DPZero: Dimension-Independent and Differentially Private Zeroth-Order Optimization

Liang Zhang
ETH Zurich & Max Planck Institute
liang.zhang@inf.ethz.ch

Kiran Koshy Thekumparampil
Amazon Search
kkt@amazon.com

Sewoong Oh
University of Washington
sewoong@cs.washington.edu

Niao He
ETH Zurich
niao.he@inf.ethz.ch

Abstract

The widespread practice of fine-tuning pretrained large language models (LLMs) on domain-specific data faces two major challenges in memory and privacy. First, as the size of LLMs continue to grow, encompassing billions of parameters, the memory demands of gradient-based training methods via backpropagation become prohibitively high. Second, given the tendency of LLMs to memorize and disclose sensitive training data, the privacy of fine-tuning data must be respected. To this end, we explore the potential of zeroth-order methods in differentially private optimization for fine-tuning LLMs. Zeroth-order methods, which rely solely on forward passes, substantially reduce memory consumption during training. However, directly combining them with standard differential privacy mechanism poses dimension-dependent complexity. To bridge the gap, we introduce DPZERO, a novel differentially private zeroth-order algorithm with nearly dimension-independent rates. Our theoretical analysis reveals that its complexity hinges primarily on the problem's intrinsic dimension and exhibits only a logarithmic dependence on the ambient dimension. This renders DPZERO a highly practical option for real-world LLMs deployments.

1 Introduction

Fine-tuning pretrained large language models (LLMs), including BERT [20, 66, 82], OPT [111], the LLaMA family [93, 94], and the GPT family [80, 11, 77, 76], has become a standard practice for achieving state-of-the-art performance in a wide array of downstream applications. However, two significant challenges persist in practical adoption: memory demands for gradient-based optimizers and the need to safeguard the privacy of domain-specific fine-tuning data.

As LLMs evolve from having millions to billions of parameters [11], the resource requirements for their training escalate significantly, creating barriers for entities with limited computational resources from deploying and further developing LLMs. Various approaches have been explored to tackle these limitations, ranging from parameter-efficient fine-tuning (PEFT) [55, 44] to the development of novel optimization algorithms [16, 60]. While these methods offer some advantages, they still require memory-intensive first-order gradient information through backpropagation, which is prohibitive for large models. A recent trend has emerged in developing algorithms for training neural networks without using backpropagation [10, 87, 42, 43, 79]. Specifically for LLMs, Malladi et al. [70] introduced zeroth-order methods for fine-tuning, thereby requiring solely forward passes and achieving a 12x memory reduction compared to first-order algorithms. For instance, utilizing

International Workshop on Federated Learning in the Age of Foundation Models in Conjunction with NeurIPS 2023 (FL@FM-NeurIPS'23). This workshop does not have official proceedings and this paper is non-archival.

a single A100 GPU, zeroth-order methods enable the training of LLMs with 30 billion parameters, whereas first-order methods, even equipped with PEFT, are limited to training models of up to 6.7 billion parameters. This greatly expands the potential for deploying and fine-tuning LLMs even on personal devices.

Besides the computational challenge, empricial studies [13, 68, 90] have highlighted the risk of LLMs inadvertently revealing sensitive information from their training datasets. For instance, Carlini et al. [13] exploited vulnerabilities in GPT-2 to disclose hundreds of verbatim text samples used during the model's training. Such privacy concerns are pronounced especially when users opt to fine-tune LLMs on datasets of their own. Notably, the expectation that machine learning models should not compromise the confidentiality of their contributing entities is codified into legal frameworks [97]. The most widely accepted mathematical framework for ensuring privacy in machine learning is the notion of differential privacy [24], which, with high probability, prevents attackers from identifying participating entities through model outputs [86]. Consequently, the development of methods that fine-tunes LLMs under differential privacy has attracted considerable attention [57, 107, 41, 21, 91]; however, most of these efforts focused on first-order algorithms.

Motivated by the interplay of the memory constraints and privacy concerns in fine-tuning LLMs, we investigate zeroth-order methods that guarantee differential privacy for solving the following stochastic optimization problem:

$$\min_{x\in\mathbb{R}^d} F_S(x) \;:=\; \frac{1}{n}\sum_{i=1}^{n} f(x;\xi_i)\;, \tag{1}$$

where $S=\{\xi_i\}_{i=1}^n$ is the training data, x is the model weight, the loss $f(x;\xi_i)$ is Lipschitz for each sample ξ_i, and the averaged loss $F_S(x)$ is smooth and possibly nonconvex. In theory, established studies on both differentially private optimization [7] and zeroth-order optimization [23] indicate a dependence on the dimension d in the convergence guarantees; such a dimension dependence becomes problematic in the context of LLMs with dimension d scaling to billions. In practice, and somewhat surprisingly, empirical studies on the fine-tuning of LLMs using zeroth-order methods [70] and DP first-order methods [107, 57, 56] have shown that the performance degradation due to the large model size is not significant. For example, Yu et al. [107] showed that the performance drop due to privacy is smaller for larger architectures. A 345 million-sized GPT-2-Medium, trained with $(\varepsilon=6.8,\delta=10^{-5})$-DP, suffers from a modest drop of 5.1 in BLEU score [78] (compared to a non-private model of the same size and architecture), whereas the 1.5 billion-sized GPT-2-XL suffers a drop of only 4.3 BLEU score under the same privacy budget.

This gap between theory and practice has been linked to the presence of low-rank structures in the fine-tuning of pretrained LLMs in previous studies [70, 56]. Empirical evidence [81, 40, 35] suggests that the training of deep neural networks occurs within a subspace governed by a small number of the leading eigenvectors of the loss's Hessian. Particularly in the context of fine-tuning LLMs, Aghajanyan et al. [2] demonstrated that commonly pretrained LLMs often possess a remarkably low intrinsic dimension [54] (200 for RoBERTa [66] with 355 million parameters) – a low dimensional reparameterization that retains efficacy in fine-tuning comparable to the full parameter space; Li et al. [56] revealed that when privately fine-tuning DistilRoBERTa [82, 66] with LoRA [44] (with 7 million parameters), the gradients are primarily influenced by a small number of principal components, and projecting these gradients onto a subspace with dimension 100 suffices to regain the original performance. Note that previous works report how performance changes when fine-tuning is restricted to some low-dimensional subspace to argue for inherent low-dimensional structure. This is because directly computing the rank of the Hessian matrix, for a model with hundreds of millions of parameters, is prohibitively expensive. Such empirical findings drive the investigation of dimension-independent rates under particular "low-rank" structures. Previous studies in [70] and [69, 56] have examined the dimension-dependence in zeroth-order and DP first-order optimization, respectively. Both works suggest that the dependence in the dimension can be effectively replaced by the trace of the loss's Hessian, which we refer to as the effective rank or intrinsic dimension.

Given the popularity of fine-tuning LLMs on domain-specific datasetes, we ask the following fundamental question: *Can we achieve a dimension-independent rate both under differential privacy and with access to only the zeroth-order oracle?* Our contributions are summarized in the following.

• We first show that the straightforward approach – that combines DP first-order methods with zeroth-order gradient estimations (Algorithm 1) – exhibits an undesirable dimension-dependence

in the convergence guarantees, even when the effective rank of the problem does not scale with the dimension (Thereom 3.2 and 3.4 in Section 3). There are two root causes. First, the standard practice of choosing the clipping threshold to be the maximum norm of the estimated sample gradient leads to an unnecessarily large threshold. Next, on top of this choice of threshold forcing a large noise to be added to ensure privacy, Algorithm 1 adds that noise in all d directions.

• We present DPZERO (Algorithm 2), the first nearly dimension-independent DP zeroth-order method for stochastic optimization. Its convergence guarantee depends on the effective rank of the problem (specified in Assumption 3.3) and exhibits logarithmic dependence on the dimension d (Theorem 4.1 in Section 4). This builds upon two insights. First, the direction of the estimated gradient is a public information and does not need to be private; it is sufficient to make only the magnitude of the estimated gradient private, which is a scalar value. Next, we introduce a tighter analysis that allows us to choose a significantly smaller clipping threshold, leveraging the fact that the typical norm of the estimated gradient is much smaller than its maximum.

1.1 Related Works

We build upon exciting advances in zeroth-order optimization and differentially private optimization, which we survey here. Notably, DPZERO is inspired by new empirical and theoretical findings showing that fine-tuning large language models do not suffer in high-dimensions when using zeroth-order methods in Malladi et al. [70] or using private first-order optimization in Li et al. [56].

Zeroth-Order Optimization. Nesterov and Spokoiny [74] pioneered the formal analysis of the convergence rate of zeroth-order methods, i.e., zeroth-order (stochastic) gradient descent (ZO-(S)GD) that replaces gradients in (S)GD by their zeroth-order estimations, under various assumptions on the objective function. This is motivated by renewed interest in adopting zeroth-order methods in industry due to, for example, fast differentiation techniques that require storing all intermediate computations reaching the memory limitations. Their findings on nonsmooth convex functions are later refined by Shamir [83]. Lin et al. [59] contributed to further advancements on nonsmooth nonconvex functions recently. Additionally, Ghadimi and Lan [34] extended the results for smooth functions into the stochastic setting. Zeroth-order methods have also been expanded to incorporate approaches such as coordinate descent [58], conditional gradient descent [6], variance reduction techniques [61, 26, 47], SignSGD [62], and minimax optimization [102]. Additionally, zeroth-order methods find applications in fields such as black-box machine learning [15, 17], bandit optimization [30, 83], and distributed learning [28, 108] to reduce communication overhead. These well-established results indicate that the norm of the zeroth-order gradient scales with the dimension d and the required stepsize is d-times smaller than that in first-order gradient-based methods, consequently leading to a d-times increase in the final time complexity. For example, the convergence rate of gradient descent for minimizing a smooth convex function $f(x)$ is $f(\bar{x}_T) - \min_{x\in\mathbb{R}^d} f(x) \leq \mathcal{O}(1/T)$ where $\bar{x}_T$ is the average of T iterates [73], while the zeroth-order method only achieves a rate $\mathcal{O}(d/T)$. It has been shown that such dimension-dependence of zeroth-order methods is inevitable without additional structures [103, 23].

There has been a recent effort to relax the dimension-dependence in zeroth-order methods leveraging problem structures. Wang et al. [101] and Cai et al. [12] assumed certain sparsity structure in the problem and applied sparse recovering algorithms, e.g. LASSO, to obtain sparse gradients from zeroth-order observations. Golovin et al. [36] analyzed the case when the objective function is $f(Px)$ for some low-rank projection matrix P. These works either require the objective or the algorithm to be modified to have a dimension-independent guarantee. Balasubramanian and Ghadimi [6] demonstrated that ZO-SGD can directly identify the sparsity of the problem and dimension-independent rate is possible when the support of gradients remains unchanged [12]. Recently, Malladi et al. [70] provided a relaxation from dependence on the dimension d to a dependence on the trace of the loss's Hessian for ZO-SGD when the smoothing parameter approaches 0.

Differentially Private Optimization. Previous works on differentially private optimization mostly center around first-order methods. For constrained convex problems, tight utility guarantees on both excess empirical [14, 7, 104, 109, 99] and population [8, 9, 29, 4, 53, 110] losses are well-understood. As an example, a typical result states that the optimal rate on the excess empirical loss for convex objectives is $\Theta(\sqrt{d\log(1/\delta)}/(n\varepsilon))$, where (ε, δ) are privacy parameters, n is the number of samples, and d is the dimension. The dimension-dependence is fundamental as both the upper bound [7], using

differentially private (stochastic) gradient descent (DP-(S)GD) introduced in [88], and the lower bound [7], using a reduction to finger printing codes, have the same dependence.

When the problem is nonconvex, which is the setting of our interest, DP-(S)GD achieves a rate of $\mathcal{O}(\sqrt{d\log(1/\delta)}/(n\varepsilon))$ on the squared norm of the gradient [99, 112]. We show that DPZERO matches this rate with access to only the zeroth-order oracle in Theorem 4.1 in the worst case. When having access to the first-order oracle, it has been recently shown that such rate can be improved to $\mathcal{O}((\sqrt{d\log(1/\delta)}/(n\varepsilon))^{4/3})$ leveraging momentum [95] or variance reduction techniques [3]. It remains an open question whether one can improve Theorem 4.1 analogously. Further, the convergence to second-order stationary points in nonconvex DP optimization is studied in [33]. Recent advancements in DP optimization have also delved into the understanding of the potential of public data [31, 67], the convergence properties of per-sample gradient clipping [105, 27, 52], and the relaxation of the dimension-dependence in the utility upper bound [69, 56].

Early works [46, 89] established that dimension-independent rates can be attained when the gradients lie in some fixed low-rank subspace. By first identifying this gradient subspace, dimension-independent algorithms [113, 49] can be designed. Closest to our result is Song et al. [89], which demonstrated that the rate of DP-(S)GD for smooth nonconvex optimization can be improved to $\mathcal{O}(\sqrt{r\log(1/\delta)}/(n\varepsilon))$ under certain structural assumptions, i.e., for generalized linear models (GLMs) with a rank-r feature matrix. We match this result with access to only the zeroth-order oracle in Theorem 4.1 for more general structures beyond low-rank GLMs. This is inspired by more recent results in Li et al. [56] that introduced a relaxed Lipschitz condition for the gradients and provide dimension-free bounds when the loss is convex and the relaxed Lipschitz parameters decay rapidly. Similalry, Ma et al. [69] suggested that the d dependence in the utility upper bound for DP stochastic convex optimization can be improved to a dependence on the trace of the Hessian.

The exploration of DP optimization algorithms that extend beyond first-order methods remains notably limited. Ganesh et al. [32] investigated the potential of second-order methods for DP convex optimization. Gratton et al. [37] proposed to use zeroth-order methods for DP-ADMM [45] in distributed learning. They state that the noise intrinsic in zeroth-order methods is enough to provide privacy guarantee and rely on the output of zeroth-order methods being Gaussian, which is unverified to the best of our knowledge. Liu et al. [63] proposed a private genetic algorithm based on zeroth-order optimization heuristics for private synthetic data generation. Du et al. [21] introduced a novel noise adding mechanism that happens in the forward pass of training. Although the algorithm is termed "DP-Forward", it is not a zeroth-order method and still requires backpropagation for training. There is also another line of research [38, 92, 84] on the design of differentially private algorithms for the stochastic bandit problem based on upper confidence bound [5]. Their algorithms are not directly applicable to our setting. As far as we are aware, no previous studies consider the problem of deriving dimension-independent rate in DP zeroth-order optimization.

2 Preliminaries

Notation. $\|\cdot\|$ is reserved for the Euclidean norm, and we let $\|v\|_W^2 := v^\top W v$ for some square matrix W. We use $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d \,|\, \|x\| = 1\}$ to denote the unit sphere in the d-dimensional Euclidean space, and $\eta\mathbb{S}^{d-1}$ is the sphere of radius $\eta > 0$. A function $p : \mathbb{R}^d \to \mathbb{R}$ is L-Lipschitz if $|p(x_1) - p(x_2)| \le L\|x_1 - x_2\|$ for x_1, x_2 in the domain of p. A function $q : \mathbb{R}^d \to \mathbb{R}$ is ℓ-smooth if it is differentiable and $\|\nabla q(x_1) - \nabla q(x_2)\| \le \ell\|x_1 - x_2\|$. The trace of a square matrix J is denoted by $\mathrm{Tr}(J)$. We say a symmetric real matrix $M \succeq 0$ if it is positive semi-definite, and $M \preceq 0$ if $-M$ is positive semi-definite. We say two symmetric real matrices satisfy that $M_1 \succeq M_2$ if $M_1 - M_2 \succeq 0$, and $M_1 \preceq M_2$ if $M_1 - M_2 \preceq 0$. The clipping operation is defined to be $\mathrm{clip}_C(x) = x\,\min\{1, C/\|x\|\}$ given $C > 0$. The notation $\tilde{\mathcal{O}}(\cdot)$ hides additional logarithmic terms.

2.1 Differential Privacy

Definition 1 (Differential Privacy [24, 25])**.** For two datasets $S = \{\xi_i\}_{i=1}^n$ and $S' = \{\xi_i'\}_{i=1}^n$, we say the pair (S, S') is *neighboring* if $\max\{|S \setminus S'|, |S' \setminus S|\} = 1$ and we denote neighboring datasets with $S \sim S'$. For an algorithm $\mathcal{A}$ and some privacy parameters $\varepsilon > 0$ and $\delta \in (0, 1)$, we say $\mathcal{A}$ satisfies (ε, δ)-*differential privacy* (DP) if $\mathbb{P}(\mathcal{A}(S) \in B) \le e^\varepsilon \mathbb{P}(\mathcal{A}(S') \in B) + \delta$ for all $S \sim S'$ and all measurable set B in the range of $\mathcal{A}$.

The design and analysis of DP algorithms typically rely on a thorough understanding of sensitivity and the composition of DP mechanisms.

Lemma 2.1. *(Advanced Composition [48, Theorem 4.3]) Let $\mathcal{A}$ be some randomized algorithm operating on S and outputting a vector in $\mathbb{R}^d$. If $\mathcal{A}$ has sensitivity $\Delta := \sup_{S \sim S'} \|\mathcal{A}(S) - \mathcal{A}(S')\| > 0$, the mechanism that adds Gaussian noise $\mathcal{N}(0, \sigma^2 \mathrm{I}_d)$ with variance $\sigma^2 = (2\Delta\sqrt{2T\log(e + (\varepsilon/\delta))}/\varepsilon)^2$ satisfies (ε, δ)-DP under T-fold adaptive composition for any $\varepsilon > 0$ and $\delta \in (0, 1)$.*

The original advanced composition theorem in Kairouz et al. [48] is stated for the case where the output of $\mathcal{A}$ is a scalar. Given the spherical symmetry properties of Gaussian noise, the results can be readily extended to multiple dimensions, as outlined in Lemma 1 of Kenthapadi et al. [51] where the basis can be selected in a way such that $\mathcal{A}(S)$ and $\mathcal{A}(S')$ differ in exactly one dimension.

2.2 Zeroth-Order Optimization

For the optimization problem in Eq. (1) where access to gradients is costly, we consider zeroth-order methods that solely require the evaluation of function values or directional derivatives which are cheaper to compute. Given a zeroth-order oracle that yields function evaluations at each querying point, we define the following two-point estimator [74, 23, 83] of the gradient for each sample ξ_i in the dataset S:

$$g_\lambda(x; \xi_i) \;\; := \;\; \frac{f(x + \lambda u; \xi_i) - f(x - \lambda u; \xi_i)}{2\lambda} u \;,$$

where the random vector u is sampled uniformly from the Euclidean sphere $\sqrt{d}\,\mathbb{S}^{d-1}$ and $\lambda > 0$ is the smoothing parameter [74, 106, 22]. A common approach [72, 71] to generate u uniformly from $\sqrt{d}\,\mathbb{S}^{d-1}$ is to set $u = \sqrt{d}\, z/\|z\|$, with z sampled from the standard multivariate Gaussian $\mathcal{N}(0, \mathrm{I}_d)$. We refer to $g_\lambda(x; \xi)$ as the *zeroth-order gradient* (estimator) in the sequel. Our approach naturally extends to other zeroth-order gradient estimators, e.g., any distribution of u satisfying $\mathbb{E}[uu^\top] = \mathrm{I}_d$ [23], the one-point estimators [30, 74], and the directional derivative in the limit $\lambda \to 0$ [74, 10].

3 DP-GD with Zeroth-Order Gradients suffers in High Dimensions

A straightforward application of the Gaussian mechanism to the standard zeroth-order method, detailed in Algorithm 1, suffers from dimension-dependence, even when the effective rank of the Hessian is bounded (Theorem 3.4). The standard approach to privately solve the optimization problem in Eq. (1) is to make each gradient update private and apply the composition theorem over the entire set of T checkpoints, which is known as DP-(S)GD [88, 1, 99]: $x_{t+1} \leftarrow x_t - \alpha((1/n)\sum_{i=1}^n \mathrm{clip}_C(\nabla f(x_t; \xi_i)) + z_t)$. While doing so, each sample gradient is first clipped within an ℓ_2 ball of a certain radius C to ensure finite sensitivity of $\Delta = 2C/n$ required by Lemma 2.1. The noise z_t sampled from $\mathcal{N}(0, (2\Delta\sqrt{2T\log(e + (\varepsilon/\delta))}/\varepsilon)^2 \mathrm{I}_d)$ ensures end-to-end privacy as per the lemma. This suggests a straightforward zeroth-order method that applies the same technique to the gradient estimators $g_\lambda(x_t; \xi_i)$, as outlined in Algorithm 1.

Assumption 3.1. *The function $f(x; \xi)$ is L-Lipschitz for every ξ. The average function $F_S(x)$ is ℓ-smooth for every given dataset S, and its minimum $F_S^* := \min_{x \in \mathbb{R}^d} F_S(x)$ is finite.*

We analyze Algorithm 1 under Assumption 3.1, which is common in the nonconvex DP optimization literature [99, 100, 112, 95, 3]. The privacy guarantee follows from standard DP-(S)GD analysis, and the utility guarantee on the squared gradient norm is derived from classical techniques [74] for analyzing zeroth-order methods. We provide a proof in Appendix B.

Theorem 3.2. *For any $\varepsilon > 0$ and $\delta \in (0, 1)$, Algorithm 1 is (ε, δ)-DP. Under Assumption 3.1, the output x_τ satisfies that*

$$\mathbb{E}[\|\nabla F_S(x_\tau)\|^2] \;\; \leq \;\; 16\Big((F_S(x_0) - F_S^*)\,\ell + 2L^2\Big)\frac{d\sqrt{d\log(e + (\varepsilon/\delta))}}{n\varepsilon}, \tag{2}$$

with the choice of parameters

$$\alpha = \frac{1}{4\ell d}, \quad T = \frac{n\varepsilon}{\sqrt{d\log(e + (\varepsilon/\delta))}}, \quad \lambda \leq \frac{4L}{\ell d}\left(\frac{\sqrt{d\log(e + (\varepsilon/\delta))}}{n\varepsilon}\right)^{1/2}, \quad C = Ld.$$

Algorithm 1 DP-GD with Zeroth-Order Gradients

Input: Dataset $S = \{\xi_1, \cdots, \xi_n\}$, initialization $x_0 \in \mathbb{R}^d$, number of iterations T, stepsize $\alpha > 0$, smoothing parameter $\lambda > 0$, clipping threshold $C > 0$, privacy parameters $\varepsilon > 0, \delta \in (0, 1)$.

1: **for** $t = 0, 1, \cdots, T-1$ **do**

2: Sample u_t uniformly at random from the Euclidean sphere $\sqrt{d}\,\mathbb{S}^{d-1}$.

3: **for** $i = 1, \cdots, n$ **do**

4: Compute the zeroth-order gradient estimator

$$g_\lambda(x_t; \xi_i) \;\leftarrow\; \frac{f(x_t + \lambda u_t; \xi_i) - f(x_t - \lambda u_t; \xi_i)}{2\lambda} u_t.$$

5: Sample $z_t \in \mathbb{R}^d$ randomly from the multivariate Gaussian distribution $\mathcal{N}(0, \sigma^2 \mathrm{I}_d)$ with variance $\sigma = 4C\sqrt{2T\log(e + (\varepsilon/\delta))}/(n\varepsilon)$ and update the parameter

$$x_{t+1} \;\leftarrow\; x_t - \alpha \left(\frac{1}{n} \sum_{i=1}^{n} \mathrm{clip}_C(g_\lambda(x_t; \xi_i)) + z_t \right).$$

Output: x_τ for τ sampled uniformly at random from $\{0, 1, \cdots, T-1\}$.

The total number of zeroth-order gradient computations is $nT = n^2\varepsilon/\sqrt{d\log(e + (\varepsilon/\delta))}$, which is of the order of $\mathcal{O}(n^2/\sqrt{d})$.

Remark 1. There are three sources of the dependence in d: the norm of the zeroth-order gradient estimator $\mathbb{E}[\|(1/n)\sum_{i=1}^n g_\lambda(x, \xi_i)\|^2] = \mathcal{O}(d\,\|\nabla F_S(x)\|^2)$, the clipping threshold $C = \mathcal{O}(d)$, and the norm of the privacy noise $\mathbb{E}[\|z_t\|^2] = \mathcal{O}(d\,C^2) = \mathcal{O}(d^3)$. Following the standard analysis of a one-step update in Eq. (11) gives

$$\mathbb{E}[F_S(x_{t+1})] \;\le\; \mathbb{E}[F_S(x_t)] - \frac{\alpha}{2}\big(1 - 2d\,\ell\alpha\big)\,\mathbb{E}[\|\nabla F_S(x_t)\|^2] + c\,\alpha^2\,d^3, \tag{3}$$

where we set $\lambda = 0$ to simplify the RHS and c is a constant that depends on problem parameters other than α and d. A small enough step size, $\alpha < 1/(2\ell d)$, is required to make the second term negative, where the dependence in d comes from $\mathbb{E}[\|(1/n)\sum_{i=1}^n g_\lambda(x, \xi_i)\|^2]$. The dependence in d^3 in the last term comes from $\mathbb{E}[\|z_t\|^2]$, which, after balancing error terms, gives the $\mathcal{O}(d^{3/2})$ dependence in Eq. (2). Remark 4 shows how these terms change under a low-rank structure, enabling Algorithm 1 to attain a reduced error.

Remark 2. The choice of the clipping threshold $C = Ld$ ensures that clipping does not happen with probability one, which is a common choice in the theoretical analysis of private optimization algorithms [7, 99, 8]. This follows from the fact that, for L-Lipschitz $f(x; \xi)$, the zeroth-order gradient is upper bounded by $\|g_\lambda(x; \xi)\| \le Ld$ almost surely. One of the main contributions of DPZERO is to provide a tighter analysis that allows a smaller choice by a factor of $d^{1/2}$, thus reducing the achievable error (Section 4). The topic on how to select clipping threshold without knowledge of the Lipschitz constant L still remains an important open question [18, 105, 27, 52].

Remark 3. In non-DP optimization, e.g., [74, 34], zeroth-order methods suffer from an $\mathcal{O}(d)$ factor larger error measured in squared gradient norm, compared with first-order methods. Analogously, Algorithm 1 achieves an $\mathcal{O}(d)$ factor larger error compared with the first-order gradient-based method, DP-GD [99], that achieves $\mathbb{E}[\|\nabla F_S(x_\tau)\|^2] = \mathcal{O}(\sqrt{d\log(1/\delta)}/(n\varepsilon))$ with stepsize $\mathcal{O}(1/\ell)$. This $\mathcal{O}(d)$ gap can be closed by a judicious choice of the clipping threshold and improved noise adding mechanism, as we show in the next section.

3.1 Improved Bounds under Low Effective Rank Structures

We are interested in the scenario of fine-tuning pretrained large language models on private data. It has been demonstrated that certain "low-rank" structure exists in popular fine-tuning scenarios as discussed in Section 1. Ideally, we hope that Algorithm 1 achieves an error independent of the dimension, d, when applied to an objective with a low effective rank structure below.

Assumption 3.3. *The function $f(x; \xi)$ is L-Lipschitz and ℓ-smooth for every ξ. The average function $F_S(x)$ is twice differentiable with $\nabla^2 F_S(x) \preceq H$ for any $x \in \mathbb{R}^d$, and its minimum*

$F_S^* := \min_{x\in\mathbb{R}^d} F_S(x)$ *is finite. Here, the real-valued* $d \times d$ *matrix* $H \succeq 0$ *satisfies that* $\|H\|_2 \le \ell$ *and* $\mathrm{Tr}(H) \le r\|H\|_2$. *We refer to* r *as the effective rank or the intrinsic dimension of the problem.*

We can slightly relax the smoothness condition to require that only the average loss $F_S(x)$ is ℓ-smooth and the same result in Theorem 3.4 holds. Assumption 3.3 is stated for a more strict condition that every instance of $f(x;\xi)$ is ℓ-smooth in order to be consistent with what we need for Theorem 4.1. Setting $r = d$, this recovers Assumption 3.1 as a special case, since $-H' \preceq \nabla^2 F_S(x) \preceq H', \forall\, x \in \mathbb{R}^d$, where $H' = \ell\, \mathrm{I}_d$ implies that $\|H'\|_2 \le \ell$ and $\mathrm{Tr}(H') \le d\|H'\|_2$. With $r < d$, this assumption restricts the Hessian upper bound to have additional structures. Note that our assumption on the Hessian is less restrictive than a strict low-rank assumption even if $r \ll d$. For example, the assumption is satisfied with $r = \mathcal{O}(\log d) \ll d$ in the case of a full-rank matrix H, with its i-th largest eigenvalue being ℓ/i for $1 \le i \le d$.

Similar assumptions have been employed to successfully relax the dimension-dependence in zeroth-order optimization in the limit $\lambda \to 0$ [70] and also for DP first-order optimization when the objective is smooth and convex [69]. A basic understanding of why the assumption works can be found in (iii) and (iv) of Lemma A.1 in the appendix: the d-dependence in both the squared norm of zeroth-order gradient, $\mathbb{E}[\|(1/n)\sum_{i=1}^n g_\lambda(x_t;\xi_i)\|^2]$, and the DP noise, $\mathbb{E}[\|z_t\|^2]$, can be refined to a dependence on $\mathrm{Tr}(H)$. However, even equipped with Assumption 3.3, the DP-GD with zeroth-order gradients (Algorithm 1) still suffers from a dependence in the dimension d, as presented in the theorem below, with a proof shown in Appendix B. This recovers Theorem 3.2 in the worst-case when $r = d$, and with smaller effective rank, r, achieves an improved error on the squared gradient norm.

Theorem 3.4. *For any* $\varepsilon > 0$ *and* $\delta \in (0,1)$, *Algorithm 1 is* (ε,δ)*-DP. Under Assumption 3.3, the output* x_τ *satisfies that*

$$\mathbb{E}[\|\nabla F_S(x_\tau)\|^2] \;\le\; 16\Big((F_S(x_0) - F_S^*)\,\ell + 2L^2\Big)\frac{d\sqrt{r\log(e + (\varepsilon/\delta))}}{n\varepsilon}, \tag{4}$$

with the choice of parameters

$$\alpha = \frac{1}{4\ell(r+2)}, \quad T = \frac{n(r+2)\varepsilon}{d\sqrt{r\log(e+(\varepsilon/\delta))}}, \quad \lambda \le \frac{4L}{\ell d}\left(\frac{\sqrt{r\log(e+(\varepsilon/\delta))}}{n\varepsilon}\right)^{1/2}, \quad C = Ld.$$

The total number of zeroth-order gradient computations is $nT = n^2(r+2)\varepsilon/(d\sqrt{r\log(e+(\varepsilon/\delta))})$, *which is of the order of* $\mathcal{O}(n^2\sqrt{r}/d)$.

Remark 4. Under the low effective rank scenario, Eq. (3) becomes

$$\mathbb{E}[F_S(x_{t+1})] \;\le\; \mathbb{E}[F_S(x_t)] - \frac{\alpha}{2}\big(1 - 2(r+2)\ell\alpha\big)\,\mathbb{E}[\|\nabla F_S(x_t)\|^2] + c\,\alpha^2\, r\, d^2, \tag{5}$$

where the smaller $2(r+2)\ell\alpha$ factor comes from a fine-grained analysis (Eq. (14) in appendix) which reveals a dependence on $\mathbb{E}[\|(1/n)\sum_{i=1}^n g_\lambda(x,\xi_i)\|_H^2] = \mathcal{O}(r/d)\,\mathbb{E}[(1/n)\sum_{i=1}^n \|g_\lambda(x,\xi_i)\|^2]$ (recall $\|v\|_H^2 := v^\top H v$), thus replacing the $\mathcal{O}(d)$ term in Eq. (3) with the $\mathcal{O}(r)$ term above. Similar fine-grained analysis for the third term reveals a dependence on $\mathbb{E}[\|z_t\|_H^2] = \mathcal{O}(r/d)\,\mathbb{E}[\|z_t\|^2]$, which replaces an $\mathcal{O}(d)$ factor in Eq. (3) with an $\mathcal{O}(r)$ factor above. However, the third term still has $\mathcal{O}(d^2)$ dependence due to the clipping threshold, $C = \mathcal{O}(d)$. Consequently, even when the effective rank, r, is bounded, the error in Eq. (4) still grows linearly in d.

4 Nearly Dimension-Independent DP Zeroth-Order Optimization

Algorithm 1 suffers from dependence in the dimension d. In this section, we introduce a novel approach, DPZERO, as detailed in Algorithm 2, and prove that it is nearly dimension-independent when the effective rank is small. This improved rate is rooted in two key insights: scalar privacy noise and a tighter clipping threshold.

First, since the direction of the update, u_t, is a public knowledge, we only need to make the "magnitude" of our update private. This can be achieved by clipping the finite-difference, $(f(x_t + \lambda u_t;\xi_i) - f(x_t - \lambda u_t;\xi_i))/(2\lambda)$, and adding a scalar noise z_t as shown in Algorithm 2. The privacy noise is only added in one direction, dramatically improving upon the guarantees of Algorithm 1 in Eq. (4) by a factor of $d^{1/2}$.

Next, another factor of $d^{1/2}$ improvement can be achieved by tightening the clipping threshold. In the worst-case, the finite-difference can be as large as $Ld^{1/2}$ (which corresponds to the worst-case clipping threshold of Ld for $\|g_\lambda(x_t, \xi_i)\|$ in Algorithm 1). However, this happens with an exponentially small probability over the randomness of u_t. As proved in Eq. (15) in Appendix C, the typical size of the finite-difference is

$$\frac{|f(x_t + \lambda u_t; \xi_i) - f(x_t - \lambda u_t; \xi_i)|}{2\lambda} \;\leq\; |u_t^\top \nabla f(x_t; \xi_i)| + \frac{\ell}{2}\lambda d,$$

where we use the assumption that each $f(x; \xi)$ is ℓ-smooth. Note that such individual smoothness assumption is common in the analysis of zeroth-order methods [74, 34]. When u_t is sampled from the sphere $\sqrt{d}\,\mathbb{S}^{d-1}$, a standard tail bound (part (ii) of Lemma A.1 in the appendix) implies that

$$\mathbb{P}\left(\left|u_t^\top \nabla f(x_t; \xi_i)\right| \geq C\right) \leq 2\sqrt{2\pi}\,\exp\left(-\frac{C^2}{8L^2}\right).$$

By selecting the smoothing parameter λ to be sufficiently small, a careful choice of $C = \tilde{\mathcal{O}}(L)$, nearly independent of d, can ensure a high probability that clipping does not occur. This is significantly smaller than the worst-case clipping threshold of $Ld^{1/2}$. The main technical challenge is that we need to analyze the algorithm given the event that clipping does not happen. The choice of drawing u_t from the uniform distribution over the sphere, together with corresponding tail bounds in Appendix A, allows us to prove the following nearly dimension-independent bound under the low effective rank structure in Assumption 3.3. A proof is provided in Appendix C.

Algorithm 2 DPZERO

Input: Dataset $S = \{\xi_1, \cdots, \xi_n\}$, initialization $x_0 \in \mathbb{R}^d$, number of iterations T, stepsize $\alpha > 0$, smoothing parameter $\lambda > 0$, clipping threshold $C > 0$, privacy parameters $\varepsilon > 0, \delta \in (0, 1)$.

1: **for** $t = 0, 1, \cdots, T-1$ **do**
2: Sample u_t uniformly at random from the Euclidean sphere $\sqrt{d}\,\mathbb{S}^{d-1}$.
3: **for** $i = 1, \cdots, n$ **do**
4: Compute the finite difference $f(x_t + \lambda u_t; \xi_i) - f(x_t - \lambda u_t; \xi_i)$.
5: Sample $z_t \in \mathbb{R}$ randomly from the univariate Gaussian distribution $\mathcal{N}(0, \sigma^2)$ with variance $\sigma = 4C\sqrt{2T\log(e + (\varepsilon/\delta))}/(n\varepsilon)$ and update the parameter

$$x_{t+1} \;\leftarrow\; x_t - \alpha\left(\frac{1}{n}\sum_{i=1}^{n} \mathrm{clip}_C\left(\frac{f(x_t + \lambda u_t; \xi_i) - f(x_t - \lambda u_t; \xi_i)}{2\lambda}\right) + z_t\right) u_t.$$

Output: x_τ for τ sampled uniformly at random from $\{0, 1, \cdots, T-1\}$.

Theorem 4.1. *For any $\varepsilon > 0$ and $\delta \in (0, 1)$, Algorithm 2 is (ε, δ)-DP. Under Assumption 3.3, its output x_τ satisfies that*

$$\mathbb{E}[\|\nabla F_S(x_\tau)\|^2] \;\leq\; \left(64\Big((F_S(x_0) - F_S^*)\,\ell + \tilde{L}^2\Big) + 2L^2\right)\frac{\sqrt{r\log(e + (\varepsilon/\delta))}}{n\varepsilon}, \tag{6}$$

where we define

$$\tilde{L}^2 = L^2\,\log\left(\frac{2\sqrt{2\pi}\,d(r+2)n^3\varepsilon^2}{r\log(e + (\varepsilon/\delta))}\right),$$

and choose the parameters to be

$$\alpha = \frac{1}{4\ell(r+2)}, \quad T = \frac{n(r+2)\varepsilon}{4\sqrt{r\log(e + (\varepsilon/\delta))}}, \quad C = 4\tilde{L},$$
$$\lambda \leq \frac{1}{\ell d}\min\left\{4(2-\sqrt{2})\tilde{L},\; \frac{L}{\sqrt{d}}\left(\frac{\sqrt{r\log(e + (\varepsilon/\delta))}}{n\varepsilon}\right)^{1/2}\right\}.$$

The total number of zeroth-order gradient computations is $nT = n^2(r+2)\varepsilon/(4\sqrt{r\log(e + (\varepsilon/\delta))})$, which is of the order of $\mathcal{O}(n^2\sqrt{r})$.

Remark 5. Algorithm 2 is nearly dimension-independent, given that its dependence on d in the final utility guarantee on the squared gradient norm is solely logarithmic. To the best of our knowledge, this is the first zeroth-order DP method that is nearly dimension-independent, a feature significantly beneficial for fine-tuning pretrained LLMs where $d \gg r$ [2, 56]. When $r = d$, the error rate $\tilde{\mathcal{O}}(\sqrt{d\log(1/\delta)}/(n\varepsilon))$ in Eq. (6) nearly matches that of the best known achievable bound of the first-order gradient-based method DP-(S)GD [99] for smooth nonconvex losses. When the effective rank, r, is smaller, this algorithm achieves $\tilde{\mathcal{O}}(\sqrt{r\log(1/\delta)}/(n\varepsilon))$ squared gradient norm. Similar dimension-free rate is established for DP-(S)GD on smooth nonconvex unconstrained generalized linear losses [89], with a dependence on the rank of the feature matrix. For general private nonconvex optimization without any low-rank assumptions, there exist first-order methods that achieve an improved rate $\mathcal{O}((\sqrt{d\log(1/\delta)}/(n\varepsilon))^{4/3})$ leveraging momentum [95] or variance reduction techniques [3]. It remains an interesting open question to study their zeroth-order extensions and whether these methods can achieve improved and dimension-independent bounds under Assumption 3.3.

Remark 6. The RHS of Eq. (6) improves upon Eq. (4) of Algorithm 1 by a factor of d. Simplifying our analysis in Eq. (21), and conditioning on the event that the clipping does not happen, we get a similar one-step update analysis as Eq. (5) (see Eq. (21) for a precise inequality). However, since we now have reduced the clipping threshold by a factor of $d^{1/2}$ and the privacy noise z_t is scalar, we have $\mathbb{E}[\|z_t u_t\|_H^2] = \tilde{\mathcal{O}}(r)$ is nearly independent of the dimension d, and thus the final error scales as $\tilde{\mathcal{O}}(r^{1/2})$.

Remark 7. The strategy of appropriately selecting the clipping threshold to ensure that clipping occurs with low probability is commonly applied in the analysis of private algorithms [27, 85]. Adaptive choices of clipping thresholds can provably improve error rates for certain problems including PCA [64] and linear regression [65]. One technical challenge in the choice of the clipping threshold in DPZERO is that we need the *expected* one-step progress to be sufficient in Eq. (21). This requires controlling the progress in the low-probability event that finite difference is clipped. The fact that $\|u_t\|$ is finite with probability one simplifies the analysis, which is the reason we choose to sample u_t uniformly at random over the sphere. We believe that the analysis extends to the commonly used spherical Gaussian random vectors with more work, which we leave as a future research direction.

Remark 8. One byproduct of our design is that the clipping is significantly more efficient. For each sample, we only need to clip the difference of function values, which is scalar. This is aligned with recent trends in [41, 57] of innovating the clipping operation to make DP-(S)GD efficient, as clipping is increasingly becoming the major bottleneck.

5 Conclusion

Motivated by the memory restricitons and privacy requirements in fine-tuning LLMs, we design novel zeroth-order methods with differentially private guarantees. Worst-case analyses of both DP optimization and zeroth-order optimization are known to suffer from dependence on the dimension, which is prohibitive for fine-tuning LLMs. However, it has been empirically observed that fine-tuning LLMs exhibit certain low-rank structures, under which zeroth-order optimization [70] and DP optimization [56] can be shown to respectively overcome the dependence on the dimension. Inspired by these advances, we present DPZERO, a differentially private and nearly dimension-independent zeroth-order algorithm. To the best of our knowledge, this is the first zeroth-order and DP approach whose error rate depends primarily on the problem's effective rank.

DPZERO uses the full batch gradient every iteration, and the anlaysis guarantees an upper bound on the the empirical average gradient assuming smooth nonconvex objectives. We defer extensions to the stochastic mini-batch setting, guarantees on the population loss leveraging the stability of zeroth-order methods [75], and considerations of other assumptions on the objective functions like convexity, PŁ inequality [50], and nonsmoothness to future research. We believe this work also opens up a plethora of other prospective directions in DP zeroth-order optimization. This includes, but is not limited to, understanding the advantages of the intrinsic noise in zeroth-order gradient estimators for DP optimization, discovering other structural assumptions like the restricted Lipschitz condition [56] for dimension-independent rates, and utilizing momentum [95] or variance reduction [3] techniques for an improved rate and computational complexity.

Acknowledgments and Disclosure of Funding

L.Z. gratefully acknowledges funding by the Max Planck ETH Center for Learning Systems (CLS). This work does not relate to the current position of K.T. at Amazon. N.H. is supported by ETH research grant funded through ETH Zurich Foundations and Swiss National Science Foundation Project Funding No. 200021-207343. S.O. is supported in part by the National Science Foundation under grant no. 2019844, 2112471, and 2229876 supported in part by funds provided by the National Science Foundation, by the Department of Homeland Security, and by IBM. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or its federal agency and industry partners.

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318, 2016.

[2] Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, pages 7319–7328, 2021.

[3] Raman Arora, Raef Bassily, Tomás González, Cristóbal A Guzmán, Michael Menart, and Enayat Ullah. Faster rates of convergence to stationary points in differentially private optimization. In *International Conference on Machine Learning*, pages 1060–1092. PMLR, 2023.

[4] Hilal Asi, Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: Optimal rates in l1 geometry. In *International Conference on Machine Learning*, pages 393–403. PMLR, 2021.

[5] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47:235–256, 2002.

[6] 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.

[7] Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *IEEE Annual Symposium on Foundations of Computer Science*, pages 464–473. IEEE, 2014.

[8] Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Guha Thakurta. Private stochastic convex optimization with optimal rates. *Advances in Neural Information Processing Systems*, 32, 2019.

[9] Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. *Advances in Neural Information Processing Systems*, 33, 2020.

[10] Atılım Güneş Baydin, Barak A Pearlmutter, Don Syme, Frank Wood, and Philip Torr. Gradients without backpropagation. *arXiv preprint arXiv:2202.08587*, 2022.

[11] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

[12] HanQin Cai, Daniel Mckenzie, Wotao Yin, and Zhenliang Zhang. Zeroth-order regularized optimization (ZORO): Approximately sparse gradients and adaptive sampling. *SIAM Journal on Optimization*, 32(2):687–714, 2022.

[13] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *USENIX Security Symposium*, pages 2633–2650, 2021.

[14] Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3), 2011.

[15] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. ZOO: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the ACM Workshop on Artificial Intelligence and Security*, pages 15–26, 2017.

[16] Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Yao Liu, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, et al. Symbolic discovery of optimization algorithms. *arXiv preprint arXiv:2302.06675*, 2023.

[17] 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.

[18] Xiangyi Chen, Steven Z Wu, and Mingyi Hong. Understanding gradient clipping in private SGD: A geometric perspective. *Advances in Neural Information Processing Systems*, 33: 13773–13782, 2020.

[19] Harald Cramér. *Mathematical methods of statistics*, volume 43. Princeton University Press, 1999.

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the North American Association for Computational Linguistics*, pages 4171–4186, 2019.

[21] Minxin Du, Xiang Yue, Sherman SM Chow, Tianhao Wang, Chenyu Huang, and Huan Sun. DP-Forward: Fine-tuning and inference on language models with differential privacy in forward pass. *arXiv preprint arXiv:2309.06746*, 2023.

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

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

[24] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, pages 265–284. Springer, 2006.

[25] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[26] Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. SPIDER: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. *Advances in Neural Information Processing Systems*, 31, 2018.

[27] Huang Fang, Xiaoyun Li, Chenglin Fan, and Ping Li. Improved convergence of differential private SGD with gradient clipping. In *International Conference on Learning Representations*, 2023.

[28] Wenzhi Fang, Ziyi Yu, Yuning Jiang, Yuanming Shi, Colin N Jones, and Yong Zhou. Communication-efficient stochastic zeroth-order optimization for federated learning. *IEEE Transactions on Signal Processing*, 70:5058–5073, 2022.

[29] Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: optimal rates in linear time. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 439–449, 2020.

[30] Abraham D Flaxman, Adam Tauman Kalai, and H Brendan McMahan. Online convex optimization in the bandit setting: Gradient descent without a gradient. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, pages 385–394, 2005.

[31] Arun Ganesh, Mahdi Haghifam, Milad Nasr, Sewoong Oh, Thomas Steinke, Om Thakkar, Abhradeep Guha Thakurta, and Lun Wang. Why is public pretraining necessary for private model training? In *International Conference on Machine Learning*, pages 10611–10627. PMLR, 2023.

[32] Arun Ganesh, Mahdi Haghifam, Thomas Steinke, and Abhradeep Thakurta. Faster differentially private convex optimization via second-order methods. *arXiv preprint arXiv:2305.13209*, 2023.

[33] Arun Ganesh, Daogao Liu, Sewoong Oh, and Abhradeep Thakurta. Private (stochastic) nonconvex optimization revisited: Second-order stationary points and excess risks. *arXiv preprint arXiv:2302.09699*, 2023.

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

[35] Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via Hessian eigenvalue density. In *International Conference on Machine Learning*, pages 2232–2241. PMLR, 2019.

[36] Daniel Golovin, John Karro, Greg Kochanski, Chansoo Lee, Xingyou Song, and Qiuyi Zhang. Gradientless descent: High-dimensional zeroth-order optimization. In *International Conference on Learning Representations*, 2020.

[37] Cristiano Gratton, Naveen KD Venkategowda, Reza Arablouei, and Stefan Werner. Privacy-preserved distributed learning with zeroth-order optimization. *IEEE Transactions on Information Forensics and Security*, 17:265–279, 2021.

[38] Abhradeep Guha Thakurta and Adam Smith. (Nearly) optimal algorithms for private online learning in full-information and bandit settings. *Advances in Neural Information Processing Systems*, 26, 2013.

[39] Arjun K Gupta and Saralees Nadarajah. *Handbook of Beta distribution and its applications*. CRC Press, 2004.

[40] Guy Gur-Ari, Daniel A Roberts, and Ethan Dyer. Gradient descent happens in a tiny subspace. *arXiv preprint arXiv:1812.04754*, 2018.

[41] Jiyan He, Xuechen Li, Da Yu, Huishuai Zhang, Janardhan Kulkarni, Yin Tat Lee, Arturs Backurs, Nenghai Yu, and Jiang Bian. Exploring the limits of differentially private deep learning with group-wise clipping. In *International Conference on Learning Representations*, 2023.

[42] Geoffrey Hinton. The forward-forward algorithm: Some preliminary investigations. *arXiv preprint arXiv:2212.13345*, 2022.

[43] Bairu Hou, Joe O'connor, Jacob Andreas, Shiyu Chang, and Yang Zhang. Promptboosting: Black-box text classification with ten forward passes. In *International Conference on Machine Learning*, pages 13309–13324. PMLR, 2023.

[44] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[45] Zonghao Huang, Rui Hu, Yuanxiong Guo, Eric Chan-Tin, and Yanmin Gong. DP-ADMM: ADMM-based distributed learning with differential privacy. *IEEE Transactions on Information Forensics and Security*, 15:1002–1012, 2019.

[46] Prateek Jain and Abhradeep Guha Thakurta. (Near) dimension independent risk bounds for differentially private learning. In *International Conference on Machine Learning*, pages 476–484. PMLR, 2014.

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

[48] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. In *International Conference on Machine Learning*, pages 1376–1385. PMLR, 2015.

[49] Peter Kairouz, Monica Ribero Diaz, Keith Rush, and Abhradeep Thakurta. (Nearly) dimension independent private ERM with adagrad rates via publicly estimated subspaces. In *Conference on Learning Theory*, pages 2717–2746. PMLR, 2021.

[50] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In *European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811, 2016.

[51] Krishnaram Kenthapadi, Aleksandra Korolova, Ilya Mironov, and Nina Mishra. Privacy via the Johnson-Lindenstrauss transform. *Journal of Privacy and Confidentiality*, 5(1):39–71, 2013.

[52] Anastasia Koloskova, Hadrien Hendrikx, and Sebastian U Stich. Revisiting gradient clipping: Stochastic bias and tight convergence guarantees. In *International Conference on Machine Learning*, 2023.

[53] Janardhan Kulkarni, Yin Tat Lee, and Daogao Liu. Private non-smooth erm and sco in subquadratic steps. *Advances in Neural Information Processing Systems*, 34, 2021.

[54] Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. In *International Conference on Learning Representations*, 2018.

[55] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the Association for Computational Linguistics and International Joint Conference on Natural Language Processing*, pages 4582–4597, 2021.

[56] Xuechen Li, Daogao Liu, Tatsunori B Hashimoto, Huseyin A Inan, Janardhan Kulkarni, Yin-Tat Lee, and Abhradeep Guha Thakurta. When does differentially private learning not suffer in high dimensions? *Advances in Neural Information Processing Systems*, 35:28616–28630, 2022.

[57] Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners. In *International Conference on Learning Representations*, 2022.

[58] Xiangru Lian, Huan Zhang, Cho-Jui Hsieh, Yijun Huang, and Ji Liu. A comprehensive linear speedup analysis for asynchronous stochastic parallel optimization from zeroth-order to first-order. *Advances in Neural Information Processing Systems*, 29, 2016.

[59] Tianyi Lin, Zeyu Zheng, and Michael Jordan. Gradient-free methods for deterministic and stochastic nonsmooth nonconvex optimization. *Advances in Neural Information Processing Systems*, 35:26160–26175, 2022.

[60] Hong Liu, Zhiyuan Li, David Hall, Percy Liang, and Tengyu Ma. Sophia: A scalable stochastic second-order optimizer for language model pre-training. *arXiv preprint arXiv:2305.14342*, 2023.

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

[62] Sijia Liu, Pin-Yu Chen, Xiangyi Chen, and Mingyi Hong. SignSGD via zeroth-order oracle. In *International Conference on Learning Representations*, 2019.

[63] Terrance Liu, Jingwu Tang, Giuseppe Vietri, and Steven Wu. Generating private synthetic data with genetic algorithms. In *International Conference on Machine Learning*, pages 22009–22027. PMLR, 2023.

[64] Xiyang Liu, Weihao Kong, Prateek Jain, and Sewoong Oh. DP-PCA: Statistically optimal and differentially private PCA. *Advances in Neural Information Processing Systems*, 35: 29929–29943, 2022.

[65] Xiyang Liu, Prateek Jain, Weihao Kong, Sewoong Oh, and Arun Sai Suggala. Near optimal private and robust linear regression. *arXiv preprint arXiv:2301.13273*, 2023.

[66] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[67] Andrew Lowy, Zeman Li, Tianjian Huang, and Meisam Razaviyayn. Optimal differentially private learning with public data. *arXiv preprint arXiv:2306.15056*, 2023.

[68] Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. Analyzing leakage of personally identifiable information in language models. *arXiv preprint arXiv:2302.00539*, 2023.

[69] Yi-An Ma, Teodor Vanislavov Marinov, and Tong Zhang. Dimension independent generalization of DP-SGD for overparameterized smooth convex optimization. *arXiv preprint arXiv:2206.01836*, 2022.

[70] Sadhika Malladi, Tianyu Gao, Eshaan Nichani, Alex Damian, Jason D Lee, Danqi Chen, and Sanjeev Arora. Fine-tuning language models with just forward passes. *arXiv preprint arXiv:2305.17333*, 2023.

[71] George Marsaglia. Choosing a point from the surface of a sphere. *The Annals of Mathematical Statistics*, 43(2):645–646, 1972.

[72] Mervin E Muller. A note on a method for generating points uniformly on n-dimensional spheres. *Communications of the ACM*, 2(4):19–20, 1959.

[73] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

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

[75] Konstantinos Nikolakakis, Farzin Haddadpour, Dionysis Kalogerias, and Amin Karbasi. Black-box generalization: Stability of zeroth-order learning. *Advances in Neural Information Processing Systems*, 35:31525–31541, 2022.

[76] OpenAI. GPT-4 Technical Report, 2023.

[77] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[78] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[79] Jason Phang, Yi Mao, Pengcheng He, and Weizhu Chen. Hypertuning: Toward adapting large language models without back-propagation. In *International Conference on Machine Learning*, pages 27854–27875. PMLR, 2023.

[80] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. *OpenAI*, 2018.

[81] Levent Sagun, Utku Evci, V Ugur Guney, Yann Dauphin, and Leon Bottou. Empirical analysis of the Hessian of over-parametrized neural networks. *arXiv preprint arXiv:1706.04454*, 2017.

[82] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[83] 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.

[84] Roshan Shariff and Or Sheffet. Differentially private contextual linear bandits. *Advances in Neural Information Processing Systems*, 31, 2018.

[85] Zebang Shen, Jiayuan Ye, Anmin Kang, Hamed Hassani, and Reza Shokri. Share your representation only: Guaranteed improvement of the privacy-utility tradeoff in federated learning. In *International Conference on Learning Representations*, 2023.

[86] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy*, pages 3–18. IEEE, 2017.

[87] David Silver, Anirudh Goyal, Ivo Danihelka, Matteo Hessel, and Hado van Hasselt. Learning by directional gradient descent. In *International Conference on Learning Representations*, 2022.

[88] Shuang Song, Kamalika Chaudhuri, and Anand D Sarwate. Stochastic gradient descent with differentially private updates. In *IEEE Global Conference on Signal and Information Processing*, pages 245–248. IEEE, 2013.

[89] Shuang Song, Thomas Steinke, Om Thakkar, and Abhradeep Thakurta. Evading the curse of dimensionality in unconstrained private GLMs. In *International Conference on Artificial Intelligence and Statistics*, pages 2638–2646. PMLR, 2021.

[90] Albert Yu Sun, Eliott Zemour, Arushi Saxena, Udith Vaidyanathan, Eric Lin, Christian Lau, and Vaikkunth Mugunthan. Does fine-tuning GPT-3 with the OpenAI API leak personally-identifiable information? *arXiv preprint arXiv:2307.16382*, 2023.

[91] Xinyu Tang, Richard Shin, Huseyin A Inan, Andre Manoel, Fatemehsadat Mireshghallah, Zinan Lin, Sivakanth Gopi, Janardhan Kulkarni, and Robert Sim. Privacy-preserving in-context learning with differentially private few-shot generation. *arXiv preprint arXiv:2309.11765*, 2023.

[92] Aristide Tossou and Christos Dimitrakakis. Algorithms for differentially private multi-armed bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

[93] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[94] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. LLAMA 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[95] Hoang Tran and Ashok Cutkosky. Momentum aggregation for private non-convex ERM. *Advances in Neural Information Processing Systems*, 35:10996–11008, 2022.

[96] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

[97] Paul Voigt and Axel Von dem Bussche. The EU general data protection regulation (GDPR). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676):10–5555, 2017.

[98] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[99] Di Wang, Minwei Ye, and Jinhui Xu. Differentially private empirical risk minimization revisited: Faster and more general. *Advances in Neural Information Processing Systems*, 30, 2017.

[100] Di Wang, Changyou Chen, and Jinhui Xu. Differentially private empirical risk minimization with non-convex loss functions. In *International Conference on Machine Learning*, pages 6526–6535. PMLR, 2019.

[101] Yining Wang, Simon Du, Sivaraman Balakrishnan, and Aarti Singh. Stochastic zeroth-order optimization in high dimensions. In *International Conference on Artificial Intelligence and Statistics*, pages 1356–1365. PMLR, 2018.

[102] Zhongruo Wang, Krishnakumar Balasubramanian, Shiqian Ma, and Meisam Razaviyayn. Zeroth-order algorithms for nonconvex–strongly-concave minimax problems with improved complexities. *Journal of Global Optimization*, pages 1–32, 2022.

[103] Andre Wibisono, Martin J Wainwright, Michael Jordan, and John C Duchi. Finite sample convergence rates of zero-order stochastic optimization methods. *Advances in Neural Information Processing Systems*, 25, 2012.

[104] Xi Wu, Fengan Li, Arun Kumar, Kamalika Chaudhuri, Somesh Jha, and Jeffrey Naughton. Bolt-on differential privacy for scalable stochastic gradient descent-based analytics. In *Proceedings of ACM International Conference on Management of Data*, pages 1307–1322, 2017.

[105] Xiaodong Yang, Huishuai Zhang, Wei Chen, and Tie-Yan Liu. Normalized/Clipped SGD with perturbation for differentially private non-convex optimization. *arXiv preprint arXiv:2206.13033*, 2022.

[106] Farzad Yousefian, Angelia Nedić, and Uday V Shanbhag. On stochastic gradient and subgradient methods with adaptive steplength sequences. *Automatica*, 48(1):56–67, 2012.

[107] Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, Sergey Yekhanin, and Huishuai Zhang. Differentially private fine-tuning of language models. In *International Conference on Learning Representations*, 2022.

[108] Eric Zelikman, Qian Huang, Percy Liang, Nick Haber, and Noah D Goodman. Just one byte (per gradient): A note on low-bandwidth decentralized language model finetuning using shared randomness. *arXiv preprint arXiv:2306.10015*, 2023.

[109] Jiaqi Zhang, Kai Zheng, Wenlong Mou, and Liwei Wang. Efficient private erm for smooth objectives. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 3922–3928, 2017.

[110] Liang Zhang, Kiran K Thekumparampil, Sewoong Oh, and Niao He. Bring your own algorithm for optimal differentially private stochastic minimax optimization. *Advances in Neural Information Processing Systems*, 35:35174–35187, 2022.

[111] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

[112] Yingxue Zhou, Xiangyi Chen, Mingyi Hong, Zhiwei Steven Wu, and Arindam Banerjee. Private stochastic non-convex optimization: Adaptive algorithms and tighter generalization bounds. *arXiv preprint arXiv:2006.13501*, 2020.

[113] Yingxue Zhou, Steven Wu, and Arindam Banerjee. Bypassing the ambient dimension: Private SGD with gradient subspace identification. In *International Conference on Learning Representations*, 2021.

A Technical Lemmas

Lemma A.1. *Let u be uniformly sampled from the Euclidean sphere $\sqrt{d}\cdot\mathbb{S}^{d-1}$, $a\in\mathbb{R}^d$ be some fixed vector independent of u, and $H\in\mathbb{R}^{d\times d}$ be some fixed matrix independent of u. We have that*

(i) $\mathbb{E}[u]=0$ *and* $\mathbb{E}[uu^\top]=\mathrm{I}_d$.

(ii) $\mathbb{E}_u[u^\top a]=0$, $\mathbb{E}_u[(u^\top a)^2]=\|a\|^2$ *and* $\forall\, C\geq 0$,

$$\mathbb{P}(|u^\top a|\geq C)\leq 2\sqrt{2\pi}\exp\left(-\frac{C^2}{8\|a\|^2}\right).$$

(iii) $\mathbb{E}_u[(u^\top a)u]=a$ *and*

$$\mathbb{E}_u[(u^\top a)^2\|u\|^2]=d\|a\|^2,$$
$$\mathbb{E}_u[(u^\top a)^2uu^\top]=\frac{d}{d+2}\left(2aa^\top+\|a\|^2\mathrm{I}_d\right).$$

(iv) $\mathbb{E}_u[u^\top Hu]=\mathrm{Tr}(H)$ *and*

$$\mathbb{E}_u[(u^\top a)^2u^\top Hu]=\frac{d}{d+2}\left(2a^\top Ha+\|a\|^2\,\mathrm{Tr}(H)\right).$$

Proof. (i) is a standard result, e.g., in Duchi et al. [23], and follows by the symmetry of the sphere. For any $u\in\sqrt{d}\cdot\mathbb{S}^{d-1}$, it must be the case that $-u\in\sqrt{d}\cdot\mathbb{S}^{d-1}$ as well, which suggests that $\mathbb{E}[u]=0$. Since $\mathbb{E}[\sum_{i=1}^d u_i^2]=\mathbb{E}\|u\|^2=d$, we immediately have that $\mathbb{E}[u_i^2]=1$ for every i by symmetry. Then for the off-diagonal terms, since for any $u=(u_1,\cdots,u_i,\cdots,u_j,\cdots,u_d)\in\sqrt{d}\cdot\mathbb{S}^{d-1}$, it must be the case that $(u_1,\cdots,u_i,\cdots,-u_j,\cdots,u_d)\in\sqrt{d}\cdot\mathbb{S}^{d-1}$ as well, which suggests that $\mathbb{E}[u_iu_j]=0$ when $i\neq j$. As a result, we can conclude that the matrix $\mathbb{E}[uu^\top]=\mathrm{I}_d$.

We then show (ii). Applying (i), we have that $\mathbb{E}_u[u^\top a]=0$, and that

$$\begin{aligned}\mathbb{E}_u[(u^\top a)^2]&=\sum_{i=1}^d a_i^2\mathbb{E}[u_i^2]+\sum_{i\neq j}a_ia_j\mathbb{E}[u_iu_j]\\&=\|a\|^2.\end{aligned}$$

The tail bound follows from Example 3.12 in Wainwright [98], where they show that for any function $h:\mathbb{S}^{d-1}\to\mathbb{R}$ such that $\forall x,y\in\mathbb{S}^{d-1}$,

$$|h(x)-h(y)|\leq\arccos(x^\top y),$$

when x is uniformly sampled from $\mathbb{S}^{d-1}$, it holds that $\forall\,\gamma\geq 0$,

$$\mathbb{P}(|h(x)-\mathbb{E}[h(x)]|\geq\gamma)\leq 2\sqrt{2\pi}\exp\left(-\frac{d\gamma^2}{8}\right). \tag{7}$$

Let $h(x)=x^\top a/\|a\|$ for $x\in\mathbb{S}^{d-1}$. First, we have that $\forall x,y\in\mathbb{S}^{d-1}$,

$$\begin{aligned}|h(x)-h(y)|^2&=\frac{|(x-y)^\top a|^2}{\|a\|^2}\\&\leq\|x-y\|^2\\&=2(1-x^\top y)\\&\leq(\arccos(x^\top y))^2,\end{aligned}$$

where we use the inequality that $\theta^2/2+\cos(\theta)-1\geq 0$ for $\theta\in[0,\pi]$ and let $x^\top y=\cos(\theta)$ such that $\arccos(x^\top y)=\theta$ for some $\theta\in[0,\pi]$. When u is uniformly sampled from $\sqrt{d}\cdot\mathbb{S}^{d-1}$, we know $u/\sqrt{d}$ is uniformly from $\mathbb{S}^{d-1}$. Applying (7) for $h(x)=x^\top a/\|a\|$ where $x\in\mathbb{S}^{d-1}$, we obtain that

$$\mathbb{P}\left(\left|\frac{u^\top a}{\sqrt{d}\|a\|}-\frac{\mathbb{E}[u^\top a]}{\sqrt{d}\|a\|}\right|\geq\gamma\right)\leq 2\sqrt{2\pi}\exp\left(-\frac{d\gamma^2}{8}\right).$$

Setting $C = \gamma\sqrt{d}\|a\|$, the proof is complete since $\mathbb{E}[u^\top a] = 0$. Similar results also exist in Theorem 5.1.4 of Vershynin [96], with all constants hidden behind some absolute c.

Next, we prove (iii). Applying (i), we have that

$$\begin{aligned}\mathbb{E}_u[(u^\top a)u_i] &= a_i\mathbb{E}[u_i^2] + \sum_{j\neq i} a_j\mathbb{E}[u_iu_j] \\ &= a_i.\end{aligned}$$

This implies that $\mathbb{E}_u[(u^\top a)u] = a$. Applying (ii), we obtain that

$$\begin{aligned}\mathbb{E}_u[(u^\top a)^2\|u\|^2] &= d\cdot\mathbb{E}_u[(u^\top a)^2] \\ &= d\|a\|^2.\end{aligned}$$

For the expectation of the matrix, we start from the diagonal terms.

$$\begin{aligned}\mathbb{E}_u[(u^\top a)^2u_i^2] &= \sum_{j=1}^d a_j^2\mathbb{E}[u_j^2u_i^2] + \sum_{j\neq k} a_ja_k\mathbb{E}[u_ju_ku_i^2] \\ &= a_i^2\mathbb{E}[u_i^4] + \sum_{j\neq i} a_j^2\mathbb{E}[u_j^2u_i^2].\end{aligned} \tag{8}$$

Here, we use the property that $\mathbb{E}[u_ju_ku_i^2] = 0$ for every i when $j \neq k$. This follows from symmetry of the sphere such that for any $u = (u_1, \cdots, u_j, \cdots, u_k, \cdots, u_d) \in \sqrt{d}\cdot\mathbb{S}^{d-1}$, it must be the case that $(u_1, \cdots, u_j, \cdots, -u_k, \cdots, u_d) \in \sqrt{d}\cdot\mathbb{S}^{d-1}$ as well. Again by symmetry, we have that $\mathbb{E}[u_i^4]$ remains the same for every i, and $\mathbb{E}[u_i^2u_j^2]$ remains the same for every $i \neq j$. Denote $w_1 = \mathbb{E}[u_i^4]$ and $w_2 = \mathbb{E}[u_i^2u_j^2]$. Since it holds that

$$\begin{aligned}\sum_{i=1}^d \mathbb{E}_u[(u^\top a)^2u_i^2] &= \mathbb{E}_u[(u^\top a)^2\|u\|^2] \\ &= d\|a\|^2,\end{aligned}$$

taking summation over (8), we can have that

$$\begin{aligned}d\|a\|^2 &= \sum_{i=1}^d a_i^2\mathbb{E}[u_i^4] + \sum_{i=1}^d\sum_{j=1,j\neq i}^d a_j^2\mathbb{E}[u_j^2u_i^2] \\ &= w_1\|a\|^2 + w_2\sum_{i=1}^d(\|a\|^2 - a_i^2) \\ &= w_1\|a\|^2 + (d-1)w_2\|a\|^2.\end{aligned}$$

This holds for arbitrary $a \in \mathbb{R}^d$, and thus we obtain that

$$w_1 + (d-1)w_2 = d. \tag{9}$$

We only compute $w_1 = \mathbb{E}[u_i^4]$ by showing that u_i^2/d actually follows the Beta distribution, and the value of w_2 can be derived from (9). First, $z/\|z\|$ is uniformly distributed on the unit sphere $\mathbb{S}^{d-1}$ for $z \in \mathbb{R}^d$ sampled from the standard multivariate Gaussian $\mathcal{N}(0, \mathrm{I}_d)$ [72, 71]. This means that z_i^2 is distributed according to the χ^2-distribution with 1 degree of freedom, and $\bar{z}_i^2 := \sum_{j\neq i} z_j^2$ is distributed according to the χ^2-distribution with degree $(d-1)$. Since χ^2-distribution is a special case of the Gamma distribution and z_i^2, $\bar{z}_i^2$ are independent, we conclude [19, 39] that $z_i^2/(z_i^2 + \bar{z}_i^2)$ has the Beta distribution with parameters $1/2$ and $(d-1)/2$. Finally, since $u/\sqrt{d}$ is uniformly distributed on $\mathbb{S}^{d-1}$, by symmetry of the sphere, we know that u_i^2/d has the same Beta distribution as $z_i^2/(z_i^2 + \bar{z}_i^2)$. The mean and variance of Beta$(1/2, (d-1)/2)$ is $1/d$ and $2(d-1)/(d^2(d+2))$. This suggests that $\mathbb{E}[u_i^2] = 1$, as already proved in (i), and that

$$\begin{aligned}w_1 &= \mathbb{E}[(u_i^2 - \mathbb{E}[u_i^2])^2] + (\mathbb{E}[u_i^2])^2 \\ &= d^2\left(\frac{2(d-1)}{d^2(d+2)} + \frac{1}{d^2}\right) \\ &= \frac{3d}{d+2}.\end{aligned}$$

By (9), we know $w_2 = d/(d+2)$. According to (8), we have that the diagonal terms

$$\begin{aligned}\mathbb{E}_u[(u^\top a)^2 u_i^2] &= w_1 a_i^2 + w_2(\|a\|^2 - a_i^2) \\ &= \frac{2d}{d+2} a_i^2 + \frac{d}{d+2}\|a\|^2.\end{aligned}$$

Then we compute the off-diagonal entries for $i \neq j$. By the same reasoning as (8), we have that

$$\begin{aligned}\mathbb{E}_u[(u^\top a)^2 u_i u_j] &= \sum_{i \neq j} a_i a_j \mathbb{E}[u_i^2 u_j^2] \\ &= \frac{2d}{d+2} a_i a_j.\end{aligned}$$

All other terms equal to 0 by symmetry of the sphere. Combining both diagonal and off-diagonal elements, we have that $\mathbb{E}_u[(u^\top a)^2 u u^\top] = (d/(d+2))(2aa^\top + \|a\|^2 \mathrm{I}_d)$. Similar results are also shown in Appendix F of Malladi et al. [70].

Finally, we give the proof of (iv). For the first statement, applying (i) in this lemma, we have that

$$\begin{aligned}\mathbb{E}_u\left[u^\top H u\right] &= \mathbb{E}\Big[\operatorname{Tr}(uu^\top H)\Big] \\ &= \operatorname{Tr}\Big(\mathbb{E}[uu^\top] \cdot H\Big) \\ &= \operatorname{Tr}(H).\end{aligned}$$

Similarly for the second statement, we apply (iii) in this lemma and obtain that

$$\begin{aligned}\mathbb{E}_u\left[(u^\top a)^2 u^\top H u\right] &= \mathbb{E}\Big[(u^\top a)^2 \cdot \operatorname{Tr}(uu^\top H)\Big] \\ &= \mathbb{E}\Big[\operatorname{Tr}\Big((u^\top a)^2 uu^\top \cdot H\Big)\Big] \\ &= \operatorname{Tr}\Big(\mathbb{E}\Big[(u^\top a)^2 uu^\top\Big] \cdot H\Big) \\ &= \frac{2d}{d+2}\operatorname{Tr}(aa^\top H) + \frac{d}{d+2}\|a\|^2 \operatorname{Tr}(H) \\ &= \frac{2d}{d+2} a^\top H a + \frac{d}{d+2}\|a\|^2 \operatorname{Tr}(H).\end{aligned}$$

This concludes the proof. □

Lemma A.2. *Let u be uniformly sampled from the Euclidean sphere $\sqrt{d}\,\mathbb{S}^{d-1}$ and v be uniformly sampled from the Euclidean ball $\sqrt{d}\,\mathbb{B}^d = \{x \in \mathbb{R}^d \mid \|x\| \leq \sqrt{d}\}$. For any function $f(x) : \mathbb{R}^d \to \mathbb{R}$ and $\lambda > 0$, we define its zeroth-order gradient estimator as $g_\lambda(x) = ((f(x+\lambda u) - f(x-\lambda u))/(2\lambda))u$, and the smoothed function $f_\lambda(x) = \mathbb{E}_v[f(x+\lambda v)]$. The following properties hold:*

- (i) *$f_\lambda(x)$ is differentiable and $\mathbb{E}_u[g_\lambda(x)] = \nabla f_\lambda(x)$.*
- (ii) *If $f(x)$ is ℓ-smooth, then we have that*

$$\begin{aligned}\|\nabla f(x) - \nabla f_\lambda(x)\| &\leq \frac{\ell}{2}\lambda d^{3/2}, \\ \mathbb{E}_u[\,\|g_\lambda(x)\|^2\,] &\leq 2d \cdot \|\nabla f(x)\|^2 + \frac{\ell^2}{2}\lambda^2 d^3.\end{aligned}$$

The above results are consistent with (iii) in Lemma A.1 when $\lambda \to 0$ and $f(x)$ is differentiable such that $g_0(x) = u^\top \nabla f(x) u$.

Proof. We first show (i). Similarly to Lemma 10 in Shamir [83], we have that

$$\mathbb{E}_{u \in \sqrt{d}\cdot\mathbb{S}^{d-1}}[g_\lambda(x)] = \mathbb{E}_{u \in \sqrt{d}\cdot\mathbb{S}^{d-1}}\left[\frac{f(x+\lambda u)u}{\lambda}\right].$$

Applying Lemma 2.1 in Flaxman et al. [30], we know

$$\mathbb{E}_{u' \in \mathbb{S}^{d-1}}[f(x+\lambda' u')u'] = \frac{\lambda'}{d} \nabla \mathbb{E}_{v' \in \mathbb{B}^d}[f(x+\lambda' v')].$$

Introducing $u = \sqrt{d}u'$, $v = \sqrt{d}v'$ and $\lambda = \lambda'/\sqrt{d}$, we thus obtain

$$\begin{aligned}
\mathbb{E}_{u \in \sqrt{d} \cdot \mathbb{S}^{d-1}} \left[\frac{f(x+\lambda u)u}{\lambda} \right] &= \mathbb{E}_{u' \in \mathbb{S}^{d-1}} \left[\frac{f(x+\lambda' u')u'd}{\lambda'} \right] \\
&= \nabla \mathbb{E}_{v' \in \mathbb{B}^d}[f(x+\lambda' v')] \\
&= \nabla \mathbb{E}_{v \in \sqrt{d} \cdot \mathbb{B}^d}[f(x+\lambda v)].
\end{aligned}$$

This suggests that $f_\lambda(x)$ is differentiable and $\mathbb{E}_u[g_\lambda(x)] = \nabla f_\lambda(x)$.

The proof of (ii) mostly follows from Nesterov and Spokoiny [74], where the results are originally obtained for the case that u is sampled from the standard multivariate Gaussian distribution. By (iii) in Lemma A.1 and (i) here, we have that for u uniformly sampled from $\sqrt{d} \cdot \mathbb{S}^{d-1}$,

$$\begin{aligned}
\|\nabla f(x) - \nabla f_\lambda(x)\| &= \left\| \mathbb{E}_u[(u^\top \nabla f(x))u] - \mathbb{E}_u \left[\frac{f(x+\lambda u) - f(x-\lambda u)}{2\lambda} u \right] \right\| \\
&\leq \mathbb{E}_u \left\| \left(\frac{f(x+\lambda u) - f(x-\lambda u)}{2\lambda} - u^\top \nabla f(x) \right) u \right\| \\
&\leq \frac{\sqrt{d}}{2\lambda} \mathbb{E}_u |f(x+\lambda u) - f(x) - \lambda u^\top \nabla f(x)| \\
&\quad + \frac{\sqrt{d}}{2\lambda} \mathbb{E}_u |f(x) - f(x-\lambda u) - \lambda u^\top \nabla f(x)| \\
&\leq \frac{\ell}{2} \lambda d^{3/2},
\end{aligned}$$

where in the last step we use smoothness of $f(x)$ such that $|f(x+\lambda u) - f(x) - \lambda u^\top \nabla f(x)| \leq \ell \lambda^2 d/2$ and the same holds for $|f(x) - f(x-\lambda u) - \lambda u^\top \nabla f(x)| = |f(x-\lambda u) - f(x) + \lambda u^\top \nabla f(x)|$. To show the last statement, similarly we have that

$$\begin{aligned}
\mathbb{E}_u[\|g_\lambda(x)\|^2] &= \frac{d}{4\lambda^2} \mathbb{E}_u[(f(x+\lambda u) - f(x-\lambda u))^2] \\
&\leq 2d \cdot \mathbb{E}_u[(u^\top \nabla f(x))^2] + \frac{d}{2\lambda^2} \mathbb{E}_u[(f(x+\lambda u) - f(x-\lambda u) - 2\lambda u^\top \nabla f(x))^2] \\
&\leq 2d \cdot \mathbb{E}_u[(u^\top \nabla f(x))^2] + \frac{d}{\lambda^2} \mathbb{E}_u[(f(x+\lambda u) - f(x) - \lambda u^\top \nabla f(x))^2] \\
&\quad + \frac{d}{\lambda^2} \mathbb{E}_u[(f(x) - f(x-\lambda u) - \lambda u^\top \nabla f(x))^2] \\
&\leq 2d \cdot \|\nabla f(x)\|^2 + \frac{\ell^2}{2} \lambda^2 d^3, \qquad (10)
\end{aligned}$$

where in the last step we use Lemma A.1 and smoothness of $f(x)$. □

B Proof of Theorem 3.2 and 3.4

Proof of Theorem 3.2. The privacy guarantees directly follow from Lemma 2.1 noticing that the sensitivity is $2C/n$. We then focus on the utility guarantee on $\mathbb{E}[\|\nabla F_S(x_\tau)\|^2]$. Since $f(x;\xi)$ is L-Lipschitz for every ξ and $\|u_t\| = \sqrt{d}$, we have that

$$\begin{aligned}
\|g_\lambda(x_t;\xi_i)\| &= \frac{|f(x_t + \lambda u_t;\xi_i) - f(x_t - \lambda u_t;\xi_i)|}{2\lambda} \|u_t\| \\
&\leq L\|u_t\|^2 \\
&= Ld.
\end{aligned}$$

This means $\text{clip}_C(g_\lambda(x_t;\xi_i)) = g_\lambda(x_t;\xi_i)$ when setting $C = Ld$. For notation simplicity, we let

$$\begin{aligned} G_\lambda(x_t) &:= \frac{1}{n}\sum_{i=1}^{n} g_\lambda(x_t;\xi_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{f(x_t+\lambda u_t;\xi_i) - f(x_t-\lambda u_t;\xi_i)}{2\lambda} u_t \\ &= \frac{F_S(x_t+\lambda u_t) - F_S(x_t-\lambda u_t)}{2\lambda} u_t. \end{aligned}$$

Algorithm 1 reduces to $x_{t+1} = x_t - \alpha(G_\lambda(x_t) + z_t)$. By smoothness of $F_S(x)$, we have that

$$\begin{aligned} F_S(x_{t+1}) &\le F_S(x_t) + \nabla F_S(x_t)^\top (x_{t+1} - x_t) + \frac{\ell}{2}\|x_{t+1} - x_t\|^2 \\ &= F_S(x_t) - \alpha \nabla F_S(x_t)^\top (G_\lambda(x_t) + z_t) + \frac{\ell}{2}\alpha^2 \|G_\lambda(x_t)\|^2 + \frac{\ell}{2}\alpha^2\|z_t\|^2 + \ell\alpha^2 z_t^\top G_\lambda(x_t). \end{aligned}$$

Since z_t is sampled from $\mathcal{N}(0, \sigma^2 \mathrm{I}_d)$ and is independent of x_t, u_t and S, we have that

$$\mathbb{E}_{z_t}[F_S(x_{t+1})] \le F_S(x_t) - \alpha \nabla F_S(x_t)^\top G_\lambda(x_t) + \frac{\ell}{2}\alpha^2 \|G_\lambda(x_t)\|^2 + \frac{\ell}{2}\alpha^2\, d\, \sigma^2.$$

Define $F_\lambda(x) := \mathbb{E}_v[F_S(x + \lambda v)]$ for v sampled uniformly from the Euclidean ball $\sqrt{d}\cdot\mathbb{B}^d$. By Lemma A.2, we know $\mathbb{E}_{u_t}[G_\lambda(x_t)] = \nabla F_\lambda(x_t)$. Since u_t is independent of x_t and S, taking expectation with respect to u_t and applying (ii) in Lemma A.2, we obtain that

$$\begin{aligned} \mathbb{E}_{z_t,u_t}[F_S(x_{t+1})] &\le F_S(x_t) - \alpha \nabla F_S(x_t)^\top \nabla F_\lambda(x_t) + \frac{\ell}{2}\alpha^2 \mathbb{E}_{u_t}[\|G_\lambda(x_t)\|^2] + \frac{\ell}{2}\alpha^2\, d\, \sigma^2 \\ &= F_S(x_t) - \frac{\alpha}{2}\|\nabla F_S(x_t)\|^2 - \frac{\alpha}{2}\|\nabla F_\lambda(x_t)\|^2 + \frac{\alpha}{2}\|\nabla F_\lambda(x_t) - \nabla F_S(x_t)\|^2 \\ &\quad + \frac{\ell}{2}\alpha^2\, \mathbb{E}_{u_t}[\|G_\lambda(x_t)\|^2] + \frac{\ell}{2}\alpha^2\, d\, \sigma^2 \\ &\le F_S(x_t) - \frac{\alpha}{2}(1 - 2d\ell\alpha)\|\nabla F_S(x_t)\|^2 + \frac{\ell^2}{8}\alpha(1+2\ell\alpha)\lambda^2 d^3 + \frac{\ell}{2}\alpha^2\, d\, \sigma^2. \end{aligned} \tag{11}$$

Choosing $\alpha = 1/(4\ell d)$ such that $1 - 2d\ell\alpha = 1/2$ and $2\ell\alpha < 1$, we obtain that

$$\begin{aligned} \mathbb{E}[\|\nabla F_S(x_t)\|^2] &< \frac{4\,\mathbb{E}[F_S(x_t) - F_S(x_{t+1})]}{\alpha} + \ell^2\lambda^2 d^3 + 2\ell\alpha\, d\sigma^2 \\ &= \frac{4\,\mathbb{E}[F_S(x_t) - F_S(x_{t+1})]}{\alpha} + \ell^2\lambda^2 d^3 + \frac{64\ell C^2\, \alpha T\, d \log(e + (\varepsilon/\delta))}{n^2\varepsilon^2} \\ &= \frac{4\,\mathbb{E}[F_S(x_t) - F_S(x_{t+1})]}{\alpha} + \ell^2\lambda^2 d^3 + \frac{64\ell L^2\, \alpha T\, d^3 \log(e + (\varepsilon/\delta))}{n^2\varepsilon^2}. \end{aligned}$$

As a result, taking summation from $t = 0$ to $T-1$ and dividing both sides by T, we have that

$$\begin{aligned} \mathbb{E}[\|\nabla F_S(x_\tau)\|^2] &= \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla F_S(x_t)\|^2] \\ &\le \frac{4(F_S(x_0) - F_S^*)}{\alpha T} + \ell^2\lambda^2 d^3 + \frac{64\ell L^2\, \alpha T\, d^3 \log(e + (\varepsilon/\delta))}{n^2\varepsilon^2} \\ &\le \frac{16(\ell(F_S(x_0) - F_S^*) + 2L^2) d\sqrt{d\log(e + (\varepsilon/\delta))}}{n\varepsilon}, \end{aligned}$$

with the choice of parameters

$$\alpha T = \frac{n\varepsilon}{4\ell d\sqrt{d\log(e + (\varepsilon/\delta))}}, \quad \lambda \le \frac{4L}{\ell d}\left(\frac{\sqrt{d\log(e + (\varepsilon/\delta))}}{n\varepsilon}\right)^{1/2}.$$

This suggests that the total number of iteration is $T = n\varepsilon/\sqrt{d\log(e + (\varepsilon/\delta))}$ and the total number of zeroth-order gradient computations is $nT = n^2\varepsilon/\sqrt{d\log(e + (\varepsilon/\delta))}$. Note that the above selection of parameters ensures scale invariance. □

Proof of Theorem 3.4. The privacy analysis remains the same as before, and we focus on the utility analysis on $\mathbb{E}\|\nabla F_S(x_\tau)\|^2$. By the same reasoning, when setting $C = Ld$, Algorithm 1 reduces to $x_{t+1} = x_t - \alpha(G_\lambda(x_t) + z_t)$ where $G_\lambda(x_t) = (F_S(x_t + \lambda u_t) - F_S(x_t - \lambda u_t))u_t/(2\lambda)$. By Taylor's theorem with remainder, for some $\theta \in (0, 1)$, we have that

$$\begin{aligned}
F_S(x_{t+1}) &= F_S(x_t) + \nabla F_S(x_t)^\top (x_{t+1} - x_t) + \frac{1}{2}(x_{t+1} - x_t)^\top \nabla^2 F_S(x_t + \theta(x_{t+1} - x_t))(x_{t+1} - x_t) \\
&\le F_S(x_t) - \alpha \nabla F_S(x_t)^\top (G_\lambda(x_t) + z_t) + \frac{\alpha^2}{2} G_\lambda(x_t)^\top H G_\lambda(x_t) + \frac{\alpha^2}{2} z_t^\top H z_t \\
&\quad + \frac{\alpha^2}{2} \left(G_\lambda(x_t)^\top H z_t + z_t^\top H G_\lambda(x_t) \right).
\end{aligned}$$

Here in the inequality, we use Assumption 3.3 such that $\nabla^2 F_S(x) \preceq H$ for any $x \in \mathbb{R}^d$. Similarly to (iv) in Lemma A.1, we have that $\mathbb{E}[z_t^\top H z_t] = \operatorname{Tr}(\mathbb{E}[z_t z_t^\top] H) = \sigma^2 \operatorname{Tr}(H)$. Since z_t is sampled from $\mathcal{N}(0, \sigma^2 \mathrm{I}_d)$ and is independent of u_t, x_t and the dataset S, taking expectation with respect to z_t, we can then obtain that

$$\begin{aligned}
\mathbb{E}_{z_t}[F_S(x_{t+1})] &\le F_S(x_t) - \alpha \nabla F_S(x_t)^\top G_\lambda(x_t) + \frac{\alpha^2}{2} G_\lambda(x_t)^\top H G_\lambda(x_t) + \frac{\alpha^2}{2} \mathbb{E}_{z_t}[z_t^\top H z_t] \\
&= F_S(x_t) - \alpha \nabla F_S(x_t)^\top G_\lambda(x_t) + \frac{\alpha^2}{2} G_\lambda(x_t)^\top H G_\lambda(x_t) + \frac{\alpha^2 \sigma^2}{2} \operatorname{Tr}(H).
\end{aligned} \tag{12}$$

Assumption 3.3 implies $F_S(x)$ is also ℓ-smooth. By a similar argument as (10) in the proof of (ii) in Lemma A.2, we have that

$$\left(\frac{F_S(x_t + \lambda u_t) - F_S(x_t - \lambda u_t)}{2\lambda} \right)^2 \le 2 \left(u_t^\top \nabla F_S(x_t) \right)^2 + \frac{\ell^2}{2} \lambda^2 d^2. \tag{13}$$

As $u_t^\top H u_t \ge 0$, by (iv) in Lemma A.1 and Assumption 3.3, we have that

$$\begin{aligned}
\mathbb{E}\left[G_\lambda(x_t)^\top H G_\lambda(x_t) \right] &= \mathbb{E}\left[\left(\frac{F_S(x_t + \lambda u_t) - F_S(x_t - \lambda u_t)}{2\lambda} \right)^2 u_t^\top H u_t \right] \\
&\le 2\, \mathbb{E}\left[\left(u_t^\top \nabla F_S(x_t) \right)^2 u_t^\top H u_t \right] + \frac{\ell^2}{2} \lambda^2 d^2\, \mathbb{E}\left[u_t^\top H u_t \right] \\
&= \frac{2d}{d+2} \left(2 \nabla F_S(x_t)^\top H \nabla F_S(x_t) + \|\nabla F_S(x_t)\|^2 \operatorname{Tr}(H) \right) + \frac{\ell^2}{2} \lambda^2 d^2 \operatorname{Tr}(H) \\
&\le 2\ell(r+2) \|\nabla F_S(x_t)\|^2 + \frac{\ell^3}{2} \lambda^2 d^2 r.
\end{aligned}$$

Taking expectation of (12) with respect to u_t, by Lemma A.2 for $F_\lambda(x) = \mathbb{E}_v[F_S(x + \lambda v)]$ with v uniformly sampled from $\sqrt{d} \cdot \mathbb{B}^d$, we have that

$$\begin{aligned}
\mathbb{E}[F_S(x_{t+1})] &\le F_S(x_t) - \alpha \nabla F_S(x_t)^\top \nabla F_\lambda(x_t) + \ell \alpha^2 (r+2) \|\nabla F_S(x_t)\|^2 + \frac{\ell^3 \alpha^2 \lambda^2 d^2 r}{4} + \frac{\ell \alpha^2\, r \sigma^2}{2} \\
&\le F_S(x_t) - \frac{\alpha}{2}(1 - 2(r+2)\ell\alpha) \|\nabla F_S(x_t)\|^2 + \frac{\alpha}{2} \|\nabla F_S(x_t) - \nabla F_\lambda(x_t)\|^2 + \frac{\ell^3 \alpha^2 \lambda^2 d^2 r}{4} + \frac{\ell \alpha^2\, r \sigma^2}{2} \\
&\le F_S(x_t) - \frac{\alpha}{2}(1 - 2(r+2)\ell\alpha) \|\nabla F_S(x_t)\|^2 + \frac{\ell^2 \alpha \lambda^2 d^2 (d + 2r\ell\alpha)}{8} + \frac{\ell \alpha^2\, r \sigma^2}{2}.
\end{aligned} \tag{14}$$

Choosing $\alpha = 1/(4\ell(r+2))$ such that $1 - 2(r+2)\ell\alpha = 1/2$ and $2\ell\alpha r < 1 \le d$, we have that

$$\begin{aligned}
\mathbb{E}[\|\nabla F_S(x_t)\|^2] &< \frac{4\, \mathbb{E}[F_S(x_t) - F_S(x_{t+1})]}{\alpha} + \ell^2 \lambda^2 d^3 + 2\ell\alpha\, r\sigma^2 \\
&= \frac{4\, \mathbb{E}[F_S(x_t) - F_S(x_{t+1})]}{\alpha} + \ell^2 \lambda^2 d^3 + \frac{64 \ell C^2\, \alpha T\, r \log(e + (\varepsilon/\delta))}{n^2 \varepsilon^2} \\
&= \frac{4\, \mathbb{E}[F_S(x_t) - F_S(x_{t+1})]}{\alpha} + \ell^2 \lambda^2 d^3 + \frac{64 \ell L^2\, \alpha T\, d^2 r \log(e + (\varepsilon/\delta))}{n^2 \varepsilon^2}.
\end{aligned}$$

As a result, taking summation from $t = 0$ to $T - 1$ and dividing both sides by T, we have that

$$\begin{aligned}\mathbb{E}[\|\nabla F_S(x_\tau)\|^2] &= \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla F_S(x_t)\|^2] \\ &\leq \frac{4(F_S(x_0) - F_S^*)}{\alpha T} + \ell^2\lambda^2 d^3 + \frac{64\ell L^2\,\alpha T\, d^2 r\log(e + (\varepsilon/\delta))}{n^2\varepsilon^2} \\ &\leq \frac{16(\ell(F_S(x_0) - F_S^*) + 2L^2)d\sqrt{r\log(e + (\varepsilon/\delta))}}{n\varepsilon},\end{aligned}$$

with the choice of parameters

$$\alpha T = \frac{n\varepsilon}{4\ell d\sqrt{r\log(e + (\varepsilon/\delta))}}, \quad \lambda \leq \frac{4L}{\ell d}\left(\frac{\sqrt{r\log(e + (\varepsilon/\delta))}}{n\varepsilon}\right)^{1/2}.$$

This suggests that the total number of iteration is $T = n(r+2)\varepsilon/(d\sqrt{r\log(e + (\varepsilon/\delta))})$ and the total number of zeroth-order gradient computations is $nT = n^2(r+2)\varepsilon/(d\sqrt{r\log(e + (\varepsilon/\delta))})$. Note that the above selection of parameters ensures scale invariance. □

C Proof of Theorem 4.1

Since u_t is independent of the dataset S, the privacy guarantees directly follow from Lemma 2.1 and post-processing [25] noticing that the sensitivity is $2C/n$. We then focus on the utility guarantee on $\mathbb{E}\|\nabla F_S(x_\tau)\|^2$. Since $f(x;\xi)$ is ℓ-smooth for every ξ, we have that

$$\begin{aligned}\frac{|f(x_t + \lambda u_t;\xi_i) - f(x_t - \lambda u_t;\xi_i)|}{2\lambda} &\leq |u_t^\top\nabla f(x_t;\xi_i)| + \frac{|f(x_t + \lambda u_t;\xi_i) - f(x_t;\xi_i) - \lambda u_t^\top\nabla f(x_t;\xi_i)|}{2\lambda} \\ &\quad + \frac{|f(x_t - \lambda u_t;\xi_i) - f(x_t;\xi_i) + \lambda u_t^\top\nabla f(x_t;\xi_i)|}{2\lambda} \\ &\leq |u_t^\top\nabla f(x_t;\xi_i)| + \frac{\ell}{2}\lambda d.\end{aligned} \tag{15}$$

Therefore, by (ii) in Lemma A.1 and Lipschitzness of $f(x;\xi)$, we have that

$$\begin{aligned}\mathbb{P}\left(\frac{|f(x_t + \lambda u_t;\xi_i) - f(x_t - \lambda u_t;\xi_i)|}{2\lambda} \geq C_0 + \frac{\ell}{2}\lambda d\right) &\leq \mathbb{P}(|u_t^\top\nabla f(x_t;\xi_i)| \geq C_0) \\ &\leq 2\sqrt{2\pi}\exp\left(-\frac{C_0^2}{8\|\nabla f(x_t;\xi_i)\|^2}\right) \\ &\leq 2\sqrt{2\pi}\exp\left(-\frac{C_0^2}{8L^2}\right).\end{aligned}$$

We define $Q_{t,i}$ to be the event that the clipping does not happen at iteration t for sample ξ_i, and $\bar{Q}_{t,i}$ to be the event that the clipping does happen. The above equation implies that if the clipping threshold is chosen to be $C \geq C_0 + \ell\lambda d/2$, then we have that $\mathbb{P}(Q_{t,i}) = 1 - \mathbb{P}(\bar{Q}_{t,i}) \geq 1 - 2\sqrt{2\pi}\exp(-C_0^2/(8L^2))$. Let Q denote the event that the clipping does not happen for every iteration $t = 0, 1, \cdots, T-1$ and every sample $1 \leq i \leq n$, and $\bar{Q}$ be the event that there exist some t and i such that the clipping does happen. By the union bound, we have that

$$\begin{aligned}\mathbb{P}(Q) &= 1 - \mathbb{P}(\bar{Q}) \\ &= 1 - \mathbb{P}\left(\bigcup_{t=0}^{T-1}\bigcup_{i=1}^{n}\bar{Q}_{t,i}\right) \\ &\geq 1 - 2\sqrt{2\pi}\cdot nT\exp\left(-\frac{C_0^2}{8L^2}\right).\end{aligned}$$

To simplify the notation, we let

$$G_\lambda(x_t) = \frac{1}{n}\sum_{i=1}^{n} \frac{f(x_t+\lambda u_t;\xi_i) - f(x_t-\lambda u_t;\xi_i)}{2\lambda} u_t$$
$$= \frac{F_S(x_t+\lambda u_t) - F_S(x_t-\lambda u_t)}{2\lambda} u_t,$$

and its per-sample clipped version as

$$\hat{G}_\lambda(x_t) = \frac{1}{n}\sum_{i=1}^{n} \text{clip}_C\left(\frac{f(x_t+\lambda u_t;\xi_i) - f(x_t-\lambda u_t;\xi_i)}{2\lambda}\right) u_t.$$

Algorithm 2 becomes $x_{t+1} = x_t - \alpha(\hat{G}_\lambda(x_t) + z_t u_t)$ under the above notation. By Taylor's theorem with remainder, for some $\theta \in (0,1)$, we have that

$$\begin{aligned} F_S(x_{t+1}) &= F_S(x_t) + \nabla F_S(x_t)^\top (x_{t+1}-x_t) + \frac{1}{2}(x_{t+1}-x_t)^\top \nabla^2 F_S(x_t+\theta(x_{t+1}-x_t))(x_{t+1}-x_t) \\ &\le F_S(x_t) - \alpha \nabla F_S(x_t)^\top \left(\hat{G}_\lambda(x_t) + z_t u_t\right) + \frac{\alpha^2}{2}\hat{G}_\lambda(x_t)^\top H \hat{G}_\lambda(x_t) + \frac{\alpha^2}{2} z_t^2 u_t^\top H u_t \\ &\quad + \frac{\alpha^2}{2} z_t \left(\hat{G}_\lambda(x_t)^\top H u_t + u_t^\top H \hat{G}_\lambda(x_t)\right). \end{aligned}$$

Here in the inequality, we use Assumption 3.3 such that $\nabla^2 F_S(x) \preceq H$ for any $x \in \mathbb{R}^d$. Note that in Algorithm 2, $z_t \in \mathbb{R}$ is a scalar. Since z_t is sampled from $\mathcal{N}(0,\sigma^2)$ and is independent of u_t, x_t and the dataset S, taking expectation with respect to z_t, we have that

$$\mathbb{E}_{z_t}[F_S(x_{t+1})] \le F_S(x_t) - \alpha \nabla F_S(x_t)^\top \hat{G}_\lambda(x_t) + \frac{\alpha^2}{2}\hat{G}_\lambda(x_t)^\top H \hat{G}_\lambda(x_t) + \frac{\alpha^2\sigma^2}{2} u_t^\top H u_t. \quad (16)$$

We then compute the expectation of each term conditioned on the event Q. When Q happens, we know that $\hat{G}_\lambda(x_t) = G_\lambda(x_t)$ for every t and thus

$$\mathbb{E}\left[\hat{G}_\lambda(x_t)^\top H \hat{G}_\lambda(x_t) \,\middle|\, Q\right] = \mathbb{E}\left[\left(\frac{F_S(x_t+\lambda u_t) - F_S(x_t-\lambda u_t)}{2\lambda}\right)^2 u_t^\top H u_t \,\middle|\, Q\right].$$

Since $H \succeq 0$, we have that $u_t^\top H u_t \ge 0$. By law of total probability, we obtain

$$\begin{aligned} &\mathbb{E}\left[\left(\frac{F_S(x_t+\lambda u_t) - F_S(x_t-\lambda u_t)}{2\lambda}\right)^2 u_t^\top H u_t\right] \\ &= \mathbb{E}\left[\left(\frac{F_S(x_t+\lambda u_t) - F_S(x_t-\lambda u_t)}{2\lambda}\right)^2 u_t^\top H u_t \,\middle|\, Q\right]\mathbb{P}(Q) \\ &\quad + \mathbb{E}\left[\left(\frac{F_S(x_t+\lambda u_t) - F_S(x_t-\lambda u_t)}{2\lambda}\right)^2 u_t^\top H u_t \,\middle|\, \bar{Q}\right]\mathbb{P}(\bar{Q}) \\ &\ge \mathbb{E}\left[\left(\frac{F_S(x_t+\lambda u_t) - F_S(x_t-\lambda u_t)}{2\lambda}\right)^2 u_t^\top H u_t \,\middle|\, Q\right]\mathbb{P}(Q). \end{aligned} \quad (17)$$

Assumption 3.3 implies $F_S(x)$ is also ℓ-smooth. Similarly to the proof of Theorem 3.4, by (13) and the fact that $u_t^\top H u_t \ge 0$, applying (iv) in Lemma A.1 and Assumption 3.3, we can then obtain that

$$\begin{aligned} \mathbb{E}\left[\hat{G}_\lambda(x_t)^\top H \hat{G}_\lambda(x_t) \,\middle|\, Q\right] &\le \frac{\mathbb{E}\left[(F_S(x_t+\lambda u_t) - F_S(x_t-\lambda u_t))^2 u_t^\top H u_t\right]}{4\lambda^2 \cdot \mathbb{P}(Q)} \\ &\le \frac{\mathbb{E}\left[2\left(u_t^\top \nabla F_S(x_t)\right)^2 u_t^\top H u_t\right]}{\mathbb{P}(Q)} + \frac{\ell^2\lambda^2 d^2}{2\,\mathbb{P}(Q)}\mathbb{E}\left[u_t^\top H u_t\right] \\ &= \frac{2d\left(2\nabla F_S(x_t)^\top H \nabla F_S(x_t) + \|\nabla F_S(x_t)\|^2 \,\text{Tr}(H)\right)}{(d+2)\,\mathbb{P}(Q)} + \frac{\ell^2\lambda^2 d^2\,\text{Tr}(H)}{2\,\mathbb{P}(Q)} \\ &\le \frac{2\ell(r+2)}{\mathbb{P}(Q)}\|\nabla F_S(x_t)\|^2 + \frac{\ell^3\lambda^2 d^2 r}{2\,\mathbb{P}(Q)}. \end{aligned} \quad (18)$$

The same as (17), we can also get that

$$\begin{aligned}\mathbb{E}\left[u_t^\top H u_t \,\middle|\, Q\right] &\leq \frac{\mathbb{E}\left[u_t^\top H u_t\right]}{\mathbb{P}(Q)} \\ &\leq \frac{r\ell}{\mathbb{P}(Q)}.\end{aligned} \tag{19}$$

For the inner-product term, we have that

$$\begin{aligned}\mathbb{E}\left[\nabla F_S(x_t)^\top \hat{G}_\lambda(x_t) \,\middle|\, Q\right] &= \mathbb{E}\left[\nabla F_S(x_t)^\top G_\lambda(x_t) \,\middle|\, Q\right] \\ &= \mathbb{E}\left[\frac{F_S(x_t+\lambda u_t) - F_S(x_t-\lambda u_t)}{2\lambda} \cdot u_t^\top \nabla F_S(x_t) \,\middle|\, Q\right].\end{aligned}$$

By law of total probability, we know that

$$\begin{aligned}\mathbb{E}\left[\nabla F_S(x_t)^\top G_\lambda(x_t) \,\middle|\, Q\right]\mathbb{P}(Q) + \mathbb{E}\left[\nabla F_S(x_t)^\top G_\lambda(x_t) \,\middle|\, \bar{Q}\right]\mathbb{P}(\bar{Q}) &= \mathbb{E}\left[\nabla F_S(x_t)^\top G_\lambda(x_t)\right] \\ &= \mathbb{E}\left[\nabla F_S(x_t)^\top \nabla F_\lambda(x_t)\right],\end{aligned}$$

where we use Lemma A.2 for $F_\lambda(x) = \mathbb{E}_v[F_S(x+\lambda v)]$ with v uniformly sampled from $\sqrt{d}\cdot\mathbb{B}^d$. Rearranging terms, we thus obtain that

$$\begin{aligned}\mathbb{E}\left[\nabla F_S(x_t)^\top G_\lambda(x_t) \,\middle|\, Q\right] &= \frac{\mathbb{E}\left[\nabla F_S(x_t)^\top \nabla F_\lambda(x_t)\right]}{\mathbb{P}(Q)} - \frac{\mathbb{E}\left[\nabla F_S(x_t)^\top G_\lambda(x_t) \,\middle|\, \bar{Q}\right]\mathbb{P}(\bar{Q})}{\mathbb{P}(Q)} \\ &= \frac{\mathbb{E}\|\nabla F_S(x_t)\|^2}{2\,\mathbb{P}(Q)} + \frac{\mathbb{E}\|\nabla F_\lambda(x_t)\|^2}{2\,\mathbb{P}(Q)} - \frac{\mathbb{E}\|\nabla F_S(x_t) - \nabla F_\lambda(x_t)\|^2}{2\,\mathbb{P}(Q)} \\ &\quad - \frac{\mathbb{E}\left[\nabla F_S(x_t)^\top G_\lambda(x_t) \,\middle|\, \bar{Q}\right]\mathbb{P}(\bar{Q})}{\mathbb{P}(Q)} \\ &\geq \frac{\mathbb{E}\|\nabla F_S(x_t)\|^2}{2\,\mathbb{P}(Q)} - \frac{\ell^2\lambda^2 d^3}{8\,\mathbb{P}(Q)} - \frac{\mathbb{E}\left[\nabla F_S(x_t)^\top G_\lambda(x_t) \,\middle|\, \bar{Q}\right]\mathbb{P}(\bar{Q})}{\mathbb{P}(Q)},\end{aligned}$$

where we apply (ii) in Lemma A.2. Assumption 3.3 implies that $F_S(x)$ is also Lipschitz, and thus

$$\begin{aligned}\nabla F_S(x_t)^\top G_\lambda(x_t) &\leq \|\nabla F_S(x_t)\|\|G_\lambda(x_t)\| \\ &\leq L^2\|u_t\|^2 \\ &= L^2 d.\end{aligned}$$

As a result, we obtain that

$$\mathbb{E}\left[\nabla F_S(x_t)^\top \hat{G}_\lambda(x_t) \,\middle|\, Q\right] \geq \frac{\mathbb{E}\|\nabla F_S(x_t)\|^2}{2\,\mathbb{P}(Q)} - \frac{\ell^2\lambda^2 d^3}{8\,\mathbb{P}(Q)} - \frac{L^2 d\,\mathbb{P}(\bar{Q})}{\mathbb{P}(Q)}. \tag{20}$$

Plugging (20), (18) and (19) back into (16), we obtain that

$$\begin{aligned}\mathbb{E}[F_S(x_{t+1})|Q] \leq \mathbb{E}[F_S(x_t)|Q] &- \frac{\alpha}{2}(1-2(r+2)\ell\alpha)\frac{\mathbb{E}\|\nabla F_S(x_t)\|^2}{\mathbb{P}(Q)} + \frac{\ell\alpha^2\, r\sigma^2}{2\,\mathbb{P}(Q)} \\ &+ \frac{\ell^2\alpha(d+2\ell\alpha r)\lambda^2 d^2}{8\,\mathbb{P}(Q)} + \frac{\alpha L^2 d\,\mathbb{P}(\bar{Q})}{\mathbb{P}(Q)}.\end{aligned} \tag{21}$$

Choosing $\alpha = 1/(4\ell(r+2))$ such that $1-2(r+2)\ell\alpha = 1/2$ and $2\ell\alpha r < 1 \leq d$, we have that

$$\begin{aligned}\mathbb{E}\|\nabla F_S(x_t)\|^2 &\leq \frac{4\,\mathbb{E}[F_S(x_t) - F_S(x_{t+1})|Q]\mathbb{P}(Q)}{\alpha} + 2\ell\alpha\, r\sigma^2 + \ell^2 d^3\lambda^2 + 4L^2 d\,\mathbb{P}(\bar{Q}) \\ &\leq \frac{4\,\mathbb{E}[F_S(x_t) - F_S(x_{t+1})|Q]\mathbb{P}(Q)}{\alpha} + \frac{64\ell C^2\,\alpha T\, r\log(e+(\varepsilon/\delta))}{n^2\varepsilon^2} \\ &\quad + \ell^2 d^3\lambda^2 + 8\sqrt{2\pi}L^2\, n\, d\, T\exp\left(-\frac{C_0^2}{8L^2}\right).\end{aligned}$$

Taking summation from $t = 0$ to $T - 1$ and dividing both sides by T, we have that

$$\begin{aligned}
\mathbb{E}\|\nabla F_S(x_\tau)\|^2 &= \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F_S(x_t)\|^2 \\
&\leq \frac{4[F_S(x_0) - F_S^*]}{\alpha T} + \frac{64\ell C^2\,\alpha T\, r\log(e + (\varepsilon/\delta))}{n^2\varepsilon^2} + \ell^2 d^3\lambda^2 \\
&\quad + 8\sqrt{2\pi}L^2\, n\, d\, T\exp\left(-\frac{C_0^2}{8L^2}\right) \\
&= \left(64\ell[F_S(x_0) - F_S^*] + 4C^2\right)\frac{\sqrt{r\log(e + (\varepsilon/\delta))}}{n\varepsilon} + \ell^2 d^3\lambda^2 \\
&\quad + \frac{2\sqrt{2\pi}L^2\, n^2 d(r+2)\varepsilon}{\sqrt{r\log(e + (\varepsilon/\delta))}}\exp\left(-\frac{C_0^2}{8L^2}\right),
\end{aligned}$$

with the choice of parameters to be

$$\alpha T = \frac{n\varepsilon}{16\ell\sqrt{r\log(e + (\varepsilon/\delta))}}, \quad \alpha = \frac{1}{4\ell(r+2)}, \quad T = \frac{n(r+2)\varepsilon}{4\sqrt{r\log(e + (\varepsilon/\delta))}}.$$

When selecting $\lambda \leq 2(\sqrt{2} - 1)C_0/(\ell d)$, we can set $C = \sqrt{2}C_0$ such that $C \geq C_0 + \ell\lambda d/2$ is satisfied. If C_0 and λ further satisfy that

$$C_0^2 = 8L^2\,\log\left(\frac{2\sqrt{2\pi}\,d(r+2)n^3\varepsilon^2}{r\log(e + (\varepsilon/\delta))}\right), \quad \lambda \leq \frac{L}{\ell d^{3/2}}\left(\frac{\sqrt{r\log(e + (\varepsilon/\delta))}}{n\varepsilon}\right)^{1/2},$$

we can then obtain that

$$\begin{aligned}
\mathbb{E}\|\nabla F_S(x_\tau)\|^2 &\leq \left(64\,\ell[F_S(x_0) - F_S^*] + 4C^2 + 2L^2\right)\frac{\sqrt{r\log(e + (\varepsilon/\delta))}}{n\varepsilon} \\
&= \left(64\,\ell[F_S(x_0) - F_S^*] + 64\,L^2\,\log\left(\frac{2\sqrt{2\pi}\,d(r+2)n^3\varepsilon^2}{r\log(e + (\varepsilon/\delta))}\right) + 2L^2\right)\frac{\sqrt{r\log(e + (\varepsilon/\delta))}}{n\varepsilon}.
\end{aligned}$$

We conclude that the clipping threshold C and smoothing parameter λ should satisfy that

$$\begin{aligned}
&C = 4L\sqrt{\log\left(\frac{2\sqrt{2\pi}\,d(r+2)n^3\varepsilon^2}{r\log(e + (\varepsilon/\delta))}\right)}, \\
&\lambda \leq \frac{L}{\ell d}\min\left\{4(2-\sqrt{2})\sqrt{\log\left(\frac{2\sqrt{2\pi}\,d(r+2)n^3\varepsilon^2}{r\log(e + (\varepsilon/\delta))}\right)}, \frac{1}{\sqrt{d}}\left(\frac{\sqrt{r\log(e + (\varepsilon/\delta))}}{n\varepsilon}\right)^{1/2}\right\}.
\end{aligned}$$

The total number of zeroth-order gradient computations is $nT = n^2(r+2)\varepsilon/(4\sqrt{r\log(e + (\varepsilon/\delta))})$, which is in the order of $\mathcal{O}(n^2\sqrt{r})$.