Privacy-preserving Prediction

Cynthia Dwork
Harvard University

Vitaly Feldman
Google Brain

Editors: Sebastien Bubeck, Vianney Perchet and Philippe Rigollet

Abstract

Ensuring differential privacy of models learned from sensitive user data is an important goal that has been studied extensively in recent years. It is now known that for some basic learning problems, especially those involving high-dimensional data, producing an accurate private model requires much more data than learning without privacy. At the same time, in many applications it is not necessary to expose the model itself. Instead users may be allowed to query the prediction model on their inputs only through an appropriate interface. Here we formulate the problem of ensuring privacy of individual predictions and investigate the overheads required to achieve it in several standard models of classification and regression.

We first describe a simple baseline approach based on training several models on disjoint subsets of data and using standard private aggregation techniques to predict. We show that this approach has nearly optimal sample complexity for (realizable) PAC learning of any class of Boolean functions. At the same time, without strong assumptions on the data distribution, the aggregation step introduces a substantial overhead. We demonstrate that this overhead can be avoided for the well-studied class of thresholds on a line and for a number of standard settings of convex regression. The analysis of our algorithm for learning thresholds relies crucially on strong generalization guarantees that we establish for all differentially private prediction algorithms.

1. Introduction and problem formulation

In machine learning tasks, the training data often consists of information collected from individuals. This data can be highly sensitive, for example in the case of medical or financial information, and therefore privacy-preserving data analysis is becoming an increasingly important area of study in machine learning, data mining and statistics (Dwork and Smith, 2009; Sarwate and Chaudhuri, 2013; Dwork and Roth, 2014). We rely on the well-studied differential privacy model of privacy that has become a de facto standard for formal understanding of privacy (Dwork et al., 2006).

The standard setting of privacy-preserving learning aims to ensure that the model learned from the data is produced in a differently private way. Thus this approach preserves privacy even when a potential adversary has complete access to the description of the predictive model. The downside of this strong guarantee is that for some learning problems, achieving the guarantee is known to have substantial additional costs. More examples are needed to achieve the same level of accuracy (or lower accuracy is achievable for a given number of examples). In addition, private learning may require new and computationally less efficient algorithms.

0. Extended abstract. Full version appears as (Dwork and Feldman, 2018, v2).

In this work we consider learning in a setting where the description of the learned model is not accessible to the (potentially adversarial) user(s). Instead the users have access to the model through an interface (often referred to as an API). For an input point the interface provides the value of the predictive model on that point. This view is appropriate for many existing applications where user privacy is a concern. For example, companies that collect data about their users usually expose only a cloud-based interface to the models they train on user data. Credit rating bureaus only allow access to their models through an electronic interface. In addition, it may enable new applications where privacy considerations are currently preventing the use of predictive models trained on sensitive user data. For example, in medical diagnostics a prediction interface would suffice for most applications.

Allowing such restricted access may appear to pose no risk to individual privacy. However, as recently demonstrated by Shokri et al. (2017), blackbox access to Amazon ML and Google prediction APIs suffice for successful membership inference attacks. Membership inference is the task in which given a user's record the goal is to infer whether the record was used for training the model. This information is known to be sensitive in several contexts. Membership inference can also be used to complete partial records revealing the values of sensitive attributes. Even more recently, Long et al. (2018) demonstrated several additional successful membership inference attacks based on blackbox access. Further, Carlini et al. (2018) proposed a more formal way to measure the degree to which sensitive information is memorized by generative sequence models and explored several techniques to extract sensitive information using black box access to such models. The use of differentially private learning algorithms to protect against such attacks has been proposed in (Shokri et al., 2017) and briefly explored in (Carlini et al., 2018).

We now describe the setting more formally. For a prediction problem over a domain X and label space Y, a prediction interface is an algorithm that has access to a dataset $S \in (X \times Y)^n$ and given a query point $x \in X$ outputs a value $y \in Y$. The algorithm can be queried multiple times and is stateful (namely, responses can depend on previous queries). We define the privacy of such an interface in the same way as usually done for interactive algorithms. Namely, for a prediction interface M and a stateful query generating algorithm Q we denote by $(Q \rightleftarrows M(S))$ the sequence of queries and responses generated in the interaction of Q and M on dataset S.

Definition 1.1 (Private prediction interface) *A prediction interface M, is (ϵ, δ)-differentially private if for every interactive query generating algorithm Q, the output $(Q \rightleftarrows M(S))$ is (ϵ, δ)-differentially private with respect to dataset S.*

While the problem setting has many facets that merit investigation, we focus on perhaps the most basic question: what is the cost of ensuring privacy of a single prediction. In other words, we focus on the problem of answering a single prediction query. Composition properties of differential privacy imply that such an algorithm can be used to answer multiple queries with privacy parameters that degrade gracefully with the number of queries (Dwork and Roth, 2014). Therefore such an algorithm is a natural building block for constructing an algorithm that can answer multiple queries. Naturally, better ways of dealing with sequences of queries might exist and the general topic of answering interactive sequences of queries has been studied extensively in the differential privacy literature (see (Dwork and Roth, 2014) for an overview).

An algorithm M that answers a single query x defines a randomized prediction at x and hence such an algorithm implicitly defines a learning algorithm that outputs a randomized predictor $h(x) = M(S, x)$.

Definition 1.2 *Let M be an algorithm that given a dataset $S \in (X \times Y)^n$ and a point x produces a value in Y. We say that M is* (ϵ, δ)-differentially private prediction *algorithm if for every $x \in X$, the output $M(S, x)$ is (ϵ, δ)-differentially private with respect to S. We use $M(S)$ to refer to the (randomized) function $M(S, \cdot)$.*

This definition allows us to treat this building block in the same way as regular learning algorithms and discuss it in the context of standard statistical learning models.

Two standard and closely related models for classification we will look at are PAC (or realizable) learning (Valiant, 1984) and agnostic (Haussler, 1992; Kearns et al., 1994) learning. In the PAC learning model the algorithm is given random examples in which each point is sampled i.i.d. from some unknown distribution over the domain and is labeled by an unknown function from a set of functions C. In the agnostic learning model the algorithm is given examples sampled i.i.d. from an arbitrary (and unknown) distribution over labeled points. The goal of the learning algorithm in both models is to output a hypothesis whose prediction error on the distribution from which examples are sampled is within additive α of the prediction error of the best function in C (which is 0 in the PAC model).

We will also consider a more general regression setting in which we are given a loss function $\ell : \mathbb{R} \times Y \to \mathbb{R}$ and the goal is to design a private prediction algorithm M that minimizes

$$\mathcal{E}_{\mathcal{P}}[\ell(M(S)] = \mathop{\mathbf{E}}_{M,(x,y)\sim\mathcal{P}}[\ell(M(S, x), y)],$$

where $\mathcal{P}$ is an unknown probability distribution over $X \times Y$.

2. Overview of the results

We first consider a natural "baseline" approach to this problem based on private aggregation of non-private learning algorithms.

2.1. Private aggregation of non-private models

To produce a prediction differentially privately we partition the dataset S into several subsamples $S_1, \ldots, S_r$ and run a non-private learning algorithm on each of those subsamples too obtain predictors $f_1, \ldots, f_r$. Now given a point x we use a differentially private aggregation technique on values $f_1(x), \ldots, f_r(x)$ and output the result. Several such subsample-and-aggregate techniques are known (Nissim et al., 2007; Dwork and Lei, 2009; Smith and Thakurta, 2013; Dwork and Roth, 2014) that carefully exploit properties of the distribution over results on subsamples. A significant advantage of this approach is that it does not require a new learning algorithm and hence is easy to implement (there is an additional computational cost that is easy to parallelize).

Obviously, using r subsamples requires more data than non-private learning and therefore it is natural to ask whether this approach is optimal and how it compares to differentially private learning in the standard setting. We discuss these questions in the context of specific problems below.

PAC Learning: For PAC learning (or realizable case) accurate models $f_1, \ldots, f_r$ have to be close to the true labeling function f (that is, they disagree with probability at most α). In particular, the fraction of points on which more than $1/4$ of the predictors output the wrong label cannot be more than 4α. Outputting the correct label with privacy is easy in this setting and we do this using a soft majority vote (or, equivalently, the exponential mechanism (McSherry and Talwar, 2007) on the

label counts). A number of other approaches would give comparable guarantees. A simple analysis shows that using $r = O(\ln(1/\alpha)/\epsilon)$ this reduction ensures ϵ-differentially private prediction for a formal statement).

As an immediate corollary of this reduction and standard bounds on the sample complexity of PAC learning we obtain the following upper bound.

Corollary 2.1 *Let C be a class of Boolean functions of VC dimension d. Then for all $\alpha, \beta, \epsilon > 0$, there exists an ϵ-differentially private prediction algorithm M that PAC learns C with error α and confidence $1 - \beta$ given $n = \tilde{O}\left(\frac{d+\log(1/\beta)}{\epsilon\alpha}\right)$ examples.*

It turns out that this simple approach is essentially optimal in the worst case. Specifically, we prove that the sample complexity of this problem is $\Omega(d/(\epsilon\alpha))$ even when δ is as large as $\epsilon/3$.

Theorem 2.2 *Let C be a class of Boolean functions of VC dimension d. Then for all $\alpha, \epsilon > 0$, any $(\epsilon, \epsilon/3)$-differentially private prediction algorithm M that PAC learns C with error α and confidence $1/12$ requires $n = \Omega(d/(\epsilon\alpha))$ examples.*

For comparison, Kasiviswanathan et al. (2011) showed that the sample complexity of differentially privately PAC learning a class C over domain X is $O(\log(|C|)/(\epsilon\alpha))$. By Sauer's lemma, $\log(|C|) = O(d \cdot \log(|X|))$ and therefore the multiplicative gap between these two measures can be as large as $\log(|X|)$. The sample complexity of ϵ-differentially private PAC learning was subsequently shown to be $\tilde{\Theta}(R/(\epsilon\alpha))$, where R is the so-called representation dimension of C (Beimel et al., 2013). However, as shown in (Feldman and Xiao, 2015), for many classes the gap between R and the VC dimension is still roughly $\log(|X|)$. For example, the representation dimension of linear threshold functions over $[N]^p$ is $p^2 \cdot \log N$ whereas the VC dimension is just p.

We remark that the technique we use to prove the lower bound in Thm. 2.2 is different from those used for proving lower bounds in the standard setting of learning with privacy.

Agnostic learning: In agnostic learning, the labels $f_1(x), \ldots, f_r(x)$ no do not necessarily agree on most points x and taking the majority vote may even reduce the accuracy. In this setting we predict by first averaging the non-private predictions to obtain $v(x) = \frac{1}{r}(f_1(x) + \cdots + f_r(x))$ and then outputting 1 with probability $v(x) + \zeta$ (truncated to range $[0, 1]$), where ζ is a Laplace noise variable. It is not hard to show that for $r = O(1/(\epsilon\alpha))$, this approach ensures that the prediction will be ϵ-differentially private and the addition of noise increases the prediction error by at most an extra α term. As a corollary of this reduction, we obtain the following upper-bound on the sample complexity in this setting.

Corollary 2.3 *Let C be a class of Boolean functions of VC dimension d. Then for all $\alpha, \beta, \epsilon > 0$ there exists an ϵ-differentially private prediction algorithm M that agnostically learns C with excess error α and confidence $1 - \beta$ given $n = \tilde{O}\left(\frac{d+\log(1/\beta)}{\epsilon\alpha^3}\right)$ examples.*

In this case the upper bound is much worse than the lower bound of $\Omega(d/\alpha^2 + d/(\epsilon\alpha))$ implied by Thm. 2.2. For comparison, ϵ-differentially private agnostic learning can be done using $\tilde{O}(d/\alpha^2 + R/(\epsilon\alpha))$ examples, where R is the representation dimension of C mentioned above (Beimel et al., 2013; Feldman and Xiao, 2015). As a result, for classes such that $R = O(d)$ a differentially private learning algorithm matches the lower bound for private prediction. This leads to a natural question of whether it is possible to match the lower bound for all classes C. While we do not answer this

question for arbitrary classes C, we give an example of an algorithm that goes beyond these two approaches. Specifically, it agnostically learns C with ϵ-private prediction using $\tilde{O}(d/\alpha^2 + d/(\epsilon\alpha))$ examples whereas learning C with privacy in the standard model requires an infinite number of examples.

Convex regression: Our analysis of agnostic learning can be seen as a special case of a more general analysis of prediction problems with convex loss functions. Specifically, the aggregation by averaging can be seen as a way to increase the *uniform prediction stability* of a learning algorithm. A learning algorithm is uniformly prediction stable with rate γ if for predictors f_S and $f_{S'}$ produced on any pair of datasets S, S' that differ on a single element and any point x, $|f_S(x) - f_{S'}(x)| \leq \gamma$. As follows immediately from this definition, a uniformly prediction stable learning algorithm can be converted to a differentially private prediction algorithm simply by adding Laplace (or Gaussian) noise to the prediction. Hence it reduces our problem to the problem of finding a uniformly prediction stable learning algorithm with sufficiently low rate of stability. Aggregation by averaging the predictors obtained by running a learning algorithm on r disjoint datasets can be seen as improving its uniform prediction stability by a factor of r. Convexity of the loss function, in turn, ensures that such averaging preserves the guarantees on the expected loss of the algorithm.

We demonstrate how this general approach can be applied to convex regression problems. Specifically, we consider problems in which we have a family of predictors $\{f(w, \cdot)\}_{w \in \mathcal{K}}$ parameterized by a vector $w \in \mathcal{K}$, where $\mathcal{K} \subset \mathbb{R}^d$ is some convex body, ℓ is a convex loss function and $\ell(f(\cdot, x), y)$ is a convex function of w over $\mathcal{K}$ for all $(x, y) \in X \times Y$. The goal is to find $\hat{w}$ such that

$$\mathop{\mathbf{E}}_{(x,y)\sim\mathcal{P}}[\ell(f(\hat{w}, x), y)] \leq \min_{w \in \mathcal{K}} \mathop{\mathbf{E}}_{(x,y)\sim\mathcal{P}}[\ell(f(w, x), y)] + \alpha,$$

where $\mathcal{P}$ is an unknown distribution over examples. This setting captures many important learning problems and has also been extensively investigated in the privacy literature (see (Chaudhuri et al., 2011; Kifer et al., 2012; Bassily et al., 2014; Talwar et al., 2015; Wang et al., 2017) and references therein). For the purpose of comparison with sample complexity bounds known in this literature we restrict our attention to a basic setting in which $\mathcal{K}$ is a subset of the unit Euclidean ball and $\ell(f(w, x), y)$ is 1-Lipschitz in w for all (x, y) in support of $\mathcal{P}$. For this setting it is known that $\tilde{O}(d/(\epsilon\alpha^2))$ samples suffice to solve the problem with ϵ-differential privacy and $\tilde{O}(\sqrt{d}\log^4(1/\delta)/(\epsilon\alpha^2))$ samples suffice for (ϵ, δ)-differential privacy (Bassily et al., 2014). Further, such dependence on the dimension is optimal in both settings (Bassily et al., 2014).

The dependence on the dimension is not necessary for non-private learning in this setting. In addition, we can exploit known stability analyses to reduce (or even eliminate) the need to use the aggregation step. By plugging the known stability results based on strong convexity and/or Bousquet and Elisseeff (2002); Shalev-Shwartz et al. (2010); Hardt et al. (2016), we demonstrate that convex regression problems of this type can be solved with ϵ-differentially private prediction using $O(1/(\epsilon\alpha^2))$ examples. We also demonstrate that smoothness of the loss function ℓ can be used to improve the dependence on ϵ. We note that stability of the optimal solution of a strongly convex problem has been used to achieve differential privacy in multiple prior works starting with the pioneering work of Chaudhuri et al. (2011). Stability of gradient descent on convex smooth functions has also been recently used to obtain privacy guarantees (Wu et al., 2017).

2.2. Beyond aggregation: learning thresholds

The class of linear thresholds Thr is defined over a subset of reals and consists of indicator functions of "$x \geq a$" for all $a \in \mathbb{R}$. Without loss of generality, we consider such functions over the set $[N] = \{1, \ldots, N\}$. While the class is very simple, learning it with privacy has proved to be rather challenging and some basic questions are still not fully resolved (Beimel et al., 2010; Chaudhuri and Hsu, 2011; Beimel et al., 2013; Feldman and Xiao, 2015; Bun et al., 2015). It is known that ϵ-differentially private PAC learning of Thr requires $\Omega(\log(N)/(\epsilon\alpha))$ examples (Feldman and Xiao, 2015) and *proper* (ϵ, δ) differentially private PAC learning requires $\Omega(\log^*(N)/(\epsilon\alpha))$ examples (Bun et al., 2015) (no lower bounds for non-proper learning and $\delta > 0$ case are known). Note that the VC dimension of this class is just 1.

We give an ϵ-differentially private prediction algorithm for agnostic learning of this class with the following guarantee:

Theorem 2.4 *For any $\alpha, \epsilon > (0, 1]$ and $N \in \mathbb{N}$, there exists an ϵ-differentially private prediction algorithm M that given $n \geq \frac{12\ln(2/\alpha)}{\alpha\epsilon}$ examples from an arbitrary distribution $\mathcal{P}$ over $[N] \times \{0, 1\}$ guarantees:*

$$\mathop{\mathbf{E}}_{S \sim \mathcal{P}^n}[\mathsf{Err}_{\mathcal{P}}(M(S))] \leq e^{\epsilon} \cdot (\mathsf{Opt}_{\mathcal{P}}(\mathsf{Thr}) + \alpha).$$

Note that this statement implies an upper bound of $n = O(\ln(1/\alpha)/(\alpha\epsilon))$ in the realizable case when $\mathsf{Opt}_{\mathcal{P}}(\mathsf{Thr}) = 0$ and also an upper bound of $n = O(\ln(1/\alpha)/(\alpha\epsilon) + \ln(1/\alpha)/\alpha^2)$ in the agnostic setting. The $\tilde{O}(1/\alpha^2)$ term arises from having to set $\epsilon < \alpha$ to ensure that the expected error is at most $\mathsf{Opt}_{\mathcal{P}}(\mathsf{Thr}) + O(\alpha)$. Our algorithm can also handle unions of k intervals (at the expense of an additional factor k in the sample complexity).

At a high level our algorithm works as follows. First, the examples are sorted. To determine the probability with which to output 1 on point x the algorithm traverses the examples on points smaller than x in increasing order. Starting from bias $1/2$ the algorithm increases or decreases the current bias by a factor of (roughly) e^{ϵ} for each example it traverses. The bias is increased if the label of the example is 1 and decreased otherwise. Importantly, the bias is projected back to the interval $[\alpha, 1 - \alpha]$ after each update. The algorithm outputs 1 with probability obtained at the end of this process. While the prediction privacy of our algorithm is relatively easy to establish, the analysis of its error is more delicate and we are not aware of similar algorithms having been proposed for this problem. Furthermore, our analysis only bounds the empirical error of this algorithm. The hypothesis produced by the algorithm is sufficiently complicated that it would not be possible to ensure generalization using VC dimension or similar techniques. Remarkably, the fact that our algorithm is prediction private allows us to prove that it generalizes.

2.3. Generalization

It has been known for a while that differential privacy is a notion of stability and hence implies bounds on the expectation of generalization error. Recent work in the context of adaptive data analysis has substantially strengthened this connection, proving that differential privacy ensures generalization with high probability (Dwork et al., 2014; Bassily et al., 2016; Feldman and Steinke, 2017). Prediction privacy can also be seen as a notion of stability that is weaker than differential privacy but stronger than uniform prediction stability. We show how to derive relatively strong generalization guarantees from this notion of stability. These guarantees are stronger than those

known for classical notions of stability (*e.g.* (Bousquet and Elisseeff, 2002; Shalev-Shwartz et al., 2010)) but not as strong as those proved for differential privacy. Specifically, our generalization results imply that for every non-negative loss function ℓ, a moment $k \geq 1$, and an ϵ-differentially private prediction algorithm M:

$$\mathop{\mathbf{E}}_{S,S' \sim \mathcal{P}^n,} \left[\left(\mathcal{E}_{S'}[\ell(M(S))] \right)^k \right] \leq e^{k^2 \epsilon} \cdot \mathop{\mathbf{E}}_{S \sim \mathcal{P}^n} \left[\left(\mathcal{E}_{S}[\ell(M(S))] \right)^k \right],$$

where $\mathcal{E}_S[\ell(M(S)]$ denotes the expected empirical loss of $M(S)$ on S. Note that on the left hand side we are bounding the average loss on an independently drawn set of examples S' which is tightly concentrated around the expected loss $\mathcal{E}_{\mathcal{P}}[\ell(M(S)]$. For comparison, ϵ-differential privacy gives a similar bound with $e^{k\epsilon}$ factor instead of $e^{k^2\epsilon}$ (Dwork et al., 2014). The bound above is stated using the $k = 1$ version of this result. However this generalization bound implies that loss is also well concentrated. We give an example of how to derive high probability bounds on the generalization error from this moment bound.

2.4. Related work

Pathak et al. (2010) consider secure and differentially private aggregation of non-private linear models held by multiple mistrusting parties. They achieve it by computing the average model and adding noise to it. They do not consider accuracy guarantees of their approach formally.

To the best of our knowledge, the privacy-preserving aggregation of non-private predictions to produce privacy-preserving predictions was first investigated by Bilenko, Dwork, Muthukrishnan, Rothblum, Thakurta and Wang in 2014[1]. Bilenko *et al.* , obtained high levels of composition by exploiting the frequently high degree of (near) consensus among the predictions of the non-private models via a variant of the sparse-vector technique Dwork and Roth (2014). Our work shares the same goal of generating differentially private predictions. At the same time we formalize the general problem of learning with differentially private predictions and focus on the sample complexity of making a single prediction. In addition, we demonstrate approaches that go beyond privacy-preserving aggregation.

Aggregation of non-private models to produce labels while preserving privacy was also used in recent works of Hamm et al. (2016) and, subsequently, Papernot et al. (2017, 2018) to give a new semi-supervised approach to differentially private learning. Specifically, their approach is predicated on availability of public *unlabeled* dataset Z. The dataset Z is labeled using differentially-private aggregation of labels provided by models trained on the sensitive dataset S. The labeled data is used to train a new model. Since differential privacy is closed under post-processing, this new model is privacy-preserving for S (but not for Z). The works of Papernot et al. (Papernot et al., 2017, 2018) deal primarily with techniques for accurately bounding the privacy parameters while ensuring accurate prediction on benchmark datasets. Hamm et al. (2016) also formally examine additional error that noisy aggregation introduces and explicitly rely on stability of strongly-convex regression problems to provide formal guarantees for their approach. Their framework and the guarantees are incomparable to ours, and, in particular, they do not avoid dependence on the dimension.

In a recent and independent work, Bassily et al. (2018) consider the formal guarantees for answering a sequence of prediction queries using differentially private aggregation techniques. They

1. This was the core of a larger project on privacy-preserving click prediction that did not survive the closing of the Silicon Valley lab.

demonstrate that given a non-private learning algorithm has error of at most α (such as in the PAC model), there exists an algorithm that answers m prediction queries for points chosen i.i.d. from the same distribution with error $O(\alpha)$ and privacy parameter ϵ scaling as $\sqrt{m\alpha} \cdot \log m$ (for comparison, a direct application of composition theorems for differential privacy implies $\sqrt{m}$ scaling for an arbitrary sequence of queries). They then analyze the sample complexity of semi-supervised (or, equivalently, label-private) learning algorithm that is obtained by labeling a public unlabeled dataset using their algorithm for answering prediction queries.

We remark that all these works do not examine the problem of private prediction itself and focus on the aggregation-based approaches. Recall that in private prediction, it is the privacy of the training data for the predictor (model) that is being protected.

References

R. Bassily, A. Smith, and A. Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *FOCS*, pages 464–473, 2014.

Raef Bassily, Kobbi Nissim, Adam D. Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *STOC*, pages 1046–1059, 2016.

Raef Bassily, Om Thakkar, and Abhradeep Thakurta. Model-agnostic private learning via stability. *CoRR*, abs/1803.05101, 2018. URL http://arxiv.org/abs/1803.05101.

Amos Beimel, Shiva Prasad Kasiviswanathan, and Kobbi Nissim. Bounds on the sample complexity for private learning and private data release. In *TCC*, pages 437–454, 2010.

Amos Beimel, Kobbi Nissim, and Uri Stemmer. Characterizing the sample complexity of private learners. In *ITCS*, pages 97–110, 2013.

Olivier Bousquet and André Elisseeff. Stability and generalization. *JMLR*, 2:499–526, 2002.

Mark Bun, Kobbi Nissim, Uri Stemmer, and Salil P. Vadhan. Differentially private release and learning of threshold functions. In *FOCS*, pages 634–649, 2015.

Nicholas Carlini, Chang Liu, Jernej Kos, Úlfar Erlingsson, and Dawn Song. The secret sharer: Measuring unintended neural network memorization & extracting secrets. *CoRR*, abs/1802.08232, 2018. URL http://arxiv.org/abs/1802.08232.

Kamalika Chaudhuri and Daniel Hsu. Sample complexity bounds for differentially private learning. In *COLT*, pages 155–186, 2011.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12:1069–1109, 2011.

C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, pages 265–284, 2006.

Cynthia Dwork and Vitaly Feldman. Privacy-preserving prediction. *CoRR*, abs/1803.10266, 2018. URL http://arxiv.org/abs/1803.10266. Extended abstract in COLT 2018.

Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *STOC*, pages 371–380, 2009.

Cynthia Dwork and Aaron Roth. *The Algorithmic Foundations of Differential Privacy*, volume 9. 2014. URL http://dx.doi.org/10.1561/0400000042.

Cynthia Dwork and Adam Smith. Differential privacy for statistics: What we know and what we want to learn. *Journal of Privacy and Confidentiality*, 1(2):135–154, 2009.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. Preserving statistical validity in adaptive data analysis. *CoRR*, abs/1411.2664, 2014. Extended abstract in STOC 2015.

Vitaly Feldman and Thomas Steinke. Generalization for adaptively-chosen estimators via stable median. In *Conference on Learning Theory (COLT)*, 2017.

Vitaly Feldman and David Xiao. Sample complexity bounds on differentially private learning via communication complexity. *SIAM J. Comput.*, 44(6):1740–1764, 2015.

Jihun Hamm, Yingjun Cao, and Mikhail Belkin. Learning privately from multiparty data. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *ICML*, volume 48 of *Proceedings of Machine Learning Research*, pages 555–563, 2016. URL http://proceedings.mlr.press/v48/hamm16.html.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *ICML*, pages 1225–1234, 2016. URL http://jmlr.org/proceedings/papers/v48/hardt16.html.

D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100(1):78–150, 1992. ISSN 0890-5401.

Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM J. Comput.*, 40(3):793–826, June 2011.

M. Kearns, R. Schapire, and L. Sellie. Toward efficient agnostic learning. *Machine Learning*, 17 (2-3):115–141, 1994.

Daniel Kifer, Adam D. Smith, and Abhradeep Thakurta. Private convex optimization for empirical risk minimization with applications to high-dimensional regression. In *COLT*, pages 25.1–25.40, 2012. URL http://www.jmlr.org/proceedings/papers/v23/kifer12/kifer12.pdf.

Yunhui Long, Vincent Bindschaedler, Lei Wang, Diyue Bu, Xiaofeng Wang, Haixu Tang, Carl A. Gunter, and Kai Chen. Understanding membership inferences on well-generalized learning models. *CoRR*, abs/1802.04889, 2018. URL http://arxiv.org/abs/1802.04889.

Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *FOCS*, pages 94–103, 2007.

Kobbi Nissim, Sofya Raskhodnikova, and Adam D. Smith. Smooth sensitivity and sampling in private data analysis. In *STOC*, pages 75–84, 2007.

Nicolas Papernot, Martín Abadi, Úlfar Erlingsson, Ian J. Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, 2017.

Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Ulfar Erlingsson. Scalable private learning with PATE. *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rkZB1XbRZ.

Manas A. Pathak, Shantanu Rane, and Bhiksha Raj. Multiparty differential privacy via aggregation of locally trained classifiers. In *NIPS*, pages 1876–1884, 2010.

Anand D. Sarwate and Kamalika Chaudhuri. Signal processing and machine learning with differential privacy: Algorithms and challenges for continuous data. *IEEE Signal Process. Mag.*, 30 (5):86–94, 2013.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 11:2635–2670, 2010.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy, SP 2017*, pages 3–18, 2017.

Adam Smith and Abhradeep Guha Thakurta. Differentially private feature selection via stability arguments, and the robustness of the lasso. In *COLT*, 2013.

Kunal Talwar, Abhradeep Thakurta, and Li Zhang. Nearly optimal private LASSO. In *NIPS*, pages 3025–3033, 2015. URL http://papers.nips.cc/paper/5729-nearly-optimal-private-lasso.

L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

Di Wang, Minwei Ye, and Jinhui Xu. Differentially private empirical risk minimization revisited: Faster and more general. In *NIPS*, pages 2719–2728, 2017.

Xi Wu, Fengan Li, Arun Kumar, Kamalika Chaudhuri, Somesh Jha, and Jeffrey F. Naughton. Bolt-on differential privacy for scalable stochastic gradient descent-based analytics. In *SIGMOD*, pages 1307–1322, 2017.