Faster Rates of Private Stochastic Convex Optimization

Jinyan Su HUIZHANG@STD.UESTC.EDU.CN
School of Mathematical Sciences
University of Electronic Science and Technology of China[*†]

Lijie Hu LIJIE.HU@KAUST.EDU.SA
Division of Computer, Electrical and Mathematical Sciences and Engineering
King Abdullah University of Science and Technology

Di Wang DI.WANG@KAUST.EDU.SA
Division of Computer, Electrical and Mathematical Sciences and Engineering
King Abdullah University of Science and Technology

Editors: Sanjoy Dasgupta and Nika Haghtalab

Abstract

In this paper, we revisit the problem of Differentially Private Stochastic Convex Optimization (DP-SCO) and provide excess population risks for some special classes of functions that are faster than the previous results of general convex and strongly convex functions. In the first part of the paper, we study the case where the population risk function satisfies the Tysbakov Noise Condition (TNC) with some parameter $\theta > 1$. Specifically, we first show that under some mild assumptions on the loss functions, there is an algorithm whose output could achieve an upper bound of $\widetilde{O}((\frac{1}{\sqrt{n}} + \frac{d}{n\epsilon})^{\frac{\theta}{\theta-1}})$ and $\widetilde{O}((\frac{1}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}}{n\epsilon})^{\frac{\theta}{\theta-1}})$ for ϵ-DP and (ϵ, δ)-DP, respectively when $\theta \geq 2$, here n is the sample size and d is the dimension of the space. Then we address the inefficiency issue, improve the upper bounds by Poly$(\log n)$ factors and extend to the case where $\theta \geq \bar{\theta} > 1$ for some known $\bar{\theta}$. Next we show that the excess population risk of population functions satisfying TNC with parameter $\theta \geq 2$ is always lower bounded by $\Omega((\frac{d}{n\epsilon})^{\frac{\theta}{\theta-1}})$ and $\Omega((\frac{\sqrt{d\log(1/\delta)}}{n\epsilon})^{\frac{\theta}{\theta-1}})$ for ϵ-DP and (ϵ, δ)-DP, respectively, which matches our upper bounds. In the second part, we focus on a special case where the population risk function is strongly convex. Unlike the previous studies, here we assume the loss function is *non-negative* and *the optimal value of population risk is sufficiently small*. With these additional assumptions, we propose a new method whose output could achieve an upper bound of $O(\frac{d\log(1/\delta)}{n^2\epsilon^2} + \frac{1}{n^{\tau}})$ and $O(\frac{d^2}{n^2\epsilon^2} + \frac{1}{n^{\tau}})$ for any $\tau > 1$ in (ϵ, δ)-DP and ϵ-DP model respectively if the sample size n is sufficiently large. These results circumvent their corresponding lower bounds in (Feldman et al., 2020) for general strongly convex functions. Finally, we conduct experiments of our new methods on real world data. Experimental results also provide new insights into established theories.

Keywords: Differential Privacy, Stochastic Convex Optimization

[*] Part of the work was done when Jinyan Su was a research intern at KAUST.
[†] Extended abstract. Full version appears as (Su et al., 2022).

1. Introduction

Preserving the privacy of training data has become an important consideration and now is a challenging task for machine learning algorithms. To address the privacy issue, Differential Privacy (DP) (Dwork et al., 2006), which roots in cryptography, is a strong mathematical scheme for privacy preserving. It allows for rich statistical and machine learning analysis, and is now becoming a de facto notation for private data analysis. Methods to guarantee differential privacy have been widely studied, and recently adopted in industry (Tang et al., 2017; Ding et al., 2017).

As one of the most important problems in Machine Learning and Differential Privacy community, the Empirical Risk Minimization problem in the DP model, *i.e.,* DP-ERM, has been studied quite well in the last decade, starting from (Chaudhuri et al., 2011), such as (Bassily et al., 2014; Wang et al., 2017, 2019a; Wu et al., 2017; Kasiviswanathan and Jin, 2016; Kifer et al., 2012; Smith et al., 2017; Wang et al., 2018, 2019b; Asi et al., 2021a). Besides DP-ERM, its population (or expected) version, namely Differentially Private Stochastic Convex Optimization (DP-SCO), has received much attention in recent years, starting from (Bassily et al., 2014). DP-SCO is defined as the follows.

Definition 1 (DP-SCO (Bassily et al., 2014)) *Given a dataset $S = \{x_1, \cdots, x_n\}$ from a data universe $\mathcal{X}$ where x_i are i.i.d. samples from some unknown distribution $\mathcal{D}$, a convex loss function $f(\cdot, \cdot)$, and a convex constraint set $\mathcal{W} \subseteq \mathbb{R}^d$, Differentially Private Stochastic Convex Optimization (DP-SCO) is to find w^{priv} so as to minimize the population risk,* i.e., $F(w) = \mathbb{E}_{x\sim\mathcal{D}}[f(w, x)]$ *with the guarantee of being differentially private. The utility of the algorithm is measured by the (expected) excess population risk, that is $\mathbb{E}_{\mathcal{A}}[F(w^{priv})] - \min_{w\in\mathcal{W}} F(w)$, where the expectation of $\mathcal{A}$ is taken over all the randomness of the algorithm. Besides the population risk, we can also measure the empirical risk of dataset S: $\bar{F}(w, S) = \frac{1}{n}\sum_{i=1}^{n} f(w, x_i)$.*

Specifically, (Bassily et al., 2019) first provides the optimal rate of DP-SCO with general convex loss functions in (ϵ, δ)-DP, which is quite different from the optimal rate in DP-ERM. Later, (Feldman et al., 2020) extends this problem to strongly convex and (or) non-smooth cases by providing a general localization technique. Moreover, their methods have linear time complexity if the loss functions are smooth. For non-smooth loss functions, (Kulkarni et al., 2021) recently proposes a new method which only need subquadratic gradient complexity. While there are already a large number of studies on DP-SCO, the problem is still far from well understood. A key observation is that, all of the previous works only focus on the the case where the loss functions are either general convex or strongly convex. However, there are also many problems that are even stronger than strongly convex functions, or fall between convex and strongly convex functions. In the non-private counterpart, various studies have attempted to get faster rates by imposing additional assumptions on the loss functions. And it has been shown that it is indeed possible to achieve rates that are faster than the rates of general convex loss functions (Yang et al., 2018; Koren and Levy, 2015; van Erven et al., 2015), or it could even achieve the same rate as in the strongly convex case even if the function is not strongly convex (Karimi et al., 2016; Liu et al., 2018; Xu et al., 2017). Motivated by this, our question is,

For the problem of DP-SCO with special classes of population risk functions, is it possible to achieve faster rates of the excess population risk than the optimal ones of general convex and (or) strongly convex cases?

In this paper, we will mainly study the case where the population risk satisfies the Tysbakov Noise Condition (TNC) (Ramdas and Singh, 2012; Liu et al., 2018) [1] , which includes strongly convex functions, SVM and linear regression as special cases. Moreover, it has been studied quite well and has been shown that it could achieve faster rates than the optimal one of general convex loss functions in the non-private case. Specifically,

Definition 2 *For a convex function $F(\cdot)$, let $\mathcal{W}_* = \arg\min_{w\in\mathcal{W}} F(w)$ denote the optimal set and for any $w \in \mathcal{W}$, let $w^* = \arg\min_{u\in\mathcal{W}_*} \|u - w\|_2$ denote the projection of w onto the optimal set $\mathcal{W}_*$. Function F satisfies (θ, λ)-TNC for some $\theta > 1$ and $\lambda > 0$ if for any $w \in \mathcal{W}$ the following inequality holds*

$$F(w) - F(w^*) \geq \lambda ||w - w^*||_2^\theta. \tag{1}$$

Our contributions can be summarized as follows.

- In the first part of the paper, we study the problem where the population risk satisfying (θ, λ)-TNC and propose three methods.
  1. When $\theta \geq 2$ and is known, we first propose a method that could achieve an excess population risk of $\widetilde{O}((\frac{1}{\sqrt{n}} + \frac{d}{n\epsilon})^{\frac{\theta}{\theta-1}})$ and $\widetilde{O}((\frac{1}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}}{n\epsilon})^{\frac{\theta}{\theta-1}})$ in ϵ-DP and (ϵ, δ)-DP model respectively under the assumption that the loss function is smooth and Lipschitz, where n is the sample size of the data and d is the dimension of the space. Our idea is based on the localization technique proposed by (Feldman et al., 2020), which provides an algorithm, namely Phased-SGD for DP-SCO with general convex loss functions. We propose an adaptive stochastic approximation algorithm. The updates are divided into m stages. At each stage, the Phased-SGD algorithm is applied with a number of samples. Each employment of the Phased-SGD algorithm is warm-started by the initial point that is returned from the last stage.
  2. However, in practice, the main difficulty on implementing the previous Algorithm is we need to implement the projection onto the intersection of two balls in each iteration of the Phased-SGD in each phase. To overcome the challenge, instead of considering the original stochastic function, we focus on the problem with an additional and adaptive strongly convex regularization in each stage. Specifically, we first divide the whole algorithm into m stages. In each stage we hope to find a private estimator w^k such that $w^k \approx \arg\min_{w\in\mathcal{W}} F(w) + \frac{1}{2\gamma_k}\|w - w_{k-1}\|_2^2$ with γ_k changing with k. Note that due to the additional ℓ_2 regularization, now the function is strongly convex. Thus, instead of using the original Phased-SGD for general convex loss, here we use a strongly convex version of Phased-SGD, which is adopted from (Feldman et al., 2020). Moreover, since now we have an additional ℓ_2-norm regularization, here we do not need the projection onto the intersection of two balls during updates compared with the previous algorithm.
  3. Finally, we propose another method to resolve the inefficiency issue under the assumption that θ is known. The idea of our algorithm is as the following: assuming that the value of θ is unknown, but θ is lower bounded by some known constant $\bar{\theta} > 1$, namely $\theta \geq \bar{\theta} > 1$. We first divide the whole dataset into $k = \lfloor(\log_{\bar{\theta}} 2) \cdot \log\log n\rfloor$ disjoint subsets, where

1. In some related work it is also called the Error Bound Condition or the Growth Condition (Liu et al., 2018; Xu et al., 2017).

the i-th subset has $n_i = 2^{i-1}n/(\log n)^{\log_{\bar{\theta}} 2}$ samples; then we repeat the Phased-SGD for k times where each phase runs on the i-th subset and is initialized at the output of the previous phase. We also show that our algorithm improves the previous upper bounds of error by $\text{Poly}(\log n)$ factors and it outperforms the previous methods practically.

- Next, we focus on the lower bounds of the excess population risk. Specifically, for any $\theta \geq 2$, we show that there is a population risk function satisfying TNC with parameter θ such that for any ϵ-DP ((ϵ, δ)-DP) algorithm, its output achieves an excess risk of $\Omega((\frac{d}{n\epsilon})^{\frac{\theta}{\theta-1}})$ ($\Omega((\frac{\sqrt{d\log(1/\delta)}}{n\epsilon})^{\frac{\theta}{\theta-1}})$) with high probability.
- In the second part of the paper, we will focus on the problem where the population risk function is strongly convex, which is a special case of TNC functions with $\theta = 2$. Unlike the previous studies, here we assume the loss function is non-negative and the optimal value of the population is sufficiently small. With these additional assumptions, we propose a new method whose output could achieve an upper bound of $O(\frac{d\log(1/\delta)}{n^2\epsilon^2} + \frac{1}{n^\tau})$ and $O(\frac{d^2}{n^2\epsilon^2} + \frac{1}{n^\tau})$ for any $\tau > 1$ in (ϵ, δ)-DP and ϵ-DP model respectively if the sample size n is sufficiently large. These rates circumvent their corresponding lower bounds for general strong convex functions in (Feldman et al., 2020), *i.e.,* $\Theta(\frac{d^2}{n^2\epsilon^2} + \frac{1}{n})$ for ϵ-DP and $\Theta(\frac{d\log(1/\delta)}{n^2\epsilon^2} + \frac{1}{n})$ for (ϵ, δ)-DP.

  There are two parts in the algorithm. In the first part, we perform the original Iterated Phased-SGD on the first half of the data to get a good solution to the optimal parameter w^*. After that we perform a new method, namely Epoch-DP-SGD on the second half of the data, which may also be used in other problems.

2. Related Work

Starting from (Chaudhuri et al., 2011), a long list of works have attacked the problems of DP-ERM from different perspectives: (Bassily et al., 2014; Iyengar et al., 2019; Zhou et al., 2020; Song et al., 2020; Wang et al., 2017; Zhang et al., 2017) studied the problems in the low dimensional case and the central model, (Kasiviswanathan and Jin, 2016; Kifer et al., 2012; Talwar et al., 2015; Wang and Gu, 2020; Cai et al., 2020) considered the problems in the high dimensional sparse case and the central model, (Smith et al., 2017; Duchi et al., 2013; Wang et al., 2020a; Duchi et al., 2018) focused on the problems in the local model. However, almost all of these works only focus the case where the empirical risk function is either general convex or strongly convex. For special class of functions, (Wang et al., 2017) studies the empirical risk functions satisfying Polyak-Lojasiewicz (PL) condition, which is weaker than strongly convexity and show that it is possible to achieve an excess empirical risk of $O(\frac{d\log(1/\delta)}{n^2\epsilon^2})$, which is the same as the strongly convex loss. And the PL condition is equivalent to TNC with parameter $\theta = 2$. Thus, in this paper we extend the result from the empirical risk to the population risk function.

For DP-SCO, besides the related work we mentioned in the previous section, there is another direction which studies some special cases of DP-SCO. For example, (Bassily et al., 2021) and (Asi et al., 2021a) consider the case where the underlying constraint set $\mathcal{W}$ has specific geometric structures, such as polyhedron. (Guzmán et al., 2021) studies the (non)smooth and (non)convex generalized linear loss. (Wang et al., 2020b) and (Kamath et al., 2021) focus on the case where the distribution of the data or the gradient of the loss function is heavy-tailed. However, none of these works study the case where the population risk satisfies TNC. (Liu et al., 2021) recently studies the theoretical guarantees of the PATE model (Papernot et al., 2016) under the assumption that the population risk function satisfies TNC and shows that it is possible to achieve faster rates than in the convex case (Bassily et al., 2018). However, since here we focus on a different problem, their results cannot be used to DP-SCO.

Concurrent Work: We notice that (Asi et al., 2021b) also studies DP-SCO with TNC population risk functions concurrently. However, compared with its results there are several critical differences.

1) The idea of Algorithm 2 in (Asi et al., 2021b) is similar to one algorithm in our paper. However, the idea of proof and the choice of parameters are quite different.

2) The same as our first algorithm, Algorithm 2 in (Asi et al., 2021b) is also inefficient and has poor performance in practice. To resolve the issue, we also develop two other algorithms.

3) For (ϵ, δ)-DP model, (Asi et al., 2021b) only shows the worst-case lower bound under the assumption $\theta \geq 1 + c$ for some constant $c > 0$ while in this paper we also extend the result to $\theta \geq 2$. Although the hard instance in (Asi et al., 2021b) is similar to ours, the proofs of lower bounds are different.

4) In this paper, we also provide experimental results on the problem which has not been studied in (Asi et al., 2021b).

5) Besides TNC population risk functions, in this paper we also provide faster rates of DP-SCO with strongly convex loss function with additional assumptions which also has not been studied in (Asi et al., 2021b).

Acknowledgments

Di Wang, Lijie Hu and Jinyan Su were support in part by the baseline funding BAS/1/1689-01-01 and funding from the AI Initiative REI/1/4811-10-01 of King Abdullah University of Science and Technology (KAUST).

References

Hilal Asi, John Duchi, Alireza Fallah, Omid Javidbakht, and Kunal Talwar. Private adaptive gradient methods for convex optimization. In *International Conference on Machine Learning*, pages 383–392. PMLR, 2021a.

Hilal Asi, Daniel Levy, and John Duchi. Adapting to function difficulty and growth conditions in private optimization. *arXiv preprint arXiv:2108.02391*, 2021b.

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

Raef Bassily, Abhradeep Guha Thakurta, and Om Dipakbhai Thakkar. Model-agnostic private learning. *Advances in Neural Information Processing Systems*, 2018.

Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Thakurta. Private stochastic convex optimization with optimal rates. *arXiv preprint arXiv:1908.09970*, 2019.

Raef Bassily, Cristóbal Guzmán, and Anupama Nandi. Non-euclidean differentially private stochastic convex optimization. *arXiv preprint arXiv:2103.01278*, 2021.

T Tony Cai, Yichen Wang, and Linjun Zhang. The cost of privacy in generalized linear models: Algorithms and minimax lower bounds. *arXiv preprint arXiv:2011.03900*, 2020.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3), 2011.

Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. Collecting telemetry data privately. In *Advances in Neural Information Processing Systems*, pages 3571–3580, 2017.

John C Duchi, Michael I Jordan, and Martin J Wainwright. Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 429–438. IEEE, 2013.

John C Duchi, Michael I Jordan, and Martin J Wainwright. Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association*, 113(521):182–201, 2018.

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.

Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: optimal rates in linear time. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 439–449, 2020.

Cristóbal Guzmán, Raef Bassily, and Michael Menart. Differentially private stochastic optimization: New results in convex and non-convex settings. *arXiv preprint arXiv:2107.05585*, 2021.

Roger Iyengar, Joseph P Near, Dawn Song, Om Thakkar, Abhradeep Thakurta, and Lun Wang. Towards practical differentially private convex optimization. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 299–316. IEEE, 2019.

Gautam Kamath, Xingtu Liu, and Huanyu Zhang. Improved rates for differentially private stochastic convex optimization with heavy-tailed data. *arXiv preprint arXiv:2106.01336*, 2021.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer, 2016.

Shiva Prasad Kasiviswanathan and Hongxia Jin. Efficient private empirical risk minimization for high-dimensional learning. In *International Conference on Machine Learning*, pages 488–497, 2016.

Daniel Kifer, Adam Smith, and Abhradeep Thakurta. Private convex empirical risk minimization and high-dimensional regression. In *Conference on Learning Theory*, pages 25–1, 2012.

Tomer Koren and Kfir Y Levy. Fast rates for exp-concave empirical risk minimization. In *NIPS*, pages 1477–1485, 2015.

Janardhan Kulkarni, Yin Tat Lee, and Daogao Liu. Private non-smooth empirical risk minimization and stochastic convex optimization in subquadratic steps. *arXiv preprint arXiv:2103.15352*, 2021.

Chong Liu, Yuqing Zhu, Kamalika Chaudhuri, and Yu-Xiang Wang. Revisiting model-agnostic private learning: Faster rates and active learning. In *International Conference on Artificial Intelligence and Statistics*, pages 838–846. PMLR, 2021.

Mingrui Liu, Xiaoxuan Zhang, Lijun Zhang, Rong Jin, and Tianbao Yang. Fast rates of erm and stochastic approximation: Adaptive to error bound conditions. *arXiv preprint arXiv:1805.04577*, 2018.

Nicolas Papernot, Martín Abadi, Ulfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. *arXiv preprint arXiv:1610.05755*, 2016.

Aaditya Ramdas and Aarti Singh. Optimal rates for first-order stochastic convex optimization under tsybakov noise condition. *arXiv preprint arXiv:1207.3012*, 2012.

Adam Smith, Abhradeep Thakurta, and Jalaj Upadhyay. Is interaction necessary for distributed private learning? In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 58–77. IEEE, 2017.

Shuang Song, Om Thakkar, and Abhradeep Thakurta. Characterizing private clipped gradient descent on convex generalized linear problems. *arXiv preprint arXiv:2006.06783*, 2020.

Jinyan Su, Lijie Hu, and Di Wang. Faster rates of private stochastic convex optimization. *arXiv preprint arXiv:2108.00331*, 2022.

Kunal Talwar, Abhradeep Thakurta, and Li Zhang. Nearly-optimal private lasso. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 2*, pages 3025–3033, 2015.

Jun Tang, Aleksandra Korolova, Xiaolong Bai, Xueqiang Wang, and XiaoFeng Wang. Privacy loss in apple's implementation of differential privacy on macos 10.12. *CoRR*, abs/1709.02753, 2017.

Tim van Erven, Peter D Grünwald, Nishant A Mehta, Mark D Reid, and Robert C Williamson. Fast rates in statistical and online learning. *Journal of Machine Learning Research*, 16:1793–1861, 2015.

Di Wang, Minwei Ye, and Jinhui Xu. Differentially private empirical risk minimization revisited: Faster and more general. In *Advances in Neural Information Processing Systems*, pages 2722–2731, 2017.

Di Wang, Marco Gaboardi, and Jinhui Xu. Empirical risk minimization in non-interactive local differential privacy revisited. In *Advances in Neural Information Processing Systems*, pages 965–974, 2018.

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

Di Wang, Adam Smith, and Jinhui Xu. Noninteractive locally private learning of linear models via polynomial approximations. In *Algorithmic Learning Theory*, pages 897–902, 2019b.

Di Wang, Marco Gaboardi, Adam Smith, and Jinhui Xu. Empirical risk minimization in the non-interactive local model of differential privacy. *Journal of Machine Learning Research*, 21(200): 1–39, 2020a.

Di Wang, Hanshen Xiao, Srinivas Devadas, and Jinhui Xu. On differentially private stochastic convex optimization with heavy-tailed data. In *International Conference on Machine Learning*, pages 10081–10091. PMLR, 2020b.

Lingxiao Wang and Quanquan Gu. A knowledge transfer framework for differentially private sparse learning. In *AAAI*, pages 6235–6242, 2020.

Xi Wu, Fengan Li, Arun Kumar, Kamalika Chaudhuri, Somesh Jha, and Jeffrey Naughton. Bolt-on differential privacy for scalable stochastic gradient descent-based analytics. In *Proceedings of the 2017 ACM International Conference on Management of Data*, pages 1307–1322. ACM, 2017.

Yi Xu, Qihang Lin, and Tianbao Yang. Stochastic convex optimization: Faster local growth implies faster global convergence. In *International Conference on Machine Learning*, pages 3821–3830. PMLR, 2017.

Tianbao Yang, Zhe Li, and Lijun Zhang. A simple analysis for exp-concave empirical minimization with arbitrary convex regularizer. In *International Conference on Artificial Intelligence and Statistics*, pages 445–453. PMLR, 2018.

Jiaqi Zhang, Kai Zheng, Wenlong Mou, and Liwei Wang. Efficient private erm for smooth objectives. *arXiv preprint arXiv:1703.09947*, 2017.

Yingxue Zhou, Zhiwei Steven Wu, and Arindam Banerjee. Bypassing the ambient dimension: Private sgd with gradient subspace identification. *arXiv preprint arXiv:2007.03813*, 2020.