Proceedings of Machine Learning Research vol xxx:1–6, 2022

Open Problem: Better Differentially Private Learning Algorithms with Margin Guarantees

Raef Bassily BASSILY.1@OSU.EDU
The Ohio State University & Google Research NY.

Mehryar Mohri MOHRI@GOOGLE.COM
Google Research & Courant Institute of Mathematical Sciences, New York, NY.

Ananda Theertha Suresh THEERTHA@GOOGLE.COM
Google Research NY.

Abstract

The design of efficient differentially private (DP) learning algorithms with dimension-independent learning guarantees has been one of the central challenges in the field of privacy-preserving machine learning. Existing algorithms either suffer from weak generalization guarantees, restrictive model assumptions, or quite large computation cost. In non-private learning, dimension-independent generalization guarantees based on the notion of *confidence margin* were shown to be the most informative and useful learning guarantees. This motivates a systematic study of DP learning algorithms with confidence-margin generalization guarantees. A recent work has started exploring this direction in the context of linear and kernel-based classification as well as certain classes of neural networks (NNs). Despite showing several positive results, a number of fundamental questions are still open. We identify two major open problems related to DP margin-based learning algorithms. The first problem relates to the design of algorithms with more favorable computational cost. The second one pertains to the question of achieving margin guarantees for NNs under DP with no explicit dependence on the network size.

1. Introduction

Preserving privacy is a crucial objective for machine learning algorithms. A widely adopted criterion in statistical data privacy is the notion of differential privacy (DP) Dwork et al. (2006); Dwork (2006); Dwork and Roth (2014), which ensures that the information gained by an adversary is roughly invariant to the presence or absence of an individual in a dataset. Despite the remarkable theoretical and algorithmic progress in differential privacy over the last decade or more, however, its application to learning still faces several obstacles. A recent series of publications have shown that differentially private PAC learning of infinite hypothesis sets is not possible, even for common hypothesis sets such as that of linear functions. In fact, this is the case for any hypothesis set containing threshold functions Bun et al. (2015); Alon et al. (2019). These results imply serious limitations for private agnostic learnability.

Another rich body of literature has studied differentially private empirical risk minimization (DP-ERM) and differentially private stochastic convex optimization (DP-SCO) (e.g., Chaudhuri et al. (2011); Jain and Thakurta (2014); Bassily et al. (2014, 2019); Feldman et al. (2020); Song et al. (2021); Bassily et al. (2021b); Asi et al. (2021); Bassily et al. (2021a)). When the underlying optimization problem is constrained (*constrained setting*), tight upper and lower bounds have been derived for the excess empirical risk of DP-ERM Bassily et al. (2014) and for the excess population risk for DP-SCO Bassily et al. (2019); Feldman et al. (2020). These results show that learning guarantees necessarily admit a dependency on the dimension d of the form $\sqrt{d}/m$, where m is the sample size. This dependency is persistent, even in the special case of *generalized linear losses* (GLLs) Bassily et al. (2014), which limits the benefit of such guarantees, since learning algorithms typically deal with high-dimensional spaces.

When the underlying optimization problem is unconstrained (*unconstrained setting*) and the loss is a generalized linear loss, the bounds given by Jain and Thakurta (2014), Song et al. (2021) and Bassily et al. (2021a) are dimension-independent but they admit a dependency on $\|w^*\|^2$, where w^* is the unconstrained minimizer of the expected loss (population risk), or $\|\widehat{w}\|^2$, where $\widehat{w}$ is the unconstrained minimizer of the empirical loss. Since the problem is unconstrained, the norm of these vectors can be very large, even for classification problems for which the minimizer of the zero-one loss admits a relatively small norm. Thus, in both the constrained and unconstrained settings, the learning guarantees derived from DP-ERM and DP-SCO are weak for hypothesis sets commonly used in machine learning.

The results just mentioned raise some fundamental questions about private learning: is differentially private learning with better, dimension-independent guarantees possible for standard hypothesis sets? Must one resort to distribution-dependent bounds instead? In view of the negative PAC-learning results and other learning bounds mentioned earlier, a natural direction to pursue is that of optimistic margin-based learning bounds. Learning bounds based on the notion of confidence margin have been shown to be the most informative and useful guarantees (Koltchinskii and Panchenko, 2002; Schapire et al., 1997; Mohri et al., 2018; Cortes et al., 2021). This motivates our study of differentially private learning algorithms with margin-based guarantees. Note that our *confidence-margin* analysis and guarantees do not require the hard-margin separability assumptions adopted in Blum et al. (2005); Le Nguyen et al. (2020), which is a strong assumption that typically does not hold in practice.

1.1. Preliminaries

We consider an input space $\mathcal{X}$, a binary output space $\mathcal{Y} = \{-1, +1\}$ and a hypothesis set $\mathcal{H}$ of functions mapping from $\mathcal{X}$ to $\mathbb{R}$. We denote by $\mathcal{D}$ a distribution over $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and denote by $R_{\mathcal{D}}(h)$ the generalization error and by $\widehat{R}_S(h)$ the empirical error of a hypothesis $h \in \mathcal{H}$:

$$R_{\mathcal{D}}(h) = \mathop{\mathbb{E}}_{z=(x,y)\sim\mathcal{D}}[1_{yh(x)\leq 0}] \qquad \widehat{R}_S(h) = \mathop{\mathbb{E}}_{z=(x,y)\sim S}[1_{yh(x)\leq 0}].$$

where we write $z \sim S$ to indicate that z is randomly drawn from the empirical distribution defined by the dataset S. Given $\rho \geq 0$, the ρ-margin loss and empirical ρ-margin loss of $h \in \mathcal{H}$ are defined as:

$$R^{\rho}_{\mathcal{D}}(h) = \mathop{\mathbb{E}}_{z=(x,y)\sim\mathcal{D}}[1_{yh(x)\leq \rho}] \qquad \widehat{R}^{\rho}_S(h) = \mathop{\mathbb{E}}_{z=(x,y)\sim S}[1_{yh(x)\leq \rho}].$$

We also consider the convex ρ-hinge loss that enables devising computationally-efficient algorithms. For any $\rho > 0$, define ρ-hinge loss as $\ell^{\rho}(u) \triangleq \max(1 - u/\rho, 0)$, $u \in \mathbb{R}$. Similar to the above definitions, given $\rho > 0$, for a dataset S, we define the ρ-hinge loss and empirical ρ-hinge loss as

$$L^{\rho}_{\mathcal{D}}(w) = \mathop{\mathbb{E}}_{z=(x,y)\sim\mathcal{D}}[\ell^{\rho}(y_i \langle w, x_i \rangle)] \qquad \widehat{L}^{\rho}_S(w) = \mathop{\mathbb{E}}_{z=(x,y)\sim S}[\ell^{\rho}(y_i \langle w, x_i \rangle)].$$

In the context of learning, differential privacy is defined as follows.

Differential Privacy: Let $\varepsilon, \delta \geq 0$. Let $\mathcal{A}: (\mathcal{X} \times \mathcal{Y})^m \to \mathcal{H}$ be a randomized algorithm. We say that $\mathcal{A}$ is (ε, δ)-*DP* if for any measurable subset $O \subset \mathcal{H}$ and all $S, S' \in (\mathcal{X} \times \mathcal{Y})^m$ that differ in one sample, the following inequality holds:

$$\mathbb{P}(\mathcal{A}(S) \in O) \leq e^{\varepsilon}\, \mathbb{P}(\mathcal{A}(S') \in O) + \delta. \tag{1}$$

If $\delta = 0$, we refer to this guarantee as *pure differential privacy*.

1.2. Existing work on DP learning with confidence-margin guarantees.

Building on the embedding idea of Le Nguyen et al. (2020), a recent paper (Bassily et al., 2022) gave DP learning algorithms with confidence-margin guarantees for learning linear classifiers, kernel classifiers, and neural-network classifiers.

1.2.1. LINEAR AND KERNEL CLASSIFIERS

Let $\mathbb{B}^d(r) \triangleq \left\{x \in \mathbb{R}^d : \|x\|_2 \leq r\right\}$ denote the Euclidean ball in $\mathbb{R}^d$ of radius r and let $\mathcal{X} \subseteq \mathbb{B}^d(r)$ denote the feature space. The class of linear predictors over $\mathcal{X}$ is defined as $\mathcal{H}_{\mathsf{Lin}} = \left\{h_w : x \mapsto \langle w, x\rangle \mid w \in \mathbb{B}^d(\Lambda)\right\}$. Here, one may view the dimension d as possibly much larger than the sample size m. Bassily et al. (2022) give an (ε, δ)-DP that outputs a linear predictor h^{Priv} over $\mathbb{R}^d$ such that with high probability over an input sample $S \sim \mathcal{D}^m$ and the algorithm's randomness,

$$R_{\mathcal{D}}(w^{\mathsf{Priv}}) \leq \min_{w \in \mathbb{B}^d(\Lambda)} \widehat{L}^{\rho}_S(w) + \widetilde{O}\left(\frac{\Lambda r}{\rho\sqrt{\min(1, \varepsilon)\, m}}\right). \tag{2}$$

Their algorithm is based on fast construction of Johnson-Lindenstrauss (JL) embedding combined with a DP-ERM algorithm. We note that the margin bound (2) nearly matches the standard, non-private analog. However, the computational cost of this algorithm is $\widetilde{O}(md)$, which is quite large and can be prohibitive from a practical stand point.
The same reference also considers the family of kernel-based classifiers w.r.t. a continuous, positive definite, shift-invariant kernel $K: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Let $\mathbb{H}$ denote the the reproducing kernel Hilbert space (RKHS) of K. Define $\mathcal{H}_{\Lambda} \triangleq \{h \in \mathbb{H} : \|w\|_{\mathbb{H}} \leq \Lambda\}$, where $\|\cdot\|_{\mathbb{H}}$ is the RKHS norm. Using finite-dimensional kernel approximation combined with their algorithm for linear classifiers, Bassily et al. (2022) give an (ε, δ)-DP algorithm that outputs a classifier h^{Priv} such that with high probability over the input sample $S \sim \mathcal{D}^m$ and the algorithm's randomness,

$$R_{\mathcal{D}}(h^{\mathsf{Priv}}) \leq \min_{h \in \mathcal{H}_{\Lambda}} \widehat{L}^{\rho}_S(h) + \widetilde{O}\left(\frac{\Lambda r}{\rho\sqrt{\min(1, \varepsilon)\, m}}\right). \tag{3}$$

Again, we note that the margin bound (3) nearly matches the standard, non-private analog. However, the computational cost of this algorithm is $\widetilde{O}(m^3 d)$, which is worse than that of linear classifiers. The authors discuss extensions to kernels that are not necessarily shift-invariant, such as polynomial kernels, however, these extensions suffer from the same computational cost.

1.2.2. FEED-FORWARD NEURAL NETWORKS

Bassily et al. (2022) initiate the study of DP learning of neural-network classifiers with margin guarantees. They give a pure DP algorithm for the family of feed-forward neural networks and prove a confidence-margin bound that is independent of the input dimension and scales only linearly with the total number of neurons in the network. More concretely, let $\mathcal{H}_{\mathsf{NN}}$ be a family of L-layer feed-forward neural networks defined over $\mathbb{B}^d(r)$. A function h in $\mathcal{H}_{\mathsf{NN}}$ can be viewed as a cascade of linear maps composed with a non-linear activation function. We define $\mathcal{H}_{\mathsf{NN}^\Lambda}$ as the subset of $\mathcal{H}_{\mathsf{NN}}$ with weight matrices that are Λ-bounded in their Frobenius norm for some $\Lambda > 0$. The width (number of neurons) in each hidden layer, denoted by N, is assumed to be the same for all the layers. Bassily et al. (2022) give a computationally inefficient ε-DP algorithm which returns a neural network $h^{\mathsf{Priv}} \in \mathcal{H}_{\mathsf{NN}}$ such that with high probability over the input sample $S \sim \mathcal{D}^m$ and the algorithm's randomness,

$$R_{\mathcal{D}}(h^{\mathsf{Priv}}) \leq \min_{h \in \mathcal{H}_{\mathsf{NN}^\Lambda}} \widehat{R}^{\rho}_S(h) + O\left(\frac{r\Lambda^L\sqrt{NL}}{\rho\sqrt{m}} + \frac{r^2(2\Lambda)^{2L}NL}{\rho^2\varepsilon m}\right) \tag{4}$$

This result entails a new analysis of an embedding-based "network compression" technique. In particular, the construction of Bassily et al. (2022) is based on using L embeddings given by data-independent JL-transform matrices to reduce the dimension of the inputs in each layer. We note that although bound (4) is more favorable than standard bounds obtained via a uniform convergence argument (which depend on d, as well as the total number of edges $\Omega(N^2)$), this bound is potentially far from optimal in the light of existing non-private margin bounds for neural-network learning (Bartlett et al., 2017). In particular, it is unclear whether the explicit dependence on NL is necessary.

2. Open Problems

Faster constructions for linear and kernel classifiers: Consider the problems of DP learning of linear and kernel-based predictors described in Section 1.2.1. Are there (ε, δ)-DP algorithms for these problems, achieving essentially the same guarantees as those of (2) and (3), with more favorable running-time complexity (in terms of their polynomial dependence on m, d) than the respective algorithms mentioned in Section 1.2.1?

Better margin guarantees for learning neural networks: Consider the problems of DP learning the family $\mathcal{H}_{\mathsf{NN}}$ of feed-forward neural networks described in Section 1.2.2. Is it possible to prove a margin-based generalization guarantee for DP learning of $\mathcal{H}_{\mathsf{NN}}$ with no explicit dependence on the network size? In particular, can we design a DP learning algorithm for $\mathcal{H}_{\mathsf{NN}}$ with the following learning guarantee?

$$R_{\mathcal{D}}(h^{\mathsf{Priv}}) \leq \min_{h \in \mathcal{H}_{\mathsf{NN}^\Lambda}} \widehat{R}^{\rho}_S(h) + O\left(\frac{r\Lambda^L}{\rho\sqrt{m}} + \frac{r^2(\Lambda)^{2L}}{\rho^2\varepsilon m}\right).$$

Acknowledgements

This research was done while RB was visiting Google, NY. RB's research at OSU is supported by NSF Award AF-1908281, NSF Award 2112471, and NSF CAREER Award 2144532.

References

Noga Alon, Roi Livni, Maryanthe Malliaris, and Shay Moran. Private PAC learning implies finite littlestone dimension. In Moses Charikar and Edith Cohen, editors, *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing, STOC 2019, Phoenix, AZ, USA, June 23-26, 2019*, pages 852–860. ACM, 2019.

Hilal Asi, Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: Optimal rates in l1 geometry. *arXiv preprint arXiv:2103.01516*, 2021.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30, 2017.

Raef Bassily, Adam D. Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *55th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2014, Philadelphia, PA, USA, October 18-21, 2014*, pages 464–473. IEEE Computer Society, 2014.

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 Michael Menart. Differentially private stochastic optimization: New results in convex and non-convex settings. *arXiv preprint arXiv:2107.05585. Appeared at NeurIPS 2021.*, 2021a.

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

Raef Bassily, Mehryar Mohri, and Ananda Theertha Suresh. Differentially private learning with margin guarantees. *arXiv preprint arXiv:2204.10376*, 2022.

Avrim Blum, Cynthia Dwork, Frank McSherry, and Kobbi Nissim. Practical privacy: the sulq framework. In Chen Li, editor, *Proceedings of the Twenty-fourth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 13-15, 2005, Baltimore, Maryland, USA*, pages 128–138. ACM, 2005.

Mark Bun, Kobbi Nissim, Uri Stemmer, and Salil P. Vadhan. Differentially private release and learning of threshold functions. In Venkatesan Guruswami, editor, *IEEE 56th Annual Symposium on Foundations of Computer Science, FOCS 2015, Berkeley, CA, USA, 17-20 October, 2015*, pages 634–649. IEEE Computer Society, 2015.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate. Differentially private empirical risk minimization. *J. Mach. Learn. Res.*, 12:1069–1109, 2011.

Corinna Cortes, Mehryar Mohri, and Ananda Theertha Suresh. Relative deviation margin bounds. In *International Conference on Machine Learning*, pages 2122–2131. PMLR, 2021.

Cynthia Dwork. Differential privacy. In Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener, editors, *Automata, Languages and Programming, 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II*, volume 4052 of *Lecture Notes in Computer Science*, pages 1–12. Springer, 2006.

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

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam D. Smith. Calibrating noise to sensitivity in private data analysis. In Shai Halevi and Tal Rabin, editors, *Theory of Cryptography, Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006, Proceedings*, volume 3876 of *Lecture Notes in Computer Science*, pages 265–284. Springer, 2006.

Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: optimal rates in linear time. In Konstantin Makarychev, Yury Makarychev, Madhur Tulsiani, Gautam Kamath, and Julia Chuzhoy, editors, *Proccedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, STOC 2020, Chicago, IL, USA, June 22-26, 2020*, pages 439–449. ACM, 2020.

Prateek Jain and Abhradeep Thakurta. (near) dimension independent risk bounds for differentially private learning. In *ICML*, 2014.

Vladmir Koltchinskii and Dmitry Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, 30, 2002.

Huy Le Nguyen, Jonathan Ullman, and Lydia Zakynthinou. Efficient private algorithms for learning large-margin halfspaces. In *Algorithmic Learning Theory*, pages 704–724. PMLR, 2020.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

Robert E. Schapire, Yoav Freund, Peter Bartlett, and Wee Sun Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. In *ICML*, pages 322–330, 1997.

Shuang Song, Thomas Steinke, Om Thakkar, and Abhradeep Thakurta. Evading the curse of dimensionality in unconstrained private glms. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 2638–2646. PMLR, 13–15 Apr 2021. URL http://proceedings.mlr.press/v130/song21a.html.