

Differentially Private Zeroth-Order Methods for Scalable Large Language Model Fine-tuning

Zhihao Liu
Zhejiang University
zhihao_liu@zju.edu.cn

Yuke Hu
Zhejiang University
yukehu@zju.edu.cn

Jian Lou
Zhejiang University
jian.lou@zju.edu.cn

Bo Li
Chicago University
bol@uchicago.edu

Kui Ren
Zhejiang University
kuiren@zju.edu.cn

Wenjie Bao
Zhejiang University
wenjie_bao@zju.edu.cn

Zhan Qin
Zhejiang University
qinzhao@zju.edu.cn

ABSTRACT

Fine-tuning on task-specific datasets is a widely-embraced paradigm of harnessing the powerful capability of pretrained LLMs for various downstream tasks. Due to the popularity of LLMs fine-tuning and its accompanying privacy concerns, differentially private (DP) fine-tuning of pretrained LLMs has been widely used to safeguarding the privacy of task-specific datasets. Lying at the design core of DP LLM fine-tuning methods is the satisfactory tradeoff among privacy, utility, and scalability. Most existing methods build upon the seminal work of DP-SGD. Despite pushing the scalability of DP-SGD to its limit, DP-SGD-based fine-tuning methods are unfortunately limited by the inherent inefficiency of SGD.

In this paper, we investigate the potential of DP zeroth-order methods for LLM pretraining, which avoids the scalability bottleneck of SGD by approximating the gradient with the more efficient zeroth-order gradient. Rather than treating the zeroth-order method as a drop-in replacement for SGD, this paper presents a comprehensive study both theoretically and empirically. First, we propose the stagewise DP zeroth-order method (DP-ZOSO) that dynamically schedules key hyperparameters. This design is grounded on the synergy between DP random perturbation and the gradient approximation error of the zeroth-order method, and its effect on fine-tuning trajectory. Second, we further enhance the scalability by reducing the trainable parameters that are identified by repurposing a data-free pruning technique requiring no additional data or extra privacy budget. We propose DP zeroth-order stagewise pruning method (DP-ZOPO) with several pruning strategies and undertake a comparative analysis of these strategies. Additionally, we identify and elucidate the superior pruning strategy.

We provide theoretical analysis for both proposed methods. We conduct extensive empirical analysis on both encoder-only masked language model and decoder-only autoregressive language model, achieving impressive results in terms of scalability and utility (compared with DPZero [47], DP-ZOPO improves 4.5% on SST-5, 5.5% on MNLI with RoBERTa-Large and 9.2% on CB, 3.9% on BoolQ with OPT-2.7B when $\epsilon = 4$).

1 INTRODUCTION

Pretrained Large Language Models (LLMs), scaling up to unprecedented sizes, have demonstrated remarkable potential in their capabilities of understanding and processing natural languages with human-like proficiency [32] [33]. This has prompted a rapid surge in demand to harness the power of pretrained LLMs, particularly open-sourced series like OPT [49], llama [39] and GPT [33], to boost performance across a wide range of downstream tasks. Fine-tuning is one of the most fundamental and dominant approaches for adapting pretrained LLMs to specific downstream tasks, which has been proven effective in [52] [37]. Fine-tuning starts with the publicly accessible checkpoint of the selected model and continues to train the model for several epochs based on the task-specific dataset (referred to as the fine-tuning dataset hereafter). While appearing a simple process at first glance, the sheer scale of nowadays LLMs introduces significant scalability issues for fine-tuning, e.g., incurring prohibitive memory footprint, which complicates the design of training methods even in nonprivate fine-tuning [48].

It is widely recognized that a finetuned LLM can leak sensitive information [9] [10] from its fine-tuning dataset, which is considered private and valuable in certain downstream application areas related to finance, healthcare [10]. Therefore, private fine-tuning of pretrained LLMs has become a pressing need due to the escalating privacy concerns associated with the growing popularity of LLMs [44]. Differential privacy (DP) [14] is a widely adopted mathematical framework that ensures a rigorous privacy protection guarantee by introducing calibrated random perturbations. A long-standing research theme in DP is the pursuit of an ideal tradeoff between privacy and utility, which arises because the random perturbation for privacy-preserving purpose will inevitably degrade the utility. DP LLM fine-tuning faces even intensified tension as it has to handle the tradeoff among privacy, utility, and scalability.

Most of the existing DP LLM fine-tuning methods [24] [20] build upon the seminal work of Differentially Private SGD (DP-SGD) [1], which clips the gradient vector per training sample, injects random Gaussian noise after clipping, and tracks the privacy budget across iterations by the moments accountant technique. These DP-SGD-based methods mitigate the scalability issue in roughly two lines of effort. The first line [26] [7] [20] aims to reduce the computational

cost of the per-sample clipping of the gradient vector that is known to incur heavy computational and memory burden. For instance, [26] introduce group clipping to replace the per-sample clip, [8] propose book keep (BK) strategy. The second line [17] incorporates DP-SGD with parameter-efficient fine-tuning techniques in LLMs, which limits the number of trainable parameters. The intuition is that the utility-privacy tradeoff degrades quickly with respect to the number of trainable parameters since the DP perturbations need to be injected into all dimensions of the gradient vector. For instance, [44] [24] consider LoRA, adapters, and compacter, which either train two smaller factors to approximate the LLM parameter update or insert additional trainable modules while freezing the pretrained LLM parameters.

Although the DP-SGD-based methods mentioned above have made the best effort to squeeze performance out of DP-SGD, the inefficiency inherent in SGD continues to hinder achieving a satisfactory “privacy-utility-scalability” tradeoff for LLM fine-tuning. That is, the gradient calculation requires caching all intermediate activations during the forward pass and gradients during the backward pass, which leads to prohibitive memory usage for LLM fine-tuning, i.e., consuming up to $12\times$ the memory usage required for inference [29]. Driven by such limitation, recent nonprivate LLM fine-tuning methods turn to zeroth-order methods, which circumvent the scalability issues associated with gradient computations by approximating them using two inferences [29] [51] [50]. In nonprivate LLM fine-tuning, the effectiveness of zeroth-order methods has been validated through both theoretical and empirical studies, which suggests a promising direction for developing the DP LLM fine-tuning method.

In this paper, we strive to achieve a better “privacy-utility-scalability” tradeoff for DP LLM fine-tuning, through the lens of zeroth-order methods that are previously under-explored in DP literature. We comprehensively investigate DP zeroth-order methods from both theoretical and empirical perspectives, rather than treating the zeroth-order method as a drop-in replacement for SGD.

We further leverage pruning to enhance DP zeroth-order LLM fine-tuning where LLM is updated on a subset of parameters with zeroth-order optimizer. We are motivated by the insights drawn from the following three factors. Firstly, a large proportion of LLM’s parameters are typically redundant, and fine-tuning only a subset of parameters can yield significant improvements in downstream tasks. Secondly, reducing the dimensionality of the parameters to be updated improves the accuracy of zeroth-order gradient estimation. Lastly, since only a subset of parameters is updated with private data, the scale of DP noise introduced will also be reduced. Moreover, we further design multiple pruning strategies to enhance model performance under private settings.

Our contributions. First, we focus on the synergy between the DP random perturbations and the gradient approximation error of zeroth-order estimation to the true gradient, across different training stages. Specifically, we propose to dynamically adjust the ZO scale parameter that controls the gradient approximation error and divide the DP LLM fine-tuning into stages. We propose DP-ZOSO, the first stagewise algorithm for differentially private zeroth-order fine-tuning to our best knowledge. In earlier stages, the gradient approximation error is controlled to be smaller, and

together with a larger learning rate, it allows the LLM fine-tuning to approach the optimum more quickly. In the later stages, we need to deliberately increase the gradient approximation error, together with a decreased learning rate, offering a stabilization effect on the fine-tuning trajectory. Our theoretical analysis demonstrates that this stagewise DP zeroth-order fine-tuning strategy provides an improved convergence rate.

Second, we are the first to propose stagewise DP zeroth-order method in conjunction with reduced trainable parameters (DP-ZOPO). Unlike existing works in DP-SGD, which introduce additional trainable modules to modify the LLM structure, or the LoRA-based method, which still entails a certain number of variables, we propose to initially identify key parameters within the given LLM. Subsequently, we treat these identified parameters as trainable while freezing the remainder. To identify the key parameters, we repurpose a data-free pruning method, which does not necessitate public data or incur additional privacy costs during the pruning stage. In conjunction with the stagewise fine-tuning process, we also propose several dynamic pruning strategies and point out the superior strategy, it is proved that implementing pruning can enhance optimization both theoretically and empirically.

Third, we conduct extensive empirical evaluations to corroborate the superiority of our proposed DP zeroth-order methods for LLM fine-tuning. We utilize four different open-source LLMs, including masked language model (RoBERTa-large) and autoregressive language model (OPT-2.7B) for downstream tasks such as sentiment classification, natural language inference, and topic classification (including datasets like SST-2, SST-5, SNLI, MNLI, and TREC). Our method exhibits strong scalability and performs well across all models and datasets. Our method has enhanced the model’s performance greatly (improves 4.5% on SST-5, 5.5% on MNLI with RoBERTa-Large and 9.2% on CB, 3.9% on BoolQ with OPT-2.7B when $\epsilon = 4$) compared with existing differentially private zeroth-order fine-tuning method.

Our main contributions can be summarized as follows:

- We are the first to conduct a comprehensive investigation of DP LLM fine-tuning from a zeroth-order perspective.
- We propose two novel stagewise DP zeroth-order algorithms DP-ZOSO and DP-ZOPO. DP-ZOSO optimizes the model stagewise on all dimensions while DP-ZOPO utilizes the pruning mask to guide the model towards updating in more important directions. We provide a comprehensive theoretical analysis of privacy and convergence for both of them. DP-ZOPO reduces the total complexity of DP-ZOSO by a factor of $1/r$ theoretically.
- We extensively experiment with our methods, providing empirical evidence of their scalability and utility. DP-ZOPO outperforms parameter-efficient method (DP prefix tuning) by 10.9% on MNLI and 7.6% on TREC when $\epsilon = 4$ with RoBERTa-large. DP-ZOPO also outperforms the zeroth-order method (DPZero) by 5.5% on MNLI and 4.4% on TREC when $\epsilon = 4$ with RoBERTa-large (see Figure 1).

In sum, we are the only work providing both theoretical and empirical studies among all concurrent works, offering more involved theoretical analysis and more comprehensive empirical analysis.

Therefore, we believe our work offers distinct and significant contributions to both fields of DP LLM fine-tuning and DP zeroth-order methods, compared to these concurrent works.

2 BACKGROUND AND RELATED WORK

2.1 Private Parameter Efficient Fine Tuning

As LLMs scale up, full-parameter fine-tuning becomes impractical due to the requirement for extensive GPU memory. Things only get worse with privacy, which leads to overheads in terms of running time, memory usage, and most importantly, accuracy. The magnitude of noise introduced to a model due to DP grows as the model size increases, which can result in poor performance of LLMs. Parameter-efficient fine-tuning methods reduce memory consumption by updating just a fraction of the network parameters. Differentially private prompt tuning [25] [13] has emerged as a simple and parameter-efficient solution for steering LLMs to downstream tasks. DP fine-tuning with Reparametrized Gradient Perturbation [45] [24] computes the low-dimension projected gradient without computing the gradient itself, significantly reducing computation cost. Besides parameter-efficient fine-tuning, there also exist other memory-efficient methods.

2.2 DP Zeroth-Order Fine-Tuning

Zeroth-order fine-tuning methods decrease memory consumption by replacing backpropagation with two inferences [29] [50], each done with the same random perturbation with flipped signs. Recently, there have been several research that explores differential private zeroth-order fine-tuning: a workshop paper [47] and a “work in progress” paper on arXiv [38]. We stress that our work differs from them in three key aspects: 1) Regarding algorithm design, all these works merely consider constant scheduling of hyperparameters and have not explored parameter sparsification to further boost scalability like us; 2) Regarding theoretical analysis: [38] did not provide any theoretical analysis for the utility. [47]’s utility analysis did not consider the dynamic scheduling of the hyperparameters that render the analysis more involved; 3) Regarding empirical analysis: [47] did not provide any empirical studies. While [38] also considered LLM finetuning, our empirical study is more comprehensive by considering both RoBERTa-large and OPT models, along with classification and multiple tasks. Moreover, targeting the sparsity inherent in LLMs, we propose DP-ZO fine-tuning with pruning and achieve better performance both theoretically and empirically.

2.3 DP-SGD with Sparsification

Due to the low-rankness or the sparsity of neural networks, pruning has been widely used as a dimensionality reduction method to improve the scalability of DP-SGD [3] [28] [30]. Adamczewski et. al. [2] proposed parameter freezing pruning method: pre-prune the network and update those selected using DP-SGD. They [3] further propose parameter selection: select which parameters to update at each step of training and update only those selected using DP-SGD. They use public data for freezing or selecting parameters to avoid privacy loss incurring in these steps. However, there has been no study about stagewise pruning method with dynamic pruning rates in differentially private zeroth-order fine-tuning.

3 PRELIMINARY

3.1 LLM Fine-Tuning

LLM fine-tuning has been a popular way of adapting a pre-trained large language model to a specific task or domain by further training it on task-specific data.

Definition 3.1 (LLM Fine-tuning). Fine-tuning a pretrained language model $f(\theta)$ on the dataset $\mathcal{D}$ of downstream task can be described as the following optimization problem:

$$\min_{\theta \in \mathcal{R}^d} \{f(\theta, \mathcal{D})\} := \frac{1}{n} \sum_{i=1}^n f(\theta, x_i), \quad (1)$$

where $x_i \in \mathcal{D}$ is the i -th training sample of the total training dataset $\mathcal{D}$ and n is the number of training samples.

Stochastic Gradient Descent (SGD) is an optimization algorithm commonly used in LLM fine-tuning. It’s a variant of the gradient descent algorithm that’s designed to handle large datasets more efficiently.

Definition 3.2 (Stochastic gradient descent). SGD is a differentially private optimizer with learning rate η that updates parameters as

$$\theta_t = \theta_{t-1} - \eta \cdot \nabla f(\theta_{t-1}, \xi_t), \quad (2)$$

where $\xi_t \in \mathcal{D}$ is the minibatch at time t and $\nabla f(\theta_{t-1}, \xi_t)$ is the average of gradients estimated by back propagation on ξ_t .

3.2 Differential Privacy

Differential privacy aims to provide a mathematical definition for the privacy guarantees of an algorithm or system.

Definition 3.3 (Differential Privacy [15]). A randomized mechanism $\mathcal{A} : \mathcal{X} \rightarrow \mathcal{S}$ is called (ϵ, δ) -differential private with respect to $\mathcal{d}$ if for every pair of adjacent datasets $X, X' \in \mathcal{X}$ satisfying $\mathcal{d}(X, X') \leq 1$ and every subset of outputs $s \in \mathcal{S}$ it holds that:

$$\mathbb{P}[\mathcal{A}(X) \in s] \leq e^\epsilon * \mathbb{P}[\mathcal{A}(X') \in s] + \delta, \quad (3)$$

where $\mathcal{d} : \mathcal{X}^2 \rightarrow [0, \infty)$ be the distance between two datasets.

Differentially Private Stochastic Gradient Descent (DP-SGD) is a privacy-preserving variant of SGD. It is designed to train models while providing strong guarantees of differential privacy.

Definition 3.4 (DP-SGD [1]). DP-SGD is a differentially private optimizer with clipping threshold C , noise scale σ and learning rate η that updates parameters as

$$\theta_t = \theta_{t-1} - \eta \cdot \frac{1}{m} \left(\sum_i \hat{g}(\theta_{t-1}, x_i) + N(0, \sigma^2 C^2 \mathbf{I}_d) \right), \quad (4)$$

where $\hat{g}(\theta_{t-1}, x_i)$ is the gradient clipped by clipping threshold C , $\hat{g}(\theta_{t-1}, x_i) = g(\theta_{t-1}, x_i) / \max(1, \frac{\|g(\theta_{t-1}, x_i)\|_2}{C})$ and $g(\theta_{t-1}, x_i)$ denote the true gradient on data point x_i .

The following three lemmas are frequently used in the privacy analysis of DP-SGD.

LEMMA 3.5 (PRIVACY AMPLIFICATION BY SUBSAMPLING [22]). Let $\mathcal{A} : \mathcal{X}^{r_2} \rightarrow \mathcal{S}$ be a (ϵ, δ) -DP mechanism. Then, the mechanism $\mathcal{A}' : \mathcal{X}^{r_1} \rightarrow \mathcal{S}$ defined by $\mathcal{A}' = \mathcal{A} \circ \text{samp}_{r_1, r_2}$ is (ϵ', δ') -DP, where

$$\epsilon' = \log(1 + q(e^\epsilon - 1)), \delta' = q\delta, q = \frac{r_2}{r_1}. \quad (5)$$

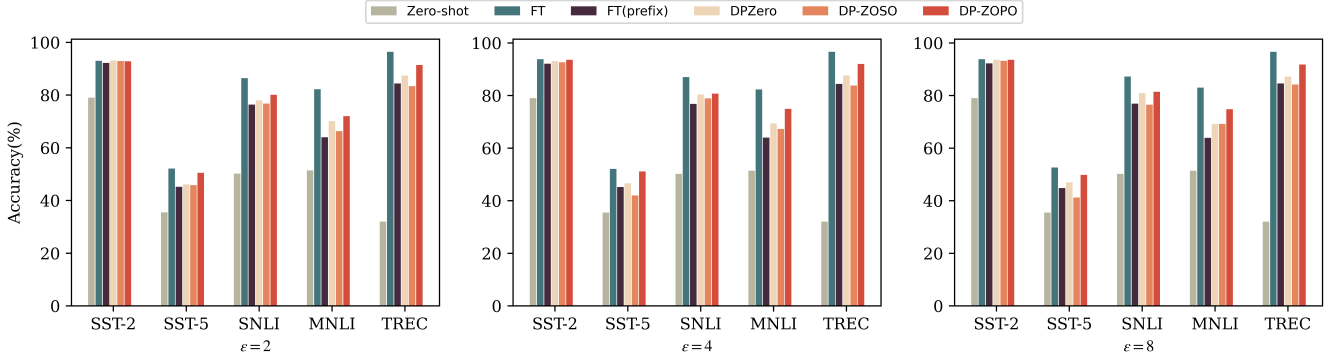


Figure 1: Experiments on RoBERTaA-large. We report zero-shot, DPZero[47], DP-ZOPO, DP-ZOSO and fine-tuning (FT) with full parameter and prefix-tuning. DP-ZOSO and DP-ZOPO both outperform zero-shot, FT (prefix) with much less memory. DP-ZOPO far outperforms DPZero, and DP-ZOSO and approaches FT (Detailed numbers in Table 2).

In particular, when $\epsilon < 1$, $\mathcal{A}'$ is $(\mathcal{O}(q\epsilon), q\delta)$ -DP.

LEMMA 3.6 (STRONG COMPOSITION [16]). Let $\mathcal{A}_1, \dots, \mathcal{A}_T$ be T adaptive (ϵ, δ) -DP mechanisms, where $\epsilon, \delta \geq 0$. Then, for any $\delta' \geq 0$, the composed mechanism $\mathcal{A} = (\mathcal{A}_1, \dots, \mathcal{A}_T)$ is $(\hat{\epsilon}, \hat{\delta})$ -DP, where

$$\hat{\epsilon} = \sqrt{2T \log\left(\frac{1}{\delta'}\right)} \epsilon + T\epsilon(e^\epsilon - 1), \hat{\delta} = T\delta + \delta'. \quad (6)$$

Moments accountant method tracks the accumulation of privacy loss over multiple iterations. By using the moments accountant, it is possible to estimate the overall privacy loss over a sequence of operations, which provides a tighter privacy analysis compared with strong composition Lemma 3.6.

LEMMA 3.7 (MOMENTS ACCOUNTANT [1]). There exist constant c_1 and c_2 so that given the sampling probability $q = m/n$ and the number of steps T , for any $\epsilon < c_1 q^2 T$, DP-SGD is (ϵ, δ) -differentially private for any $\delta > 0$ if we choose

$$\sigma \geq c_2 \frac{q\sqrt{T \log(1/\delta)}}{\epsilon}. \quad (7)$$

3.3 Zeroth-Order Optimization

Zeroth-order gradient estimation does not rely on explicit gradient information to update the parameters of a model. In zeroth-order methods, function evaluations at different points in the parameter space are used to approximate the gradient.

Definition 3.8 (Zeroth-order gradient estimation [36]). Given a model with parameters $\theta \in \mathcal{R}^d$ and a loss function f , SPSS estimates the gradient on a minibatch ξ as

$$\nabla f_\beta(\theta, \xi) = \frac{1}{2\beta} (f(\theta + \beta \mathbf{v}, \xi) - f(\theta - \beta \mathbf{v}, \xi)) \mathbf{v}, \quad (8)$$

where $\mathbf{v} \in \mathcal{R}^d$ with $\mathbf{v} \sim N(0, \mathbf{I}_d)$ and β is the ZO scale parameter.

Gaussian Smoothing: Gaussian smoothing is a well-known technique that converts a possibly non-smooth function to a smooth approximation [31]. Given a differentiable function $f : \mathcal{R}^d \rightarrow \mathcal{R}$ and $\beta \geq 0$, we define the Gaussian smoothing of f as $f_\beta(\theta) = \mathbb{E}_{\mathbf{v} \sim N(0, \mathbf{I}_d)} [f(\theta + \beta \mathbf{v})]$ where $N(0, \mathbf{I}_d)$ is a standard Gaussian distribution. Gaussian smoothing has the following properties.

LEMMA 3.9. Suppose $f : \mathcal{R}^d \rightarrow \mathcal{R}$ is differentiable and L -Lipschitz. Then (i) f_β is L -Lipschitz; (ii) $\|f_\beta(\theta) - f(\theta)\| \leq L\beta\sqrt{d}$; (iii) f_β is differentiable and $\frac{\sqrt{d}L}{\beta}$ -smooth; (iv)

$$\nabla f_\beta(\theta) = \mathbb{E}_{\mathbf{v} \sim N(0, \mathbf{I}_d)} \left[\frac{1}{2\beta} (f(\theta + \beta \mathbf{v}) - f(\theta - \beta \mathbf{v})) \mathbf{v} \right]. \quad (9)$$

Weakly-convex function is a class of “nice-behaved” non-convex objectives that are close to a convex function which is well studied as non-convexity assumption in [46] [5].

Definition 3.10 (Weakly-convex). The loss function $f(\theta)$ is ρ -weakly-convex if for every θ, θ' , there exist constant ρ that satisfies:

$$f(\theta') \geq f(\theta) + \langle \nabla f(\theta), \theta' - \theta \rangle - \frac{\rho}{2} \|\theta - \theta'\|^2 \quad (10)$$

Assumption 1. The function $f_\beta(\theta, x)$ is L -Lipschitz and $\frac{\sqrt{d}L}{\beta}$ -smooth for every x . There exist γ that for every $x \in \mathcal{D}$, we have

$$\mathbb{E} \left[\|\nabla F_\beta(\theta) - \nabla f_\beta(\theta, x)\|^2 \right] \leq \gamma^2. \quad (11)$$

It is notable that many papers have proposed and analyzed deterministic/stochastic optimization algorithms under the PL condition [21] [4].

Assumption 2. The function $f_\beta(\theta, \mathcal{D})$ satisfies μ -PL condition if for any $\theta \in \mathcal{R}^d$ we have

$$\|\theta - \theta^*\|^2 \leq \frac{1}{2\mu} (f_\beta(\theta, \mathcal{D}) - f_\beta(\theta^*, \mathcal{D})). \quad (12)$$

4 METHODOLOGY

In Section 4.1, we introduce differential private zeroth-order fine-tuning and describe our novel DP noise injection method for zeroth-order. In Section 4.2, we propose zeroth-order gradient estimation with dynamic ZO scale parameter. In Section 4.3, we propose DP-ZOSO, the first stagewise DP-ZO fine-tuning method to our knowledge, and provide theoretical analysis for both privacy and convergence. In Section 4.4, we propose DP-ZOPO, the first stagewise DP-ZO fine-tuning method with data-free pruning to guide the model to update on important directions, whose schematic diagram is shown in Figure 2.

4.1 Differential Private Zeroth-Order Fine-Tuning

The zeroth-order gradient is estimated by calculating the difference of the function on two points $\theta + \beta \mathbf{v}$ and $\theta - \beta \mathbf{v}$. Instead of clipping both the loss function $f(\theta + \beta \mathbf{v})$ and $f(\theta - \beta \mathbf{v})$, we clip the difference of the function on two points since the absolute value of the difference is much smaller than the loss function itself ($3.2\text{e-}5$ compared to 0.45 in average).

For privacy guarantee, Gaussian noise is injected into the clipped gradient during training. However, in the zeroth-order gradient descent algorithm, the direction of the approximate gradient $\nabla f_\beta(\theta)$, denoted by $\mathbf{v}$, is sampled from the standard Gaussian distribution independent of the private data.

Compared to adding Gaussian to all dimensions of the gradient in DP-SGD, we propose a novel method of noise injection by adding noise $N(0, \sigma^2 C^2)$ to $\frac{1}{2\beta} (f(\theta + \beta \mathbf{v}) - f(\theta - \beta \mathbf{v}))$, achieving (ϵ, δ) -DP guarantee. By standard post processing, the approximate gradient $\nabla f_\beta(\theta)$ is also (ϵ, δ) -DP since the direction $\mathbf{v}$ is independent of private data.

4.2 Zeroth-Order Gradient Estimation with Dynamic ZO scale parameter

The zeroth-order method suffers from utility loss owing to the bias between the actual gradient and the zeroth-order gradient. Zeroth-order gradient is estimated by calculating the difference of the function on two points $\theta + \beta \mathbf{v}$ and $\theta - \beta \mathbf{v}$ where β is the ZO scale parameter. The choice of ZO scale parameter β has only been seen as a hyperparameter in [29] [19]. Shi et al. [34] pointed out that the ZO scale parameter should be as small as possible but can be set too small in practice since the accuracy of the computing system is limited. However, when the model is close to convergence, the difference of the loss function on two points can be quite small such that the two-point estimated gradient can be quite large when the ZO scale parameter is small, contrary to the nature of the model approaching convergence. Thus, we propose dynamic zeroth-order gradient estimation with an increasing ZO scale parameter, reducing the bias of the zeroth-order gradient. More results are detailed in Table 5.

Furthermore, we find that zeroth-order gradients with bigger ZO scale parameters have. The lower the possibility of zeroth-order gradient getting clipped (detailed in Table 1).

Table 1: Clipping bias with different ZO scale parameters. $\mathbb{P}(\text{Clip})$ denotes the possibility of getting clipped.

β	1e-6	2e-6	4e-6	6e-6
$\mathbb{P}(\text{Clip})$	13.40%	12.20%	11.70%	11.60%

4.3 DP-ZO Stagewise Fine-Tuning

First, let us propose algorithm Differentially Private Zeroth-Order Stagewise Optimizer (DP-ZOSO) in Algorithm 1. At the s -th stage, a regularized function $\phi_s(\theta)$ is constructed that consists of the original objective $f_{\beta_s}(\theta)$ and a quadratic regularizer $\frac{1}{2\lambda} \|\theta - \theta^{s-1}\|_2^2$.

The reference point θ^{s-1} is a returned solution from the previous stage, which is also used for an initial solution for the current stage. Adding the strongly convex regularizer at each stage helps convert the weakly-convex loss function $f_{\beta_s}(\theta)$ to convex or strongly-convex $\phi_s(\theta)$. For each regularized problem, the SGD with a constant stepsize and ZO scale parameter is employed for several iterations with an appropriate returned solution. The procedure is described in Algorithm 2.

Algorithm 1: DP-ZOSO (Differentially Private Zeroth-Order Stagewise Optimizer)

input : Initial point $\theta^0 \in \mathcal{R}^d$, initial stepsize $\eta_0 \geq 0$, initial ZO scale parameter β_0 , regularization parameter λ , initial iteration number T_0 , number of epoch S , dataset $\mathcal{D}$

1 **for** $s = 1$ **to** S **do**

2 $\beta_s = k \cdot \beta_{s-1}$, $T_s = 2 \cdot T_{s-1}$, $\eta_s = \eta_{s-1}/2$, $\lambda = \lambda_{s-1}/2$;

3 $\phi_s(\theta) = f_{\beta_s}(\theta) + \frac{1}{2\lambda} \|\theta - \theta^{s-1}\|_2^2$;

4 $\theta^s = \text{DP-ZOO}(\phi_s(\theta), \theta^{s-1}, \beta_s, T_s, \eta_s, N(0, \mathbf{I}_d), \mathcal{D})$

5 **end**

6 **return** final model parameter θ^S

Algorithm 2 is a Differentially Private Zeroth-Order Optimizer (DP-ZOO) that updates parameters in specific distribution $\mathbb{V}$. In the step t of Algorithm 2, a random set of ξ_t is sampled from dataset $\mathcal{D}$. Random P vectors $\{\mathbf{v}_t^p\}_{p=1}^P \in \mathbb{V}$ are sampled from distribution $\mathbb{V}$. For vector $\mathbf{v}_t^p$, we estimate the gradient step $\text{gra}_{f,t}^{i,p}$ of original objective $f(\theta_{t-1})$ on every data point $\mathbf{x}_t^i$ in line 7 and then clip it to $\hat{\text{gra}}_{f,t}^{i,p}$ in L2-norm. The gradient step of the quadratic regularizer is added to the clipped gradient step as $\text{gra}_t^{i,p}$ in line 9. Gaussian noise is added to the sum of $\text{gra}_t^{i,p}$ and then average. Timing the direction $\mathbf{v}_t^p$ is the gradient $g_p(\theta_{t-1})$ on direction $\mathbf{v}_t^p$. The gradient in iteration t is $g(\theta_{t-1})$ by taking average over P directions. Parameters are then updated by $g(\theta_{t-1})$ with learning rate η_s .

THEOREM 4.1 (PRIVACY ANALYSIS OF DP-ZOSO). *In every stage of Algorithm 1 and in iteration t of DP-ZOO, a random set ξ_t of m samples are sampled out of dataset $\mathcal{D}$ and P random vectors are sampled from standard Gaussian distribution. Gaussian noise is added to the sum of zeroth-order on dataset ξ_t in direction $\mathbf{v}_t^p$ after clipping $|\text{gra}_{f,t}^{i,p}|$ to C . There exist constants c_1 and c_2 , for any $\epsilon < c_1 m^2 T / n^2$, the overall algorithm is (ϵ, δ) -DP if*

$$\sigma^2 \geq \frac{c_2^2 P^2 m^2 T \log(P/\delta)}{\epsilon^2 n^2}. \quad (13)$$

PROOF. In the algorithm, the sensitivity of the sum of ZO estimated $\sum_{i=1}^m \hat{\text{gra}}_{f,t}^{i,p}$ is clipped to C . The zeroth-order gradient on data point $\mathbf{x}_t^i$ is estimated in P random directions such that the privacy budget is divided equally into P . Vectors are randomly sampled from standard Gaussian distribution, thus by post-processing, the privacy of the zeroth-order gradient is guaranteed. $\square$

Algorithm 2: DP-ZOO($\phi_s(\theta), \theta^{s-1}, \beta_s, T_s, \eta_s, \mathbb{V}, \mathcal{D}$)

input : stepsize $\eta_s \geq 0$, ZO scale parameter β_s , parameter dimension d , sample dataset ξ_t , sample size m , clipping threshold C , vector distribution $\mathbb{V}$ and iteration number T_s

- 1 Set initial parameter $\theta_0 = \theta^{s-1}$;
- 2 **for** $t = 1$ to T_s **do**
- 3 Randomly sample $\xi_t = \{x_t^i\}_{i=1}^m$ from dataset $\mathcal{D}$;
- 4 Sample P random vectors $\{\mathbf{v}_t^p\}_{p=1}^P \in \mathbb{V}$;
- 5 **for** $p = 1$ to P **do**
- 6 **for** $i = 1$ to m **do**
- 7 $\text{gra}_{f,t}^{i,p} = \frac{1}{2\beta_s} (f(\theta_{t-1} + \beta_s \mathbf{v}_t^p, x_t^i) - f(\theta_{t-1} - \beta_s \mathbf{v}_t^p, x_t^i))$;
- 8 Clip : $\text{gra}_{f,t}^{i,p} \leftarrow \text{gra}_{f,t}^{i,p} / \max \left\{ 1, \frac{|\text{gra}_{f,t}^{i,p}|}{C} \right\}$;
- 9 $\text{gra}_t^{i,p} = \text{gra}_{f,t}^{i,p} + \frac{1}{\lambda} \|\theta_{t-1} - \theta_0\|$;
- 10 **end**
- 11 $g_p(\theta_{t-1}) = \frac{1}{m} \left(\sum_{i=1}^m \text{gra}_t^{i,p} + N(0, \sigma^2 C^2) \right) \cdot \mathbf{v}_t^p$;
- 12 **end**
- 13 $g(\theta_{t-1}) = \frac{1}{P} \sum_{p=1}^P g_p(\theta_{t-1})$;
- 14 $\theta_t = \theta_{t-1} - \eta_s \cdot g(\theta_{t-1})$;
- 15 **end**
- 16 **return** θ_{T_s}

LEMMA 4.2. In DP-ZOSO, under assumption 1 and the dimension of the parameters to be updated is d with clipping bound C . The error between the gradient $\nabla F_\beta(\theta)$ on the total dataset and the actual update gradient $\nabla \hat{f}_\beta(\theta_t, \xi_{t+1}) = \frac{1}{P} \sum_{p=1}^P \frac{1}{m} \sum_{i=1}^m \text{CLIP} \left(\nabla f_{\beta_s}^p(\theta_t, x_{t+1}^i) \right) + \mathbf{z}_t^p$ is bounded by four terms:

$$\mathbb{E}[\|\nabla F_\beta(\theta_t) - \nabla \hat{f}_\beta(\theta_t, \xi_{t+1})\|^2] \leq \frac{dc_2^2 C^2 PT \log(P/\delta)}{\epsilon^2 n^2} + \frac{8dC^2}{ePm} + \frac{64d\beta_s^2 L^4}{Pm} + \frac{\gamma^2}{Pm}, \quad (14)$$

where the first two terms are due to DP noise and clipping and the last two terms are due to zeroth-order error and data sampling error.

LEMMA 4.3. In DP-ZOSO, under assumption 1 and $f(\theta, x)$ is a ρ -weakly-convex function of θ , by applying DP-ZOO (Algorithm 2) with $\eta_s \leq \frac{\beta_s}{\sqrt{dL}}$, for any $\theta \in \mathcal{R}^d$, we have

$$\mathbb{E}[\phi_s(\theta_{T_s}) - \phi_s(\theta)] \leq \left(\frac{1}{2\eta_s T_s} + \frac{1}{2T_s \lambda} \right) \|\theta_0 - \theta\|^2 + \eta_s \left(\frac{dc_2^2 C^2 PT \log(P/\delta)}{\epsilon^2 n^2} + \frac{64d\beta_s^2 L^4}{Pm} + \frac{\gamma^2}{Pm} + \frac{8dC^2}{ePm} \right). \quad (15)$$

THEOREM 4.4. In DP-ZOSO, suppose assumption 1 holds and the loss function $f(\theta, x)$ is ρ -weakly-convex of θ . Then by setting $\eta_s = \alpha_s \cdot \min \left\{ \frac{Pm}{7\gamma^2}, \frac{Pme}{56dC^2}, \frac{\epsilon^2 n^2}{7dc_2^2 C^2 PT \log(P/\delta)}, \frac{Pm}{448d\beta_s^2 L^4} \right\}$ and $\lambda = \frac{7}{2\mu}, \eta_s T_s = \frac{7}{2\mu}$, after $S = \lceil \log(\alpha_0/\alpha) \rceil$ stages, we have

$$\phi_s(\theta^S) - \phi_s(\theta^*) \leq \alpha. \quad (16)$$

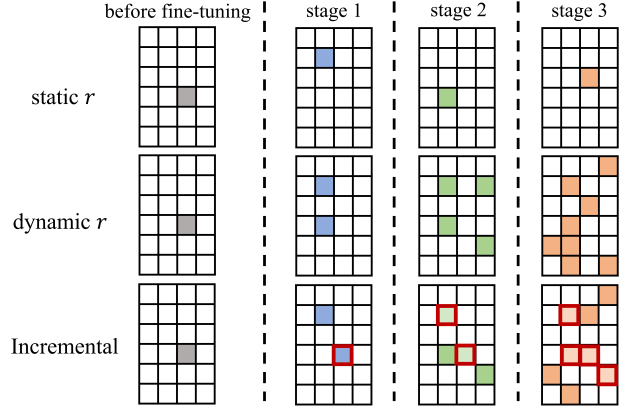


Figure 2: Different pruning strategies of dynamic pruning. The colored squares represent the parameters that need to be fine-tuned in the current stage. In incremental strategy, the highlighted squares indicate the parameters that were fine-tuned in the previous round and will be trained in the current stage.

The total ZO oracle complexity of DP-ZOSO is

$$\mathcal{O} \left(\left(\gamma^2 + dC^2 + d\beta_s^2 L^4 + \frac{dC^2 P^2 m \log(P/\delta)}{\epsilon^2 n^2} \right) \cdot \frac{1}{\mu\alpha} \right). \quad (17)$$

4.4 DP-ZO Stagewise Fine-Tuning with pruning

The success in parameter-efficient fine-tuning motivates us that fine-tune a subset of parameters LLMs can significantly improve the performance on downstream tasks. Pruning helps the model to update on more important directions, especially in the setting of differentially private zeroth-order fine-tuning. Compared with first-order optimizer, zeroth-order optimizer in conjunction with pruning yields greater enhancement in private settings (more details in Section 5.5).

We propose Differentially Private Stagewise Pruning Optimizer (DP-ZOPO) in Algorithm 5. In DP-ZOPO, data-free pruning is employed to guide the model to optimize on more important directions sampled from $\mathbb{V}$. Pruning mask can be calculated at initial and remains static during the whole fine-tuning process, or can be updated during stages with constant or dynamic (increasing) pruning rate. Furthermore, we also propose incremental pruning strategy which is a more conservative pruning strategy.

We define the zeroth-order gradient estimation with pruning $\nabla_{\mathbb{V}} f_\beta(\theta) \stackrel{\text{def}}{=} (\nabla f_\beta(\theta))^{(\mathbb{V})} = \mathbb{E}_{\mathbf{v} \sim \mathbb{V}} \left[\frac{1}{2\beta} (f(\theta + \beta \mathbf{v}) - f(\theta - \beta \mathbf{v})) \mathbf{v} \right]$.

4.4.1 Constant Pruning Mask. Algorithm 3 is a differentially private ZO fine-tuning framework with constant pruning mask calculated at initial. In Algorithm 3, a new loss function $\mathcal{L}(\theta)$ is defined, where $\mathbb{1}$ is the all ones vector and $|\theta^{[l]}|$ is the element-wise absolute value of parameters in the l -th layer. Synaptic saliency scores $\nabla \mathcal{L}_p \odot \theta$ are estimated. The scores will be used in Algorithm 4 to update the vector distribution which helps with convergence.

Parameter-efficient fine-tuning (PEFT) has been widely used in DP fine-tuning, which only updates a fraction of full parameters and can yield comparable performance. In DP fine-tuning, data-free pruning will not cause extra privacy concerns since no private data

Algorithm 3: Data-free ZO pruning($\theta, r, type$)

input : Model parameter θ , pruning rate r and matrix property $type$

- 1 Define $\mathcal{L}(\theta) \leftarrow \mathbb{1}^\top \left(\prod_{l=1}^L |\theta^{[l]}| \right) \mathbb{1}$;
- 2 Initialize binary mask: $mask = \mathbb{1}$;
- 3 Sample P random vectors $\{\mathbf{v}_p\}_{p=1}^P \in N(0, \mathbf{I}_d)$;
- 4 **for** $p = 1$ to P **do**
- 5 $\nabla \mathcal{L}_p = \frac{1}{2\beta} (\mathcal{L}(\theta + \beta \mathbf{v}_p) - \mathcal{L}(\theta - \beta \mathbf{v}_p)) \mathbf{v}_p$
- 6 **end**
- 7 $Score \leftarrow \frac{1}{P} \sum_{p=1}^P \nabla \mathcal{L}_p \odot \theta$;
- 8 **Update vector distribution** $\mathbb{V}$;
- 9 $\mathbb{V} \leftarrow \text{Importance Matrix}(Score, r, type) \cdot N(0, \mathbf{I}_d)$

is used. Pruning freezes most of the pre-trained parameters and modifies the pre-trained model with small trainable parameters.

The pruning mask is a sparse vector due to the pruning rate being set at around 1%. To avoid additional GPU memory overhead, we use `to_sparse()` in NumPy to store the indices and values of non-zero coordinates rather than the entire sparse vector, reducing GPU memory from d to $(d + \log d) * r$.

Pruning can be seen as radically setting the direction of the gradient for frozen parameters to be sampled from $N(0, 0)$. We further propose **Importance Matrix** to guide the training of the remaining parameters. Parameters with high scores are sampled from $N(0, x) (x > 1)$ with larger weight, tuned with more effort. The experiment result is shown in Table 2. We propose rank-based important matrix whose standard deviation follows the uniform distribution from A to B based on the rank of parameters remained.

Algorithm 4: Importance Matrix($Score, r, type$)

input : Previous pruning matrix M_0 , pruning rate r , Matrix type $type$, list $Score$, upper bound A, lower bound B

- 1 Initial importance matrix $M = 0_{[d \times d]}$;
- 2 **if** $type == \text{Incremental}$ **then**
- 3 $Score[i] = -\infty$ if $M_0[i][i] \neq 0$
- 4 **end**
- 5 $Score = Score.sort()$;
- 6 $Score = Score[r \cdot \text{len}(Score) :]$;
- 7 **for** i -th dimension θ_i of parameter θ **do**
- 8 **if** $\theta_i.score \in Score$ **then**
- 9 **if** $type == \text{pruning} - \text{only}$ **then**
- 10 $M[i][i] = 1$
- 11 **end**
- 12 **if** $type == \text{rank} - \text{based}$ **then**
- 13 $rank = Score.index(\theta_i.score)$;
- 14 $M[i][i] = A - \frac{(A-B) \cdot rank}{r \cdot \text{len}(Score)}$
- 15 **end**
- 16 **end**
- 17 **end**
- 18 **return** importance matrix M

4.4.2 Dynamic Pruning Mask. As iterations progress, the direction in which parameters require the most updates keeps changing, making the dynamic pruning mask essential. To tackle the above problem, we propose a dynamic pruning approach with an increasing pruning rate across stages. At the beginning of each stage, the pruning mask is updated based on the score calculated from the current parameters and will remain unchanged throughout the current stage. The pruning rate will increase as the stage progresses, meaning that more parameters will be fine-tuned, further exploiting the potential of the model. In the initial stage, DP zeroth-order fine-tuning with a smaller pruning rate accelerates the model to converge, not only because it reduces the scale of DP noise introduced but also because it diminishes the gradient approximation error as the dimensionality decreases. In the following stages, the continuously increasing pruning rate overcomes the limitation of the static pruning strategy, which can only fine-tune a fixed subset of parameters, failing to explore the full potential of the large language model. In Algorithm 5, the dynamic pruning strategy differs from the static pruning strategy in lines 1 and 3 where the vector distribution is updated during stages.

Algorithm 5: DP-ZOPO (Differentially Private Stagewise Pruning Optimizer)

input : Initial point $\theta^0 \in \mathcal{R}^d$, initial stepsize $\eta_0 \geq 0$, initial ZO scale parameter β_0 , dynamic pruning rate r_s , regularization parameter λ , initial iteration number T_0 , number of epoch S

- 1 $\mathbb{V} \leftarrow \text{Algorithm3}(\theta^0, r, type)$ if $strategy == \text{static}$;
- 2 **for** $s = 1$ to S **do**
- 3 $\mathbb{V}_s \leftarrow \text{Algorithm3}(\theta^{s-1}, r_s, type)$ if $strategy == \text{dynamic}$;
- 4 $\beta_s = k \cdot \beta_{s-1}$, $T_s = 2 \cdot T_{s-1}$, $\eta_s = \eta_{s-1}/2$;
- 5 $\phi_s(\theta) = f_{\beta_s}(\theta) + \frac{1}{2\lambda} \|\theta - \theta^{s-1}\|_2^2$;
- 6 $\theta^s = \text{DP-ZOO}(\phi_s(\theta), \theta^{s-1}, \beta_s, T_s, \eta_s, \mathbb{V}_s, D)$
- 7 **end**
- 8 **return** final model parameter θ^S

Due to the introduction of noise in DP fine-tuning, it becomes challenging for the pruning strategy to find the optimal pruning mask. Furthermore, it is deeply concerning that an unreasonable pruning mask can lead to the catastrophic collapse of the entire model. To overcome the above problem, we further propose an incremental dynamic pruning strategy. Incremental dynamic pruning strategy encourages conservative updates of the pruning mask. The parameters requiring fine-tuning in the current stage consist of two parts of equal quantity. One part comprises the parameters that needed updating from the previous stage, while the other part consists of additional parameters calculated based on the scores for the current stage. However, the incremental dynamic pruning strategy will fine-tune fewer parameters compared to the standard dynamic pruning strategy since the parameters that were fine-tuned in the previous stage are not necessarily selected for fine-tuning in the current stage based on the pruning strategy.

The incremental pruning strategy ensures that the parameters requiring fine-tuning from the previous stage are included in the

current stage by setting their scores to negative infinity which is detailed in lines 2-3 of Algorithm 4.

We present the procedure code of stagewise DP zeroth-order fine tuning with pruning (DP-ZOPO) in Algorithm 5. If *strategy* == *static*, the pruning mask is calculated at initial and fixed for the whole fine-tuning process. If *strategy* == *dynamic*, the pruning mask is dynamically updated for each stage based on the current pruning rate (increasing pruning rate is more optimal). If it is an incremental pruning strategy, the parameters to be updated in the current stage will include parameters that were updated in the previous stage.

DP-ZOPO can be divided into two phases. First, we employ data-free pruning to find the important parameters of θ (the parameters to be fine-tuned). Next, we use a ZO-based fine-tuning method to optimize the model on $\mathbb{V}$. We denote $\nabla_{\mathbb{V}} f_{\beta}(\theta)$ as the zeroth-order gradient on the direction sampled from the distribution $\mathbb{V}$.

THEOREM 4.5 (PRIVACY ANALYSIS OF DP-ZOPO). *Since pruning in every stage of DP-ZOPO does not require private data, the calculated vector distribution $\mathbb{V}$ will not leak any information about private data. Thus, DP-ZOPO shares the same privacy guarantee with Algorithm DP-ZOSO.*

THEOREM 4.6. *In DP-ZOPO, suppose assumption 1 holds and loss function $f(\theta, x)$ is ρ -weakly-convex of θ . The dimensionality of parameters to be updated in stage s is $d \cdot r_s$. Then by setting $\eta_s = \alpha_s \cdot \min \left\{ \frac{Pm}{7\gamma^2}, \frac{Pme}{56dr_s C^2}, \frac{\epsilon^2 n^2}{7dr_s C^2 PT \log(P/\delta)}, \frac{Pm}{448dr_s \beta_S^2 L^4} \right\}$ and $\lambda = \frac{7}{2\mu}$, $\eta_s T_s = \frac{7}{2\mu}$, after $S = \lceil \log(\alpha_0/\alpha) \rceil$ stages, we have*

$$\phi_s(\theta^{r^S}) - \phi_s(\theta^*) \leq \alpha. \quad (18)$$

d The total ZO oracle complexity of DP-ZOPO is

$$\mathcal{O} \left(\left(\gamma^2 + dr_s C^2 + dr_s \beta_S^2 L^4 + \frac{dr_s C^2 P^2 m \log(P/\delta)}{\epsilon^2 n^2} \right) \cdot \frac{1}{\mu \alpha} \right). \quad (19)$$

Remark 1. *Compared to the total complexity of DP-ZOSO in Theorem 4.4, the total complexity of DP-ZOPO is reduced by a factor of r_S theoretically. In the following Section 5, DP-ZOPO is also proven to obtain better convergence results than DP-ZOSO empirically.*

5 EXPERIMENT

We conduct comprehensive experiments on both medium-sized masked LM (RoBERTa-large, 350M [27]) and large autoregressive LM (OPT-2.7B [49]) in few-shot settings. We also explore both full-parameter tuning and parameter-efficient fine-tuning like prefix-tuning.

5.1 Experimental Setup

We report the metric of accuracy on downstream tasks that run with random seed=42. For RoBERTa-large, we conduct experiments under the hyperparameters detailed in Table 11. We use the hyperparameters in Table 12 for our experiments on OPT-2.7B. For privacy concerns, we do not use early stopping in our experiments. The final model is saved and tested on 1000 test examples. **Datasets:** For RoBERTa-large, we consider classification datasets: SST-2[35], SST-5[35], SNLI[6], MNLI[43] and TREC[40]. We randomly sample 1000 examples for testing and have 512 examples per class for

Table 2: Experiments on RoBERTa-large. Our method outperforms zero-shot, DPZero and approaches FT with much less memory.

Task Type	SST-2 —sentiment—	SST-5	SNLI NL inference	MNLI	TREC topic
Zero-shot	79.0	35.5	50.2	51.4	32.0
Small Privacy budget: $\epsilon = 2$					
FT	93.0	52.1	86.4	82.2	96.4
FT-prefix	92.2	45.2	76.4	64.0	84.4
DPZero	93.1	46.1	78.0	70.1	87.4
DP-ZOSO	92.9	45.8	76.8	66.3	83.4
DP-ZOPO	92.8	50.5	80.1	72.0	91.4
Medium Privacy budget: $\epsilon = 4$					
FT	93.8	52.1	87.0	82.3	96.6
FT-prefix	92.1	45.2	76.8	64.0	84.4
DPZero	93.1	46.6	80.4	69.4	87.6
DP-ZOSO	92.6	42.0	78.9	67.3	83.8
DP-ZOPO	93.6	51.1	80.7	74.9	92.0
Large Privacy budget: $\epsilon = 8$					
FT	93.8	52.6	87.2	83.0	96.6
FT-prefix	92.2	44.8	76.9	63.9	84.6
DPZero	93.4	47.0	80.9	69.2	87.2
DP-ZOSO	93.2	41.2	76.5	69.2	84.2
DP-ZOPO	93.6	49.8	81.6	74.8	91.8

both training and validation. For OPT experiments, we consider the SuperGLUE dataset collection[42] including: CB[12], BoolQ[11] and MultiRC[23]. We also include SST-2[35] in our experiments. We randomly sample 1024 examples for training, 500 examples for validation, and 1000 examples for testing.

Models: We conduct experiments on both masked language model (RoBERTa-large, 350M [27]) and autoregressive language model (OPT-2.7B[49]) in few-shot settings. We present our main results in Table 2.3. We include a range of ablation studies, including varying the privacy budget, the pruning strategy and the pruning rate. We also explore both full-parameter tuning and parameter-efficient fine-tuning like prefix-tuning.

Privacy Budgets: We consider various privacy levels with $\epsilon = [2, 4, 8]$ and dynamic $\delta = 1/n$ where n is the number of private data for (ϵ, δ) -DP. We also include $\epsilon = \infty$ baseline that is trained without adding DP noise. In our experiments, we set the clipping threshold to $C = 30$. We include the zero-shot does not incur any privacy loss because we evaluate the pre-trained model directly without fine-tuning on private data.

5.2 Main Results

We conduct experiments with RoBERTa-large on sentiment classification, natural language inference, and topic classification. We follow [29] [18] to study the few-shot and many-shot settings, sampling k examples per class for $k = 512$. DPZero [47] is tested with moment accountant just for fairness. We summarize the results from Table 2.

DP-ZOPO works significantly better than zero-shot and other memory-equivalent methods. On all five diverse tasks, our method can optimize the pre-trained model and consistently perform better than zero-shot and prefix-tuning. We also show for several tasks that DP-ZOPO can outperform DPZero up to 5% even with the same privacy analysis (improve 4.4% on TREC, 5.5% on MNLI with $\epsilon = 4$). DP-ZOPO outperforms 1.6% on SST-2 with $\epsilon = 4$ compared with the results shown in [38].

Pruning improves model performance in private settings. We compare DP-ZOPO (with pruning) and DP-ZOSO (without pruning). We show the performance of DP-ZOPO (static pruning mask) among six different pruning rates in Figure 4. Pruning improves performance greatly in all five tasks in all private settings ($\epsilon = 2, 4, 8$). As a detailed result, the best performance of DP-ZOPO is 74.9% on MNLI, outperforming DP-ZOSO by 7.6% with $\epsilon = 4$. In different private settings, the optimal pruning rate varies which will be discussed in Section 5.6. With the promising results from RoBERTa-large, we extend our method to the OPT-2.7B [49]. We select SuperGLUE tasks [42] (including classification and multiple-choice). We randomly sample 1024 examples for training and 1000 test examples for each dataset.

Table 3: Experiments on OPT-2.7B with privacy budget $\epsilon = 4$, $\delta = 1/1024$. ICL: in-context learning; FT-prefix for prefix-tuning. DP-ZOSO and DP-ZOPO performs better than zero-shot, ICL, and DPZero with almost the same memory consumption.

Task	SST-2	CB	BoolQ	MultiRC
Zero-shot	56.3	50.0	48.0	44.3
ICL	77.6	62.5	57.9	47.8
DPZero	91.5	69.6	63.6	61.9
DP-ZOSO	90.0	62.0	63.7	50.0
DP-ZOPO (static)	93.2	69.7	63.7	61.9
DP-ZOPO (dynamic)	92.5	78.8	67.5	65.0

DP-ZOPO works significantly better than zero-shot and ICL. On both classification and multiple-choice tasks, DP-ZOPO exhibits strong performance. On a 2.7B-parameter scale, DP-ZOPO outperforms zero-shot and ICL across all tasks with equivalent memory consumption (details in Table 3). DP-ZOPO outperforms DPZero [47] in all five tasks on autoregressive language OPT-2.7B, we improve the accuracy by 3.9% on BoolQ and 9.2% on CB with $\epsilon = 4$. We believe that the quality of data-free pruning can greatly influence optimization and we leave how to enhance the performance of pruning as future work.

Dynamic pruning outperforms Static pruning. In dynamic pruning, we select which parameters to update at each step of training and update only those selected using DP-ZOO. We achieve an improvement of 9.1% on CB and 3.8% on BoolQ with OPT-2.7B. Same conclusions are also drawn on the RoBERTa-large in Table 7. Moreover, incremental pruning strategy is superior to other pruning strategies in most scenarios.

5.3 Memory usage

In this section, we profile the memory usage of zero-shot, ICL, FT, FT-prefix, DP-ZOSO and DP-ZOPO. We test OPT-1.3B and OPT-2.7B with Nvidia A100 GPUs (40GB memory) on SST-2 and report the GPU memory consumption in Table 4.

Table 4: GPU memory consumption with OPT-1.3B and OPT-2.7B (fine-tuned on SST-2). DP-ZOPO takes $r = 1\%$ as the constant pruning rate.

GPU memory (MB)	OPT-1.3B	OPT-2.7B
Zero-shot	2517.73	5066.17
ICL	2517.73	5066.17
DP-ZOSO	2517.73	5066.17
DP-ZOPO	2737.03 (1.08 $\times$)	5514.65 (1.08 $\times$)
FT-prefix	2528.67	5080.55
FT	7545.08 (3 $\times$)	20250.93 (4 $\times$)

DP-ZOSO costs the same GPU memory consumption as Zero-shot and ICL which only use inference for updates. DP-ZOPO uses pruning mask (the same dimensionality as the model) to freeze model parameters during training which costs quite large GPU memory. However, the pruning rate is usually set quite small which means the pruning mask is a sparse vector. We store non-zero ids of the mask instead of the whole vector to reduce memory consumption overhead. Pruning costs a slight amount of (8%) GPU memory while improving the model utility greatly.

As shown in Table 4, DP-ZOPO exhibits almost the same (1.08 $\times$) memory consumption which offers memory savings of up to 4 times compared to standard FT. Moreover, if we pre-prune the network and update the remaining parameters using DP-ZOO, the pruning mask can be calculated offline and stored locally, without incurring additional GPU memory overhead.

5.4 Dynamic ZO scale parameter

We fine-tune RoBERTa-large on SST-2 with both fixed and dynamic ZO scale parameter to demonstrate the superiority of dynamic ZO scale parameter. We present the final training loss of the fine-tuned model with different ZO scale parameter in Table 5.

Table 5: Training loss of RoBERTa-large on SST-2 with both fixed and dynamic ZO scale parameter. Dynamic ZO scale parameter is reduced during stages. We denote M1 as the scale parameter decreasing from $1e-5 \sim 1e-4$, M2 as $1e-6 \sim 1e-5$ and M3 as $1e-6 \sim 1e-4$.

Fix ZO scale parameter			Dynamic ZO scale parameter		
1e-4	1e-5	1e-6	M1	M2	M3
0.31532	0.31530	0.31639	0.31812	0.31434	0.31368

All experiments are done under $\epsilon = 4$ with 6k steps. M3 represents the algorithm with ZO scale parameter increasing from $1e-6$ to $1e-4$ as the stage advances. It is shown in Table 5 that dynamic ZO scale parameter method M3 ($1e-6 \sim 1e-4$) exhibits the lowest loss on the training set compared with both fixed ZO scale parameter $1e-4$ and $1e-6$. Dynamic M2 and M3 both perform better than all fixed ZO scale parameter, further elaborating on the superiority of the dynamic ZO scale parameter approach. For the

following experiments of RoBERTa-large, we set the dynamic ZO scale parameter from $1e-6$ to $1e-4$.

5.5 Zeroth-Order Pruning VS First-Order Pruning

In this section, we compare the effectiveness of pruning in both zeroth-order and first-order methods. In first-order method, the gradient is calculated by backpropagation, and only r of the parameters are updated. First-order method in conjunction with pruning reduces the scale of DP noise but also diminishes the model’s utility since only a subset of parameters is fine-tuned.

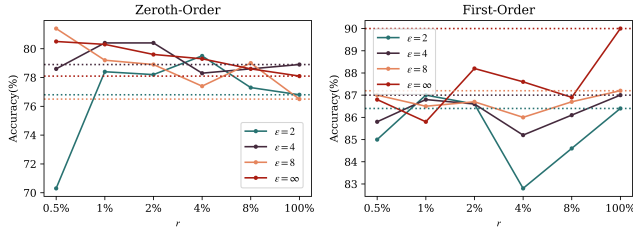


Figure 3: Results of fine-tuned RoBERTa-large on SNLI with zeroth-order and first-order method. In zeroth-order method, fine-tuning with pruning helps with optimization under all private settings.

In Figure 3, we find that in first-order method, pruning helps with fine-tuning when the privacy budget is small and does not significantly harm fine-tuning, only reducing the accuracy by 0.2% with larger privacy budget. However, in zeroth-order method, pruning not only scales down the DP noise but also diminishes the gradient approximation error as the dimensionality decreases. Under all private settings ($\epsilon = 2, 4, 8, \infty$), pruning helps with zeroth-order fine-tuning and improves 2.7% when $\epsilon = 2$ (more detailed numbers in Table 14). The success of the zeroth-order method with pruning motivates us to investigate the impact of different pruning strategies on DP zeroth-order fine-tuning. We will present more experiment results of different pruning strategies in the following section.

5.6 Data-free Pruning with static mask

We first conduct experiments with different learning rates among $\{1e-5, 2e-5, \dots, 1e-4, 2e-4\}$, aiming to find appropriate learning rates for different pruning rates. We conduct our following experiments under appropriate learning rates. We further study the trends in optimal pruning rate selection under different privacy settings $\epsilon = 2, 4, 8, \infty$.

Pruning works well in private ZO fine-tuning. We show our results in Figure 4 that under all private settings, zeroth-order fine-tuning with pruning achieves better performance (above 2%, more detailed numbers in Table 14). Especially when the privacy budget is large like $\epsilon = 8, \infty$, zeroth-order fine-tuning with pruning outperforms without pruning by 4 ~ 5%.

Optimal pruning rate changes with privacy budget. In our experiments, the pruning rate is seen as a hyperparameter. However, the choice of pruning rate affects the model utility on downstream tasks (see Figure 4). When the privacy budget is large (like $\epsilon = 8$), useful parameters can be tuned well, small pruning rate can be

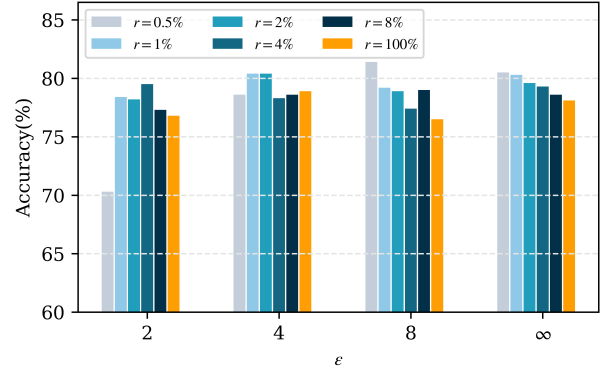


Figure 4: Fine-tuning RoBERTa-large model on dataset SNLI with different pruning rates. Pruning rate r denotes the ratio of parameters to be tuned. Optimal r scales up with the decrease of privacy budget. Detailed numbers in Table 14.

Table 6: Rank-based important matrix with different intervals on RoBERTa-large. [1.0-1.0] performs best with small privacy budget $\epsilon = 2$.

Interval	[0.7-1.3]	[0.8-1.2]	[0.9-1.1]	[1.0-1.0]
Small Privacy budget: $\epsilon = 2$				
SST-2	91.7	92.1	92.4	92.7
SST-5	47.3	48.3	48.1	48.8
SNLI	78.6	78.8	78.9	78.9
MNLI	68.7	70.2	70.1	71.4
TREC	88.4	89.0	88.2	89.6
Medium Privacy budget: $\epsilon = 4$				
SST-2	92.2	93.3	93.2	93.0
SST-5	49.2	49.2	51.1	50.5
SNLI	80.7	80.4	80.1	80.4
MNLI	69.2	71.1	70.6	71.5
TREC	90.2	90.4	88.8	88.2
Large Privacy budget: $\epsilon = 8$				
SST-2	90.8	93.5	93.5	93.2
SST-5	49.6	48.5	49.1	49.8
SNLI	78.7	80.1	78.0	81.4
MNLI	70.8	70.2	69.4	73.1
TREC	89.2	89.4	91.4	90.8

chosen to lower the scale of DP noise discussed in Section 4.4. When privacy budget is small (like $\epsilon = 2$), 0.5% parameters can not be tuned well for large DP noise, thus bigger pruning rate $r = 4\%$ should be chosen for better convergence.

We further conduct experiments on pruning-only method and rank-based important matrix with different settings of intervals (upper bound and lower bound in Algorithm 4). It is shown in Table 6 that rank-based important matrix outperforms pruning-only method under medium and large privacy budgets while behaving poorly under small privacy budgets.

We point out that the optimal interval varies with privacy budget ϵ . Taking the interval [0.8-1.2] as an example, the step of zeroth-order gradient on the most important parameter will be sampled

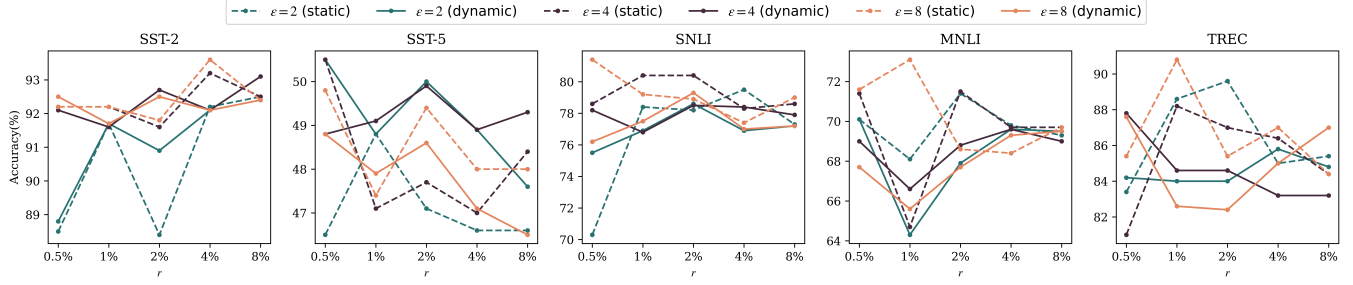


Figure 5: RoBERTa-large on five datasets. Dynamic pruning with a constant pruning rate fails to improve optimization.

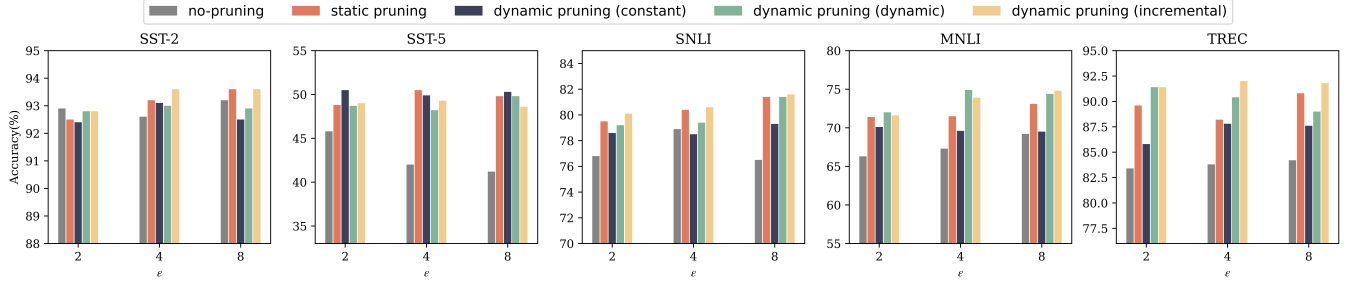


Figure 6: RoBERTa-large on five datasets with different pruning strategies. The incremental pruning strategy is superior to other pruning strategies.

from $N(0, 1.2)$, thus the faster to converge but at the same time the noise introduced is also larger than other parameters in expectation which makes the parameter hard to find its saddle point. Things get extremely worse with small privacy budget. When $\epsilon = 2$, parameters can not be tuned well in the first place, fine-tuning with important matrix makes it even worse. Thus, the optimal interval of important matrix is $[1.0-1.0]$ with small privacy budget $\epsilon = 2$. However, optimal interval under different private settings remains an open question to discuss.

5.7 Data-free Pruning with dynamic mask

In this section, we present two parts of experiments regarding different dynamic pruning strategies. The first part of the experiments follows the pruning strategy where the pruning mask is dynamically updated for each stage, while the pruning rate remains constant. In the second part of the experiments, the pruning mask is dynamically updated for each stage with an increasing pruning rate. We also conduct experiments on the incremental pruning strategy which is a more conservative pruning approach.

Dynamic pruning with constant pruning rate. We compare dynamic and constant pruning strategies under the same (constant) pruning rate. We have found that dynamic updating of the pruning mask does not necessarily lead to an improvement in accuracy with the same (fixed) pruning rate. In Table 7, we present the best accuracy results of both pruning strategies across all pruning rates. Figure 5 shows the accuracy under static and dynamic pruning strategy (constant pruning rate) varying pruning rate on five datasets.

We blame the failure on two main reasons: Firstly, in dynamic pruning, the parameters that were fine-tuned in the previous stage were not fully optimized and may not be selected for further fine-tuning in the next stage. As a result, the model may not be well-optimized. Secondly, the introduction of DP noise leads to the low quality of the update of the pruning mask. In Table 8, we show

Table 7: Experiments on RoBERTa-large (350M parameters). Dynamic pruning with a constant pruning rate fails to improve optimization.

Task	SST-2	SST-5	SNLI	MNLI	TREC
Small Privacy budget: $\epsilon = 2$					
static mask	92.5	48.8	79.5	71.4	89.6
dynamic mask	92.4 ↓	50.5 ↑	78.6 ↓	70.1 ↓	85.8 ↓
Medium Privacy budget: $\epsilon = 4$					
static mask	93.2	50.5	80.4	71.5	88.2
dynamic mask	93.1 ↓	49.9 ↓	78.5 ↓	69.6 ↓	87.8 ↓
Large Privacy budget: $\epsilon = 8$					
static mask	93.6	49.8	81.4	73.1	90.8
dynamic mask	92.5 ↓	50.3 ↑	79.3 ↓	69.5 ↓	87.6 ↓

the proportion of the parameters fine-tuned in the current stage that are selected for fine-tuning in the next stage. Additionally, we provide the proportion of all parameters involved in the fine-tuning process compared to the total number of model parameters and the final accuracy of the model.

It is shown in Table 8 that only a minuscule proportion of parameters are selected for fine-tuning in the next stage. For example, in the first stage of 0.5-0.5-0.5%, only 0.54% of the parameters that are updated will continue to be fine-tuned in the next stage, which results in parameters not being well-tuned and they will not be fine-tuned furthermore in the following fine-tuning process. This explains why dynamic pruning with a constant rate may diminish the utility of the model compared to static pruning. However, the above issue is effectively addressed with dynamic (increasing) pruning rate and incremental strategy, which further enhance the model’s accuracy. As shown in Table 8, dynamic pruning with increasing dynamic $r=1-2-4\%$ outperforms dynamic pruning with

Table 8: RoBERTa-large on SNLI with $\epsilon = 4$. “Stage 1” denotes the proportion of parameters fine-tuned in stage 1 that will be fine-tuned in the next stage (stage 2), “Total” denotes the proportion of parameters fine-tuned in the whole fine-tuning process, and “Acc” denotes the accuracy of the model. “1-2-4%” * denotes the incremental pruning strategy with increasing pruning rate 1-2-4%.

Method	Type	Stage 1	Stage 2	Total	Acc
constant r	0.5-0.5-0.5%	0.54%	0.86%	0.75%	78.2%
	1-1-1%	1.74%	2.56%	1.47%	76.8%
	2-2-2%	3.30%	4.79%	2.89%	78.5%
	4-4-4%	8.98%	11.11%	5.48%	78.4%
	8-8-8%	17.60%	19.13%	10.15%	77.9%
dynamic r	0.5-1-2%	1.01%	2.86%	1.73%	79.2%
	0.5-1-2%*	100%	100%	2%	78.4%
	1-2-4%	3.19%	8.27%	3.38%	79.4%
	1-2-4%*	100%	100%	4%	80.1%
	2-4-8%	6.00%	15.00%	6.55%	78.7%
	2-4-8%*	100%	100%	8%	80.6%

Table 9: Experiments of four pruning strategies on RoBERTAa-large. All the results are the best result across all pruning rates. “static*” denotes the best result when the pruning mask is calculated before fine-tuning and remains static during the whole fine-tuning process. “dynamic(constant r)” and “dynamic(dynamic r)” denote the dynamic pruning strategy with constant and dynamic pruning rate.

Task	SST-2	SST-5	SNLI	MNLI	TREC
Small Privacy budget: $\epsilon = 2$					
static*	92.5	48.8	79.5	71.4	89.6
dynamic(constant r)	92.4	50.5	78.6	70.1	85.8
dynamic(dynamic r)	92.8	48.7	79.2	72.0	91.4
dynamic(incremental)	92.8	49.0	80.1	71.6	91.4
Medium Privacy budget: $\epsilon = 4$					
static*	93.2	50.5	80.4	71.5	88.2
dynamic(constant r)	93.1	49.9	78.5	69.6	87.8
dynamic(dynamic r)	93.0	48.2	79.4	74.9	90.4
dynamic(incremental)	93.6	49.3	80.6	73.9	92.0
Large Privacy budget: $\epsilon = 8$					
static*	93.6	49.8	81.4	73.1	90.8
dynamic(constant r)	92.5	50.3	79.3	69.5	87.6
dynamic(dynamic r)	92.9	49.8	81.4	74.4	89.0
dynamic(incremental)	93.6	48.6	81.6	74.8	91.8

constant pruning rate of all three $r=1\%$, 2% , and 4% (improving accuracy by 2.6%, 0.9% and 1%). Incremental dynamic pruning strategy further improves the model accuracy by 0.7%.

Dynamic pruning with dynamic pruning rate and incremental strategy. We compare the best results of all four pruning strategies to explore the potential of different pruning strategies in Table 9. As shown in Figure 6, dynamic pruning with dynamic (increasing) pruning rate outperforms dynamic pruning with constant pruning rate. For example, on SNLI with $\epsilon = 4$, “1-2-4%” outperforms “1-1-1%” by 2.6% for the reason that “1-2-4%” fine-tunes more

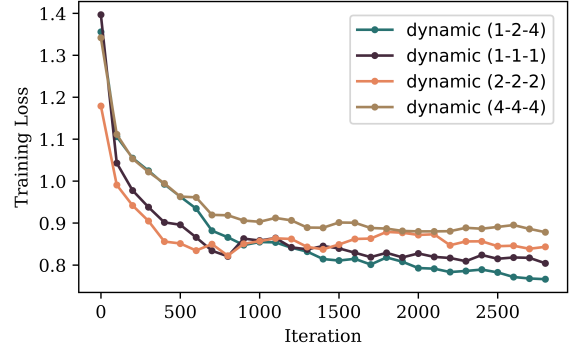


Figure 7: The training-set loss of RoBERTa-large on SNLI with privacy budget $\epsilon = 4$. Dynamic pruning with increasing pruning rate “1-2-4%” performs better than both “1-1-1%” and “4-4-4%”.

Table 10: Experiments on RoBERTAa-large. 1-2-4% * and 1-2-4% denote the incremental pruning and dynamic pruning strategy separately, both with dynamic pruning rate r increasing from 1% to 2% and 4%. Incremental pruning strategy outperforms dynamic pruning strategy in most dataset under all private settings.

Task	SST-2	SST-5	SNLI	MNLI	TREC
Small Privacy budget: $\epsilon = 2$					
0.5-1-2%	89.4	48.7	79.2	71.3	86.6
0.5-1-2% *	88.5	49.0	76.7	71.6	88.4
1-2-4%	92.8	48.2	77.6	71.3	85.2
1-2-4% *	92.2	47.7	80.1	70.7	91.2
2-4-8%	89.6	46.4	75.3	72.0	91.4
2-4-8% *	92.8	48.6	77.5	70.6	91.4
Medium Privacy budget: $\epsilon = 4$					
0.5-1-2%	93.0	48.0	79.2	71.3	86.6
0.5-1-2% *	93.6	49.1	80.0	73.1	90.4
1-2-4%	92.8	48.2	79.4	71.2	87.4
1-2-4% *	92.9	49.3	80.1	71.6	92.0
2-4-8%	92.0	47.9	78.7	74.9	88.6
2-4-8% *	92.8	48.6	80.6	73.9	88.6
Large Privacy budget: $\epsilon = 8$					
0.5-1-2%	92.2	49.8	80.8	71.9	88.0
0.5-1-2% *	93.6	48.5	79.2	74.8	89.2
1-2-4%	92.5	48.8	79.3	72.0	89.0
1-2-4% *	92.9	48.1	79.4	72.4	88.0
2-4-8%	92.9	46.0	81.4	74.4	89.0
2-4-8% *	92.8	48.6	81.6	74.8	91.8

parameters (1.47% compared with 3.38%) in the whole fine-tuning stage. Compared with “4-4-4%”, “1-2-4%” fine-tunes fewer parameters (5.48% compared with 3.38%) and outperforms by 1.0% for the reason that fine-tuning with a smaller pruning rate in the early stages can accelerate the convergence. We show in Figure 7 the training-set loss of “1-2-4%”, “1-1-1%”, “2-2-2%” and “4-4-4%”. “1-1-1%” converges faster in the initial stages compared to “1-2-4%”. However, since “1-2-4%” fine-tunes more parameters, it achieves better convergence with lower training-set loss.

Incremental pruning strategy is a more conservative pruning strategy, it inherits the advantages of the static pruning method, fine-tuning the parameters calculated before DP fine-tuning in every stage. For example in Table 10, “2-4-8%*” outperforms static pruning with $r = 2\%$ by 0.2% since more parameters are fine-tuned. Compared with static pruning with $r = 8\%$, “2-4-8%*” achieves a 2% improvement for fast convergence in the early stage. In summary, incremental pruning strategy performs better in most scenarios when it comes to private zeroth-order fine-tuning. We further present our experimental results on OPT-2.7B in Table 15.

6 CONCLUSION

We have provided both theoretical and empirical studies among all concurrent works. Detailed theoretical proof of both privacy and convergence for both our algorithms are provided. We have shown that the stagewise DP zeroth-order method with pruning is memory-efficient and can effectively optimize large LMs across many tasks and scales. Further experiments suggest that the optimal pruning rate varies with the privacy budget ϵ . As a limitation, we did not explore the optimal selection of pruning rates for different tasks and different models, which may be regarded as hyperparameter fine-tuning. Moreover, the adaptive choice of optimal pruning rate across stages and the number of stages is worth studying in the future, as are more effective pruning algorithms and pruning strategies.

REFERENCES

- [1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. 308–318.
- [2] Kamil Adamczewski, Yingchen He, and Mijung Park. 2023. Pre-Pruning and Gradient-Dropping Improve Differentially Private Image Classification. *arXiv preprint arXiv:2306.11754* (2023).
- [3] Kamil Adamczewski and Mijung Park. 2023. Differential privacy meets neural network pruning. *arXiv preprint arXiv:2303.04612* (2023).
- [4] Raef Bassily, Mikhail Belkin, and Siyuan Ma. 2018. On exponential convergence of sgd in non-convex over-parametrized learning. *arXiv preprint arXiv:1811.02564* (2018).
- [5] Raef Bassily, Cristóbal Guzmán, and Michael Menart. 2021. Differentially private stochastic optimization: New results in convex and non-convex settings. *Advances in Neural Information Processing Systems* 34 (2021), 9317–9329.
- [6] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326* (2015).
- [7] Zhiqi Bu, Yu-Xiang Wang, Sheng Zha, and George Karypis. 2022. Automatic clipping: Differentially private deep learning made easier and stronger. *arXiv preprint arXiv:2206.07136* (2022).
- [8] Mark Bun, Jonathan Ullman, and Salil Vadhan. 2014. Fingerprinting codes and the price of approximate differential privacy. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*. 1–10.
- [9] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*. 267–284.
- [10] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*. 2633–2650.
- [11] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044* (2019).
- [12] Marie-Catherine De Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The commitmentbank: Investigating projection in naturally occurring discourse. In *proceedings of Sinn und Bedeutung*, Vol. 23. 107–124.
- [13] Haonan Duan, Adam Dziedzic, Nicolas Papernot, and Franziska Boenisch. 2023. Flocks of Stochastic Parrots: Differentially Private Prompt Learning for Large Language Models. *arXiv preprint arXiv:2305.15594* (2023).
- [14] Cynthia Dwork. 2006. Differential privacy. In *International colloquium on automata, languages, and programming*. Springer, 1–12.
- [15] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*. Springer, 265–284.
- [16] Cynthia Dwork, Aaron Roth, et al. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science* 9, 3–4 (2014), 211–407.
- [17] Jun Feng, Laurence T Yang, Bocheng Ren, Deqing Zou, Mianxiong Dong, and Shunli Zhang. 2023. Tensor recurrent neural network with differential privacy. *IEEE Trans. Comput.* (2023).
- [18] Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723* (2020).
- [19] Saeed Ghadimi and Guanghui Lan. 2013. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization* 23, 4 (2013), 2341–2368.
- [20] Jiyan He, Xuechen Li, Da Yu, Huishuai Zhang, Janardhan Kulkarni, Yin Tat Lee, Arturs Backurs, Nenghai Yu, and Jiang Bian. 2022. Exploring the limits of differentially private deep learning with group-wise clipping. *arXiv preprint arXiv:2212.01539* (2022).
- [21] Hamed Karimi, Julie Nutini, and Mark Schmidt. 2016. Linear convergence of gradient and proximal-gradient methods under the polyak-lojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19–23, 2016, Proceedings, Part I* 16. Springer, 795–811.
- [22] Shiva Prasad Kasiviswanathan, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. 2011. What can we learn privately? *SIAM J. Comput.* 40, 3 (2011), 793–826.
- [23] Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 252–262.
- [24] Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. 2021. Large language models can be strong differentially private learners. *arXiv preprint arXiv:2110.05679* (2021).
- [25] Yansong Li, Zhixing Tan, and Yang Liu. 2023. Privacy-preserving prompt tuning for large language model services. *arXiv preprint arXiv:2305.06212* (2023).
- [26] Haolin Liu, Chenyu Li, Bochao Liu, Pengju Wang, Shiming Ge, and Weiping Wang. 2021. Differentially private learning with grouped gradient clipping. In *ACM Multimedia Asia*. 1–7.
- [27] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).
- [28] Zelun Luo, Daniel J Wu, Ehsan Adeli, and Li Fei-Fei. 2021. Scalable differential privacy with sparse network finetuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5059–5068.
- [29] Sadhika Malladi, Tianyu Gao, Eshaan Nichani, Alex Damian, Jason D Lee, Danqi Chen, and Sanjeev Arora. 2023. Fine-Tuning Language Models with Just Forward Passes. *arXiv preprint arXiv:2305.17333* (2023).
- [30] Fatemehsadat Mireshghallah, Arturs Backurs, Huseyin A Inan, Lukas Wutschitz, and Janardhan Kulkarni. 2022. Differentially private model compression. *Advances in Neural Information Processing Systems* 35 (2022), 29468–29483.
- [31] Yurii Nesterov and Vladimir Spokoiny. 2017. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics* 17 (2017), 527–566.
- [32] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. (2018).
- [33] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.
- [34] Wanli Shi, Hongchang Gao, and Bin Gu. 2022. Gradient-Free Method for Heavily Constrained Nonconvex Optimization. In *International Conference on Machine Learning*. PMLR, 19935–19955.
- [35] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*. 1631–1642.
- [36] James C Spall. 1992. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE transactions on automatic control* 37, 3 (1992), 332–341.
- [37] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification?. In *Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings* 18. Springer, 194–206.
- [38] Xinyu Tang, Ashwinee Panda, Milad Nasr, Saeed Mahloujifar, and Prateek Mittal. 2024. Private Fine-tuning of Large Language Models with Zeroth-order Optimization. *arXiv preprint arXiv:2401.04343* (2024).

- [39] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
- [40] Ellen M Voorhees and Dawn M Tice. 2000. Building a question answering test collection. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*. 200–207.
- [41] Martin J Wainwright. 2019. *High-dimensional statistics: A non-asymptotic viewpoint*. Vol. 48. Cambridge university press.
- [42] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems* 32 (2019).
- [43] Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426* (2017).
- [44] Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, et al. 2021. Differentially private fine-tuning of language models. *arXiv preprint arXiv:2110.06500* (2021).
- [45] Da Yu, Huishuai Zhang, Wei Chen, Jian Yin, and Tie-Yan Liu. 2021. Large scale private learning via low-rank reparametrization. In *International Conference on Machine Learning*. PMLR, 12208–12218.
- [46] Zhuoning Yuan, Yan Yan, Rong Jin, and Tianbao Yang. 2019. Stagewise training accelerates convergence of testing error over SGD. *Advances in Neural Information Processing Systems* 32 (2019).
- [47] Liang Zhang, Kiran Koshy Thekumparampil, Sewoong Oh, and Niao He. 2023. DPZero: Dimension-Independent and Differentially Private Zeroth-Order Optimization. *International Workshop on Federated Learning in the Age of Foundation Models in Conjunction with NeurIPS 2023* (2023).
- [48] Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. 2023. Llama-adaptor: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199* (2023).
- [49] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068* (2022).
- [50] Yihua Zhang, Pingzhi Li, Junyuan Hong, Jiaxiang Li, Yimeng Zhang, Wenqing Zheng, Pin-Yu Chen, Jason D Lee, Wotao Yin, Mingyi Hong, et al. 2024. Revisiting Zeroth-Order Optimization for Memory-Efficient LLM Fine-Tuning: A Benchmark. *arXiv preprint arXiv:2402.11592* (2024).
- [51] Yanjun Zhao, Sizhe Dang, Haishan Ye, Guang Dai, Yi Qian, and Ivor W Tsang. 2024. Second-Order Fine-Tuning without Pain for LLMs: A Hessian Informed Zeroth-Order Optimizer. *arXiv preprint arXiv:2402.15173* (2024).
- [52] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593* (2019).

A APPENDIX

In Sec A.1, we provide the proof for Lemma 4.2 and the oracle complexity of both DP-ZOSO and DP-ZOPO. In Sec A.2, we will provide the experiment settings and more experiment results.

A.1 Complete Proof

In this section, we will provide detailed proof for Lemma 4.2, 4.3 and Theorem 4.4, 4.6.

LEMMA A.1 (RESTATEMENT OF LEMMA 4.2). *In DP-ZOSO, under assumption 1 and the dimension of the parameters to be updated is d with clipping bound C . The error between the gradient $\nabla F_{\beta}(\theta)$ on the total dataset and the actual update gradient $\nabla \hat{f}_{\beta}(\theta_t, \xi_{t+1}) = \frac{1}{P} \sum_{p=1}^P \frac{1}{m} \sum_{i=1}^m \text{CLIP}(\nabla f_{\beta_s}^p(\theta_t, x_{t+1}^i)) + \mathbf{z}_t^p$ is bounded by:*

$$\mathbb{E}[\|\nabla F_{\beta}(\theta_t) - \nabla \hat{f}_{\beta}(\theta_t, \xi_{t+1})\|^2] \leq \frac{dc_2^2 C^2 PT \log(P/\delta)}{\epsilon^2 n^2} + \frac{8dC^2}{ePm} + \frac{64d\beta^2 L^4}{Pm} + \frac{\gamma^2}{Pm}. \quad (20)$$

PROOF OF LEMMA 4.2. Due to the introduction of DP noise, we have

$$\begin{aligned} & \mathbb{E} \left[\left\| \nabla F_{\beta_s}(\theta_t) - \nabla \hat{f}_{\beta_s}(\theta_t, \xi_{t+1}) \right\|^2 \right] \\ &= \mathbb{E} \left[\left\| \nabla F_{\beta_s}(\theta_t) - \frac{1}{P} \sum_{p=1}^P \frac{1}{m} \sum_{i=1}^m \text{CLIP}(\nabla f_{\beta_s}^p(\theta_t, x_{t+1}^i)) + \mathbf{z}_t^p \right\|^2 \right] \\ &= \frac{1}{P^2} \sum_{p=1}^P \mathbb{E} \left[\left\| \nabla F_{\beta_s}(\theta_t) - \frac{1}{m} \sum_{i=1}^m \text{CLIP}(\nabla f_{\beta_s}^p(\theta_t, x_{t+1}^i)) \right\|^2 \right] \\ &\quad + \frac{dc_2^2 C^2 PT \log(P/\delta)}{\epsilon^2 n^2} \\ &= \left\| \mathbb{E}_{\mathbf{v}} [\nabla F_{\beta_s}(\theta_t)] - \frac{1}{P} \sum_{p=1}^P \frac{1}{m} \sum_{i=1}^m \text{CLIP}(\nabla f_{\beta_s}^p(\theta_t, x_{t+1}^i)) \right\|^2 \\ &\quad + \frac{dc_2^2 C^2 PT \log(P/\delta)}{\epsilon^2 n^2}. \end{aligned} \quad (21)$$

Using the elementary inequality $(a - b)^2 \leq 2a^2 + 2b^2$, we have

$$\begin{aligned} & \mathbb{E} \left[\left\| \frac{1}{2\beta} (f(\theta_{t-1} + \beta \mathbf{v}, x) - f(\theta_{t-1} - \beta \mathbf{v}, x)) \right\|^2 \right] \\ &= \frac{1}{4\beta^2} \mathbb{E}[\|f(\theta_{t-1} + \beta \mathbf{v}, x) - \mathbb{E}_{\mathbf{v}}[f(\theta_{t-1} + \beta \mathbf{v}, x)]\|^2] \\ &\quad + \mathbb{E}_{\mathbf{v}}[f(\theta_{t-1} + \beta \mathbf{v}, x) - f(\theta_{t-1} - \beta \mathbf{v}, x)]^2 \\ &\leq \frac{1}{2\beta^2} \mathbb{E}[\|f(\theta_{t-1} + \beta \mathbf{v}, x) - \mathbb{E}_{\mathbf{v}}[f(\theta_{t-1} + \beta \mathbf{v}, x)]\|^2] \\ &\quad + \frac{1}{2\beta^2} \mathbb{E}[\|\mathbb{E}_{\mathbf{v}}[f(\theta_{t-1} + \beta \mathbf{v}, x)] - f(\theta_{t-1} - \beta \mathbf{v}, x)\|^2]. \end{aligned} \quad (22)$$

Since $\mathbf{v}$ has a symmetric distribution around the origin, we have

$$\begin{aligned} & \mathbb{E}[\|f(\theta_{t-1} + \beta \mathbf{v}, x) - \mathbb{E}_{\mathbf{v}}[f(\theta_{t-1} + \beta \mathbf{v}, x)]\|^2] \\ &= \mathbb{E}[\|f(\theta_{t-1} - \beta \mathbf{v}, x) - \mathbb{E}_{\mathbf{v}}[f(\theta_{t-1} + \beta \mathbf{v}, x)]\|^2]. \end{aligned} \quad (23)$$

Define $h(\theta) = f(\theta + \beta \mathbf{v})$. Since f is L -Lipschitz and $\mathbf{v} \in \mathcal{R}^d$ is sampled from standard Gaussian distribution. Then, by [[41] Proposition 3.2], we have

$$\mathbb{P}(|h(\theta) - \mathbb{E}[h(\theta)]| \geq c) \leq 2e^{-\frac{c^2}{2\beta^2 L^2}}. \quad (24)$$

Then, we have

$$\begin{aligned} & \mathbb{E}[(h(\theta) - \mathbb{E}[h(\theta)])^2] = \int_0^{+\infty} \mathbb{P}(|h(\theta) - \mathbb{E}[h(\theta)]| \geq \sqrt{c}) dc \\ &= \int_0^{+\infty} \mathbb{P}(|h(\theta) - \mathbb{E}[h(\theta)]| \geq \sqrt{c}) dc \leq 2 \int_0^{+\infty} e^{-\frac{c}{2\beta^2 L^2}} dc \\ &= 4\beta^2 L^2. \end{aligned} \quad (25)$$

By the definition of h , we have

$$\mathbb{E} \left[\left\| \frac{1}{2\beta} (f(\theta_{t-1} + \beta \mathbf{v}, x) - f(\theta_{t-1} - \beta \mathbf{v}, x)) \right\|^2 \right] \leq 4L^4. \quad (26)$$

Furthermore, we can obtain that

$$\begin{aligned} & \mathbb{E} \left[\left\| \nabla f_{\beta}(\theta) - \mathbb{E}_{\mathbf{v}} [\nabla f_{\beta}(\theta)] \right\|^2 \right] \\ & \leq 2d \mathbb{E} \left[\left\| \frac{1}{2\beta} (f(\theta + \beta \mathbf{v}) - \mathbb{E}_{\mathbf{v}} [f(\theta + \beta \mathbf{v})]) \right\|^2 \right] \\ & \quad + 2d \mathbb{E} \left[\left\| \frac{1}{2\beta} (f(\theta - \beta \mathbf{v}) - \mathbb{E}_{\mathbf{v}} [f(\theta - \beta \mathbf{v})]) \right\|^2 \right] \leq 64d\beta^2 L^4. \end{aligned} \quad (27)$$

Following assumption 1 that $\mathbb{E} [\left\| \nabla F_{\beta}(\theta) - \nabla f_{\beta}(\theta, x) \right\|^2] \leq \gamma^2$

$$\begin{aligned} & \left\| \mathbb{E}_{\mathbf{v}} [\nabla F_{\beta_s}(\theta_t)] - \frac{1}{P} \sum_{p=1}^P \frac{1}{m} \sum_{i=1}^m \text{CLIP} \left(\nabla f_{\beta_s}^p(\theta_t, x_{t+1}^i) \right) \right\|^2 \\ & \leq \frac{1}{P^2 m^2} \sum_{p=1}^P \sum_{i=1}^m \left\| \nabla F_{\beta_s}^p(\theta_t) - \text{CLIP} \left(\nabla f_{\beta_s}^p(\theta_t, x_{t+1}^i) \right) \right\|^2 \\ & \quad + \frac{64d\beta_s^2 L^4}{Pm} \\ & \leq \frac{1}{P^2 m^2} \sum_{p=1}^P \sum_{i=1}^m \left\| \nabla f_{\beta_s}^p(\theta_t, x_{t+1}^i) - \text{CLIP} \left(\nabla f_{\beta_s}^p(\theta_t, x_{t+1}^i) \right) \right\|^2 \\ & \quad + \frac{64d\beta_s^2 L^4}{Pm} + \frac{\gamma^2}{Pm}. \end{aligned} \quad (28)$$

Furthermore, we can bound the possibility of clipping happens

$$\mathbb{P} \left(\frac{1}{2\beta} |f(\theta_{t-1} + \beta \mathbf{v}, x) - f(\theta_{t-1} - \beta \mathbf{v}, x)| \geq C \right) \leq 2e^{-\frac{C^2}{2L^2}}. \quad (29)$$

We define Q_{t+1}^i to be the event that the clipping does not happen at iteration $t+1$ for sample x_{t+1}^i , and $\overline{Q_{t+1}^i}$ to be the event that the clipping does happen. When event Q_{t+1}^i happens, clipping does not happen at iteration $t+1$ for sample x_{t+1}^i such that $\left\| \nabla f_{\beta_s}(\theta_t, x_{t+1}^i) - \text{CLIP}(\nabla f_{\beta_s}(\theta_t, x_{t+1}^i)) \right\|^2 = 0$, we can obtain that

$$\begin{aligned} & \mathbb{E} \left[\left\| \nabla f_{\beta_s}(\theta_t, x_{t+1}^i) - \text{CLIP} \left(\nabla f_{\beta_s}(\theta_t, x_{t+1}^i) \right) \right\|^2 \right] \\ & = \left\| \nabla f_{\beta_s}(\theta_t, x_{t+1}^i) - C \right\|^2 \mathbb{P}(\overline{Q_{t+1}^i}) \\ & \leq (2dC^2 + 4dL^2) \cdot \exp \left(-\frac{C^2}{2L^2} \right). \end{aligned} \quad (30)$$

By setting $C^2 = 2L^2$, we have

$$\mathbb{E} \left[\left\| \nabla f_{\beta_s}(\theta_t, x_{t+1}^i) - \text{CLIP} \left(\nabla f_{\beta_s}(\theta_t, x_{t+1}^i) \right) \right\|^2 \right] \leq \frac{8dC^2}{e}. \quad (31)$$

Combining inequality 21, 28 and 31, we can bound the following term

$$\begin{aligned} & \mathbb{E} [\left\| \nabla F_{\beta_s}(\theta_t) - \nabla \hat{f}_{\beta_s}(\theta_t, \xi_{t+1}) \right\|^2] \\ & \leq \frac{dc_2^2 C^2 PT \log(P/\delta)}{\epsilon^2 n^2} + \frac{64d\beta_s^2 L^4}{Pm} + \frac{8dC^2}{ePm} + \frac{\gamma^2}{Pm}. \end{aligned} \quad (32)$$

□

LEMMA A.2 (RESTATEMENT OF LEMMA 4.3). *Under Assumption 1 and $f(\theta, x)$ is a ρ -weakly-convex function of θ , by applying DP-ZOO*

(Algorithm 2) with $\eta_s \leq \frac{\beta_s}{\sqrt{dL}}$, for any $\theta \in \mathcal{R}^d$, we have

$$\begin{aligned} & \mathbb{E} [\phi_s(\hat{\theta}_{T_s}) - \phi_s(\theta)] \leq \left(\frac{1}{2\eta_s T_s} + \frac{1}{2T_s \lambda} \right) \|\theta^{s-1} - \theta\|^2 \\ & \quad + \eta_s \left(\frac{dc_2^2 C^2 PT \log(P/\delta)}{\epsilon^2 n^2} + \frac{64d\beta_s^2 L^4}{Pm} + \frac{8dC^2}{ePm} + \frac{\gamma^2}{Pm} \right). \end{aligned} \quad (33)$$

Proof for Lemma 4.3: Recall that $\phi_s(\theta) = f_{\beta_s}(\theta) + \frac{1}{2\lambda} \|\theta_t - \theta^s\|^2$, let $F_{\beta_s}(\theta) = f_{\beta_s}(\theta, \mathcal{D})$, $r_s(\theta) = \frac{1}{2\lambda} \|\theta - \theta^s\|^2$. Due to the ρ -weak-convexity of $f(\theta)$, the $\frac{1}{\lambda}$ -strong-convexity of $r_s(\theta)$ and the $\frac{\sqrt{dL}}{\beta_s}$ -smoothness of $f_{\beta_s}(\theta)$, we have the following three inequality

$$F_{\beta_s}(\theta) \geq F_{\beta_s}(\theta_t) + \langle \nabla F_{\beta_s}(\theta_t), \theta - \theta_t \rangle - \frac{\rho}{2} \|\theta_t - \theta\|^2$$

$$r_s(\theta) \geq r_s(\theta_{t+1}) + \langle \partial r_s(\theta_{t+1}), \theta - \theta_{t+1} \rangle + \frac{1}{2\lambda} \|\theta - \theta_{t+1}\|^2$$

$$F_{\beta_s}(\theta_t) \geq F_{\beta_s}(\theta_{t+1}) - \langle \nabla F_{\beta_s}(\theta_t), \theta_{t+1} - \theta_t \rangle - \frac{\sqrt{dL}}{2\beta_s} \|\theta_t - \theta_{t+1}\|^2. \quad (34)$$

combing them together, we have

$$\begin{aligned} & F_{\beta_s}(\theta_{t+1}) + r_s(\theta_{t+1}) - (F_{\beta_s}(\theta) + r_s(\theta)) \\ & \leq \langle \nabla F_{\beta_s}(\theta_t) + \partial r_s(\theta_{t+1}), \theta_{t+1} - \theta \rangle \\ & \quad + \frac{\rho}{2} \|\theta_t - \theta\|^2 + \frac{\sqrt{dL}}{2\beta_s} \|\theta_t - \theta_{t+1}\|^2 - \frac{1}{2\lambda} \|\theta - \theta_{t+1}\|^2. \end{aligned} \quad (35)$$

If we set the gradient in θ_{t+1} to 0, there exists $\partial r_s(\theta_{t+1})$ such that

$$\partial r_s(\theta_{t+1}) = \frac{1}{\eta_s} (\theta_t - \theta_{t+1}) - \nabla \hat{f}_{\beta_s}(\theta_t, \xi_{t+1}). \quad (36)$$

where $\nabla \hat{f}_{\beta_s}(\theta_t, \xi_{t+1}) = \frac{1}{P} \sum_{p=1}^P \frac{1}{m} \sum_{i=1}^m \text{CLIP} \left(\nabla f_{\beta_s}^p(\theta_t, x_{t+1}^i) \right) + \mathbf{z}_t^p$ and $\mathbf{z}_t^p = \mathbf{z}_t^p \cdot \mathbf{v}_t^p \left(\mathbf{z}_t^p \sim N(0, \sigma^2 C^2) \right)$. Plugging the above equation into 35 and setting $\hat{\theta}_{t+1}$ be the updated parameter on dataset $\mathcal{D}$ without DP noise in iteration $t+1$. Since η is decreasing and β is increasing, $\eta_s \leq \frac{\beta_s}{\sqrt{dL}}$ holds true for all $s \in [0, S]$ when $\eta_1 \leq \frac{\beta_1}{\sqrt{dL}}$, by taking $\eta_s \leq \frac{\beta_s}{\sqrt{dL}}$, we have

$$\begin{aligned} & F_{\beta_s}(\theta_{t+1}) + r_s(\theta_{t+1}) - (F_{\beta_s}(\theta) + r_s(\theta)) \\ & \leq \langle \nabla F_{\beta_s}(\theta_t) - \nabla \hat{f}_{\beta_s}(\theta_t, \xi_{t+1}), \theta_{t+1} - \theta \rangle + \left\langle \frac{1}{\eta_s} (\theta_t - \theta_{t+1}), \theta_{t+1} - \theta \right\rangle \\ & \quad + \frac{\rho}{2} \|\theta_t - \theta\|^2 + \frac{\sqrt{dL}}{2\beta_s} \|\theta_t - \theta_{t+1}\|^2 - \frac{1}{2\lambda} \|\theta - \theta_{t+1}\|^2 \\ & = \langle \nabla F_{\beta_s}(\theta_t) - \nabla \hat{f}_{\beta_s}(\theta_t, \xi_{t+1}), \theta_{t+1} - \hat{\theta}_{t+1} + \hat{\theta}_{t+1} - \theta \rangle \\ & \quad + \left(\frac{1}{2\eta_s} + \frac{\rho}{2} \right) \|\theta_t - \theta\|^2 - \left(\frac{1}{2\eta_s} + \frac{1}{2\lambda} \right) \|\theta - \theta_{t+1}\|^2 \\ & = \langle \nabla F_{\beta_s}(\theta_t) - \nabla \hat{f}_{\beta_s}(\theta_t, \xi_{t+1}), \hat{\theta}_{t+1} - \theta \rangle \\ & \quad + \eta_s \left\| \nabla F_{\beta_s}(\theta_t) - \nabla \hat{f}_{\beta_s}(\theta_t, \xi_{t+1}) \right\|^2 + \left(\frac{1}{2\eta_s} + \frac{\rho}{2} \right) \|\theta_t - \theta\|^2 \\ & \quad - \left(\frac{1}{2\eta_s} + \frac{1}{2\lambda} \right) \|\theta - \theta_{t+1}\|^2. \end{aligned} \quad (37)$$

Taking expectation on both sides, we have

$$\begin{aligned} \mathbb{E} [\phi_s(\theta_{t+1}) - \phi_s(\theta)] &\leq \eta_s \mathbb{E} \left[\left\| \nabla F_{\beta_s}(\theta_t) - \nabla \hat{f}_{\beta_s}(\theta_t, \xi_{t+1}) \right\|^2 \right] \\ &+ \left(\frac{1}{2\eta_s} + \frac{\rho}{2} \right) \|\theta_t - \theta\|^2 - \left(\frac{1}{2\eta_s} + \frac{1}{2\lambda} \right) \|\theta - \theta_{t+1}\|^2. \end{aligned} \quad (38)$$

Following Lemma 4.2 and by taking summation of the above inequality from $t = 0$ to $T_s - 1$, $\lambda < 1/\rho$, we have

$$\begin{aligned} \sum_{t=0}^{T_s-1} \phi_s(\theta_{t+1}) - \phi_s(\theta) &\leq \left(\frac{1}{2\eta_s} + \frac{\rho}{2} \right) \|\theta^{s-1} - \theta\|^2 \\ &+ \eta_s T_s \left(\frac{dc_2^2 C^2 PT \log(P/\delta)}{\epsilon^2 n^2} + \frac{64d\beta_s^2 L^4}{Pm} + \frac{8dC^2}{ePm} + \frac{\gamma^2}{Pm} \right) \\ &- \left(\frac{1}{2\eta_s} + \frac{1}{2\lambda} \right) \|\theta - \theta_{T_s}\|^2. \end{aligned} \quad (39)$$

By employing Jensens' inequality, denoting the output of stage s by $\theta^s = \hat{\theta}_{T_s} = \frac{1}{T_s} \sum_{t=1}^{T_s} \theta_t$, since $\rho \leq 1/\lambda$ we can obtain that

$$\begin{aligned} \phi_s(\hat{\theta}_{T_s}) - \phi_s(\theta) &\leq \left(\frac{1}{2\eta_s T_s} + \frac{1}{2T_s \lambda} \right) \|\theta^{s-1} - \theta\|^2 \\ &+ \eta_s \left(\frac{dc_2^2 C^2 PT \log(P/\delta)}{\epsilon^2 n^2} + \frac{64d\beta_s^2 L^4}{Pm} + \frac{\gamma^2}{Pm} + \frac{8dC^2}{ePm} \right). \end{aligned} \quad (40)$$

THEOREM A.3 (RESTATATION OF THEOREM 4.4). *In DP-ZOSO, suppose assumption 1 holds and the loss function $f(\theta, x)$ is ρ -weakly-convex of θ . Then by setting $\eta_s = \alpha_s \cdot \min\left\{\frac{Pm}{7\gamma^2}, \frac{Pme}{56dC^2}, \frac{\epsilon^2 n^2}{7dc_2^2 C^2 PT \log(P/\delta)}, \frac{Pm}{448d\beta_s^2 L^4}\right\}$ and $\lambda = \frac{7}{2\mu}$, $\eta_s T_s = \frac{7}{2\mu}$, after $S = \lceil \log(\alpha_0/\alpha) \rceil$ stages, we have*

$$\phi_s(\theta^S) - \phi_s(\theta^*) \leq \alpha. \quad (41)$$

The total ZO oracle complexity is

$$\mathcal{O} \left(\left(\gamma^2 + dC^2 + d\beta_s^2 L^4 + \frac{dC^2 P^2 m \log(P/\delta)}{\epsilon^2 n^2} \right) \cdot \frac{1}{\mu\alpha} \right). \quad (42)$$

PROOF OF THEOREM 4.4. By taking $\theta = \theta^*$, we have

$$\begin{aligned} f_{\beta_s}(\hat{\theta}_{T_s}) - f_{\beta_s}(\theta^*) &\leq \left(\frac{1}{2\lambda} + \frac{1}{2\eta_s T_s} + \frac{1}{2T_s \lambda} \right) \|\theta^{s-1} - \theta^*\|^2 \\ &+ \eta_s \left(\frac{dc_2^2 C^2 PT \log(P/\delta)}{\epsilon^2 n^2} + \frac{64d\beta_s^2 L^4}{Pm} + \frac{\gamma^2}{Pm} + \frac{8dC^2}{ePm} \right). \end{aligned} \quad (43)$$

Since f_{β} satisfies μ -PL condition, we will prove by induction that $f_{\beta_s}(\theta^s) - f_{\beta_s}(\theta^*) \leq \alpha_s$, where $\alpha_s = \alpha_0/2^s$, which is true for $s = 0$,

we have

$$\begin{aligned} f_{\beta_s}(\theta^s) - f_{\beta_s}(\theta^*) &\leq \left(\frac{1}{4\lambda\mu} + \frac{1}{4\eta_s T_s \mu} + \frac{1}{4T_s \lambda \mu} \right) (f_{\beta_{s-1}}(\theta^{s-1}) - f_{\beta_{s-1}}(\theta^*)) \\ &+ \eta_s \left(\frac{dc_2^2 C^2 PT \log(P/\delta)}{\epsilon^2 n^2} + \frac{64d\beta_s^2 L^4}{Pm} + \frac{\gamma^2}{Pm} + \frac{8dC^2}{ePm} \right) \\ &\leq \left(\frac{1}{4\lambda\mu} + \frac{1}{4\eta_s T_s \mu} + \frac{1}{4T_s \lambda \mu} \right) \alpha_{s-1} \\ &+ \eta_s \left(\frac{dc_2^2 C^2 PT \log(P/\delta)}{\epsilon^2 n^2} + \frac{64d\beta_s^2 L^4}{Pm} + \frac{\gamma^2}{Pm} + \frac{8dC^2}{ePm} \right). \end{aligned} \quad (44)$$

By setting $\eta_s = \alpha_s \cdot \min\left\{\frac{Pm}{7\gamma^2}, \frac{Pme}{56dC^2}, \frac{\epsilon^2 n^2}{7dc_2^2 C^2 PT \log(P/\delta)}, \frac{Pm}{448d\beta_s^2 L^4}\right\}$ and $\lambda = \frac{7}{2\mu}$, $\eta_s T_s = \frac{7}{2\mu}$, we have

$$f_{\beta_s}(\theta^s) - f_{\beta_s}(\theta^*) \leq \alpha_s. \quad (45)$$

By induction, after $S = \lceil \log(\alpha_0/\alpha_S) \rceil$ stages, we have

$$f_{\beta_S}(\theta^S) - f_{\beta_S}(\theta^*) \leq \alpha_S. \quad (46)$$

The total ZO oracle complexity of DP-ZOSO is

$$\mathcal{O} \left(\left(\gamma^2 + dC^2 + d\beta_s^2 L^4 + \frac{dC^2 P^2 m \log(P/\delta)}{\epsilon^2 n^2} \right) \cdot \frac{1}{\mu\alpha} \right). \quad (47)$$

□

LEMMA A.4. *In DP-ZOPO, under assumption 1, the dimension of parameters to be fine-tuned in stage s is reduced to $d * r_s$. The error between the gradient $\nabla_{\mathbb{V}} F_{\beta}(\theta)$ on the total dataset and the actual update gradient $\nabla_{\mathbb{V}} \hat{f}_{\beta}(\theta_t, \xi_{t+1})$ is bounded by:*

$$\begin{aligned} \mathbb{E} [\|\nabla_{\mathbb{V}} F_{\beta}(\theta_t) - \nabla_{\mathbb{V}} \hat{f}_{\beta}(\theta_t, \xi_{t+1})\|^2] &\leq \frac{dr_s c_2^2 C^2 PT \log(P/\delta)}{\epsilon^2 n^2} + \frac{8dr_s C^2}{ePm} + \frac{64dr_s \beta_s^2 L^4}{Pm} + \frac{\gamma^2}{Pm}. \end{aligned} \quad (48)$$

PROOF OF LEMMA A.4. In DP-ZOPO, the dimension of parameters to be fine-tuned in stage s is $d * r_s$ ($r_s \leq 1$). First, the scale of DP noise introduced is reduced by the current pruning rate r_s . Then, the d in the clipping error and the zeroth-order estimation error (second and third terms) is replaced with $d * r_s$ since the l_2 -norm of the direction vector $\|\mathbf{v}\|$ ($\mathbf{v} \sim \mathbb{V}$) is reduced to $d * r_s$. □

THEOREM A.5 (RESTATATION OF THEOREM 4.6). *In DP-ZOPO, suppose assumption 1 holds and loss function $f(\theta, x)$ is ρ -weakly-convex of θ . The dimensionality of parameters to be updated in stage s is $d * r_s$.*

Then by setting $\eta_s = \alpha_s \cdot \min\left\{\frac{Pm}{7\gamma^2}, \frac{Pme}{56dr_s C^2}, \frac{\epsilon^2 n^2}{7dr_s c_2^2 C^2 PT \log(P/\delta)}, \frac{Pm}{448dr_s \beta_s^2 L^4}\right\}$ and $\lambda = \frac{7}{2\mu}$, $\eta_s T_s = \frac{7}{2\mu}$, after $S = \lceil \log(\alpha_0/\alpha) \rceil$ stages, we have

$$\phi_s(\theta^{rS}) - \phi_s(\theta^*) \leq \alpha. \quad (49)$$

The total ZO oracle complexity of DP-ZOPO is

$$\mathcal{O} \left(\left(\gamma^2 + rdr_s C^2 + rdr_s \beta_s^2 L^4 + \frac{rdr_s C^2 P^2 m \log(P/\delta)}{\epsilon^2 n^2} \right) \cdot \frac{1}{\mu\alpha} \right). \quad (50)$$

PROOF OF THEOREM 4.6. Similar in Theorem 4.4, we can obtain that

$$\begin{aligned}
f_{\beta_s}(\theta^s) - f_{\beta_s}(\theta^*) &\leq \left(\frac{1}{4\lambda\mu} + \frac{1}{4\eta_s T_s \mu} + \frac{1}{4T_s \lambda \mu} \right) (f_{\beta_{s-1}}(\theta^{s-1}) - f_{\beta_{s-1}}(\theta^*)) \\
&\quad + \eta_s \left(\frac{dr_s c_2^2 C^2 PT \log(P/\delta)}{\epsilon^2 n^2} + \frac{64dr_s \beta_s^2 L^4}{Pm} + \frac{8dr_s C^2}{ePm} + \frac{\gamma^2}{Pm} \right) \\
&\leq \left(\frac{1}{4\lambda\mu} + \frac{1}{4\eta_s T_s \mu} + \frac{1}{4T_s \lambda \mu} \right) \alpha_{s-1} \\
&\quad + \eta_s \left(\frac{dr_s c_2^2 C^2 PT \log(P/\delta)}{\epsilon^2 n^2} + \frac{64dr_s \beta_s^2 L^4}{Pm} + \frac{8dr_s C^2}{ePm} + \frac{\gamma^2}{Pm} \right). \tag{51}
\end{aligned}$$

By setting $\eta_s = \alpha_s \cdot \min \left\{ \frac{Pm}{7\gamma^2}, \frac{Pme}{56dr_s C^2}, \frac{\epsilon^2 n^2}{7dr_s c_2^2 C^2 PT \log(P/\delta)}, \frac{Pm}{448dr_s \beta_s^2 L^4} \right\}$ and $\lambda = \frac{7}{2\mu}$, $\eta_s T_s = \frac{7}{2\mu}$, we have

$$f_{\beta_s}(\theta^s) - f_{\beta_s}(\theta^*) \leq \alpha_s. \tag{52}$$

By induction, after $S = \lceil \log(\alpha_0/\alpha_S) \rceil$ stages, we have

$$f_{\beta_S}(\theta^S) - f_{\beta_S}(\theta^*) \leq \alpha_S. \tag{53}$$

The total ZO oracle complexity of DP-ZOPO is

$$\mathcal{O} \left(\left(\gamma^2 + dr_s C^2 + dr_s \beta_s^2 L^4 + \frac{dr_s C^2 P^2 m \log(P/\delta)}{\epsilon^2 n^2} \right) \cdot \frac{1}{\mu\alpha} \right). \tag{54}$$

□

A.2 More Experiment Results

In this section, we will provide the experiment settings and more experiment results that are not included in the main body of this paper.

Table 11: The hyperparameters grids used for RoBERTa-large experiments. All experiments use 6K steps for training.

Experiment	Hyperparameters	Values
DPZero	Batch size	64
	Learning rate	$\{2e-5, 1e-5, 5e-6\}$
	ZO scale parameter	1e-6
DP-ZOSO, DP-ZOPO	Batch size	64
	Learning rate	$\{2e-4, 1e-4, 4e-5, 2e-5, 1e-5\}$
	ZO scale parameter	1e-6 to 1e-5
	Pruning rate (r)	$\{0.5\%, 1\%, 2\%, 4\%, 8\%, 100\%\}$
	Stage size	3
	Regulation parameter	5e-4
	Regulation parameter	5e-4
DP-SGD	Batch size	64
	Learning rate	$\{1e-4, 5e-4, 1e-3, 5e-3\}$
DP-prefix	Batch size	64
	Learning rate	$\{1e-2, 3e-2, 5e-2\}$
	prefix tokens	5

Table 12: The hyperparameters grids used for OPT experiments. All experiments use 6K steps for training.

Experiment	Hyperparameters	Values
ICL	Examples	32
	Batch size	16
	Learning rate	$\{1e-6\}$
DP-ZOSO, DP-ZOPO	ZO scale parameter	1e-4
	Batch size	16
	Learning rate	$\{4e-5, 3e-5, 2e-5, 1e-5, 5e-6\}$
	ZO scale parameter	1e-4 to 4e-4
	Pruning rate (r)	$\{0.5\%, 1\%, 2\%, 4\%, 8\%\}$
	Stage size	3
	Regulation parameter	5e-4

Table 13: Experiments on RoBERTa-large (350M parameters). $r = 0.5\%$ denotes the static pruning with pruning rate $r = 0.5\%$ before fine-tuning and $r = 0.5\%^*$ denotes the dynamic pruning with constant pruning rate $r = 0.5\%$

Task	SST-2	SST-5	SNLI	MNLI	TREC
Small Privacy budget: $\epsilon = 2$					
$r = 0.5\%$	88.5	46.5	70.3	70.1	83.4
$r = 0.5\%^*$	88.8	50.5	75.5	70.1	84.2
$r = 1\%$	91.7	48.8	78.4	68.1	88.6
$r = 1\%^*$	91.7	48.8	76.9	64.3	84.0
$r = 2\%$	88.4	47.1	78.2	71.4	89.6
$r = 2\%^*$	90.9	50.0	78.6	67.9	84.0
$r = 4\%$	92.2	46.6	79.5	69.8	85.0
$r = 4\%^*$	92.1	48.9	76.9	69.6	85.8
$r = 8\%$	92.5	46.6	77.3	69.3	85.4
$r = 8\%^*$	92.4	47.6	77.2	69.5	84.8
Medium Privacy budget: $\epsilon = 4$					
$r = 0.5\%$	92.2	50.5	78.6	71.4	81.0
$r = 0.5\%^*$	92.1	48.8	78.2	69.0	87.8
$r = 1\%$	92.2	47.1	80.4	64.7	88.2
$r = 1\%^*$	91.6	49.1	76.8	66.6	84.6
$r = 2\%$	91.6	47.7	80.4	71.5	87.0
$r = 2\%^*$	92.7	49.9	78.5	68.8	84.6
$r = 4\%$	93.2	47.0	78.3	69.7	86.4
$r = 4\%^*$	92.1	48.9	78.4	69.6	83.2
$r = 8\%$	92.5	48.4	78.6	69.7	84.4
$r = 8\%^*$	93.1	49.3	77.9	69.0	83.2
Large Privacy budget: $\epsilon = 8$					
$r = 0.5\%$	92.2	49.8	81.4	71.6	85.4
$r = 0.5\%^*$	92.5	48.8	76.2	67.7	87.6
$r = 1\%$	92.2	47.4	79.2	73.1	90.8
$r = 1\%^*$	91.7	47.9	77.5	65.6	82.6
$r = 2\%$	91.8	49.4	78.9	68.6	85.4
$r = 2\%^*$	92.5	48.6	79.3	67.7	82.4
$r = 4\%$	93.6	48.0	77.4	68.4	87.0
$r = 4\%^*$	92.1	47.1	77.0	69.3	85.0
$r = 8\%$	92.4	48.0	79.0	69.7	84.4
$r = 8\%^*$	92.4	46.5	77.2	69.5	87.0

Table 14: Experiments of RoBERTa-large **zeroth-order** and **first-order** fine-tuned on SNLI with different pruning rates.

(a) Zeroth-order					(b) First-order				
Pruning Rate	$\epsilon = 2$	$\epsilon = 4$	$\epsilon = 8$	$\epsilon = \infty$	Pruning Rate	$\epsilon = 2$	$\epsilon = 4$	$\epsilon = 8$	$\epsilon = \infty$
$r = 0.5\%$	70.3	78.6	81.4	80.5	$r = 0.5\%$	85.0	85.8	87.0	86.8
$r = 1\%$	78.4	80.4	79.2	80.3	$r = 1\%$	87.0	86.8	86.5	85.8
$r = 2\%$	78.2	80.4	78.9	79.6	$r = 2\%$	86.6	86.6	86.7	88.2
$r = 4\%$	79.5	78.3	77.4	79.3	$r = 4\%$	82.8	85.2	86.0	87.6
$r = 8\%$	77.3	78.6	79.0	78.6	$r = 8\%$	84.6	86.1	86.7	86.9
$r = 100\%$	76.8	78.9	76.5	78.1	$r = 100\%$	86.4	87.0	87.2	90.0

Table 15: Experiments on OPT-2.7B with privacy budget $\epsilon = 4$. 1-2-4% * denotes the incremental pruning strategy of with $r = 1-2-4\%$. Dynamic pruning outperforms static pruning in all five datasets.

Task	SST-2	CB	BoolQ	MultiRC
static*	90.0	62.0	63.7	50.0
0.5 – 1 – 2%	92.5	76.0	67.5	65.0
0.5 – 1 – 2%*	92.2	77.2	69.2	61.6
1 – 2 – 4%	88.1	77.2	67.3	55.8
1 – 2 – 4%*	92.3	78.8	67.3	58.3
2 – 4 – 8%	90.0	76.8	66.7	54.0
2 – 4 – 8%*	89.0	76.8	66.7	57.0

Table 18: Experiments on RoBERTa-large (350M parameters) with privacy budget $\epsilon = 8$.

Task	SST-2	SST-5	SNLI	MNLI	TREC
$r = 0.5\%$	92.2	49.8	81.4	71.6	85.4
$r = 1\%$	92.2	47.4	79.2	73.1	90.8
$r = 2\%$	91.8	49.4	78.9	68.6	85.4
$r = 4\%$	93.6	48.0	77.4	68.4	87.0
$r = 8\%$	92.4	48.0	79.0	69.7	84.4
$r = 100\%$	93.2	41.2	76.5	69.2	84.2

Table 16: Experiments on RoBERTa-large (350M parameters) with privacy budget $\epsilon = 2$.

Task	SST-2	SST-5	SNLI	MNLI	TREC
$r = 0.5\%$	88.5	46.5	70.3	70.1	83.4
$r = 1\%$	91.7	48.8	78.4	68.1	88.6
$r = 2\%$	88.4	47.1	78.2	71.4	89.6
$r = 4\%$	92.2	46.6	79.5	69.8	85.0
$r = 8\%$	92.5	46.6	77.3	69.3	85.4
$r = 100\%$	92.9	45.8	76.8	66.3	83.4

Table 17: Experiments on RoBERTa-large (350M parameters) with privacy budget $\epsilon = 4$.

Task	SST-2	SST-5	SNLI	MNLI	TREC
$r = 0.5\%$	92.2	50.5	78.6	71.4	81.0
$r = 1\%$	92.2	47.1	80.4	64.7	88.2
$r = 2\%$	91.6	47.7	80.4	71.5	87.0
$r = 4\%$	93.2	47.0	78.3	69.7	86.4
$r = 8\%$	92.5	48.4	78.6	69.7	84.4
$r = 100\%$	92.6	42.0	78.9	67.3	83.8