# Learning with Differential Privacy: Stability, Learnability and the Sufficiency and Necessity of ERM Principle

**Yu-Xiang Wang**
Machine Learning Department
Carnegie Mellon University
yuxiangw@cs.cmu.edu

**Jing Lei**
Department of Statistics
Carnegie Mellon University
jinglei@andrew.cmu.edu

**Stephen E. Fienberg**
Department of Statistics
Carnegie Mellon University
fienberg@stat.cmu.edu

## Abstract

While machine learning has proven to be a powerful data-driven solution to many real-life problems, its use in sensitive domains that involve human subjects has been limited due to privacy concerns. The cryptographic approach known as "differential privacy" offers provable privacy guarantees. In this paper we study the learnability under Vapnik's general learning setting with differential privacy constraint, and reveal some intricate relationships between privacy, stability and learnability.

In particular, we show that a problem is privately learnable *if an only if* there is a private algorithm that asymptotically minimizes the empirical risk (AERM). This is rather surprising because for non-private learning, AERM alone is not sufficient for learnability. This result suggests that when searching for private learning algorithms, we can restrict the search to algorithms that are AERM. In light of this, we propose a conceptual procedure that always finds a universally consistent algorithm whenever the problem is learnable under privacy constraint. We also propose a generic and practical algorithm and show that under very general conditions it privately learns a wide class of learning problems.

## 1 Introduction

A major challenge in developing privacy-preserving learning methods is to quantify formally the amount of privacy leakage, given all possible and unknown auxiliary information the attacker may have, a challenge in part addressed by the notion of *differential privacy* [8, 11]. Differential privacy has three main advantages over other approaches: (1) it rigorously quantifies the privacy property of any data analysis mechanism; (2) it also guarantees the amount of privacy leakage regardless of the attacker's resource or knowledge, (3) it has useful interpretations from the perspectives of Bayesian inference and statistical hypothesis testing, and hence fits naturally in the general framework of statistical machine learning, e.g., see [10, 29, 22, 17, 27], as well as applications involving regression [7, 23] and GWAS data [30], etc.

In this paper we focus on the following fundamental question about differential privacy and machine learning: *What problems can we learn with differential privacy?* Most literature focuses on designing differentially private learning algorithms in a case by case fashion, where the methods depend crucially on the specific context and differ vastly in nature. But with the privacy constraint, we have less choice in developing learning and data analysis algorithms. It remains unclear how such a constraint affects our ability of learning, and if it is possible to design a generic privacy-preserving analysis mechanism that is applicable to a wide class of learning problems.

**Our Contributions** We provide a general answer to the relationship between learnability and differential privacy under Vapnik's General Learning Setting [26] in three aspects.

1. We characterize the subset of problems in the General Learning Setting that can be learned under differential privacy. Specifically, we show that a sufficient and necessary condition for a problem to be privately learnable is the existence of an algorithm that is differentially private and asymptotically minimizes the empirical risk. This characterization generalizes previous studies of the subject that focus on binary classification in the PAC learning setting. Our characterization reveals insight on the relationship between differential privacy and a variant of algorithmic stability that is shown to be necessary for learnability [21]. We show that privacy by definition implies stability.

2. We also introduce a weaker notion of learnability, which only requires consistency for a class of distributions $\mathfrak{D}$. Problems that are not privately learnable (a surprisingly large class that includes simple problems such as 0-1 loss binary classification in continuous feature domain [6]) are usually private $\mathfrak{D}$-learnable for some "nice" distribution class $\mathfrak{D}$. We characterize $\mathfrak{D}$-learnable problems using conditions analogous to those in distribution-free private learning.

3. Inspired by the equivalence between privacy learnability and private AERM, we propose a generic (but impractical) procedure that characterizes a consistent and private algorithm for any privately learnable (or $\mathfrak{D}$-learnable) problems. We also propose a generic of the algorithm that aims at minimizing the empirical risk while preserving the privacy. We show that under a sufficient condition that relies on the geometry of the hypothesis space and the data distribution, this algorithm is able to privately learn (or $\mathfrak{D}$-learn) a large range of learning problems including classification, regression, clustering, density estimation and etc, and it is computationally efficient when the problem is convex. In fact, this generic learning algorithm learns any privately learnable problems in the PAC learning setting [4]. It remains an open problem whether the second algorithm also learns any privately learnable problem in the General Learning Setting.

Our primary objective is to understand the conceptual impact of differential privacy and learnability under a general framework and the rates of convergence obtained in the analysis may be suboptimal. Although we do provide some discussion on polynomial time approximations to the proposed algorithm, learnability under computational constraints is beyond the scope of this paper.

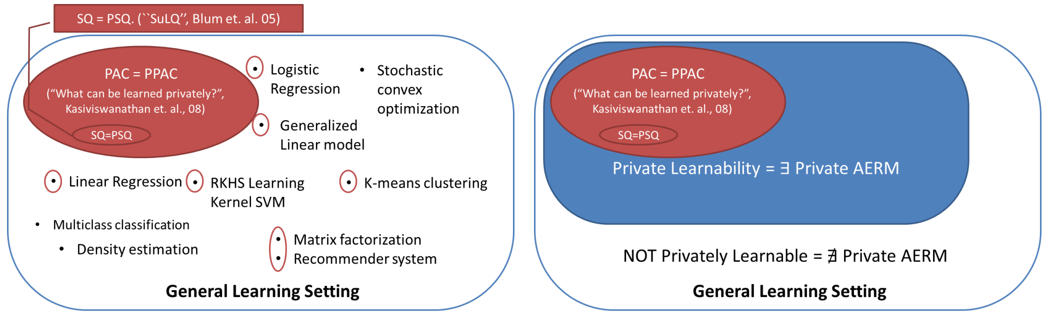
Due to space limit of this extended abstract, we will present only the characterization of private learnability. We invite readers to check out the remaining results and examples in our full paper [28].

**Related work** Private learnability has been studied only in a few special cases, e.g., Kasiviswanathan et al. [13], Chaudhuri & Hsu [6], Beimel et al. [4], which mainly concerns PAC learning (supervised learning, sometimes only for discrete problems with finite hypothesis space). A key difference of our work from previous works is that we consider a more general setting and provide a proper treatment in a statistical learning framework. This allows us to cover many more important learning problems (see Figure 1(a) and Table 1).

Our characterization of private learnability (and private $\mathfrak{D}$-learnability) extends the same characterization in (non-private) learnability given by Shalev-Shwartz et al. [21]. We also borrow ideas from McSherry & Talwar [18], Chaudhuri & Hsu [6], Beimel et al. [4] in other parts of our technical results.

Our claim that "privacy implies stability" and "privacy implies generalization" are known as folklores for some time in the differential privacy community [1]. Our result formalizes the folklore, and provides substantial new understanding in that, we placed differential privacy formally in a very general learning theoretic framework and provided not only sufficient but also necessary condition of learnability under privacy constraints. An different argument that leads to the same formalization of this folklore appeared in Dwork et al. [9], but our proof is simpler and leads to slightly better rate of convergence. We refer readers to our full paper for an more extensive survey of related work.

[1] For instance, Moritz Hardt discussed the connection of differential privacy to stability and generalization in his blog post http://blog.mrtz.org/2014/01/13/false-discovery. Kunar Talwar described in his blog an example of exploiting differential privacy for measure concentration http://windowsontheory.org/2014/02/04/differential-privacy-for-measure-concentration/. Also, a similar claim was made in Appendix F of Bassily et al. [2] but no proofs or other form of justifications were provided.


SQ = PSQ. (``SuLQ'', Blum et. al. 05)
PAC = PPAC
("What can be learned privately?", Kasiviswanathan et. al., 08)
SQ=PSQ
Logistic Regression
Stochastic convex optimization
Generalized Linear model
Linear Regression
RKHS Learning Kernel SVM
K-means clustering
Multiclass classification
Density estimation
Matrix factorization
Recommender system
General Learning Setting
PAC = PPAC
("What can be learned privately?", Kasiviswanathan et. al., 08)
SQ=PSQ
Private Learnability = ∃ Private AERM
NOT Privately Learnable = ∄ Private AERM
General Learning Setting

(a) Illustration of general learning setting. Examples of known DP extensions are circled in **maroon**.

(b) Our characterization of private learnable problems in the general learning setting (in **blue**).

Figure 1: The Big Picture: illustration of general learning setting and our contribution in understanding differentially private learnability.

## 2 Background

### 2.1 Learnability under the General Learning Setting

In the General Learning Setting of Vapnik [26], a learning problem is characterized by a triplet $(\mathcal{Z}, \mathcal{H}, \ell)$. Here $\mathcal{Z}$ is the sample space (with a $\sigma$-algebra). The hypothesis space $\mathcal{H}$ is a collection of models such that each $h \in \mathcal{H}$ describes some structures of the data. The loss function $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$ measures how well the hypothesis $h$ explains the data instance $z \in \mathcal{Z}$. For example, in supervised learning problems $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ where $\mathcal{X}$ is the feature space and $\mathcal{Y}$ is the label space; $\mathcal{H}$ defines a collection of mapping $h : \mathcal{X} \to \mathcal{Y}$; and $\ell(h, z)$ measures how well $h$ predicts the feature-label relationship $z = (x, y) \in \mathcal{Z}$. This setting includes problems with continuous input/output in potentially infinite dimensional spaces (e.g. RKHS methods), hence is much more general than PAC learning. In addition, the general learning setting also covers a variety of unsupervised learning problems, including clustering, density estimation, principal component analysis (PCA) and variants ( e.g., Sparse PCA, Robust PCA), dictionary learning, matrix factorization and even Latent Dirichlet Allocation (LDA). Details of these examples are given in Table 1 (the first few are extracted from Shalev-Shwartz et al. [21]).

To account for the randomness in the data, we are primarily interested in the case where the data $Z = \{z_1, ..., z_n\} \in \mathcal{Z}^n$ are independent samples drawn from an unknown probability distribution $\mathcal{D}$ on $\mathcal{Z}$. We denote such a random sample by $Z \sim \mathcal{D}^n$. For a given distribution $\mathcal{D}$, let $R(h)$ be the expected loss of hypothesis $h$ and $\hat{R}(h, Z)$ the empirical risk from a sample $Z \in \mathcal{Z}^n$:

$$R(h) = \mathbb{E}_{z \sim \mathcal{D}} \ell(h, z), \qquad \hat{R}(h, Z) = \frac{1}{n} \sum_{i=1}^{n} \ell(h, z_i) .$$

The optimal risk $R^* = \inf_{h \in \mathcal{H}} R(h)$ and we assume that it is achieved by an optimal $h^* \in \mathcal{H}$. Similarly, the minimal empirical risk $\hat{R}^*(Z) = \inf_{h \in \mathcal{H}} \hat{R}(h, Z)$ is achieved by $\hat{h}^*(Z) \in \mathcal{H}$. For a possibly randomized algorithm $\mathcal{A} : \mathcal{Z}^n \to \mathcal{H}$ that learns some hypothesis $\mathcal{A}(Z) \in \mathcal{H}$ given data sample $Z$, we say $\mathcal{A}$ is *consistent* if

$$\lim_{n \to \infty} \mathbb{E}_{Z \sim \mathcal{D}^n} \left( \mathbb{E}_{h \sim \mathcal{A}(Z)} R(h) - R^* \right) = 0. \tag{1}$$

In addition, we say $\mathcal{A}$ is consistent with rate $\xi(n)$ if

$$\mathbb{E}_{Z \sim \mathcal{D}^n} \left( \mathbb{E}_{h \sim \mathcal{A}(Z)} R(h) - R^* \right) \le \xi(n), \text{ where } \lim_{n \to \infty} \xi(n) \to 0. \tag{2}$$

Since the distribution $\mathcal{D}$ is unknown, we cannot adapt the algorithm $\mathcal{A}$ to $\mathcal{D}$, especially when privacy is a concern. Also, even if $\mathcal{A}$ is pointwise consistent for any distribution $\mathcal{D}$, it may have different rates for different $\mathcal{D}$ and potentially be arbitrarily slow for some $\mathcal{D}$. This makes it hard to evaluate whether $\mathcal{A}$ indeed learns the learning problem and forbids the study of the learnability problem. In this study, we adopt the stronger notion of learnability considered in Shalev-Shwartz et al. [21], which is a direct

| Problem | Hypothesis class $\mathcal{H}$ | $\mathcal{Z}$ or $\mathcal{X} \times \mathcal{Y}$ | Loss function $\ell$ |
|---|---|---|---|
| PAC Learning | $\mathcal{H} \subset \{f : \{0,1\}^d \to \{0,1\}\}$ | $\{0,1\}^d \times \{0,1\}$ | $1(h(x) \neq y)$ |
| Regression | $\mathcal{H} \subset \{f : [0,1]^d \to \mathbb{R}\}$ | $[0,1]^d \times \mathbb{R}$ | $\lvert h(x) - y\rvert^2$ |
| Density Estimation | Bounded distributions on $\mathcal{Z}$ | $\mathcal{Z} \subset \mathbb{R}^d$ | $-\log(h(z))$ |
| K-means Clustering | $\{S \subset \mathbb{R}^d : \lvert S\rvert = k\}$ | $\mathcal{Z} \subset \mathbb{R}^d$ | $\min\limits_{c \in h} \lVert c - z\rVert^2$ |
| RKHS classification | Bounded RKHS | RKHS$\times\{0,1\}$ | $\max\{0, 1 - y\langle x, h\rangle\}$ |
| RKHS regression | Bounded RKHS | RKHS$\times\mathbb{R}$ | $\lvert\langle x, h\rangle - y\rvert^2$ |
| Sparse PCA | Rank-$r$ projection matrices | $\mathbb{R}^d$ | $\lVert hz - z\rVert^2 + \lambda\lVert h\rVert_1$ |
| Robust PCA | All subspaces in $\mathbb{R}^d$ | $\mathbb{R}^d$ | $\lVert\mathcal{P}_h(z) - z\rVert_1 + \lambda\text{rank}(h)$ |
| Matrix Completion | All subspaces in $\mathbb{R}^d$ | $\mathbb{R}^d \times \{1,0\}^d$ | $\min\limits_{b \in h} \lVert y \circ (b - x)\rVert^2 + \lambda\text{rank}(h)$ |
| Dictionary Learning | All dictionaries $\in \mathbb{R}^{d\times r}$ | $\mathbb{R}^d$ | $\min\limits_{b \in \mathbb{R}^r} \lVert hb - z\rVert^2 + \lambda\lVert b\rVert_1$ |
| Non-negative MF | All dictionaries $\in \mathbb{R}_+^{d\times r}$ | $\mathbb{R}^d$ | $\min\limits_{b \in \mathbb{R}_+^r} \lVert hb - z\rVert^2$ |
| Subspace Clustering | A set of $k$ rank-$r$ subspaces | $\mathbb{R}^d$ | $\min\limits_{b \in h} \lVert\mathcal{P}_b(z) - z\rVert^2$ |
| Topic models (LDA) | $\{\mathbb{P}(\text{word}\mid\text{topic})\}$ | Documents | $-\max\limits_{b \in \{\mathbb{P}(\text{Topic})\}} \sum\limits_{w \in z} \log \mathbb{P}_{b,h}(w)$ |

Table 1: An illustration of problems in the General Learning setting.

generalization of PAC-learnability [24] and agnostic PAC-learnability [15] to the General Learning Setting as studied by Haussler [12].

**Definition 1** (Learnability [21])**.** *A learning problem $(\mathcal{Z}, \mathcal{H}, \ell)$ is learnable if there exists an algorithm $\mathcal{A}$ and rate $\xi(n)$, such that $\mathcal{A}$ is consistent with rate $\xi(n)$ for any distribution $\mathcal{D}$ defined on $\mathcal{Z}$.*

This definition requires consistency to hold universally for any distribution $\mathcal{D}$ with a distribution-independent rate $\xi(n)$. This type of problem is often called *distribution-free learning* [24], and an algorithm is said to be *universally consistent* with rate $\xi(n)$ if it realizes the criterion.

### 2.2 Differential privacy

Differential privacy requires that if we arbitrarily perturb a database by only one data point, the output should not differ much. Therefore, if one conducts a statistical test for whether any individual is in the database or not, the false positive and false negative probabilities cannot both be small [29]. Formally, define "edit distance" (also called "Hamming distance")

$$d(Z, Z') := \#\{i = 1, ..., n : z_i \neq z'_i\}\,. \tag{3}$$

**Definition 2** ($\epsilon$-Differential Privacy [8])**.** *We say an algorithm $\mathcal{A}$ is $\epsilon$-differentially private, if*

$$\mathbb{P}(\mathcal{A}(Z) \in H) \leq \exp(\epsilon)\mathbb{P}(\mathcal{A}(Z') \in H)$$

*for $\forall\ Z,\ Z'$ obeying $d(Z, Z') = 1$ and any measurable subset $H \subseteq \mathcal{H}$.*

There are weaker notions of differential privacy. For example $(\epsilon, \delta)$-differential privacy allows for a small probability $\delta$ where the privacy guarantee does not hold. In this paper, we will work with the stronger $\epsilon$-differential privacy.

Our objective is to understand whether there is a gap between learnable problems and privately learnable problems in the general learning setting, and to quantify the tradeoff required to protect privacy. To achieve this objective, we need to show the existence of an algorithm that learns a class of problems while preserving differential privacy. More formally, we define

**Definition 3** (Private learnability)**.** *A learning problem is $\epsilon(n)$-differential-privately learnable with rate $\xi(n)$ if there exists an algorithm $\mathcal{A}$ that satisfies both universal consistency (as in Definition 1) with rate $\xi(n)$ and $\epsilon(n)$-differential privacy with privacy parameter $\epsilon(n) \to 0$ as $n \to \infty$.*

We can view the consistency requirement Definition 3 as a measure of utility. This utility is not a function of the observed data, however, but rather of how the results generalize to unseen data.

It may seem that this definition of private learnability requires a stronger privacy guarantee than usual. Indeed it is more common to only require that $\epsilon$ is some small constant. The following lemma explains that they are in fact equivalent for the problem of learnability.

**Lemma 4.** *If there is an $\epsilon$-DP algorithm that is consistent with rate $\xi(n)$ for any constant $0 < \epsilon < \infty$, then there is a $\frac{1}{\sqrt{n}}(e^{\epsilon} - e^{-\epsilon})$-DP algorithm that is consistent with rate $2\xi(\sqrt{n})$.*

The proof uses a subsampling theorem adapted from Beimel et al. [3, Lemma 4.4].

There are many approaches to design differentially private algorithms, such as noise perturbation using Laplace noise [8, 11] and the Exponential Mechanism [18]. Our construction of generic differentially private learning algorithms applies the Exponential Mechanism to penalized empirical risk minimization. Our argument will make use of a general characterization of learnability described below.

### 2.3 Stability and Asymptotic ERM

An important breakthrough in learning theory is a full characterization of all learnable problems in the General Learning Setting in terms of stability and empirical risk minimization [21]. Without assuming uniform convergence of empirical risk, Shalev-Shwartz et al. showed that a problem is learnable if and only if there exists a "strongly uniform-RO stable" and "always asymptotically empirical risk minimization" (Always AERM) randomized algorithm that learns the problem. Here "RO" stands for "replace one". Also, any strongly uniform-RO stable and "universally" AERM (weaker than "always" AERM) learning rule learns the problem consistently. Here we give detailed definitions.

**Definition 5** (Universally/Always AERM [21])**.** *A (possibly randomized) learning rule $\mathcal{A}$ is Universally AERM if for any distribution $\mathcal{D}$ defined on $\mathcal{Z}$*

$$\mathbb{E}_{Z\sim\mathcal{D}^n}\left[\mathbb{E}_{h\sim\mathcal{A}(Z)}\hat{R}(h,Z) - \hat{R}^*(Z)\right] \to 0, \quad \text{as } n \to \infty$$

*where $\hat{R}^*(Z)$ is the minimum empirical risk for data set $Z$. We say $\mathcal{A}$ is Always AERM, if in addition,*

$$\sup_{Z\in\mathcal{Z}^n} \mathbb{E}_{h\sim\mathcal{A}(Z)}\hat{R}(h,Z) - \hat{R}^*(Z) \to 0, \quad \text{as } n \to \infty\,.$$

**Definition 6** (Strongly Uniform RO-Stability [21])**.** *An algorithm $\mathcal{A}$ is strongly uniform RO-stable if*

$$\sup_{z\in\mathcal{Z}} \sup_{\substack{Z,Z'\in\mathcal{Z}^n,\\ d(Z,Z')=1}} \left|\mathbb{E}_{h\sim\mathcal{A}(Z)}\ell(h,z) - \mathbb{E}_{h\sim\mathcal{A}(Z')}\ell(h,z)\right| \to 0 \text{ as } n\to\infty.$$

*where $d(Z, Z')$ is defined in (3), in other word, $Z$ and $Z'$ can differ by at most one data point.*

Since we will not deal with other variants of algorithmic stability in this paper (e.g., hypothesis stability [14], uniform stability [5] and leave-one-out (LOO) stability in Mukherjee et al. [19]), we simply call Definition 6 stability or uniform stability. Likewise, we will refer to $\epsilon$-differential privacy as just "privacy" although there are several other notions of privacy in the literature.

## 3 Main results: characterization of private learnability

We are now ready to state our main result. The only assumption we make is the uniform boundedness of the loss function. This is also assumed in Shalev-Shwartz et al. [21] for the learnability problem without privacy constraints. Without loss of generality, we can assume $0 \le \ell(h,z) \le 1$.

**Theorem 7.** *Given a learning problem $(\mathcal{Z}, \mathcal{H}, \ell)$, the following statements are equivalent.*

1. *The problem is privately learnable.*
2. *There exists a differentially private universal AERM algorithm.*
3. *There exists a differentially private always AERM algorithm.*

The proof is simple yet revealing, we will present the arguments for 2 ⇒ 1 (sufficiency of AERM) in Section 3.1 and 1 ⇒ 3 (necessity of AERM) in Section 3.2. 3 ⇒ 2 follows trivially from the definition of "always" and "universal" AERM.

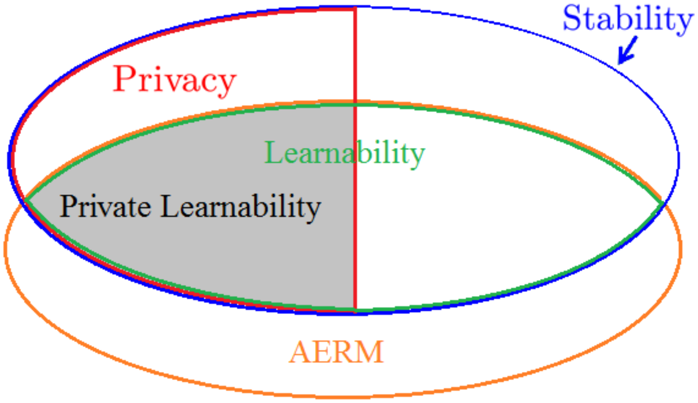


Figure 2: A summary of the relationships of various notions revealed by our analysis.

The theorem is conceptually very interesting. It says that we can stick to ERM-like algorithms for private learning, despite that ERM may fail for some problems in the (non-private) general learning setting [21]. Thus a standard procedure for finding universally consistent and differentially private algorithms would be to approximately minimize the empirical risk using some differentially private procedures [7, 16, 2]. If the utility analysis reveals that the method is AERM, we do not need to worry about generalization as it is guaranteed by privacy. This consistency analysis is considerably simpler than non-private learning problems where one typically needs to control generalization error either via uniform convergence (VC-dimension, Rademacher complexity, metric entropy, etc) or to adopt the stability argument [21].

This result does not imply that privacy is helping the algorithm to learn in any sense, as the simplicity is achieved at the cost of having a smaller class of learnable problems. A concrete example of a problem being learnable but not privately learnable is given in [6] and we will revisit it in Section 3.4. For some problems where ERM fails, it may not be possible to make it AERM while preserving privacy. In particular, we were not able to privatize the problem in Section 4.1 of Shalev-Shwartz et al. [21].

To avoid any potential misunderstanding, we stress that Theorem 7 is a characterization of learnability, *not* learning algorithms. It does not prevent the existence of a universally consistent learning algorithm that is private but not AERM. Also, the characterization given in Theorem 7 is about consistency, and it does not claim anything on sample complexity. The algorithm that is AERM may be suboptimal in terms of convergence rate.

### 3.1 Sufficiency: Privacy implies stability

A key ingredient in the proof of sufficiency is that differential privacy by definition implies uniform stability, which is useful in its own right.

**Lemma 8** (Privacy $\Rightarrow$ Stability)**.** *Assume boundedness* $0 \leq \ell(h, z) \leq 1$*, any* $\epsilon$*-differential private algorithm satisfies* $(e^{\epsilon} - 1)$*-stability. Moreover if* $\epsilon \leq 1$ *it satisfies* $2\epsilon$*-stability.*

The proof of this lemma comes directly from the definition of differential privacy so it is algorithm independent. The converse, however, is not true in general (e.g., a non-trivial deterministic algorithm can be stable, but not differentially private.)

**Corollary 9** (Privacy + Universal AERM $\Rightarrow$ Consistency)**.** *If a learning algorithm* $\mathcal{A}$ *is* $\epsilon(n)$*-differentially private and* $\mathcal{A}$ *is universally AERM with rate* $\xi(n)$*, then* $\mathcal{A}$ *is universally consistent with rate* $\xi(n) + 2\epsilon(n)$*.*

The proof of Corollary 9, which we provide in the Appendix, combines Lemma 8 and the fact that consistency is implied by stability and AERM, where the proof of AERM is based on Theorem 8 in Shalev-Shwartz et al. [21]. In fact, Corollary 9 can be stated in a stronger per distribution form, since universality is not used in the proof. We will revisit this point when we discuss a weaker notion of private learnability below.

If for a problem privacy and always AERM cannot coexist, then the problem is not privately learnable. This is what we will show next.

### 3.2 Necessity: Consistency implies Always AERM

To prove that the existence of an always AERM learning algorithm is necessary for the any private learnable problems, it suffices to construct such a learning algorithm for any universally consistent learning algorithm.

**Lemma 10** (Consistency + Privacy $\Rightarrow$ Private Always AERM)**.** *If $\mathcal{A}$ is a universally consistent learning algorithm satisfying $\epsilon$-DP with any $\epsilon > 0$ and consistent with rate $\xi(n)$, then there is another universally consistent learning algorithm $\mathcal{A}'$ that is always AERM with rate $2\xi(\sqrt{n})$ and satisfies $\frac{1}{\sqrt{n}}(e^{\epsilon} - e^{-\epsilon})$-DP.*

The intuition of the necessity makes use of the distribution-free requirement of the learnability, and constructs an algorithm that is AERM. Detailed proofs are given in the full paper [28].

### 3.3 A generic learning algorithm

The characterization of private learnability suggests a generic (but impractical) procedure that learns all privately learnable problems (in the same flavor as the generic algorithm in Shalev-Shwartz et al. [21] that learns all learnable problems). This is to solve

$$\underset{\substack{(\mathcal{A},\epsilon):\\ \mathcal{A}:\mathcal{Z}^n\to\mathcal{H},\\ \mathcal{A}\text{ is }\epsilon\text{-DP}}}{\operatorname{argmin}} \left[\epsilon + \sup_{Z\in\mathcal{Z}^n}\left(\mathbb{E}_{h\sim\mathcal{A}(Z)}\hat{R}(h,Z) - \inf_{h\in\mathcal{H}}\hat{R}(h,Z)\right)\right], \tag{4}$$

or to privately $\mathfrak{D}$-learn the problem when (4) is not feasible

$$\underset{\substack{(\mathcal{A},\epsilon):\\ \mathcal{A}:\mathcal{Z}^n\to\mathcal{H},\\ \mathcal{A}\text{ is }\epsilon\text{-DP}}}{\operatorname{argmin}} \left[\epsilon + \sup_{\mathcal{D}\in\mathfrak{D}}\mathbb{E}_{Z\sim\mathcal{D}^n}\left(\mathbb{E}_{h\sim\mathcal{A}(Z)}\hat{R}(h,Z) - \inf_{h\in\mathcal{H}}\hat{R}(h,Z)\right)\right]. \tag{5}$$

**Theorem 11.** *Assume the problem is learnable. If the problem is private learnable,* (4) *will always output a universally consistent private learning algorithm.*

*Proof.* If the problem is private learnable, by Theorem 7 there exists an algorithm $\mathcal{A}$ that is $\epsilon(n)$-DP and always AERM with rate $\xi(n)$ and $\epsilon(n) + \xi(n) \to 0$. This $\mathcal{A}$ is a witness in the optimization so we know that any minimizer of (4) will have a objective value that is no greater than $\epsilon(n) + \xi(n)$ for any $n$. Corollary 9 concludes its universal consistency. ■

It is of course impossible to minimize the supremum over any data $Z$, nor is it possible to efficiently search over the space of all algorithms, let alone DP algorithms, but conceptually, this formulation may be of great interest to a number of theoretical questions related to the search of private learning algorithms and the fundamental limit of machine learning under privacy constraint.

### 3.4 Private Learnability vs. Non-private Learnability

Now we have a characterization of all privately learnable problems, and a conceptual procedure that always produces an private algorithm to learn a privately learnable problem. A natural question to ask is how large is the set of privately learnable problems in the space of all (non-privately) learnable problems. Specifically, is any learnable problem also privately learnable?

Turns out that the answer is "yes" for learning in Statistical Query (SQ)-model and PAC Learning model with finite hypothesis space, and is "no" for continuous hypothesis space.

This can already be seen from Chaudhuri & Hsu [6], where the gap between learning and private learning is established. We revisit Chaudhuri & Hsu's interesting example in our notation under the general learning setting and produce an alternative proof by showing that differential privacy contradicts *always AERM*, then invoking Theorem 7 to show the problem is not learnable.

**Proposition 12** (Theorem 5 in [6])**.** *There is a learning problem that is learnable by a non-private algorithm, but not privately learnable. In particular, any private algorithm cannot be* always AERM *in this problem.*

We describe the counterexample and re-establish the impossibility of private learning for this problem using the contrapositive of Theorem 7, which suggests that if privacy and always AERM algorithm cannot coexist for some problem, then the problem is not privately learnable.

Consider the binary classification problem with $\mathcal{X} = [0, 1]$, $\mathcal{Y} = \{0, 1\}$ and 0-1 loss function. Let $\mathcal{H}$ be the collection of threshold functions that output $h(x) = 1$ if $x > h$ and $h(x) = 0$ otherwise. This class has VC-dimension 1, and hence the problem is learnable.

Next we will construct $K = \lceil \exp(\epsilon_n n) \rceil$ data sets such that if $K - 1$ of them obeys AERM, the remaining one cannot be. Let $\eta = 1/\exp(\epsilon n)$, $K := \lceil 1/\eta \rceil$. Let $h_1, h_2, ..., h_K$ be a disjoint thresholds such that they are at least $\eta$ apart and $[h_i - \eta/3, h_i + \eta/3]$ are disjoint intervals.

If we take $Z_i \subseteq [h_i - \eta/3, h_i + \eta/3]$ with half of the points in $[h_i - \eta/3, h_i)$ and the other half in $(h_i, h_i + \eta/3]$ and we label each data point in it with $\mathbf{1}(z > h_i)$, then empirical risk $\hat{R}(h_i, Z_i) = 0 \; \forall i = 1, ..., K$. So for any AERM learning rule, $\mathbb{E}_{h\sim\mathcal{A}(Z_i)}\hat{R}(h, Z_i) \to 0$ for all $i$. For some sufficiently large $n$, $\mathbb{E}_{h\sim\mathcal{A}(Z_i)}\hat{R}(h, Z_i) < 0.1$.

Now consider $Z_1$,

$$\mathbb{P}(\mathcal{A}(Z_1) \notin [h_1 - \eta/3, h_1 + \eta/3]) \geq \sum_{i=2}^{K} \mathbb{P}(\mathcal{A}(Z_1) \in [h_i - \eta/3, h_i + \eta/3]),$$

since these intervals are disjoint. Then by the definition of $\epsilon$-DP,

$$\mathbb{P}(\mathcal{A}(Z_1) \in [h_i - \eta/3, h_i + \eta/3]) \geq \exp(-\epsilon n)\mathbb{P}(\mathcal{A}(Z_i) \in [h_i - \eta/3, h_i + \eta/3]).$$

It follows that $\mathbb{P}(\mathcal{A}(Z_i) \in [h_i - \eta/3, h_i + \eta/3]) > 0.9$ otherwise $\mathbb{E}_{h\sim\mathcal{A}(Z_i)}\hat{R}(h, Z_i) \geq 0.1$, therefore

$$\mathbb{P}(\mathcal{A}(Z_1) \notin [h_1 - \eta/3, h_1 + \eta/3]) \geq K \exp(-\epsilon n) 0.9 \geq 0.9,$$

and $\mathbb{E}_{h\sim\mathcal{A}(Z_i)}\hat{R}(h, Z_i) \geq 0.9 \times 1 = 0.9$, which violates the "always AERM" condition that requires $\mathbb{E}_{h\sim\mathcal{A}(Z_1)}\hat{R}(h, Z_1) < 0.1$.

The above example suggests that even very simple learning problems may not be privately learnable, which motivates our proposal of Private $\mathfrak{D}$-learnability that requires only consistency on a class $\mathfrak{D}$ of "nice" distributions. With an additional technical assumption, we are able to characterize this larger class of problems too. Details of the results are presented in the full paper [28].

## 4 Conclusion

In this paper, we revisited the question *"What can we learned privately?"* and considered a much broader class of statistical machine learning problems than those studied previously. Specifically, we characterized the learnability under privacy constraint by showing any privately learnable problems can be learned by a private algorithm that asymptotically minimizes the empirical risk for any data, and the problem is not privately learnable otherwise. Our analysis reveals an interesting insight: differential privacy implies a notion of stability that partially characterizes learnable problems. Building upon the characterization, we provided a conceptual procedure that learns any privately learnable problem.

However, the set of all privately learnable problems is a strict subset of all learnable problems and in many cases exclude even very simple learning problems. Our fix to this issue, we propose a relaxed notion of private learnability called private $\mathfrak{D}$-learnability. To see detailed descriptions of this fix, and other technical results and examples, please refer to [28].

# References

[1] Applegate, D., & Kannan, R. (1991). Sampling and integration of near log-concave functions. In *Proceedings of 23rd ACM Symposium on Theory of Computing*, (pp. 156–163).

[2] Bassily, R., Smith, A., & Thakurta, A. (2014). Private empirical risk minimization, revisited. *rem*, *3*, 17.

[3] Beimel, A., Brenner, H., Kasiviswanathan, S. P., & Nissim, K. (2014). Bounds on the sample complexity for private learning and private data release. *Machine learning*, *94*(3), 401–437.

[4] Beimel, A., Nissim, K., & Stemmer, U. (2013). Characterizing the sample complexity of private learners. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, (pp. 97–110). ACM.

[5] Bousquet, O., & Elisseeff, A. (2002). Stability and generalization. *Journal of Machine Learning Research*, *2*, 499–526.

[6] Chaudhuri, K., & Hsu, D. (2011). Sample complexity bounds for differentially private learning. In *COLT*, vol. 19, (pp. 155–186).

[7] Chaudhuri, K., Monteleoni, C., & Sarwate, A. D. (2011). Differentially private empirical risk minimization. *Journal of Machine Learning Research*, *12*, 1069–1109.

[8] Dwork, C. (2006). Differential privacy. In *Automata, Languages and Programming*, (pp. 1–12). Springer.

[9] Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., & Roth, A. (2014). Preserving statistical validity in adaptive data analysis. *arXiv preprint arXiv:1411.2664*.

[10] Dwork, C., & Lei, J. (2009). Differential privacy and robust statistics. In *Proceedings of 41st ACM Symposium on Theory of Computing*, (pp. 371–380). ACM.

[11] Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography*, (pp. 265–284). Springer.

[12] Haussler, D. (1992). Decision theoretic generalizations of the pac model for neural net and other learning applications. *Information and computation*, *100*(1), 78–150.

[13] Kasiviswanathan, S. P., Lee, H. K., Nissim, K., Raskhodnikova, S., & Smith, A. (2011). What can we learn privately? *SIAM Journal on Computing*, *40*(3), 793–826.

[14] Kearns, M., & Ron, D. (1999). Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural Computation*, *11*(6), 1427–1453.

[15] Kearns, M. J., Schapire, R. E., & Sellie, L. M. (1992). Toward efficient agnostic learning. In *Proceedings of the fifth annual workshop on Computational learning theory*, (pp. 341–352). ACM.

[16] Kifer, D., Smith, A., & Thakurta, A. (2012). Private convex empirical risk minimization and high-dimensional regression. *Journal of Machine Learning Research*, *1*, 41.

[17] Lei, J. (2011). Differentially private *m*-estimators. In *NIPS*, (pp. 361–369).

[18] McSherry, F., & Talwar, K. (2007). Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science, 2007 (FOCS'07)*, (pp. 94–103).

[19] Mukherjee, S., Niyogi, P., Poggio, T., & Rifkin, R. (2006). Learning theory: Stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics*, *25*(1-3), 161–193.

[20] Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning*, *5*(2), 197–227.

[21] Shalev-Shwartz, S., Shamir, O., Srebro, N., & Sridharan, K. (2010). Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, *11*, 2635–2670.

[22] Smith, A. (2011). Privacy-preserving statistical estimation with optimal convergence rates. In *Proceedings of 43rd Annual ACM symposium on Theory of Computing*, (pp. 813–822).

[23] Thakurta, A. G., & Smith, A. (2013). Differentially private feature selection via stability arguments, and the robustness of the lasso. In *Conference on Learning Theory*, (pp. 819–850).

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

[25] Vapnik, V. (1998). *Statistical Learning Theory*. Wiley.

[26] Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer.

[27] Wang, Y.-X., Fienberg, S. E., & Smola, A. (2015). Privacy for free: Posterior sampling and stochastic gradient monte carlo. In *ICML'15*.

[28] Wang, Y.-X., Lei, J., & Fienberg, S. E. (2015). Learning with differential privacy: Stability, learnability and the sufficiency and necessity of erm principle. *arXiv preprint arXiv:1502.06309*.

[29] Wasserman, L., & Zhou, S. (2010). A statistical framework for differential privacy. *Journal of the American Statistical Association*, *105*(489), 375–389.

[30] Yu, F., Fienberg, S. E., Slavković, A., & Uhler, C. (2014). Scalable privacy-preserving data sharing methodology for genome-wide association studies. *Journal of Biomedical Informatics*, (p. in press).