Open Problem: Better Differentially Private Learning Algorithms with Margin Guarantees

Raef Bassily  
The Ohio State University & Google Research NY.

Mehryar Mohri  
Google Research & Courant Institute of Mathematical Sciences, New York, NY.

Ananda Theertha Suresh  
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 $X$, a binary output space $Y = \{-1, +1\}$ and a hypothesis set $H$ of functions mapping from $X$ to $R$. We denote by $D$ a distribution over $Z = X \times Y$ and denote by $R_D(h)$ the generalization error and by $R_S(h)$ the empirical error of a hypothesis $h \in H$:

$$R_D(h) = \mathbb{E}_{z \sim D} [1_{y^h(x) \leq 0}], \quad R_S(h) = \mathbb{E}_{z \sim S} [1_{y^h(x) \leq 0}],$$

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 $\rho$-margin loss and empirical $\rho$-margin loss of $h \in H$ are defined as:

$$R^\rho_D(h) = \mathbb{E}_{z \sim D} [1_{y^h(x) \leq \rho}], \quad R^\rho_S(h) = \mathbb{E}_{z \sim S} [1_{y^h(x) \leq \rho}],$$

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

$$L^\rho_D(w) = \mathbb{E}_{z \sim D} [\ell^\rho(y_i \langle w, x_i \rangle)], \quad L^\rho_S(w) = \mathbb{E}_{z \sim S} [\ell^\rho(y_i \langle w, x_i \rangle)].$$

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 $(\varepsilon, \delta)$-DP if for any measurable subset $O \subseteq \mathcal{H}$ and all $S, S' \in (\mathcal{X} \times \mathcal{Y})^m$ that differ in one sample, the following inequality holds:

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

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

$$R_D(w^\text{Priv}) \leq \min_{w \in B^d(\Lambda)} \mathcal{T}_S^\rho(w) + O\left(\frac{\Lambda r}{\rho \sqrt{\min(1, \varepsilon)} m}\right). \quad (2)$$

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 $O(md)$, which is quite large and can be prohibitive from a practical standpoint.

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 $\mathcal{H}$ denote the reproducibility kernel Hilbert space (RKHS) of $K$. Define $\mathcal{H}_\Lambda = \{h \in \mathcal{H} : \|h\|_{\mathcal{H}} \leq \Lambda\}$, where $\|\cdot\|_{\mathcal{H}}$ is the RKHS norm. Using finite-dimensional kernel approximation combined with their algorithm for linear classifiers, Bassily et al. (2022) give an $(\varepsilon, \delta)$-DP algorithm that outputs a classifier $h^\text{Priv}$ such that with high probability over the input sample $S \sim \mathcal{D}^m$ and the algorithm’s randomness,

$$R_D(h^\text{Priv}) \leq \min_{h \in \mathcal{H}_\Lambda} \mathcal{L}_S^p(h) + O\left(\frac{\Lambda r}{\rho \sqrt{\min(1, \varepsilon)} m}\right). \quad (3)$$

Again, we note that the margin bound (3) nearly matches the standard, non-private analog. However, the computational cost of this algorithm is $O(m^d)$, 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 $\mathcal{H}_{NN}$ be a family of $L$-layer
feed-forward neural networks defined over $B^d(r)$. A function $h$ in $\mathcal{H}_{NN}$ can be viewed as a cascade
of linear maps composed with a non-linear activation function. We define $\mathcal{H}_{NN,\Lambda}$ as the subset
of $\mathcal{H}_{NN}$ with weight matrices that are $\Lambda$-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 $\varepsilon$-DP algorithm which returns a
neural network $h^{Priv} \in \mathcal{H}_{NN}$ such that with high probability over the input sample $S \sim \mathcal{D}^m$ and the
algorithm’s randomness,

$$R_D(h^{Priv}) \leq \min_{h \in \mathcal{H}_{NN,\Lambda}} \overline{R}^\rho_{S}(h) + O\left(\frac{r\Lambda^L \sqrt{NL}}{\rho \sqrt{m}} + \frac{r^2(2\Lambda)^{2L} NL}{\rho^2 \varepsilon m}\right)$$

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-
dependent 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 $(\varepsilon, \delta)$-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 $\mathcal{H}_{NN}$ 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 $\mathcal{H}_{NN}$ with no explicit dependence on
the network size? In particular, can we design a DP learning algorithm for $\mathcal{H}_{NN}$ with the following
learning guarantee?

$$R_D(h^{Priv}) \leq \min_{h \in \mathcal{H}_{NN,\Lambda}} \overline{R}^\rho_{S}(h) + O\left(\frac{r\Lambda^L}{\rho \sqrt{m}} + \frac{r^2(\Lambda)^{2L}}{\rho^2 \varepsilon m}\right).$$
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


