Private Stochastic Non-convex Optimization with Improved Utility Rates

Qiuchen Zhang[1] , **Jing Ma**[1] , **Jian Lou**[1,2*] and **Li Xiong**[1]
[1]Emory University
[2]Xidian University
{qiuchen.zhang, jing.ma, jian.lou, lxiong}@emory.edu, jlou@xidian.edu.cn

Abstract

We study the differentially private (DP) stochastic nonconvex optimization with a focus on its understudied utility measures in terms of the expected excess empirical and population risks. While the excess risks are extensively studied for convex optimization, they are rarely studied for nonconvex optimization, especially the expected population risk. For the convex case, recent studies show that it is possible for private optimization to achieve the same order of excess population risk as to the nonprivate optimization under certain conditions. It still remains an open question for the nonconvex case whether such ideal excess population risk is achievable.

In this paper, we progress towards an affirmative answer to this open problem: DP nonconvex optimization is indeed capable of achieving the same excess population risk as to the nonprivate algorithm in most common parameter regimes, under certain conditions (i.e., well-conditioned nonconvexity). We achieve such improved utility rates compared to existing results by designing and analyzing the stagewise DP-SGD with early momentum algorithm. We obtain both excess empirical risk and excess population risk to achieve differential privacy. Our algorithm also features the first known results of excess and population risks for DP-SGD with momentum. Experiment results on both shallow and deep neural networks when respectively applied to simple and complex real datasets corroborate the theoretical results.

1 Introduction

Many machine learning models have the underlying goal of minimizing the population loss of the form $\min_{\boldsymbol{\omega}\in\mathbb{R}^d} F_{\mathbb{E}}(\boldsymbol{\omega}) := \mathbb{E}_{\mathbf{z}\sim\mathcal{Z}}[f(\boldsymbol{\omega},\mathbf{z})]$, where f is a nonconvex loss function of the d-dimension variable $\boldsymbol{\omega}$, and $\mathbf{z}$ is data sample from the distribution $\mathcal{Z}$. It is often known as the stochastic nonconvex optimization (SNCO) problem. In practice, SNCO is approached by solving the nonconvex empirical risk minimization (ERM) problem $\min_{\boldsymbol{\omega}\in\mathbb{R}^d} F_{\mathbb{S}}(\boldsymbol{\omega}) := \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{\omega},\mathbf{z}_i)$, where $\mathbf{z}_i \in \mathbb{S}$ for $i = 1, ..., n$ are drawn from $\mathcal{Z}$ and $\mathbb{S}$ is the n-size training dataset.

The privacy breach of sensitive information contained by the data samples in the training dataset has become a growing concern. To provide rigorous privacy protection, differential privacy (DP) [Dwork *et al.*, 2006] has become a standard technique in privacy-preserving SNCO and ERM training [Abadi *et al.*, 2016], abbreviated as DP-SNCO and DP-ERM hereafter. DP works by injecting additional perturbation to the training process to hide in probability the presence or absence of any single data sample. For example, DP-SGD [Abadi *et al.*, 2016], one of the most popular DP algorithms for deep learning, injects calibrated Gaussian noise to the stochastic gradient in each iteration. The DP perturbation inevitably degrades the utility of the private model compared to the nonprivate counterpart. Therefore, quantifying the utility of the DP-SNCO and DP-ERM becomes an important problem for understanding their capability and limitation, which will be the focus of this paper. The expected excess empirical risk (empirical risk for short) and expected excess population risk (population risk) are common utility measures for DP-ERM and DP-SNCO, respectively, which are summarized in the following definition.

Definition 1.1. (Utility measures and their relationship [Hardt *et al.*, 2016]) The definitions of the expected excess population risk, expected excess empirical risk, along with testing error, and generalization error, as well as their relationship are presented as follows,

$$\begin{aligned}&\overbrace{\mathbb{E}_{\mathcal{A},\mathbb{S}}[F_{\mathbb{E}}(\boldsymbol{\omega}_{\mathbb{S}})] - \min_{\boldsymbol{\omega}} F_{\mathbb{E}}(\boldsymbol{\omega})}^{\text{expected excess population risk}}\\ &\leq \overbrace{\mathbb{E}_{\mathcal{A},\mathbb{S}}[F_{\mathbb{E}}(\boldsymbol{\omega}_{\mathbb{S}})] - \mathbb{E}_{\mathbb{S}}[F_{\mathbb{S}}(\boldsymbol{\omega}^*_{\mathbb{S}})]}^{\text{testing error}}\\ &\leq \mathbb{E}_{\mathbb{S}}\underbrace{\mathbb{E}_{\mathcal{A}}[F_{\mathbb{S}}(\boldsymbol{\omega}_{\mathbb{S}}) - F_{\mathbb{S}}(\boldsymbol{\omega}^*_{\mathbb{S}})]}_{\text{expected excess empirical risk}} + \underbrace{\mathbb{E}_{\mathcal{A},\mathbb{S}}[F_{\mathbb{E}}(\boldsymbol{\omega}_{\mathbb{S}}) - F_{\mathbb{S}}(\boldsymbol{\omega}_{\mathbb{S}})]}_{\text{generalization error}},\end{aligned} \tag{1}$$

where $\boldsymbol{\omega}^*_{\mathbb{S}} \in \arg\min_{\boldsymbol{\omega}} F_{\mathbb{S}}(\boldsymbol{\omega}_{\mathbb{S}})$ denotes a minimizer of the empirical risk $F_{\mathbb{S}}$, $\mathcal{A}$ denotes the randomized optimization algorithm, $\mathbb{E}_{\mathcal{A}}$ denotes the expectation with respect to $\mathcal{A}$; and $\mathbb{S}$ denotes the training set randomly drawn from the population distribution $\mathcal{Z}$, $\mathbb{E}_{\mathbb{S}}$ denotes the expectation with respect to $\mathbb{S}$; $\mathbb{E}_{\mathcal{A},\mathbb{S}}$ denotes expectation with both randomness of $\mathcal{A}$ and $\mathbb{S}$.

*Corresponding Author

For the convex loss function, the empirical risk has been extensively studied [Bassily *et al.*, 2014; Wang *et al.*, 2017], while the population risk is much less. Very recently, [Bassily *et al.*, 2019] improves the result in [Bassily *et al.*, 2014] and shows that the population risk of DP stochastic convex optimization (DP-SCO) is the larger of the nonprivate population loss of order $O(\frac{1}{\sqrt{n}})$ and the optimal excess empirical loss of order $O(\frac{\sqrt{d\log(1/\delta)}}{n\epsilon})$ under common assumptions. It indicates that it is possible for DP-SCO to obtain the same excess population risk as its nonprivate counterpart in most parameter regime. [Feldman *et al.*, 2020] shows the same population loss performance with lower computational complexity. For the nonconvex loss function, despite their practical popularity and success in various applications, the theoretical understanding of the utility of the DP-ERM and DP-SNCO is much less, among which the results of the excess risks are even rare, especially for the population risk. There are some work that use different utility measures [Wang *et al.*, 2017; Wang and Xu, 2019; Wang *et al.*, 2019a; Wang *et al.*, 2019b] (e.g., the gradient norm), which are not as meaningful as the excess risks, as pointed out by [Wang *et al.*, 2019a]. To the best of our knowledge, [Wang *et al.*, 2019a] is the only existing work that studies the population risk and [Wang *et al.*, 2017; Wang *et al.*, 2019a] are the only known results for the empirical risk, which are detailed as follows.

1.1 Related Works and Gaps

Existing Excess Empirical Risk for Nonconvex Loss

For the **E**mpirical risk, E-1) [Wang *et al.*, 2019a] analyzes the DP gradient langevin dynamics (GLD) and obtains $O(\frac{1}{n\epsilon^2} + \frac{d}{\log(n)})$ for general nonconvex and smooth loss, under $T \to +\infty$ and $\log(n) \geq d$ to be meaningful. In addition, they improve the excess empirical risk to $\frac{C_0(d)}{n^\tau\epsilon^\tau}$ under the number of iterations $T \to +\infty$ for $\tau \in (0, 1)$ by a finer analysis of the same DP-GLD algorithm. E-2) With the additional assumption of the loss satisfying the Polyak-Łojasiewicz (PL) condition, [Wang *et al.*, 2017] obtains $O(\frac{d\log^2(n)\log(1/\delta)}{n^2\epsilon^2})$ by analyzing the DP full gradient descent. E-3) Although [Wang *et al.*, 2019a] shows that there exists algorithm that provides the $\tilde{O}(\frac{d}{n\epsilon})$ excess empirical risk for general non-convex and smooth loss, which matches the rate under general convex setting, it is inapplicable in practice since it has an exponential computational complexity of $O((1 + \frac{n}{d})^d n)$.

Existing Excess Population Risk for Nonconvex Loss

For the **P**opulation risk, P-1) [Wang *et al.*, 2019a] obtains $O(\frac{1}{n\epsilon^2} + \frac{d}{\log(n)})$ for general non-convex and smooth loss by analyzing DP-GLD. P-2) For two special non-convex losses of the generalized linear model [Foster *et al.*, 2018] and the robust regression [Loh and Wainwright, 2015], [Wang *et al.*, 2019a] obtains $O(\frac{d^{\frac{1}{4}}}{\sqrt{n\epsilon}})$ population loss by analyzing the DP Frank-Wolfe algorithm, which can be improved to $O(\frac{\sqrt{\log(nd)}}{\sqrt{n\epsilon}})$ if the constraint is the ℓ_1 norm ball.

Gaps

In addition to the under-studied situation, there are some key limitations of the above existing work, which raise the following three gaps.

The first gap lies between the private algorithms analyzed in existing theory and the ones applied in practice. The algorithms analyzed above may not best suit large-scale machine learning problems nowadays, e.g., deep neural networks (DNN). In fact, these algorithms all require full gradient computation, which results in high per-iteration computation and poor scalability. In practice, SGD-based optimization algorithms are more popular in such large-scale problems. For example, DP-SGD has been offered by Tensorflow for private deep neural network training. For the convex case, it is possible to obtain the optimal excess risks by analyzing DP-SGD [Bassily *et al.*, 2019]. *It is unknown in theory how the DP-SGD-based algorithms perform in terms of the excess risks under the nonconvex loss setting, despite its scalability and practical popularity.*

The second gap lies in the SGD algorithm design differences between the private and nonprivate settings. For the nonprivate SGD, there are many popular algorithm designs that further improve its performance. For one example, it is popular for the nonprivate SGD to utilize the stagewise learning rate scheduling (also known as the geometric step decay learning rate) [Yuan *et al.*, 2019; Ge *et al.*, 2019; Davis *et al.*, 2019; Zhao *et al.*, 2020], which decreases the learning rate by a certain factor after certain iterations. Such stagewise step-size scheduling has been popularly adopted in practice (e.g., offered by Tensorflow, PyTorch) and has recently been analyzed in theory [Yuan *et al.*, 2019; Zhao *et al.*, 2020; Davis *et al.*, 2019; Chen *et al.*, 2019], where $O(\frac{1}{\sqrt{n}})$ population risk [Yuan *et al.*, 2019; Zhao *et al.*, 2020] is obtained under well-conditioned nonconvexity assumptions known to be held by some of deep neural networks. However, all existing excess risks analysis of SGD/GD-based algorithms in both convex and nonconvex settings only analyzes the constant step-size setting. For another example, the momentum technique [Polyak, 1964] is also popular in nonprivate SGD for DNN training, but there is only a few private counterpart [Bu *et al.*, 2019] providing merely empirical utility results under constant learning rate. *It is unknown how the designs of the stagewise step-size and momentum affect the DP-SGD and whether it is possible to obtain improved excess risks in theory given these designs.*

The third gap lies between the existing excess risks of the private algorithms and the ones for nonprivate algorithms. Under the high-dimensional setting of $n = \Theta(d)$ considered by [Bassily *et al.*, 2019; Feldman *et al.*, 2020], all existing private population risks (i.e., P-1,P-2) are worse than the nonprivate risks of $O(\frac{1}{\sqrt{n}})$ [Yuan *et al.*, 2019; Zhao *et al.*, 2020] obtained under well-conditioned nonconvexity assumptions. However, for the convex losses, the private population risk can match the nonprivate one in most parameter regimes [Bassily *et al.*, 2019; Feldman *et al.*, 2020]. For the empirical risks, although E-2 has achieved a smaller rate, it is still worse than the $O(\frac{d\log(1/\delta)}{n^2\epsilon^2})$ rate achievable for strongly convex losses [Zhang *et al.*, 2017]. *It is unknown whether it is*

possible to improve the excess risks to match the nonprivate ones at least for well-conditioned nonconvexity problems.

1.2 Our Contributions

Overview

In this paper, we progress towards answering the open problem that for nonconvex loss, whether the private optimization has a chance to achieve the same excess population risk as the nonprivate optimization. We provide much improved rates of both the expected excess empirical and population risks for the non-convex loss compared to existing works. We theoretically analyze the DP-SGD with stagewise learning rate and momentum under the same assumptions used by nonprivate optimization [Yuan *et al.*, 2019; Zhao *et al.*, 2020; Ramezani-Kebrya *et al.*, 2018]. We also conduct experiments on both shallow (2-layer convolution neural network [Lawrence *et al.*, 1997]) and deep neural networks (residual net ResNet-20 [He *et al.*, 2016]) for simple (MNIST [LeCun *et al.*, 1998]) and complex (CIFAR-10 [Krizhevsky *et al.*, 2009]) datasets, respectively.

Algorithm Design

We propose a **D**ifferentially **P**rivate st**age**wise **S**GD with **E**arly **M**omentum (**DpageEM**). To close Gap 2, we consider 1) stagewise learning rate scheduling in DP-SGD, which decreases the learning rate by a factor after certain iterations; 2) the early momentum strategy [Ramezani-Kebrya *et al.*, 2018], which switches the momentum on during the early iterations and switches it off for the later iterations in each stage. The early momentum not only facilitates the need to strike a trade-off between the training efficiency and generalization performance, but also includes the DP-SGD without momentum as a special case. Thus, our theoretical results obtained for DpageEM can be easily applied to DP-SGD.

Assumptions

We analyze its excess empirical and population risks under the same set of assumptions adopted by nonprivate optimization, which include common mild assumptions like Lipschitz smoothness and continuity in Assumption 2.1 & 2.2, together with the Polyak-Łojasiewicz (PL) condition in Assumption 2.4 and one point weakly quasi-convex Assumption 2.5. The latter two have been observed and proved for shallow and deep neural networks as detailed in Remark 1.

Excess Risks

To the best of our knowledge, our excess risks are the first known theoretical results for the momentum technique under the private setting, especially for nonconvex losses. Analyzing both excess risks is a nontrivial task. Neither the convergence nor the uniform stability is known for the stagewise early momentum SGD even without the private restriction. As a result, it requires us to develop the following two new results as the stepping stone towards the private excess risks: 1) We provide the convergence of DpageEM in Theorem 3.2, which is general enough to include the momentum always-on and always-off as special cases. 2) We provide the uniform stability result of DpageEM in Thoerem 3.3, which shows the tradeoff between the number of iterations of momentum-on and the generalization error. By neglecting the DP noise, these two results can be of independent interest for nonprivate stagewise early momentum SGD for DNN optimization.

We obtain $O(\frac{d\log(1/\delta)}{n^2\epsilon^2})$ excess empirical risk for DpageEM with (ϵ, δ)-DP, which is the same order in the dependence of the training data size n and model dimension d with the strongly convex loss setting [Zhang *et al.*, 2017]. However, our assumptions are much weaker than [Zhang *et al.*, 2017] as discussed in Remark **??**.

We obtain $O(\max\{\frac{d\log(1/\delta)}{n^2\epsilon^2}, \frac{1}{\sqrt{n}}\})$ excess population risk, where the $O(\frac{d\log(1/\delta)}{n^2\epsilon^2})$ term is the same as its private excess empirical risk and the $O(\frac{1}{\sqrt{n}})$ term matches the population risk of the nonprivate stagewise SGD [Yuan *et al.*, 2019; Zhao *et al.*, 2020]. Thus, it indicates that it is possible for the private SNCO to have the same rate of the population risk as the nonprivate SNCO in most parameter regimes in practice. In the same high-dimensional regime considered in [Bassily *et al.*, 2019; Feldman *et al.*, 2020] where $n = \Theta(d)$, our population risk is better than the $O(\frac{d^{1/4}}{\sqrt{n\epsilon}}) = O(\frac{1}{n^{\frac{1}{4}}\epsilon^{\frac{1}{2}}})$ rate (i.e., P-2) in [Wang *et al.*, 2019a], which is obtained for the specific generalized linear model [Foster *et al.*, 2018] and robust regression [Loh and Wainwright, 2015] problems by the DP Frank-Wolfe algorithm that is not popular for DNN.

2 Preliminaries

2.1 Stochastic Nonconvex Optimization

To show that private optimization has the same order of utility with nonprivate optimization, we invoke the same set of assumptions for both by following [Yuan *et al.*, 2019; Zhao *et al.*, 2020; Ramezani-Kebrya *et al.*, 2018]. For the nonconvex loss function $f(\boldsymbol{\omega}, \mathbf{z})$, we assume it is bounded and without loss of generality, $|f(\boldsymbol{\omega}, \mathbf{z})| \leq 1$. We also make the following common assumptions.

Assumption 2.1. ***(Lipschitz continuous)*** *$f(\boldsymbol{\omega}, \mathbf{z})$ is L-Lipschitz continuous in $\boldsymbol{\omega}$ for any $\mathbf{z}$, i.e., $\forall \boldsymbol{\omega}, \boldsymbol{\omega}' \in \mathbb{R}^d$, we have $|f(\boldsymbol{\omega}, \mathbf{z}) - f(\boldsymbol{\omega}', \mathbf{z})| \leq L\|\boldsymbol{\omega} - \boldsymbol{\omega}'\|_2$.*

Assumption 2.2. ***(Lipschitz smoothness)*** *$f(\boldsymbol{\omega}, \mathbf{z})$ is β-Lipschitz smooth in $\boldsymbol{\omega}$ for any $\mathbf{z}$, $\forall \boldsymbol{\omega}, \boldsymbol{\omega}' \in \mathbb{R}^d$, we have $\|\nabla f(\boldsymbol{\omega}, \mathbf{z}) - \nabla f(\boldsymbol{\omega}', \mathbf{z})\|_2 \leq \beta\|\boldsymbol{\omega} - \boldsymbol{\omega}'\|_2$.*

Assumption 2.3. ***(Bounded variance)*** *The variance of the stochastic gradient is bounded, i.e., $\mathbb{E}_i[\|\nabla f(\boldsymbol{\omega}, \mathbf{z}_i) - \nabla F_{\mathbf{S}}(\boldsymbol{\omega})\|_2^2] \leq \sigma^2$, for any $\boldsymbol{\omega}$.*

In addition, the loss function is well-conditioned nonconvex as depicted by the following two assumptions.

Assumption 2.4. ***(Polyak-Łojasiewicz (PL) condition)*** *$F_{\mathbf{S}}(\boldsymbol{\omega})$ satisfies μ-Polyak-Łojasiewicz (PL) condition, i.e., $2\mu(F_{\mathbf{S}}(\boldsymbol{\omega}) - \min_{\boldsymbol{\omega}} F_{\mathbf{S}}(\boldsymbol{\omega})) \leq \|\nabla F_{\mathbf{S}}(\boldsymbol{\omega})\|_2^2$.*

Assumption 2.5. ***(One point weakly quasi-convex)*** *$F_{\mathbf{S}}(\boldsymbol{\omega})$ is one point θ-weakly quasi convex, i.e., for any $\boldsymbol{\omega}$, $\langle \nabla F_{\mathbf{S}}(\boldsymbol{\omega}), \boldsymbol{\omega} - \boldsymbol{\omega}_{\mathbf{S}}^* \rangle \geq \theta(F_{\mathbf{S}}(\boldsymbol{\omega}) - F_{\mathbf{S}}(\boldsymbol{\omega}_{\mathbf{S}}^*))$, where $\boldsymbol{\omega}_{\mathbf{S}}^*$ is the optimal solution.*

Remark 1. (Consistency of the assumptions with some deep neural networks) Both assumptions are satisfied with some deep learning losses. For the PL condition, it has been empirically observed or theoretically proved for shallow (e.g.,

two-layer) and deep neural networks (DNN) [Hardt and Ma, 2017]. In fact, throughout the paper, the assumption only needs to hold locally instead of in the entire space. That is, [Allen-Zhu *et al.*, 2019] shows that this relaxed local PL condition holds for two-layer and deep overparameterized neural networks in a ball that contains the global optimum and centers around a random initial solution. For the one point weakly quasi-convex assumption, [Yuan *et al.*, 2019] shows that it is satisfied by some deep neural networks (e.g., ResNet [He *et al.*, 2016]) together with the PL condition. As for the range of the parameters, we consider $\mu \ll 1$ (i.e., ill-conditioned) and $\theta \approx 1$ which are consistent with the empirical observations about DNN [Yuan *et al.*, 2019].

2.2 Differential Privacy

Definition 2.1. (Neighboring datasets) Let the data universe be $\mathcal{Z}$ and $\mathcal{S}, \mathcal{S}'$ be two datasets of n observed data samples drawn from $\mathcal{Z}$ with distribution $\mathbb{P}_{\mathbf{z}}$. Then, $\mathcal{S}$ and $\mathcal{S}'$ are called neighboring datasets if they differ in exactly one data sample.

Definition 2.2. (Differential privacy) A randomized algorithm $\mathcal{A}$ is (ϵ, δ)-differentially private, if for all neighboring datasets $\mathcal{S}, \mathcal{S}'$, and for any events $\mathcal{O}$ in the output range of $\mathcal{A}$, we have $\mathbb{P}[\mathcal{A}(\mathcal{S}) \in \mathcal{O}] \leq e^{\epsilon} \cdot \mathbb{P}[\mathcal{A}(\mathcal{S}') \in \mathcal{O}] + \delta$, where the probability is over the random coins of $\mathcal{A}$.

Definition 2.3. (Gaussian mechanism) Given a query $q : \mathcal{Z}^n \to \mathcal{R}^d$, the Gaussian mechanism is defined as: $\mathcal{M}(\mathcal{S}, q) = q(\mathcal{S}) + \xi$, where ξ is drawn from Gaussian distribution $\mathcal{N}(0, \pi^2 \mathbb{I}_d)$.

Theorem 2.1. *([Abadi* et al.*, 2016]) For an L-Lipschitz continuous loss function f, there exists constants c_1 and c_2 so that given the sampling probability $q = \frac{B}{n}$ and the number of total iterations T, for any $\epsilon < c_1 q^2 T$, a stochastic gradient-based algorithm with batch size B and stochastic gradient injected with noise from Gaussian mechanism with standard deviation $L\pi$, is (ϵ, δ)-differentially private for any $\delta > 0$ if $\pi \geq c_2 \frac{q\sqrt{T \log(1/\delta)}}{\epsilon}$.*

2.3 Stability and Generalization

The uniform stability is a well-known technique exploited to study the generalization performance of the SGD algorithms [Hardt *et al.*, 2016].

Definition 2.4. (Uniform stability) A randomized algorithm $\mathcal{A}$ is called to satisfy α-uniform stability, if for any neighboring datasets $\mathcal{S}, \mathcal{S}' \in \mathcal{Z}^n$ the following holds,

$$\sup_{\mathbf{z}} \mathbb{E}_{\mathcal{A}}[f(\mathcal{A}(\mathcal{S}), \mathbf{z}) - f(\mathcal{A}(\mathcal{S}'), \mathbf{z})] \leq \alpha. \tag{2}$$

Theorem 2.2. ***(Generalization error by uniform stability, Theorem 2.2 in [Hardt* et al.*, 2016])*** *If an algorithm $\mathcal{A} : \mathcal{Z}^n \to \mathcal{W}$ satisfies α-uniform stability with respect to loss $f : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}$, its generalization error is bounded by α:*

$$\mathbb{E}_{\mathcal{A},\mathcal{S}}[F_{\mathbb{E}}(\mathcal{A}(\mathcal{S})) - F_{\mathcal{S}}(\mathcal{A}(\mathcal{S}))] \leq \alpha. \tag{3}$$

Algorithm 1 DPageEM

Input: Dataset $\mathcal{S} = \{\mathbf{z}_1, ..., \mathbf{z}_n\}$, step-size series $\{\eta^k\}$, number of inner iterations $\{T^k\}$, number of iterations with momentum on $\{t^k_{Mon}\}$, momentum parameter ϱ, DP noise parameter π, batch size B, initialization $\boldsymbol{\omega}^0$;

1: **for** $k = 1, ..., K$ **do**
2: $\quad \boldsymbol{\omega}^k = \text{DP-EMSGD}(F_{\mathcal{S}}, \boldsymbol{\omega}^{k-1}, \boldsymbol{\omega}^{k-1}, \eta^k, T^k, t^k_{Mon}, \varrho)$;
3: **end for**

Output: $\boldsymbol{\omega}^K$;

Algorithm 2 DP-EMSGD$(F_{\mathcal{S}}, \boldsymbol{\omega}_0, \boldsymbol{\omega}_{-1}, \eta, T, t_{Mon}, \varrho)$

1: **for** $t = 1, ..., T$ **do**
2: $\quad$ Draw mini-batch with index $\mathcal{I}_t \in \{n\}$ uniformly random, where $|\mathcal{I}_t| = B$;
3: $\quad$ Draw DP noise $\xi_t \sim \mathcal{N}(0, \pi^2 \mathbb{I}_d)$;
4: $\quad \boldsymbol{\omega}_{t+1} = \boldsymbol{\omega}_t - \eta(\frac{1}{B}\sum_{i \in \mathcal{I}_t} \nabla f(\boldsymbol{\omega}_t, \mathbf{z}_i) + \xi_t) + \varrho_t(\boldsymbol{\omega}_t - \boldsymbol{\omega}_{t-1})$, where $\varrho_t = \varrho \cdot \mathrm{T}rue(t \in \{1, \ldots, t_{Mon}\})$;
5: **end for**

Output: $\boldsymbol{\omega}_O$ randomly sampled from $\boldsymbol{\omega}_1, ..., \boldsymbol{\omega}_T$;

3 DP Stagewise Early Momentum SGD with Step Decay Step-size

3.1 Algorithm Description

Algorithm 1 presents the DP-SGD with stagewise learning rate and early momentum (**DpageEM**), which runs T^k inner iterations under the learning rate η^k at stage k and iterates for K outer iterations. Within each stage k, we use the DP early momentum SGD (Algorithm 2), which switches the momentum on for iterations $t = 1, ..., t^k_{Mon}$ by setting the momentum parameter $\varrho_t = \varrho$, and switches the momentum off for iterations $t = t^k_{Mon} + 1, ..., T^k$ by setting $\varrho_t = 0$ (Mon is short for Momentum on).

3.2 Algorithm Analysis

Privacy Analysis

The following theorem provides the privacy guarantee of Algorithm 1, which follows 1) the moments accountant analysis in Theorem 3.1 from [Abadi *et al.*, 2016]; and 2) the privacy amplification result in [Balle *et al.*, 2018] that the amplification ratio is the same under the uniform sampling without replacement and under the Poisson sampling when $\frac{B}{n} = q$.

Theorem 3.1. ***(Differential privacy of Algorithm 1)*** *Suppose Assumptions 2.1 and 2.2 hold. There exist some constants $c_1, c_2 > 0$ such that given the sampling ratio $\frac{B}{n}$ and the number of total iterations $\sum_{k=1}^K T^k$, for any $\epsilon < c_1(\frac{B}{n})^2 \sum_{k=1}^K T^k$, Algorithm 1 is (ϵ, δ)-differentially private for any $\delta > 0$ under the choice of*

$$\pi^2 \geq c_2^2 \frac{\sum_{k=1}^K T^k L^2 \log(1/\delta)}{n^2 \epsilon^2}. \tag{4}$$

Remark 2. We emphasize that the privacy design and DP analysis are not the main tasks of this paper, so we follow this widely accepted DP technique in deep learning. It enables a direct comparison with the existing works on excess risks [Wang

et al., 2019a] which also utilize moments accountant for DP analysis. It is possible to use more advanced DP analysis for a tighter privacy amplification result derived for the uniform sampling without replacement in [Wang *et al.*, 2019c], as well as tighter privacy composition, e.g., the Gaussian DP [Dong *et al.*, 2019] used in [Bu *et al.*, 2019].

Excess Empirical Risk

In the following, we provide the excess empirical risk and excess population risk results of DpageEM. To obtain the excess risks, we analyze the convergence and uniform stability of the SGD with early momentum. As the early momentum has not been analyzed with stagewise step-size before, both results are new and the analysis is also nontrivial even without the DP consideration. The detailed proof (and all proofs for the remaining results) can be found in Supplement[1].

We first analyze the excess empirical risk of DpageEM, which has better rate than existing ones (E-1 to E-3).

Theorem 3.2. ***(Excess empirical risk)*** *Suppose Assumptions 2.1–2.5 hold. By setting π in eq.(4), $\eta^k = (\frac{1}{2})^k\eta^0$, $T^k = (2)^kT^0$, $t^k_{Mon} = (2)^k t^0_{Mon}$, where $\eta^0 \leq \max\{\frac{\theta(1-\varrho)^2}{2\beta}, 1\}$ and $\eta^0((T^0 - t^0_{Mon}) + \frac{1}{2(1-\varrho)}t^0_{Mon}) = \frac{1}{c_4\theta\mu}$ for some $0 < c_4 < 2$ (which gives $\eta^0(T^0 - t^0_{Mon} + t^0_{Mon}\frac{1}{2(1-\varrho)}) > \frac{1}{2\theta\mu}$), in Algorithm 1, with $\Gamma = \frac{1}{4\theta\mu\eta^0}\frac{1}{(T^0-t^0_{Mon})+\frac{t^0_{Mon}}{2(1-\varrho)}} < \frac{1}{2}$,*

$$\begin{aligned}\mathbb{E}[F_{\mathcal{S}}(\boldsymbol{\omega}^K) - F_{\mathcal{S}}(\boldsymbol{\omega}^*_{\mathcal{S}})] \leq (\frac{1}{2})^K\Big(&\mathbb{E}[F_{\mathcal{S}}(\boldsymbol{\omega}^0) - F_{\mathcal{S}}(\boldsymbol{\omega}^*_{\mathcal{S}})] \\ &+ \frac{1}{1-2\Gamma}\frac{(T^0 - t^0_{Mon}) + \frac{t^0_{Mon}}{(1-\varrho)^3}}{(T^0 - t^0_{Mon}) + \frac{t^0_{Mon}}{2(1-\varrho)}} \cdot \frac{\eta^0\sigma^2}{2\theta B}\Big) \\ &+ \frac{1}{1-2\Gamma}\frac{(T^0 - t^0_{Mon}) + \frac{t^0_{Mon}}{(1-\varrho)^3}}{(T^0 - t^0_{Mon}) + \frac{t^0_{Mon}}{2(1-\varrho)}}\frac{2c_2^2dL^2\log(1/\delta)\eta^0}{c_4\theta^2\mu n^2\epsilon^2}.\end{aligned} \tag{5}$$

By setting $K = \Omega(\log(\frac{n^2\epsilon^2\theta^2\mu}{dL^2\log(1/\delta)}))$, we obtain the excess empirical risk: $\mathbb{E}[F_{\mathcal{S}}(\boldsymbol{\omega}^K) - F_{\mathcal{S}}(\boldsymbol{\omega}^*_{\mathcal{S}})] \leq O(\frac{dL^2\log(1/\delta)}{\theta^2\mu n^2\epsilon^2})$.

Remark 3. The above theorem shows that the excess empirical risk of DpageEM is $O(\frac{d\log(1/\delta)}{n^2\epsilon^2})$, which matches the result in [Zhang *et al.*, 2017]. However, we obtain it under assumptions of weak quasi-convexity and PL-conditions, which are much weaker than their strongly convex assumption. Compared with nonconvex settings under different utility measures and different assumptions, our result is also much better. It indicates that it is possible to achieve good excess empirical risk by the practical step-size scheduling.

Remark 4. By the generality of the analysis, the above result also implies that the two special cases, i.e., stagewise DP-SGD with momentum always on and always off, also have the same order of excess empirical risk. Thus, our result can also be flexibly applied to DP-SGD with stagewise learning rate.

Excess Population Risk

In this part, we provide the excess population risk of DpageEM. We utilize the uniform stability technique to obtain the generalization error (i.e., **Thm.2.2**) [Hardt *et al.*, 2016]. The following **Theorem 3.3** presents the α-uniform stability of DpageEM, which is nontrivial because it not only deals with the additional DP noise, but also extends the stability result from the constant step-size SGD with early momentum [Ramezani-Kebrya *et al.*, 2018] to the stagewise setting with step decay step-size scheduling. Although the momentum technique has been considered in the stagewise setting [Chen *et al.*, 2019; Zhao *et al.*, 2020], they only provide the convergence result, while we also study the stability, generalization and population risk. To the best of our knowledge, this is the first uniform stability and generalization result for the momentum SGD with stagewise step-size.

Theorem 3.3. ***(α-Uniform stability)*** *With the same assumptions and parameter settings as in Theorem 3.2, Algorithm 1 is α-uniformly stable with*

$$\begin{aligned}\alpha = &\frac{2L^2}{c_4\theta\mu n}(T^K)^q e^{2\varrho(t^K_{Mon}-1)}(\log(1+\frac{1}{2\varrho}) - \frac{1}{2}\log(1+\frac{1}{\varrho t^K_{Mon}})) \\ &+ \frac{2L^2}{(n-1)\beta}(T^K)^q + \frac{(T^K+1)B}{n},\ \text{where } q = (1-\frac{1}{n})\frac{\beta}{c_4\theta\mu}.\end{aligned}$$

Remark 5. Theorem 3.3 reflects the tradeoff between the stability, generalization performance, excess population risk against the training efficiency. According the first term in Theorem 3.3, α increases when t^K_{Mon} becomes larger, i.e., the stability of the algorithm gets worse when the momentum switches on for more iterations. By Theorem 2.2, the generalization performance is negatively affected by the momentum, which further implies that the population risk will be larger.

Theorem 3.4. ***(Excess population risk)*** *Under the same assumptions and parameter setting as in **Theorem 3.2**, we have the excess population risk of Algorithm 1,*

$$\mathbb{E}_{\mathcal{A},\mathcal{S}}[F_{\mathbb{E}}(\boldsymbol{\omega}^K) - \min_{\boldsymbol{\omega}} F_{\mathbb{E}}(\boldsymbol{\omega})] \leq O(\max\{\frac{dL^2\log(1/\delta)}{\theta^2\mu n^2\epsilon^2}, \sqrt{\frac{1}{\theta\mu n}}\}).$$

Remark 6. The above theorem shows that the excess population risk of DpageEM is the larger one between the excess empirical risk of DpageEM and the excess population risk of the nonprivate stagewise SGD. Thus, we have shown that DP-SGD-based method is able to match the nonprivate population risk in most parameter regimes for the nonconvex losses with the same assumptions with the nonprivate algorithms. This result greatly improves the $O(\frac{\log(1/\delta)}{n\epsilon^2} + \frac{d}{\log(n)})$ in P-1 which is for general nonconvex loss and obtained by DP-GLD with constant step-size and full gradient descent. It also improves over the $O(\frac{(d\log(1/\delta))^{\frac{1}{4}}}{\sqrt{n\epsilon}})$ and $O(\frac{\sqrt{\log(nd)}}{\sqrt{n\epsilon}})$ in P-2, which are obtained for the special nonconvex generalized linear model [Foster *et al.*, 2018] and nonconvex robust regression [Loh and Wainwright, 2015] by the DP Frank-Wolfe algorithm.

4 Numerical Experiments

4.1 Experiments Setup

Datasets and Models We conduct experiments on two real datasets: MNIST and CIFAR-10. For the simpler MNIST dataset, we consider a shallow convolution neural network (2-layer CNN); For the more complex CIFAR-10 dataset, we use a deeper neural network of ResNet-20. ELU activation

[1]https://www.dropbox.com/sh/ldtrfa3ihx51dkz/AADWQD1qvNEeiLFYgRn5I30ba?dl=0

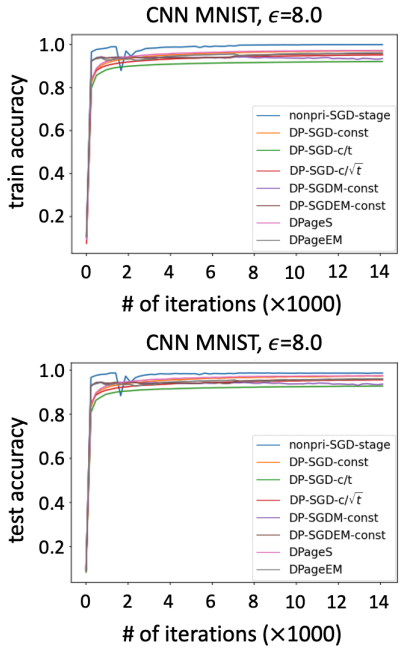

Figure 1: Training/Testing accuracy of CNN on MNIST. $\epsilon = 8$.

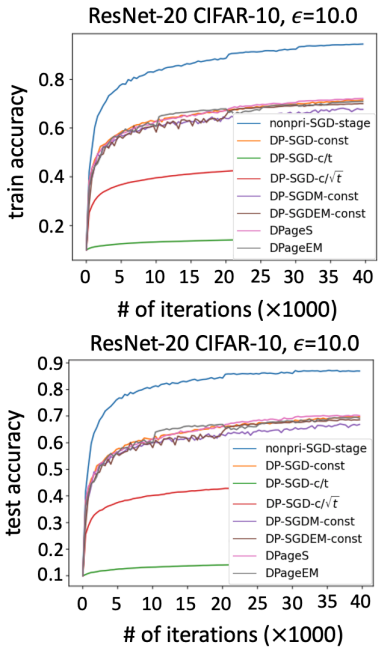

Figure 2: Training/Testing accuracy of ResNet-20 on CIFAR-10. $\epsilon = 10$.

function is used. The assumptions made for theoretical analysis are (approximately) satisfied by the empirical settings [Yuan *et al.*, 2019]. The experiment details and additional experiment results can be found in Supplement.

Compared Methods We compare the proposed methods: DPageEM with momentum always off (DPageS) and DP early momentum SGD with stagewise learning rate (DpageEM) with the following methods: 1) nonprivate SGD with stagewise learning rate (nonpri-SGD-stage); 2) DP-SGD with constant learning rate (DP-SGD-const); 3) DP-SGD with linear decayed learning rate with respect to the training iteration t (DP-SGD-c/t); 4) DP-SGD with polynomial decayed learning rate with respect to the square root of the training iteration (DP-SGD-$c/\sqrt{t}$); 5) DP momentum SGD with constant learning rate (DP-SGDM-const); 6) DP early momentum SGD with constant learning rate (DP-SGDEM-const).

4.2 Experiments Results and Discussions

We plot the training and testing accuracy versus the number of iterations for all algorithms under different privacy parameters ϵ in Figure 1-2. The key observations and discussions are summarized as follows:
1) DPageS achieves the best training and testing accuracy among the private algorithms in all experiment settings, which indicates that stagewise learning rate indeed improves the training performance. Thus, this empirical observation corroborates the improved excess empirical and population risks.
2) DpageEM has better accuracy than DP-SGDM-const and DP-SGDEM-const in all experiment settings. Also, DP-SGDM-const performs the worst compared to the early momentum methods. The early momentum methods perform worse than DP-SGD without any momentum. Thus, it indicates that momentum trades the utility for efficiency and early momentum can help make an explicit trade-off by adjusting the ratio of the momentum-on iterations.
3) The performance gap between private and nonprivate models corresponding to the training and testing accuracy is smaller for shallow neural networks (2-layer CNN) applied to simpler dataset (MNIST), while becoming larger for deeper neural networks (ResNet-20) applied to more complex dataset (CIFAR-10). It indicates opportunities of future improvements for DP optimization targeted at deeper neural networks.

5 Conclusion

In this paper, we studied the expected empirical and population risks of nonconvex DP-ERM and DP-SNCO by designing and analyzing DP-SGD-based algorithms. In order to reduce the gap between the nonprivate algorithms with designs popular in practice and the private algorithms analyzed in theory, we introduced and analyzed the DP-SGD with the stagewise step-size and momentum designs. Under the same assumptions that are observed and proved for nonprivate nonconvex learning, the proposed algorithm is able to reach improved excess risks over existing results and the excess population risk can match the nonprivate setting. Experiments on both shallow and deep neural networks when respectively applied to simple and complex datasets corroborate the theoretical results. In the future, we will study the excess risks for other types of nonconvex optimization algorithms and under more general assumptions, e.g., nonsmooth loss function.

Acknowledgements

This work is partially supported by the NSF under CNS-1952192, IIS-1838200, and NIH under R01GM118609 and CTSA Award UL1TR002378.

References

[Abadi *et al.*, 2016] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *CCS*, 2016.

[Allen-Zhu *et al.*, 2019] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *ICML*, 2019.

[Balle *et al.*, 2018] Borja Balle, Gilles Barthe, and Marco Gaboardi. Privacy amplification by subsampling: tight analyses via couplings and divergences. In *NeurIPS*, 2018.

[Bassily *et al.*, 2014] Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *FOCS*. IEEE, 2014.

[Bassily *et al.*, 2019] Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Guha Thakurta. Private stochastic convex optimization with optimal rates. In *NeurIPS*, 2019.

[Bu *et al.*, 2019] Zhiqi Bu, Jinshuo Dong, Qi Long, and Weijie J Su. Deep learning with gaussian differential privacy. *arXiv preprint arXiv:1911.11607*, 2019.

[Chen *et al.*, 2019] Zaiyi Chen, Zhuoning Yuan, Jinfeng Yi, Bowen Zhou, Enhong Chen, and Tianbao Yang. Universal stagewise learning for non-convex problems with convergence on averaged solutions. In *ICLR*, 2019.

[Davis *et al.*, 2019] Damek Davis, Dmitriy Drusvyatskiy, and Vasileios Charisopoulos. Stochastic algorithms with geometric step decay converge linearly on sharp functions. *arXiv preprint arXiv:1907.09547*, 2019.

[Dong *et al.*, 2019] Jinshuo Dong, Aaron Roth, and Weijie J Su. Gaussian differential privacy. *arXiv preprint arXiv:1905.02383*, 2019.

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

[Feldman *et al.*, 2020] Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: Optimal rates in linear time. In *STOC*, 2020.

[Foster *et al.*, 2018] Dylan J Foster, Ayush Sekhari, and Karthik Sridharan. Uniform convergence of gradients for non-convex learning and optimization. In *NeurIPS*, 2018.

[Ge *et al.*, 2019] Rong Ge, Sham M Kakade, Rahul Kidambi, and Praneeth Netrapalli. The step decay schedule: A near optimal, geometrically decaying learning rate procedure for least squares. In *NeurIPS*, 2019.

[Hardt and Ma, 2017] Moritz Hardt and Tengyu Ma. Identity matters in deep learning. In *ICML*, 2017.

[Hardt *et al.*, 2016] Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *ICML*, 2016.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[Krizhevsky *et al.*, 2009] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Technical Report, University of Toronto*, 2009.

[Lawrence *et al.*, 1997] Steve Lawrence, C Lee Giles, Ah Chung Tsoi, and Andrew D Back. Face recognition: A convolutional neural-network approach. *IEEE Transactions on Neural Networks*, 8(1):98–113, 1997.

[LeCun *et al.*, 1998] 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.

[Loh and Wainwright, 2015] Po-Ling Loh and Martin J Wainwright. Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *The Journal of Machine Learning Research*, 16(1):559–616, 2015.

[Polyak, 1964] Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

[Ramezani-Kebrya *et al.*, 2018] Ali Ramezani-Kebrya, Ashish Khisti, and Ben Liang. On the stability and convergence of stochastic gradient descent with momentum. *arXiv preprint arXiv:1809.04564*, 2018.

[Wang and Xu, 2019] Di Wang and Jinhui Xu. Differentially private empirical risk minimization with smooth non-convex loss functions: A non-stationary view. In *AAAI*, 2019.

[Wang *et al.*, 2017] Di Wang, Minwei Ye, and Jinhui Xu. Differentially private empirical risk minimization revisited: Faster and more general. In *NeurIPS*, 2017.

[Wang *et al.*, 2019a] Di Wang, Changyou Chen, and Jinhui Xu. Differentially private empirical risk minimization with non-convex loss functions. In *ICML*, 2019.

[Wang *et al.*, 2019b] Lingxiao Wang, Bargav Jayaraman, David Evans, and Quanquan Gu. Efficient privacy-preserving nonconvex optimization. *arXiv preprint arXiv:1910.13659*, 2019.

[Wang *et al.*, 2019c] Yu-Xiang Wang, Borja Balle, and Shiva Prasad Kasiviswanathan. Subsampled rényi differential privacy and analytical moments accountant. In *AISTATS*. PMLR, 2019.

[Yuan *et al.*, 2019] Zhuoning Yuan, Yan Yan, Rong Jin, and Tianbao Yang. Stagewise training accelerates convergence of testing error over sgd. In *NeurIPS*, 2019.

[Zhang *et al.*, 2017] Jiaqi Zhang, Kai Zheng, Wenlong Mou, and Liwei Wang. Efficient private erm for smooth objectives. In *IJCAI*, 2017.

[Zhao *et al.*, 2020] Shen-Yi Zhao, Yin-Peng Xie, and Wu-Jun Li. Stagewise enlargement of batch size for sgd-based learning. *arXiv preprint arXiv:2002.11601*, 2020.