \mathcal{N}(0, I_d)$ and return $\widehat{\nabla}_2 f_\xi(w) := \frac{f(w+\xi Z)-f(w)}{\xi} Z$.*

3. *$SP_\xi(f, w)$: Sample $Z \sim \mathcal{N}(0, I_d)$ and return $\widehat{\nabla}_3 f_\xi(w) := \frac{d}{\xi} f(w + \xi Z) Z$*

The above mentioned estimators are widely used in various zeroth order optimization and control algorithms. They are unbiased estimators of the gradient of a smoothed version of the function f, defined as $f_\xi(x) = \mathbb{E}_{Z \sim \mathcal{N}(0, I_d)}[f(x + \xi Z)]$. In other words, $\mathbb{E}\left[\widehat{\nabla} f_\xi(w)\right] = \nabla f_\xi(x)$ (Nesterov and Spokoiny, 2015; Nemirovsky and Yudin, 1983).

Since the aforementioned estimators are randomized, several works, such as Zhang et al. (2024b), Malladi et al. (2023), and Duchi et al. (2015), have focused on strategies to reduce the variance of the updates obtained through these estimators, by taking the mean of different samples. Building on these randomized estimators, multi-point estimators can be defined aggregate information across multiple points. We provide an example of such aggregation below.

Example 2.4. *For any randomized estimator $\mathcal{E}$, the mean extension of the estimator $M_m^{\mathcal{E}}(f, w)$ works as follows: Sample i.i.d. $U_1, U_2, ..., U_m \sim \mathcal{E}(f, w)$ and return $\frac{1}{m}\sum_{i=1}^m U_i$.*

Example 2.5. *Using the construction mentioned in Example 2.4, one can further define M_m^{SPSA} (Malladi et al., 2023; Zhang et al., 2024b), M_m^{FD} (Duchi et al., 2015), and M_m^{SP}.*

Algorithm 1 Projected Stochastic ZO Descent

```
Given number of steps $T$, initialization distribution $\mathcal{R}_{init}$, ZO oracle $\mathcal{O}$, and constraint set $\mathcal{D}$
Sample $w_0 \sim \mathcal{R}_{init}$
for $t \leftarrow 1, ..., T$ do
    Draw $\widehat{\nabla}\mathcal{L}(w_{t-1}; \mathcal{X})$ from the distribution $\mathcal{O}(\mathcal{L}(\cdot; \mathcal{X}), w_{t-1})$
    $w_t \leftarrow \Pi_{\mathcal{D}}\left(w_{t-1} - \eta \widehat{\nabla}\mathcal{L}(w_{t-1}; \mathcal{X})\right)$
end
```

1. *$M_m^{SPSA_\xi}(f, w)$: Sample i.i.d. $Z_1, ..., Z_m \sim \mathcal{N}(0, I_d)$ and return $\widehat{\nabla}_1^m f_\xi(w) := \frac{1}{m}\sum_{i=1}^m \frac{f(w+\xi Z_i)-f(w-\xi Z_i)}{2\xi} Z_i$.*

2. *$M_m^{FD_\xi}(f, w)$: Sample i.i.d. $Z_1, ..., Z_m \sim \mathcal{N}(0, I_d)$ and return $\widehat{\nabla}_2^m f_\xi(w) := \frac{1}{m}\sum_{i=1}^m \frac{f(w+\xi Z_i)-f(w)}{\xi} Z_i$.*

3. *$M_m^{SP_\xi}(f, w)$: Sample i.i.d. $Z_1, Z_2, ..., Z_m \sim \mathcal{N}(0, I_d)$ and return $\widehat{\nabla}_3^m f_\xi(w) := \frac{1}{m}\sum_{i=1}^m \frac{d}{\xi} f(w + \xi Z_i) Z_i$*

It is important to note that most work in ZO convex optimization relies on these estimators. Consequently, it becomes essential to define a subclass of randomized zeroth order oracles that generalizes this family of estimators and also *provides a clearer understanding of their privacy implications.* Understanding the properties of such oracles is crucial for analyzing the privacy aspects of ZO methods that reveal their intermediate states.

3 Privacy of Randomized Zeroth-Order Oracles

In this section, we define and discuss a subclass of randomized zeroth-order oracles as defined in Definition 2.2, and discuss privacy properties of this subclass. Let us start by defining zero-preserving noisy oracles:

Definition 3.1. *An oracle $\mathcal{O}$ is a zero-preserving noisy oracle if for any $f : \mathbb{R}^d \to \mathbb{R}$, it satisfies the following properties,*

1. *$\mathcal{O}(f, w)$ returns $\mathbf{0}_d$ when $f(w) = 0$ for all $w \in \mathbb{R}^d$ i.e. if $U \sim \mathcal{O}(0, w)$ then $\mathbb{P}[U = \mathbf{0}_d] = 1$.*

2. *For $a > 0$, if $f(u) = \frac{a}{2}\|u\|_2^2$, then $\mathcal{O}(f, w)$ is a continuous probability measure for $w \neq \mathbf{0}_d$ i.e. if $U \sim \mathcal{O}(f, w)$ and $w \neq \mathbf{0}_d$ then for all $c \in \mathbb{R}^d, \mathbb{P}[U = c] = 0$.*

It is important to note that, a zero-preserving noisy oracle does not necessarily imply a zeroth order oracle. For instance, consider the estimator: Sample $Z \sim \mathcal{N}(0, I_d)$ and return $\widehat{\nabla} f(u) = (Z^T \nabla f(u))Z$. We see that this estimator satisfies the property of being a zero-preserving noisy oracle, even though it still uses first order information.

Nonetheless, zero-preserving noisy oracles capture many popular ZO estimators used in the literature. In the following lemma, we show that several of the estimators discussed in Example 2.3 satisfy Definition 3.1.

Lemma 3.2. *SPSA, FD, SP, M_m^{SPSA}, M_m^{FD} and M_m^{SP} are zero-preserving noisy oracles.*

We defer the proof to appendix B.1. We also give example of another estimator proposed in Zhang et al. (2024a) and show that it is also a zero-preserving noisy oracle.

The following lemma shows that any zero-preserving noisy oracle is not differentially private.

Theorem 3.3. *If $\mathcal{O}$ is a zero-preserving noisy oracle (as defined in Definition 3.1), then there exists an L-Lipschitz strongly convex loss function over the set $L\mathbb{B}^d$ and a pair of datasets such that $\mathcal{O}$ is not (ε, δ)-differentially private for any $\varepsilon < \infty$ and any $\delta < 1$.*

Proof. Let $\mathcal{X} = \{x_1, x_2, ..., x_n\}$ be the database with $x_i \in \mathbb{R}^k$ and $\|x_i\|_2 \leq 1$, $\forall i$. Given this database, consider the function

$$\mathcal{L}(w; \mathcal{X}) = \frac{1}{n} \sum_{i=1}^{n} \|x_i\|_2 \|w\|_2^2,$$

with parameter $w \in L\mathbb{B}^d$. Consider the neighboring databases $\mathcal{X} = \{x_1, x_2, ..., x_{n-1}, x_n\}$ and $\mathcal{X}' = \{x_1, x_2, ..., x_{n-1}, x_n'\}$ differing at the last entry (WLOG). Assigning $x_1, ..., x_{n-1}$ to be 0_k. For the last points take, $x_n = 0$ and $x_n' = \frac{1}{\sqrt{k}} 1_k$. With this construction, we have $\mathcal{L}(w; \mathcal{X}) = 0$ and $\mathcal{L}(w; \mathcal{X}') = \frac{L}{n} \|w\|_2^2$. Let $R_{\mathcal{X}} \sim \mathcal{O}(\mathcal{L}(\cdot; \mathcal{X}), w)$ and $R_{\mathcal{X}'} \sim \mathcal{O}(\mathcal{L}(.; \mathcal{X}'), w)$. By the property of a zero-preserving noisy oracle, we have $\mathbb{P}[R_{\mathcal{X}} = \mathbf{0}_d] = 1$, while $R_{\mathcal{X}'}$ would be a continuous random variable. Hence, for the singleton set $S = \{\mathbf{0}_s\}$, we have $Pr[R_{\mathcal{X}} \in S] = 1$ and $Pr[R_{\mathcal{X}'} \in S] = 0$ because $R_{\mathcal{X}'}$ is a continuous random variable with unbounded support. Clearly, $\mathbb{P}[R_{\mathcal{X}} \in S] > e^{\varepsilon} \mathbb{P}[R_{\mathcal{X}'} \in S] + \delta$ for any $\varepsilon < \infty$ and $\delta < 1$, contradicting (ε, δ)-differential privacy definition. □

A few key observations about this result are

1. Abadi et al. (2016); Charles et al. (2024) showed that for GD and SGD algorithms to be differentially private, each parameter update must be differentially private if the attacker has access to the all the intermediate states. Thus, understanding the privacy preserving properties of distribution of updates (e.g. ZO oracles) helps us in understanding the privacy of the algorithm itself.

2. Theorem 3.3 demonstrates that algorithms involving the sharing of gradient estimates or parameter updates sampled from zero-preserving noisy oracles between parties are not private. For instance, in many federated learning mechanisms (Lowy and Razaviyayn, 2023; Lowy et al., 2023; Aldaghri et al., 2023; Gao et al., 2024), gradients (estimates) are shared from silos to a central server. In the absence of a trusted server, the lack of privacy of the updates, from a particular silo, poses a significant risk to the confidentiality of the data within the silo.

It is also important to note that Theorem 3.3 does not dismiss the privacy guarantee of zeroth order estimators with *independent* additive noise as discussed in Zhang et al. (2024a) and Zhang et al. (2024b). This is because gradient estimators with *independent* additive noise do not satisfy the first property of Definition 3.1. Consider the gradient estimator with an additive noise to be $\widehat{\nabla} f(w) = \widehat{\nabla}' f(w) + \gamma$ where $\widehat{\nabla}' f(w) : \mathbb{R}^d \to \mathbb{R}^d$ represents an arbitrary estimator and $\gamma \sim \mathcal{N}(0, \sigma^2 I_d)$ is an independent noise for a σ which satisfies (ϵ, δ) DP (assuming that $\widehat{\nabla}' f(w)$ has bounded sensitivity). We see that (in the worst case) when $\widehat{\nabla}' f(w)$ is a constant $\widehat{\nabla} f(w)$ is still a continuous distribution since γ is an independent noise added.

4 Privacy of ZO SGD

At this point, we have given a partial answer to the question asked in Zhang et al. (2024a); Tang et al. (2024) with respect to privacy of *their zeroth order oracles.* However, this result does not account for the fact when we have no knowledge about the intermediate states of the algorithm. Chourasia et al. (2021); Altschuler and Talwar (2022) have proven that when one considers the case that the attacker has no access to hidden states, then after a small burn-in period, Projected Noisy SGD on strongly convex and convex functions incurs no additional loss in privacy as T increases. Thus, it is possible for some iterates to not be private individually, but the noise due to the zeroth order oracle "accumulates" over time and gives certain privacy guarantees for the final iterate. Therefore, a natural question is as follows:

> *Is the inherent noise of ZO Stochastic Gradient Descent (SGD) with a constant initialization sufficient to preserve privacy given access to the final iterate only?*

The following theorem answers this question.

Theorem 4.1. *Consider running T steps of Algorithm 1 using a zero-preserving noisy oracle $\mathcal{O}$, as defined in Definition 3.1, with $D > 0$ and $\mathcal{R}_{init}$ as a fixed constant $w_0 \in \mathbb{R}^d$ such that $\|w_0\|_2 < D$. Assume that the algorithm only returns the final iterate. Then, there exists an L-Lipschitz linear loss function over the set $[-D, D]^d \subset \mathbb{R}^d$ such that for any $T \geq 1$ the output of Algorithm 1 is not (ε, δ)-differentially private for any $\varepsilon < \infty$, $\delta < 1$*

Proof. Consider the following function for a database $\mathcal{X} = \{x_1, x_2, ..., x_n\}$ where for all $i \in [n]$ $x_i \in \mathbb{R}^k$ such that $\|x_i\|_2 \leq 1$ with parameter $w \in L\mathbb{B}^d$

$$\mathcal{L}(w; \mathcal{X}) = \frac{1}{n} \sum_{i=1}^{n} \|x_i\|_2 \|w\|_2^2 .$$

Consider the neighboring databases differing at the last entry (WLOG) $\mathcal{X} = \{x_1, x_2, ..., x_{n-1}, x_n\}$ and $\mathcal{X}' = \{x_1, x_2, ..., x_{n-1}, x_n'\}$. Assign $x_1, ..., x_{n-1}$ to be 0_k. For the last points take, $x_n = 0$ and $x_n' = \frac{1}{\sqrt{k}} 1_k$. With the above construction, we have the following two functions under $\mathcal{X}$ and $\mathcal{X}'$, $\mathcal{L}(w; \mathcal{X}) = 0$ and $\mathcal{L}(w; \mathcal{X}') = \frac{L}{n} \|w\|_2^2$.

Take the random variables for iterates corresponding to running Algorithm 1 on χ and χ' to be W_t^{χ} and $W_t^{\chi'}$ respectively. Since $\mathcal{L}(w; \chi) = 0$, then by the first property of $\mathcal{O}$ along with the fact that $\mathcal{O}(f, w) = \mathbf{0}_d$ for a constant f, we have that $\mathbb{P}\left[\mathcal{O}(\mathcal{L}(\cdot, \chi), w) = \mathbf{0}_d\right] = 1$. This implies that the iterate would not change on each step and since $w_0 \in D\mathbb{B}^d$, projection would be identity at each step, which implies $W_t^{\chi} = w_0$.

On the other hand, $W_t^{\chi'} = \Pi_{D\mathbb{B}^d}\left(W_{t-1}^{\chi'} - \eta \widehat{\nabla} \mathcal{L}(W_{t-1}; \mathcal{X}')\right)$. Since the minima of $\mathcal{L}(w; \mathcal{X}')$ lies strictly at $w = 0$, it implies that $\mathcal{L}(w_0; \mathcal{X}') \neq 0$ and since $\mathcal{L}(w; \mathcal{X}')$ is not a constant, then by the second property in Definition 3.1 it implies that $\widehat{\nabla} \mathcal{L}(W_{t-1}^{\chi'}, \chi')$ is a continuous distribution, which implies that $W_{t-1}^{\chi'} - \eta \widehat{\nabla} \mathcal{L}(W_{t-1}; \mathcal{X}')$ is a continuous distribution. However, since we are projecting into a ball at every iteration, then it implies that for $w \in D\mathbb{B}^d$ such that $\|w\|_2 < D$, $\mathbb{P}[W_t^{\chi'} = w] = \mathbb{P}[W_{t-1}^{\chi'} - \eta \widehat{\nabla} \mathcal{L}(W_{t-1}; \mathcal{X}') = w] = 0$ while for $\|w\|_2 = D$, $\mathbb{P}[W_t^{\chi'} = D] = \mathbb{P}[W_{t-1}^{\chi'} - \eta \widehat{\nabla} \mathcal{L}(W_{t-1}; \mathcal{X}') \geq D]$. Thus, if we consider the singleton set $S = \{w_0\}$, then we would get that $Pr[W_T^{\chi} \in S] = 1$ and $Pr[W_T^{\chi'} \in S] = 0$ because $\|w_0\|_2$ is strictly less than D. Hence, $\mathbb{P}[W_T^{\chi} \in S] > e^{\varepsilon} \mathbb{P}[W_T^{\chi} \in S] + \delta$ for any $\varepsilon < \infty$ and $\delta < 1$. □

This result answers the open question proposed by Tang et al. (2024) and Zhang et al. (2024a) on the privacy of zeroth order (S)GD with a fixed initialization. Thus, it dismisses the promise of privacy of (S)GD under such oracles. It is important to note that our framework does not capture certain algorithms like Stochastic Zeroth Order Conditional Gradient Descent (Balasubramanian and Ghadimi, 2022) or Mirror Descent (Duchi et al., 2015). It would be interesting to see if such a result can be generalized for other classes of optimization algorithms discussed in Duchi et al. (2015) or Balasubramanian and Ghadimi (2022).

Notably, in modern machine learning tasks, model parameters are initialized randomly (He et al., 2015). Thus, if we consider the optimization of the functions corresponding to two neighbouring datasets as defined in Definition 2.1, then we cannot comment about the distribution of the final iterate. Therefore, a natural extension to our previous question emerges as follows:

> *Is the inherent noise of ZO Stochastic Gradient Descent (SGD)* ***with random initialization*** *sufficient to preserve privacy given access to the final iterate only?*

It is important to note that the distribution of the initial iterate is continuous. Thus, by proof of Theorem 4.1, the continuity of the zero-preserving noisy oracle is no longer sufficient to ensure privacy. We define a new class of oracles below for which we analyse privacy.

Definition 4.2. *An update oracle $\mathcal{O}$ is C_s-anti-concenterated (AC) if there exists an index $i^* \in [d]$ such that if we consider the function class $\mathcal{F} = \{\langle g\mathbf{e}_{i^*}, w\rangle : -L \leq g \leq 0\}$, then for any $h \in \mathcal{F}$ and $\overline{w} \in \mathbb{R}^d$, if $U \sim \mathcal{O}(h, \overline{w})$ then,*

1. *$\{\nabla h(\overline{w})\}_{i^*} = 0$ implies $U = \mathbf{0}_d$ w.p. 1 i.e. $\mathbb{P}\left[U = \mathbf{0}_d\right] = 1$*

2. $\{\nabla h(\overline{w})\}_{i^*} \neq 0$ *implies* $\mathbb{P}[\{U\}_{i^*} < 0] = 1$

3. $\mathbb{E}[\{U\}_{i^*}] \leq \{\nabla h(\overline{w})\}_{i^*}$

4. *For any set of* $\{w_1, w_2, ..., w_N\} \subset \mathbb{R}^d$*, let* $U_j \sim \mathcal{O}(h, w_j)$ *independently for all* $j \in [N]$*. Then* $\mathbb{E}\left[\left(\sum_{j=1}^N \{U_j\}_{i^*}\right)^2\right] \leq C_s \mathbb{E}\left[\sum_{j=1}^N \{U_j\}_{i^*}\right]^2$

This oracle roughly gives us a stronger guarantee (than zero-preserving noisy oracles) that updates for a certain class of functions (with non-zero gradients) would be strictly increasing while ensuring a zero-preserving property (like zero-preserving noisy oracles) on functions which are zero over $\mathbb{R}^d$. Further its distributional properties, like the condition on expectation and variance, are necessary for ensuring an "anti-concenteration" phenomena over the updates (hence the name). This suggests that the updates from AC oracles would shift the iterates away from the initial point with a good probability, even when it is randomized.

Similar to the definition of zero-preserving noisy oracles, this seemingly specific oracle class is able to capture the two point oracles we have discussed so far. Specifically, we show that the $SPSA$ and FD estimators are 3-AC oracles.

Lemma 4.3. *SPSA and FD are 3-AC.*

Proof. We defer the complete proof to Appendix D. (Sketch) The proof comes from fixing $i^* = 1$ and analyzing the output of the estimator for any function in the function class ($\mathcal{F}$) corresponding to i^* as specified in Definition 4.2. Due to the difference form of the estimators, the distribution of the first index of the outputs of these estimators turns out to be a chi-squared random variable with the degree of freedom 1 scaled by the parameter g defined in Definition 4.2. Hence, this distribution satisfies all the properties mentioned in Definition 4.2, thus completing the proof. □

We further prove that the two-point oracle proposed by Duchi et al. (2015) is also 3-AC. Moreover, the property of being an AC oracle is preserved under the mean extension, as shown by the following lemma.

Lemma 4.4. *If an oracle* $\mathcal{E}$ *is* C_s*-AC then* $M_m^{\mathcal{E}}$ *is* C_s*-AC.*

Proof. We defer the complete proof to Appendix D. (Sketch) The argument here is, we are taking mean of i.i.d. random variables $U_i^{(f,w_j)} \sim \mathcal{E}(f, w_j)$ for all $i \in [m]$ for $\mathcal{E}$ which are C_s anti-concentrated. Hence, the "strict" properties (namely zero-preserving property and the strict negativity under non-zero gradient of AC oracles) of samples of $\mathcal{E}(f, w_j)$ also comply to the mean of i.i.d. samples from distribution satisfying property 1 and property 2. Due to linearity of expectation, the same upper bound on the expectation holds for the mean satisfying property 3. Due to reordering of RV and Young's inequality, property 4 also complies to the mean. □

This lemma directly implies that M_m^{FD} and M_m^{SPSA} are 3-AC.

If we restrict ourselves to the class of zeroth order oracles, then we make an interesting observation: All the estimators discussed above that use at least *two function evaluations* satisfy the properties mentioned in Definition 4.2. On the other hand, for the SP estimator defined in Example 2.3, we observe that it does not satisfy the third property of Definition 4.2. Thus, restricted to zeroth order oracles, the idea of an AC oracle roughly captures an important distinction between popular single-point and two-point estimator(s).

Using the definition of an AC oracle, we show that even with unknown (but random) initialization, the model loses privacy for a large enough diameter of the constraint set.

Theorem 4.5. *Consider running* T *steps of Algorithm 1 using a* C_s*-AC oracle* $\mathcal{O}$*, as defined in Definition 4.2, with* $D > \frac{\eta L}{2n}$ *and* $\mathcal{R}_{init}$ *is* $\mathcal{N}(0, \sigma^2 I)$*. Assume that the algorithm only returns the final iterate. Then, there exists an L-Lipschitz linear loss function over the set* $[-D, D]^d \subset \mathbb{R}^d$ *such that for any* $T \geq 1$ *the output of Algorithm 1 is not* (ε, δ)*-differentially private for any* ε, δ *satisfying* $\delta \leq \frac{1}{16C_s} \max\left\{\frac{1}{T^{2/3}}, \frac{\eta L}{2nD}\right\}$ *and*

$$\varepsilon \leq \min\left\{\frac{\eta^2 L^2 T^{4/3}}{8n^2\sigma^2}, \frac{D^2}{2\sigma^2}\right\} + \ln\left(\frac{\sqrt{2\pi} L\eta}{64 C_s n\sigma}\right).$$

We defer the proof to Appendix C.1. If we take the values of $n \in [100, 1000]$, $\eta \in [0.001, 0.01]$, $C_s = 3$, $L = 100$, and $\sigma = \frac{1}{1020}$, we see that for a large enough $D\left(\geq \frac{T^{2/3}}{200}\right)$ this result gives us that privacy is not possible (roughly) for $\epsilon \leq 896 \cdot T^{4/3}$ and $\delta \leq \frac{1}{96T^{2/3}}$. Notably, we get the same flavor of result as obtained in the lower bound construction of Altschuler and Talwar (2022). Hence, it would be interesting to see if a matching upper bound is attainable for Projected ZO-GD.

5 Conclusion and Open Problems

In this work, we demonstrated that projected (stochastic) ZO-gradient descent cannot ensure differential privacy without incorporating additional (additive) noise. While our findings apply to a broad class of zeroth-order oracles, the algorithmic framework we used is limited to the class of projected SGD. Potential future directions for this problem include:

1. Can we prove Theorem 4.5 for other function classes (e.g., strongly convex loss functions)?
2. Can we extend our (lower-bound) analysis beyond projected (stochastic) ZO-GD?
3. Although zeroth-order estimators do not offer inherent privacy on their own, can they amplify the privacy of other (additive) noisy methods?

The answer to any of these questions will help understand the limitations and potential of zeroth-order optimization for private optimization.

Acknowledgments

The work of MR and DG was supported by a gift from Meta, a gift from Google, and a gift from Amazon. VS was supported by NSF CAREER Award CCF-2239265 and an Amazon Research Award. The work was partly done when VS was visiting the Simons Institute in Berkeley.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. (2016). Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, page 308–318, New York, NY, USA. Association for Computing Machinery.

Aldaghri, N., Mahdavifar, H., and Beirami, A. (2023). Federated learning with heterogeneous differential privacy.

Altschuler, J. and Talwar, K. (2022). Privacy of noisy stochastic gradient descent: More iterations without more privacy loss. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K., editors, *Advances in Neural Information Processing Systems*.

Arora, R., Bassily, R., González, T., Guzmán, C., Menart, M., and Ullah, E. (2023). Faster rates of convergence to stationary points in differentially private optimization. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Asi, H., Feldman, V., Koren, T., and Talwar, K. (2021). Private stochastic convex optimization: Optimal rates in l1 geometry. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 393–403. PMLR.

Balasubramanian, K. and Ghadimi, S. (2022). Zeroth-Order nonconvex stochastic optimization: Handling constraints, high dimensionality, and saddle points. *Foundations of Computational Mathematics*, 22(1):35–76.

Bassily, R., Feldman, V., Talwar, K., and Guha Thakurta, A. (2019). Private stochastic convex optimization with optimal rates. *Advances in neural information processing systems*, 32.

Bassily, R., Smith, A., and Thakurta, A. (2014). Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 464–473.

Block, A., Bun, M., Desai, R., Shetty, A., and Wu, S. (2024). Oracle-efficient differentially private learning with public data.

Carmon, Y., Jambulapati, A., Jin, Y., Lee, Y. T., Liu, D., Sidford, A., and Tian, K. (2023). Resqueing parallel and private stochastic convex optimization. In *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 2031–2058.

Charles, Z., Ganesh, A., McKenna, R., McMahan, H. B., Mitchell, N., Pillutla, K., and Rush, K. (2024). Fine-tuning large language models with user-level differential privacy.

Chaudhuri, K. and Hsu, D. (2011). Sample complexity bounds for differentially private learning. In Kakade, S. M. and von Luxburg, U., editors, *Proceedings of the 24th Annual Conference on Learning Theory*, volume 19 of *Proceedings of Machine Learning Research*, pages 155–186, Budapest, Hungary. PMLR.

Chaudhuri, K., Monteleoni, C., and Sarwate, A. D. (2011). Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(29):1069–1109.

Chourasia, R., Ye, J., and Shokri, R. (2021). Differential privacy dynamics of langevin diffusion and noisy gradient descent. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 14771–14781. Curran Associates, Inc.

Cohen, E., Lyu, X., Nelson, J., Sarlós, T., and Stemmer, U. (2023). Optimal differentially private learning of thresholds and quasi-concave optimization. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, STOC 2023, page 472–482, New York, NY, USA. Association for Computing Machinery.

Dagan, Y. and Shamir, O. (2018). Detecting correlations with little memory and communication. In Bubeck, S., Perchet, V., and Rigollet, P., editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 1145–1198. PMLR.

Ding, B., Kulkarni, J., and Yekhanin, S. (2017). Collecting telemetry data privately. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In Halevi, S. and Rabin, T., editors, *Theory of Cryptography*, pages 265–284, Berlin, Heidelberg. Springer Berlin Heidelberg.

Dwork, C., Rothblum, G. N., and Vadhan, S. (2010a). Boosting and differential privacy. In *Proceedings of the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, FOCS '10, page 51–60, USA. IEEE Computer Society.

Dwork, C., Rothblum, G. N., and Vadhan, S. (2010b). Boosting and differential privacy. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 51–60.

Erlingsson, U., Pihur, V., and Korolova, A. (2014). Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, CCS '14, page 1054–1067, New York, NY, USA. Association for Computing Machinery.

Feldman, V., Koren, T., and Talwar, K. (2020). Private stochastic convex optimization: optimal rates in linear time. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2020, page 439–449, New York, NY, USA. Association for Computing Machinery.

Flaxman, A. D., Kalai, A. T., and McMahan, H. B. (2005). Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '05, page 385–394, USA. Society for Industrial and Applied Mathematics.

Gao, C., Lowy, A., Zhou, X., and Wright, S. J. (2024). Private heterogeneous federated learning without a trusted server revisited: Error-optimal and communication-efficient algorithms for convex losses.

Gonen, A., Hazan, E., and Moran, S. (2019). Private learning implies online learning: An efficient reduction. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Gopi, S., Lee, Y. T., and Liu, D. (2022). Private convex optimization via exponential mechanism. In Loh, P.-L. and Raginsky, M., editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 1948–1989. PMLR.

Gopi, S., Lee, Y. T., Liu, D., Shen, R., and Tian, K. (2023). Private convex optimization in general norms. In *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 5068–5089. SIAM.

Guo, W., Long, J., Zeng, Y., Liu, Z., Yang, X., Ran, Y., Gardner, J. R., Bastani, O., Sa, C. D., Yu, X., Chen, B., and Xu, Z. (2024). Zeroth-order fine-tuning of llms with extreme sparsity.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034.

Hu, E. J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2022). LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Kulkarni, J., Lee, Y. T., and Liu, D. (2021). Private non-smooth erm and sco in subquadratic steps. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 4053–4064. Curran Associates, Inc.

Lowy, A., Ghafelebashi, A., and Razaviyayn, M. (2023). Private non-convex federated learning without a trusted server. In Ruiz, F., Dy, J., and van de Meent, J.-W., editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 5749–5786. PMLR.

Lowy, A. and Razaviyayn, M. (2023). Private federated learning without a trusted server: Optimal algorithms for convex losses.

Lowy, A., Ullman, J., and Wright, S. (2024). How to make the gradients small privately: Improved rates for differentially private non-convex optimization. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F., editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 32904–32923. PMLR.

Malladi, S., Gao, T., Nichani, E., Damian, A., Lee, J. D., Chen, D., and Arora, S. (2023). Fine-tuning language models with just forward passes. *Advances in Neural Information Processing Systems*, 36:53038–53075.

Nemirovsky, A. S. and Yudin, D. B. (1983). Problem complexity and method efficiency in optimization. *Wiley*.

Nesterov, Y. and Spokoiny, V. G. (2015). Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17:527 – 566.

Paley, R. E. and Zygmund, A. (1932). On some series of functions,(3). In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 28, pages 190–205. Cambridge University Press.

Rogers, R., Subramaniam, S., Peng, S., Durfee, D., Lee, S., Kancha, S. K., Sahay, S., and Ahammad, P. (2021). Linkedin's audience engagements api: A privacy preserving data analytics system at scale. *Journal of Privacy and Confidentiality*, 11(3).

Shamir, O. (2013). On the complexity of bandit and derivative-free stochastic convex optimization. In Shalev-Shwartz, S. and Steinwart, I., editors, *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pages 3–24, Princeton, NJ, USA. PMLR.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. (2017). Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE.

Spall, J. (1992). Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control*, 37(3):332–341.

Tang, X., Panda, A., Nasr, M., Mahloujifar, S., and Mittal, P. (2024). Private fine-tuning of large language models with zeroth-order optimization.

Wibisono, A., Wainwright, M. J., Jordan, M., and Duchi, J. C. (2012). Finite sample convergence rates of zero-order stochastic optimization methods. In Pereira, F., Burges, C., Bottou, L., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

Yousefpour, A., Shilov, I., Sablayrolles, A., Testuggine, D., Prasad, K., Malek, M., Nguyen, J., Ghosh, S., Bharadwaj, A., Zhao, J., Cormode, G., and Mironov, I. (2022). Opacus: User-friendly differential privacy library in pytorch.

Yue, P., Yang, L., Fang, C., and Lin, Z. (2023). Zeroth-order optimization with weak dimension dependency. In Neu, G. and Rosasco, L., editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 4429–4472. PMLR.

Zhang, L., Li, B., Thekumparampil, K. K., Oh, S., and He, N. (2024a). DPZero: Private fine-tuning of language models without backpropagation. In *International Conference on Machine Learning*. PMLR.

Zhang, Q., Tran, H., and Cutkosky, A. (2024b). Private zeroth-order nonsmooth nonconvex optimization. In *The Twelfth International Conference on Learning Representations*.

Zhang, Y., Li, P., Hong, J., Li, J., Zhang, Y., Zheng, W., Chen, P.-Y., Lee, J. D., Yin, W., Hong, M., Wang, Z., Liu, S., and Chen, T. (2024c). Revisiting zeroth-order optimization for memory-efficient llm fine-tuning: A benchmark.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes/No/Not Applicable] **Yes**
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes/No/Not Applicable] **No**
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes/No/Not Applicable] **Not Applicable**

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes/No/Not Applicable] **Yes**
   (b) Complete proofs of all theoretical results. [Yes/No/Not Applicable] **Yes**
   (c) Clear explanations of any assumptions. [Yes/No/Not Applicable] **Yes**

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes/No/Not Applicable] **Not Applicable**
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes/No/Not Applicable] **Not Applicable**
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes/No/Not Applicable] **Not Applicable**
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes/No/Not Applicable] **Not Applicable**

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes/No/Not Applicable] **Not Applicable**
   (b) The license information of the assets, if applicable. [Yes/No/Not Applicable] **Not Applicable**
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes/No/Not Applicable] **Not Applicable**
   (d) Information about consent from data providers/curators. [Yes/No/Not Applicable] **Not Applicable**
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes/No/Not Applicable] **Not Applicable**

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Yes/No/Not Applicable] **Not Applicable**
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Yes/No/Not Applicable] **Not Applicable**
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Yes/No/Not Applicable] **Not Applicable**

A Useful Inequalities

Lemma A.1 (Gaussian Concentration (Dagan and Shamir, 2018))**.** *Let W be a random variable (RV) distributed normally with mean 0 and variance σ^2. Then, for any $w > 0$,*

$$\mathbb{P}\left[W \geq w\right] \leq \frac{\sigma e^{-w^2/2\sigma^2}}{w\sqrt{2\pi}}.$$

Lemma A.2 ((Paley and Zygmund, 1932))**.** *Let Z be a non-negative random variable (RV) and let $\alpha \in [0,1]$, then*

$$\mathbb{P}\left[Z \geq \alpha \mathbb{E}\left[Z\right]\right] \geq (1-\alpha)^2 \frac{(\mathbb{E}\left[Z\right])^2}{\mathbb{E}\left[Z^2\right]}.$$

B Proofs and Discussions of Section 3

B.1 Proof of Lemma 3.2

Proof. By the definition of SPSA estimator for $f(w)$, $\mu > 0$ and $Z \sim \mathcal{N}(0, I_d)$, we have

$$SPSA(f,w) = \frac{f(w+\mu Z) - f(w - \mu Z)}{2\mu} Z.$$

If $f(w) = 0$ for all $w \in \mathbb{R}^d$, then $SPSA(f,w) = \mathbf{0}_d$ satisfying property 1. Moreover, for the function $f(w) = \frac{a}{2}\left\|w\right\|_2^2$ for $a > 0$, we have:

$$\begin{aligned} SPSA(f,w) &= \frac{a}{2}\frac{\left\|w+\mu Z\right\|_2^2 - \left\|w - \mu Z\right\|_2^2}{2\mu} Z \\ &= \frac{a}{2}\frac{\left(\left\|w\right\|_2^2 + \left\|\mu Z\right\|_2^2 + 2\mu w^T Z\right) - \left(\left\|w\right\|_2^2 + \left\|\mu Z\right\|_2^2 - 2\mu w^T Z\right)}{2\mu} Z \\ &= aZZ^T w. \end{aligned}$$

Then for any $c \in \mathbb{R}^d$, consider its first index $\{c\}_1$, then the event $aZZ^Tw = c$ implies that $\{aZZ^Tw\}_1 = \{c\}_1$ i.e. $\mathbb{P}\left[aZZ^Tw = c\right] \leq \mathbb{P}\left[\{aZZ^Tw\}_1 = \{c\}_1\right]$. Now, we can write $\{aZZ^Tw\}_1 = aw_1\{Z\}_1^2 + a\sum_{i=2}^d w_i\{Z\}_i\{Z\}_1$. Since $Z \sim \mathcal{N}(0, I_d)$, it implies that $\{aZZ^Tw\}_1$ is a non-constant polynomial in $\{Z\}_1, ..., \{Z\}_d$ which means that $\{aZZ^Tw\}_1$ is continuous, implying that $\mathbb{P}\left[\{aZZ^Tw\}_1 = \{c\}_1\right] = 0$ for any $\{c\}_1 \in \mathbb{R}$, implying that $\mathbb{P}\left[aZZ^Tw = c\right] = 0$. Hence, $SPSA(f,w)$ is a continuous RV satisfying property 2.

Similarly, by definition of the FD estimator for $f(w)$, $\mu > 0$ and $Z \sim \mathcal{N}(0, I_d)$, we get that

$$FD(f,w) = \frac{f(w+\mu Z) - f(w)}{2\mu} Z$$

If $f(w) = 0$ for all $w \in \mathbb{R}^d$, then $FD(f,w) = \mathbf{0}_d$ satisfying property 1. If $f(w) = \frac{a}{2}\left\|w\right\|_2^2$ for $a > 0$, we have that:

$$\begin{aligned} FD(f,w) &= \frac{a}{2}\frac{\left\|w+\mu Z\right\|_2^2 - \left\|w\right\|_2^2}{\mu} Z \\ &= \frac{a}{2}\frac{\left(\left\|w\right\|_2^2 + \left\|\mu Z\right\|_2^2 + 2\mu w^T Z\right) - \left\|w\right\|_2^2}{\mu} Z \\ &= \frac{a}{2}\left(\mu\left\|Z\right\|_2^2 + 2w^T Z\right) Z. \end{aligned}$$

Considering a similar argument as the SPSA estimator, we evaluate $\{FD(f,w)\}_1$, which is equal to $\frac{a\mu}{2}\{Z\}_1^3 + \frac{a\mu}{2}\sum_{i=2}^{d}\{Z\}_i^2\{Z\}_1 + aw_1\{Z\}_1^2 + a\sum_{i=2}^{d} w_i\{Z\}_i\{Z\}_1$. Since $\{FD(f,w)\}_1$ is a non-constant polynomial in $\{Z\}_1, ..., \{Z\}_d$, $\{FD(f,w)\}_1$ is continuous. Hence, $FD(f,w)$ is a continuous RV satisfying property 2.

Similarly, by definition of the SP estimator for $f(w)$, $\mu > 0$ and $Z \sim \mathcal{N}(0, I_d)$, we get that

$$SP(f,w) = \frac{d}{\mu} f(w + \mu Z)Z$$

If $f(w) = 0$ for all $w \in \mathbb{R}^d$, then $FD(f,w) = \mathbf{0}_d$ satisfying property 1. If $f(w) = \frac{a}{2}\|w\|_2^2$ for $a > 0$, we have that:

$$\begin{aligned} SP(f,w) &= \frac{d}{\mu} f(w + \mu Z)Z \\ &= \frac{ad}{2\mu}\|w + \mu Z\|_2^2 Z \\ &= \frac{a}{2}\left(\|w\|_2^2 + \mu^2\|Z\|_2^2 + 2w^T Z\right) Z. \end{aligned}$$

Considering similar arguments as the SPSA and FD estimator, we see that $\{SP(f,w)\}_1$ is a non-constant polynomial in $\{Z\}_1, ..., \{Z\}_d$ which implies that $\{SP(f,w)\}_1$ is continuous. Hence, $SP(f,w)$ is a continuous RV satisfying property 2.

For the mean extensions of the given oracles, we use the definition of the mean oracle

$$M^{\mathcal{E}}(f,w,m) = \frac{1}{m}\sum_{i=1}^{m} U_i,$$

where U_i drawn i.i.d. from $\mathcal{E}(f,w)$. Each U_i is $\mathbf{0}_d$ when $f(w) = 0$ for all $w \in \mathbb{R}^d$. Then, we get that for $f(w) = 0$ for all $w \in \mathbb{R}^d$, $M^{\mathcal{E}}(f,w,m) = \mathbf{0}_d$, satisfying property 1. In the other case, since $\mathcal{E}_i(f,w)$ is continuous for all $i \in [m]$. We utilize the fact that if multiple continuous and independent random variables are added then their resultant addition is also a continuous random variable. Thus, the results extend for M^{SPSA}, M^{FD}, and M^{SP}. □

B.2 Further Discussion on Zeroth Order Estimators

Consider the estimator given by Duchi et al. (2015) for minimization of non-smooth functions.

Definition B.1 (Duchi et al. (2015))**.** *If $Z_1, Z_1 \sim \mathcal{N}(0, I_d)$, then the estimator is given by*

$$G_{\mu_1,\mu_2}(f,w) = \frac{f(w + \mu_1 Z_1 + \mu_2 Z_2) - f(w + \mu_1 Z_1)}{\mu_2} Z_2.$$

Lemma B.2. *The oracle defined in Definition B.1 is a zero-preserving noisy oracle.*

Proof. Using identical arguments from the proof of Theorem 3.2, we see that with $f(w) = 0$ for all $w \in \mathbb{R}^d$, $G_{\mu_1,\mu_2}(f,w) = \mathbf{0}_d$. Calculating for $f(w) = \frac{A}{2}\|w\|_2^2$, we get that

$$\begin{aligned} G_{\mu_1,\mu_2}(f,w) = &= \frac{A}{2}\frac{\|w + \mu_1 Z_1 + \mu_2 Z_2\|_2^2 - \|w + \mu_1 Z_1\|_2^2}{\mu_2} Z_2 \\ &= \frac{A}{2}\frac{\left(\|w\|_2^2 + \|\mu_1 Z_1 + \mu_2 Z_2\|_2^2 + 2w^T(\mu_1 Z_1 + \mu_2 Z_2)\right) - \|w\|_2^2}{\mu_2} Z_2 \\ &= \frac{A}{2\mu_2}\left(\|\mu_1 Z_1 + \mu_2 Z_2\|_2^2 + 2w^T(\mu_1 Z_1 + \mu_2 Z_2)\right) Z_2. \end{aligned}$$

Thus, using arguments similar to the proof of Lemma 3.2 $G_{\mu_1,\mu_2}(f,w)$ would have a continuous distribution, proving our claim. □

The above lemma also implies that $\mathcal{M}^{G_{\mu_1,\mu_2}}$ would also be a continuous distribution. It is evident that almost every zeroth order estimator used in the literature (yet) satisfies the properties of zero-preserving noisy oracles.

C Proofs of Section 4

C.1 Proof of Theorem 4.5

Definition C.1. *(Restated Definition 4.2) An update oracle $\mathcal{O}$ is C_s-anti-concenterated if there exists an index $i^* \in [d]$ such that if we consider the function class $\mathcal{F} = \{\langle g\mathbf{e}_{i^*}, w\rangle : -L \leq g \leq 0\}$, then for any $h \in \mathcal{F}$ and $\overline{w} \in \mathbb{R}^d$, if $U \sim \mathcal{O}(h, \overline{w})$ then,*

1. *$\{\nabla h(\overline{w})\}_{i^*} = 0$ implies $U = \mathbf{0}_d$ w.p. 1 i.e. $\mathbb{P}[U = \mathbf{0}_d] = 1$*
2. *$\{\nabla h(\overline{w})\}_{i^*} \neq 0$ implies $\mathbb{P}[\{U\}_{i^*} < 0] = 1$*
3. *$\mathbb{E}[\{U\}_{i^*}] \leq \{\nabla h(\overline{w})\}_{i^*}$*
4. *For any set of $\{w_1, w_2, ..., w_N\} \subset \mathbb{R}^d$, let $U_j \sim \mathcal{O}(h, w_j)$ independently for all $j \in [N]$. Then $\mathbb{E}\left[\left(\sum_{j=1}^N \{U_j\}_{i^*}\right)^2\right] \leq C_s \mathbb{E}\left[\sum_{j=1}^N \{U_j\}_{i^*}\right]^2$*

Theorem C.2. *(Restated Theorem 4.5) Consider running T steps of Algorithm 1 using a C_s-AC oracle $\mathcal{O}$, as defined in Definition 4.2, with $D > \frac{\eta L}{2n}$ and $\mathcal{R}_{init}$ is $\mathcal{N}(0, \sigma^2 I)$. Assume that the algorithm only returns the final iterate. Then, there exists an L-Lipschitz linear loss function over the set $[-D, D]^d \subset \mathbb{R}^d$ such that for any $T \geq 1$ the output of Algorithm 1 is not (ε, δ)-differentially private for any ε, δ satisfying $\delta \leq \frac{1}{16C_s} \max\left\{\frac{1}{T^{2/3}}, \frac{\eta L}{2nD}\right\}$ and*

$$\varepsilon \leq \min\left\{\frac{\eta^2 L^2 T^{4/3}}{8n^2\sigma^2}, \frac{D^2}{2\sigma^2}\right\} + \ln\left(\frac{\sqrt{2\pi} L\eta}{64 C_s n\sigma}\right).$$

Proof. Consider the following loss function for a database $\mathcal{X} = \{x_1, x_2, ..., x_n\}$

$$\mathcal{L}(w; \mathcal{X}) = \frac{-1}{n}\left\langle w, \sum_{i=1}^n x_i\right\rangle,$$

where $\|x_i\|_2 \leq L$ for all $i \in [n]$ and $x_i \in \mathbb{R}^d$ with $w \in [-D, D]^d$. Consider the neighbouring databases $\chi = \{x_1, x_2, ..., x_{n-1}, x_n\}$ and $\chi' = \{x_1, x_2, ..., x_{n-1}, x'_n\}$ differing only at the last entry. Let $x_1, ..., x_{n-1}, x_n$ to be $\mathbf{0}_d$. Given the index i^* defined in Defintion 4.2 for C_s-AC oracles, let $x'_n = L\mathbf{e}_{i^*}$. With this construction, we have $\mathcal{L}(w; \mathcal{X}) = 0$ and $\mathcal{L}(w; \mathcal{X}') = \frac{-L}{n}\langle w, \mathbf{e}_{i^*}\rangle$. Let $W_t^{\mathcal{X}}$ and $W_t^{\mathcal{X}'}$ be the t^{th} iterate of Algorithm 1 is run on $\mathcal{L}(\cdot, \mathcal{X})$ and $\mathcal{L}(\cdot, \mathcal{X}')$, respectively.

To show that the algorithm is differentially private, we need to define a measurable set S so that Definition 2.1 fails. Let

$$S = \left\{w \in \mathbb{R}^d : \{w\}_{i^*} \geq \min\left\{\frac{\eta L}{2n} T^{2/3}, D\right\}\right\}.$$

We will show that for this set S, $\mathbb{P}\left[W_T^{\mathcal{X}'} \in S\right] \geq e^{\varepsilon_0}\mathbb{P}\left[W_T^{\mathcal{X}} \in S\right] + \delta_0$ for some ε_0 and δ_0.

Note: The two adjacent functions have been designed in a manner such that in one case, the iterate stays near origin with high probability. In the other case, it shifts away from its original point with a good probability, enough to ensure separation between the high probability regions of the two iterates.

Based on our definition of S, we divide our analysis into two cases:

- **Case 1:** $T \leq \left(\frac{2nD}{\eta L}\right)^{3/2}$ or equivalently $\frac{\eta L}{2n}T^{2/3} \leq D$

  Computing $\mathbb{P}\left[W_T^{\mathcal{X}'} \in S\right]$ Since we have assumed the constraint space to be a hypercube, projection corresponds to coordinate wise clipping. Hence, it would suffice to analyze the setting for one coordinate without affecting the other coordinates and vice versa. Due to the second property in Definition 4.2, we have that $\{W_t^{\mathcal{X}'}\}_{i^*}$ is monotonic, i.e. $\{W_{t-1}^{\mathcal{X}'}\}_{i^*} \leq \{W_t^{\mathcal{X}'}\}_{i^*}$ for $t \geq 2$. Notice that if $W_T^{\mathcal{X}'} \notin S$ then $\{W_T^{\mathcal{X}'}\}_{i^*} < D$ which implies that no projection occurred in the i^{*th} coordinate on the right side of the interval $[-D, D]$ in T iterations. Hence, $\{W_T^{\mathcal{X}'}\}_{i^*} = \max\left\{-D, \{W_0^{\mathcal{X}'}\}_{i^*} - \eta\{U_1^{\mathcal{X}'}\}_{i^*}\right\} - \eta\sum_{j=2}^{T}\{U_j^{\mathcal{X}'}\}_{i^*}$ where $U_j^{\mathcal{X}'} \sim \mathcal{O}\left(\mathcal{L}(\cdot;\mathcal{X}'), W_{j-1}^{\mathcal{X}'}\right)$. Hence,

$$\begin{aligned}
\mathbb{P}\left[W_T^{\mathcal{X}'} \in S\right] &= 1 - \mathbb{P}\left[\max\left\{-D, \{W_0^{\mathcal{X}'}\}_{i^*} - \eta\{U_1^{\mathcal{X}'}\}_{i^*}\right\} - \eta\sum_{j=2}^{T}\{U_j^{\mathcal{X}'}\}_{i^*} < \frac{\eta L}{2n}T^{2/3}\right] \\
&\geq 1 - \mathbb{P}\left[\{W_0^{\mathcal{X}'}\}_{i^*} - \eta\sum_{j=1}^{T}\{U_j^{\mathcal{X}'}\}_{i^*} < \frac{\eta L}{2n}T^{2/3}\right] \\
&= \mathbb{P}\left[\{W_0^{\mathcal{X}'}\}_{i^*} - \eta\sum_{j=1}^{T}\{U_j^{\mathcal{X}'}\}_{i^*} \geq \frac{\eta L}{2n}T^{2/3}\right], \qquad \text{(a)}
\end{aligned}$$

where the inequality is due to the fact $\{W_T^{\mathcal{X}'}\}_{i^*} \geq \{W_0^{\mathcal{X}'}\}_{i^*} - \eta\sum_{j=1}^{T}\{U_j^{\mathcal{X}'}\}_{i^*}$. Next, we obtain a lower bound on $\mathbb{P}\left[\{W_0^{\mathcal{X}'}\}_{i^*} - \eta\sum_{j=1}^{T}\{U_j^{\mathcal{X}'}\}_{i^*} \geq \frac{\eta L}{2n}T^{2/3}\right]$. Take $Z = -\sum_{j=1}^{T}\{U_j^{\mathcal{X}'}\}_{i^*}$. We have

$$\begin{aligned}
\mathbb{P}\left[\{W_0^{\mathcal{X}'}\}_{i^*} + \eta Z \geq \frac{\eta L}{2n}T^{2/3}\right] &\geq \mathbb{P}\left[\eta Z \geq \frac{\eta L}{2n}T^{2/3} - \{W_0^{\mathcal{X}'}\}_{i^*} \middle| \{W_0^{\mathcal{X}'}\}_{i^*} \geq 0\right]\mathbb{P}\left[\{W_0^{\mathcal{X}'}\}_{i^*} \geq 0\right] \\
&\geq \frac{1}{2}\mathbb{P}\left[Z \geq \frac{L}{2n}T^{2/3}\right] \\
&\geq \frac{1}{2}\mathbb{P}\left[Z \geq \frac{1}{2T^{1/3}}\frac{LT}{n}\right] \\
&\geq \frac{1}{2}\mathbb{P}\left[Z \geq \frac{1}{2T^{1/3}}\mathbb{E}\left[Z\right]\right]. \qquad (1)
\end{aligned}$$

In the first inequality, we used the fact that $\{W_0^{\mathcal{X}'}\}_{i^*} \sim \mathcal{N}(0, \sigma^2)$ and therefore $\mathbb{P}\left[\{W_0^{\mathcal{X}'}\}_{i^*} \geq 0\right] = \frac{1}{2}$. In the fourth inequality, we used the third property of the C_s-AC oracle in Definition 4.2 which (with linearity of expectation) implies that $\mathbb{E}\left[Z\right] = \mathbb{E}\left[-\sum_{j=1}^{T}\{U_j\}_{i^*}\right] \leq -\sum_{j=1}^{T}\{\nabla\mathcal{L}(W_j^{\chi'};\chi')\}_{i^*}$ and the fact that $\{\nabla\mathcal{L}(w;\chi')\}_{i^*} = -\frac{L}{n}$ for all $w \in \mathbb{R}^d$, by our construction.

Using, the second property of C_s-AC oracle, we have $Z = -\sum_{j=1}^{T}\{U_j\}_{i^*} \geq 0$. Thus, applying Paley-Zygmund (Lemma A.2) on random variable Z for $\alpha = \frac{1}{2T^{1/3}}$, we get

$$\begin{aligned}
\mathbb{P}\left[Z \geq \frac{1}{2T^{1/3}}\mathbb{E}\left[Z\right]\right] &\geq \left(1 - \frac{1}{2T^{1/3}}\right)^2\frac{(\mathbb{E}\left[Z\right])^2}{\mathbb{E}\left[Z^2\right]} \\
&\geq \frac{1}{4C_sT^{2/3}}. \qquad (2)
\end{aligned}$$

In the second inequality, we use the fourth property of C_s-AC oracles as defined in Definition 4.2 to get $\mathbb{E}\left[\left(\sum_{j=1}^{N}\{U_j\}_{i^*}\right)^2\right] \leq C_s\mathbb{E}\left[\sum_{j=1}^{N}\{U_j\}_{i^*}\right]^2$ which implies that $\mathbb{E}\left[Z^2\right] \leq C_s(\mathbb{E}\left[Z\right])^2$ and $T \geq 1$. Combining inequalities 1 and 2, we get that,

$$\mathbb{P}\left[\{W_0^{\mathcal{X}'}\}_{i^*} - \eta\sum_{j=1}^{T}\{U_j^{\mathcal{X}'}\}_{i^*} \geq \frac{\eta L}{2n}T^{2/3}\right] \geq \frac{1}{8C_sT^{2/3}}. \qquad \text{(b)}$$

Combining inequalities (a) and (b), we get

$$\mathbb{P}\left[W_T^{\mathcal{X}'} \in S\right] \geq \frac{1}{8C_s T^{2/3}}. \tag{C1}$$

So far, we computed a lower bound on $\mathbb{P}\left[W_T^{\mathcal{X}'} \in S\right]$. Next, we compute an upper bound on $\mathbb{P}\left[W_T^{\mathcal{X}} \in S\right]$.

Computing $\mathbb{P}\left[W_T^{\mathcal{X}} \in S\right]$ Using the first property of the C_s-AC oracle in Definition 4.2, we have that $W_T^{\mathcal{X}} = W_0 \sim \mathcal{N}(0, I_d)$. Since the projection operator simply projects any value outside the interval to the edge, it does not change the inverse CDF on points which are within the interval. Then, we get that

$$\begin{aligned}\mathbb{P}\left[W_T^{\mathcal{X}} \in S\right] &= \mathbb{P}\left[\{w_0\}_{i^*} \geq \frac{\eta L}{2n} T^{2/3}\right]\\ &\leq \frac{2n\sigma}{\sqrt{2\pi}\eta L T^{2/3}} e^{-\frac{\eta^2 L^2 T^{4/3}}{8n^2\sigma^2}},\end{aligned} \tag{C2}$$

where the inequality is due to Gaussian Concenteration (Lemma A.1) . Using inequalities (C1) and (C2), we get that

$$\frac{\mathbb{P}\left[W_t^{\mathcal{X}'} \in S\right] - \frac{1}{16C_s T^{2/3}}}{\mathbb{P}\left[W_T^{\mathcal{X}} \in S\right]} \geq \frac{\sqrt{2\pi}\eta L}{32 n\sigma C_s} e^{\frac{\eta^2 L^2 T^{4/3}}{8n^2\sigma^2}}.$$

Therefore,

$$\mathbb{P}\left[W_t^{\mathcal{X}'} \in S\right] \geq e^{\frac{\eta^2 L^2 T^{4/3}}{8n^2\sigma^2} + \ln\left(\frac{\sqrt{2\pi}L\eta}{32C_s n\sigma}\right)} \mathbb{P}\left[W_T^{\mathcal{X}} \in S\right] + \frac{1}{16C_s T^{2/3}}.$$

- **Case 2:** $T \geq \left(\frac{2nD}{\eta L}\right)^{3/2}$ or equivalently $\frac{\eta L}{2n} T^{2/3} \geq D$

  The approach to this case is the same Case 1. It only differs in the computation of the constants and dependence on the respective variables.

  Computing $\mathbb{P}\left[W_T^{\mathcal{X}'} \in S\right]$ Using the same argument as that in case 1, we get that

$$\mathbb{P}\left[W_T^{\mathcal{X}'} \in S\right] \geq \mathbb{P}\left[\{W_0^{\mathcal{X}'}\}_{i^*} - \eta \sum_{j=1}^{T} \{U_j^{\mathcal{X}'}\}_{i^*} \geq D\right] \tag{d}$$

  Now, to obtain the lower bound on $\mathbb{P}\left[\{W_0^{\mathcal{X}'}\}_{i^*} - \eta \sum_{j=1}^{T} \{U_j^{\mathcal{X}'}\}_{i^*} \geq D\right]$, we use the same series of steps as those in Case 1.

$$\begin{aligned}\mathbb{P}\left[\{W_0^{\mathcal{X}'}\}_{i^*} - \eta \sum_{j=1}^{T} \{U_j^{\mathcal{X}'}\}_{i^*} \geq D\right] &\geq \frac{1}{2}\mathbb{P}\left[-\sum_{j=1}^{T} \{U_j^{\mathcal{X}'}\}_{i^*} \geq \frac{D}{\eta}\right]\\ &\geq \frac{1}{2}\mathbb{P}\left[-\sum_{j=1}^{T} \{U_j^{\mathcal{X}'}\}_{i^*} \geq \frac{Dn}{\eta LT}\frac{LT}{n}\right]\\ &\geq \frac{1}{2}\mathbb{P}\left[-\sum_{j=1}^{T} \{U_j^{\mathcal{X}'}\}_{i^*} \geq \frac{Dn}{\eta LT}\mathbb{E}\left[-\sum_{j=1}^{T} \{U_j^{\mathcal{X}'}\}_{i^*}\right]\right]\\ &\geq \frac{1}{2}\left(1 - \frac{Dn}{\eta LT}\right)^2 \frac{1}{C_s}\\ &\geq \frac{1}{2}\left(1 - \frac{1}{2}\sqrt{\frac{\eta L}{2Dn}}\right)^2 \frac{1}{C_s}\\ &\geq \frac{1}{16}\frac{\eta L}{Dn}\frac{1}{C_s}.\end{aligned} \tag{e}$$

In the first equality, we simply used the fact that $\{W_0^{\mathcal{X}'}\}_{i^*} \sim \mathcal{N}(0, \sigma^2)$ and therefore $\mathbb{P}\left[\{W_0^{\mathcal{X}'}\}_{i^*} \geq 0\right] = \frac{1}{2}$. In the third inequality, we used the third property of the C_s-AC oracle where $\{\nabla\mathcal{L}(\cdot;\chi')\}_{i^*}T \geq \mathbb{E}\left[\sum_{j=1}^{T}\{U_j\}_{i^*}\right]$ and $\{\nabla\mathcal{L}(\cdot;\chi')\}_{i^*} = -\frac{L}{n}$, by our construction. In the fourth inequality, using the second property of the C_s-AC oracle, we get $-\sum_{j=1}^{T}\{U_j\}_{i^*} \geq 0$, $T \geq 1$. Hence, we apply Paley-Zygmund on $-\sum_{j=1}^{N}\{U_j\}_{i^*}$ and then use the fourth property of the C_s-AC oracle implying that $\mathbb{E}\left[\left(\sum_{j=1}^{N}\{U_j\}_{i^*}\right)^2\right] \leq C_s\mathbb{E}\left[\sum_{j=1}^{N}\{U_j\}_{i^*}\right]^2$. In the fifth inequality, we use the fact that $T \geq \left(\frac{2nD}{\eta L}\right)^{3/2}$ and in the sixth inequality, we use the fact that $\frac{1}{2}\sqrt{\frac{\eta L}{2Dn}} \leq \frac{1}{2}$. Thus, combining inequalities (d) and (e), we get that

$$\mathbb{P}\left[W_t^{\mathcal{X}'} \notin S\right] \geq \frac{1}{16}\frac{\eta L}{Dn}\frac{1}{C_s} \tag{F1}$$

Computing $\mathbb{P}\left[W_T^{\mathcal{X}} \in S\right]$ This argument follows exactly from the first case. We get that

$$\mathbb{P}\left[W_T^{\mathcal{X}} \in S\right] = \mathbb{P}\left[\{w_0\}_{i^*} \geq D\right] \leq \frac{\sigma}{\sqrt{2\pi}D}e^{-\frac{D^2}{2\sigma^2}} \tag{F1}$$

Hence, using inequalities F1 and F2, we get

$$\frac{\mathbb{P}\left[W_t^{\mathcal{X}'} \in S\right] - \frac{\eta L}{32C_sDn}}{\mathbb{P}\left[W_T^{\mathcal{X}} \in S\right]} \geq \frac{\sqrt{2\pi}\eta L}{32C_sn\sigma}e^{\frac{D^2}{2\sigma^2}}$$

Therefore,

$$\mathbb{P}\left[W_t^{\mathcal{X}'} \in S\right] \geq e^{\frac{D^2}{2\sigma^2}+\ln\left(\frac{\sqrt{2\pi}\eta L}{32C_sn\sigma}\right)}\mathbb{P}\left[W_T^{\mathcal{X}} \in S\right] + \frac{Dn}{32C_s\eta L}$$

Combining the above two cases proves our claim. □

D Discussion of Oracles

Lemma D.1. *SPSA, FD, and estimator defined in Definition B.1 are 3 AC.*

Proof. Take $i^* = 1$. Consider any $f \in \mathcal{F}$. Then, the expression of $f = \langle ge_1, w\rangle$, which means that $|\{\nabla f\}_{i^*}| = g$. Then, for $Z_{SPSA}, Z_{FD}, Z_{G_1}Z_{G_2} \sim \mathcal{N}(0, I_d)$,

$$\begin{aligned}
SPSA(f,w) &= \frac{\langle ge_1, w+\mu Z_{SPSA}\rangle - \langle ge_1, w-\mu Z_{SPSA}\rangle}{2\mu}Z_{SPSA} \\
&= g\{Z_{SPSA}\}_1 Z_{SPSA} \\
FD(f,w) &= \frac{\langle ge_1, w+\mu Z_{FD}\rangle - \langle ge_1, w\rangle}{\mu}Z_{FD} \\
&= g\{Z_{FD}\}_1 Z_{FD} \\
G_{\mu_1,\mu_2}(f,w) &= \frac{\langle ge_1, w+\mu_1 Z_{G_1}+\mu_2 Z_{G_2}\rangle - \langle ge_1, w+\mu_1 Z_{G_1}\rangle}{\mu_2}Z_{G_2} \\
&= g\{Z_{G_2}\}_1 Z_{G_2}
\end{aligned}$$

Since $Z_{SPSA}, Z_{FD}, Z_{G_2}$ are i.i.d. random variables, it implies that the given estimators follow the same distribution. Take $\mathcal{E}$ to be any one of the oracles, and for $Z \sim \mathcal{N}(0, I_d)$, we have that $\mathcal{E}(f,w) = g\{Z\}_1 Z$. Thus, we have that $\{\mathcal{E}(f,w)\}_1 = g\{Z\}_1^2$ and $\{Z\}_1^2 \sim \chi^2(1)$. $\{U^{(f)}\}_1 = gV$ where $V \sim \chi^2(1)$. Now, we verify the properties

1. Observe that when $|\{\nabla f\}_{i^*}| = g = 0$ which implies that $\mathcal{E}(f, w) = \mathbf{0}_d$, satisfying property 1.
2. For the second property, $V \geq 0$ which implies that $gV \leq 0$ for all .
3. $\mathbb{E}\left[V\right] = 1$ which implies that $\mathbb{E}\left[\{U^{(f)}\}_{i^*}\right] = g\mathbb{E}\left[V\right] = g = |\{\nabla f\}_{i^*}|$.
4. $R = \sum_{j=1}^{N} V_j \sim \chi^2(N)$. Hence, we know that $\mathbb{E}\left[R\right] = N$ and $Var[R] = 2N$, which implies that $\mathbb{E}\left[R^2\right] = N^2 + 2N$. We also have that $\sum_{j=1}^{N}\{U_j^{(f)}\}_{i^*} = g\sum_{j=1}^{N} V_j$, which implies that $\mathbb{E}\left[\left(\sum_{j=1}^{N}\{U_j^{(f)}\}_{i^*}\right)^2\right] = g^2(N^2 + 2N)$ and $\mathbb{E}\left[\left(\sum_{j=1}^{N}\{U_j^{(f)}\}_{i^*}\right)\right]^2 = g^2N^2$. Using $N \geq 1$ gives the value of $C_s = 3$ proving our given claim.

□

Lemma D.2. *(Restated) If an oracle $\mathcal{E}$ is C_s-AC then $M_m^{\mathcal{E}}$ is C_s-AC.*

Proof. If $\mathcal{E}$ is C_s-AC, then there exists an $i^* \in [N]$ which satisfies the properties mentioned in Definition 4.2. Using definition of mean extensions of estimators, we have that for $f \in \mathcal{F}$ (as defined in Definition 4.2)

$$M_m^{\mathcal{E}}(f, w) = \frac{1}{m}\sum_{j=1}^{m} U_j^{(f,w)},$$

where $U_j^{(f,w)}$ is drawn i.i.d. from $\mathcal{E}(f, w)$.

Using property 1 of $\mathcal{E}$, $|\{\nabla f\}_{i^*}| = 0$ implies $U_j^{(f,w)} = \mathbf{0}_d$ for all $j \in [m]$. Thus, we get $\{M_m^{\mathcal{E}}(f, w)\}_{i^*} = \frac{1}{m}\sum_{j=1}^{m}\{U_j^{(f,w)}\}_{i^*} < 0$, satisfying property 2.

Similarly, for property 2, $|\{\nabla f\}_{i^*}| \neq 0$ implies that $\{U_j^{(f,w)}\}_{i^*} < 0$ for all $j \in [m]$. Thus, we get $\{M_m^{\mathcal{E}}(f, w)\}_{i^*} = \frac{1}{m}\sum_{j=1}^{m}\{U_j^{(f,w)}\}_{i^*} < 0$, satisfying property 2.

For the third property, consider $\mathbb{E}\left[\{M_m^{\mathcal{E}}(f, w)\}_{i^*}\right]$. Using linearity of expectation, we get

$$\begin{aligned}\mathbb{E}\left[\frac{1}{m}\sum_{j=1}^{m}\{U_j^{(f,w)}\}_{i^*}\right] &= \frac{1}{m}\sum_{j=1}^{m}\mathbb{E}\left[\{U_j^{(f,w)}\}_{i^*}\right]\\ &\leq \{\nabla h\}_{i^*}.\end{aligned}$$

Hence, $M_m^{\mathcal{E}}$ satisfies property 3.

Consider the set $\{w_c\}_{c=1}^{N}$. By the definition of $M_m^{\mathcal{E}}$, we have

$$\mathbb{E}\left[\left(\frac{1}{m}\sum_{k=1}^{N}\sum_{j=1}^{m}\{U_j^{(f,w_k)}\}_{i^*}\right)^2\right] = \mathbb{E}\left[\left(\frac{1}{m}\sum_{j=1}^{m}\sum_{k=1}^{N}\{U_j^{(f,w_k)}\}_{i^*}\right)^2\right]$$

Let $S_j = \sum_{k=1}^{N}\{U_j^{(f,w_k)}\}_i$. Since $U_j^{(f,w_k)}$ are sampled i.i.d. from $\mathcal{E}(f, w_k)$ for all $j \in [m]$, it implies that $S_j = \sum_{k=1}^{N}\{U_j^{(f,w_k)}\}_{i^*}$ is identically distributed for all $j \in [m]$. Thus, we can take $\mathbb{E}\left[S_u\right] = K_f$, and $\mathbb{E}\left[S_u^2\right] = K_s$ for all $u \in [m]$. Applying Young's inequality, we get that

$$\begin{aligned}\mathbb{E}\left[\left(\frac{1}{m}\sum_{u=1}^{m} S_u\right)^2\right] &\leq \mathbb{E}\left[\frac{1}{m}\sum_{u=1}^{m} S_u^2\right]\\ &= K_s\end{aligned}$$

Using property 4 of $\mathcal{E}$, we have $K_s \leq C_s K_f^2$. Thus, substituting the original terms, we have

$$\mathbb{E}\left[\left(\frac{1}{m}\sum_{k=1}^{N}\sum_{j=1}^{m}\{U_j^{(f,w_k)}\}_{i^*}\right)^2\right] \leq C_s \mathbb{E}\left[\sum_{k=1}^{N}\{U_u^{(f,w_k)}\}_{i^*}\right]^2 \tag{g}$$

Thus, using the fact that $U_j^{(f,w_k)}$ is sampled i.i.d. for a $k \in [N]$, we see that $\mathbb{E}\left[\sum_{k=1}^{N}\{U_u^{(f,w_k)}\}_{i^*}\right] = \mathbb{E}\left[\frac{1}{m}\sum_{j=1}^{m}\sum_{k=1}^{N}\{U_j^{(f,w_k)}\}_{i^*}\right]$. Substituting this in the RHS of (g), we get that $M_m^{\mathcal{E}}$ satisfies the final property. □