To be private and robust: Differentially Private Optimizers Can Learn Adversarially Robust Models

Anonymous authors
Paper under double-blind review

Abstract

Machine learning models have shone in a variety of domains and attracted increasing attention from both the security and the privacy communities. One important yet worrying question is: will training models under the differential privacy (DP) constraint unfavorably impact on the adversarial robustness? While previous works have postulated that privacy comes at the cost of worse robustness, we give the first theoretical analysis that DP models can indeed be robust and accurate, even sometimes more robust than their naturally-trained non-private counterparts. We observe three key factors that influence the privacy-robustness-accuracy tradeoff: (1) hyperparamters for DP optimizers are critical; (2) pre-training on public data significantly mitigates the accuracy and robustness gap; (3) choice of DP optimizers makes a difference. With these factors set properly, we achieve 90% natural accuracy, 72% robust accuracy (+9% than the non-private model) under $l_2(0.5)$ attack, and 69% robust accuracy (+16% than the non-private model) with pre-trained SimCLRv2 model under $l_\infty(4/255)$ attack on CIFAR10 with $\epsilon = 2$. In fact, DP models can be, theoretically and empirically, Pareto optimal in terms of accuracy and robustness. The robustness of DP models is consistently observed on MNIST, Fashion MNIST and CelebA, with ResNet and Vision Transformer. We believe our results are an encouraging step towards training models that are private as well as robust.

1 Introduction

Machine learning models trained on large amount of data can be vulnerable to privacy attacks and leak sensitive information. For example, Carlini et al. (2021) shows that attackers can extract text input from the training set via GPT2 (Radford et al., 2019), that contains private information such as address, phone number, name and so on; Zhu et al. (2019) shows that attackers can recover both the image input and the label from gradients of ResNet (He et al., 2016) trained on CIFAR100 (Krizhevsky et al., 2009), SVHN (Netzer et al., 2011), and LFW (Huang et al., 2008).

To protect against the privacy risk rigorously, the differential privacy (DP) is widely applied in various deep learning tasks (Abadi et al., 2016; McMahan et al., 2017; Bu et al., 2020; Nori et al., 2021; Li et al., 2021), including but not limited to computer vision, natural language processing, recommendation system, and so on. At high level, DP is realized via DP optimizers such as DP-SGD and DP-Adam, while allowing the privacy-preserving models to achieve high accuracy. Therefore, the privacy concerns have been largely alleviated by switching from regular optimizers to DP ones.

An equally important concern from the security community is that, many models such as deep neural networks are known to be vulnerable against adversarial attacks. This robustness risk is severe when the attackers can successfully fool models to make the wrong prediction, through negligible perturbation on the input data. An example from Engstrom et al. (2019) shows that a strong ResNet50 trained on ImageNet (Deng et al., 2009) with 76.13% accuracy can degrade to 0.00% accuracy, even if the input image is merely perturbed by $4/255$.

However, at the intersection of these two concerns, previous works have empirically observed an upsetting privacy-robustness tradeoff under some scenarios, implying the implausibility of achieving both robustness and privacy at the same time. In Song et al. (2019) and Mejia et al. (2019), adversarially trained models are shown to be more vulnerable to privacy attacks, such as the membership inference attack, than naturally-trained models. In Tursynbek et al. (2020) and Boenisch et al. (2021), DP trained models were more vulnerable to robustness attacks than naturally-trained models on MNIST and CIFAR10.

On the contrary, we show that DP models can be adversarially robust, sometimes even more robust than the naturally-trained models. Indeed, we illustrate that DP models can be Pareto optimal. That is, under some conditions, any model with higher accuracy than DP ones must have worse robustness. In sharp contrast to the empirical nature of previous arts, we enhance our understanding about the adversarial robustness of DP models from a theoretical angle. Our analysis shows that DP classifier (without adversarial training) can be the most adversarially robust. Motivated by our theoretical analysis, we claim that the hyperparameter tuning is vital to successfully learning a robust and private model, where the optimal choice is to use small clipping norms and large learning rates. Interestingly, such a hyperparameter choice is also observed to be the most effective in learning naturally accurate models under DP (Li et al., 2021; Kurakin et al., 2022; De et al., 2022; Bu et al., 2022b). Additionally, we advocate pretraining and selecting proper optimizers for DP training, which allow us to max out the performance on MNIST (LeCun et al., 1998), Fashion MNIST(Xiao et al., 2017), CIFAR10(Krizhevsky et al., 2009), and CelebA(Liu et al., 2015).

2 Preliminaries

Notation. We denote the model as f, the data as $\{\boldsymbol{x}_i\} \in \mathbb{R}^d$ and the labels as $\{y_i\}$, following i.i.d. from the distributions $\boldsymbol{x}$ and y respectively. We denote the gradient of the i-th sample at step t as $g_t(\boldsymbol{x}_i, y_i; \mathbf{w}, b)$, where $\mathbf{w}$ is the weights and b is the bias of model f.

To start, we introduce the definition of (ϵ, δ)-DP (Dwork et al., 2014).

Definition 2.1. A randomized algorithm M is (ε, δ)-DP if for any neighboring datasets S, S' that differ by one arbitrary sample, and for any event E, it holds that

$$\mathbb{P}[M(S) \in E] \leqslant \mathrm{e}^{\varepsilon} \mathbb{P}\left[M\left(S'\right) \in E\right] + \delta. \tag{1}$$

In words, DP guarantees in the worst case that adding or removing one single datapoint $(\boldsymbol{x}_i, y_i)$ does not affect the model much, as quantified by the small constants (ϵ, δ). Therefore DP limits the information possibly leaked about such datapoint.

Algorithmically, DP is guaranteed by privatizing the gradient in two steps: (1) the per-sample gradient clipping Abadi et al. (2016) (specified by the clipping norm R, to bound the sensitivity of $\sum_i \boldsymbol{g}(\boldsymbol{x}_i, y_i)$); (2) the random noising (specified by the noise multiplier σ_{DP}, to randomize the outcome so that each sample's contribution is indistinguishable). In words, DP training simply applies any optimizer (SGD/Adam/...) on the private gradients instead of on the regular gradients.

$$\begin{aligned} &\text{Non-DP training on regular gradient:} && \textstyle\sum_i \boldsymbol{g}(\boldsymbol{x}_i, y_i) \\ &\text{DP training on private gradient:} && \textstyle\sum_i \boldsymbol{g}(\boldsymbol{x}_i, y_i) \cdot \min\{R/||\boldsymbol{g}_i||_2, 1\} + \sigma_{\text{DP}} R \cdot \mathcal{N}(\mathbf{0}, \mathbf{I}) \end{aligned} \tag{2}$$

We are interested in the adversarial robustness of models trained by DP optimizers: we consider the adversarially robust classification error and the natural classification error as

$$\mathcal{R}_\gamma(f) := \mathbb{P}(\exists ||\boldsymbol{p}||_\infty < \gamma, \text{ s.t. } f(\boldsymbol{x} + \boldsymbol{p}) \neq y), \quad \mathcal{R}_0(f) := \mathbb{P}(f(\boldsymbol{x}) \neq y), \tag{3}$$

where $\boldsymbol{p} \in \mathbb{R}^d$ is the adversarial perturbation, γ is the attack magnitude, and l_∞ attack is considered here. Notice that when $\gamma = 0$, the robust error in (3) reduces to the natural error.

Remark 2.2. It is important to note that ϵ is determined by σ_{DP}, but not R, e.g. $\epsilon = \sqrt{2\ln(1.25/\delta)}/\sigma_{\text{DP}}$ Dwork et al. (2014); Abadi et al. (2016); Dong et al. (2021). Hence, DP classifiers with any R share exactly the same privacy guarantee, but some are more robust and/or some are more accurate. We highlight in Section 3.2 that DP itself does not imply robustness, but its optimization (with correct hyperparameter) might.

3 Theoretical Analysis on Linear Classifiers

To theoretically understand the adversarial robustness of DP learning, we study the robustness of the DP and non-DP linear models on a binary classification problem. We consider a mixed Gaussian distribution, where the positive class $y=+1$ has a larger variance (i.e. it is more difficult to be classified correctly[1]) than the negative class $y=-1$:

$$\boldsymbol{x} \sim \begin{cases} \mathcal{N}\left(\boldsymbol{\theta}_d, K^2\sigma^2\mathbf{I}_d\right) & \text{if } y=+1 \\ \mathcal{N}\left(-\boldsymbol{\theta}_d, \sigma^2\mathbf{I}_d\right) & \text{if } y=-1 \end{cases} \tag{4}$$

where $y \overset{\text{unif}}{\sim} \{-1,+1\}$, $\boldsymbol{\theta}_d = (\theta, \cdots, \theta) \in \mathbb{R}^d$, $\sigma > 0$ and $K > 1$.

The setting[2] in (4) is analyzable because of the data symmetry along the diagonal axis $\mathbb{E}x_1 = \cdots = \mathbb{E}x_d$ (see Figure 1). We show that this symmetry leads to an explicit decision hyperplane of linear classifiers in Theorem 1, which further characterizes the strongest adversarial perturbation $\boldsymbol{p}^* \equiv \arg\sup_{\|\boldsymbol{p}\|_\infty<\gamma} \mathbb{P}(f(\boldsymbol{x}+\boldsymbol{p}) \neq y)$ explicitly.

In the following analysis, we focus on the linear classifiers with weights w_j and bias b,

$$f(\boldsymbol{x}; \mathbf{w}, b) = \text{sign}(\sum_{j=1}^{d} w_j x_j + b).$$

Within the family of linear classifiers, by the symmetry of data in (4), it can be rigorously shown by (1) that the optimal weights with respect to the natural and robust errors are always $w_1 = \cdots = w_d$. That is, the weights do not distinguish between the robust and natural models. Consequently, the key to the adversarial robustness lies in the intercept b, which is analyzed in the subsequent sections.

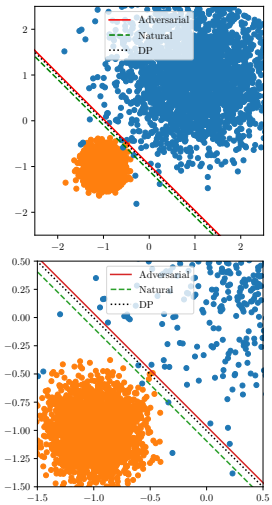

Figure 1: Decision boundaries of linear classifiers for (4), $K = 4, \sigma = 0.2, \theta = 1$.

3.1 Optimal Robust and Natural Linear Classifiers

We firstly review the robust error of robust classifier and the explicit formula of its intercept.

Theorem 1 (Extended from Theorem 2 in Xu et al. (2021)). *For data distribution $(\boldsymbol{x}, y)$ in Equation (4) and under the γ attack magnitude, the optimal robust linear classifier is*

$$f_\gamma = \underset{f \text{ is linear}}{\arg\min}\ \mathbb{P}(\exists ||\boldsymbol{p}||_\infty < \gamma,\ s.t.\ f(\boldsymbol{x}+\boldsymbol{p}) \neq y) = \underset{f \text{ is linear}}{\arg\min}\ \mathcal{R}_\gamma(f).$$

The optimal robust error is

$$\mathcal{R}_\gamma\left(f_\gamma\right) = \frac{1}{2}\Phi\left(B(K,\gamma) - K\sqrt{B(K,\gamma)^2 + q(K)}\right) + \frac{1}{2}\Phi\left(-KB(K,\gamma) + \sqrt{B(K,\gamma)^2 + q(K)}\right),$$

where Φ is the cumulative distribution function of standard normal, $B(K,\gamma) = \frac{2}{K^2-1}\frac{\sqrt{d}(\theta-\gamma)}{\sigma}$ and $q(K) = \frac{2\log K}{K^2-1}$. Furthermore, by the symmetry of the data distribution, we have

$$1, \cdots, 1, b_\gamma = \underset{\mathbf{w},b}{\arg\min} \mathcal{R}_\gamma(f(\cdot; \mathbf{w}, b))$$

in which

$$b_\gamma = \frac{K^2+1}{K^2-1}d(\theta-\gamma) - K\sqrt{\frac{4d^2(\theta-\gamma)^2}{\left(K^2-1\right)^2} + d\sigma^2 q(K)}. \tag{5}$$

Theorem 1 gives the closed form of the optimal robust classifier f_γ, or equivalently its intercept b_γ, and the optimal robust error. The special case of natural error can be easily recovered with $\gamma = 0$:

$$1, \cdots, 1, b_0 = \arg\min_{\mathbf{w},b} \mathcal{R}_0(f(\cdot; \mathbf{w}, b)).$$

[1] The fact that larger variance indicates lower intra-class accuracy is rigorously proven by Xu et al. (2021, Theorem 1).

[2] This setting is also studied in Xu et al. (2021), which focuses on the robustness-fairness tradeoff, and thus is different to our interest.

We know for sure from Theorem 1 that there exists a tradeoff between the robustness and accuracy: it is impossible for the natural classifier f_0 to be optimally robust, or for the robust classifier f_γ to be optimally accurate, since $b_0 \neq b_\gamma$ (c.f. Figure 2 and Fact 3.1).

Fact 3.1. *b_γ in (5) is strictly decreasing in γ, ranging from b_0 to $-\infty$.*

Proof of Fact 3.1. Proof 5 in Xu et al. (2021) shows that $\frac{db_\gamma}{d\gamma} \leq -\frac{K-1}{K+1}d < 0$, thus b_γ is strictly decreasing in γ. Therefore, the range of b_γ is $(b_\infty, b_0]$. Finally, we note that $b_\gamma < \frac{K^2+1}{K^2-1}(\theta-\gamma)$, hence $b_\infty = -\infty$. □

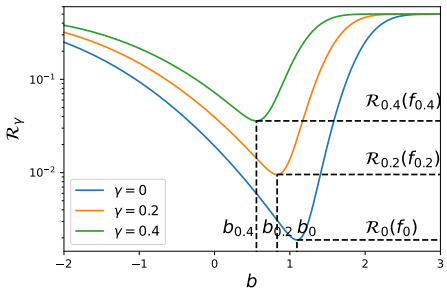

Figure 2: Intercepts and robust/natural error under (4), with $K = 4, \sigma = 0.2, \theta = 1$.

3.2 Adversarially Robust Errors of Private Linear classifiers

We focus on a specific linear classifier $f(\cdot; \mathbf{1}, b)$ or simply $f(\cdot; b)$ where only the intercept b is learned and DP-protected. This bias-only optimization is commonly used in transfer learning Zaken et al. (2022); He et al. (2021). Note that this linear classifier is as private as those optimizing both weights and bias, since σ_{DP} is the same.

For this classifier and any b, the robust error ($\gamma \neq 0$) and the natural error ($\gamma = 0$) are:

$$\begin{aligned}
\mathcal{R}_\gamma(f) =& \mathbb{P}(\exists \|\boldsymbol{p}\|_\infty \leq \gamma \text{ s.t. } f(\boldsymbol{x}+\boldsymbol{p}) \neq y) = \max_{\|\boldsymbol{p}\|_\infty \leq \gamma} \mathbb{P}(f(\boldsymbol{x}+\boldsymbol{p}) \neq y) \\
=& \frac{1}{2}\mathbb{P}(f(\boldsymbol{x}+\boldsymbol{\gamma}_d) \neq -1 \mid y=-1) + \frac{1}{2}\mathbb{P}(f(\boldsymbol{x}-\boldsymbol{\gamma}_d) \neq +1 \mid y=+1) \\
=& \frac{1}{2}\mathbb{P}\left(\sum_{i=1}^d w_j\,(x_j+\gamma) + b > 0 \mid y=-1\right) + \frac{1}{2}\mathbb{P}\left(\sum_{i=1}^d w_j\,(x_j-\gamma) + b < 0 \mid y=+1\right) \\
=& \frac{1}{2}\Phi\left(-\frac{\sqrt{d}(\theta-\gamma)}{\sigma} + \frac{1}{\sqrt{d}\sigma}\cdot b\right) + \frac{1}{2}\Phi\left(-\frac{\sqrt{d}(\theta-\gamma)}{K\sigma} - \frac{1}{K\sqrt{d}\sigma}\cdot b\right) \qquad (6)
\end{aligned}$$

where $\boldsymbol{\gamma}_d \equiv (\gamma, \cdots, \gamma)$. With (6), we can analyze the robust and natural errors for any intercept b (private or not, robust or natural) and any attack magnitude γ.

DP classifiers can be the most adversarially robust Our next result answers the following question: fixing a DP classifier $f_{\text{DP}} := f(\cdot; b_{\text{DP}})$, or equivalently its intercept b_{DP}, under which attack magnitude is it robust? We show that, it is possible for some attack magnitude γ^* that $b_{\text{DP}} = b_{\gamma^*}$, and thus the DP classifier is the most robust classifier among all linear classifiers.

Theorem 2. *For data distribution $(\boldsymbol{x}, y)$ in Equation (4) and for any $b_{\text{DP}} < b_0$, there exists $\gamma^* > 0$ such that $b_{\gamma^*} = b_{\text{DP}}$, and therefore*

$$\min_{f \text{ is linear}} \mathcal{R}_{\gamma^*}(f) \equiv \mathcal{R}_{\gamma^*}(f_{\gamma^*}) = \mathcal{R}_{\gamma^*}(f_{\text{DP}})$$

In words, the DP classifier is optimally robust under attack magnitude γ^*.

Proof of Theorem 2. By Fact 3.1, $b_\gamma - b_{\text{DP}}$ is decreasing in γ, ranging from $b_0 - b_{\text{DP}}$ to $-\infty$. By the intermediate value theorem, there exists $\gamma^* > 0$ such that $b_{\gamma^*} = b_{\text{DP}}$, i.e. $f_{\text{DP}} = f_{\gamma^*}(\cdot; b_{\gamma^*})$ is the most adversarially robust. □

By Theorem 2, as long as the DP intercept is sufficiently small, the DP classifier must be the most robust under some attack magnitude. We visualize in Figure 3 that for small b_{DP}, we can inversely find $\gamma^* = 0.075$ such that indeed $b_{\text{DP}} \approx b_{\gamma^*}$ (grey solid line).

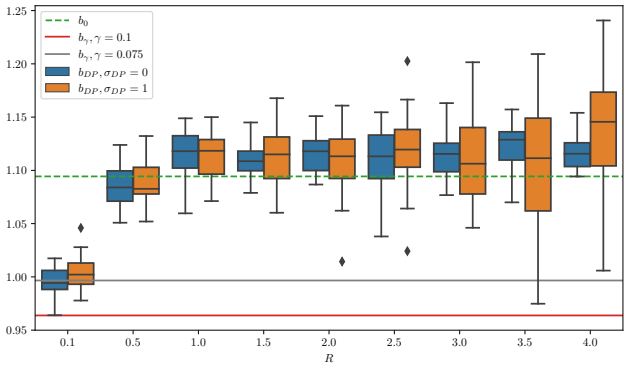

Figure 3: Intercepts for (4), same setup as Figure 2. For DP classifiers, we use DP-SGD with $\eta = 8$, epochs=50, batch size=1000, sample size=10000, (ϵ, δ)=(15,1e-4).

DP optimization can lead to robustness The achievability of $b_{\text{DP}} < b_0$, for Theorem 2 to hold, can be explained by the per-sample gradient clipping in DP optimizers. We temporarily ignore the noise ($\sigma = 0$)[3] and denote $C(\boldsymbol{x}_i, y_i; b) := g(\boldsymbol{x}_i, y_i; b) \min\{R/|\boldsymbol{g}_i|, 1\}$ as the clipped per-sample gradient. By definition of stationary point, $\sum C(\boldsymbol{x}_i, +1; b_{\text{DP}}) + \sum C(\boldsymbol{x}_i, -1; b_{\text{DP}}) = 0$ and $\sum g(\boldsymbol{x}_i, +1; b_0) + \sum g(\boldsymbol{x}_i, -1; b_0) = 0$. For large R, the clipping happens: $C(\boldsymbol{x}_i, y_i; b) \approx g(\boldsymbol{x}_i, y_i; b)$ and $b_{\text{DP}} \approx b_0$, i.e. DP classifier is as accurate/robust as the natural classifier; for small R (i.e. $|g(\boldsymbol{x}_i, y_i; b_{\text{DP}}| > R$), $-C(\boldsymbol{x}_i, +1; b_{\text{DP}}) = C(\boldsymbol{x}_i, -1; b_{\text{DP}}) = R$ which indicates the gradients from $+1$ class balance against the other class. Hence, we can have $b_{\text{DP}} < b_0$ and, by Fact 3.1, DP classifier is more robust than the natural classifier.

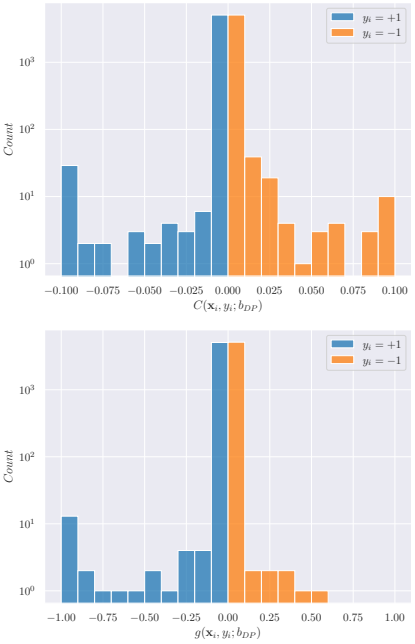

Figure 4: Per-sample gradients at $b = b_{0.075}$ before and after clipping for (4), $K = 4, \sigma = 0.2, \theta = 1$, and $R = 0.1$.

DP classifiers can be Pareto optimal in robustness and accuracy Next, suppose we only require the DP classifier to be more robust than the natural classifier, not necessarily the most robust among all linear classifiers. We can answer the question: fixing the attack magnitude γ, under which condition is f_{DP} more robust than the natural classifier f_0?

Theorem 3. *Fixing the attack magnitude γ, if data distribution in Equation (4) satisfies $\frac{K^2+1}{2K}\gamma < |\theta - \gamma| + |\theta|$, then whenever $b_\gamma < b_{\text{DP}} < b_0$, we have*

$$\min_{f \text{ is linear}} \mathcal{R}_\gamma(f) \equiv \mathcal{R}_\gamma(f_\gamma) < \mathcal{R}_\gamma(f_{\text{DP}}) < \mathcal{R}_\gamma(f_0).$$

Furthermore, any b with better natural accuracy than b_{DP} must have worse robustness:

$$\mathcal{R}_0(f) < \mathcal{R}_0(f_{\text{DP}}) \Longrightarrow \mathcal{R}_\gamma(f) > \mathcal{R}_\gamma(f_{\text{DP}}).$$

That is, DP linear classifier can be Pareto optimal in terms of the robust and natural accuracy.

Theorem 3 shows that, under some conditions, DP models are more robust than natural models, and cannot be dominated in the Pareto optimal sense. We visualize the premise $b_\gamma < b_{\text{DP}} < b_0$ in Figure 3 and Figure 1, and that $\mathcal{R}_\gamma(f_\gamma) < \mathcal{R}_\gamma(f_{\text{DP}}) < \mathcal{R}_\gamma(f_0)$ in Figure 2.

[3]Noise only adds variance to the gradient but does not affect the mean (see Figure 3), and when learning rate is small, σ has little effect on convergence (Bu et al., 2021).

We emphasize that our results on the l_∞ attacks are extendable to l_2 attacks (see Corollary 3.2 and experiments in Appendix C). Put differently, we show that DP models can be adversarially robust and Pareto optimal under both l_∞ and l_2 attacks.

Corollary 3.2. *All theorems hold for l_2 attacks by changing $\gamma \to \gamma/\sqrt{d}$.*

Remark 3.3. Theorem 2 gives sufficient and necessary condition for the DP classifier to be more robust than all classifiers at one attack magnitude γ^*; Theorem 3 gives sufficient but not necessary condition for the DP classifier to be more robust than one classifier (the natural one) at many attack magnitudes.

4 Training Private and Robust Deep Neural Networks

We extend our investigation beyond the linear classifiers in Theorem 2 and Theorem 3, and study the robustness of DP neural networks. Notably, several state-of-the-art advances are actually achieved by linear classifiers within the deep neural networks (Mehta et al., 2022; Tramer & Boneh, 2020), i.e. by finetuning only the last linear layer of neural networks, such as Wide ResNet and SimCLR. In what follows, we finetune all parameters in all layers.

By experimenting with real datasets MNIST, CIFAR10 and CelebA, we corroborate our claim that DP models can be adversarially robust on deep neural networks. We use one GPU of Nvidia GTX 1080 Ti.

4.1 Hyperparameters are keys to robustness: small clipping norm

In DP deep learning, the clipping norm R and learning rate η are crucial for high natural accuracy. On one hand, R has to be small to achieve state-of-the-art natural accuracy. In Li et al. (2021); De et al. (2022); Mehta et al. (2022); Bu et al. (2022a), large models such as ResNet, ViT and GPT2 are optimally trained at $R < 1$, even though the gradient's dimension is of hundreds of millions. On the other hand, DP training empirically benefits from large learning rate, usually 10 times larger than the non-DP training. This pattern is observed for DP-Adam (Li et al., 2021) and for DP-SGD (Kurakin et al., 2022), over text and image datasets.

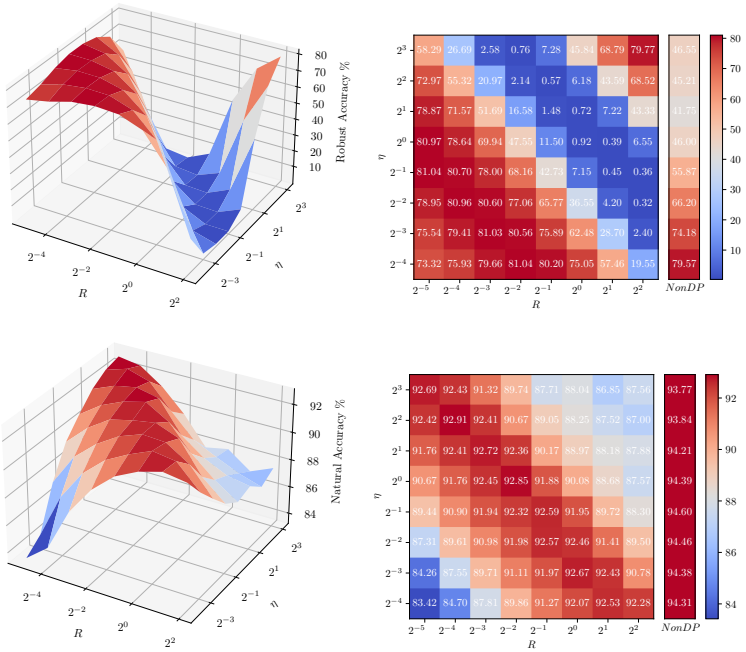

Figure 5: Robust and natural accuracy by (R, η) on CIFAR10. Same setting as Tramer & Boneh (2020): pretraining SimCLR on ImageNet and privately training using DP-SGD (momentum=0.9), under (ϵ, δ)=(2,1e-5) and attacked by 20 steps of $l_\infty(2/255)$ PGD.

Interestingly, we also observe such choice of (R, η) performs strongly in the adversarial robustness context (though not exactly the same hyperparameters). From the ablation study in Figure 5 for CIFAR10 and in Appendix B for MNIST, Fashion MNIST and CelebA, it is clear that robust accuracy and natural accuracy have distinctively different landscapes over (R, η). We observe that the optimal (R, η) should be carefully selected along the diagonal ridge for DP-SGD to obtain high robustness and high natural accuracy. Otherwise, even small deviation can lead to a sharp drop in the robustness, though the natural accuracy may remain similar (see upper right corner of 2D plots in Figure 5).

Our ablation study demonstrates that the most robust DP network, with 81.04% robust accuracy and 89.86% natural accuracy, can be more robust than the most robust naturally-trained network (79.59% robust accuracy and 94.31% natural accuracy). If we trade some robustness off the DP network, we can achieve the same level of robustness as the most robust non-DP network (80.20% robust accuracy and 91.27% natural accuracy).

While Figure 5 presents the result of a single attack magnitude, we further study the influence of hyperparameters under different attack magnitudes. We illustrate on CIFAR10 the l_∞ attack performance in Table 1 and the l_2 one in Table 5.

	SimCLRv2 pre-trained on unlabelled ImageNet								ResNet50	
attack magnitude	DP $\epsilon = 2$ robust	DP $\epsilon = 2$ accurate	DP $\epsilon = 4$ robust	DP $\epsilon = 4$ accurate	DP $\epsilon = 8$ robust	DP $\epsilon = 8$ accurate	Non-DP $\epsilon = \infty$ robust	Non-DP $\epsilon = \infty$ accurate	Non-DP $\epsilon = \infty$ adv $^8/_{255}$	Non-DP $\epsilon = \infty$ accurate
$\gamma = 0$	89.69%	92.87%	90.91%	93.41%	91.22%	93.64%	94.29%	94.55%	87.03%	95.25%
$\gamma = {}^2/_{255}$	81.05%	33.21%	82.53%	57.80%	83.02%	68.90%	79.79%	59.56%	–	–
$\gamma = {}^4/_{255}$	68.85%	0.16%	70.21%	9.69%	71.08%	28.09%	53.56%	15.99%	–	–
$\gamma = {}^8/_{255}$	39.63%	0.00%	38.39%	0.00%	39.28%	0.01%	8.14%	0.00%	53.49%	0.00%
$\gamma = {}^{16}/_{255}$	1.20%	0.00%	0.65%	0.00%	0.91%	0.00%	0.00%	0.00%	18.13%	0.00%

Table 1: Natural and robust accuracy of SimCLRv2 (Chen et al., 2020) and ResNet50 from (Engstrom et al., 2019) on CIFAR10 under 20 steps l_∞ PGD attack. See the explanation of *robust* or *accurate* versions below and detailed hyperparameters in Appendix D.

We evaluate the robust and natural accuracy on the state-of-the-art DP models in Tramer & Boneh (2020), considering two groups of hyperparameters: the *accurate* one reproduced from Tramer & Boneh (2020) that has the highest natural accuracy, and the *robust* one from a grid search on (R, η) for the highest robust accuracy. From Table 1 and Table 5, we see that even under the same privacy constraint, the robustness from different hyperparameters can be fundamentally different. For example, DP SimCLR at $\epsilon = 2$ can be either very robust ($\approx 70\%$ accuracy at $\gamma = 4/255$) or not robust at all (0.16% accuracy). Consequently, our results may explain the misunderstanding of previous researches by the improper choice of the hyperparameters.

Scrutinizing the robust version of hyperparameters, we see that, across all l_∞ attack magnitudes $\gamma = \{2/255, 4/255, 8/255, 16/255\}$and l_2 ones $\gamma = \{0, 0.25, 0.5, 1.0, 2.0\}$, DP SimCLR can be more robust than the non-DP SimCLR, in fact comparable to the adversarially trained ResNet50 that is benchmarked in Engstrom et al. (2019). To be assured, we demonstrate that our choice of small R and large η is consistently robust on Fashion MNIST, CIFAR10 and CelebA in Appendix C.

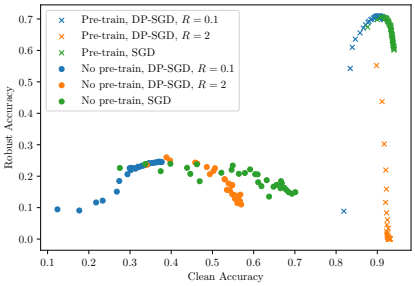

Figure 6: Robust and natural accuracy on CIFAR10 at different iterations. Dots are CNN from Papernot et al. (2020). Crosses are SimCLR from Tramer & Boneh (2020). See details in Appendix D.

4.2 Pareto optimality in accuracy and robustness

In the standard non-DP regime, the tradeoff between the accuracy and the robustness is well-known Engstrom et al. (2019). We extend the Pareto statement in Theorem 3 to DP deep learning, thus adding the privacy dimension into the privacy-accuracy-robustness

tradeoff. In Figure 6, we show that two state-of-the-art DP models on CIFAR10 (one pre-trained, the other not) achieve Pareto optimality with proper hyperparamters, and thus cannot be dominated by any natural classifiers. This can be observed by the fact that no green cross (or dot) is to the top right of all blue crosses (or dots), meaning that any natural classifier may have better robustness or higher accuracy, but not both. Therefore, our observation supports the claim that DP neural networks can be Pareto optimal in terms of the robustness and the accuracy.

4.3 DP neural networks can be robust against general attacks

Following the claim in Section 4.1 that DP neural networks can be robust against l_2 and l_∞ PGD attacks, we now demonstrate the transferability of DP neural networks' robustness against different attacks. We consider FGSM(Goodfellow et al., 2014), BIM(Kurakin et al., 2018), PGD_∞(Madry et al., 2017), APGD_∞(Croce & Hein, 2020), PGD_2(Madry et al., 2017) and APGD_2(Croce & Hein, 2020). This is interesting in the sense that DP mechanism does no intentionally defend against any adversarial attack, while the adversarial training (Goodfellow et al., 2014) usually specifically targets a particular attack, e.g. PGD attack is defensed by PGD adversarial training. In Table 2, we attack on the robust models from Table 1, with the *robust* parameters. We consistently observe that DP models can be more adversarially robust than the non-DP ones on MNIST/Fashion MNIST/CelebA in Appendix C, if the hyperparameters (R, η) are set properly.

	Natural	FGSM	BIM	PGD_∞	APGD_∞	PGD_2	APGD_2
DP , $\epsilon = 2$	89.86%	69.73%	68.85%	68.85%	68.85%	72.10%	71.97%
DP , $\epsilon = 4$	90.91%	71.13%	70.21%	70.21%	70.21%	72.79%	72.63%
DP , $\epsilon = 8$	91.22%	71.88%	71.08%	71.08%	71.06%	73.08%	72.99%
Non-DP	94.29%	56.24%	53.56%	53.56%	53.56%	63.32%	62.93%

Table 2: Natural and robust accuracy on CIFAR10 under general l_2/l_∞ adversarial attacks. Same model as Table 1 with the *robust* parameters. See attack hyperparameters in Appendix D.

4.4 Robustness and accuracy landscapes of DP optimizers

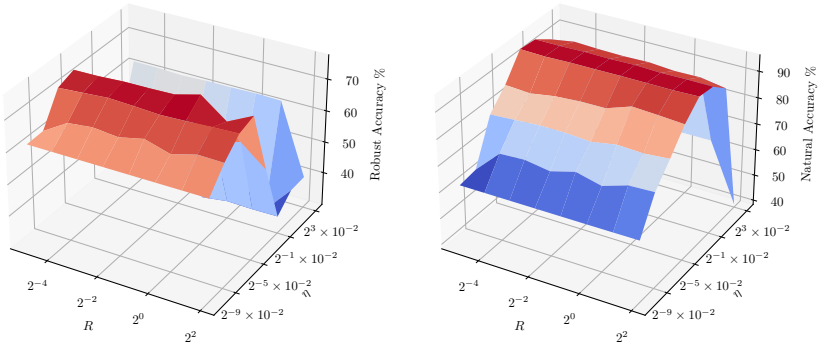

Figure 7: Robust and natural accuracy by η and R on CelebA with label 'Male'. We train a simple CNN with DP-Adam and test under 20 steps of $l_\infty(2/255)$ PGD attack. See details in Appendix D.

We claim that the diagonal pattern of the accuracy landscapes observed in Figure 5 (CIFAR10 & DP-Heavyball), Figure 11 (CIFAR10 & DP-SGD), Figure 12 (MNIST & DP-SGD) and Figure 13 (Fashion MNIST & DP-SGD) is not universal. For example, in Figure 7, we show that adaptive optimizers are much less sensitive to the clipping norm R, as the landscapes are characterized by the row-wise pattern instead of the diagonal pattern. This pattern is particularly obvious in the small R regime, where the robust and natural accuracy are high (see right panel in Figure 8).

To analyze the insensitivity to the clipping norm, we take the RMSprop(Tieleman et al., 2012) as an example, which complements the analysis of Adam in Bu et al. (2022b). For sufficiently small R, the private gradient in (2) becomes $\tilde{\boldsymbol{g}}_t = R \cdot \left(\sum_i \frac{\boldsymbol{g}_t(\boldsymbol{x}_i)}{||\boldsymbol{g}_t(\boldsymbol{x}_i)||_2} + \sigma \mathcal{N}(0, \mathbf{I}) \right) := R \cdot \hat{\boldsymbol{g}}_t$.

With private gradient $\tilde{\boldsymbol{g}}_t$, DP-RMSprop updates parameters $\boldsymbol{\theta}_t$ by $\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - \eta_t \frac{\tilde{\boldsymbol{g}}_t}{\sqrt{\tilde{\boldsymbol{v}}_t}}$, where $\tilde{\boldsymbol{v}}_t = \alpha \tilde{\boldsymbol{v}}_{t-1} + (1-\alpha)\tilde{\boldsymbol{g}}_t^2 = \sum_s^t (1-\alpha)\alpha^{t-s} \tilde{\boldsymbol{g}}_s^2 = R^2 \cdot \sum_s^t (1-\alpha)\alpha^{t-s} \hat{\boldsymbol{g}}_s^2$. Overall, we obtain an updating rule that is independent of the clipping norm R,

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - \eta_t \frac{R \cdot \hat{\boldsymbol{g}}_t}{\sqrt{R^2 \cdot \sum_s^t (1-\alpha)\alpha^{t-s} \hat{\boldsymbol{g}}_s^2}} = \boldsymbol{\theta}_{t-1} - \eta_t \frac{\hat{\boldsymbol{g}}_t}{\sqrt{\sum_s^t (1-\alpha)\alpha^{t-s} \hat{\boldsymbol{g}}_s^2}}.$$

As a result, DP optimizers can have fundamentally different landscapes with respect to the hyperparameters (R, η), and thus affect the accuracy and robustness (see Figure 9).

4.5 Large scale experiments on CelebA face datasets

We further validate our claims on CelebA (Liu et al., 2015), a public high-resolution (178×218 pixels) image dataset, consisting of 200000 real human faces that are supposed to be protected against privacy risks. We train ResNet18 (He et al. (2016), 11 million parameters) and Vision Transformer (ViT, Dosovitskiy et al. (2020), 6 million parameters) with DP-RMSprop. Both models are implemented by Wightman (2019) and pretrained on ImageNet.

	ResNet18				ViT			
attack magnitude	DP $\epsilon = 2$	DP $\epsilon = 4$	DP $\epsilon = 8$	Non-DP $\epsilon = \infty$	DP $\epsilon = 2$	DP $\epsilon = 4$	DP $\epsilon = 8$	Non-DP $\epsilon = \infty$
$\gamma = 0$	80.10%	85.10%	88.48%	91.91%	92.30%	92.33%	92.09%	92.87%
$\gamma = 2/255$	1.26%	0.47%	1.03%	1.19%	1.42%	2.02%	10.35%	0.08%
$\gamma = 4/255$	0.01%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
$\gamma = 8/255$	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
$\gamma = 16/255$	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

Table 3: Natural and robust accuracy on CelebA with label 'Smiling', DP-RMSprop, under 20 steps l_∞ PGD attack. Here the hyperparameters have not been carefully searched for the best robustness. See details in Appendix D.

	Natural	FGSM	BIM	PGD_∞	APGD_∞	PGD_2	APGD_2
DP , $\epsilon = 2$	80.10%	24.47%	1.24%	1.26%	1.18%	47.02%	46.25%
DP , $\epsilon = 4$	85.10%	24.40%	0.45%	0.47%	0.41%	56.92%	56.09%
DP , $\epsilon = 8$	88.48%	29.32%	0.97%	1.03%	0.41%	57.40%	56.69%
Non-DP	91.91%	22.94%	1.09%	1.19%	0.68%	66.89%	66.13%

Table 4: Natural and robust accuracy on CelebA with label 'Smiling' under general l_2/l_∞ adversarial attacks. Same ResNet18 as Table 3. See attack hyperparameters in Appendix D.

In Table 3 and Table 4, we observe that DP ResNet18 and ViT are almost as adversarially robust as their non-DP counterparts, if not more robust. These observations are consistent with those of simpler models on tiny images (c.f. Table 1 and Table 2).

5 Discussion

Through the lens of theoretical analysis and extensive experiments, we have shown that differentially private models can be adversarially robust and sometimes even more robust than the naturally-trained models. This phenomenon holds for various attacks with different magnitudes, from linear models to large vision networks, from grey-scale images to real face datasets, and from SGD to adaptive optimizers. We not only are the first to reveal this possibility of achieving privacy and robustness simultaneously, but also are the first to offer practical guidelines to address this important goal. We hope that our insights will excite the practitioners and boost their confidence to protect both the privacy and the robustness in real-world applications.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

Franziska Boenisch, Philip Sperl, and Konstantin Böttinger. Gradient masking and the underestimated robustness threats of differential privacy in deep learning. *arXiv preprint arXiv:2105.07985*, 2021.

Zhiqi Bu, Jinshuo Dong, Qi Long, and Weijie J Su. Deep learning with gaussian differential privacy. *Harvard data science review*, 2020(23), 2020.

Zhiqi Bu, Hua Wang, Qi Long, and Weijie J Su. On the convergence of deep learning with differential privacy. *arXiv e-prints*, pp. arXiv–2106, 2021.

Zhiqi Bu, Jialin Mao, and Shiyun Xu. Scalable and efficient training of large convolutional neural networks with differential privacy. *arXiv preprint arXiv:2205.10683*, 2022a.

Zhiqi Bu, Yu-Xiang Wang, Sheng Zha, and George Karypis. Automatic clipping: Differentially private deep learning made easier and stronger. *arXiv preprint arXiv:2206.07136*, 2022b.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pp. 2206–2216. PMLR, 2020.

Soham De, Leonard Berrada, Jamie Hayes, Samuel L Smith, and Borja Balle. Unlocking high-accuracy differentially private image classification through scale. *arXiv preprint arXiv:2204.13650*, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Jinshuo Dong, Aaron Roth, and Weijie Su. Gaussian differential privacy. *Journal of the Royal Statistical Society*, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407, 2014.

Logan Engstrom, Andrew Ilyas, Hadi Salman, Shibani Santurkar, and Dimitris Tsipras. Robustness (python library), 2019. URL `https://github.com/MadryLab/robustness`.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. In *International Conference on Learning Representations*, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

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

Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pp. 99–112. Chapman and Hall/CRC, 2018.

Alexey Kurakin, Steve Chien, Shuang Song, Roxana Geambasu, Andreas Terzis, and Abhradeep Thakurta. Toward training at imagenet scale with differential privacy. *arXiv preprint arXiv:2201.12328*, 2022.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners. *arXiv preprint arXiv:2110.05679*, 2021.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

H Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. *arXiv preprint arXiv:1710.06963*, 2017.

Harsh Mehta, Abhradeep Thakurta, Alexey Kurakin, and Ashok Cutkosky. Large scale transfer learning for differentially private image classification. *arXiv preprint arXiv:2205.02973*, 2022.

Felipe A Mejia, Paul Gamble, Zigfried Hampel-Arias, Michael Lomnitz, Nina Lopatina, Lucas Tindall, and Maria Alejandra Barrios. Robust or private? adversarial training makes models more vulnerable to privacy attacks. *arXiv preprint arXiv:1906.06449*, 2019.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

Harsha Nori, Rich Caruana, Zhiqi Bu, Judy Hanwen Shen, and Janardhan Kulkarni. Accuracy, interpretability, and differential privacy via explainable boosting. In *International Conference on Machine Learning*, pp. 8227–8237. PMLR, 2021.

Nicolas Papernot, Abhradeep Thakurta, Shuang Song, Steve Chien, and Ulfar Erlingsson. Tempered sigmoid activations for deep learning with differential privacy. *arXiv preprint arXiv:2007.14191*, pp. 10, 2020.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Liwei Song, Reza Shokri, and Prateek Mittal. Privacy risks of securing machine learning models against adversarial examples. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pp. 241–257, 2019.

Tijmen Tieleman, Geoffrey Hinton, et al. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning*, 4(2):26–31, 2012.

Florian Tramer and Dan Boneh. Differentially private learning needs better features (or much more data). *arXiv preprint arXiv:2011.11660*, 2020.

Nurislam Tursynbek, Aleksandr Petiushko, and Ivan Oseledets. Robustness threats of differential privacy. *arXiv preprint arXiv:2012.07828*, 2020.

Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

Han Xu, Xiaorui Liu, Yaxin Li, Anil Jain, and Jiliang Tang. To be robust or to be fair: Towards fairness in adversarial training. In *International Conference on Machine Learning*, pp. 11492–11501. PMLR, 2021.

Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1–9, 2022.

Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. *Advances in Neural Information Processing Systems*, 32, 2019.