Improved Rates of Differentially Private Nonconvex-Strongly-Concave Minimax Optimization

Ruijia Zhang [*1,2], Mingxi Lei [*3], Meng Ding [3], Zihang Xiang [4], Jinhui Xu [3], Di Wang [4,5]

[1] The Chinese University of Hong Kong, Shenzhen
[2] Johns Hopkins University
[3] State University of New York at Buffalo
[4] King Abdullah University of Science and Technology CEMSE
[5] Center of Excellence for Generative AI, KAUST

Abstract

In this paper, we study the problem of (finite sum) minimax optimization in the Differential Privacy (DP) model. Unlike most of the previous studies on the (strongly) convex-concave settings or loss functions satisfying the Polyak-Lojasiewicz condition, here we mainly focus on the nonconvex-strongly-concave one, which encapsulates many models in deep learning such as deep AUC maximization. Specifically, we first analyze a DP version of Stochastic Gradient Descent Ascent (SGDA) and show the utility bound in terms of the Euclidean norm of the gradient for the empirical risk function. We then propose a new method with less gradient noise variance and improve the upper bound to the best-known result for DP Empirical Risk Minimization with non-convex loss. We also discussed several lower bounds of private minimax optimization. Finally, experiments on AUC maximization, generative adversarial networks, and temporal difference learning with real-world data support our theoretical analysis.

Introduction

In recent years, minimax optimization has received great attention as it encompasses several basic machine learning and deep learning models such as generative adversarial networks (GANs) (Goodfellow et al. 2014; Creswell et al. 2018), deep AUC maximization (Yang and Ying 2022), distributionally robust optimization (Levy et al. 2020), and reinforcement learning (Sutton 1988), which have been widely used in different applications such as biomedicine and healthcare (Ling et al. 2022; Chen et al. 2022). The wide applications of minimax optimization also present privacy challenges in this problem as they always involve data with sensitive information. Differential Privacy (DP), introduced by (Dwork et al. 2006), has gained widespread recognition as a method for preserving privacy by adding a controlled amount of random noise to the data or query responses, thereby effectively concealing the details of any individual.

Recently, DP (finite sum) minimax optimization has been widely studied (see the related work section for details). However, compared to DP Empirical Risk Minimization (Wang, Ye, and Xu 2017; Wang et al. 2021; Wang and Xu 2019b), DP Minimax optimization is still in its early stages of development. Specifically, most of the previous work focuses on the case where the loss is either (strongly)-convex-(strongly)-concave (Yang et al. 2022; Zhang et al. 2022; Boob and Guzmán 2024; Bassily, Guzmán, and Menart 2023; González, Guzmán, and Paquette 2024; Zhou and Bassily 2024) or nonconvex but satisfying the Polyak-Łojasiewicz (PL) condition (Yang et al. 2022). However, compared to these settings, nonconvex minimax optimization is more widespread in deep neural networks, and all these methods are based on stability analysis and are hard to extend to the non-convex minimax problem. Thus, there is still lacking understanding when the loss is nonconvex, which motivates the study in this paper.

Recently, (Zhao et al. 2023) presented the first study on DP temporal difference learning, which can be formalized as a specific nonconvex-strong-concave minimax problem. However, several challenges remain: First, compared to the utility metrics of DP Empirical Risk Minimization, which always use first order or second order gradient of the objective function (Wang, Chen, and Xu 2019; Wang and Xu 2019a, 2021), the metric in (Zhao et al. 2023) cannot directly measure the stationariness of a model in general, indicating that it is hard to be explained whether the private model is good or not. Moreover, their utility metric has not been widely used in other related work for both minimax optimization and reinforcement learning, making it hard to compare with the non-private case and hard to use in general minimax optimization problems (see Theorem 5.2 in (Zhao et al. 2023) for details). Second, although in the ideal case (Zhao et al. 2023) shows that their utility will be close to the l_2-norm gradient of the objective function, they show a utility bound of $\tilde{O}(\frac{d^{1/8}}{(n\epsilon)^{1/4}})$, where $d = \max\{d_1, d_2\}$ with d_1 and d_2 are model dimensions and n is the sample size. It still has a gap with the best-known result $\tilde{O}(\frac{d^{1/3}}{(n\epsilon)^{2/3}})$ for DP Empirical Risk Minimization with non-convex loss (Murata and Suzuki 2023; Tran and Cutkosky 2022). Finally, their approach is only tailored for temporal difference learning, and it is unknown whether it can be extended to general minimax problems.

To address the aforementioned issues, this paper revisits the DP minimax optimization problem in the nonconvex-strong-concave (NC-SC) setting, offering a more general and

*These authors contributed equally.

enhanced analysis. Our contributions can be summarized as follows:

1. When the loss function is Lipschitz and smooth, we first show that by modifying the classical Stochastic Gradient Descent Ascent (SGDA) algorithm, it is possible to get an (ϵ, δ)-DP model whose l_2-norm of the gradient for the empirical risk function is upper bounded by $\tilde{O}(\frac{d^{1/4}}{(n\epsilon)^{1/2}})$.
2. The primary weakness of DP-SGDA is that it relies on using noise of the same scale to ensure differential privacy, which results in excessive variance and an unsatisfactory utility bound. To address this issue, we leverage the gradient difference between the current and previous models to adjust the noise scale. This approach allows us to add less noise as the iterations progress since the gradient difference tends to diminish. Specifically, we propose a novel method called PrivateDiff Minimax and demonstrate that its output can achieve an upper bound of $\tilde{O}(\frac{d^{1/3}}{(n\epsilon)^{2/3}})$, which matches the best-known result for DP Empirical Risk Minimization with non-convex loss.
3. We also provide a preliminary study on the lower bounds of private minimax optimization. Specifically, for finite sum minimax problems, we show that there exists an instance such that for any (ϵ, δ)-DP model, its l_2-norm gradient is lower bounded by $\Omega(\frac{\sqrt{d}}{n\epsilon})$. Moreover, for the group distributional robust optimization problem, its utility is lower bounded by $\Omega(\frac{d\sqrt{d}}{n\epsilon})$.
4. Finally, we conduct experiments on AUC maximization, generative adversarial networks, and temporal difference learning with real-world data. Our results demonstrate that our method, PrivateDiff Minimax, outperforms other approaches across various datasets and privacy budgets, providing empirical support for our theoretical analysis.

Related Work

DP Minimax Optimization. (Yang et al. 2022) provides the first study on DP stochastic minimax optimization. Specifically, for the convex-(strongly)-concave case, they provide upper bounds in terms of weak primal-dual population risk, which match the optimal rates for DP Stochastic Convex Optimization (Su, Hu, and Wang 2024; Su, Zhao, and Wang 2023; Hu et al. 2022; Huai et al. 2020; Wang et al. 2020; Xue et al. 2021; Tao et al. 2022). They further consider the NC-SC case where the loss satisfies the PL condition. However, as their analysis is based on algorithmic stability, it is difficult to extend to general NC-SC loss, which is studied in this paper. (Zhang et al. 2022) also studies the convex-(strongly)-concave case and provides a linear-time algorithm, which can also achieve optimal rates. (Boob and Guzmán 2024) considers both convex-concave minimax optimization and stochastic variational inequality, it provides both strong and weak primal-dual population risks. Recently (Bassily, Guzmán, and Menart 2023) justifies that the (strong) primal-dual gap is a more meaningful and challenging efficiency estimate for DP convex-concave minimax optimization. Very recently, (González, Guzmán, and Paquette 2024) considers the convex-concave case where the constrain sets are polyhedral; it provides utility bounds that are independent of the polynomial of the model dimension. (Zhou and Bassily 2024) considers the DP worst-group risk minimization with convex loss, which is a specific instance of minimax optimization, and provides both upper and lower bounds of the problem.

To the best of our knowledge, (Zhao et al. 2023) is the only paper that studies the general NC-SC case of stochastic minimax optimization. However, as mentioned previously, their utility has been only used in reinforcement learning rather than in other minimax optimization problems. In our paper, we consider the gradient norm as the utility, which is more natural and has been widely used in both non-private studies and the DP nonconvex case (Wang, Chen, and Xu 2019; Xiao et al. 2023; Wang and Xu 2019a; Wang et al. 2023; Murata and Suzuki 2023; Tran and Cutkosky 2022).

Nonconvex Minimax Optimization. As there is a long list of work on minimax optimization, here we only focus on the ones that consider the NC-SC setting. Previous work mainly focuses on improving the gradient complexity or number of loops (Nouiehed et al. 2019; Lin, Jin, and Jordan 2020a,b; Lu et al. 2020; Zhang, Aybat, and Gurbuzbalaban 2022; Boţ and Böhm 2023; Sharma et al. 2022; Guo et al. 2021; Yan et al. 2020; Xu et al. 2023; Luo et al. 2020). For example, (Lin, Jin, and Jordan 2020a) shows the local convergence of SGDA w.r.t. the gradient norm if the stepsizes are chosen appropriately, which motivates our first algorithm DP-SGDA. (Luo et al. 2020) provides a variance reduction-based approach to accelerate SGDA further. It is notable that our second method is quite different from all these non-private methods. Specifically, our approach is still based on SGDA. However, we use the gradient difference between the current and previous models to reduce the variance of added noise. This makes us add less noise as the iteration increases since the gradient difference tends to be zero. Thus, even from the optimization point of view, our method is still of interest.

Preliminaries

Differential Privacy

Definition 1 (Differential Privacy (Dwork et al. 2006))**.** *Given a data universe $\mathcal{Z}$, we say that two datasets $D, D' \subseteq \mathcal{Z}$ are neighbors if they differ by only one entry, which is denoted as $D \sim D'$. A randomized algorithm $\mathcal{A}$ is (ϵ, δ)-differentially private (DP) if for all neighboring datasets D, D' and for all events E in the output space of $\mathcal{A}$, the following holds*

$$\mathbb{P}(\mathcal{A}(D) \in E) \leq e^{\epsilon}\mathbb{P}(\mathcal{A}(D') \in E) + \delta.$$

If $\delta = 0$, we call algorithm $\mathcal{A}$ is ϵ-DP.

In this paper, we focus on (ϵ, δ)-DP and mainly use the Gaussian mechanism and moment accountant (Abadi et al. 2016) to guarantee the DP property.

Definition 2 (l_2-sensitivity)**.** *Given a function $q : \mathcal{Z} \to \mathbb{R}^d$, we say q has $\Delta_2(q)$ l_2-sensitivity if for any neighboring datasets D, D' we have $\|q(D) - q(D')\|_2 \leq \Delta_2(q)$.*

Definition 3 (Gaussian Mechanism)**.** *Given any function $q : \mathcal{Z} \to \mathbb{R}^d$, the Gaussian mechanism is defined as $q(D) + \xi$*

where $\xi \sim \mathcal{N}(0, \frac{8\Delta_2^2(q)\log(1.25/\delta)}{\epsilon^2}\mathbb{I}_d)$, Gaussian mechanism preserves (ϵ, δ)-DP for $0 < \epsilon, \delta \leq 1$.

Definition 4. *For an (randomized) algorithm $\mathcal{A}$ and neighboring datasets D, D', the λ-th moment is given as*

$$\alpha_{\mathcal{A}}(\lambda, D, D') = \log \mathbb{E}_{O\sim\mathcal{A}(D)}[(\frac{\mathbb{P}[\mathcal{A}(D) = O]}{\mathbb{P}[\mathcal{A}(D') = O]})^{\lambda}].$$

The moment accountant is then defined as

$$\alpha_{\mathcal{A}}(\lambda) = \sup_{D,D'} \alpha_{\mathcal{A}}(\lambda, D, D').$$

Lemma 1. *(Abadi et al. 2016) Consider a sequence of mechanisms $\{\mathcal{A}_t\}_{t\in[T]}$ and the composite mechanism $\mathcal{A} = (\mathcal{A}_1, \cdots, \mathcal{A}_T)$. We have the following properties:*
(a) [Composability] For any λ,

$$\alpha_{\mathcal{A}}(\lambda) = \sum_{t=1}^{T} \alpha_{\mathcal{A}_t}(\lambda).$$

(b) [Tail bound] For any ϵ, the mechanism $\mathcal{A}$ is (ϵ, δ) differentially private for

$$\delta = \min_{\lambda} \alpha_{\mathcal{A}}(\lambda) - \lambda\epsilon.$$

Lemma 2 (Privacy Amplification via Subsampling (Balle, Barthe, and Gaboardi 2018))**.** *Consider a sequence of mechanisms $\mathcal{A}_t = q_t(D_t) + \xi_t$ where $\xi_t \sim \mathcal{N}(0, \sigma^2\mathbb{I})$. Here each function $q_t : \mathcal{Z} \to \mathbb{R}^d$ has l_2-sensitivity of 1. And each D_t is a subsample of size m obtained by uniform sampling without replacement from space $\mathcal{Z}$, i.e. $D_t \sim (\mathrm{Unif}(D))^m$. Then we have*

$$\alpha_{\mathcal{A}_t}(\lambda) \leq \frac{m^2 n\lambda(\lambda + 1)}{n^2(n - m)\sigma^2} + \mathcal{O}(\frac{m^3\lambda^3}{n^3\sigma^3}).$$

Minimax Optimization

Given a dataset $D = \{z_1, \cdots, z_n\} \in \mathcal{Z}^n$ and a loss function $f : \mathcal{X} \times \mathcal{Y} \times \mathcal{Z} \mapsto \mathbb{R}$, a (finite sum) minimax optimization problem aims to optimize the following empirical risk function:

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \hat{L}(x, y; D) := \frac{1}{n}\sum_{i=1}^{n} f(x, y; z_i), \qquad (1)$$

where $\mathcal{X}$ and $\mathcal{Y}$ are the constrained sets. If each z_i is i.i.d. sampled from some underlying distribution $\mathcal{Z}$, then we further aim to optimize the population risk:

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} L(x, y; D) := \mathbb{E}_{\mathcal{Z}}[L(x, y; z)]. \qquad (2)$$

In this paper, we mainly focus on the empirical risk function.

Recall that the minimax problem (1) is equivalent to minimizing the function $\Phi(\cdot) = \max_{y\in\mathcal{Y}} \hat{L}(\cdot, y)$. For nonconvex strongly concave minimax problems in which $\hat{L}(x, \cdot)$ is strongly concave in y for each $x \in \mathcal{X}$, the maximization problem $\max_{y\in\mathcal{Y}} \hat{L}(x, y)$ can be solved efficiently and provides useful information about Φ. However, it is NP-hard to find the global minimum of Φ in general when Φ is nonconvex, which is considered in our paper. In this work, we hope to find an approximate first-order stationary point instead, which has been widely adopted in previous literature (Lin, Jin, and Jordan 2020a).

Definition 5. *A point x is an ϵ-stationary point ($\epsilon \geq 0$) of a differentiable function Φ if $\|\nabla\Phi(x)\| \leq \epsilon$. If $\epsilon = 0$, then x is a stationary point.*

Note that there are also other metrics for stationary points (Lu et al. 2020; Nouiehed et al. 2019); however, these notions are weaker than $\|\nabla\Phi(\cdot)\|$. From the above definitions, it is clear that DP minimax optimization aims to develop an (ϵ, δ)-DP algorithm whose output $(x^{\text{priv}}, y^{\text{priv}})$ makes $\|\nabla\Phi(x^{\text{priv}})\|$ be as small as possible. In this paper, we focus on the nonconvex-strongly-convex (NC-SC) setting and we impose the following assumptions.

Definition 6. *A function g is G-Lipschitz if for $\forall x, x' \in \mathcal{X}$, we have $\|g(x) - g(x')\| \leq G\|x - x'\|$.*

Definition 7. *A function g is l-smooth if for $\forall x, x' \in \mathcal{X}$, we have $\|\nabla g(x) - \nabla g(x')\| \leq l\|x - x'\|$.*

Definition 8. *A function g is μ-strongly convex if for $\forall x, x' \in \mathcal{X}$, we have $\langle\nabla g(x) - \nabla g(x'), x - x'\rangle \geq \mu\|x - x'\|_2^2$. A function g is μ-strongly concave if $-g$ is μ-strongly convex.*

Assumption 1. *For any fixed $x \in \mathcal{X}$, $\hat{L}(x, \cdot; D)$ is μ-strongly concave in y. Moreover, we assume $\mathcal{X} = \mathbb{R}^{d_1}$ and $\mathcal{Y} \subseteq \mathbb{R}^{d_2}$ is a convex and bounded set with diameter Λ (we denote $d = \max\{d_1, d_2\}$). We also assume $f(\cdot, \cdot; z_i) \leq M$.*

Assumption 2. *There exist G_x, G_y such that, for any $x \in \mathcal{X}, y \in \mathcal{Y}$, function $f(\cdot, y; z_i)$ is G_x-Lipschitz and function $f(x, \cdot; z_i)$ is G_y-Lipschitz. Denote $G = \max\{G_{\mathbf{w}}, G_{\mathbf{v}}\}$.*

Assumption 3. *There exists a constant l_x and l_y such that for any $x \in \mathcal{X}, y \in \mathcal{Y}$, function $\hat{L}(\cdot, y; D)$ is l_x-smooth and function $\hat{L}(x, \cdot; D)$ is l_y-smooth. Denote $l = \max\{l_x, l_y\}$.*

Assumption 4. *For randomly drawn $j \in [n]$, the gradients $\nabla_x f(x, y; z_j)$ and $\nabla_y f(x, y; z_j)$ have bounded variances B_x and B_y respectively. Let $\mathcal{B} = \max\{B_x, B_y\}$.*

We present a technical lemma on the structure of function Φ, which is essential for the convergence analysis.

Lemma 3 ((Lin, Jin, and Jordan 2020a))**.** *Under Assumption 1 and 3, $\Phi(\cdot) = \max_{y\in\mathcal{Y}} \hat{L}(\cdot, y)$ is $(l + \kappa l)$-smooth, where $\kappa = \frac{l}{\mu}$ is the condition number. Moreover, for any $x \in \mathcal{X}$, $\nabla\Phi(x) = \nabla_x\hat{L}(x, y^\star(x))$, where $y^\star(x) = \arg\max_{y\in\mathcal{Y}} \hat{L}(x, y)$ and $y^\star(\cdot)$ is κ-Lipschitz.*

A Preliminary Exploration

An Upper Bound via DP-SGDA

In the non-private case, a natural approach to solving the minimax problem is the gradient descent ascent (GDA). However, when privacy is a concern, directly applying GDA can lead to significant privacy risks. To address this, we explore a differentially private version of stochastic GDA (DP-SGDA) in this section, providing a preliminary analysis of our problem while ensuring privacy is maintained.

DP-SGDA (Algorithm 1) was proposed by (Yang et al. 2022). Their analysis relies on the algorithm's stability within the convex-concave setting, which can not extend to our nonconvex-strongly-concave (NC-SC) case. In the following, we provide a more general analysis tailored for the NC-SC setting.

Algorithm 1: Differentially Private Stochastic Gradient Descent Ascent (DP-SGDA)

Require: Dataset D, privacy budget ϵ, δ, iteration number T, learning rates $\{\eta_x, \eta_y\}$, initialization (x_0, y_0), clipping thresholds C_1, C_2.

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

Draw a collection of i.i.d. data samples $\{z_t^j\}_{j=1}^m$ uniformly without replacement.

Sample independent noises $\xi_t \sim \mathcal{N}(0, \sigma_x^2 I_{d_1})$ and $\zeta_t \sim \mathcal{N}(0, \sigma_y^2 I_{d_2})$.

Update x_{t+1}:
$x_{t+1} = x_t - \eta_x(\frac{1}{m}\sum_{j=1}^m \text{Clipping}(\nabla_x f(x_t, y_t; z_t^j), C_1) + \xi_t)$.

Update y_{t+1}:
$y_{t+1} = \Pi_{\mathcal{Y}}(y_t + \eta_y(\frac{1}{m}\sum_{j=1}^m \text{Clipping}(\nabla_y f(x_t, y_t; z_t^j), C_2) + \zeta_t)$.

end for

return $(x^{\text{priv}}, y^{\text{priv}}) \in \{(x_0, y_0), \cdots, (x_T, y_T)\}$ where the tuple is uniformly sampled.

Algorithm 2: Clipping (x, C)

Require: x and clipping threshold $C > 0$.

$\hat{x} = \min\{\frac{C}{\|x\|_2}, 1\}x$

return $\hat{x}$.

Theorem 1. *There exist constants c_1, c_2 and $c_3 > 0$ such that given the mini-batch size m and total iterations T, for any $\epsilon < c_1 m^2 T/n^2$ and $0 < \delta < 1$, Algorithm 1 is (ϵ, δ)-DP if we set*

$$\sigma_x = \frac{c_2 C_1 \sqrt{T \log(1/\delta)}}{n\epsilon}, \sigma_y = \frac{c_3 C_2 \sqrt{T \log(1/\delta)}}{n\epsilon}. \quad (3)$$

In practice, a set of parameters applicable to Theorem 1 is provided by (Yang et al. 2022; Abadi et al. 2016) to ensure the privacy guarantee. By setting $\epsilon \leq 1, \delta \leq 1/n^2$ and $m = \max(1, n\sqrt{\epsilon/(4T)})$, the explicit values for the variances are given as $\sigma_x = \frac{8\sqrt{T\log(1/\delta)}}{n\epsilon}, \sigma_y = \frac{8\sqrt{T\log(1/\delta)}}{n\epsilon}$.

Next, we show an improved utility bound of Algorithm 1.

Theorem 2. *Suppose Assumptions 1-4 hold. If we choose parameters satisfying: iterations $T = \Theta(\frac{n\epsilon}{\sqrt{d\log(1/\delta)}})$, clipping thresholds $C_1 \geq G_x, C_2 \geq G_y$, step sizes $\eta_x = O(\frac{1}{l\kappa^2})$, $\eta_y = O(\frac{1}{l})$ and batch size $m = O(\frac{n\epsilon}{\sqrt{d\log(1/\delta)}})$, then the output of DP-SGDA satisfies*

$$\mathbb{E}\|\nabla\Phi(x^{priv})\| \leq O\left(\frac{(d\log(1/\delta))^{1/4}}{\sqrt{n\epsilon}}\right), \quad (4)$$

where O hides other terms related to $G, \ell, \mathcal{B}, \mu$ and κ.

Technical Overview Although the idea of DP-SGDA is natural, our utility analysis is highly non-trivial. Specifically, our proof needs to set a pair of stepsizes (η_x, η_y), which updates $\{y_t\}_{t\geq 1}$ significantly faster than that of $\{x_t\}_{t\geq 1}$. Recall that $y^\star(\cdot)$ is κ-Lipschitz in Lemma 3:

$$\|y^\star(x_1) - y^\star(x_2)\| \leq \kappa\|x_1 - x_2\|.$$

Consequently, if $\{x_t\}_{t\geq 1}$ changes slowly, it follows that its corresponding sequence $y^*(x_t)$ also evolves gradually. Therefore, This allows us to perform gradient descent analysis on the strongly concave function $\hat{L}(x_t, \cdot; D)$, albeit it is changing slowly. Additionally, by defining the error as $\theta_t = \|y^\star(x_t) - y_t\|^2$, we can first apply the descent lemma to $\Phi(x)$. Then, by performing telescoping, we obtain the following inequality:

$$\mathbb{E}\Phi(x_{T+1}) - \Phi(x_0) \leq -\Omega(\eta_x)(\sum_{t=0}^T \mathbb{E}\|\nabla\Phi(x_t)\|^2)$$
$$+ O(\eta_x) + O(\eta_x)[\sum_{t=0}^T \mathbb{E}\|\xi_t\|_2^2 + \mathbb{E}\|\zeta_t\|_2^2] + O(\frac{T\eta_x}{m}).$$

Thus, $\sum_{t=0}^T \|\nabla\Phi(x_t)\|^2$ can be upper bounded by the last term on the right-hand side, which is the desired utility bound.

Remark 1. *Note that when there is no variable y, then DP-SGDA will be reduced to DP-SGD in (Wang, Ye, and Xu 2017). Moreover, the bound $\tilde{O}(\frac{\sqrt[4]{d}}{\sqrt{n\epsilon}})$ aligns with the bounds provided in previous work on DP Empirical Risk Minimization with non-convex loss, such as (Wang, Ye, and Xu 2017; Wang et al. 2023). While (Yang et al. 2022) considered DP-SGDA for non-convex loss under the PL condition, our approach differs in the choice of stepsize: we use a constant stepsize throughout all iterations, whereas (Yang et al. 2022) require the stepsize to decay with respect to the iteration number.*

Lower Bounds of the DP-Minimax Problem

We now show a lower bound $\Omega(\frac{d\log(1/\delta)}{n^2\epsilon^2})$ for the utility under differential privacy in the case where $\mathcal{X} = \mathbb{R}^{d_1}$ and $\mathcal{Y}$ is a bounded convex set. Our lower bound matches the current best-known lower bound for DP-ERM with non-convex loss (Arora et al. 2023) and holds even for convex functions.

Theorem 3. *Given $n, \epsilon = O(1)$, $2^{-\Omega(n)} \leq \delta \leq 1/n^{1+\Omega(1)}$, there exists a convex set $\mathcal{Y} \subseteq R^{d_2}$, a loss function $\hat{L} : \mathbb{R}^{d_1} \times \mathcal{Y} \times \mathcal{Z}^n \mapsto \mathbb{R}$ satisfying Assumption 1-3 with $\mu, G, l = O(1)$ and a dataset D of n samples such as for any (ϵ, δ)-DP algorithm with output (x^{priv}, y^{priv}) satisfies*

$$\|\nabla\Phi(x^{priv})\| \geq \Omega(\min\{1, \frac{\sqrt{d\log(1/\delta)}}{n\epsilon}\}).$$

It is notable that this result implies the same lower bound (up to logarithmic factors) for the population gradient using the technique in (Bassily et al. 2019). Furthermore, the aforementioned lower bound applies specifically to minimax problems in finite-sum form, as described in (1). However, different lower bounds may be derived for specific problems that cannot be expressed in this form. For instance, consider the (regularized) worst-group risk minimization problem:

$$\min_{x\in\mathbb{R}^{d_2}} \max_{y\in\Delta_{d_2}} \hat{L}(x, y; D) = \sum_{i=1}^{d_2} y_i \hat{L}_{D_i}(x) - \frac{1}{2}\|y\|_2^2, \quad (5)$$

where $\Delta_{d_2} = \{y \in [0,1]^{d_2} : \|y\|_1 = 1\}$, $D = \bigcup D_i$, $D_i \bigcap D_j = \emptyset$ for $i \neq j$, and $\hat{L}_{D_i}(x) = \frac{1}{|D_i|}\sum_{z\in D_i} \hat{L}(x; z)$.

Theorem 4. *Given $n, \epsilon = O(1)$, $2^{-\Omega(n)} \leq \delta \leq 1/n^{1+\Omega(1)}$, there exists a convex set $\mathcal{Y} \subseteq R^{d_2}$, a Lipschitz and smooth loss function $\hat{L} : \mathbb{R}^{d_1} \times \mathcal{Z} \mapsto \mathbb{R}$ and a dataset S of n samples such as for any (ϵ, δ)-DP algorithm with output (x^{priv}, y^{priv}) satisfies*

$$\|\nabla \Phi(x^{priv})\| \geq \Omega(\min\{1, \frac{d\sqrt{d\log(1/\delta)}}{n\epsilon}\}).$$

Improved Rate via PrivateDiff Minimax

As discussed in the previous section, there remains a gap of $\widetilde{O}(\frac{d^{1/4}}{\sqrt{n\epsilon}})$ between the upper and lower bounds. In this section, we aim to bridge this gap. Specifically, our goal is to develop a method that achieves a rate of $\widetilde{O}(\frac{d^{1/3}}{(n\epsilon)^{2/3}})$.

Our key observation is that the utility of Algorithm 1 heavily depends on the noise variance we add in each iteration. Notably, the scale of its noise variance is proportional to the l_2-norm sensitivity of the gradient, which is upper bounded by the smoothness constant of the function. Thus, by using the composition theorem, adding the same scale of noise in each iteration in Algorithm 1 can guarantee DP. From the utility side, this is fine for variable y as $\hat{L}(x, \cdot; D)$ is strongly concave, and it is known that DP-SGD with the same scale of noise in each iteration can achieve the optimal rate (Bassily et al. 2019). However, such a strategy is only sub-optimal for variable x, which corresponds to a nonconvex loss $\hat{L}(\cdot, y; D)$. As a result, we propose an algorithm called PrivateDiff Minimax, which focuses on improving the performance for x.

Main Idea: In essence, PrivateDiff Minimax updates variable y and variable x alternatively within each iteration. Suppose in the r-th iteration, we have $(x_r, \tilde{y}_r)$ after update. For variable y, due to the strong convexity on the maximization side, we can directly update it at the beginning of each iteration and get a temporary $\tilde{y}_{r+1}$. Subsequently, our algorithm involves building a private estimator $\widetilde{v}_r$ to approximate the $\nabla_x \hat{L}(x_r, \tilde{y}_{r+1}; D)$. Generally speaking, $\widetilde{v}_r$ accumulates stochastic gradient differences between two consecutive iterations. In detail, we begin with the following equation

$$\begin{aligned}\nabla_x \hat{L}(x_r, \tilde{y}_{r+1}; D) &= \nabla_x \hat{L}(x_r, \tilde{y}_{r+1}; D) \\ &- \nabla_x \hat{L}(x_{r-1}, \tilde{y}_r; D) + \nabla_x \hat{L}(x_{r-1}, \tilde{y}_r; D).\end{aligned}$$

We use a stochastic gradient ascent algorithm to update $\tilde{y}_r$ to $\tilde{y}_{r+1}$. In doing so, we can approximate $\nabla_x \hat{L}(x_r, \tilde{y}_{r+1}; D)$ and $\nabla_x \hat{L}(x_{r-1}, \tilde{y}_r; D)$ by $\nabla\Phi(x_r)$ and $\nabla\Phi(x_{r-1})$ respectively. This approximation is accurate up to some controlled error term by Lemma 3 and the convergence rate of SGDA. Moreover, since $\Phi(\cdot)$ is smooth, indicating that $|\nabla\Phi(x_r) - \nabla\Phi(x_{r-1})| \leq O(1)\|x_r - x_{r-1}\|$. This means that we can add a noise whose variance ξ_{x_r} is proportional to $\|x_r - x_{r-1}\|$ to ensure the differential privacy. In total, we have

$$\widetilde{v}_r \approx (\nabla\Phi(x_r) - \nabla\Phi(x_{r-1}) + \xi_{x_r}) + \widetilde{v}_{r-1}. \qquad (6)$$

with initial $\widetilde{v}_0 := 0$. Previously, the l_2-sensitivity of the private estimator in Algorithm 1 is bounded by the whole gradient's Lipschitz constant. It is now bounded by the distance of x_r and x_{r-1}. Therefore, it can be much smaller than the gradient's l_2-sensitivity when x_r and x_{r-1} are near enough. Hence, our algorithm's gradient differences accumulation design breeds the ability to add smaller noise variance while preserving privacy.

Algorithm Layouts: Algorithm 3 is the detailed implementation of our above idea. In each iteration, we first leverage a clipped version of Mini-batch SGA for strongly concave loss function $\hat{L}(x_r, \cdot; D)$ with initialization y_r to get a non-private version of y_{r+1} (step 3). We then need to construct a private estimator $\widetilde{v}_{r+1}$ to approximate the gradient $\nabla_x \hat{L}(x_r, y_{r+1}; D)$. To get this, our framework restarts every T round, where the length of T is carefully controlled in pursuit of optimal utility bound in our analysis. Specifically, for every T iteration, we will calculate a subsampled gradient $\mathbf{d_r}$, which is a base state analogous to the initial differential private gradient estimator $\tilde{v}_1 = \mathbf{d_0} = \frac{1}{m}\sum_{j=1}^{m} \text{Clipping}(\nabla_x f(x_0, y_1; z_0^j), C_1)$. Note that such a restart mechanism is essential as it can significantly reduce the noise we add since we just need to add noise with whose scale depends on the Lipschitz constant every T iterations.

If $r\%T \neq 0$, we then leverage (6) to recursively update v_{r+1} via adding v_r with the gradient difference $\mathbf{d_r} = \frac{1}{m}\sum_{j=1}^{m}\text{Clipping} \; (\nabla_x f(x_r, y_{r+1}; z_r^j) - \nabla_x f(x_{r-1}, y_r; z_{r-1}^j), C_{2,r})$(step 7). Subsequently, We add noise to v_{r+1} to ensure DP. Note that when $r\%T \neq 0$ the noise scale only depends on $\|x_r - x_{r-1}\|$ and a small constant C_3 that corresponds to the convergence rate of SGA.

The private estimator $\widetilde{v}_r$ is then utilized by performing gradient descent on x_r to get new x_{r+1}. After that, we perform output perturbation on y_r to get the final private version $\tilde{y}_{r+1}$. In the following, we provide privacy and utility guarantees.

Theorem 5. *Under Assumption 1, there exist constants c_4, c_5, c_6 and $c_7 > 0$ so that given the mini-batch size m, restart interval T and total iterations R, for any $\epsilon < c_4 m^2 T/n^2$, Algorithm 3, is (ϵ, δ)-differentially private for any $\delta > 0$ if we choose*

$$\sigma_{x_1} = \frac{c_5\sqrt{\frac{R}{T}\log(1/\delta)}}{n\epsilon}, \sigma_{x_2} = \frac{c_6\sqrt{R\log(1/\delta)}}{n\epsilon}$$
$$\text{and} \quad \sigma_y = c_7\frac{(2C_0^2 + \beta M)\sqrt{R\log(1/\delta)}}{n\epsilon}.$$

Remark 2. *We give a set of parameters applicable to Theorem 5 here in practice. By setting $\epsilon \leq 1, \delta \leq 1/n^2$ and $m = \max(1, n\sqrt{\epsilon/(8T)})$, then explicit values for the variances are: $\sigma_{x_1} = \frac{4\sqrt{\frac{R}{T}\log(1/\delta)}}{n\epsilon}, \sigma_{x_2} = \frac{4\sqrt{R\log(1/\delta)}}{n\epsilon}, \sigma_y = \frac{4(2C_0^2+\beta M)\sqrt{R\log(1/\delta)}}{\mu n\epsilon}$.*

Since our mechanism can reduce the noise scale and thus the variance of the private gradient estimator $\widetilde{v}_r$. Therefore, we expect a better utility bound than the standard DP-SGDA, which is formally stated as the following.

Theorem 6. *Let $\varepsilon \in (0, \frac{1}{e})$ and suppose Assumptions 1-4 hold. In Algorithm 3 and under the choices of $\sigma_{x_1}^2$, $\sigma_{x_2}^2$, and σ_y in Theorem 5, if we further set $C_0 \geq G_y$,* Algorithm 3: PrivateDiff Minimax

Require: Initial Point $x_0, \tilde{y}_0$, dateset D, learning rates η_x, noise variance $\sigma_{x_1}^2, \sigma_{x_2}^2, \sigma_y^2$, clipping radius C_0, C_1, C_2 and C_3, iteration number T_1, T_2 and R, batch size m.

for $r = 0, 1, 2, 3 \ldots, R$ **do**
  Draw a collection of i.i.d. data samples $\{z_r^j\}_{j=1}^m$ uniformly without replacement.
  $y_r = \tilde{y}_r$
  $y_{r+1} =$ Mini-batch SGA$(\hat{L}(x_r, y_r; D), T_2, C_0)$
  if r % T=0 then
    $\mathbf{d_r}$= $\frac{1}{m}\sum_{j=1}^m$ Clipping$(\nabla_x f(x_r, y_{r+1}; z_r^j), C_1)$.
    Set $\sigma_x = \sigma_{x_1}, C = C_1$ and $\widetilde{v}_r = 0$;
  else:
    $\mathbf{d_r} = \frac{1}{m}\sum_{j=1}^m$Clipping $(\nabla_x f(x_r, y_{r+1}; z_r^j) - \nabla_x f(x_{r-1}, y_r; z_{r-1}^j), C_{2,r})$
    Set $\sigma_x = \sigma_{x_2}$and $C = C_{2,r} = C_2\|x_r - x_{r-1}\| + C_3$.
  end if
  Set $v_{r+1} = \mathbf{d_r} + \widetilde{v}_r$ and $\widetilde{v}_{r+1} = v_{r+1} + \xi_{x_{r+1}}$, where $\xi_{x_{r+1}} \sim N(0, \sigma_x^2 C^2 I_{d_1})$.
  $x_{r+1} = x_r - \eta_x \widetilde{v}_{r+1}$.
  $\tilde{y}_{r+1} = y_{r+1} + \zeta$, where $\zeta \sim \mathcal{N}(0, \sigma_y^2 I_{d_2})$.
end for
return $(x^{\text{priv}}, y^{\text{priv}}) \in \{(x_1, \tilde{y}_1), \cdots, (x_R, \tilde{y}_R)\}$ where the tuple is uniformly sampled.

Algorithm 4: Mini-batch Stochastic Gradient Ascent (Mini-batch SGA)

Require: Fixed x, step size η_{y_i}, initial point $y_0' = y$, number of iterations T_2, clipping threshold C_0.

for $i = 0, 1, 2, 3 \ldots, T_2$ **do**
  Draw a collection of i.i.d. data samples $\{z_i^j\}_{j=1}^m$ uniformly without replacement.
  Update y_{i+1}' as $y_{i+1}' = \Pi_{\mathcal{Y}}(y_i' + \frac{\eta_{y_i}}{m}\sum_{j=1}^m$ Clipping$(\nabla_y f(x, y_i'; z_i^j), C_0))$.
end for
Return y_{T_2}'.

$C_1 \geq G_x$, $C_2 \geq l + \kappa l$ and $C_3 = \tilde{O}(\frac{1}{\sqrt{T_2}})$; the stepsizes $\eta_x = O(\min\{\frac{1}{l+\kappa l}, \frac{1}{\sqrt{T} l \sigma_{x_2}\sqrt{d}}\}), \eta_{y_i} = \frac{1}{\mu i}$; the restart interval $T = \Theta\left((\frac{\sqrt{d}}{n\epsilon})^{2/3} R\right)$, total number of rounds $R = \widetilde{\Theta}\left(\max\{\frac{1}{\varepsilon_{\text{opt}}}, (\frac{d}{n^2\epsilon^2\varepsilon_{\text{opt}}^2}\}\right)$ with $\varepsilon_{\text{opt}} := O(\frac{d^{\frac{2}{3}}}{(n\epsilon)^{\frac{4}{3}}})$, number of iterations of Mini-batch SGA $T_2 = O\left(\max\{\frac{(n\epsilon)^{4/3}}{d^{2/3}}, TR \cdot \frac{d^{1/3}}{(n\epsilon)^{2/3}}\}\right)$ and the batch size $m = O(\frac{(n\epsilon)^{4/3}}{d^{2/3}})$, with probability at least $1 - \vartheta$, the utility bound of PrivateDiff Minimax satisfies

$$\mathbb{E}\|\nabla\Phi(x^{priv})\| \leq \widetilde{O}(\frac{(d\log\frac{1}{\delta})^{1/3}}{(n\epsilon)^{\frac{2}{3}}}).$$

The obtained utility is significantly better than the best-known utility bound $\widetilde{O}(d^{\frac{1}{4}}/\sqrt{n\epsilon})$ when $n \geq \Omega(\sqrt{d}/(G\epsilon))$. Note that by some appropriate choice of the thresholds, one can show that the clipping has no effect. Moreover, we can see there are two terms in $C_{2,r}$ where the first term corresponds to the upper bound of $|\nabla\Phi(x_r) - \nabla\Phi(x_{r-1})|$ and the second one is the convergence error caused by y_{r+1}. Thus, when T_2 is large enough, the noise σ_{x_2} could be very small if x_r is close enough to x_{r-1}.

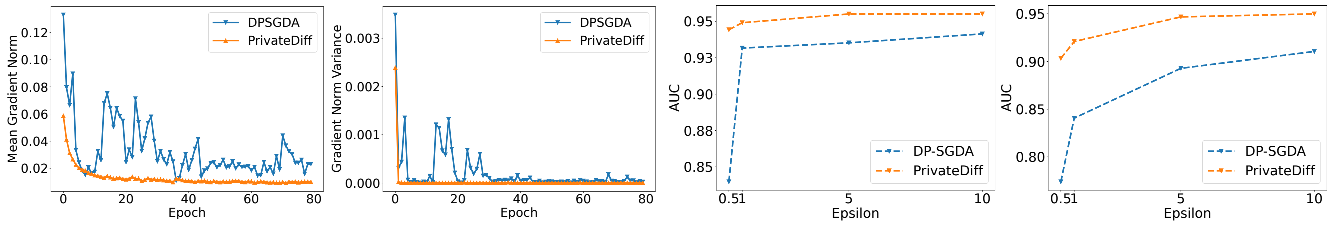
Experiments

In this section, we evaluate the effectiveness of our proposed PrivateDiff Minimax method. Due to space constraints, we focus on the AUC maximization experiment here. Additional experiments, including reinforcement learning and generative adversarial networks, are provided in the appendix.

Experimental Setup We first conduct experiments on the problem of the Area under the curve (AUC) maximization with the least squares loss (Yuan et al. 2021) to evaluate the DP-SGDA and PrivateDiff (Minimax) algorithms. AUC, ranging from 0 to 1, is a widely used metric to evaluate the performance of binary classification models. It is particularly valuable in situations where the class distribution is imbalanced because it captures the trade-offs between true positive and false positive rates. A good classifier should achieve AUC scores close to one. Maximizing AUC was demonstrated to be equivalent to a minimax problem. More detailed introductions to AUC are included in the appendix.

Our experiments are based on two common datasets, MNIST and FashionMNIST, which are transformed into binary classes by randomly partitioning the data into two groups. Following this, we create imbalanced conditions, setting an imbalance ratio of 0.1 for training, where minority classes are underrepresented, and 0.5 for testing. We chose to evaluate an imbalanced dataset because the evaluation metric, AUC scores, is particularly well-suited for assessing small or imbalanced datasets, providing a clearer indication of the algorithm's performance.

We set privacy budget $\epsilon = \{0.5, 1, 5, 10\}$ and $\delta = \frac{1}{n^{1.1}}$. A two-layer multilayer perceptron is used, consisting of 256 and 128 neurons, respectively. For other hyperparameters, we either used a grid search to select the best one or followed



(a) Gradient Norm (b) Gradient Variance (c) Fashion-MNIST (d) MNIST

Figure 1: Comparison of Gradient Norm, Gradient Variance, and AUC Performance between DP-SGDA and PrivateDiff.

Dataset	Fashion-MNIST		MNIST		Imbalanced Fashion-MNIST		Imbalanced MNIST	
	DP-SGDA ↑	PrivateDiff ↑	DP-SGDA ↑	PrivateDiff ↑	DP-SGDA ↑	PrivateDiff ↑	DP-SGDA ↑	PrivateDiff ↑
Non-private	0.9661	0.9659	0.9901	0.9901	0.9567	0.9569	0.9588	0.9593
$\epsilon = 0.5$	0.9203	0.9569	0.8837	0.9608	0.8398	0.9442	0.7739	0.9033
$\epsilon = 1$	0.9403	0.9609	0.9022	0.9729	0.9317	0.9491	0.8406	0.9209
$\epsilon = 5$	0.9412	0.9657	0.9544	0.9860	0.9352	0.9551	0.8928	0.9467
$\epsilon = 10$	0.9426	0.9660	0.9532	0.9878	0.9414	0.9551	0.9105	0.9499

Table 1: Comparison of AUC performance in DP-SGDA and PrivateDiff Minimax on various datasets.

our previous theorems.

General AUC Performance vs Privacy Table 1 demonstrates that PrivateDiff Minimax consistently achieves higher AUC scores than DP-SGDA across all dataset and privacy budget combinations. It shows that PrivateDiff consistently outperforms DP-SGDA across various datasets (Fashion-MNIST, MNIST, Imbalanced Fashion-MNIST, and Imbalanced MNIST). The performance gap is most significant at lower privacy budgets ($\epsilon = 0.5$ and 1), particularly in the MNIST and Imbalanced MNIST datasets. As the privacy budget increases, the gap narrows, but PrivateDiff still maintains a higher AUC across all scenarios, demonstrating its robustness and effectiveness in preserving utility under strong privacy constraints.

We also compare the performance of DP-SGDA and PrivateDiff across various privacy budgets (ϵ) on the Fashion-MNIST and MNIST datasets. The results in Figures 1c and 1d highlight the following observations:**1**) Performance Across Datasets: On both the Fashion-MNIST and MNIST datasets, PrivateDiff consistently outperforms DP-SGDA across all values of ϵ. This suggests that PrivateDiff is more robust in maintaining a higher AUROC score, indicating better classification performance even under stronger privacy constraints. **2**) Impact of Epsilon on AUROC: As ϵ increases, the AUROC for both DP-SGDA and PrivateDiff improves, reflecting the typical trade-off between privacy and utility in differential privacy frameworks. With higher ϵ, the privacy guarantee becomes weaker, allowing the models to achieve higher AUROC values. **3**) Comparison of Improvements: The relative improvement in AUROC with increasing ϵ is more pronounced for DP-SGDA, particularly in the MNIST dataset (Figure 1d). This might suggest that DP-SGDA's performance is more sensitive to changes in the privacy budget than that of PrivateDiff.

Robustness of PrivateDiff PrivateDiff consistently maintains lower gradient norm variance throughout the training process, as seen in Figure (1b). This reduced variance indicates a more consistent optimization trajectory, minimizing the stochastic fluctuations and contributing to a more robust training process. In contrast, DP-SGDA shows higher variance early in the training process, which indicates initial instability. An increase in variance leads to more unstable updates, which may result in overshooting or oscillating around the optimal solution. Note that a similar phenomenon has also appeared at DP Empirical Risk Minimization with nonconvex loss (Wang, Chen, and Xu 2019).

Moreover, Figure 1a illustrates that PrivateDiff achieves a stable decrease in the mean gradient norm over epochs, exhibiting fewer fluctuations compared to DP-SGDA. The steady reduction in mean gradient norm and low variance associated with PrivateDiff suggest a more reliable convergence behavior, crucial for steadily approaching the optimal solution without divergence or instability. Conversely, DP-SGDA's convergence is less reliable due to its higher variance and instability, which can lead to convergence to suboptimal solutions. These observations align with our theoretical conclusions that PrivateDiff can effectively reduce variance and offer a more stable and consistent optimization process.

Conclusions

We studied the finite sum minimax optimization problem in the Differential Privacy (DP) model where the loss function is nonconvex-(strongly)-concave. Specifically, we first analyzed DP-SGDA, which was studied previously only for convex-concave or the loss satisfying the PL condition. We then discussed several lower bounds. To further fill in the gap between lower and upper bounds, we then proposed a novel variance reduction-based algorithm. Experiments on AUC maximization, generative adversarial networks and temporal difference learning supported our theoretical analysis.

Acknowledgments

Di Wang and Zihang Xiang are supported in part by the funding BAS/1/1689-01-01, URF/1/4663-01-01, REI/1/5232-01-01, REI/1/5332-01-01, and URF/1/5508-01-01 from KAUST, and funding from KAUST - Center of Excellence for Generative AI, under award number 5940.

References

Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H. B.; Mironov, I.; Talwar, K.; and Zhang, L. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 308–318.

Arora, R.; Bassily, R.; González, T.; Guzmán, C. A.; Menart, M.; and Ullah, E. 2023. Faster rates of convergence to stationary points in differentially private optimization. In *International Conference on Machine Learning*, 1060–1092. PMLR.

Balle, B.; Barthe, G.; and Gaboardi, M. 2018. Privacy Amplification by Subsampling: Tight Analyses via Couplings and Divergences. arXiv:1807.01647.

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

Bassily, R.; Guzmán, C.; and Menart, M. 2023. Differentially private algorithms for the stochastic saddle point problem with optimal rates for the strong gap. In *The Thirty Sixth Annual Conference on Learning Theory*, 2482–2508. PMLR.

Boob, D.; and Guzmán, C. 2024. Optimal algorithms for differentially private stochastic monotone variational inequalities and saddle-point problems. *Mathematical Programming*, 204(1): 255–297.

Boţ, R. I.; and Böhm, A. 2023. Alternating proximal-gradient steps for (stochastic) nonconvex-concave minimax problems. *SIAM Journal on Optimization*, 33(3): 1884–1913.

Chen, Y.; Yang, X.-H.; Wei, Z.; Heidari, A. A.; Zheng, N.; Li, Z.; Chen, H.; Hu, H.; Zhou, Q.; and Guan, Q. 2022. Generative adversarial networks in medical image augmentation: a review. *Computers in Biology and Medicine*, 144: 105382.

Creswell, A.; White, T.; Dumoulin, V.; Arulkumaran, K.; Sengupta, B.; and Bharath, A. A. 2018. Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1): 53–65.

Dwork, C.; McSherry, F.; Nissim, K.; and Smith, A. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, 265–284. Springer.

González, T.; Guzmán, C.; and Paquette, C. 2024. Mirror Descent Algorithms with Nearly Dimension-Independent Rates for Differentially-Private Stochastic Saddle-Point Problems. *arXiv preprint arXiv:2403.02912*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Guo, Z.; Xu, Y.; Yin, W.; Jin, R.; and Yang, T. 2021. A novel convergence analysis for algorithms of the adam family and beyond. *arXiv preprint arXiv:2104.14840*.

Hu, L.; Ni, S.; Xiao, H.; and Wang, D. 2022. High dimensional differentially private stochastic optimization with heavy-tailed data. In *Proceedings of the 41st ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, 227–236.

Huai, M.; Wang, D.; Miao, C.; Xu, J.; and Zhang, A. 2020. Pairwise learning with differential privacy guarantees. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 694–701.

Levy, D.; Carmon, Y.; Duchi, J. C.; and Sidford, A. 2020. Large-scale methods for distributionally robust optimization. *Advances in Neural Information Processing Systems*, 33: 8847–8860.

Lin, T.; Jin, C.; and Jordan, M. 2020a. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, 6083–6093. PMLR.

Lin, T.; Jin, C.; and Jordan, M. I. 2020b. Near-optimal algorithms for minimax optimization. In *Conference on Learning Theory*, 2738–2779. PMLR.

Ling, Z.; Hu, F.; Zhang, H.; and Han, Z. 2022. Age-of-information minimization in healthcare IoT using distributionally robust optimization. *IEEE Internet of Things Journal*, 9(17): 16154–16167.

Lu, S.; Tsaknakis, I.; Hong, M.; and Chen, Y. 2020. Hybrid block successive approximation for one-sided non-convex min-max problems: algorithms and applications. *IEEE Transactions on Signal Processing*, 68: 3676–3691.

Luo, L.; Ye, H.; Huang, Z.; and Zhang, T. 2020. Stochastic recursive gradient descent ascent for stochastic nonconvex-strongly-concave minimax problems. *Advances in Neural Information Processing Systems*, 33: 20566–20577.

Murata, T.; and Suzuki, T. 2023. DIFF2: Differential private optimization via gradient differences for nonconvex distributed learning. *arXiv preprint arXiv:2302.03884*.

Nouiehed, M.; Sanjabi, M.; Huang, T.; Lee, J. D.; and Razaviyayn, M. 2019. Solving a class of non-convex min-max games using iterative first order methods. *Advances in Neural Information Processing Systems*, 32.

Sharma, P.; Panda, R.; Joshi, G.; and Varshney, P. 2022. Federated minimax optimization: Improved convergence analyses and algorithms. In *International Conference on Machine Learning*, 19683–19730. PMLR.

Su, J.; Hu, L.; and Wang, D. 2024. Faster Rates of Differentially Private Stochastic Convex Optimization. *Journal of Machine Learning Research*, 25(114): 1–41.

Su, J.; Zhao, C.; and Wang, D. 2023. Differentially private stochastic convex optimization in (non)-Euclidean space revisited. In *Uncertainty in Artificial Intelligence*, 2026–2035. PMLR.

Sutton, R. S. 1988. Learning to predict by the methods of temporal differences. *Machine learning*, 3: 9–44.

Tao, Y.; Wu, Y.; Cheng, X.; and Wang, D. 2022. Private Stochastic Convex Optimization and Sparse Learning with Heavy-tailed Data Revisited. In *IJCAI*, 3947–3953. ijcai.org.

Tran, H.; and Cutkosky, A. 2022. Momentum aggregation for private non-convex erm. *Advances in Neural Information Processing Systems*, 35: 10996–11008.

Wang, D.; Chen, C.; and Xu, J. 2019. Differentially private empirical risk minimization with non-convex loss functions. In *International Conference on Machine Learning*, 6526–6535. PMLR.

Wang, D.; Xiao, H.; Devadas, S.; and Xu, J. 2020. On differentially private stochastic convex optimization with heavy-tailed data. In *International Conference on Machine Learning*, 10081–10091. PMLR.

Wang, D.; and Xu, J. 2019a. Differentially private empirical risk minimization with smooth non-convex loss functions: A non-stationary view. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 1182–1189.

Wang, D.; and Xu, J. 2019b. On sparse linear regression in the local differential privacy model. In *International Conference on Machine Learning*, 6628–6637. PMLR.

Wang, D.; and Xu, J. 2021. Escaping saddle points of empirical risk privately and scalably via dp-trust region method. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part III*, 90–106. Springer.

Wang, D.; Ye, M.; and Xu, J. 2017. Differentially private empirical risk minimization revisited: Faster and more general. *Advances in Neural Information Processing Systems*, 30.

Wang, L.; Jayaraman, B.; Evans, D.; and Gu, Q. 2023. Efficient privacy-preserving stochastic nonconvex optimization. In *Uncertainty in Artificial Intelligence*, 2203–2213. PMLR.

Wang, Y.; Ma, X.; Bailey, J.; Yi, J.; Zhou, B.; and Gu, Q. 2021. On the convergence and robustness of adversarial training. *arXiv preprint arXiv:2112.08304*.

Xiao, H.; Xiang, Z.; Wang, D.; and Devadas, S. 2023. A theory to instruct differentially-private learning via clipping bias reduction. In *2023 IEEE Symposium on Security and Privacy (SP)*, 2170–2189. IEEE.

Xu, Z.; Zhang, H.; Xu, Y.; and Lan, G. 2023. A unified single-loop alternating gradient projection algorithm for nonconvex–concave and convex–nonconcave minimax problems. *Mathematical Programming*, 1–72.

Xue, Z.; Yang, S.; Huai, M.; and Wang, D. 2021. Differentially Private Pairwise Learning Revisited. In *IJCAI*, 3242–3248. ijcai.org.

Yan, Y.; Xu, Y.; Lin, Q.; Liu, W.; and Yang, T. 2020. Optimal epoch stochastic gradient descent ascent methods for min-max optimization. *Advances in Neural Information Processing Systems*, 33: 5789–5800.

Yang, T.; and Ying, Y. 2022. AUC maximization in the era of big data and AI: A survey. *ACM computing surveys*, 55(8): 1–37.

Yang, Z.; Hu, S.; Lei, Y.; Vashney, K. R.; Lyu, S.; and Ying, Y. 2022. Differentially private sgda for minimax problems. In *Uncertainty in Artificial Intelligence*, 2192–2202. PMLR.

Yuan, Z.; Yan, Y.; Sonka, M.; and Yang, T. 2021. Large-Scale Robust Deep AUC Maximization: A New Surrogate Loss and Empirical Studies on Medical Image Classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 3040–3049.

Zhang, L.; Thekumparampil, K. K.; Oh, S.; and He, N. 2022. Bring your own algorithm for optimal differentially private stochastic minimax optimization. *Advances in Neural Information Processing Systems*, 35: 35174–35187.

Zhang, X.; Aybat, N. S.; and Gurbuzbalaban, M. 2022. Sapd+: An accelerated stochastic method for nonconvex-concave minimax problems. *Advances in Neural Information Processing Systems*, 35: 21668–21681.

Zhao, C.; Ze, Y.; Dong, J.; Wang, B.; and Li, S. 2023. Differentially Private Temporal Difference Learning with Stochastic Nonconvex-Strongly-Concave Optimization. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, 985–993.

Zhou, X.; and Bassily, R. 2024. Differentially Private Worst-group Risk Minimization. *arXiv preprint arXiv:2402.19437*.