Journal of Machine Learning Research 17 (2016) 1-40 Submitted 6/15; Revised 3/16; Published 4/16

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

Yu-Xiang Wang[1,2] YUXIANGW@CS.CMU.EDU
Jing Lei[2] JINGLEI@ANDREW.CMU.EDU
Stephen E. Fienberg[1,2] FIENBERG@STAT.CMU.EDU
[1] *Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213*
[2] *Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213*

Editor: Moritz Hardt

Abstract

While machine learning has proven to be a powerful data-driven solution to many real-life problems, its use in sensitive domains has been limited due to privacy concerns. A popular approach known as *differential privacy* offers provable privacy guarantees, but it is often observed in practice that it could substantially hamper learning accuracy. In this paper we study the learnability (whether a problem can be learned by any algorithm) 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). In contrast, 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. Lastly, we extend some of the results to the more practical (ϵ, δ)-differential privacy and establish the existence of a phase-transition on the class of problems that are approximately privately learnable with respect to how small δ needs to be.

Keywords: *differential privacy, learnability, characterization, stability, privacy-preserving machine learning*

1. Introduction

Increasing public concerns regarding data privacy have posed obstacles in the development and application of new machine learning methods as data collectors and curators may no longer be able to share data for research purposes. In addition to addressing the original goal of information extraction, privacy-preserving learning also requires the learning procedure to protect sensitive information of individual data entries. For example, the second Netflix Prize competition was canceled in response to a lawsuit and Federal Trade Commission privacy concerns, and the National Institute of Health decided in August 2008 to remove

aggregate Genome-Wide Association Studies (GWAS) data from the public web site, after learning about a potential privacy risk.

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* (Dwork, 2006; Dwork et al., 2006b). Differential privacy has three main advantages over other approaches: (1) it rigorously quantifies the privacy property of any data analysis mechanism; (2) it controls 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 (Dwork and Lei, 2009; Wasserman and Zhou, 2010; Smith, 2011; Lei, 2011; Wang et al., 2015), as well as applications involving regression (Chaudhuri et al., 2011; Thakurta and Smith, 2013) and GWAS data (Yu et al., 2014), 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 extensions of various learning algorithms, 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 to learn, 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 (Vapnik, 1995) in four 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 (Kasiviswanathan et al., 2011; Beimel et al., 2013a) that focus on binary classification in discrete domain under the PAC learning model. Technically, the result relies on the now well-known intuitive observation that "privacy implies algorithmic stability" and the argument in Shalev-Shwartz et al. (2010) that shows a variant of algorithmic stability is necessary for learnability.

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 (Chaudhuri and Hsu, 2011)) are usually private $\mathfrak{D}$-learnable for some "nice" distribution class $\mathfrak{D}$. We characterize the subset of private $\mathfrak{D}$-learnable problems that are also (non-privately) learnable 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 always finds a consistent and private algorithm for any privately learnable (or $\mathfrak{D}$-learnable) problems. We also study a specific 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 (Beimel et al., 2013a). It remains an open problem whether the second algorithm also learns any privately learnable problem in the General Learning Setting.

4. Lastly, we provide a preliminary study of learnability under the more practical (ϵ, δ)-differential privacy. Our results reveal that whether there is separation between learnability and approximate private learnability depends on how fast δ is required to go to 0 with respect to the size of the data. Finding where the exact phase transition occurs is an open problem of future interest.

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.

Related work While a large amount of work has been devoted to finding consistent (and rate optimal) differentially private learning algorithms in various settings (e.g., Chaudhuri et al., 2011; Kifer et al., 2012; Jain and Thakurta, 2013; Bassily et al., 2014), the characterization of privately learnable problems were only studied in a few special cases.

Kasiviswanathan et al. (2011) showed that, for binary classification with a finite discrete hypothesis space, anything that is non-privately learnable is privately learnable under the agnostic Probably Approximately Correct (PAC) learning framework, therefore "finite VC-dimension" characterizes the set of private learnable problems in this setting. Beimel et al. (2013a) extends Kasiviswanathan et al. (2011) by characterizing the sample complexity of the same class of problems, but the result only applies to the realizable (non-agnostic) case. Chaudhuri and Hsu (2011) provided a counter-example showing that for continuous hypothesis space and data space, there is a gap between learnability and learnability under privacy constraint. They proposed to fix this issue by either weakening the privacy requirement to labels only or by restricting the class of potential distribution. While meaningful in some cases, these approaches do not resolve the learnability problem in general.

A key difference of our work from Kasiviswanathan et al. (2011); Chaudhuri and Hsu (2011); Beimel et al. (2013a) is that we consider a more general class of learning problems and provide a proper treatment in a statistical learning framework. This allows us to capture a wider collection of important learning problems (see Figure 1(a) and Table 1).

It is important to note that despite its generality, Vapnik's general learning setting still does not nearly cover the full spectrum of private learning. In particular, our results do not apply to improper learning (learning using a different hypothesis class) as considered in Beimel et al. (2013a) or to structural loss minimization (the loss function jointly take all data points as input) considered in Beimel et al. (2013b). Also, our results do not address the sample complexity problem, which remains open in the general learning setting even for learning without privacy constraints.

Our characterization of private learnability (and private $\mathfrak{D}$-learnability) in Section 3 uses a recent advance in the characterization of general learnability given by Shalev-Shwartz et al. (2010). Roughly speaking, they showed that a problem is learnable if and only if there exists an algorithm that (i) is stable under small perturbation of training data, and (ii) behaves like empirical risk minimization (ERM) asymptotically. We also makes use of a folklore observation that "Privacy $\Rightarrow$ Stability $\Rightarrow$ Generalization". The connection of privacy and stability appeared as early as 2008 in a conference version of Kasiviswanathan et al. (2011). Further connection to "generalization" recently appeared in blog posts[1], stated as a theorem in Appendix F of Bassily et al. (2014), and was shown to hold with strong concentration in Dwork et al. (2015b).

Dwork et al. (2015b) is part of an independent line of work (Hardt and Ullman, 2014; Bassily et al., 2015; Dwork et al., 2015a; Blum and Hardt, 2015) on adaptive data analysis, which also stems from the observation that privacy implies stability and generalization. Comparing to adaptive data analysis works, our focus is quite different. Adaptive data analysis work focus on the impact of k on how fast the maximum absolute error of k-adaptively chosen queries goes to 0 as a function of n, while this paper is concerned with whether the error can go to 0 at all for each learning problem when we require the learning algorithm be differentially private with $\epsilon < \infty$. Nonetheless, we acknowledge that Theorem 7 in Dwork et al. (2015b) provides an interesting alternative proof for "differentially private learners have small generalization error", when choosing the statistical query as evaluating a loss function at a privately learned hypothesis. The connection is not quite obvious and we provide a more detailed explanation in Appendix B.

The main tool used in the construction of our generic private learning algorithm in Section 4 is the Exponential Mechanism (McSherry and Talwar, 2007), which provides a simple and differentially-private approximation to the maximizer of a score function among a candidate set. In the general learning context, we use the negative empirical risk as the utility function, and apply the exponential mechanism to a possibly pre-discretized hypothesis space. This exponential mechanism approach was used in Bassily et al. (2014) for minimizing convex and Lipschitz functions. The sample discretization procedure has been considered in Chaudhuri and Hsu (2011) and Beimel et al. (2013a). Our scope and proof techniques are different. Our strategy is to show that, under some general regularity conditions, the exponential mechanism is stable and behaves like ERM. Our sublevel set condition has the same flavor

1. For instance, Frank McSherry described in a blog post an example of exploiting differential privacy for measure concentration `http://windowsontheory.org/2014/02/04/differential-privacy-for-measure-concentration/`; 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`.

as that in the proof of Bassily et al. (2014, Theorem 3.2), although we do not need the loss function to be convex or Lipschitz.

Stability, privacy and generalization were also studied in Thakurta and Smith (2013) with different notions of stability. More importantly, their stability is used as an assumption rather than a consequence, so their result is not directly comparable to ours.

2. Background

2.1 Learnability under the General Learning Setting

In the General Learning Setting of Vapnik (1995), a learning problem is characterized by a triplet $(\mathcal{Z}, \mathcal{H}, \ell)$. Here $\mathcal{Z}$ is the sample space (with a σ-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. (2010)).

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 $Z \in \mathcal{Z}^n$:

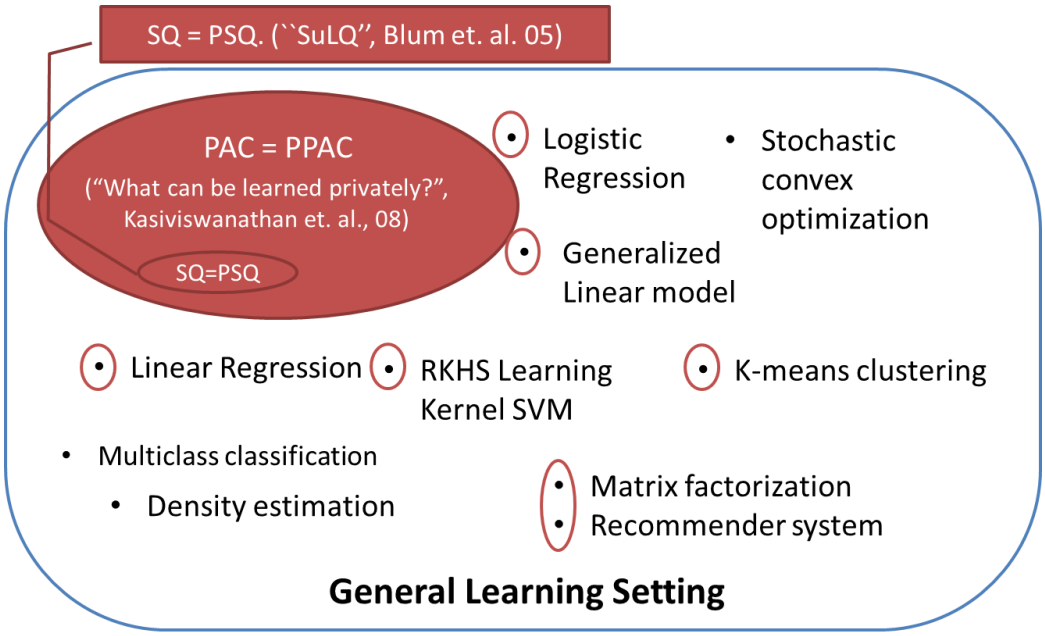
$$R(h) = \mathbb{E}_{z\sim\mathcal{D}}\ell(h, z), \qquad \hat{R}(h, Z) = \frac{1}{n}\sum_{i=1}^{n} \ell(h, z_i)\,.$$

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

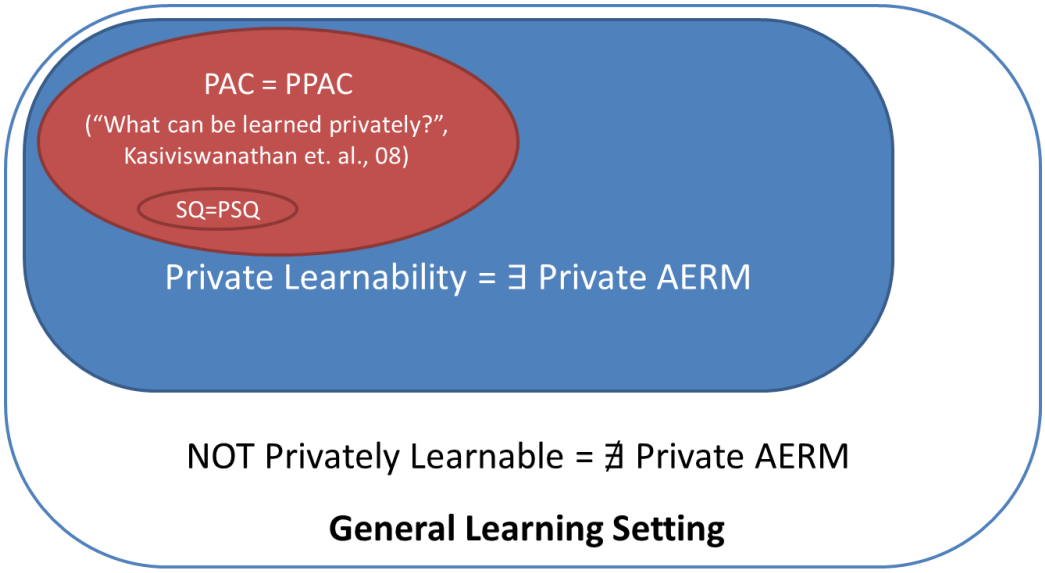
$$\lim_{n\to\infty} \mathbb{E}_{Z\sim\mathcal{D}^n}\left(\mathbb{E}_{h\sim\mathcal{A}(Z)}R(h) - R^*\right) = 0. \tag{1}$$

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

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



(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.

Problem	Hypothesis class $\mathcal{H}$	$\mathcal{Z}$ or $\mathcal{X} \times \mathcal{Y}$	Loss function ℓ
Binary classification	$\mathcal{H} \subset \{f : \{0,1\}^d \to \{0,1\}\}$	$\{0,1\}^d \times \{0,1\}$	$1(h(x) \neq y)$
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	$\{S \subset \mathbb{R}^d : \lvert S\rvert = k\}$	$\mathcal{Z} \subset \mathbb{R}^d$	$\min_{c \in h} \lVert c - z\rVert^2$
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 $\mathbb{R}^d$	$\mathbb{R}^d \times \{1,0\}^d$	$\min_{b \in h} \lVert y \circ (b - x)\rVert^2 + \lambda\text{rank}(h)$
Dictionary Learning	All dictionaries $\in \mathbb{R}^{d \times r}$	$\mathbb{R}^d$	$\min_{b \in \mathbb{R}^r} \lVert hb - z\rVert^2 + \lambda\lVert b\rVert_1$
Non-negative MF	All dictionaries $\in \mathbb{R}_+^{d \times r}$	$\mathbb{R}^d$	$\min_{b \in \mathbb{R}_+^r} \lVert hb - z\rVert^2$
Subspace Clustering	A set of k rank-r subspaces	$\mathbb{R}^d$	$\min_{b \in h} \lVert\mathcal{P}_b(z) - z\rVert^2$
Topic models (LDA)	$\{\mathbb{P}(\text{word}\lvert\text{topic})\}$	Documents	$-\max_{b \in \{\mathbb{P}(\text{Topic})\}} \sum_{w \in z} \log \mathbb{P}_{b,h}(w)$

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

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. (2010), which is a direct generalization of PAC-learnability (Valiant, 1984) and agnostic PAC-learnability (Kearns et al., 1992) to the General Learning Setting as studied by Haussler (1992).

Definition 1 (Learnability, Shalev-Shwartz et al., 2010) *A learning problem 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 uniform (distribution-independent) rate $\xi(n)$. This type of problem is often called *distribution-free learning* (Valiant, 1984), 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 (Wasserman and Zhou, 2010). Formally, define "Hamming distance"

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

Definition 2 (ϵ-Differential Privacy, Dwork, 2006) *An algorithm $\mathcal{A}$ is ϵ-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 (ϵ, δ)-differential privacy allows for a small probability δ where the privacy guarantee does not hold. In this paper, we will mainly work with the stronger ϵ-differential privacy. In Section 6 we discuss the problem of (ϵ, δ)-differential privacy and extend some of the results to this setting.

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 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 ϵ-differential privacy with privacy parameter $\epsilon < \infty$.*

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 how the results generalize to unseen data.

The following lemma shows that the above definition of private learnability is actually equivalent to a seemingly much stronger condition with a vanishing privacy loss ϵ.

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

The proof, given in Appendix A.1, uses a subsampling theorem adapted from Beimel et al. (2014, Lemma 4.4).

There are many approaches to design differentially private algorithms, such as noise perturbation using Laplace noise (Dwork, 2006; Dwork et al., 2006b) and the Exponential Mechanism (McSherry and Talwar, 2007). 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 (Shalev-Shwartz et al., 2010). 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, Shalev-Shwartz et al., 2010) *A (possibly randomized) learning rule $\mathcal{A}$ is Universally AERM if for any distribution $\mathcal{D}$ defined on domain $\mathcal{Z}$*

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

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

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

Definition 6 (Strongly Uniform RO-Stability, Shalev-Shwartz et al., 2010) *An algorithm $\mathcal{A}$ is strongly uniform RO-stable if*

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

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 (Kearns and Ron, 1999), uniform stability (Bousquet and Elisseeff, 2002) and leave-one-out (LOO) stability in Mukherjee et al. (2006)), we simply call Definition 6 stability or uniform stability. Likewise, we will refer to ϵ-differential privacy as just "privacy" although there are several other notions of privacy in the literature.

3. 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. (2010) 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 universally AERM algorithm.*
3. *There exists a differentially private always AERM algorithm.*

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

The theorem 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 (Shalev-Shwartz et al., 2010). Thus a standard procedure for finding universally consistent and

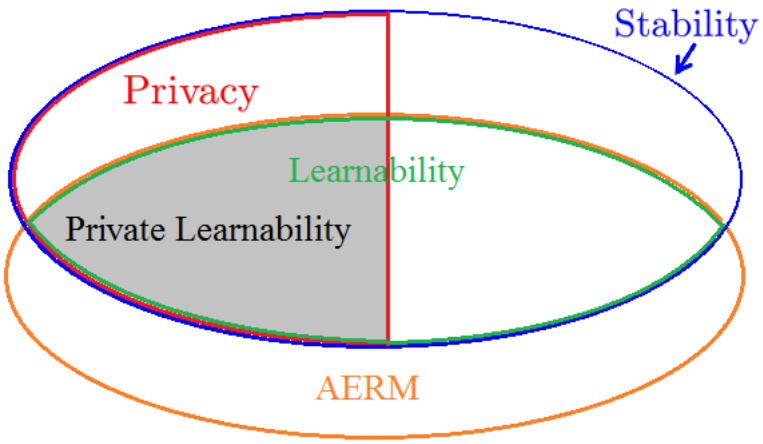


Figure 2: A summary of the relationships of various notions revealed by our analysis.

differentially private algorithms would be to approximately minimize the empirical risk using some differentially private procedures (Chaudhuri et al., 2011; Kifer et al., 2012; Bassily et al., 2014). 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 (Shalev-Shwartz et al., 2010).

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 (Chaudhuri and Hsu, 2011) and we will revisit it in Section 3.3. 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. (2010).

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. An 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 a well-known heuristic observation that differential privacy by definition implies uniform stability, which is useful in its own right.

Lemma 8 (Privacy $\Rightarrow$ Stability) *Assume $0 \le \ell(h, z) \le 1$, any ϵ-differentially private algorithm satisfies $(e^\epsilon - 1)$-stability. Moreover if $\epsilon \le 1$ it satisfies 2ϵ-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 ⇒ 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 $\xi(n) + e^{\epsilon(n)} - 1 = O(\xi(n) + \epsilon(n))$.*

The proof of Corollary 9, provided in the Appendix, combines Lemma 8 and the fact that consistency is implied by stability and AERM (Theorem 28). Our Theorem 28 is based on minor modifications of Theorem 8 in Shalev-Shwartz et al. (2010). 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.

Lemma 4 and Corollary 9 together establishes $2 \Rightarrow 1$ in Theorem 7.

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 any private learnable problems, it suffices to construct such a learning algorithm from

or each learnable problem. any universally consistent learning algorithm.

Lemma 10 (Consistency + Privacy ⇒ Private Always AERM) *If $\mathcal{A}$ is a universally consistent learning algorithm satisfying ϵ-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 $\xi(\sqrt{n})$ and satisfies $\frac{2}{\sqrt{n}}(e^{\epsilon} - e^{-\epsilon})$-DP.*

Lemma 10 is proved in Appendix A.2. The proof idea is to run $\mathcal{A}$ on a size $O(\sqrt{n})$ random subsample of Z, which will be universally consistent with a slower rate, differentially private with $\epsilon(n) \to 0$ (Lemma 27), and at the same time always AERM. The last part uses an argument in Lemma 24 of Shalev-Shwartz et al. (2010) which appeals to the universality of $\mathcal{A}$'s consistency on a specific discrete distribution supported on the given data set Z.

As pointed out by an anonymous reviewer, there is a simpler proof by invoking Theorem 10 of Shalev-Shwartz et al. (2010) that says any consistent and generalizing algorithm must be AERM and a result (e.g., Bassily et al., 2014, Appendix F) that says "privacy ⇒ generalization". This is a valid observation. But their Theorem 10 is proven using a detour through "generalization", which leads to a slower rate than what we are able to obtain in Lemma 10 using a more direct argument.

3.3 Private Learnability vs. Non-private Learnability

Now we have a characterization of all privately learnable problems, a natural question to ask is that whether any learnable problem is also privately learnable. The answer is "yes"

for learning in Statistical Query (SQ)-model and PAC Learning model (binary classification) with finite hypothesis space, and is "no" for continuous hypothesis space (Chaudhuri and Hsu, 2011).

By definition, all privately learnable problems are learnable. But now that we know that privacy implies generalization, it is tempting to hope that privacy can help at least some problem to learn better than any non-private algorithm. In terms of learnability, the question becomes: Could there be a (learnable) problem that is *exclusively* learnable through private algorithms? We now show that such a problem does not exist.

Proposition 11 *If a learning problem is learnable by an ϵ-DP algorithm $\mathcal{A}$, then it is also learnable by a non-private algorithm.*

The proof is given in Appendix A.3. The idea is that $\mathcal{A}(Z)$ defines a distribution over $\mathcal{H}$. Pick an $z \in \mathcal{Z}$. If $z \notin Z$, algorithm $\mathcal{A}' = \mathcal{A}$. Otherwise, $\mathcal{A}'(Z)$ samples from a slightly different distribution than $\mathcal{A}(Z)$ that does not affect the expectation much.

On the other hand, not all learnable problems are privately learnable. This can already be seen from Chaudhuri and Hsu (2011), where the gap between learning and private learning is established. We revisit Chaudhuri and Hsu's 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 privately learnable.

Proposition 12 (Chaudhuri and Hsu, 2011, Theorem 5) *There exists a 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 obey 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 η 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 ϵ-DP,

$$\mathbb{P}(\mathcal{A}(Z_1) \in [h_i - \eta/3, h_i + \eta/3]) \geq \exp(-\epsilon n)\mathbb{P}(\mathcal{A}(Z_i) \in [h_i - \eta/3, h_i + \eta/3]). \tag{4}$$

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

$$\mathbb{P}(\mathcal{A}(Z_1) \notin [h_1 - \eta/3, h_1 + \eta/3]) \geq K \exp(-\epsilon n)0.9 \geq 0.9, \tag{5}$$

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 $\mathbb{E}_{h\sim\mathcal{A}(Z_1)}\hat{R}(h, Z_1) < 0.1$. Therefore, the problem is not privately learnable.

As is pointed out by an anonymous reviewer, the same conclusion of this impossibility result of privately learning thresholds on $[0, 1]$ can be drawn numerically through the characterization of the sample complexity (Beimel et al., 2013a), via the bound that depends logarithmically on the $\log(|\mathcal{H}|)$ and on $[0, 1]$ this number is infinite. The above analysis provides different insights about the problem. We will be using it again for understanding the separation of learnability and learnability under (ϵ, δ)-Differential Privacy later in Section 6.

3.4 Private $\mathfrak{D}$-learnability

The above example implies that even very simple learning problems may not be privately learnable. To fix this caveat, note that most data sets of practical interest have nice distributions. Therefore, it makes sense to consider a smaller class of distributions, e.g., smooth distributions that have bounded kth order derivative, or those having bounded total variation. These are common assumptions in non-parametric statistics, such as kernel density estimation, smoothing spline regression and mode clustering. Similarly, in high dimensional statistics, there are often assumptions on the structures of the underlying distribution, such as sparsity, smoothness, and low-rank conditions.

Definition 13 ((Private) $\mathfrak{D}$-learnability) *We say a learning problem $(\mathcal{Z}, \mathcal{H}, \ell)$ is $\mathfrak{D}$-learnable if there exists a learning algorithm $\mathcal{A}$ that is consistent for every unknown distribution $\mathcal{D} \in \mathfrak{D}$. If in addition, the problem is $\mathfrak{D}$-learnable under ϵ-differential privacy for some $0 \leq \epsilon < \infty$, then we say the problem is privately $\mathfrak{D}$-learnable.*

Almost all of our arguments hold in a per distribution fashion, therefore they also hold for any such subclass $\mathfrak{D}$. The only exception is the necessity of "always AERM" (Lemma 10), where we used the universal consistency on an arbitrary discrete uniform distribution in the proof. The characterization still holds if the class $\mathfrak{D}$ contains all finite discrete uniform distributions. For general distribution classes, we characterize private $\mathfrak{D}$-learnability using a weaker "universally AERM" (instead of "always AERM") under the assumption that the problem itself is learnable in a distribution-free setting without privacy constraints.

Lemma 14 (private $\mathfrak{D}$-learnability $\Rightarrow$ private $\mathfrak{D}$-universal AERM) *If an ϵ-DP algorithm $\mathcal{A}$ is $\mathfrak{D}$-universally consistent with rate $\xi(n)$ and the problem itself is learnable in a distribution-free sense with rate $\xi'(n)$, then there exists a $\mathfrak{D}$-universally consistent learning algorithm $\mathcal{A}'$ that is $\mathfrak{D}$-universally AERM with rate $12\xi'(n^{1/4}) + \frac{37}{\sqrt{n}} + \xi(\sqrt{n})$ and satisfies $\frac{2}{\sqrt{n}}(e^{\epsilon} - e^{-\epsilon})$-DP.*

The proof, given in Appendix A.4, shows that the algorithm $\mathcal{A}'$ that applies $\mathcal{A}$ to a random subsample of size $\lfloor\sqrt{n}\rfloor$ is AERM for any distribution in the class $\mathfrak{D}$.

Theorem 15 (Characterization of private $\mathfrak{D}$-learnability) *A problem is privately $\mathfrak{D}$-learnable* if *there exists an algorithm that is $\mathfrak{D}$-universally AERM and differentially private with privacy loss $\epsilon(n) \to 0$. If in addition, the problem is (distribution-free and non-privately) learnable, then the converse is also true.*

Proof The "if" part is exactly the same as the argument in Section 3.1, since both Lemma 8 and Lemma 9 holds for each distribution independently. Under the additional assumption that the problem itself is learnable (distribution-free and non-privately), the "only if" part is given by Lemma 14. ■

This result may appear to be unsatisfactory due to the additional assumption of learnability. It is clearly a strong assumption because many problems that are $\mathfrak{D}$-learnable for a practically meaningful $\mathfrak{D}$ are not actually learnable. We provide one such example here.

Example 1 *Let the data space be $[0, 1]$, the hypothesis space be the class of all* finite *subset of $[0, 1]$ and the loss function $\ell(h, z) = 1_{z \notin h}$. This problem is not learnable, and not even $\mathfrak{D}$-learnable when $\mathfrak{D}$ is the class of all discrete distributions with finite number of possible values. But it is $\mathfrak{D}$-learnable when $\mathfrak{D}$ is further restricted with an upper bound on the total number of possible values.*

Proof For any discrete distribution with a finite support set, there is an $h \in \mathcal{H}$ such that the optimal risk is 0. Assume the problem is learnable with rate $\xi(n)$, then for some n $\xi(n) < 0.5$. However, we can always construct a uniform distribution over $3n$ elements and it is information-theoretically impossible for any estimators based on n samples from the distribution to achieve a risk better than $2/3$. The problem is therefore not learnable. When we assume an upper bound N on the maximum number of bins of the underlying distribution, then the ERM which outputs just the support of all observed data will be universally consistent with rate $\xi(n) = N/n$. ■

It turns out that we cannot hope to *completely* remove the assumption from Theorem 15. The following example illustrates that some form of qualification (implied by the learnability assumption) is necessary for the converse statement to be true.

Example 2 *Consider the learning problem in Example 1. Let $\mathfrak{D}$ be the class of all continuous distributions. There is a learning problem that is s privately $\mathfrak{D}$-learnable but no private AERM algorithm exists.*

Proof Let the learning problem be that in Example 1 and $\mathfrak{D}$ be the class of all continuous distributions defined on $[0, 1]$. Consider The learning algorithm $\mathcal{A}(Z)$ always returns $h = \emptyset$. The optimal risk for any continuous distribution is 1 because any finite subset is of measure 0, output $\emptyset$ is 0-consistent and 0-generalizing, but not AERM, since the minimum empirical risk is 0. $\mathcal{A}$ is also 0-differentially private, therefore the problem is privately $\mathfrak{D}$-learnable for $\mathfrak{D}$ being the set of all continuous distributions.
However, it is not privately $\mathfrak{D}$-learnable via an AERM, i.e., no private AERM algorithm exists for this problem. We prove this by contradiction. Assume an ϵ-DP AERM algorithm exists, the subsampling lemma ensures the existence of an $\epsilon(n)$-DP AERM algorithm $\mathcal{A}'$ with $\epsilon(n) \to 0$. $\mathcal{A}'$ is therefore generalizing by stability, and it follows that the $\mathcal{A}'$ has risk $\mathbb{E}_{h\sim\mathcal{A}'(Z)}R(h)$ converging to 0. But there is no $h \in \mathcal{H}$ such that $R(h) < 1$, giving the contradiction. ■

Interestingly, this problem is $\mathfrak{D}$-learnable via a non-private AERM algorithm, which always outputs $h = Z$. This is 0-consistent, 0-AERM but not generalizing. This example suggests that $\mathfrak{D}$-learnability and learnability are quite different because for learnable problems, if an algorithm is consistent and AERM, then it must also be generalizing (Shalev-Shwartz et al., 2010, Theorem 10).

3.5 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. (2010) that learns all learnable problems). This is to solve

$$\underset{\substack{(\mathcal{A},\epsilon):\\ \mathcal{A}:\mathcal{Z}^n\to\mathcal{H},\\ \mathcal{A}\text{ is }\epsilon\text{-DP}}}{\operatorname{argmin}} \left[\epsilon + \sup_{Z\in\mathcal{Z}^n}\left(\mathbb{E}_{h\sim\mathcal{A}(Z)}\hat{R}(h,Z) - \inf_{h\in\mathcal{H}}\hat{R}(h,Z)\right)\right], \tag{6}$$

or to privately $\mathfrak{D}$-learn the problem when (6) is not feasible

$$\underset{\substack{(\mathcal{A},\epsilon):\\ \mathcal{A}:\mathcal{Z}^n\to\mathcal{H},\\ \mathcal{A}\text{ is }\epsilon\text{-DP}}}{\operatorname{argmin}} \left[\epsilon + \sup_{\mathcal{D}\in\mathfrak{D}}\mathbb{E}_{Z\sim\mathcal{D}^n}\left(\mathbb{E}_{h\sim\mathcal{A}(Z)}\hat{R}(h,Z) - \inf_{h\in\mathcal{H}}\hat{R}(h,Z)\right)\right]. \tag{7}$$

Theorem 16 *Assume the problem is learnable. If the problem is private learnable,* (6) *will always output a universally consistent private learning algorithm. If the problem is private $\mathfrak{D}$-learnable,* (7) *will always output a $\mathfrak{D}$-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 (6) will have a objective value that is no greater than $\epsilon(n) + \xi(n)$ for any n. Corollary 9 concludes its universal consistency. The second claim follows from the characterization of private $\mathfrak{D}$-learnability in Theorem 15. ■

Algorithm 1 Exponential Mechanism for regularized ERM

Input: Data points $Z = \{z_1, ..., z_n\} \in \mathcal{Z}^n$, loss function ℓ, regularizer g_n, privacy parameter $\epsilon(n)$ and a hypothesis space $\mathcal{H}$.

1. Construct utility function $q(h, Z) := -\frac{1}{n}\sum_{i=1}^{n} \ell(h, z_i) - g_n(h)$, and its sensitivity $\Delta q := \sup_{h\in\mathcal{H}, d(Z,Z')=1} |q(h, Z) - q(h, Z')| \leq \frac{2}{n} \sup_{h\in\mathcal{H}, z\in\mathcal{Z}} |\ell(h, z)|$.
2. Sample $h \in \mathcal{H}$ with probability $\mathbb{P}(h) \propto \exp(\frac{\epsilon(n)}{2\Delta q} q(h, Z))$.

Output: h.

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 interest to theoretical questions related to the search of private learning algorithms and the fundamental limit of machine learning under privacy constraints.

4. Private learning for penalized ERM

Now we describe a generic and practical class of private learning algorithms, based on the idea of minimizing the empirical risk under privacy constraint:

$$\underset{h\in\mathcal{H}}{\text{minimize}}\, F(Z, h) = \frac{1}{n}\sum_{i=1}^{n} \ell(h, z_i) + g_n(h). \tag{8}$$

The first term is empirical risk and the second term vanishes as n increases so that this estimator is asymptotically ERM. The same formulation has been studied before in the context of differentially private machine learning (Chaudhuri et al., 2011; Kifer et al., 2012), but our focus is more generic and does not require the objective function to be convex, differentiable, continuous, or even have a finite dimensional Euclidean space embedding, hence covers a larger class of learning problems.

Our generic algorithm for differentially private learning is summarized in Algorithm 1. It applies the exponential mechanism (McSherry and Talwar, 2007) to penalized ERM. We note that this algorithm implicitly requires that $\int_{\mathcal{H}} \exp(\frac{\epsilon(n)}{2\Delta q} q(h, Z))dh < \infty$, otherwise the distribution is not well-defined and it does not make sense to talk about differential privacy. In general, if $\mathcal{H}$ is a compact set with a finite volume (with respect to a base measure, such as the Lebesgue measure or counting measure), then such a distribution always exists. We will revisit this point and discuss the practicality of this assumption in the Section 5.3.

Using the characterization results developed so far, we are able to give sufficient conditions for consistency of private learning algorithms without having to establish uniform convergence. Define the sublevel set as

$$\mathcal{S}_{Z,t} = \{h \in \mathcal{H} \mid F(Z, h) \leq t + \inf_{h\in\mathcal{H}} F(Z, h)\}, \tag{9}$$

where $F(h, Z)$ is the regularized empirical risk function defined in (8). In particular, we assume the following conditions:

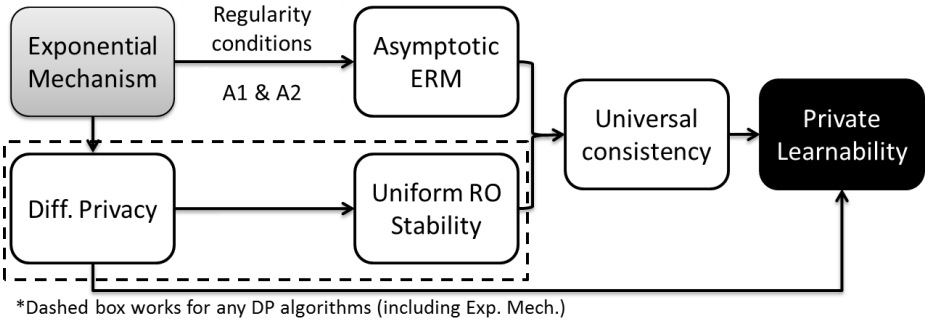


Figure 3: Illustration of Theorem 17: conditions for private learnability in general learning setting.

A1. Bounded loss function: $0 \leq \ell(h, z) \leq 1$ for any $h \in \mathcal{H}, z \in \mathcal{Z}$.

A2. Sublevel set condition: There exist constant positive integer n_0, positive real number t_0, and a sequence of regularizer g_n satisfying $\sup_{h \in \mathcal{H}} |g_n(h)| = o(n)$, such that for any $0 < t < t_0$, $n > n_0$

$$\mathbb{E}_{Z \sim \mathcal{D}^n} \left(\frac{\mu(\mathcal{H})}{\mu(\mathcal{S}_{Z,t})} \right) \leq K \left(\frac{1}{t} \right)^{\rho}, \tag{10}$$

where $K = K(n), \rho = \rho(n)$ satisfy $\log K + \rho \log n = o(n)$. Here the measure μ may depend on context, such as Lebesgue measure ($\mathcal{H}$ is continuous) or counting measure ($\mathcal{H}$ is discrete).

The first condition of boundedness is common. It is assumed in Vapnik's characterization for ERM learnability and Shalev-Shwartz et al.'s general characterization of all learnable problems. In fact, we can always consider $\mathcal{H}$ to be a sublevel set such that the boundedness condition holds. For the second condition, the intuition is that we require the sublevel set to be large enough such that the sampling procedure will return a good hypothesis with large probability. $\mu(\mathcal{S}_t)$ is a critical parameter in the utility guarantee for the exponential mechanism (McSherry and Talwar, 2007). Also, it is worth pointing out that A2 implies that the exponential distribution is well-defined.

Theorem 17 (General private learning) *Let $(\mathcal{Z}, \mathcal{H}, \ell)$ be any problem in the general learning setting. Suppose we can choose g_n such that A.1 and A.2 are satisfied with (ρ, K, g_n, n_0, t_0) for a distribution $\mathcal{D}$, then Algorithm 1 satisfies $\epsilon(n)$-privacy and is consistent with rate*

$$\xi(n) = \frac{9[\log K + (\rho + 2) \log n]}{n\epsilon(n)} + 2\epsilon(n) + \sup_{h \in \mathcal{H}} |g_n(h)|. \tag{11}$$

In particular, if $\epsilon(n) = o(1)$, $\sup_{h \in \mathcal{H}} |g_n(h)| = o(1)$ and $\log K + \rho \log n = o(n\epsilon(n))$ for all $\mathcal{D}$ (in $\mathfrak{D}$) Algorithm 1 privately learns ($\mathfrak{D}$-learns) the problem.

We give an illustration of the proof in Figure 3. The detailed proof, based on the stability argument (Shalev-Shwartz et al., 2010), is deferred to Appendix A.5.

To see that Theorem 17 actually contains a large number of problems in the general learning setting. We provide concrete examples that satisfy A1 and A2 below for both privately learnable and privately $\mathfrak{D}$-learnable problems that can be learned using Algorithm 1.

4.1 Examples of privately learnable problems

We start from a few cases where Algorithm 1 is universally consistent for all distributions.

Example 3 (Finite discrete $\mathcal{H}$) *Suppose $\mathcal{H}$ can be fully encoded by M-bits, then*

$$\mu(\mathcal{S}_t)/\mu(\mathcal{H}) \geq |\mathcal{H}|^{-1} = 2^{-M},$$

since there are at least 1 optimal hypothesis for each function and now μ is the counting measure. In other word, we can take $K = 2^M$ and $\rho = 0$ in the (11). Plug this into the expression and take $g_n \equiv 0$, $\epsilon(n) = \sqrt{(M + \log n)/n}$, we get a rate of consistency $\xi(n) = O(\frac{M+\log n}{\sqrt{n}})$. In addition, if we can find a data-independent covering set for a continuous space, then we can discretize the space and the result same results follow. This observation will be used in the construction of many private learning algorithms below.

Example 4 (Lipschitz functions/Hölder class) *Let $\mathcal{H}$ be a compact, β_p-regular subset of $\mathbb{R}^d$ satisfying $\mu(B \cap \mathcal{H}) \geq \beta_p \mu(B)$ for any ℓ_p ball $B \subset \mathbb{R}^d$ that is small enough. Assume that $F(Z, \cdot)$ is L-Lipschitz on $\mathcal{H}$: for any $h, h' \in \mathcal{H}$,*

$$|F(Z, h) - F(Z, h')| \leq L\|h - h'\|_p .$$

Then for sufficiently small t, we have Lebesgue measure

$$\mu(\mathcal{S}_t) \geq \beta_p \left(t/L\right)^d$$

and Condition A.2 holds with $K = \mu(\mathcal{H})\beta_p^{-1}L^d$, $\rho = d$. Furthermore, if we take $\epsilon(n) = \sqrt{\frac{d(\log L + \log n) + \log(\mu(\mathcal{H})/\beta_p)}{n}}$, the algorithm is $O\left(\sqrt{\frac{d(\log L+\log n)+\log(\mu(\mathcal{H})/\beta_p)}{n}} + \sup_{h\in\mathcal{H}}|g_n(h)|\right)$-consistent.

This shows that condition A2 holds for a large class of low-dimensional problems of interest in machine learning and one can learn the problem privately without actually needing to find a covering set algorithmically. Specifically, the example includes many practically used methods such as logistic regression, linear SVM, ridge regression, even multi-layer neural networks, since the loss functions in these methods are jointly bounded in (Z, h) and Lipschitz in h.

The example also raises an interesting observation that while differentially private classification is not possible in a distribution-free setting for 0-1 loss function (Chaudhuri and Hsu, 2011), it is learnable under smoother surrogate loss, e.g., logistic loss or hinge loss. In other words, private learnability and computational tractability both benefit from the same relaxation.

The Lipschitz condition still requires the dimension of the hypothesis space to be $o(n)$. Thus it does not cover high-dimensional machine learning problems where $d \gg n$, nor does it contain the example of Shalev-Shwartz et al. (2010) that ERM fails.

For high dimensional problems where d grows with n, typically some assumptions or restrictions need to be made either on the data or on the hypothesis space (so that it becomes essentially low-dimensional). We give one example here for the problem of sparse regression.

Example 5 (Best subset selection) *Consider $\mathcal{H} = \{h \in \mathbb{R}^d : \|h\|_0 < s, \|h\|_2 \leq 1\}$ and let $\ell(h, z)$ be an L-Lipschitz loss function. The solution can only be chosen from $\binom{d}{s} < d^s$ different s-dimensional subspaces. We can apply Algorithm 1 twice to first sample a support set S with utility function being the $-\min_{h \in \mathcal{H}_S} F(Z, h)$, and then sample a solution in the chosen s-dimensional subspace. By the composition theorem this two-stage procedure is differentially private. Moreover, by the arguments in Example 3 and Example 4 respectively, we have an $\mu(\mathcal{S}_t) \geq \left(\frac{1}{d}\right)^s$ for the subset selection and $\mu(\mathcal{S}_t) \geq (\frac{t}{L})^s$ for the low-dimensional regression. Note that $\rho = 0$ in both cases and the dependency on the ambient dimension d is on the logarithm. The first stage ensures that for the chosen support set $\mathcal{S}$, $\min_{h \in \mathcal{H}_S} F(Z, h)$ is close to $\min_{h \in \mathcal{H}} F(Z, h)$ by $O(\frac{s \log d + \log n}{n\epsilon(n)})$ in expectation and (the second stage ensures that the sampled hypothesis from $\mathcal{H}_S$ would have objective function close to $\min_{h \in \mathcal{H}_S} F(Z, h)$ by $O(\frac{s \log L + s \log n + \log(\mu(\mathcal{H}_S)/\beta_p)}{n\epsilon(n)})$. This leads to an overall rate of consistency (they simply add up) of $O(\frac{s(\log d + \log n + L) + \log(\mu(\mathcal{H}_S)/\beta_p)}{\sqrt{n}})$ if we choose $\epsilon(n) = 1/\sqrt{n}$.*

4.2 Examples of privately $\mathfrak{D}$-learnable problems.

For problems where private learnability is impossible to achieve, we may still apply Theorem 17 to prove the weaker private $\mathfrak{D}$-learnability for some specific class of distributions.

Example 6 (Finite Representation Dimension in the General Learning Setting) *For binary classification problems with* 0-1 *loss (PAC learning), this has been well-studied. In particular, Beimel et al. (2013a) characterized the sample complexity of privately learnable problems using a combinatorial condition they call a "Probabilistic Representation", which basically involves finding a finite, data-independent set of hypotheses to approximate any hypothesis in the class. Their claim is that if the "representation dimension" is finite, then the problem is privately learnable, otherwise it is not. We can extend the notion of probabilistic representation beyond the finite discrete and countably infinite hypothesis class considered in Beimel et al. (2013a) to cases when the problem is not privately learnable (e.g, learning threshold functions on* $[0, 1]$*). The existence of probabilistic representation for all distributions in $\mathfrak{D}$ would lead to a $\mathfrak{D}$-universally private learning algorithm.*

Another way to define a class of distribution $\mathfrak{D}$ is to assume the existence of a reference distribution that is close to any distribution of interest as in Chaudhuri and Hsu (2011).

Example 7 (Existence of a public reference distribution) *To deal with the* 0-1 *loss classification problems on a continuous hypothesis domain, Chaudhuri and Hsu (2011) assume*

that there exists a data-independent reference distribution $\mathcal{D}^*$, which by multiplying a fixed constant on its density, uniformly dominates any distributtion of interest. This essentially produces a subset of distributions $\mathfrak{D}$. The consequence is that one can build an ϵ-net of $\mathcal{H}$ with metric defined on the risk under $\mathcal{D}^*$ and this will also be a (looser) covering set of any distribution $\mathcal{D} \in \mathfrak{D}$, thereby learning the problem for any distribution in the set.

The same idea can be applied to the general learning setting. For any fixed reference distribution $\mathcal{D}^*$ defined on $\mathcal{Z}$ and constant c,

$$\mathfrak{D} = \{\mathcal{D} = (\mathcal{Z}, \mathcal{F}, \mathbb{P}) \mid \mathbb{P}_{\mathcal{D}}(z \in A) \leq c\mathbb{P}_{\mathcal{D}^*}(z \in A) \text{ for } \forall A \in \mathcal{F}\}$$

is a valid set of distributions and we are able to $\mathfrak{D}$-privately learn this problem whenever we can construct a sufficiently small cover set with respect to $\mathcal{D}^*$ and reduce the problem to Example 3. This class of problems includes high-dimensional and infinity dimensional problems such as density estimation, nonparametric regression, kernel methods and essentially any other problems that are strictly learnable (Vapnik, 1998), since they are characterized by one-sided uniform convergence (and the corresponding entropy condition).

4.3 Discussion on uniform convergence and private learnability

Uniform convergence requires that $\mathbb{E}_{Z\sim\mathcal{D}^n} \sup_{h\in\mathcal{H}} |\hat{R}(h, Z) - R(h)| \to 0$ for any distribution $\mathcal{D}$ with a distribution independent rate. Most machine learning algorithms rely on uniform convergence to establish consistency result (e.g., through complexity measure such as VC-dimension, Rademacher Complexity, covering and bracketing numbers and so on). In fact, the learnability of ERM algorithm is characterized by the one-sided uniform convergence (Vapnik, 1998), which is only slightly weaker than requiring uniform convergence on both sides.

A key point in Shalev-Shwartz et al. (2010) is that the learnability (by any algorithm) in general learning setting is no longer characterized by variants of uniform convergence. However, the class of privately learnable problems is much smaller. Clearly, uniform convergence is not sufficient for a problem to be privately learnable (see Section 3.3), but is it necessary?

In binary classification with discrete domain (agnostic PAC Learning), since VC-dimension being finite characterizes the class of privately PAC learnable problems, the necessity of uniform convergence is clear. This could also be more explicitly seen from Beimel et al. (2013a) where the *probabilistic representation dimension* is a form of uniform convergence on its own.

In the general learning setting, the problem is still open. We were not able to prove that private learnability implies uniform convergence, but we could not construct a counter example either. All our examples in this section do implicitly or explicitly uses uniform convergence, which seems to hint at a positive answer.

5. Practical concerns

5.1 High confidence private learning via boosting

We have stated all results so far in expectation. We can easily convert these to the high-confidence learning paradigm by applying Markov's inequality, since convergence in expectation to the minimum risk implies convergence in probability to the minimum risk. While the $1/\delta$ dependence on the failure probability δ is not ideal, we can apply a similar meta-algorithm "boosting"(Schapire, 1990) as in Shalev-Shwartz et al. (2010, Section 7) to get a $\log(1/\delta)$ rate. The approach is similar to cross-validation. Given a pre-chosen positive integer a, the original boosting algorithm randomly partitions the data into $(a+1)$ subsamples of size $n/(a+1)$, and applies Algorithm 1 on the first a partitions, obtaining a candidate hypotheses. The method then returns the one hypothesis with smallest validation error, calculated using the remaining subsample. To ensure differential privacy, our method instead uses the exponential mechanism to sample the best candidate hypothesis, where the logarithm of sampling probability is proportional to the negative validation error.

Theorem 18 (High-confidence private learning) *If an algorithm $\mathcal{A}$ privately learns a problem with rate $\xi(n)$ and privacy parameter $\epsilon(n)$, then the boosting algorithm $\mathcal{A}'$ with $a = \log\frac{3}{\delta}$ is $\max\left\{\epsilon\left(\frac{n}{\log(3/n)+1}\right), \frac{\log(3/\delta)+1}{\sqrt{n}}\right\}$-differentially private, its output h obeys*

$$R(h) - R^* \leq e\xi\left(\frac{n}{\log(3/\delta)+1}\right) + C\sqrt{\frac{\log(3/\delta)}{n}}$$

for an absolute constant C with probability at least $1-\delta$.

5.2 Efficient sampling algorithm for convex problems

Our proposed exponential sampling based algorithm is to establish a more explicit geometric condition upon which AERM holds, hence the algorithm may not be computationally tractable. Ignoring the difficulty of constructing the ϵ-covering set of an exponential number of elements, sampling from the set alone is not a polynomial time algorithm. But we can solve a subset of the continuous version of our Algorithm 1 described in Theorem 17 in polynomial time to arbitrary accuracy (see also Bassily et al. (2014, Theorem 3.4)).

Proposition 19 *If $n^{-1}\sum_{i=1}^n \ell(h, z_i) + g_n(h)$ is convex in h and $\mathcal{H}$ is a convex set, then the sampling procedure in Algorithm 1 can be solved in polynomial time.*

Proof When $n^{-1}\sum_{i=1}^n \ell(h, z_i) + g_n(h)$ is convex, the utility function $q(h, Z)$ is concave in h. The density to be sampled from in Algorithm 1 is proportional to $\exp(\frac{\epsilon n q(h,Z)}{B})$ and is log-concave. The Markov chain sampling algorithm in Applegate and Kannan (1991) is guaranteed to produce a sample from a distribution that is arbitrarily close to the target distribution (in the total variation sense) in polynomial time. ■

5.3 Exponential mechanism in infinite domain

As we mention earlier, the results in Section 4 based on the exponential mechanism implicitly assumes certain regularity conditions that ensures the existence of a probability distribution.

When $\mathcal{H}$ is finite, the existence is trivial. On the other hand, an infinite set $\mathcal{H}$ is tricky in that there may not exist a proper distribution that satisfies $\mathbb{P}(h) \propto e^{\frac{\epsilon}{2\Delta q} q(Z,h)}$ for at least some $q(Z,h)$. For instance, if $\mathcal{H} = \mathbb{R}$ and $q(Z,h) \equiv 1$ then $\int_{\mathbb{R}} e^{\frac{\epsilon}{2\Delta q} q(Z,h)} dh = \infty$. Such distributions that are only defined up to scale with no finite normalization constants are called improper distributions. In case of finite dimensional non-compact set, this translates into an additional assumption on the loss function and the regularization term.

Things get even trickier when $\mathcal{H}$ is an infinite dimensional space, such as a subset of a Hilbert space. While probability measures can still be defined, no density function can be defined on such spaces. Therefore, we cannot use exponential mechanism to define a valid probability distribution.

The practical implication is that exponential mechanism is really only applicable to cases when the hypothesis space $\mathcal{H}$ allows for definitions of densities in the usual sense, or then $\mathcal{H}$ can be approximated by such a space. For example, a separable Hilbert space can be studied by finite-dimensional projections. Also, we can approximate RKHS induced by translation invariant kernels via random Fourier features (Rahimi and Recht, 2007).

6. Results for learnability under (ϵ, δ)-differential privacy

Another way to weaken the definition of private learnability is through (ϵ, δ)-approximate differential privacy.

Definition 20 (Dwork et al., 2006a) *An algorithm $\mathcal{A}$ obeys (ϵ, δ)-differential privacy if for any Z, Z' such that $d(Z, Z') \leq 1$, and for any measurable set $\mathcal{S} \subset \mathcal{H}$*

$$\mathbb{P}_{h \sim \mathcal{A}(Z)}(h \in \mathcal{S}) \leq e^{\epsilon} \mathbb{P}_{h \sim \mathcal{A}(Z')}(h \in \mathcal{S}) + \delta.$$

We define a version of the problem to be

Definition 21 (Approximately Private Learnability) *We say a learning problem is $\Delta(n)$-approximately privately learnable for some pre-specified family of rate $\Delta(n)$ if for some $\epsilon < \infty$, $\delta(n) \in \Delta(n)$, there exists a universally consistent algorithm that is $(\epsilon, \delta(n))$-DP.*

This is a completely different subject to study and the class of approximately privately learnable problems could be substantially larger than the pure privately learnable problems. Moreover, the picture may vary with respect to how small $\delta(n)$ is required to be. In this section, we present our preliminary investigation on this problem.

Specifically, we will consider two questions:

1. Does the existence of an (ϵ, δ)-DP always AERM algorithm characterize the class of approximately private learnable problems?

2. Are all learnable problems approximately privately learnable for different choices of $\Delta(n)$?

The minimal requirement in the same flavor of Definition 3 would be to require $\Delta(n) = \{\delta(n)|\delta(n) \to 0\}$. The learnability problem turns out to be trivial under this definition due to the following observation.

Lemma 22 *For any algorithm $\mathcal{A}$ that acts on Z, $\mathcal{A}'$ that runs $\mathcal{A}$ on a randomly chosen subset of Z of size $\sqrt{n}$ is $(0, \frac{1}{\sqrt{n}})$-DP.*

Proof Let Z and Z' be adjacent datasets that differs only in data point i. For any i and any $S \in \sigma(\mathcal{H})$.

$$\begin{aligned}
\mathbb{P}(\mathcal{A}'(Z) \in S) &= \mathbb{P}_I(\mathcal{A}(Z_I) \in S | i \in I)\mathbb{P}(i \in I) + \mathbb{P}_I(\mathcal{A}(Z_I) \in S | i \notin I)\mathbb{P}(i \notin I) \\
&= \mathbb{P}_I(\mathcal{A}(Z_I) \in S | i \in I)\mathbb{P}(i \in I) + \mathbb{P}_I(\mathcal{A}(Z'_I) \in S | i \notin I)\mathbb{P}(i \notin I) \\
&= \mathbb{P}(\mathcal{A}'(Z') \in S) + [\mathbb{P}_I(\mathcal{A}(Z_I) \in S | i \in I) - \mathbb{P}_I(\mathcal{A}'(Z_I) \in S | i \in I)]\mathbb{P}(i \in I) \\
&\leq \mathbb{P}(\mathcal{A}'(Z') \in S) + \mathbb{P}(i \in I) \\
&= e^0 \mathbb{P}(\mathcal{A}'(Z') \in S) + \frac{1}{\sqrt{n}}.
\end{aligned}$$

This verifies the $(0, 1/\sqrt{n})$-DP of algorithm $\mathcal{A}'$. ■

The above lemma suggests that if $\delta(n) = o(1)$ is all we need for the *approximately private learnability*, then any consistent learning algorithm can be made approximately DP by simply subsampling. In other words, any learnable problem is also learnable under approximate differential privacy.

To get around this triviality, we need to specify a sufficiently fast rate of $\delta(n)$ going to 0. While it is common to require that $\delta(n) = o(1/\text{poly}(n))$ [2] for cryptographically strong privacy protection, requiring $\delta(n) = o(1/n)$ is already enough to invalidate the above subsampling argument and makes the problem of learnability a non-trivial one.

Again, the question is whether AERM characterizes approximately private learnability and whether there is a gap between the class of learnable and approximately privately learnable problems.

Here we show that the "folklore" Lemma 8 and subsampling lemma (Lemma 27) can be extended to work with (ϵ, δ)-DP and then we provide a positive answer to the first question.

Lemma 23 (Stability of (ϵ, δ)-DP) *If $\mathcal{A}$ is (ϵ, δ)-DP, and $0 \leq \ell(h, z) \leq 1$, then $\mathcal{A}$ is $(e^\epsilon - 1 + \delta)$-Strongly Uniform RO-stable.*

2. Here the notation "$o(1/\text{poly}(n))$" means "decays faster than any polynomial of n". A sequence $a(n) = o(1/\text{poly}(n))$ if and only if $a(n) = o(n^{-r})$ for any $r > 0$.

Proof For any Z, Z' such that $d(Z, Z') \leq 1$ and for any $z \in \mathcal{Z}$. Let the event $E = \{h | p(h) \geq p'(h)\}$,

$$\begin{aligned}
&\left|\mathbb{E}_{h\sim\mathcal{A}(Z)}\ell(h,z) - \mathbb{E}_{h\sim\mathcal{A}(Z')}\ell(h,z)\right| = \left|\int_h \ell(h,z)p(h)dh - \int_h \ell(h,z)p'(h)dh\right| \\
\leq &\sup_{h,z}\ell(h,z)\int_E p(h) - p'(h)dh \leq \int_E p(h) - p'(h)dh = \mathbb{P}_{h\sim\mathcal{A}(Z)}(h \in E) - \mathbb{P}_{h\sim\mathcal{A}(Z')}(h \in E) \\
\leq &(e^\epsilon - 1)\mathbb{P}_{h\sim\mathcal{A}(Z')}(h \in E) + \delta \leq e^\epsilon - 1 + \delta.
\end{aligned}$$

The last line applies the definition of (ϵ, δ)-DP. ■

Lemma 24 (Subsampling Lemma of (ϵ, δ)-DP) *If $\mathcal{A}$ is (ϵ, δ)-DP, then $\mathcal{A}'$ that acts on a random subsample of Z of size γn obeys (ϵ', δ')-DP with $\epsilon' = \log(1 + \gamma e^\epsilon(e^\epsilon - 1))$ and $\delta' = \gamma e^\epsilon \delta$.*

Proof For any event $E \in \sigma(\mathcal{H})$, let i be the coordinate where Z and Z' differs

$$\begin{aligned}
&\mathbb{P}_{h\sim\mathcal{A}'(Z)}(h \in E) = \gamma\mathbb{P}_{h\sim A(Z_I)}(h \sim E | i \in I) + (1-\gamma)\mathbb{P}_{h\sim A(Z_I)}(h \sim E | i \notin I) \\
=&\gamma\mathbb{P}_{h\sim A(Z_I)}(h \sim E | i \in I) + (1-\gamma)\mathbb{P}_{h\sim A(Z'_I)}(h \sim E | i \notin I) \\
=&\gamma\mathbb{P}_{h\sim A(Z_I)}(h \sim E | i \in I) - \gamma\mathbb{P}_{h\sim A(Z'_I)}(h \sim E | i \in I) + \gamma\mathbb{P}_{h\sim A(Z'_I)}(h \sim E | i \in I) \\
&+ (1-\gamma)\mathbb{P}_{h\sim A(Z'_I)}(h \sim E | i \notin I) \\
=&\mathbb{P}_{h\sim\mathcal{A}'(Z')}(h \in E) + \gamma[\mathbb{P}_{h\sim A(Z_I)}(h \sim E | i \in I) - \mathbb{P}_{h\sim A(Z'_I)}(h \sim E | i \in I)] \\
\leq &\mathbb{P}_{h\sim\mathcal{A}'(Z')}(h \in E) + \gamma(e^\epsilon - 1)\mathbb{P}_{h\sim A(Z'_I)}(h \sim E | i \in I) + \gamma\delta, \qquad (12)
\end{aligned}$$

where in last line, we apply (ϵ, δ)-DP of $\mathcal{A}$.

It remains to show that $\mathbb{P}_{h\sim A(Z'_I)}(h \sim E | i \in I)$ is similar to $\mathbb{P}_{h\sim\mathcal{A}'(Z')}(h \in E)$. First,

$$\mathbb{P}_{h\sim\mathcal{A}'(Z')}(h \in E) = \gamma\mathbb{P}_{h\sim\mathcal{A}(Z'_I)}(h \in E | i \in I) + (1-\gamma)\mathbb{P}_{h\sim\mathcal{A}(Z'_I)}(h \in E | i \notin I). \qquad (13)$$

Denote $\mathcal{I}_1 = \{I | i \in I\}$, $\mathcal{I}_2 = \{I | i \notin I\}$. We known $|\mathcal{I}_1| = \binom{n-1}{\gamma n - 1}$, and $|\mathcal{I}_2| = \binom{n-1}{\gamma n}$ and $|\mathcal{I}_1|/|\mathcal{I}_2| = \gamma n/(n - \gamma n)$. For every $I \in \mathcal{I}_2$ there are precisely γn elements $J \in \mathcal{I}_1$ such that $d(I, J) = 1$. Likewise, for every $J \in \mathcal{I}_1$, there are $n - \gamma n$ elements $I \in \mathcal{I}_2$ such that $d(I, J) = 1$. It follows by symmetry that if we apply (ϵ, δ)-DP to $1/\gamma n$ of each $I \in \mathcal{I}_2$ and change I to their corresponding $J \in \mathcal{I}_1$, then each $J \in \mathcal{I}_1$ will receive $(n - \gamma n)/\gamma n$ "contribution" in total from the sum over all $I \in \mathcal{I}_2$.

$$\begin{aligned}
&\mathbb{P}_{h\sim\mathcal{A}(Z'_I)}(h \in E | i \notin I) = \frac{1}{|\mathcal{I}_2|}\sum_{I\in\mathcal{I}_2}\mathbb{P}_{h\sim\mathcal{A}(Z'_I)}(h \in E) \\
=&\frac{1}{|\mathcal{I}_2|}\sum_{I\in\mathcal{I}_2}\sum_{j=1}^{\gamma n}\frac{1}{\gamma n}\mathbb{P}_{h\sim\mathcal{A}(Z'_I)}(h \in E) \\
\geq &\frac{|\mathcal{I}_1|}{|\mathcal{I}_2|}\frac{1}{|\mathcal{I}_1|}\sum_{J\in\mathcal{I}_1}\frac{n-\gamma n}{\gamma n}e^{-\epsilon}(\mathbb{P}_{h\sim\mathcal{A}(Z'_J)}(h \in E) - \delta) \\
=&\frac{1}{|\mathcal{I}_1|}\sum_{J\in\mathcal{I}_1}e^{-\epsilon}(\mathbb{P}_{h\sim\mathcal{A}(Z'_J)}(h \in E) - \delta) = e^{-\epsilon}\mathbb{P}_{h\sim\mathcal{A}(Z'_I)}(h \in E | i \in I) - e^{-\epsilon}\delta
\end{aligned}$$

Substitute into (13), we get

$$\mathbb{P}_{h\sim\mathcal{A}(Z'_I)}(h \in E | i \in I) \leq \frac{1}{\gamma + (1-\gamma)e^{-\epsilon}} \mathbb{P}_{h\sim\mathcal{A}'(Z')}(h \in E) + \frac{(1-\gamma)e^{-\epsilon}}{\gamma + (1-\gamma)e^{-\epsilon}}\delta.$$

We further relax the upper bound to a simple form $e^{\epsilon}\mathbb{P}_{h\sim\mathcal{A}'(Z')}(h \in E) + \delta$ and substitute into (12), we have

$$\mathbb{P}_{h\sim\mathcal{A}'(Z)}(h \in E) \leq (1 + \gamma e^{\epsilon}(e^{\epsilon} - 1))\mathbb{P}_{h\sim\mathcal{A}'(Z')}(h \in E) + \gamma\delta + \gamma(e^{\epsilon} - 1)\delta,$$

which concludes the proof. ■

Using the above two lemmas, we are able to establish the same result which says that AERM characterizes the approximate private learnability for certain classes of $\Delta(n)$.

Theorem 25 *A problem is $\Delta(n)$-approximately privately learnable implies that there exists an always AERM algorithm that is $(\epsilon(n), n^{-1/2}e^{\epsilon}\delta(\sqrt{n}))$-DP for some $\epsilon(n) \to 0$ and $\delta(\sqrt{n}) \in \Delta(n)$. The converse is also true if $n^{-1/2}e^{\epsilon}\delta(\sqrt{n}) \in \Delta(n)$.*

Proof If we have an always AERM algorithm with $\xi_{erm}(n)$ that is $(\epsilon(n), \delta(n))$-DP for $\delta(n) \in \Delta(n)$. Then by Lemma 23, this algorithm is strongly uniform RO-stable with rate $e^{\epsilon(n)} - 1 + \delta(n)$. By Theorem 28, the algorithm is universally consistent with rate $\xi_{erm}(n) + e^{\epsilon(n)} - 1 + \delta(n)$. This establishes the "if" part.

To see the "only if" part, by definition if a problem is $\Delta(n)$-approximately privately learnable with ϵ and $\delta(n) \in \Delta(n)$. Then by Lemma 24 with $\gamma = 1/\sqrt{n}$, we get an algorithm that obeys the privacy condition. It remains to prove always AERM, which requires exactly the same arguments in the proof of Lemma 10. Details are omitted. ■

Note that the results above suggest that in the two canonical settings $\Delta(n) = o(1/n)$ or $\Delta(n) = o(1/\text{poly}(n))$, existence of a private AERM algorithm that satisfies the stronger constraint $\epsilon(n) = o(1)$ characterizes the learnability.

The next question that whether any learnable problems are also approximately privately learnable would depend on how fast $\delta(n)$ is required to decay. We know that when we only have $\Delta(n) = o(1)$, all learnable problems are approximately privately learnable, and when we have $\Delta(n) = \{0\}$, only a strict subset of these problems is privately learnable. The following result establishes that when $\delta(n)$ needs to go to 0 with a sufficiently fast rate, there is separation between learnability and approximately private learnability.

Proposition 26 *Let $\Delta(n) = \{\delta(n) | \delta(n) \leq \tilde{\delta}(n)\}$ for some sequence $\tilde{\delta}(n) \to 0$. The following statements are true.*

- *All learnable problems are $\Delta(n)$-approximately privately learnable, if $\tilde{\delta}(n) = \omega(1/n)$.*
- *There exists a problem that is learnable but not $\Delta(n)$-approximately privately learnable, if $\tilde{\delta}(n) \leq \frac{\exp(-\epsilon(n^2)n^2)}{n}$*

Proof The first claim follows from the same argument in Lemma 22. If a problem is learnable, there exists a universally consistent learning algorithm $\mathcal{A}$. The algorithm that

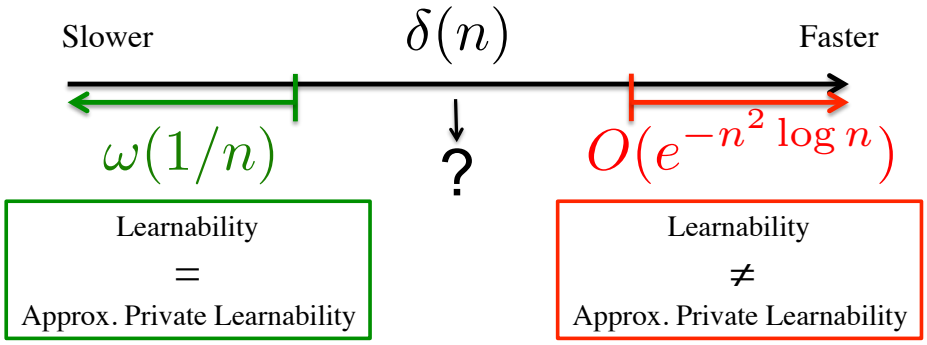


Figure 4: Illustration of Proposition 26 and the open problem.

applies $\mathcal{A}$ on a $\tilde{\delta}(n)$-fraction random subsample of the dataset is $(0, \tilde{\delta}(n))$-DP and universally consistent with rate $\xi(n\tilde{\delta}(n))$. Since $\tilde{\delta}(n) = \omega(1/n)$, $n\tilde{\delta}(n) \to \infty$.

We now show that when we require a fast decaying $\delta(n)$, then suddenly the example in Section 3.3 due to Chaudhuri and Hsu (2011) becomes not approximately privately learnable even for (ϵ, δ)-DP. Let Z, Z' be two completely different data sets, by repeatedly applying the definition of (ϵ, δ)-DP, for any set $\mathcal{S} \subset \mathcal{H}$

$$\mathbb{P}(\mathcal{A}(Z) \in \mathcal{S}) \leq e^{n\epsilon}\mathbb{P}(\mathcal{A}(Z) \in \mathcal{S}) + \sum_{i=1}^{n} e^{(i-1)\epsilon}\delta \leq e^{n\epsilon}\mathbb{P}(\mathcal{A}(Z') \in \mathcal{S}) + ne^{(n-1)\epsilon}\delta.$$

When we shift the inequality around, we get

$$\mathbb{P}(\mathcal{A}(Z') \in \mathcal{S}) \leq e^{-n\epsilon}\mathbb{P}(\mathcal{A}(Z') \in \mathcal{S}) - e^{-\epsilon}n\delta.$$

Consider the same example in Section 3.3 where we hope to learn a threshold on $[0, 1]$. Assuming there exists an algorithm $\mathcal{A}$ that is universally AERM and $(\epsilon(n), \delta(n))$-DP for $\epsilon(n) < \infty$ and $\delta(n) \leq 0.4ne^{-\epsilon n}$.

Everything up to (4) remains exactly the same. Now, apply the above implication of (ϵ, δ)-DP, we can replace (4) for each $i = 2, ..., K$, by

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

Then (5) becomes

$$\mathbb{P}(\mathcal{A}(Z_1) \notin [h_1 - \eta/3, h_1 + \eta/3]) \geq K\exp(-\epsilon n)0.9 - Ke^{-\epsilon}n\delta(n) \geq 0.9 \geq 0.5,$$

where the last inequality follows by $K > \exp(\epsilon n)$ and $\delta(n) \leq 0.4ne^{-\epsilon n}$. This yields the same contradiction to always AERM of $\mathcal{A}$ on Z_1, which requires $\mathbb{P}(\mathcal{A}(Z_1) \notin [h_1 - \eta/3, h_1 + \eta/3]) < 0.1$. Therefore, such AERM does not exist. By the contrapositive of Theorem 25, the problem is not approximately privately learnable for $\tilde{\delta}(n) \leq \frac{\exp(-\epsilon(n^2)n^2)}{n}$. ■

The bound can be further improved to $\exp(-\epsilon(n)n)/n$ if we directly work with universal consistency on various distributions rather than through always AERM on specific data points. Even that is likely to be suboptimal as there might be more challenging problems and less favorable packings to consider.

The point of this exposition, however, is to illustrate that (ϵ, δ)-DP alone does not close the gap between learnability and private learnability. Additional relaxation on the specified rate of decay on δ does. We now know that the phase transition occurs when $\delta(n)$ is somewhere between $\Omega(\exp(-n^2 \log n))$ and $O(1/n)$; but there is still a substantial gap between the upper and lower bounds. [3].

7. Conclusion and future work

In this paper, we revisited the question *"What can we learned privately?"* and considered a 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. This allows us to construct a conceptual procedure that privately learns any privately learnable problem. We also propose a relaxed notion of private learnability called private $\mathfrak{D}$-learnability, which requires the existence of an algorithm that is consistent for any the distribution within a class of distributions $\mathfrak{D}$. We characterized private $\mathfrak{D}$-learnability too with a weaker notion of AERM. For problems that can be formulated as penalized empirical risk minimization, we provide a sampling algorithm with a set of meaningful sufficient conditions on the geometry of the hypothesis space and demonstrate that it covers a large class of problems. In addition, we further extended the characterization to learnability under (ϵ, δ)-differential privacy and provided a preliminary analysis which establishes the existence of a phase transition from all learnable problems being approximately private learnable to some learnable problems being not approximately private learnable at some non-trivial rate of decay on $\delta(n)$.

Future work includes understanding the conditions under which privacy and AERM are contradictory (recall that we only have one example on learning thresholding functions due to Chaudhuri and Hsu 2011), characterizing the rate of convergence, searching for practical algorithms that generically learns all privately learnable problems, and better understanding the gap between learnability and approximate private learnability.

3. After the paper was accepted for publication, we became aware that the phase transition occurs sharply at $O(1/n)$. The result follows from a sharp lower bound of sample complexity in learning threshold functions in Bun (2016, Theorem 4.5.2), which improves over a previously published result that requires $O(n^{-1-\alpha})$ for any $\alpha > 0$ in Bun et al. (2015). The consequence is that the general learning setting is hard for (ϵ, δ)-DP too unless δ becomes meaninglessly large for privacy purposes.

Acknowledgment

We thank the AE and the anonymous reviewers for their comments that lead to significant improvement of this paper. The research was partially supported by NSF Award BCS-0941518 to the Department of Statistics at Carnegie Mellon University, and a grant by Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office.

Appendix A. Proofs of technical results

In this appendix, we provide detailed proofs to the technical results that in the main text.

A.1 Privacy in subsampling

Proof [Proof of Lemma 4] Let $\mathcal{A}$ be the consistent ϵ-DP algorithm. Consider $\mathcal{A}'$ that apply $\mathcal{A}$ to a random subsample of $\lfloor\sqrt{n}\rfloor$ data points. By Lemma 27 with $\gamma = \frac{\lfloor\sqrt{n}\rfloor}{n} \leq \frac{1}{\sqrt{n}}$, we get the privacy claim. For the consistency claim, note that the given sample is an iid sample of size $\sqrt{n}$ from the original distribution. ■

Lemma 27 (Subsampling theorem) *If Algorithm $\mathcal{A}$ is ϵ-DP for $Z \in \mathcal{Z}^n$ for any $n = 1, 2, 3, ...$, then the algorithm $\mathcal{A}'$ that output the result of $\mathcal{A}$ to a random subsample of size γn data points preserves $2\gamma(e^\epsilon - e^{-\epsilon})$-DP.*

Proof [Proof of Lemma 27 (Subsampling theorem)] This is a corollary of Lemma 4.4 in Beimel et al. (2014). To be self-contained, we reproduce the proof here in our notation.

Recall that $\mathcal{A}'$ is the algorithm that first randomly subsample γn data points then apply $\mathcal{A}$. Let Z and Z' be any neighboring databases and assume they differ on the ith data point. Let $\mathcal{S} \subset [n]$ be the indices of the random subset of the entries that are selected, and $\mathcal{R} \subset [n]\backslash\{i\}$ be a index size of size $\gamma n - 1$. We apply the law of total expectation twice and argue that for any adjacent Z, Z', any event $E \subset \mathcal{H}$,

$$\frac{\mathbb{P}_{h\sim\mathcal{A}'(Z)}(h \in E)}{\mathbb{P}_{h\sim\mathcal{A}'(Z')}(h \in E)} = \frac{\gamma\mathbb{P}_{h\sim\mathcal{A}(Z_S)}(h \in E | i \in \mathcal{S}) + (1-\gamma)\mathbb{P}_{h\sim\mathcal{A}(Z_S)}(h \in E | i \notin \mathcal{S})}{\gamma\mathbb{P}_{h\sim\mathcal{A}(Z'_S)}(h \in E | i \in \mathcal{S}) + (1-\gamma)\mathbb{P}_{h\sim\mathcal{A}(Z'_S)}(h \in E | i \notin \mathcal{S})}$$
$$= \frac{\sum_{\mathcal{R}\in[n]\backslash\{i\}} \mathbb{P}(\mathcal{R}) \left[\gamma\mathbb{P}_{h\sim\mathcal{A}(Z_S)}\left(h \in E | \mathcal{S} = \mathcal{R}\cup\{i\}\right) + (1-\gamma)\mathbb{P}_{h\sim\mathcal{A}(Z_S)}\left(h \in E | \mathcal{S} = \mathcal{R}\cup\{j\}, j \neq i\right)\right]}{\sum_{\mathcal{R}\in[n]\backslash\{i\}} \mathbb{P}(\mathcal{R}) \left[\gamma\mathbb{P}_{h\sim\mathcal{A}(Z'_S)}\left(h \in E | \mathcal{S} = \mathcal{R}\cup\{i\}\right) + (1-\gamma)\mathbb{P}_{h\sim\mathcal{A}(Z'_S)}\left(h \in E | \mathcal{S} = \mathcal{R}\cup\{j\}, j \neq i\right)\right]}$$

By the given condition that $\mathcal{A}$ is ϵ-DP, we can replace $\mathcal{R}\cup\{i\}$ with $\mathcal{R}\cup\{j\}$ for an arbitrary j with bounded changes in the probability and the above likelihood ratio can be upper bounded by

$$\frac{(\gamma e^\epsilon + 1 - \gamma)\mathbb{E}_{\mathcal{R}\in[n]\backslash\{i\}, j\neq i}\mathbb{P}_{h\sim\mathcal{A}(Z_\mathcal{S})}(h \in E | \mathcal{S} = \mathcal{R}\cup\{j\})}{(\gamma e^{-\epsilon} + 1 - \gamma)\mathbb{E}_{\mathcal{R}\in[n]\backslash\{i\}, j\neq i}\mathbb{P}_{h\sim\mathcal{A}(Z_\mathcal{S})}(h \in E | \mathcal{S} = \mathcal{R}\cup\{j\})} = \frac{\gamma e^\epsilon + 1 - \gamma}{\gamma e^{-\epsilon} + 1 - \gamma} = \frac{1 + \gamma(e^\epsilon - 1)}{1 + \gamma(e^{-\epsilon} - 1)}.$$

By definition, the privacy loss of the algorithm $\mathcal{A}'$ is therefore

$$\epsilon' \leq \log\left(1+\gamma[e^{\epsilon}-1]\right) - \log\left(1+\gamma\left[e^{-\epsilon}-1\right]\right).$$

Note that $\epsilon > 0$ implies that $-1 \leq e^{-\epsilon} - 1 < 0$ and $0 < e^{\epsilon} - 1 < \infty$. The result follows by applying the property of the natural logarithm:

$$\log(1+x) \leq \frac{x}{2}\frac{2+x}{1+x} \leq x \qquad \text{for } 0 \leq x < \infty$$
$$\log(1+x) \geq \frac{x}{2}\frac{2+x}{1+x} \geq \frac{x}{1+x} \qquad \text{for } -1 \leq x \leq 0$$

to upper bound the expression. ■

A.2 Characterization of private learnability

Privacy implies stability Lemma 8 says that an ϵ-differentially private algorithm is $(e^{\epsilon}-1)$-stable (and also 2ϵ-stable if $\epsilon < 1$).

Proof [Proof of Lemma 8] Construct Z' by replacing an arbitrary data point in Z with z' and let the probability density/mass defined by $\mathcal{A}(Z)$ and $\mathcal{A}(Z')$ be $p(h)$ and $p'(h)$ respectively, then we can bound the stability as follows

$$\begin{aligned}
&\left|\mathbb{E}_{h\sim\mathcal{A}(Z)}\ell(h,z) - \mathbb{E}_{h\sim\mathcal{A}(Z')}\ell(h,z)\right| \\
=&\left|\int_h \ell(h,z)p(h)dh - \int_h \ell(h,z)p'(h)dh\right| = \left|\int_h \ell(h,z)(p(h)-p'(h))dh\right| \\
\leq&\sup_{h,z}|\ell(h,z)| \int_{p(h)\geq p'(h)} p(h)-p'(h)dh \leq 1\cdot\int_{p(h)\geq p'(h)} p'(h)(\frac{p(h)}{p'(h)}-1)dh \\
\leq&(e^{\epsilon}-1)\int_{p(h)\geq p'(h)} p'(h)dh \leq (e^{\epsilon}-1).
\end{aligned}$$

For $\epsilon < 1$ we have $\exp(\epsilon) - 1 < 2\epsilon$.

■

Stability + AERM ⇒ consistency

Theorem 28 (Randomized version of Shalev-Shwartz et al. 2010, Theorem 8) *If any algorithm is $\xi_1(n)$-stable and $\xi_2(n)$-AERM then it is consistent with rate $\xi(n) = \xi_1(n) + \xi_2(n)$.*

Proof

We will show the following the two steps as in Shalev-Shwartz et al. (2010)

1. Uniform RO stability $\Rightarrow$ On average stability $\Leftrightarrow$ On average generalization
2. AERM + On average generalization $\Rightarrow$ consistency

The definition of these quantities is self-explanatory.

To show that "stability implies generalization", we have

$$\begin{aligned}
&\left|\mathbb{E}_{Z\sim\mathcal{D}^n}\left(\mathbb{E}_{h\sim\mathcal{A}(Z)}R(h)-\mathbb{E}_{h\sim\mathcal{A}(Z)}\hat{R}(h,Z)\right)\right| \\
=&\left|\mathbb{E}_{Z\sim\mathcal{D}^n}\left(\mathbb{E}_{z\sim\mathcal{D}}\mathbb{E}_{h\sim\mathcal{A}(Z)}\ell(h,z)-\frac{1}{n}\mathbb{E}_{h\sim\mathcal{A}(Z)}\sum_{i=1}^{n}\ell(h,z_i)\right)\right| \\
=&\left|\mathbb{E}_{Z\sim\mathcal{D}^n,\{z_1',\ldots,z_n'\}\sim\mathcal{D}^n}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{h\sim\mathcal{A}(Z)}\ell(h,z_i')-\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{h\sim\mathcal{A}(Z^{(i)})}\ell(h,z_i')\right)\right| \\
\leq&\sup_{Z,Z^{(i)}\in\mathcal{Z}^n,d(Z,Z^{(i)})=1,z'\in\mathcal{Z}}\left|\mathbb{E}_{h\sim\mathcal{A}(Z)}\ell(h,z')-\mathbb{E}_{h\sim\mathcal{A}(Z^{(i)})}\ell(h,z')\right|\leq\xi_1(n)\,,
\end{aligned}$$

where $Z^{(i)}$ is obtained by replacing the ith entry of Z with z_i'. Next, we show that "generalization and AERM implies consistency". Let $h^*\in\arg\inf_{h\in\mathcal{H}}R(h)$. By definition, we have $\mathbb{E}_{Z\sim\mathcal{D}^n}\hat{R}(h^*,Z)=R^*$. It follows that

$$\begin{aligned}
&\mathbb{E}_{Z\sim\mathcal{D}^n}[\mathbb{E}_{h\in\mathcal{A}(Z)}R(h)-R^*]=\mathbb{E}_{Z\sim\mathcal{D}^n}[\mathbb{E}_{h\in\mathcal{A}(Z)}R(h)-\hat{R}(h^*,Z)] \\
=&\,\mathbb{E}_{Z\sim\mathcal{D}^n}[\mathbb{E}_{h\in\mathcal{A}(Z)}R(h)-\mathbb{E}_{h\in\mathcal{A}(Z)}\hat{R}(h,Z)]+\mathbb{E}_{Z\sim\mathcal{D}^n}[\mathbb{E}_{h\in\mathcal{A}(Z)}\hat{R}(h,Z)-\hat{R}(h^*,Z)] \\
\leq&\,\mathbb{E}_{Z\sim\mathcal{D}^n}[\mathbb{E}_{h\in\mathcal{A}(Z)}R(h)-\mathbb{E}_{h\in\mathcal{A}(Z)}\hat{R}(h,Z)]+\mathbb{E}_{Z\sim\mathcal{D}^n}[\mathbb{E}_{h\in\mathcal{A}(Z)}\hat{R}(h,Z)-\hat{R}^*(Z)] \\
\leq&\,\xi_1(n)+\xi_2(n)\,.
\end{aligned}$$

■

Privacy + AERM $\Rightarrow$ consistency **Proof** [Proof of Corrollary 9] It follows by combining Lemma 8 and Theorem 28. ■

Necessity **Proof** [Proof of Lemma 10] We construct an algorithm $\mathcal{A}'$ by subsampling the data points using a random subset of $\sqrt{n}$ and then running $\mathcal{A}$. The privacy claim follows from Lemma 27 directly.

To prove the "always AERM" claim, we adapt the proof of Lemma 24 in Shalev-Shwartz et al. (2010). For any fixed data set $Z\in\mathcal{Z}^n$,

$$\begin{aligned}
\hat{R}(\mathcal{A}'(Z),Z)-\hat{R}^*(Z)&=\mathbb{E}_{Z'\subset Z,|Z'|=\lfloor\sqrt{n}\rfloor}\left[\hat{R}(\mathcal{A}(Z'),Z)-\hat{R}^*(Z)\right] \\
&=\mathbb{E}_{Z'\sim\mathrm{Unif}(Z)^{\lfloor\sqrt{n}\rfloor}}\left[\hat{R}(\mathcal{A}(Z'),Z)-\hat{R}^*(Z)|\text{ no duplicates}\right] \\
&\leq\frac{\mathbb{E}_{Z'\sim\mathrm{Unif}(Z)^{\lfloor\sqrt{n}\rfloor}}\left[\hat{R}(\mathcal{A}(Z'),Z)-\hat{R}^*(Z)\right]}{\mathbb{P}(\text{no duplicates})},
\end{aligned}$$

where Unif(Z) is the uniform distribution defined on the n points in Z. We need to condition on the event that there are no duplicates for the second equality to hold because Z' is a subsample taken without replacements. The last inequality is by the law of total expectation and the non-negativity of the conditional expectation. But $\mathbb{P}(\text{no duplicates}) = \prod_{i=0}^{\lfloor\sqrt{n}\rfloor-1}(1-i/n) \geq 1 - \sum_{i=0}^{\lfloor\sqrt{n}\rfloor-1} i/n \geq 1/2$. By universal consistency, $\mathcal{A}$ is consistent on the discrete uniform distribution defined on Z, so

$$\hat{R}(\mathcal{A}'(Z), Z) - \hat{R}^*(Z) \leq 2\mathbb{E}_{Z'\sim \text{Unif}(Z)^{\lfloor n\rfloor}}\left[\hat{R}(\mathcal{A}(Z'), Z) - \hat{R}^*(Z)\right] \leq 2\xi(\sqrt{n}).$$

It is obvious that $\mathcal{A}'$ is consistent with rate $\sqrt{n}$ as it applies $\mathcal{A}$ on a random sample of size $\sqrt{n}$. By Lemma 4, $\mathcal{A}'$ is $2n^{-1/2}(e^\epsilon - e^{-\epsilon})$ differentially private. By Corollary 9, the new algorithm $\mathcal{A}'$ is universally consistent. ■

A.3 Proofs for Section 3.3

Proof [Proof of Proposition 11] If $\mathcal{A}(Z)$ is a continuous distribution, we can pick $h \in \mathcal{H}$ at any point where $\mathcal{A}(Z)$ has finite density and set $\mathcal{A}'(Z)|z \in Z$ to be h with probability $1/n$ and the same as $\mathcal{A}(Z)$ with probability $1 - 1/n$. This breaks privacy because conditioned on two databases with z or without z, $\mathcal{A}$, the probability ratio of outputting h is ∞.

If $\mathcal{A}(Z)$ is a discrete distribution or a mixed distribution, it must have the same support of the point mass for all Z. Otherwise it violates DP because we need $\frac{\mathbb{P}_{h\in\mathcal{A}(Z)}(h)}{\mathbb{P}_{h\in\mathcal{A}(Z')}} \leq \exp(n\epsilon)$ for any $Z, Z' \in \mathcal{Z}^n$. Specifically, let the discrete set of point mass be $\tilde{\mathcal{H}}$ if $\mathcal{H}\backslash\tilde{\mathcal{H}} \neq \emptyset$, then we can use the same technique as in the continuous case by adding a small probability $1/n$ on $\mathcal{H}\backslash\tilde{\mathcal{H}}$ when $z \in Z$.

If $\tilde{\mathcal{H}} = \mathcal{H}$, then $\mathcal{H}$ is a discrete set, if $|\mathcal{H}| < n$, then by boundedness and Hoeffding, ERM is a deterministic algorithm that learns any learnable problem. On the other hand, if $|\mathcal{H}| > n$, then by pigeon hole principle, there always exists a hypothesis h that has probability smaller than $1/n$ in $\mathcal{A}(Z)$ for any $Z \in \mathcal{Z}^n$ and we can construct $\mathcal{A}'$ by outputting a sample of $\mathcal{A}(Z)$ if z is not observed and outputting a sample $\mathcal{A}(Z)|\mathcal{A}(Z) \neq h$ whenever z is observed.

The consistency of $\mathcal{A}'$ follows easily as its risk is at most $1/n$ larger than that of $\mathcal{A}$. ■

A.4 Proofs for characterization of private $\mathfrak{D}$-learnability

Proof [Proof of Lemma 14] Let $\mathcal{A}'$ be the algorithm that applies $\mathcal{A}$ to a random subsample of size $\lfloor\sqrt{n}\rfloor$. If we can show that, for any $\mathcal{D} \in \mathfrak{D}$,

(a) the empirical risk of $\mathcal{A}'$ converges to the the optimal population risk R^* in expectation;

(b) the empirical risk of the ERM learning rule also converges to R^* in expectation,

then by triangle inequality, the empirical risk of $\mathcal{A}'$ must also converge to the empirical risk of ERM, i.e., $\mathcal{A}'$ is $\mathfrak{D}$-universal AERM.

We will start with (a). For any distribution $\mathcal{D} \in \mathfrak{D}$, we have

$$\begin{aligned}
&\mathbb{E}_{Z\sim\mathcal{D}^n}\hat{R}(\mathcal{A}'(Z),Z) = \mathbb{E}_{Z\sim\mathcal{D}^n}\left[\mathbb{E}_{Z'\subset Z,|Z'|=\lfloor\sqrt{n}\rfloor}\hat{R}(\mathcal{A}(Z'),Z)\right]\\
=&\mathbb{E}_{Z'\sim\mathcal{D}^{\lfloor\sqrt{n}\rfloor}}\left[\frac{\lfloor\sqrt{n}\rfloor}{n}\hat{R}(\mathcal{A}(Z'),Z') + \mathbb{E}_{Z''\sim\mathcal{D}^{n-\lfloor\sqrt{n}\rfloor}}\left(\frac{n-\lfloor\sqrt{n}\rfloor}{n}\hat{R}(\mathcal{A}(Z'),Z'')\right)\right]\\
=&\mathbb{E}_{Z'\sim\mathcal{D}^{\lfloor\sqrt{n}\rfloor}}\left[\frac{\lfloor\sqrt{n}\rfloor}{n}\hat{R}(\mathcal{A}(Z'),Z') + \frac{n-\lfloor\sqrt{n}\rfloor}{n}R(\mathcal{A}(Z'))\right] \le \frac{1}{\sqrt{n}} + R^* + \xi(\sqrt{n}).
\end{aligned} \tag{14}$$

The last inequality uses the boundedness of the loss function to get $\hat{R}(\mathcal{A}(Z'),Z') \le 1$ and the $\mathfrak{D}$-consistency of $\mathcal{A}$ to bound the excess risk of $\mathbb{E}_{Z'}R(\mathcal{A}(Z'))$.

To show (b), we need to exploit the assumption that the problem is (non-privately) learnable. By Shalev-Shwartz et al. (2010, Theorem 7), the problem being learnable implies that there exists a universally consistent algorithm $\mathcal{B}$ (not restricted to $\mathfrak{D}$), that is universally AERM with rate $3\xi'(n^{\frac{1}{4}}) + \frac{8}{\sqrt{n}}$ and stable with rate $\frac{2}{\sqrt{n}}$. Moreover, by Shalev-Shwartz et al. (2010, Theorem 8), $\mathcal{B}$'s stability and AERM implies that $\mathcal{B}$ is also generalizing, with rate $6\xi'(n^{\frac{1}{4}}) + \frac{18}{\sqrt{n}}$. Here the term "generalizing" means that the empirical risk is close to the population risk. Therefore, we can establish (b) via the following chain of approximations

$$\mathbb{E}_{Z\sim\mathcal{D}^n}\hat{R}^*(Z) \underset{\substack{\uparrow\\ \text{AERM of } \mathcal{B}}}{\approx} \mathbb{E}_{Z\sim\mathcal{D}^n}\hat{R}(\mathcal{B}(Z),Z) \overset{\substack{\text{Generalization of } \mathcal{B}\\ \downarrow}}{\approx} R(\mathcal{B}(Z)) \underset{\substack{\uparrow\\ \text{Consistency of } \mathcal{B}}}{\approx} R^*.$$

More precisely,

$$\begin{aligned}
&\left|\mathbb{E}_{Z\sim\mathcal{D}^n}\hat{R}^*(Z) - R^*\right|\\
\le&\left|\mathbb{E}_{Z\sim\mathcal{D}^n}\hat{R}^*(Z) - \mathbb{E}_{Z\sim\mathcal{D}^n}\hat{R}\right| + \left|\mathbb{E}_{Z\sim\mathcal{D}^n}\hat{R} - R(\mathcal{B}(Z),Z)\right| + |R(\mathcal{B}(Z),Z) - R^*|\\
\le&[3\xi'(n^{\frac{1}{4}}) + \frac{8}{\sqrt{n}}] + [6\xi'(n^{\frac{1}{4}}) + \frac{18}{\sqrt{n}}] + [3\xi'(n^{\frac{1}{4}}) + \frac{10}{\sqrt{n}}] = 12\xi(n^{\frac{1}{4}}) + \frac{36}{\sqrt{n}}.
\end{aligned} \tag{15}$$

Combine (14) and (15), we obtain the AERM of $\mathcal{A}'$ with rate $12\xi'(n^{1/4}) + \frac{37}{\sqrt{n}} + \xi(\sqrt{n})$ as required. The privacy of $\mathcal{A}'$ follows from Lemma 27. ■

A.5 Proof for Theorem 17

We first present the proof for Theorem 17. Recall that the roadmap of the proof is summarized in Figure 3.

For readability, we denote $\epsilon(n)$ by simply ϵ.

Recall that the objective function is $F(h, Z) = \frac{1}{n}\sum_{i=1}^{n} \ell(h, z_i) + g_n(h)$ and the corresponding utility function $q(h, Z) = -F(h, Z)$. By the boundedness assumption, it is easy to show that if we replace one data point in any Z with something else, then sensitivity

$$\Delta q = \sup_{h\in\mathcal{H}, d(Z,Z')=1} \left|q(Z, h) - q(Z', h)\right| \leq \frac{2}{n}. \tag{16}$$

Then by McSherry and Talwar (2007, Theorem 6), Algorithm 1 that outputs $h \in \mathcal{H}$ with $\mathbb{P}(h) \propto \exp(\frac{\epsilon}{2\Delta q} q(h, Z))$ naturally ensures ϵ-differential privacy.

Denote shorthand $F^* := \inf_{f\in\mathcal{H}} F(Z, h)$ and $q^* := -F^*$, we can state an analog of the utility theorem of the exponential mechanism in (McSherry and Talwar, 2007).

Lemma 29 (Utility) *Assuming $\epsilon < \log n$ (otherwise the privacy protection is meaningless anyway), if assumption A1, A2 hold for distribution $\mathcal{D}$, then*

$$\mathbb{E}_{Z\sim\mathcal{D}^n}\mathbb{E}_{h\sim\mathcal{A}(Z)} q(Z, h) \geq -\mathbb{E}_{Z\sim\mathcal{D}^n} F^* - \frac{9[(\rho+2)\log n + \log K]}{n\epsilon}. \tag{17}$$

Proof By the boundedness of ℓ and g

$$q(Z, h) = -\frac{1}{n}\sum_i \ell(h, z_i) - g_n(h) \geq -(1 + \zeta(n)).$$

By Lemma 7 in McSherry and Talwar (2007) (translated to our case),

$$\mathbb{P}_{h\sim\mathcal{A}(Z)}\left[q(Z, h) < -F^* - 2t\right] \leq \frac{\mu(\mathcal{H})}{\mu(\mathcal{S}_t)} e^{-\frac{\epsilon}{2\Delta q}t}, \tag{18}$$

Apply (16), take expectation over the data distribution on both sides, and applying assumption A2, we get

$$\mathbb{E}_{Z\sim\mathcal{D}^n}\mathbb{P}_{h\sim\mathcal{A}(Z)}\left[q(Z, h) < -F^* - 2t\right] \leq Kt^{-\rho}e^{-\frac{\epsilon n t}{4}} = e^{-\frac{\epsilon n t}{4} + \log K - \rho\log t} := e^{-\gamma}. \tag{19}$$

Take $t = \frac{4\left[(\rho+2)\log n + \log(K)\right]}{\epsilon n}$, by the assumption that $\epsilon < \log n$, we get $\log(nt) > 0$. Substitute t into the expression of γ we obtain

$$\gamma = \frac{\epsilon n}{4}t - \log K + \rho\log t = 2\log n + \rho\log(nt) \geq 2\log n,$$

and therefore

$$\mathbb{E}_{Z\sim\mathcal{D}^n}\mathbb{P}_{h\sim\mathcal{A}(Z)}\left[q(Z, h) < -F^* - 2t\right] \leq n^{-2}.$$

Denote $\mathbb{P}_{h\sim\mathcal{A}(Z)}[q(Z, h) < -F^* - 2t] =: p$, we can then bound the expectation from below as follows:

$$\begin{aligned}
\mathbb{E}_{Z\sim\mathcal{D}^n}\mathbb{E}_{h\sim\mathcal{A}(Z)} q(Z, h) \geq& \mathbb{E}_{Z\sim\mathcal{D}^n}(-F^* - 2t)(1-p) + \min_{h\in\mathcal{H}, Z\in\mathcal{Z}^n} q(Z, h)\mathbb{E}_{Z\sim\mathcal{D}^n} p \\
\geq& \mathbb{E}_{Z\sim\mathcal{D}^n}(-F^* - 2t) + (-1 - \zeta(n))\, n^{-2} \\
\geq& -\mathbb{E}_{Z\sim\mathcal{D}^n} F^* - \frac{8\left[(\rho+2)\log n + \log(K)\right]}{\epsilon n} - (1 + \zeta(n))\, n^{-2} \\
\geq& -\mathbb{E}_{Z\sim\mathcal{D}^n} F^* - \frac{9\left[(\rho+2)\log n + \log(K)\right]}{\epsilon n}.
\end{aligned}$$

■

Now we can say something about the learning problem. In particular, the AERM follows directly from the utility result and stability follows from the definition of differential privacy.

Lemma 30 (Universal AERM) *Assume A1 and A2, and $\epsilon \leq \log n$ (so Lemma 29 holds), then*

$$\mathbb{E}_{Z\sim\mathcal{D}^n}\left[\mathbb{E}_{h\sim\mathcal{A}(Z)}\hat{R}(h,Z) - \hat{R}^*(Z)\right] \leq \frac{9[(\rho+2)\log n + \log(1/K)]}{n\epsilon} + \zeta(n).$$

Proof This is a simple consequence of boundedness and Lemma 29.

$$\begin{aligned}
&\mathbb{E}_{Z\sim\mathcal{D}^n}\left[\mathbb{E}_{h\sim\mathcal{A}(Z)}\hat{R}(h,Z) - \hat{R}^*(Z)\right] \\
=&\mathbb{E}_{Z\sim\mathcal{D}^n}\mathbb{E}_{h\sim\mathcal{A}(Z)}\frac{1}{n}\sum_i \ell(h,z_i) - \mathbb{E}_{Z\sim\mathcal{D}^n}\inf_h \frac{1}{n}\sum_i \ell(h,z_i) \\
\leq&\mathbb{E}_{Z\sim\mathcal{D}^n}\mathbb{E}_{h\sim\mathcal{A}(Z)}\left[\frac{1}{n}\sum_i \ell(h,z_i) + g_n(h)\right] - \mathbb{E}_{h\sim\mathcal{A}(Z)}g_n(h) \\
&- \mathbb{E}_{Z\sim\mathcal{D}^n}\inf_h\left[\frac{1}{n}\sum_i \ell(h,z_i) + g_n(h)\right] + \sup_h(g_n(h)) \\
=&\mathbb{E}_{Z\sim\mathcal{D}^n}(-F^* - \mathbb{E}_{h\sim\mathcal{A}(Z)}q(Z,h)) + \sup_h g_n(h) - \mathbb{E}_{h\sim\mathcal{A}(Z)}g_n(h) \\
\leq&\frac{9[(\rho+2)\log n + \log(1/K)]}{n\epsilon} + 2\zeta(n).
\end{aligned}$$

The last step applies Lemma 29 and $\sup_h |g_n(h)| \leq \zeta(n)$ as in Assumption A2 by using the fact that $\sup_h g_n(h) - \mathbb{E}g_n(h) \leq 2\sup_h |g_n(h)|$ for any distribution of h the expectation is taken over. ■

The above theorem shows that Algorithm 1 is asymptotic ERM. By Theorem 8, the fact that this algorithm is ϵ-differential private implies that it is 2ϵ-stable. Now the proof follows by applying Theorem 28 which says that stability and AERM of an algorithm certify its consistency. Noting that this holds for any distribution $\mathcal{D}$ completes our proof for learnability in Theorem 17.

A.6 Proofs of other technical results

High confidence private learning. Proof [Proof of Theorem 18] The algorithm $\mathcal{A}$ privately learns the problem with rate $\xi(n)$ implies that

$$\mathbb{E}_{Z\in\mathcal{D}^n}\mathbb{E}_{h\sim\mathcal{A}(Z)}R(h) - R^* \leq \xi(n).$$

Let $h \sim \mathcal{A}(Z)$ and $Z \sim \mathcal{D}^n$, by Markov's inequality, with probability at least $1 - 1/e$,

$$R(h) - R^* \leq e\xi(n).$$

If we split the data randomly into $a+1$ parts of size $n/(a+1)$ and run $\mathcal{A}$ on the first a partitions, then we get $h_j \sim \mathcal{A}(Z_j)$. Then with probability at lest $1-(1/e)^a$, at least one of them has risk

$$\min_{j\in[a]} R(h_j) - R^* \leq e\xi(\frac{n}{a+1}). \tag{20}$$

Since the $(a+1)$th partition are iid data, and ℓ is bounded, we can apply Hoeffding's inequality and union bound, so that with probability $1-\delta_1$ for all $j = 1, ..., a+1$

$$\hat{R}(h_j, Z_{a+1}) - R(h_j) \leq \sqrt{\frac{\log(2a/\delta_1)}{2n}}. \tag{21}$$

This means that if exponential mechanism picked the one with the best validation risk it will be almost as good as the one with the best risk. Assume h_1 is the one that achieves the best validation risk.

Now it remains to bound the probability that exponential mechanism pick an $h \in \{h_1, ..., h_a\}$ that is much worse than h_1.

Recall that the utility function is the negative validation risk which depends only on the last partition I_{a+1}.

$$q(X, h) = \frac{1}{n/(a+1)} \sum_{i\in I_{a+1}} \ell_i(z_i, h).$$

This is in fact a random function of the data because we are picking the the validation set I_{a+1} randomly from the data. Suppose we arbitrarily replace one data point j from the dataset, the distribution of the output of function $q(Z, h)$ is a mixture of the two cases: $j \in I_{a+1}$ and $j \notin I_{a+1}$. Since in the first case, $q(Z, h) = q(Z', h)$ for all h, sensitivity for this case is 0. In the second case, by the boundedness assumption, the sensitivity is at most $2(a+1)/n$. For the exponential mechanism guarantee ϵ differential privacy, it suffices to take the sensitivity parameter to be $2(a+1)/n$.

By the utility theorem of the exponential mechanism,

$$\mathbb{P}\left[\hat{R}(h) > \hat{R}(h_1) + \frac{8(\eta \log n + \log a)}{\epsilon n/(a+1)}\right] \leq n^{-\eta}. \tag{22}$$

Combine (20)(21) and(22) we get

$$\mathbb{P}\left[R(h) - R^* > e\xi(\frac{n}{a+1}) + \sqrt{\frac{\log(2a/\delta_1)}{2n}} + \frac{8(\eta \log n + \log a)}{\epsilon n/(a+1)}\right] \leq n^{-\eta} + \delta_1 + e^{-a}.$$

Now by appropriately choosing $\eta = \log(3/\delta)/\log n$, $a = \log(3/\delta)$, $\delta_1 = \delta/3$, we get

$$\mathbb{P}\Bigg[R(h) - R^* > e\xi(\frac{n}{\log(3/\delta)+1}) + \sqrt{\frac{\log(2\log(3/\delta)) + \log(3/\delta)}{2n}} + \frac{8(\log(3/\delta) + \log\log(3/\delta))}{\epsilon n/(\log(3/\delta)+1)}\Bigg] \leq \delta$$

combine the terms and take $\epsilon = \frac{\log(3/\delta)+1}{\sqrt{n}}$, we get the bound of the excess risk in the theorem.

To get the privacy claim, note that we are applying $\mathcal{A}$ on disjoint partitions of the data so the privacy parameter does not aggregate. Take the worst over all partitions, we get the overall privacy loss $\max\left\{\epsilon\left(\frac{n}{\log(3/n)+1}\right), \frac{\log(3/\delta)+1}{\sqrt{n}}\right\}$ as stated in the theorem. ■

The Lipschitz example. Proof [Proof of Example 4] Let $h^* \in \operatorname{argmin}_{h\in\mathcal{H}} F(Z,h)$, the Lipschitz condition dictates that for any h,

$$|F(h) - F(h^*)| \le L\|h - h^*\|_p.$$

Choose a small enough $t < t_0$ such that h is in the small neighborhood of h^*, and we can construct a function $\tilde{F}$ that within the sublevel set $\mathcal{S}_t$, such that the above inequality (when we replace F with $\tilde{F}$) is equality, then for any $h \in \mathcal{S}_{t_0}$, $\tilde{F}(h) \ge F(Z,h)$. Verify that the sublevel set of $\tilde{F}(h)$, denoted by $\tilde{\mathcal{S}}_t$ always contains $\mathcal{S}_t$. In addition, we can compute the measure $\mu(\tilde{\mathcal{S}}_t)$ explicitly, since the function is a cone and

$$L\|h - h^*\|_p = |\tilde{F}(h) - \tilde{F}(h^*)| = \tilde{F}(h) - \tilde{F}(h^*) \le t,$$

therefore

$$\tilde{\mathcal{S}}_t = \{h \mid L\|h - h^*\|_p \le t\}.$$

Since $\mathcal{H}$ is β_p-regular, $\mu(B \cap \mathcal{H}) \ge \beta_p \mu(B)$ for any ℓ_p ball $B \subset \mathbb{R}^d$, the measure of the sublevel set can be lower bounded by β_p times the volume of the ℓ_p ball with radius t/L and since $\tilde{\mathcal{S}}_t \subseteq \mathcal{S}_t$, we have

$$\mu(\mathcal{S}_t) \ge \mu(\tilde{\mathcal{S}}_t) \ge \beta_p \mu\left(B(t/L)\right) = \beta_p \left(t/L\right)^d$$

as required. ■

Appendix B. Alternative proof of Corollary 9 via Dwork et al. (2015b, Theorem 7)

In this Appendix, we describe how the results in Dwork et al. (2015b) can be used to obtain the forward direction of our characterization without going through a stability argument. We first restate the result here in our notation:

Lemma 31 (Theorem 7 in Dwork et al. 2015b) *Let $\mathcal{B}$ be an ϵ-DP algorithm such that given a dataset Z, $\mathcal{B}$ outputs a function from $\mathcal{Z}$ to $[0,1]$. For any distribution $\mathcal{D}$ over $\mathcal{Z}$ and random variable $Z \sim \mathcal{D}^n$, we let $\phi \sim \mathcal{B}(Z)$. Then for any $\beta > 0$, $\tau > 0$ and $n \ge 12\log(4/\beta)/\tau^2$, setting $\epsilon < \tau/2$ ensures*

$$\mathbb{P}_{\phi\sim\mathcal{B}(Z), Z\sim\mathcal{D}^n}\left[\left|\mathbb{E}_{z\sim\mathcal{D}}\phi(z) - \frac{1}{n}\sum_{z\in Z}\phi(z)\right| \ge \tau\right] \le \beta.$$

This lemma was originally stated to prove the claim that privately generated mechanisms for answering statistical queries always generalize.

For statistical learning problems, we can simply take the statistical query ϕ to be the loss function $\ell(h, \cdot)$ parameterized by $h \in \mathcal{H}$. If an algorithm $\mathcal{A}$ that samples from a distribution on $\mathcal{H}$ upon observing data Z is ϵ-DP, then $\mathcal{B} : Z \to \ell(\mathcal{A}(Z), \cdot)$ is also ϵ-DP. The result therefore reduces to that the empirical risk and population risk are close with high probability. Due to the boundedness assumption, we can translate the high probability result to the expectation form, which verifies the definition of "generalization".

However, "generalization" alone still does not imply "consistency", as we also need

$$\mathbb{E}_{\phi \sim \mathcal{B}(Z)} \frac{1}{n} \sum_{z \in Z} \phi(z) \to R^* = \min_{\phi \in \Phi} \mathbb{E}_{z \sim \mathcal{D}} \phi(z)$$

as Z gets large, which does not hold for all DP-output ϕ. But when $\phi = \ell(h, \cdot)$, it can be obtained if we assume $\mathcal{A}$ is AERM. This is shown via the following inequality

$$\mathbb{E}_{Z \in \mathcal{D}^n} \mathbb{E}_{\phi \sim \mathcal{B}(Z)} \frac{1}{n} \sum_{z \in Z} \phi(z) \to \mathbb{E}_{Z \in \mathcal{D}^n} \min_{\phi \in \Phi} \frac{1}{n} \sum_{z \in Z} \phi(z) \leq \mathbb{E}_{Z \in \mathcal{D}^n} \frac{1}{n} \sum_{z \in Z} \phi^*(z) = \mathbb{E}\phi^*(z) = R^*,$$

where $\phi^* = \ell(h^*, \cdot)$ and h^* is an optimal hypothesis function. This wraps up the proof of consistency.

The above proof of "consistency" via Lemma 31 and "AERM", however, leads to a looser bound comparing to our result (Corollary 9) when the additional assumption on n and τ (equivalently ϵ) is active, i.e., when $\frac{\epsilon(n)}{\log(1/\epsilon(n))} < O\left(\frac{1}{\sqrt{n}}\right)$. In this case it only implies a $\xi(n) + \frac{\log n}{\sqrt{n}}$ bound due to that ϵ-DP implies ϵ'-DP for any $\epsilon' > \epsilon$. Our proof of Corollary 9 is considerably simpler and more general in that it does not require any assumption on the number of data points n.

This can easily lead to worse overall error bound for very simple learning problems with sufficiently fast rate. For example, in the problem of learning the mean of $X \in [0, 1]$, let the loss function be $|x-h|^{10}$. Consider the $\epsilon(n)$-DP algorithm that outputs ERM+Laplace($\frac{2}{\epsilon(n)n}$) where $\epsilon(n)$ is chosen to be $n^{-9/10}$. This algorithm is AERM with rate $\xi(n) = \frac{10!2!}{(\epsilon(n)n)^{10}} = O(n^{-1})$. By Corollary 9 we get an overall rate of $O(n^{-9/10})$ while through Lemma 31 and the argument that follows, we only get $\tilde{O}(n^{-1/2})$.

References

David Applegate and Ravi Kannan. Sampling and integration of near log-concave functions. In *ACM Symposium on Theory of Computing (STOC-91)*, pages 156–163, 1991.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization, revisited. *arXiv preprint arXiv:1405.7085*, 2014.

Raef Bassily, Kobbi Nissim, Adam Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. *arXiv preprint arXiv:1511.02513*, 2015.

Amos Beimel, Kobbi Nissim, and Uri Stemmer. Characterizing the sample complexity of private learners. In *Conference on Innovations in Theoretical Computer Science (ITCS-13)*, pages 97–110. ACM, 2013a.

Amos Beimel, Kobbi Nissim, and Uri Stemmer. Private learning and sanitization: Pure vs. approximate differential privacy. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 363–378. Springer, 2013b.

Amos Beimel, Hai Brenner, Shiva Prasad Kasiviswanathan, and Kobbi Nissim. Bounds on the sample complexity for private learning and private data release. *Machine learning*, 94 (3):401–437, 2014.

Avrim Blum and Moritz Hardt. The ladder: A reliable leaderboard for machine learning competitions. In *International Conference on Machine Learning (ICML-15)*, pages 1006–1014, 2015.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

Mark Bun, Kobbi Nissim, Uri Stemmer, and Salil Vadhan. Differentially private release and learning of threshold functions. In *IEEE Symposium on Foundations of Computer Science (FOCS-15)*, pages 634–649. IEEE, 2015.

Mark Mar Bun. *New Separations in the Complexity of Differential Privacy*. PhD thesis, Harvard University Cambridge, Massachusetts, 2016.

Kamalika Chaudhuri and Daniel Hsu. Sample complexity bounds for differentially private learning. In *Conference on Learning Theory (COLT-11)*, volume 19, 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.

Cynthia Dwork. Differential privacy. In *Automata, Languages and Programming*, pages 1–12. Springer, 2006.

Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *ACM Symposium on Theory of Computing (STOC-09)*, pages 371–380. ACM, 2009.

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Advances in Cryptology-EUROCRYPT 2006*, pages 486–503. Springer, 2006a.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography*, pages 265–284. Springer, 2006b.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349 (6248):636–638, 2015a.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Leon Roth. Preserving statistical validity in adaptive data analysis. In *ACM Symposium on Theory of Computing (STOC-15)*, pages 117–126. ACM, 2015b.

Moritz Hardt and Jonathan Ullman. Preventing false discovery in interactive data analysis is hard. In *IEEE Symposium on Foundations of Computer Science (FOCS-14)*, pages 454–463. IEEE, 2014.

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

Prateek Jain and Abhradeep Thakurta. Differentially private learning with kernels. In *International Conference on Machine Learning (ICML-13)*, pages 118–126, 2013.

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

Michael Kearns and Dana Ron. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural Computation*, 11(6):1427–1453, 1999.

Michael J Kearns, Robert E Schapire, and Linda M Sellie. Toward efficient agnostic learning. In *Workshop on Computational learning theory (COLT-92)*, pages 341–352. ACM, 1992.

Daniel Kifer, Adam Smith, and Abhradeep Thakurta. Private convex empirical risk minimization and high-dimensional regression. *Journal of Machine Learning Research*, 1:41, 2012.

Jing Lei. Differentially private m-estimators. In *Advances in Neural Information Processing Systems (NIPS-11)*, pages 361–369, 2011.

Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *IEEE Symposium on Foundations of Computer Science, 2007 (FOCS-07)*, pages 94–103, 2007.

Sayan Mukherjee, Partha Niyogi, Tomaso Poggio, and Ryan Rifkin. 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, 2006.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems (NIPS-07)*, pages 1177–1184, 2007.

Robert E Schapire. The strength of weak learnability. *Machine Learning*, 5(2):197–227, 1990.

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

Adam Smith. Privacy-preserving statistical estimation with optimal convergence rates. In *ACM Symposium on Theory of Computing (STOC-11)*, pages 813–822, 2011.

Abhradeep Guha Thakurta and Adam Smith. Differentially private feature selection via stability arguments, and the robustness of the lasso. In *Conference on Learning Theory (COLT-13)*, pages 819–850, 2013.

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

Vladimir Vapnik. *Statistical Learning Theory*. Wiley, 1998.

Vladimir N Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

Yu-Xiang Wang, Stephen E Fienberg, and Alex Smola. Privacy for free: Posterior sampling and stochastic gradient monte carlo. In *International Conference on Machine Learning (ICML-15)*, 2015.

Larry Wasserman and Shuheng Zhou. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389, 2010.

Fei Yu, Stephen E Fienberg, Aleksandra B Slavković, and Caroline Uhler. Scalable privacy-preserving data sharing methodology for genome-wide association studies. *Journal of biomedical informatics*, 50:133–141, 2014.