Differentially Private Non-convex Learning: From Generalized Linear Models to Multi-layer Neural Networks

Extended Abstract

Hanpu Shen
University of California, Irvine
Irvine, United States
hanpus1@uci.edu

Cheng-Long Wang
King Abdullah University of Science and Technology
Thuwal, Saudi Arabia
chenglong.wang@kaust.edu.sa

Zihang Xiang
University of California, Los Angeles
Los Angeles, United States
zihangxiang@ucla.edu

Yiming Ying
University of Sydney
Sydney, Australia
yiming.ying@sydney.edu.au

Di Wang
King Abdullah University of Science and Technology
Thuwal, Saudi Arabia
di.wang@kaust.edu.sa

ABSTRACT

This paper focuses on the problem of Differentially Private Stochastic Optimization for (multi-layer) fully connected neural networks with a single output node. In the first part, we examine cases with no hidden nodes, i.e., Generalized Linear Models (GLMs). We investigate the well-specific model where the random noise possesses a zero mean, and the link (activation) function is both bounded and Lipschitz. We propose several algorithms and our analysis demonstrates the feasibility of achieving an excess population risk that remains invariant to the data dimension. We then delve into the scenario involving the ReLU link function, and our findings mirror those of the bounded link function. Moreover, we extend our ideas to two-layer neural networks with sigmoid or ReLU activation functions in the well-specified model. We conclude this section by contrasting well-specified and misspecified models, using ReLU regression as a representative example. In the second part, we study the theoretical guarantees of DP-SGD in Abadi et al. (2016) for multi-layer neural networks. By utilizing recent advances in Neural Tangent Kernel theory, we provide the first excess population risk when both the sample size and the width of the network are sufficiently large. Additionally, we discuss the role of some parameters in DP-SGD regarding their utility, both theoretically and empirically. [1]

KEYWORDS

Differential Privacy; Empirical Risk Minimization

ACM Reference Format:
Hanpu Shen, Cheng-Long Wang, Zihang Xiang, Yiming Ying, and Di Wang. 2026. Differentially Private Non-convex Learning: From Generalized Linear Models to Multi-layer Neural Networks: Extended Abstract. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/JRBH5415

[1] Hanpu Shen and Cheng-Long Wang have equal contributions. Di Wang is the Corresponding Author.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/JRBH5415

1 BACKGROUND

In machine learning, extracting knowledge from data consisting of sensitive attributes is an evolving concern. Such a task mandates algorithms that can proficiently interpret the data while upholding established privacy benchmarks. Differential privacy (DP) [4] has gained traction as a seminal framework for statistical data protection. Recognized widely in contemporary research, DP ensures that individual data remains non-retrievable post-analysis, offering a robust defense mechanism against privacy infractions. This underscores a burgeoning interest in devising learning architectures where DP considerations are intrinsically woven into the analytic process.

Stochastic Optimization (SO) and its empirical form, Empirical Risk Minimization (ERM), are the most fundamental models in machine learning and statistics. They have numerous applications in fields such as medicine, finance, genomics, and social science. However, these applications often involve sensitive data, making it essential to design differentially private algorithms for SO and ERM, corresponding to the problems of DP-SO and DP-ERM, respectively. Specifically, they are fined as follows.

Definition 1.1 (DP [4]). Given a data universe $\mathcal{X}$, we say that two datasets $D, D' \subseteq \mathcal{X}$ are neighbors if they differ by only one data record, 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 S in the output space of $\mathcal{A}$, we have

$$\Pr(\mathcal{A}(D) \in S) \leq e^{\epsilon}\Pr(\mathcal{A}(D') \in S) + \delta.$$

Definition 1.2 (DP-SO [3]). Given a dataset $D = \{z_1, \cdots, z_n\}$ from a data universe $\mathcal{Z}$ where each $z_i = (x_i, y_i)$ with a feature vector x_i and a label/response y_i is i.i.d. sampled from some unknown distribution $\mathcal{P}$, a convex constraint set $\mathcal{W} \subseteq \mathbb{R}^d$, and a (non-convex) loss function $\ell : \mathcal{W} \times \mathcal{Z} \mapsto \mathbb{R}$. Differentially Private Stochastic Optimization (DP-SO) is to find a model w^{priv} to minimize the population risk, *i.e.*, $L_{\mathcal{P}}(w) = \mathbb{E}_{(x,y)\sim\mathcal{P}}[\ell(w; x, y)]$ with the guarantee of

being differentially private. The utility of w^{priv} is measured by the (expected) excess population risk $\mathbb{E}L_{\mathcal{P}}(w^{\text{priv}}) - \min_{w\in\mathcal{W}} L_{\mathcal{P}}(w)$, where the expectation is taken over the randomness of the algorithm and the input data. Besides the population risk, we can also measure the *empirical risk* of dataset D: $\hat{L}(w, D) = \frac{1}{n}\sum_{i=1}^{n} \ell(w, z_i)$.

It is notable that besides the error bound, to better demonstrate our results, we may also consider the sample complexity to achieve a fixed error α to measure the utility of DP algorithms.

While DP-SO and DP-ERM have been extensively studied for more than a decade, most of the existing work considers the case where the loss function is convex. The problem of DP-SO and DP-ERM with non-convex loss functions remains far from well-understood due to their complex nature. Although there is some preliminary work, such as [5–8], there are still two critical issues. Firstly, most of the existing work adopts the gradient norm of the population risk function to measure the utility, which is quite different from the convex case where we use the excess population risk instead. However, using the gradient norm is inadequate for indicating how close the private model is to the optimal solution [2]. Secondly, while recently there has been some work considering the excess population risk for non-convex loss functions [6], most research has narrowly focused on general non-convex loss functions, overlooking the specificities of neural network structures.

2 OUR CONTRIBUTIONS

To address these issues, this paper provides the first comprehensive and theoretical study of DP Fully Connected Neural Networks (with a single output node) and presents several bounds of excess population risk. Specifically, our contributions can be summarized as follows:

Well-specified Model: Bounded Link Function Case. In the first part of the paper, we focus on the simplest neural network structure: neural networks without hidden nodes. i.e., Generalized Linear Models (GLMs). We first address the well-specified model that is characterized by zero-mean random noise, combined with bounded and Lipschitz link (activation) functions, ,eaning that the Bayes optimal classifier satisfies $\mathbb{E}[y|x] = \sigma(\langle w^*, x\rangle)$ for some underlying parameter $w^* \in \mathbb{R}^d$ and non-convex link (activation) function σ:

$$y = \sigma(\langle w^*, x\rangle) + \zeta, \tag{1}$$

where ζ is random noise with zero mean.

For this setup, we introduce an (ϵ, δ)-DP algorithm and demonstrate its efficacy with an upper bound $\tilde{O}(\frac{1}{\sqrt{n}} + \min\{\frac{1}{(n\epsilon)^{\frac{2}{3}}}, \frac{\sqrt{\theta}}{n\epsilon}\})$ for the output. Here $\theta \le n$ is an upper bound on the expected rank of the data matrix and n is the sample size.

Well-specified Model: ReLU Case. We then broaden our study to cases with unbounded link functions, specifically when employing the ReLU activation function. In this scenario, we show that an upper bound of $\tilde{O}(\frac{1}{\sqrt{n}} + \min\{\frac{\sqrt{d}}{n\epsilon}, \frac{1}{(n\epsilon)^{\frac{2}{3}}}\})$ for the output is feasible. We also extend our above ideas to the problem of privately learning two-layer neural networks and establish the sample complexity for the cases where the activation functions are either sigmoid or ReLU.

Misspecified Model: ReLU Case. Subsequently, our attention pivots to the misspecified model. To delineate its nuances vis-à-vis the well-specified model, we spotlight the ReLU activation function as a representative case. Within this scope, we innovate a distinct version of DP Gradient Descent, showcasing a sample complexity of $\tilde{O}(\max\{\frac{d\sqrt{d}}{\epsilon\alpha}, \frac{d}{\alpha^2}\})$. This sample complexity guarantees that the difference between the population risk of our private estimator and $c \cdot \text{opt}$ is no more than α, where opt is the optimal value and $c > 0$ is some constant.

DP-SGD for Multi-layer Fully Connected Neural Networks. In previous sections, we examined GLMs and one-hidden layer neural networks, but there are three critical issues with those results: (1) While we proposed several new algorithms, DP-SGD based methods [1] are preferred in practice for private neural network training. Can we obtain utility guarantees for vanilla DP-SGD in [1]? Alternatively, how do different factors such as the number of nodes, clipping threshold, and iteration number impact the utility theoretically? (2) Most of the aforementioned results rely on the well-specified model assumption and the squared loss in population risk, which can be too stringent in practice. Can we provide utility analysis without these assumptions? (3) Previous methods for one-hidden layer networks heavily depend on their specific forms and cannot be extended to general multi-layer structures. To address these issues, we study the utility of the projected version of DP-SGD for general multi-layer neural networks.

Drawing upon recent advancements in the Neural Tangent Kernel (NTK), we present the inaugural excess population risk bound for networks where both the width of each layer and the sample size are sufficiently large:

$$\underbrace{\sqrt{\frac{\log(\frac{1}{\xi})}{T}}}_{\textit{Convergence rate}} + \underbrace{\inf_{f\in\mathcal{F}(\mathbf{W}^{(0)}, \frac{R}{\sqrt{m}})} \{\frac{1}{T}\sum_{i=1}^{T} \ell(f(\mathbf{x}_i), y_i)\}}_{\textit{Approximation error}}$$
$$+ SL^{\frac{3}{2}} R \cdot \underbrace{\widetilde{O}(\frac{\max(L, \frac{d}{m})\log(\frac{1}{\gamma})\log(\frac{1}{\delta})m^2\sqrt{T}}{n^2\epsilon^2})}_{\textit{Privacy error}}, \tag{2}$$

where $\mathcal{F}(\mathbf{W}^{(0)}, \frac{R}{\sqrt{m}})$ is the Neural Tangent Random Feature function class with radius $\frac{R}{\sqrt{m}}$. In essence, this bound is composed of three elements: an approximation error attributable to NTK, an error arising from the Gaussian noise introduced in every iteration, and a combined term representing the convergence rate and sampling error. Building on our theoretical framework, we then delve into the intricate interplay and trade-offs between various parameters. We also provide experimental studies to corroborate our theoretical findings.

ACKNOWLEDGMENTS

Di Wang and Cheng-long Wang are supported in part by the funding BAS/1/1689-01-01,RGC/3/7125-01-01, and King Abdullah University of Science and Technology (KAUST) – Center of Excellence for Generative AI, under award number 5940 and a gift from Google.

REFERENCES

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

[2] Naman Agarwal, Zeyuan Allen-Zhu, Brian Bullins, Elad Hazan, and Tengyu Ma. 2017. Finding approximate local minima faster than gradient descent. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*. 1195–1199.

[3] Raef Bassily, Adam Smith, and Abhradeep Thakurta. 2014. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on*. IEEE, 464–473.

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

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

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

[7] Di Wang and Jinhui Xu. 2019. Differentially private empirical risk minimization with smooth non-convex loss functions: A non-stationary view. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 1182–1189.

[8] Lingxiao Wang, Bargav Jayaraman, David Evans, and Quanquan Gu. 2019. Efficient privacy-preserving stochastic nonconvex optimization. *arXiv preprint arXiv:1910.13659* (2019).