Sample Complexity Bounds for Differentially Private Learning

Kamalika Chaudhuri
UC San Diego
kchaudhuri@ucsd.edu

Daniel Hsu
Rutgers University and University of Pennsylvania
djhsu@rci.rutgers.edu

Abstract

We study the problem of privacy-preserving classification – namely, learning a classifier from sensitive data, while still preserving the privacy of individuals in the training set. In particular, we require that our learning algorithm guarantees differential privacy, a very strong notion of privacy that has gained significant attention over the past few years.

A natural question to ask is: what is the sample requirement of a learning algorithm that guarantees a certain level of privacy and accuracy? In this paper, we study this question in the context of infinite hypothesis classes when the data is drawn from a continuous distribution. We show that even for very simple hypothesis classes, any algorithm which uses a finite number of examples and guarantees differential privacy, fails to guarantee classification accuracy for at least one unlabeled data distribution. This result is unlike the case of finite hypothesis classes and hypothesis classes on discrete data domains that were studied by Kasiviswanathan et al. (2008).

We then propose two approaches to differentially private learning that get around this lower bound. The first approach is to use some prior knowledge about the unlabeled data distribution in the form of a reference distribution $\mathcal{U}$ that is chosen independently of the sensitive data. Given such a reference $\mathcal{U}$, we provide an upper bound on the sample requirement which depends (among other things) on a measure of closeness between $\mathcal{U}$ and the unlabeled data distribution. Our upper bound applies to the non-realizable as well as the realizable case. The second approach is to relax the privacy requirement, by requiring only label-privacy – namely, that the labels, and not the unlabeled parts of the examples be considered sensitive information. An upper bound on the sample requirement of learning with label privacy was shown by Chaudhuri et al. (2006); in this paper, we show a lower bound.

1 Introduction

As increasing amounts of personal data is collected, stored and mined by companies and government agencies, the question of how to learn from sensitive datasets while still maintaining the privacy of individuals in the data has become very important. Over the last few years, the notion of differential privacy (Dwork et al., 2006) has received a significant amount of attention, and has become the de facto standard for privacy-preserving computation. In this paper, we study the problem of learning a classifier from a dataset, while simultaneously guaranteeing differential privacy of the training data.

The key issue in differentially-private computation is that given a certain amount of resources, there is usually a tradeoff between privacy and utility. In classification, a natural measure of utility is the classification accuracy, and data is a scarce resource. Thus, a key question in differentially-private learning is: how many examples does a learning algorithm need to guarantee a certain level of privacy and accuracy? In this paper, we study this question from an information-theoretic perspective – namely, we are concerned with the sample complexity, and not the computational complexity of the learner.

This question was first considered by Kasiviswanathan et al. (2008), who studied the case of finite hypothesis classes, as well as the case of discrete data domains. They showed that in these two cases, one can obtain any given privacy guarantee and generalization error, regardless of the unlabeled data distribution with a modest increase in the worst-case sample requirement.

In this paper, we consider the sample complexity of differentially private learning in the context of *infinite hypothesis classes on continuous data distributions*. This is a very general class of learning problems, and

includes many popular machine-learning tasks such as learning linear classifiers when the examples have real-valued features, which cannot be modeled by finite hypothesis classes or hypothesis classes over discrete data domains.

Surprisingly, we show that the results of Kasiviswanathan et al. (2008) do not extend to infinite hypothesis classes on continuous data distributions. As an example, consider the class of thresholds on the unit interval. This simple learning problem has VC dimension 1, and thus for all unlabeled data distributions, it can be learnt (non-privately) with error ϵ given at most $\tilde{O}(\frac{1}{\epsilon})$ examples[1]. We show that even for this very simple hypothesis class, any algorithm which uses a bounded number of examples and guarantees differential privacy, fails to guarantee classification accuracy for at least one unlabeled data distribution.

The key intuition behind our proof is that if most of the unlabeled data is concentrated in a small region around the best classifier, then, even slightly perturbing the best classifier will result in a large classification error. As the process of ensuring differential privacy necessarily involves some perturbation – see, for example, Dwork et al. (2006), unless the algorithm has some prior public knowledge about the data distribution, the number of samples required to learn privately grows with growing concentration of the data around the best classifier.

How can we then learn privately in infinite hypothesis classes over continuous data distributions? One approach is to use some prior information about the data distribution which is known independently of the sensitive data. Another approach is to relax the privacy requirements. In this paper, we examine both approaches.

First, we consider the case when the learner has access to some prior information on the unlabeled data. In particular, the learner knows a reference distribution $\mathcal{U}$ that is close to the unlabeled data distribution. Similar assumptions are common in Bayesian learning, and PAC-Bayes style bounds have also been studied in the learning theory literature, for example, by McAllester (1998).

Under this assumption, we provide an algorithm for learning with α-privacy, excess generalization error ϵ, and confidence $1-\delta$, using $\tilde{O}(d_{\mathcal{U}} \log(\kappa/\epsilon)(\frac{1}{\epsilon^2} + \frac{1}{\epsilon\alpha}))$ samples. Here α is a privacy parameter (where, lower α implies a stronger privacy guarantee), $\mathcal{U}$ is the reference distribution, $d_{\mathcal{U}}$ is the doubling dimension of its disagreement metric (Bshouty et al., 2009), and κ is a smoothness parameter that we define. The quantity $d_{\mathcal{U}}$ measures the complexity of the hypothesis class with respect to $\mathcal{U}$ (see (Bshouty et al., 2009) for a discussion), and we assume that it is finite. The smoothness parameter measures how close the unlabeled data distribution is to $\mathcal{U}$ (smaller κ means closer), and is motivated by notions of closeness used in Dasgupta (2005) and Freund et al. (1997). Thus the sample requirement of our algorithm grows with increasing distance between $\mathcal{U}$ and the unlabeled data distribution. Our algorithm works in the non-realizable case, that is, when no hypothesis in the class has zero error; using standard techniques, a slightly better bound of $\tilde{O}(\frac{d_{\mathcal{U}} \log(\kappa/\epsilon)}{\epsilon\alpha})$ can be obtained in the realizable setting. However, like the results of Kasiviswanathan et al. (2008), our algorithm is computationally inefficient in general.

The main difficulty in reducing the differentially-private learning algorithms of Kasiviswanathan et al. (2008) to infinite hypothesis classes on continuous data distributions is in finding a suitable finite cover of the class with respect to the unlabeled data. This issue is specific to our particular problem: for non-private learning, a finite cover can always be computed based on the (sensitive) data, and for finite hypothesis classes, the entire class is a cover. The main insight behind our upper bound is that when the unlabeled distribution $\mathcal{D}$ is close to the reference distribution $\mathcal{U}$, then a cover of $\mathcal{U}$ is also a possibly coarser cover of $\mathcal{D}$. Since one can compute a private cover of $\mathcal{U}$ independent of the sensitive data, we simply compute a finer cover of $\mathcal{U}$, and learn over this fine cover using standard techniques such as the exponential mechanism (McSherry and Talwar, 2007).

Next we relax the privacy requirement by requiring only *label privacy*. In other words, we assume that the unlabeled part of the examples are not sensitive, and the only private information is the labels. This setting was considered by Chaudhuri et al. (2006). An example when this may be applicable is in predicting income from public demographic information. Here, while the label (income) is private, the demographic information of individuals, such as education, gender, and age may be public.

In this case, we provide lower bounds to characterize the sample requirement of label-private learning. We show two results, based on the value of α and ϵ. For small ϵ and α (that is, for high privacy and accuracy) we show that any algorithm which learns all hypotheses in a given class with α-label privacy and ϵ accuracy requires at least $\Omega(\frac{d}{\alpha\epsilon})$ examples. Here d is the doubling dimension of the disagreement metric at a certain scale, and is a measure of the complexity of the hypothesis class on the unlabelled data distribution. This bound holds when the hypothesis class has finite VC dimension. For larger α and ϵ, our bounds are weaker but more general; we show a lower bound of $\Omega(\frac{d'}{\alpha})$ on the sample requirement, which holds for any α and ϵ, and do not require the VC dimension of the hypothesis class to be finite. Here d' is the doubling dimension of the disagreement metric at a certain scale.

[1] Here the $\tilde{O}$ notation hides factors logarithmic in $1/\epsilon$

The main idea behind our stronger label privacy lower bounds is to show that differentially private learning algorithms necessarily perform poorly when there is a large set of hypotheses such that every pair in the set labels approximately $1/\alpha$ examples differently. We then show that such large sets can be constructed when the doubling dimension of the disagreement metric of the hypothesis class with respect to the data distribution is high.

How do these results fit into the context of non-private learning? For non-private learning, sample requirement bounds based on the doubling dimension of the disagreement metric has been extensively studied by (Bshouty et al., 2009); in the realizable case, they show an upper bound of $\tilde{O}(\frac{\bar{d}}{\epsilon})$ for learning with accuracy ϵ, where $\bar{d}$ is again the doubling dimension of the disagreement metric at a certain scale. These bounds are incomparable to ours in general, as the doubling dimensions in the two bounds are with respect to different scales; however, we can compare them for hypothesis classes and data distributions for which the doubling dimension of the disagreement metric is equal at all scales. An example is learning halfspaces with respect to the uniform distribution on the sphere. For such problems, on the upper bound side, we need a factor of $O(\frac{\log(\kappa/\epsilon)}{\alpha})$ times more examples to learn with α-privacy. On the other hand, our lower bounds indicate that for small α and ϵ, even if we only want α-label privacy, the sample requirement can be as much as a factor of $\Omega(\frac{1}{\alpha})$ more than the upper bound for non-private learning.

Finally, one may be tempted to think that we can always discretize a data domain or a hypothesis class, and therefore in practice we are likely to only learn finite hypothesis classes or over discrete data domains. However, there are several issues with such discretization. First, if we discretize either the hypothesis class or the data, then the sample requirement of differentially private learning algorithms will grow as the discretization grows finer, instead of depending on intrinsic properties of the problem. Second, as our α-privacy lower bound example shows, indiscriminate discretization without prior knowledge of the data can drastically degrade the performance of the best classifier in a class. Finally, infinite hypothesis classes and continuous data domains provide a natural abstraction for designing many machine learning algorithms, such as those based on convex optimization or differential geometry. Understanding the limitations of differentially private learning on such hypothesis classes and data domains is useful in designing differentially private approximations to these algorithms.

The rest of our paper is organized as follows. In Section 2, we define some preliminary notation, and explain our privacy model. In Section 3, we present our α-privacy lower bound. Our α-privacy upper bound is provided in Section 4. In Section 5, we provide some lower bounds on the sample requirement of learning with α-label privacy. Finally, the proofs of most of our results are in the appendix.

1.1 Related work

The work most related to ours is Kasiviswanathan et al. (2008), Blum et al. (2008) and Beimel et al. (2010), each of which deals with either finite hypothesis classes or discrete data domains.

Kasiviswanathan et al. (2008) initiated the study of the sample requirement of differentially-private learning. They provided a (computationally inefficient) α-private algorithm that learns any finite hypothesis class $\mathcal{H}$ with error at most ϵ using at most $\tilde{O}(\frac{\log|\mathcal{H}|}{\alpha\epsilon})$ examples in the realizable case. For the non-realizable case, they provided an algorithm with a sample requirement of $\tilde{O}(\log|\mathcal{H}|\cdot(\frac{1}{\alpha\epsilon}+\frac{1}{\epsilon^2}))$. Moreover, using a result from Blum et al. (2008), they provided a computationally inefficient α-private algorithm that learns a hypothesis class with VC-dimension V and data dimension n with at most $\tilde{O}(\frac{nV}{\alpha\epsilon^3})$ examples, *provided the data domain is* $\{-1,1\}^n$. None of these results apply when the data is drawn from a continuous distribution; moreover, their results cannot be directly extended to the continuous case.

The first work to study lower bounds on the sample requirement of differentially private learning was Beimel et al. (2010). They show that any α-private algorithm that selects a hypothesis from a specific set C_ϵ requires at least $\tilde{\Omega}(\log(|C_\epsilon|)/\alpha)$ samples to achieve error ϵ. Here C_ϵ is an ϵ-cover as well as an ϵ-packing of the hypothesis class $\mathcal{H}$ with respect to *every* distribution over the discrete data domain. They also show an upper bound of $\tilde{O}(\log(|C_\epsilon|)/(\alpha\epsilon))$. Such a cover C_ϵ does not exist for continuous data domains; as a result their upper bounds do not apply to our setting. Moreover, unlike our lower bounds, their lower bound only applies to algorithms of a specific form (namely, those that output a hypothesis in C_ϵ), and it also does not apply when we only require the labels to be private.

For the setting of label privacy, Chaudhuri et al. (2006) show an upper bound for PAC-learning in terms of the VC dimension of the hypothesis class. We show a result very similar to theirs in the appendix for completeness, and we show lower bounds for learning with label-privacy which indicate that their bounds are almost tight, in terms of the dependence on α and ϵ.

Zhou et al. (2009) study some issues in defining differential privacy when dealing with continuous outcomes; however, they do not consider the question of learning classifiers on such data.

Finally, a lot of our work uses tools from the theory of generalization bounds. In particular, some of our upper and lower bounds are inspired by Bshouty et al. (2009), which bounds the sample complexity of

(non-private) classification in terms of the doubling dimension of the disagreement metric.

Other related work on privacy. The issue of privacy in data analysis of sensitive information has long been a source of problems for curators of such data, and much of this is due to the realization that many simple and intuitive mechanisms designed to protect privacy are simply ineffective. For instance, the work of Narayanan and Shmatikov (2008) showed that an anonymized dataset released by Netflix revealed enough information so that an adversary, by knowing just a few of the movies rated by a particular user, would be able to uniquely identify such a user in the data set and determine *all* of his movie ratings. Similar attacks have been demonstrated on private data in other domains as well including social networks (Backstrom et al., 2007) and search engine query logs (Jones et al., 2007). Even releasing coarse statistics without proper privacy safeguards can be problematic. This was recently shown by Wang et al. (2009) in the context of genetic data, where a correlation matrix of genetic markers compiled from a group of individuals contained enough clues to uniquely pinpoint individuals in the dataset and learn of their private information, such as whether or not they had certain diseases.

In order to reason about privacy guarantees (or lack thereof), we need a formal definition of what it means to preserve privacy. In our work, we adopt the notion of *differential privacy* due to Dwork et al. (2006), which has over the last few years gained much popularity. Differential privacy is known to be a very strong notion of privacy: it has strong semantic guarantees (Kasiviswanathan and Smith, 2008) and is resistant to attacks that many earlier privacy definitions are susceptible to (Ganta et al., 2008b).

There has been a significant amount of work on differential privacy applied to a wide variety of data analysis tasks (Dwork et al., 2006, Chaudhuri and Mishra, 2006, Nissim et al., 2007, Barak et al., 2007, McSherry and Mironov, 2009). Some work that is relevant to ours include Blum et al. (2008), which provides a general method for publishing datasets on discrete data domains while preserving differential privacy so that the answers to queries from a function class with bounded VC dimension will be approximately preserved after the applying the sanitization procedure. More work on this line includes Roth (2010) and Gupta et al. (2011). A number of learning algorithms have also been suitably modified to guarantee differential privacy. For instance, both the classes of statistical query algorithms and the class of methods based on L_2-regularized empirical risk minimization with certain types of convex losses can be made differentially private (Blum et al., 2005, Chaudhuri et al., 2011).

There has also been some prior work on providing lower bounds on the loss of accuracy that any differentially private mechanism would suffer; much of this work is in the context of releasing answers to some a of queries made on a database of n individuals. The first such work is by (Blum et al., 2008), which shows that no differentially private mechanism can hope to release with a certain amount of accuracy the answer to a number of median queries when the data lies on a real line. This result is similar in spirit to our Theorem 1, but applies to a much harder problem, namely data release. Other relevant work includes (Hardt and Talwar, 2010), which uses a packing argument similar to ours to provide a lower bound on the amount of noise any differentially private mechanism needs to add to the answer to k linear queries on a database of n people.

There has also been a significant amount of prior work on privacy-preserving data mining (Agrawal and Srikant, 2000, Evfimievski et al., 2003, Sweeney, 2002, Machanavajjhala et al., 2006), which spans several communities and uses privacy models other than differential privacy. Many of the models used have been shown to be susceptible to various attacks, such as *composition attacks*, where the adversary has some amount of prior knowledge (Ganta et al., 2008a). An alternative line of privacy work is in the Secure Multiparty Computation setting due to Yao (1982), where the sensitive data is split across several adversarial databases, and the goal is to compute a function on the union of these databases. This is in contrast with our setting, where a single centralized algorithm can access the entire dataset.

2 Preliminaries

2.1 Privacy model

We use the *differential privacy* model of Dwork et al. (2006). In this model, a private database $\mathsf{DB} \subseteq \mathcal{Z}$ consists of m sensitive entries from a domain $\mathcal{Z}$; each entry in DB is a record about an individual (*e.g.*, their medical history) that one wishes to keep private.

The database DB is accessed by users through a sanitizer M. The sanitizer, a randomized algorithm, is said to preserve differential privacy if the value of any one individual in the database does not significantly alter the output distribution of M.

Definition 1. A randomized mechanism M guarantees α-*differential privacy* if, for all databases DB_1 and DB_2 that differ by the value of at most one individual, and for every set G of possible outputs of M,

$$\Pr_M[M(\mathsf{DB}_1) \in G] \leq \Pr_M[M(\mathsf{DB}_2) \in G] \cdot e^{\alpha}.$$

We emphasize that the probability in the definition above is only with respect to the internal randomization of the algorithm; it is independent of all other random sources, including any that may have generated the values of the input database.

Differential privacy is a strong notion of privacy (Dwork et al., 2006, Kasiviswanathan and Smith, 2008, Ganta et al., 2008b). In particular, if a sanitizer M ensures α-differential privacy, then, an adversary who knows the private values of all the individuals in the database except for one and has arbitrary prior knowledge about the value of the last individual, cannot gain additional confidence about the private value of the last individual by observing the output of a differentially private sanitizer. The level of privacy is controlled by α, where a lower value of α implies a stronger guarantee of privacy.

2.2 Learning model

We consider a standard probabilistic learning model for binary classification. Let $\mathcal{P}$ be a distribution over $\mathcal{X} \times \{\pm 1\}$, where $\mathcal{X}$ is the data domain and $\{\pm 1\}$ are the possible labels. We use $\mathcal{D}$ to denote the marginal of $\mathcal{P}$ over the data domain $\mathcal{X}$.

We assume that our learning algorithms are given m labeled examples $S := \{(x_1, y_1), \ldots, (x_m, y_m)\} \subseteq \mathcal{X} \times \{\pm 1\}$ which are drawn independently from $\mathcal{P}$. This can equivalently be seen as drawing an unlabeled sample $X := \{x_1, \ldots, x_m\}$ from the marginal $\mathcal{D}$, and then, for each $x \in X$, drawing the corresponding label y from the induced conditional distribution.

Given a hypothesis h, the *classification error* of h with respect to a data distibution $\mathcal{P}$ is defined as:

$$\Pr_{(x,y)\sim\mathcal{P}}[h(x) \neq y].$$

A learning algorithm is given as input a labeled sample $S \sim \mathcal{P}^m$, a target accuracy parameter $\epsilon \in (0, 1)$, and target confidence parameter $\delta \in (0, 1)$. Its goal is to return a hypothesis $h : \mathcal{X} \to \{\pm 1\}$ such that its *excess generalization error* with respect to a specified hypothesis class $\mathcal{H}$

$$\Pr_{(x,y)\sim\mathcal{P}}[h(x) \neq y] - \inf_{h' \in \mathcal{H}} \Pr_{(x,y)\sim\mathcal{P}}[h'(x) \neq y]$$

is at most ϵ with probability at least $1 - \delta$ over the random choice of the sample $S \sim \mathcal{P}^m$, as well as any internal randomness of the algorithm.

We also occasionally adopt the *realizable assumption* (with respect to $\mathcal{H}$). The realizable assumption states that there exists some $h^* \in \mathcal{H}$ such that $\Pr_{(x,y)\sim\mathcal{P}}[h^*(x) \neq y] = 0$. In this case, the excess generalization error of a hypothesis h is simply its classification error. If the realizable assumption does not hold, then there is no classifier in the hypothesis class $\mathcal{H}$ with classification error 0, and we are in the non-realizable case.

2.3 Privacy-preserving classification

In privacy-preserving classification, we assume that the database is a training dataset, which is drawn in an i.i.d manner from some data distribution $\mathcal{P}$, and the sanitization mechanism is a learning algorithm, which outputs a classifier based on the training data. In this paper, we consider two possible privacy requirements on our learning algorithms.

Definition 2. A randomized learning algorithm $\mathcal{A}$ guarantees α-*label privacy* ($\mathcal{A}$ is α-label private) if, for any two datasets $S_1 = \{(x_1, y_1), \ldots, (x_{m-1}, y_{m-1}), (x_m, y_m)\}$ and $S_2 = \{(x_1, y_1), \ldots, (x_{m-1}, y_{m-1}), (x_m, y'_m)\}$ differing in at most one label y'_m, and any set of outputs G of $\mathcal{A}$,

$$\Pr_{\mathcal{A}}[\mathcal{A}(S_1) \in G] \leq \Pr_{\mathcal{A}}[\mathcal{A}(S_2) \in G] \cdot e^{\alpha}.$$

Definition 3. A randomized learning algorithm $\mathcal{A}$ guarantees α-*privacy* ($\mathcal{A}$ is α-private) if, for any two datasets $S_1 = \{(x_1, y_1), \ldots, (x_{m-1}, y_{m-1}), (x_m, y_m)\}$ and $S_2 = \{(x_1, y_1), \ldots, (x_{m-1}, y_{m-1}), (x'_m, y'_m)\}$ differing in at most one example (x'_m, y'_m), and any set of outputs G of $\mathcal{A}$,

$$\Pr_{\mathcal{A}}[\mathcal{A}(S_1) \in G] \leq \Pr_{\mathcal{A}}[\mathcal{A}(S_2) \in G] \cdot e^{\alpha}.$$

Note that if the input dataset S is a random variable, then for any value $S' \subseteq \mathcal{X} \times \{\pm 1\}$ in the range of S, the conditional probability distribution of $\mathcal{A}(S) \mid S = S'$ is determined only by the algorithm $\mathcal{A}$ and the value S'; it is independent of the distribution of the random variable S. Therefore, for instance,

$$\Pr_{S,\mathcal{A}}[\mathcal{A}(S) \in G \mid S = S'] = \Pr_{\mathcal{A}}[\mathcal{A}(S') \in G].$$

for any $S' \subseteq \mathcal{X} \times \{\pm 1\}$ and any set of outputs G.

The difference between the two notions of privacy is that for α-label privacy, the two databases can differ only in the label of one example; whereas for α-privacy, the two databases differ can differ in a complete example (both labeled and unlabeled parts). Thus, α-label privacy only ensures the privacy of the label

component of each example; it makes no guarantees about the unlabeled part. If a classification algorithm guarantees α-privacy, then it also guarantees α-label privacy. Thus α-label privacy is a weaker notion of privacy than α-privacy.

The notion of label privacy was also considered by Chaudhuri et al. (2006), who provided an algorithm for learning with label privacy. For strict privacy, one would require the learning algorithm to guarantee α-privacy; however, label privacy may also be an useful notion. For example, if the data x represents public demographic information (*e.g.*, age, zip code, education), while the label y represents income level, an individual may consider the label to be private but may not mind if others can infer her demographic information (which could be relatively public already) by her inclusion in the database.

Thus, the goal of a α-private (resp. α-label private) learning algorithm is as follows. Given a dataset S of size m, a privacy parameter α, a target accuracy ϵ, and a target confidence parameter δ:

1. guarantee α-privacy (resp. α-label privacy) of the training dataset S;
2. with probability at least $1-\delta$ over both the random choice of $S \sim \mathcal{P}^m$ and the internal randomness of the algorithm, return a hypothesis $h : \mathcal{X} \to \{\pm 1\}$ with excess generalization error

$$\Pr_{(x,y)\sim\mathcal{P}}[h(x) \neq y] - \inf_{h' \in \mathcal{H}} \Pr_{(x,y)\sim\mathcal{P}}[h'(x) \neq y] \leq \epsilon.$$

2.4 Additional definitions and notation

We now present some additional essential definitions and notation.

Metric spaces, doubling dimension, covers, and packings. A metric space $(\mathcal{Z}, \rho)$ is a tuple, where $\mathcal{Z}$ is a set of elements, and ρ is a distance function from $\mathcal{Z} \times \mathcal{Z}$ to $\{0\} \cup \mathbb{R}^+$. Let $(\mathcal{Z}, \rho)$ be any arbitrary metric space. For any $z \in \mathcal{Z}$ and $r > 0$, let $B(z, r) = \{z' \in \mathcal{Z} : \rho(z, z') \leq r\}$ denote the ball centered at z of radius r.

The *diameter* of $(\mathcal{Z}, \rho)$ is $\sup\{\rho(z, z') : z, z' \in \mathcal{Z}\}$, the longest distance in the space. An *ε-cover* of $(\mathcal{Z}, \rho)$ is a set $C \subseteq \mathcal{Z}$ such that for all $z \in \mathcal{Z}$, there exists some $z' \in C$ such that $\rho(z, z') \leq \varepsilon$. An *$\varepsilon$-packing* of $(\mathcal{Z}, \rho)$ is a set $P \subseteq \mathcal{Z}$ such that $\rho(z, z') > \varepsilon$ for all distinct $z, z' \in P$. Let $\mathcal{N}_\varepsilon(\mathcal{Z}, \rho)$ denote the size of the smallest ε-cover of $(\mathcal{Z}, \rho)$.

We define the *doubling dimension of $(\mathcal{Z}, \rho)$ at scale ε*, denoted as $\mathrm{ddim}_\varepsilon(\mathcal{Z}, \rho)$, as the smallest number d such that each ball $B(z, \varepsilon) \subseteq \mathcal{Z}$ of radius ε can be covered by at most $\lfloor 2^d \rfloor$ balls of radius $\varepsilon/2$, *i.e.* there exists $z_1, \ldots, z_{\lfloor 2^d \rfloor} \in \mathcal{Z}$ such that $B(z, \varepsilon) \subseteq B(z_1, \varepsilon/2) \cup \ldots \cup B(z_{\lfloor 2^d \rfloor}, \varepsilon/2)$. Notice that $\mathrm{ddim}_\varepsilon(\mathcal{Z}, \rho)$ may increase or decrease with ε. The *doubling dimension of* $(\mathcal{Z}, \rho)$ is $\sup\{\mathrm{ddim}_r(\mathcal{Z}, \rho) : r > 0\}$.

Disagreement metrics. The *disagreement metric* of a hypothesis class $\mathcal{H}$ with respect to a data distribution $\mathcal{D}$ over $\mathcal{X}$ is the metric $(\mathcal{H}, \rho_\mathcal{D})$, where $\rho_\mathcal{D}$ is the following distance function:

$$\rho_\mathcal{D}(h, h') := \Pr_{x\sim\mathcal{D}}[h(x) \neq h'(x)].$$

The *empirical disagreement metric* of a hypothesis class $\mathcal{H}$ with respect to a data distribution $\mathcal{D}$ over $\mathcal{X}$ is the metric $(\mathcal{H}, \rho_X)$, where ρ_X is the following distance function:

$$\rho_X(h, h') := \frac{1}{|X|} \sum_{x \in X} I[h(x) \neq h'(x)].$$

The disagreement metric (or empirical disagreement metric) is the proportion of unlabeled examples on which h and h' disagree with respect to $\mathcal{D}$ (or the uniform distribution over X). We use the notation $B_\mathcal{D}(h, r)$ to denote the ball centered at h of radius r with respect to $\rho_\mathcal{D}$, and $B_X(h, r)$ to denote the ball centered at h of radius r with respect to ρ_X.

Datasets and empirical error. For an unlabeled dataset $X \subseteq \mathcal{X}$ and a hypothesis $h : \mathcal{X} \to \{\pm 1\}$, we denote by $S_{X,h} := \{(x, h(x)) : x \in X\}$ the labeled dataset induced by labeling X with h. The *empirical error* of a hypothesis $h : \mathcal{X} \to \{\pm 1\}$ with respect to a labeled dataset $S \subseteq \mathcal{X} \times \{\pm 1\}$ is $\mathrm{err}(h, S) := (1/|S|) \sum_{(x,y)\in S} \mathbb{1}[h(x) \neq y]$ the average number of mistakes that h makes on S; note that $\rho_X(h, S_{X,h'}) = \mathrm{err}(h, S_{X,h'})$. Finally, we informally use the $\tilde{O}(\cdot)$ notation to hide $\log(1/\delta)$ factors, as well as factors that are logarithmic in those that do appear.

3 Lower bounds for learning with α-privacy

In this section, we show a lower bound on the sample requirement of learning with α-privacy. In particular, we show an example which illustrates that when the data is drawn from a continuous distribution, for any M, all α-private algorithms that are supplied with at most M examples fail to output a good classifier for at least one unlabeled data distribution.

Our example hypothesis class is the class of thresholds on $[0, 1]$. This simple class has VC dimension 1, and thus can be learnt non-privately with classification error ϵ given only $\tilde{O}(1/\epsilon)$ examples, regardless of the unlabeled data distribution. However, Theorem 1 shows that even in the realizable case, for every α-private algorithm that is given a bounded number of examples, there is at least one unlabeled data distribution on which the learning algorithm produces a classifier with error $\geq \frac{1}{5}$, with probability at least $1/2$ over its own random coins.

The key intuition behind our example is that if most of the unlabeled data is concentrated in a small region around the best classifier, then, even slightly perturbing the best classifier will result in a large classification error. As the process of ensuring differential privacy necessarily involves some perturbation, unless the algorithm has some prior public knowledge about the data distribution, the number of samples required to learn privately grows with growing concentration of the data around the best classifier. As illustrated by our theorem, this problem is not alleviated if the support of the unlabeled distribution is known; even if the data distribution has large support, a large fraction of the data can still lie in a region close to the best classifier.

Before we describe our example in detail, we first need a definition.

Definition 4. The class of *thresholds* on the unit interval is the class of functions $h_w : [0, 1] \to \{-1, 1\}$ such that:

$$h_w(x) := \begin{cases} 1 & \text{if } x \geq w \\ -1 & \text{otherwise.} \end{cases}$$

Theorem 1. *Let $M > 2$ be any number, and let $\mathcal{H}$ be the class of thresholds on the unit interval $[0, 1]$. For any α-private algorithm A that outputs a hypothesis $h \in \mathcal{H}$, there exists a distribution $\mathcal{P}$ on labelled examples with the following properties:*

1. *There exists a threshold $h^* \in \mathcal{H}$ which has classification error 0 with respect to $\mathcal{P}$.*
2. *For all samples S of size $m \leq M$ drawn from $\mathcal{P}$, with probability at least $1/2$ over the random coins of A, the hypothesis output by $A(S)$ has classification error at least $\frac{1}{5}$ with respect to $\mathcal{P}$.*
3. *The marginal $\mathcal{D}$ of $\mathcal{P}$ over the unlabeled data has support $[0, 1]$.*

Proof. Let $\eta = \frac{1}{6+4\exp(\alpha M)}$, and let $\mathcal{U}$ denote the uniform distribution over $[0, 1]$. Let $Z = \{\eta, 2\eta, \ldots, K\eta\}$, where $K = \lfloor 1/\eta \rfloor - 1$. We let $G_z = [z - \eta/3, z + \eta/3]$ for $z \in Z$, and let $\mathcal{G}_z \subset \mathcal{H}$ be the subset of thresholds: $\mathcal{G}_z = \{h_\tau | \tau \in G_z\}$. We note that $G_z \subseteq [0, 1]$ for all $z \in Z$.

For each $z \in Z$, we define a distribution $\mathcal{P}_z$ over labelled examples as follows. First, we describe the marginal $\mathcal{D}_z$ of $\mathcal{P}_z$ over the unlabeled data. A sample from $\mathcal{D}_z$ is drawn as follows. With probability $\frac{1}{2}$, x is drawn from $\mathcal{U}$; with probability $\frac{1}{2}$, it is drawn from uniformly from G_z. Now, an unlabeled example x drawn from $\mathcal{D}_z$ is labelled positive if $x \geq z$, and negative otherwise. We observe that for every such distribution $\mathcal{P}_z$, there exists a threshold, namely, h_z that has classification error 0; in addition, the support of $\mathcal{D}_z$ is $[0, 1]$. Moreover, there are $\lfloor \frac{1}{\eta} \rfloor - 1$ such distributions $\mathcal{P}_z$ in all, and $\lfloor \frac{1}{\eta} \rfloor - 1 \geq 5$.

We say that an α-private algorithm A *succeeds* on a sample S with respect to a distribution $\mathcal{P}$ if with probability $\frac{1}{2}$ over the random coins of A, the hypothesis output by $A(S)$ has classification error $< \frac{1}{5}$ over $\mathcal{P}$.

Suppose for the sake of contradiction, there exists an α-private algorithm A^* such that for all distributions $\mathcal{P}$, there exists at least one sample S of size $\leq M$ drawn from $\mathcal{P}$ such that A^* succeeds on S with respect to $\mathcal{P}$. Then, for all $z \in Z$, and for all $\mathcal{P}_z$, there exists some sample S_z of size $m \leq M$ drawn from $\mathcal{P}_z$ such that A^* succeeds on S_z with respect to $\mathcal{P}_z$.

As the G_z's are disjoint, we can write:

$$\Pr_{A^*}[A^*(S_z) \notin \mathcal{G}_z] \geq \sum_{z' \in Z \setminus \{z\}} \Pr_{A^*}[A^*(S_z) \in \mathcal{G}_{z'}]. \tag{1}$$

Any S_z differs from $S_{z'}$ by at most m labelled examples. As A^* is α-private, for any z', from Lemma 11, we can write:

$$\Pr_{A^*}[A^*(S_z) \in \mathcal{G}_{z'}] \geq e^{-\alpha m} \Pr_{A^*}[A^*(S_{z'}) \in \mathcal{G}_{z'}]. \tag{2}$$

If $A^*(S_{z'})$ lies outside $\mathcal{G}_{z'}$, $A^*(S_{z'})$ classifies at least $1/4$ fraction of the examples from $\mathcal{P}_{z'}$ incorrectly, and thus A^* cannot succeed on $S_{z'}$ with respect to $\mathcal{P}_{z'}$; therefore, by property of A^*, for any z',

$$\Pr_{A^*}[A^*(S_{z'}) \in \mathcal{G}_{z'}] \geq \frac{1}{2}. \tag{3}$$

Combining Equations (1), (2), and (3), we can write:

$$\Pr_{A^*}[A^*(S_z) \notin \mathcal{G}_z] \geq e^{-\alpha m} \cdot \sum_{z' \in Z \setminus \{z\}} \frac{1}{2} \geq e^{-\alpha m} \cdot \left(\frac{1}{\eta} - 2 \right) \cdot \frac{1}{2}.$$

Since $m \leq M$, the quantity on the RHS of the above equation is more than $\frac{2}{3}$. A^* therefore does not succeed on S_z with respect to $\mathcal{P}_z$, thus leading to a contradiction. $\square$

4 Upper bounds for learning with α-privacy

In this section, we show an upper bound on the sample requirement of learning with α-privacy by presenting a learning algorithm that works on infinite hypothesis classes over continuous data domains, under certain conditions on the hypothesis class and the data distribution. Our algorithm works in the non-realizable case, that is, when there may be no hypothesis in the target hypothesis class with 0 classification error.

A natural way to extend the algorithm of Kasiviswanathan et al. (2008) to an infinite hypothesis class $\mathcal{H}$ is to compute a suitable finite subset $\mathcal{G}$ of $\mathcal{H}$ which contains a hypothesis with low excess generalization error, and then use the exponential mechanism of McSherry and Talwar (2007) on $\mathcal{G}$. To ensure that a hypothesis with low error is indeed in $\mathcal{G}$, we would like $\mathcal{G}$ to be an ϵ-cover of the disagreement metric $(\mathcal{H}, \rho_{\mathcal{D}})$. In a non-private or label-private learning, we can compute such a $\mathcal{G}$ directly based on the unlabeled training examples; in our setting, the training examples themselves are sensitive, and this approach does not directly apply.

The key idea behind our algorithm is that instead of using the sensitive data to compute $\mathcal{G}$, we can use a reference distribution $\mathcal{U}$ that is known independently of the sensitive data. For instance, if the domain of the unlabeled data is bounded, then a reasonable choice for $\mathcal{U}$ is the uniform distribution over the domain. Our key observation is that if $\mathcal{U}$ is close to the unlabeled data distribution $\mathcal{D}$ according to a certain measure of closeness inspired by Dasgupta (2005) and Freund et al. (1997), then a cover of the disagreement metric on $\mathcal{H}$ with respect to $\mathcal{U}$ is a (possibly coarser) cover of the disagreement metric on $\mathcal{H}$ with respect to $\mathcal{D}$. Thus we can set $\mathcal{G}$ to be a fine cover of $(\mathcal{H}, \rho_{\mathcal{U}})$, and this cover can be computed privately as it is independent of the sensitive data.

Our algorithm works when the doubling dimension of $(\mathcal{H}, \rho_{\mathcal{U}})$ is finite; under this condition, there is always such a finite cover $\mathcal{G}$. We note that this is a fairly weak condition, which is satisfied by many hypothesis classes and data distributions. For example, any hypothesis class with finite VC dimension will satisfy this condition for any unlabeled data distribution $\mathcal{U}$.

Finally, it may be tempting to think that one can further improve the sample requirement of our algorithm by using the sensitive data to privately refine a cover of $(\mathcal{H}, \rho_{\mathcal{U}})$ to a cover of $(\mathcal{H}, \rho_{\mathcal{D}})$. However, our calculations show that naively refining such a cover leads to a much higher sample requirement.

We now define our notion of closeness.

Definition 5. We say that a data distribution $\mathcal{D}$ is *κ-smooth with respect to a distribution $\mathcal{U}$* for some $\kappa \geq 1$, if for all measurable sets $A \subseteq \mathcal{X}$,

$$\Pr_{x\sim\mathcal{D}}[x \in A] \leq \kappa \cdot \Pr_{x\sim\mathcal{U}}[x \in A].$$

This notion of smoothness is very similar to, but weaker than the notions of closeness between distributions that have been used by (Dasgupta, 2005, Freund et al., 1997). We note that if $\mathcal{D}$ is absolutely continuous with respect to $\mathcal{U}$ (*i.e.*, $\mathcal{U}$ assigns zero probability to a set only if $\mathcal{D}$ does also), then $\mathcal{D}$ is κ-smooth with respect to $\mathcal{U}$ for some finite κ.

4.1 Algorithm

Our main algorithm $\mathcal{A}_1$ is given in Figure 1. The first step of the algorithm calculates the distance scale at which it should construct a cover of $(\mathcal{H}, \rho_{\mathcal{U}})$. This scale is a function of $|S|$, the size of the input data set S, and can be computed privately because $|S|$ is not sensitive information. A suitable cover of $(\mathcal{H}, \rho_{\mathcal{U}})$, which is also a suitable packing of $(\mathcal{H}, \rho_{\mathcal{U}})$ is then constructed; note that such a set always exists because of Lemma 2. In the final step, an exponential mechanism (McSherry and Talwar, 2007) is used to select a hypothesis from the cover with low error. As this step of the algorithm is the only one that uses the input data, the algorithm is α-private as long as this last step guarantees α-privacy.

4.2 Privacy and learning guarantees

Our first theorem states the privacy guarantee of Algorithm $\mathcal{A}_1$.

Theorem 2. *Algorithm $\mathcal{A}_1$ preserves α-privacy.*

Proof. The algorithm only accesses the private dataset S in the final step. Because changing one labeled example in S changes $\mathrm{err}(g, S)$ by at most 1, this step is guarantees α-privacy (McSherry and Talwar, 2007). □

The next theorem provides an upper bound on the sample requirement of Algorithm $\mathcal{A}_1$. This bound depends on the doubling dimension $d_{\mathcal{U}}$ of $(\mathcal{H}, \rho_{\mathcal{U}})$ and the smoothness parameter κ, as well as the privacy and learning parameters α, ϵ, δ.

Theorem 3. *Let $\mathcal{P}$ be a distribution over $\mathcal{X} \times \{\pm 1\}$ whose marginal over $\mathcal{X}$ is $\mathcal{D}$. There exists a universal constant $C > 0$ such that for any $\alpha, \epsilon, \delta \in (0, 1)$, the following holds. If*

Algorithm $\mathcal{A}_1$.
Input: private labeled dataset $S \subseteq \mathcal{X} \times \{\pm 1\}$, public reference distribution $\mathcal{U}$ over $\mathcal{X}$, privacy parameter $\alpha \in (0,1)$, accuracy parameter $\epsilon \in (0,1)$, confidence parameter $\delta \in (0,1)$.
Output: $h_{\mathcal{A}} \in \mathcal{H}$.

1. Solve the following equation to compute $\hat{\kappa} > 0$:

$$|S| = C \cdot \left(\frac{1}{\alpha\epsilon} + \frac{1}{\epsilon^2}\right) \cdot \left(d_{\mathcal{U}} \cdot \log \frac{\hat{\kappa}}{\epsilon} + \log \frac{1}{\delta}\right),$$

   where C is the constant from Theorem 3; let $\varepsilon_0 := \epsilon/(4\hat{\kappa})$.

2. Let $\mathcal{G}$ be an ε_0-packing of $(\mathcal{H}, \rho_{\mathcal{U}})$ that is also an ε_0-cover.

3. Randomly choose $h_{\mathcal{A}} \in \mathcal{G}$ according to the distribution $(p_g : g \in \mathcal{G})$, where $p_g \propto \exp(-\alpha|S|\operatorname{err}(g,S)/2)$ for each $g \in \mathcal{G}$, and return $h_{\mathcal{A}}$.

Figure 1: Learning algorithm for α-privacy.

1. *the doubling dimension $d_{\mathcal{U}}$ of $(\mathcal{H}, \rho_{\mathcal{U}})$ is finite,*
2. *$\mathcal{D}$ is κ-smooth with respect to $\mathcal{U}$,*
3. *$S \subseteq \mathcal{X} \times \{\pm 1\}$ is an i.i.d. random sample from $\mathcal{P}$ such that*

$$|S| \geq C \cdot \left(\frac{1}{\alpha\epsilon} + \frac{1}{\epsilon^2}\right) \cdot \left(d_{\mathcal{U}} \cdot \log \frac{\kappa}{\epsilon} + \log \frac{1}{\delta}\right), \tag{4}$$

then with probability at least $1-\delta$, the hypothesis $h_{\mathcal{A}} \in \mathcal{H}$ returned by $\mathcal{A}_1(S, \mathcal{U}, \alpha, \epsilon, \delta)$ satisfies

$$\Pr_{(x,y)\sim\mathcal{P}}[h_{\mathcal{A}}(x) \neq y] \leq \inf_{h' \in \mathcal{H}} \Pr_{(x,y)\sim\mathcal{P}}[h'(x) \neq y] + \epsilon.$$

The proof of Theorem 3 is stated in Appendix C. If we have prior knowledge that some hypothesis in $\mathcal{H}$ has zero error (the realizability assumption), then the sample requirement can be improved with a slightly modified version of Algorithm $\mathcal{A}_1$. This algorithm, called Algorithm $\mathcal{A}_{1r}$, is given in Figure 3 in Appendix C.

Theorem 4. *Let $\mathcal{P}$ be any probability distribution over $\mathcal{X} \times \{\pm 1\}$ whose marginal over $\mathcal{X}$ is $\mathcal{D}$. There exists a universal constant $C > 0$ such that for any $\alpha, \epsilon, \delta \in (0,1)$, the following holds. If*

1. *the doubling dimension $d_{\mathcal{U}}$ of $(\mathcal{H}, \rho_{\mathcal{U}})$ is finite,*
2. *$\mathcal{D}$ is κ-smooth with respect to $\mathcal{U}$,*
3. *$S \subseteq \mathcal{X} \times \{\pm 1\}$ is an i.i.d. random sample from $\mathcal{P}$ such that*

$$|S| \geq C \cdot \frac{1}{\alpha\epsilon} \cdot \left(d_{\mathcal{U}} \cdot \log(\kappa/\epsilon) + \log \frac{1}{\delta}\right), \tag{5}$$

4. *there exists $h^* \in \mathcal{H}$ for which $\Pr_{(x,y)\sim\mathcal{P}}[h^*(x) \neq y] = 0$,*

then with probability at least $1-\delta$, the hypothesis $h_{\mathcal{A}} \in \mathcal{H}$ returned by $\mathcal{A}_{1r}(S, \mathcal{U}, \alpha, \epsilon, \delta)$ satisfies

$$\Pr_{(x,y)\sim\mathcal{P}}[h_{\mathcal{A}}(x) \neq y] \leq \epsilon.$$

Again, the proof of Theorem 4 is in Appendix C.

4.3 Examples

In this section, we give some examples that illustrate the sample requirement of Algorithm $\mathcal{A}_1$.

First, we consider the example from the lower bound given in the proof of Theorem 1.

Example 1. The domain of the data is $\mathcal{X} := [0,1]$, and the hypothesis class is $\mathcal{H} := \mathcal{H}_{\text{thresholds}} = \{h_t : t \in [0,1]\}$ (recall, $h_t(x) = 1$ if and only if $x \geq t$). A natural choice for the reference distribution $\mathcal{U}$ is the uniform distribution over $[0,1]$; the doubling dimension of $(\mathcal{H}, \rho_{\mathcal{U}})$ is 1 because every interval can be covered

by two intervals of half the length. Fix some $M > 0$ and $\alpha \in (0,1)$, and let $\eta := 1/(6 + 4\exp(\alpha M))$. For $z \in [\eta, 1-\eta]$, let $\mathcal{D}_z$ be the distribution on $[0,1]$ with density

$$p_{\mathcal{D}_z}(x) := \begin{cases} \frac{1}{2} + \frac{3}{4\eta} & \text{if } z - \eta/3 \le x \le z + \eta/3, \\ \frac{1}{2} & \text{if } 0 \le x < z - \eta/3 \text{ or } z + \eta/3 < x \le 1. \end{cases}$$

Clearly, $\mathcal{D}_z$ is κ-smooth with respect to $\mathcal{U}$ for $\kappa = \frac{1}{2} + \frac{3}{4\eta} = O(\exp(\alpha M))$. Therefore the sample requirement of Algorithm $\mathcal{A}_1$ to learn with α-privacy and excess generalization error ϵ is at most

$$C \cdot \left(\frac{1}{\epsilon\alpha} + \frac{1}{\epsilon^2}\right) \cdot \left(\alpha M + \log\frac{1}{\delta}\right)$$

which is $\tilde{O}(M)$ for constant ϵ, matching the lower bound from Theorem 1 up to constants.

Next, we consider two examples in which the domain of the unlabeled data $\mathcal{X} := \mathbb{S}^{n-1}$ is the uniform distribution on the unit sphere in $\mathbb{R}^n$:

$$\mathbb{S}^{n-1} := \{x \in \mathbb{R}^n : \|x\| = 1\}$$

and the target hypothesis class $\mathcal{H} := \mathcal{H}_{\text{linear}}$ is the class of linear separators that pass through the origin in $\mathbb{R}^n$:

$$\mathcal{H}_{\text{linear}} := \{h_w : w \in \mathbb{S}^{n-1}\} \quad \text{where } h_w(x) = 1 \text{ if and only if } w \cdot x \ge 0.$$

The examples will consider two different distributions over $\mathcal{X}$.

A natural reference data distribution in this setting is the uniform distribution over $\mathbb{S}^{n-1}$; this will be our reference distribution $\mathcal{U}$. It is known that $d_{\mathcal{U}} := \sup\{\text{ddim}_r(\mathcal{H}, \rho_{\mathcal{U}}) : r \ge 0\} = O(n)$ (Bshouty et al., 2009).

Example 2. We consider a case where the unlabelled data distribution $\mathcal{D}$ is concentrated near an equator of $\mathbb{S}^{n-1}$. More formally, for some vector $u \in \mathbb{S}^{n-1}$, and $\gamma \in (0,1)$, we let $\mathcal{D}$ be uniform over $W := \{x \in \mathbb{S}^{n-1} : |u \cdot x| \le \gamma\}$; in other words, the unlabelled data lies in a small band of width γ around the equator.

By Lemma 9 (see Appendix C), $\mathcal{D}$ is κ-smooth with respect to $\mathcal{U}$ for $\kappa = \frac{1}{1 - 2\exp(-n\gamma^2/2)}$. Thus the sample requirement of Algorithm $\mathcal{A}_1$ to learn with α-privacy and excess excess generalization error ϵ is at most

$$C \cdot \left(\frac{1}{\alpha\epsilon} + \frac{1}{\epsilon^2}\right) \cdot \left(n \cdot \log\left(\frac{1}{\epsilon} \cdot \frac{1}{1 - 2\exp(-n\gamma^2/2)}\right) + \log\frac{1}{\delta}\right).$$

When n is large and $\gamma \ge 1/\sqrt{n}$, this bound is $\tilde{O}(\frac{n}{\alpha\epsilon} + \frac{n}{\epsilon^2})$, where the $\tilde{O}$ notation hides factors logarithmic in $1/\delta$ and $1/\epsilon$.

Example 3. Now we consider the case where the unlabeled data lies on two diametrically opposite spherical caps. More formally, for some vector $u \in \mathbb{S}^{n-1}$, and $\gamma \in (0,1)$, we now let $\mathcal{D}$ be uniform over $\mathbb{S}^{n-1} \setminus W$, where $W := \{x \in \mathbb{S}^{n-1} : |u \cdot x| \le \gamma\}$; in other words, the unlabeled data lies *outside* a band of width γ around the equator.

By Lemma 10 (see Appendix C), $\mathcal{D}$ is κ-smooth with respect to $\mathcal{U}$ for $\kappa = \left(\frac{2}{1-\gamma}\right)^{\frac{n-1}{2}}$. Thus the sample requirement of Algorithm $\mathcal{A}_1$ is to learn with α-privacy and excess generalization error ϵ is at most:

$$C \cdot \left(\frac{1}{\alpha\epsilon} + \frac{1}{\epsilon^2}\right) \cdot \left(n^2 \cdot \log\frac{2}{1-\gamma} + n \cdot \log\frac{1}{\epsilon} + \log\frac{1}{\delta}\right).$$

Thus, for large n and constant $\gamma < 1$, the sample requirement of Algorithm $\mathcal{A}_1$ is $\tilde{O}(\frac{n^2}{\epsilon^2} + \frac{n^2}{\alpha\epsilon})$. So, even though the smoothness parameter κ is exponential in the dimension n, the sample requirement remains polynomial in n.

5 Lower bounds for learning with α-label privacy

In this section, we provide two lower bounds on the sample complexity of learning with α-label privacy. Our first lower bound holds when α and ϵ are small (that is, high privacy and high accuracy), and when the hypothesis class has bounded VC dimension V. If these conditions hold, then we show a lower bound of $\Omega(d/\epsilon\alpha)$ where d is the doubling dimension of the disagreement metric $(\mathcal{H}, \rho_{\mathcal{D}})$ at some scale.

The main idea behind our bound is to show that differentially private learning algorithms necessarily perform poorly when there is a large set of hypotheses such that every pair in the set labels approximately $1/\alpha$ examples differently. We then show that such large sets can be constructed when the doubling dimension of the disagreement metric $(\mathcal{H}, \rho_{\mathcal{D}})$ is high.

5.1 Main results

Theorem 5. *There exists a constant $c > 0$ such that the following holds. Let $\mathcal{H}$ be a hypothesis class with VC dimension $V < \infty$, $\mathcal{D}$ be a distribution over $\mathcal{X}$, X be an i.i.d. sample from $\mathcal{D}$ of size m, and $\mathcal{A}$ be a learning algorithm that guarantees α-label privacy and outputs a hypothesis in $\mathcal{H}$. Let $d := \mathrm{ddim}_{12\epsilon}(\mathcal{H}, \rho_{\mathcal{D}}) > 2$, and $d' := \inf\{\mathrm{ddim}_{12r}(\mathcal{H}, \rho_{\mathcal{D}}) : \epsilon \le r < \Delta/6\} > 2$. If*

$$\epsilon < c \cdot \left(\frac{\Delta}{V(1+\log(1/\Delta))}\right), \quad \alpha \le c \cdot \left(\frac{d'}{V\log(1/\epsilon)}\right), \quad \text{and} \quad m < c \cdot \left(\frac{d}{\alpha\epsilon}\right)$$

where Δ is the diameter of $(\mathcal{H}, \rho_{\mathcal{D}})$, then there exists a hypothesis $h^* \in \mathcal{H}$ such that with probability at least $1/8$ over the random choice of X and internal randomness of $\mathcal{A}$, the hypothesis $h_{\mathcal{A}}$ returned by $\mathcal{A}(S_{X,h^*})$ has classification error

$$\Pr_{x\sim\mathcal{D}}\left[h_{\mathcal{A}}(x) \neq h^*(x)\right] > \epsilon.$$

We note that the conditions on α and ϵ can be relaxed by replacing the VC dimension with other (possibly distribution-dependent) quantities that determine the uniform convergence of ρ_X to $\rho_{\mathcal{D}}$; we used a distribution-free parameter to simplify the argument. Moreover, the condition on ϵ can be reduced to $\epsilon < c$ for some constant $c \in (0,1)$ provided that there exists a lower bound of $\Omega(V/\epsilon)$ to (non-privately) learn $\mathcal{H}$ under the distribution $\mathcal{D}$.

The proof of Theorem 5, which is in Appendix D relies on the following lemma (possibly of independent interest) which gives a lower bound on the empirical error of the hypothesis returned by an α-label private learning algorithm.

Lemma 1. *Let $X \subseteq \mathcal{X}$ be an unlabeled dataset of size m, $\mathcal{H}$ be a hypothesis class, $\mathcal{A}$ be a learning algorithm that guarantees α-label privacy, and $s > 0$. Pick any $h_0 \in \mathcal{H}$. If P is an s-packing of $B_X(h_0, 4s) \subseteq \mathcal{H}$, and*

$$m < \frac{\log\left(\frac{|P|}{2} - 1\right)}{8\alpha s},$$

then there exists a subset $Q \subseteq P$ such that

1. *$|Q| \ge |P|/2$;*
2. *for all $h \in Q$, $\Pr_{\mathcal{A}}[\mathcal{A}(S_{X,h}) \notin B_X(h, s/2)] \ge 1/2$.*

The proof of Lemma 1 is in Appendix D. The next theorem shows a lower bound without restrictions on ϵ and α. Moreover, this bound also applies when the VC dimension of the hypothesis class is unbounded. However, we note that this bound is weaker in that it does not involve a $1/\epsilon$ factor, where ϵ is the accuracy parameter.

Theorem 6. *Let $\mathcal{H}$ be a hypothesis class, $\mathcal{D}$ be a distribution over $\mathcal{X}$, X be an i.i.d. sample from $\mathcal{D}$ of size m, and $\mathcal{A}$ be a learning algorithm that guarantees α-label privacy and outputs a hypothesis in $\mathcal{H}$. Let $d'' := \mathrm{ddim}_{4\epsilon}(\mathcal{H}, \rho_{\mathcal{D}}) \ge 1$. If $\epsilon \le \Delta/2$ and*

$$m \le \frac{(d''-1)\log 2}{\alpha}$$

where Δ is the diameter of $(\mathcal{H}, \rho_{\mathcal{D}})$, then there exists $h^* \in \mathcal{H}$ such that with probability at least $1/2$ over the random choice of X and internal randomness of $\mathcal{A}$, the hypothesis $h_{\mathcal{A}}$ returned by $\mathcal{A}(S_{X,h^*})$ has classification error

$$\Pr_{x\sim\mathcal{D}}\left[h_{\mathcal{A}}(x) \neq h^*(x)\right] > \epsilon.$$

In other words, any α-label private algorithm for learning a hypothesis in $\mathcal{H}$ with error at most $\epsilon \le \Delta/2$ must use at least $(d''-1)\log(2)/\alpha$ examples. Theorem 6 uses ideas similar to those in (Beimel et al., 2010), but the result is stronger in that it applies to α-label privacy and continuous data domains. A detailed proof is provided in Appendix D.

5.2 Example: linear separators in $\mathbb{R}^n$

In this section, we show an example that illustrates our label privacy lower bounds. Our example hypothesis class $\mathcal{H} := \mathcal{H}_{\text{linear}}$ is the class of linear separators over $\mathbb{R}^n$ that pass through the origin, and the unlabeled data distribution $\mathcal{D}$ is the uniform distribution over the unit sphere $\mathbb{S}^{n-1}$. By Lemma 14 (see Appendix D), the doubling dimension of $(\mathcal{H}, \rho_{\mathcal{D}})$ at any scale r is at least $n-2$. Therefore Theorem 5 implies that if α and ϵ are small enough, any α-label private algorithm $\mathcal{A}$ that correctly learns all hypotheses $h \in \mathcal{H}$ with error $\le \epsilon$ requires at least $\Omega(\frac{n}{\epsilon\alpha})$ examples. (In fact, the condition on ϵ can be relaxed to $\epsilon \le c$ for some constant $c \in (0,1)$, because $\Omega(n)$ examples are needed to even non-privately learn in this setting (Long, 1995).) We also observe that this bound is tight (except for a $\log(1/\delta)$ factor): as the doubling dimension of $\mathcal{D}$ is at most n, in the realizable case, Algorithm $\mathcal{A}_{1r}$ using $\mathcal{U} := \mathcal{D}$ learns linear separators with α-label privacy given $\tilde{O}(\frac{n}{\alpha\epsilon})$ examples.

Acknowledgements

KC would like to thank NIH U54 HL108460 for research support. DH was partially supported by AFOSR FA9550-09-1-0425, NSF IIS-1016061, and NSF IIS-713540.

References

R. Agrawal and R. Srikant. Privacy-preserving data mining. *SIGMOD Rec.*, 29(2):439–450, 2000. ISSN 0163-5808. doi: http://doi.acm.org/10.1145/335191.335438.

Lars Backstrom, Cynthia Dwork, and Jon M. Kleinberg. Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography. In Carey L. Williamson, Mary Ellen Zurko, Peter F. Patel-Schneider, and Prashant J. Shenoy, editors, *WWW*, pages 181–190. ACM, 2007. ISBN 978-1-59593-654-7.

K. Ball. An elementary introduction to modern convex geometry. In Silvio Levy, editor, *Flavors of Geometry*, volume 31. 1997.

B. Barak, K. Chaudhuri, C. Dwork, S. Kale, F. McSherry, and K. Talwar. Privacy, accuracy, and consistency too: a holistic solution to contingency table release. In *PODS*, pages 273–282, 2007.

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

A. Blum, C. Dwork, F. McSherry, and K. Nissim. Practical privacy: the sulq framework. In *PODS*, pages 128–138, 2005.

A. Blum, K. Ligett, and A. Roth. A learning theory approach to non-interactive database privacy. In R. E. Ladner and C. Dwork, editors, *STOC*, pages 609–618. ACM, 2008. ISBN 978-1-60558-047-0.

Nader H. Bshouty, Yi Li, and Philip M. Long. Using the doubling dimension to analyze the generalization of learning algorithms. *J. Comput. Syst. Sci.*, 75(6):323–335, 2009.

K. Chaudhuri, C. Dwork, S. Kale, F. McSherry, and K. Talwar. Learning concept classes with privacy. Manuscript, 2006.

K. Chaudhuri, C. Monteleoni, and A. Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 2011.

Kamalika Chaudhuri and Nina Mishra. When random sampling preserves privacy. In Cynthia Dwork, editor, *CRYPTO*, volume 4117 of *Lecture Notes in Computer Science*, pages 198–213. Springer, 2006. ISBN 3-540-37432-9.

Sanjoy Dasgupta. Coarse sample complexity bounds for active learning. In *NIPS*, 2005.

C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *3rd IACR Theory of Cryptography Conference,*, pages 265–284, 2006.

A. Evfimievski, J. Gehrke, and R. Srikant. Limiting privacy breaches in privacy preserving data mining. In *PODS*, pages 211–222, 2003.

Yoav Freund, H. Sebastian Seung, Eli Shamir, and Naftali Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28(2-3):133–168, 1997.

Srivatsava Ranjit Ganta, Shiva Prasad Kasiviswanathan, and Adam Smith. Composition attacks and auxiliary information in data privacy. In *KDD*, pages 265–273, 2008a.

Srivatsava Ranjit Ganta, Shiva Prasad Kasiviswanathan, and Adam Smith. Composition attacks and auxiliary information in data privacy. In *KDD*, pages 265–273, 2008b.

Anupam Gupta, Robert Krauthgamer, and James R. Lee. Bounded geometries, fractals, and low-distortion embeddings. In *FOCS*, pages 534–543, 2003.

Anupam Gupta, Moritz Hardt, Aaron Roth, and Jonathan Ullman. Privately releasing conjunctions and the statistical query barrier. In *STOC*, 2011.

Moritz Hardt and Kunal Talwar. On the geometry of differential privacy. In *STOC*, pages 705–714, 2010.

Rosie Jones, Ravi Kumar, Bo Pang, and Andrew Tomkins. "i know what you did last summer": query logs and user privacy. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 909–914, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-803-9. doi: http://doi.acm.org/10.1145/1321440.1321573.

S. A. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith. What can we learn privately? In *Proc. of Foundations of Computer Science*, 2008.

Shiva Prasad Kasiviswanathan and Adam Smith. A note on differential privacy: Defining resistance to arbitrary side information. *CoRR*, abs/0803.3946, 2008.

A. Kolmogorov and V. Tikhomirov. ϵ-entropy and ϵ-capacity of sets in function spaces. *Translations of the American Mathematical Society*, 17:277–364, 1961.

P. M. Long. On the sample complexity of pac learning halfspaces against the uniform distribution. *IEEE Transactions on Neural Networks*, 6(6):1556–1559, 1995.

A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. In *Proc. of ICDE*, 2006.

David A. McAllester. Some pac-bayesian theorems. In *COLT*, pages 230–234, 1998.

Frank McSherry and Ilya Mironov. Differentially private recommender systems: building privacy into the net. In *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 627–636, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-495-9. doi: http://doi.acm.org/10.1145/1557019.1557090.

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

Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *IEEE Symposium on Security and Privacy*, pages 111–125, Oakland, CA, USA., May 2008. IEEE Computer Society.

Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In David S. Johnson and Uriel Feige, editors, *STOC*, pages 75–84. ACM, 2007. ISBN 978-1-59593-631-8.

D. Pollard. *Convergence of Stochastic Processes*. Springer-Verlag, 1984.

Aaron Roth. Differential privacy and the fat-shattering dimension of linear queries. In *APPROX-RANDOM*, pages 683–695, 2010.

L. Sweeney. k-anonymity: a model for protecting privacy. *Int. J. on Uncertainty, Fuzziness and Knowledge-Based Systems*, 2002.

V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, 16(2):264–280, 1971.

Rui Wang, Yong Fuga Li, XiaoFeng Wang, Haixu Tang, and Xiao yong Zhou. Learning your identity and disease from research papers: information leaks in genome wide association study. In *ACM Conference on Computer and Communications Security*, pages 534–544, 2009.

Andrew Chi-Chih Yao. Protocols for secure computations (extended abstract). In *FOCS*, pages 160–164, 1982.

Shuheng Zhou, Katrina Ligett, and Larry Wasserman. Differential privacy with compression. In *Proc. of ISIT*, 2009.

A Metric spaces

Lemma 2 (Kolmogorov and Tikhomirov, 1961)**.** *For any metric space $(\mathcal{Z}, \rho)$ with diameter Δ, and any $\varepsilon \in (0, \Delta)$, there exists an ε-packing of $(\mathcal{Z}, \rho)$ that is also an ε-cover.*

Lemma 3 (Gupta, Krauthgamer, and Lee, 2003)**.** *For any $\varepsilon > 0$ and $r > 0$, if a metric space $(\mathcal{Z}, \rho)$ has doubling dimension d and $z \in \mathcal{Z}$, then every ε-packing of $(B(z, r), \rho)$ has cardinality at most $(4r/\varepsilon)^d$.*

Lemma 4. *Let $(\mathcal{Z}, \rho)$ be a metric space with diameter Δ, and $r \in (0, 2\Delta)$. If $\mathrm{ddim}_r(\mathcal{Z}, \rho) \geq d$, then there exists $z \in \mathcal{Z}$ such that $B(z, r)$ has an $(r/2)$-packing of size at least 2^d.*

Proof. Fix $r \in (0, 2\Delta)$ and a metric space $(\mathcal{Z}, \rho)$ with diameter Δ. Suppose that for every $z \in \mathcal{Z}$, every $(r/2)$-packing of $B(z, r)$ has size less than 2^d. For each $z \in \mathcal{Z}$, let P_z be an $(r/2)$-packing of $(B(z, r), \rho)$ that is also an $(r/2)$-cover—this is guaranteed to exist by Lemma 2. Therefore, for each $z \in \mathcal{Z}$, $B(z, r) \subseteq \bigcup_{z' \in P_z} B(z, r/2)$, and $|P_z| < 2^d$. This implies that $\mathrm{ddim}_r(\mathcal{Z}, \rho)$ is less than d. □

B Uniform convergence

Lemma 5 (Vapnik and Chervonenkis, 1971)**.** *Let $\mathcal{F}$ be a family of measurable functions $f : \mathcal{Z} \to \{0, 1\}$ over a space $\mathcal{Z}$ with distribution $\mathcal{D}_{\mathcal{Z}}$. Denote by $\mathbb{E}_Z[f]$ the empirical average of f over a subset $Z \subseteq \mathcal{Z}$. Let $\varepsilon_m := (4/m)(\log(\mathcal{S}_{\mathcal{F}}(2m)) + \log(4/\delta))$, where $\mathcal{S}_{\mathcal{F}}(n)$ is the n-th VC shatter coefficient with respect to $\mathcal{F}$. Let Z be an i.i.d. sample of size m from $\mathcal{D}_{\mathcal{Z}}$. With probability at least $1 - \delta$, for all $f \in \mathcal{F}$,*

$$\mathbb{E}[f] \;\geq\; \mathbb{E}_Z[f] - \min\left\{\sqrt{\mathbb{E}_Z[f]\varepsilon_m},\; \sqrt{\mathbb{E}[f]\varepsilon_m} + \varepsilon_m\right\}.$$

Also, with probability at least $1 - \delta$, for all $f \in \mathcal{F}$,

$$\mathbb{E}[f] \;\leq\; \mathbb{E}_Z[f] + \min\left\{\sqrt{\mathbb{E}[f]\varepsilon_m},\; \sqrt{\mathbb{E}_Z[f]\varepsilon_m} + \varepsilon_m\right\}.$$

Lemma 6. *Let $\mathcal{H}$ be a hypothesis class with VC dimension V. Fix any $\delta \in (0, 1)$, and let X be an i.i.d. sample of size $m \geq V/2$ from $\mathcal{D}$. Let $\varepsilon_m := (8V \log(2em/V) + 4\log(4/\delta))/m$. With probability at least $1 - \delta$, for all pairs of hypotheses $\{h, h'\} \subseteq \mathcal{H}$,*

$$\rho_{\mathcal{D}}(h, h') \;\geq\; \rho_X(h, h') - \min\left\{\sqrt{\rho_X(h, h')\varepsilon_m},\; \sqrt{\rho_{\mathcal{D}}(h, h')\varepsilon_m} + \varepsilon_m\right\}.$$

Also, with probability at least $1 - \delta$, for all pairs of hypotheses $\{h, h'\} \subseteq \mathcal{H}$,

$$\rho_X(h, h') \;\geq\; \rho_{\mathcal{D}}(h, h') - \sqrt{\rho_{\mathcal{D}}(h, h')\varepsilon_m}.$$

Proof. This is an immediate consequence of Lemma 5 as applied to the function class $\mathcal{F} := \{x \mapsto \mathbb{1}[h(x) \neq h'(x)] : h, h' \in \mathcal{H}\}$, which has VC shatter coefficients $\mathcal{S}_{\mathcal{F}}(2m) \leq \mathcal{S}_{\mathcal{H}}(2m)^2 \leq (2em/V)^{2V}$ by Sauer's Lemma. □

C Proofs from Section 4

C.1 Some lemmas

We first give two simple lemmas. The first one, Lemma 7 states some basic properties of the exponential mechanism.

Lemma 7 (McSherry and Talwar, 2007)**.** *Let I be a finite set of indices, and let $a_i \in \mathbb{R}$ for all $i \in I$. Define the probability distribution $p := (p_i : i \in I)$ where $p_i \propto \exp(-a_i)$ for all $i \in I$. If $j \in I$ is drawn at random according to p, then the following holds for any element $i_0 \in I$ and any $t \in \mathbb{R}$.*

1. *Let $i \in I$. If $a_i \geq t$, then $\Pr_{j \sim p}[j = i] \leq \exp(-(t - a_{i_0}))$.*
2. *$\Pr_{j \sim p}[a_j \geq a_{i_0} + t] \leq |I| \exp(-t)$.*

Proof. Fix any $i_0 \in I$ and $t \in \mathbb{R}$. To show the first part of the lemma, note that for any $i \in I$ with $a_i \geq t$, we have

$$\Pr_{j \sim p}[j = i] = \frac{\exp(-a_i)}{\sum_{i' \in I} \exp(-a_{i'})} \leq \frac{\exp(-t)}{\exp(-a_{i_0})} = \exp(-(t - a_{i_0})).$$

For the second part, we apply the inequality from the first part to all $i \in I$ such that $a_i \geq a_{i_0} + t$, so

$$\Pr_{j \sim p}[a_j \geq a_{i_0} + t] = \sum_{i \in I} \mathbb{1}[a_i \geq a_{i_0} + t] \cdot \Pr_{j \sim p}[j = i] \leq \sum_{i \in I} \mathbb{1}[a_i \geq a_{i_0} + t] \cdot \exp(-t) \leq |I| \exp(-t). \;\square$$

The next lemma is consequences of smoothness between distributions $\mathcal{D}$ and $\mathcal{U}$.

Lemma 8. *If $\mathcal{D}$ is κ-smooth with respect to $\mathcal{U}$, then for all $\varepsilon > 0$, every ε-cover of $(\mathcal{H}, \rho_{\mathcal{U}})$ is a $\kappa\varepsilon$-cover of $(\mathcal{H}, \rho_{\mathcal{D}})$.*

Proof. Suppose C is an ε-cover of $(\mathcal{H}, \rho_{\mathcal{U}})$. Then, for any $h \in \mathcal{H}$, there exists $h' \in C$ such that $\rho_{\mathcal{U}}(h, h') \leq \varepsilon$. Fix such a pair h, h', and let $A := \{x \in \mathcal{X} : h(x) \neq h'(x)\}$ be the subset of $\mathcal{X}$ on which h and h' disagree. As $\mathcal{D}$ is κ-smooth with respect to $\mathcal{U}$, by definition of smoothness,

$$\rho_{\mathcal{D}}(h, h') = \Pr_{x\sim\mathcal{D}}[x \in A] \leq \kappa \cdot \Pr_{x\sim\mathcal{D}}[x \in A] = \kappa \cdot \rho_{\mathcal{U}}(h, h') \leq \kappa\varepsilon,$$

and thus C is a $\kappa\varepsilon$-cover of $(\mathcal{H}, \rho_{\mathcal{D}})$. □

C.2 Proof of Theorem 3

First, because of the lower bound on $m := |S|$ from (4), the computed value of $\hat{\kappa}$ in the first step of the algorithm must satisfy $\hat{\kappa} \geq \kappa$. Therefore, $\mathcal{D}$ is also $\hat{\kappa}$-smooth with respect to $\mathcal{U}$. Combining this with Lemma 8, $\mathcal{G}$ is an $(\epsilon/4)$-cover of $(\mathcal{H}, \rho_{\mathcal{D}})$. Moreover, as $\mathcal{G}$ is also an $(\epsilon/4\hat{\kappa})$-packing of $\mathcal{U}$, from Lemma 3, the cardinality of $\mathcal{G}$ is at most $|\mathcal{G}| \leq (16\hat{\kappa}/\epsilon)^{d_{\mathcal{U}}}$.

Define $\operatorname{err}(h) := \Pr_{(x,y)\sim\mathcal{P}}[h(x) \neq y]$. Suppose that $h^* \in \mathcal{H}$ minimizes $\operatorname{err}(h)$ over $h \in \mathcal{H}$. Let $g_0 \in \mathcal{G}$ be an element of $\mathcal{G}$ such that $\rho_{\mathcal{D}}(h^*, g_0) \leq \epsilon/4$; g_0 exists as $\mathcal{G}$ is an $(\epsilon/4)$-cover of $(\mathcal{H}, \rho_{\mathcal{D}})$. By the triangle inequality, we have that:

$$\operatorname{err}(g_0) \leq \operatorname{err}(h^*) + \rho_{\mathcal{D}}(h^*, g_0) \leq \operatorname{err}(h^*) + \epsilon/4 \tag{6}$$

Let E be the event that $\max_{g\in\mathcal{G}} |\operatorname{err}(g) - \operatorname{err}(g, S)| > \epsilon/4$, and $\bar{E}$ be its complement. By Hoeffding's inequality, a union bound, and the lower bound on $|S|$, we have that for a large enough value of the constant C in Equation (4),

$$\Pr_{S\sim\mathcal{P}^m}[E] \leq |\mathcal{G}| \max_{g\in\mathcal{G}} \Pr_{S\sim\mathcal{P}^m}\left[|\operatorname{err}(g) - \operatorname{err}(g, S)| > \frac{\epsilon}{4}\right] \leq 2|\mathcal{G}| \exp\left(-\frac{|S|\epsilon^2}{32}\right) \leq \frac{\delta}{2}.$$

In the event $\bar{E}$, we have $\operatorname{err}(h_{\mathcal{A}}) \geq \operatorname{err}(h_{\mathcal{A}}, S) - \epsilon/4$ and $\operatorname{err}(g_0) \leq \operatorname{err}(g_0, S) + \epsilon/4$ because both $h_{\mathcal{A}}$ and g_0 are in $\mathcal{G}$. Therefore,

$$\begin{aligned}
\Pr_{S\sim\mathcal{P}^m,\mathcal{A}_1}[\operatorname{err}(h_{\mathcal{A}}) > \operatorname{err}(h^*) + \epsilon] &\leq \Pr_{S\sim\mathcal{P}^m,\mathcal{A}_1}\left[\operatorname{err}(h_{\mathcal{A}}) > \operatorname{err}(g_0) + \frac{3\epsilon}{4}\right] \\
&\leq \Pr_{S\sim\mathcal{P}^m}[E] + \Pr_{\mathcal{A}_1}\left[\operatorname{err}(h_{\mathcal{A}}) > \operatorname{err}(g_0) + \frac{3\epsilon}{4} \mid \bar{E}\right] \\
&\leq \frac{\delta}{2} + \Pr_{\mathcal{A}_1}\left[\operatorname{err}(h_{\mathcal{A}}, S) > \operatorname{err}(g_0, S) + \frac{\epsilon}{4} \mid \bar{E}\right] \\
&\leq \frac{\delta}{2} + |\mathcal{G}| \exp\left(-\frac{\alpha|S|\epsilon}{8}\right) \\
&\leq \frac{\delta}{2} + \left(\frac{16\hat{\kappa}}{\epsilon}\right)^{d_{\mathcal{U}}} \exp\left(-\frac{\alpha|S|\epsilon}{8}\right) \\
&\leq \delta
\end{aligned}$$

Here, the first step follows from (7), and the final three inequalities follow from Lemma 7 (using $a_g = \alpha|S| \operatorname{err}(g, S)/2$ for $g \in \mathcal{G}$), the upper bound on $|\mathcal{G}|$, and the lower bound on m in (4).

C.3 Proof of Theorem 4

The proof is very similar to the proof of Theorem 3.

First, because of the lower bound on $m := |S|$ from (5), the computed value of $\hat{\kappa}$ in the first step of the algorithm must satisfy $\hat{\kappa} \geq \kappa$. Therefore, $\mathcal{D}$ is also $\hat{\kappa}$-smooth with respect to $\mathcal{U}$. Combining this with Lemma 8, as $\mathcal{G}$ is an $(\epsilon/4\hat{\kappa})$-cover of $\mathcal{U}$, $\mathcal{G}$ is an $(\epsilon/4)$-cover of $(\mathcal{H}, \rho_{\mathcal{D}})$. Moreover, as $\mathcal{G}$ is also an $(\epsilon/4\hat{\kappa})$-packing of $\mathcal{U}$, from Lemma 3, the cardinality of $\mathcal{G}$ is at most $|\mathcal{G}| \leq (16\hat{\kappa}/\epsilon)^{d_{\mathcal{U}}}$.

Define $\operatorname{err}(h) := \Pr_{(x,y)\sim\mathcal{P}}[h(x) \neq y]$. Suppose that $h^* \in \mathcal{H}$ minimizes $\operatorname{err}(h)$ over $h \in \mathcal{H}$. Recall that from the realizability assumption, $\operatorname{err}(h^*) = 0$. Let $g_0 \in \mathcal{G}$ be an element of $\mathcal{G}$ such that $\rho_{\mathcal{D}}(h^*, g_0) \leq \epsilon/4$; g_0 exists as $\mathcal{G}$ is an $(\epsilon/4)$-cover of $(\mathcal{H}, \rho_{\mathcal{D}})$. By the triangle inequality, we have that:

$$\operatorname{err}(g_0) \leq \operatorname{err}(h^*) + \rho_{\mathcal{D}}(h^*, g_0) \leq \epsilon/4 \tag{7}$$

Algorithm $\mathcal{A}_{1r}$.
Input: private labeled dataset $S \subseteq \mathcal{X} \times \{\pm 1\}$, public reference distribution $\mathcal{U}$ over $\mathcal{X}$, privacy parameter $\alpha \in (0,1)$, accuracy parameter $\epsilon \in (0,1)$, confidence parameter $\delta \in (0,1)$.
Output: $h_{\mathcal{A}} \in \mathcal{H}$.

1. Solve the equation $|S| = f_{\text{realizable}}(\alpha, \epsilon, \delta, \hat{\kappa})$ for $\hat{\kappa} > 0$, where
$$f_{\text{realizable}}(\alpha, \epsilon, \delta, \hat{\kappa}) = C \cdot \frac{1}{\alpha\epsilon} \cdot \left(d_{\mathcal{U}} \cdot \log(\hat{\kappa}/\epsilon) + \log \frac{1}{\delta} \right)$$
is the function from (5), and let $\varepsilon_0 := \epsilon/(4\hat{\kappa})$.
2. Let $\mathcal{G}$ be an ε_0-packing of $(\mathcal{H}, \rho_{\mathcal{U}})$ that is also an ε_0-cover; the existence of such a set is guaranteed by Lemma 2.
3. Randomly choose $h_{\mathcal{A}} \in \mathcal{G}$ according to the distribution $(p_g : g \in \mathcal{G})$, where $p_g \propto \exp(-\alpha |S| \operatorname{err}(g, S)/2)$ for each $g \in \mathcal{G}$, and return $h_{\mathcal{A}}$.

Figure 2: Learning algorithm for α-privacy under the realizable assumption.

We define two events E_1 and E_2. Let $\mathcal{G}_1 \subset G$ be the set of all $g \in \mathcal{G}$ for which $\operatorname{err}(g) \geq \epsilon$. The event E_1 is the event that $\min_{g \in \mathcal{G}_1} \operatorname{err}(g, S) > 9\epsilon/10$, and let $\bar{E}_1$ be its complement. Applying the multiplicative Chernoff bounds, for a specific $g \in \mathcal{G}_1$,

$$\Pr_{S \sim \mathcal{P}^m} \left[\operatorname{err}(g, S) < \frac{9}{10} \operatorname{err}(g) \right] \leq e^{-|S| \operatorname{err}(g)/400} \leq e^{-|S|\epsilon/400}.$$

The quantity on the right hand side is at most $\frac{\delta}{4|\mathcal{G}|} \leq \frac{\delta}{4|\mathcal{G}_1|}$ for a large enough constant C in Equation (5). Applying an union bound over all $g \in \mathcal{G}_1$, we get that

$$\Pr_{S \sim \mathcal{P}^m}[\bar{E}_1] \leq \delta/4. \tag{8}$$

We define E_2 as the event that $\operatorname{err}(g_0, S) \leq 3\epsilon/4$, and $\bar{E}_2$ as its complement. From a standard multiplicative Chernoff bound, with probability at least $1 - \delta/4$,

$$\operatorname{err}(g_0, S) \leq \operatorname{err}(g) + \sqrt{\frac{3 \operatorname{err}(g) \ln(4/\delta)}{|S|}} \leq \frac{\epsilon}{4} + \sqrt{\frac{3\epsilon}{4} \cdot \frac{\ln(4/\delta)}{|S|}} \leq \frac{\epsilon}{4} + \sqrt{\frac{3\epsilon}{4} \cdot \frac{\epsilon}{3}} = \frac{3\epsilon}{4}$$

Thus, if $|S| \geq (3/\epsilon)\log(4/\delta)$, which is the case due to Equation (5),

$$\Pr_{S \sim \mathcal{P}^m}[\bar{E}_2] = \Pr_{S \sim \mathcal{P}^m} \left[\operatorname{err}(g_0, S) > \frac{3\epsilon}{4} \right] \leq \frac{\delta}{4}. \tag{9}$$

Therefore, we have

$$\begin{aligned}
\Pr_{S \sim \mathcal{P}^m, \mathcal{A}}[\operatorname{err}(h_{\mathcal{A}}) > \epsilon] &\leq \Pr_{S \sim \mathcal{P}^m, \mathcal{A}}[\operatorname{err}(h_{\mathcal{A}}) > \epsilon \mid E_1 \cap E_2] + \Pr_{S \sim \mathcal{P}^m}[\bar{E}_1] + \Pr_{S \sim \mathcal{P}^m}[\bar{E}_2] \\
&\leq \Pr_{S \sim \mathcal{P}^m, \mathcal{A}} \left[\operatorname{err}(h_{\mathcal{A}}, S) > \operatorname{err}(g_0, S) + \left(\frac{9}{10} - \frac{3}{4} \right) \epsilon \mid E_1 \cap E_2 \right] + \delta/4 + \delta/4 \\
&\leq \Pr_{S \sim \mathcal{P}^m, \mathcal{A}} \left[\operatorname{err}(h_{\mathcal{A}}, S) > \operatorname{err}(g_0, S) + \frac{3\epsilon}{20} \mid E_1 \cap E_2 \right] + \delta/2 \\
&\leq |\mathcal{G}| \exp\left(-\frac{3\epsilon|S|}{20} \right) + \delta/2 \\
&\leq \left(\frac{16\hat{\kappa}}{\epsilon} \right)^{d_{\mathcal{U}}} \exp\left(-\frac{3\epsilon|S|}{20} \right) + \delta/2 \\
&\leq \delta/2 + \delta/2 = \delta.
\end{aligned}$$

Here, the second step follows from the definition of events E_1 and E_2 and from Equations (8) and (9), the third step follows from simple algebra, the fourth step follows from Lemma 7, the fifth step from the bound on $|\mathcal{G}|$ and the final step from Equation (5).

C.4 Examples

Lemma 9. *Let $\mathcal{U}$ be uniform over the unit sphere $\mathbb{S}^{n-1}$, and let $\mathcal{D}$ be defined as in Example 2. Then, $\mathcal{D}$ is*

$$\frac{1}{1-2\exp(-n\gamma^2/2)}\text{-smooth}$$

with respect to $\mathcal{U}$.

Proof. From (Ball, 1997), we know that $\Pr_{x\sim\mathcal{U}}[x \in W] \geq 1-2\exp(-n\gamma^2/2)$. Thus, for any set $A \subseteq \mathbb{S}^{n-1}$, we have

$$\Pr_{x\sim\mathcal{D}}[x \in A] = \Pr_{x\sim\mathcal{D}}[x \in A \cap W] = \frac{\Pr_{x\sim\mathcal{U}}[x \in A \cap W]}{\Pr_{x\sim\mathcal{U}}[x \in W]} \leq \frac{\Pr_{x\sim\mathcal{U}}[x \in A]}{1-2\exp(-n\gamma^2/2)}.$$

This means $\mathcal{D}$ is κ-smooth with respect to $\mathcal{U}$ for

$$\kappa = \frac{1}{1-2\exp(-n\gamma^2/2)}.$$ □

Lemma 10. *Let $\mathcal{U}$ be uniform over the unit sphere $\mathbb{S}^{n-1}$ and let $\mathcal{D}$ be defined as in Example 3. Then, $\mathcal{D}$ is*

$$\left(\frac{2}{1-\gamma}\right)^{\frac{n-1}{2}}\text{-smooth}$$

with respect to $\mathcal{U}$.

Proof. From (Ball, 1997), we know that $\Pr_{x\sim\mathcal{U}}[x \in \mathbb{S}^{n-1} \setminus W] = \Pr_{x\sim\mathcal{U}}[x \notin W] \geq ((1-\gamma)/2)^{(n-1)/2}$. Therefore, for any $A \subseteq \mathbb{S}^{n-1}$, we have

$$\Pr_{x\sim\mathcal{D}}[x \in A] = \Pr_{x\sim\mathcal{D}}[x \in A \setminus W] = \frac{\Pr_{x\sim\mathcal{U}}[x \in A \setminus W]}{\Pr_{x\sim\mathcal{U}}[x \in \mathbb{S}^{n-1} \setminus W]} \leq \frac{\Pr_{x\sim\mathcal{U}}[x \in A]}{\left(\frac{1-\gamma}{2}\right)^{\frac{n-1}{2}}}.$$

This means $\mathcal{D}$ is κ-smooth with respect to $\mathcal{U}$ for

$$\kappa = \left(\frac{2}{1-\gamma}\right)^{\frac{n-1}{2}}.$$ □

D Proofs from Section 5

D.1 Some lemmas

Lemma 11. *Let $S := \{(x_1, y_1), \ldots, (x_m, y_m)\} \subseteq \mathcal{X} \times \{\pm 1\}$ be a labeled dataset of size m, $\alpha \in (0,1)$, and $k \geq 0$.*

1. *If a learning algorithm $\mathcal{A}$ guarantees α-privacy and outputs a hypothesis from $\mathcal{H}$, then for all $S' := \{(x'_1, y'_1), \ldots, (x'_m, y'_m)\} \subseteq \mathcal{X} \times \{\pm 1\}$ with $(x_i, y_i) = (x'_i, y'_i)$ for at least $|S| - k$ such examples,*

$$\forall \mathcal{G} \subseteq \mathcal{H} \text{ . } \Pr_{\mathcal{A}}[\mathcal{A}(S) \in \mathcal{G}] \geq \Pr_{\mathcal{A}}[\mathcal{A}(S') \in \mathcal{G}] \cdot \exp(-k\alpha).$$

2. *If a learning algorithm $\mathcal{A}$ guarantees α-label privacy and outputs a hypothesis from $\mathcal{H}$, then for all $S' := \{(x_1, y'_1), \ldots, (x_m, y'_m)\} \subseteq \mathcal{X} \times \{\pm 1\}$ with $y_i = y'_i$ for at least $|S| - k$ such labels,*

$$\forall \mathcal{G} \subseteq \mathcal{H} \text{ . } \Pr_{\mathcal{A}}[\mathcal{A}(S) \in \mathcal{G}] \geq \Pr_{\mathcal{A}}[\mathcal{A}(S') \in \mathcal{G}] \cdot \exp(-k\alpha).$$

Proof. We prove just the first part, as the second part is similar. For a labeled dataset S' that differs from S in at most k pairs, there exists a sequence of datasets $S^{(0)}, \ldots, S^{(\ell)}$ with $\ell \leq k$ such that $S^{(0)} = S'$, $S^{(\ell)} = S$, and $S^{(j)}$ differs from $S^{(j+1)}$ in exactly one example for $1 \leq j < \ell$. In this case, if $\mathcal{A}$ guarantees α-privacy, then for all $\mathcal{G} \subseteq \mathcal{H}$,

$$\Pr_{\mathcal{A}}\left[\mathcal{A}(S^{(0)}) \in \mathcal{G}\right] \leq \Pr_{\mathcal{A}}\left[\mathcal{A}(S^{(1)}) \in \mathcal{G}\right] \cdot e^{\alpha} \leq \Pr_{\mathcal{A}}\left[\mathcal{A}(S^{(2)}) \in \mathcal{G}\right] \cdot e^{2\alpha} \leq \ldots \leq \Pr_{\mathcal{A}}\left[\mathcal{A}(S^{(\ell)}) \in \mathcal{G}\right] \cdot e^{\ell\alpha}$$

and therefore

$$\Pr_{\mathcal{A}}[\mathcal{A}(S) \in \mathcal{G}] \geq \Pr_{\mathcal{A}}[\mathcal{A}(S') \in \mathcal{G}] \cdot e^{-\ell\alpha} \geq \Pr_{\mathcal{A}}[\mathcal{A}(S') \in \mathcal{G}] \cdot e^{-k\alpha}.$$ □

Lemma 12. *There exists a constant $C > 1$ such that the following holds. Let $\mathcal{H}$ be a hypothesis class with VC dimension V, and let $\mathcal{D}$ be distribution over $\mathcal{X}$. Fix any $r \in (0,1)$, and let X be an i.i.d. sample of size m from $\mathcal{D}$. If*

$$m \geq \frac{CV}{r} \log \frac{C}{r},$$

then the following holds with probability at least $1/2$:

1. *every pair of hypotheses $\{h, h'\} \subseteq \mathcal{H}$ for which $\rho_X(h, h') > 2r$ has $\rho_{\mathcal{D}}(h, h') > r$;*
2. *for all $h_0 \in \mathcal{H}$, every $(6r)$-packing of $(B_{\mathcal{D}}(h_0, 12r), \rho_{\mathcal{D}})$ is a $(4r)$-packing of $(B_X(h_0, 16r), \rho_X)$.*

Proof. This is a consequence of Lemma 6. To show the first part, we plug in Lemma 6 with $\varepsilon_m = r/2$.

To show the second part, we use two applications of Lemma 6. Let h and h' be any two hypotheses in any $(6r)$-packing of $(B_{\mathcal{D}}(h_0, 12r), \rho_{\mathcal{D}})$; we first use Lemma 6 with $\varepsilon_m = r/3$ to show that for all such h and h', $\rho_X(h, h') > 4r$. Next we need to show that all h in any $(6r)$-packing of $(B_{\mathcal{D}}(h_0, 12r), \rho_{\mathcal{D}})$ has $\rho_X(h, h_0) \leq 16r$; we show this through a second application of Lemma 6 with $\varepsilon_m = r/3$. □

D.2 Proof of Theorem 6

We prove the contrapositive: that if $\epsilon \leq \Delta/2$ and $\Pr_{X \sim \mathcal{D}^m, \mathcal{A}}[\mathcal{A}(S_{X,h^*}) \in B_{\mathcal{D}}(h^*, \epsilon)] > 1/2$ for all $h^* \in \mathcal{H}$, then $m > \log(2^{d''-1})/\alpha$. So pick any $\epsilon \leq \Delta/2$. By Lemma 4, there exists an $h_0 \in \mathcal{H}$ and $P \subseteq \mathcal{H}$ such that P is a (2ϵ)-packing of $(B_{\mathcal{D}}(h_0, 4\epsilon), \rho_{\mathcal{D}})$ of size $\geq 2^{d''}$. For any $h, h' \in P$ such that $h \neq h'$, we have $B_{\mathcal{D}}(h, \epsilon) \cap B_{\mathcal{D}}(h', \epsilon) = \emptyset$ by the triangle inequality. Therefore for any $h \in P$ and any $X' \subseteq \mathcal{X}$ of size m,

$$\begin{aligned}
\Pr_{\mathcal{A}}\left[A(S_{X',h}) \notin B_{\mathcal{D}}(h, \epsilon)\right] &\geq \sum_{h' \in P \setminus \{h\}} \Pr_{\mathcal{A}}\left[A(S_{X',h}) \in B_{\mathcal{D}}(h', \epsilon)\right] \\
&\geq \sum_{h' \in P \setminus \{h\}} \Pr_{\mathcal{A}}\left[A(S_{X',h'}) \in B_{\mathcal{D}}(h', \epsilon)\right] \cdot e^{-\alpha m},
\end{aligned}$$

where the second inequality follows by Lemma 11 because $S_{X',h}$ and $S_{X',h'}$ can differ in at most (all) m labels. Now integrating both sides with respect to $X' \sim \mathcal{D}^m$ shows that if $\Pr_{X \sim \mathcal{D}^m, \mathcal{A}}[\mathcal{A}(S_{X,h^*}) \in B_{\mathcal{D}}(h^*, \epsilon)] > 1/2$ for all $h^* \in \mathcal{H}$, then for any $h \in P$,

$$\begin{aligned}
\frac{1}{2} > \Pr_{X \sim \mathcal{D}^m, \mathcal{A}}\left[A(S_{X,h}) \notin B_{\mathcal{D}}(h, \epsilon)\right] &\geq \sum_{h' \in P \setminus \{h\}} \Pr_{X \sim \mathcal{D}^m, \mathcal{A}}\left[A(S_{X',h'}) \in B_{\mathcal{D}}(h', \epsilon)\right] \cdot e^{-\alpha m} \\
&> (|P| - 1) \cdot \frac{1}{2} \cdot e^{-\alpha m}
\end{aligned}$$

which in turn implies $m > \log(|P| - 1)/\alpha \geq \log(2^{d''} - 1)/\alpha \geq \log(2^{d''-1})/\alpha$, as d'' is always ≥ 1.

D.3 Proof of Lemma 1

Let $h_0 \in \mathcal{H}$ and P be an s-packing of $B_X(h_0, 4s) \subseteq \mathcal{H}$. Say the algorithm $\mathcal{A}$ is *good for* h if $\Pr_{\mathcal{A}}[\mathcal{A}(S_{X,h}) \in B_X(h, s/2)] \geq 1/2$. Note that $\mathcal{A}$ is not good for $h \in P$ if and only if $\Pr_{\mathcal{A}}[\mathcal{A}(S_{X,h}) \notin B_X(h, s/2)] > 1/2$. Therefore, it suffices to show that if $\mathcal{A}$ is good for at least $|P|/2$ hypotheses in P, then $m \geq (\log((|P|/2) - 1)/(8\alpha s)$.

By the triangle inequality and the fact that P is an s-packing, $B_X(h, s/2) \cap B_X(h', s/2) = \emptyset$ for all $h, h' \in P$ such that $h \neq h'$. Therefore for any $h \in P$,

$$\Pr_{\mathcal{A}}\left[\mathcal{A}(S_{X,h}) \notin B_X(h, s/2)\right] \geq \sum_{h' \in P \setminus \{h\}} \Pr_{\mathcal{A}}\left[\mathcal{A}(S_{X,h}) \in B_X(h', s/2)\right].$$

Moreover, for all $h, h' \in P$, we have $\rho_X(h, h') \leq \rho_X(h_0, h) + \rho_X(h_0, h') \leq 8s$ by the triangle inequality, so $S_{X,h}$ and $S_{X,h'}$ differ in at most $8sm$ labels. Therefore Lemma 11 implies

$$\Pr_{\mathcal{A}}\left[\mathcal{A}(S_{X,h}) \in B_X(h', s/2)\right] \geq \Pr_{\mathcal{A}}\left[\mathcal{A}(S_{X,h'}) \in B_X(h', s/2)\right] \cdot e^{-8sm}$$

for all $h, h' \in P$. If $\mathcal{A}$ is good for at least $|P|/2$ hypotheses $h' \in P$, then for any $h \in P$ such that $\mathcal{A}$ is good for h, we have

$$\begin{aligned}
\frac{1}{2} \geq \Pr_{\mathcal{A}}\left[\mathcal{A}(S_{X,h}) \notin B_X(h, s/2)\right] &\geq \sum_{h' \in P \setminus \{h\}} \mathbb{1}[\mathcal{A} \text{ is good for } h] \cdot \Pr_{\mathcal{A}}\left[\mathcal{A}(S_{X,h'}) \in B_X(h', s/2)\right] \cdot e^{-8sm} \\
&\geq \sum_{h' \in P \setminus \{h\}} \mathbb{1}[\mathcal{A} \text{ is good for } h] \cdot \frac{1}{2} \cdot e^{-8sm} \geq \left(\frac{|P|}{2} - 1\right) \cdot \frac{1}{2} \cdot e^{-8sm}
\end{aligned}$$

which in turn implies $m \geq \log((|P|/2) - 1)/(8s)$.

D.4 Proof of Theorem 5

We need the following lemma.

Lemma 13. *There exists a constant $C > 1$ such that the following holds. Let $\mathcal{H}$ be a hypothesis class with VC dimension V, $\mathcal{D}$ be a distribution over $\mathcal{X}$, X be an i.i.d. sample from $\mathcal{D}$ of size m, $\mathcal{A}$ be a learning algorithm that guarantees α-label privacy and outputs a hypothesis in $\mathcal{H}$, and Δ be the diameter of $(\mathcal{H}, \rho_{\mathcal{D}})$. If $r \in (0, \Delta/6)$ and*

$$\frac{CV}{r}\log\frac{C}{r} \leq\ m\ < \frac{\log\left(2^{d-1}-1\right)}{32\alpha r}$$

where $d := \mathrm{ddim}_{12r}(\mathcal{H}, \rho_{\mathcal{D}})$, then there exists a hypothesis $h^* \in \mathcal{H}$ such that

$$\Pr_{X\sim\mathcal{D}^m,\mathcal{A}}\left[\mathcal{A}(S_{X,h^*}) \notin B_{\mathcal{D}}(h^*, r)\right] \geq \frac{1}{8}.$$

Proof. First, assume r and m satisfy the conditions in the lemma statement, where C is the constant from Lemma 12. Also, let $h_0 \in \mathcal{H}$ and $P \subseteq \mathcal{H}$ be such that P is a $(6r)$-packing of $(B_{\mathcal{D}}(h_0, 12r), \rho_{\mathcal{D}})$ of size $|P| \geq 2^d$; the existence of such an h_0 and P is guaranteed by Lemma 4.

We first define some events in the sample space of X and $\mathcal{A}$. For each $h \in \mathcal{H}$, and a sample X, let $E_1(h, X)$ be the event that

$\mathcal{A}(S_{X,h})$ makes more than $2rm$ mistakes on $S_{X,h}$ (*i.e.*, $\rho_X(h, \mathcal{A}(S_{X,h})) > 2r$).

Given a sample X, let $\phi(X)$ be a $0/1$ random variable which is 1 when the following conditions hold:

1. every pair of hypotheses $\{h, h'\} \subseteq \mathcal{H}$ for which $\rho_X(h, h') > 2r$ has $\rho_{\mathcal{D}}(h, h') > r$; and
2. for all $h_0 \in \mathcal{H}$, every $(6r)$-packing of $(B_{\mathcal{D}}(h_0, 12r), \rho_{\mathcal{D}})$ is a $(4r)$-packing of $(B_X(h_0, 16r), \rho_X)$

(*i.e.*, the conclusion of Lemma 12). Note that conditioned on $E_1(h, X)$ and $\phi(X) = 1$, we have $\rho_X(h, \mathcal{A}(S_{X,h})) > 2r$ and thus $\rho_{\mathcal{D}}(h, \mathcal{A}(S_{X,h})) > r$, so $\Pr_{X\sim\mathcal{D}^m,\mathcal{A}}[E_1(h, X), (\phi(X) = 1)] \leq \Pr_{X\sim\mathcal{D}^m,\mathcal{A}}[\rho_{\mathcal{D}}(h, \mathcal{A}(S_{X,h})) > r]$. Therefore it suffices to show that there exists $h^* \in \mathcal{H}$ such that $\Pr_{X\sim\mathcal{D}^m,\mathcal{A}}[E_1(h^*, X), (\phi(X) = 1)] \geq 1/8$.

The lower bound on m and Lemma 12 ensure that

$$\Pr_{X\sim\mathcal{D}^m}[\phi(X) = 1] \geq \frac{1}{2}. \tag{10}$$

Also, if the unlabeled sample X is such that $\phi(X) = 1$ holds, then the set P is a $(4r)$-packing of $(B_X(h_0, 16r), \rho_X)$. Therefore, the upper bound on m and Lemma 1 (with $s = 4r$) imply that for all such X, there exists $Q \subseteq P$ of size at least $|P|/2$ such that $\Pr_{\mathcal{A}}[E_1(h, X) \mid \phi(X) = 1] \geq 1/2$ for all $h \in Q$. In other words,

$$\sum_{h\in P}\Pr_{\mathcal{A}}\left[E_1(h, X) \mid \phi(X) = 1\right] = \sum_{h\in P}\mathbb{E}_{\mathcal{A}}\left[\mathbb{1}[E_1(h, X)] \mid \phi(X) = 1\right] \geq \frac{|P|}{4}. \tag{11}$$

Combining (10) and (11) gives

$$\begin{aligned}
\sum_{h\in P}\Pr_{X\sim\mathcal{D}^m,\mathcal{A}}\left[E_1(h, X), \phi(X) = 1\right] &= \sum_{h\in P}\mathbb{E}_{X\sim\mathcal{D}^m,\mathcal{A}}\left[\mathbb{1}[E_1(h, X)]|\phi(X) = 1\right]\cdot\Pr_{X\sim\mathcal{D}^m}[\phi(X) = 1] \\
&= \sum_{h\in P}\mathbb{E}_{X\sim\mathcal{D}^m}\mathbb{E}_{\mathcal{A}}\left[\mathbb{1}[E_1(h, X)]|\phi(X) = 1\right]\cdot\Pr_{X\sim\mathcal{D}^m}[\phi(X) = 1] \\
&= \mathbb{E}_{X\sim\mathcal{D}^m}\left[\sum_{h\in P}\mathbb{E}_{\mathcal{A}}\left[\mathbb{1}[E_1(h, X)] \mid \phi(X) = 1\right]\right]\cdot\Pr_{X\sim\mathcal{D}^m}[\phi(X) = 1] \\
&\geq \frac{|P|}{4}\Pr_{X\sim\mathcal{D}^m}\left[\phi(X) = 1\right] \\
&\geq \frac{|P|}{8}.
\end{aligned}$$

Here the first step follows because $\phi(X)$ is a $0/1$ random variable, the fourth step follows from Equation (11) and the fifth step follows from Equation (10).

Therefore there exists some $h^* \in P$ such that $\Pr_{X\sim\mathcal{D}^m,\mathcal{A}}[E_1(h^*, X), \phi(X) = 1] \geq 1/8$. □

Proof of Theorem 5. Assume

$$\epsilon < \frac{\Delta}{24CV\log(6C/\Delta)}, \quad \alpha \leq \frac{\log\left(2^{d'-1}-1\right)}{32CV\log(C/\epsilon)}, \quad \text{and} \quad m < \frac{\log\left(2^{d-1}-1\right)}{32\alpha\epsilon}$$

where C is the constant from Lemma 13. The proof is by case analysis, based on the value of m.

Case 1: $m < 1/(4\epsilon)$.

Since $\epsilon < \Delta/2$, Lemma 4 implies that there exists a pair $\{h, h'\} \subseteq \mathcal{H}$ such that $\rho_{\mathcal{D}}(h, h') > 2\epsilon$ but $\rho_{\mathcal{D}}(h, h') \leq 4\epsilon$. Using the bound on m and the fact $\epsilon \leq 1/5$, we have

$$\Pr_{X\sim\mathcal{D}^m}[\rho_X(h, h') = 0] \geq (1-4\epsilon)^m \geq (1-4\epsilon)^{\frac{1}{4\epsilon}} > \frac{1}{8}.$$

This means that $\Pr_{X\sim\mathcal{D}^m}[h_{\mathcal{A}} := \mathcal{A}(S_{X,h}) = \mathcal{A}(S_{X,h'})] \geq 1/8$. By the triangle inequality, $B_{\mathcal{D}}(h,\epsilon) \cap B_{\mathcal{D}}(h',\epsilon) = \emptyset$. So if, say, $\Pr_{X\sim\mathcal{D}^m,\mathcal{A}}[h_{\mathcal{A}} \in B_{\mathcal{D}}(h,\epsilon)] \geq 1/8$, then $\Pr_{X\sim\mathcal{D}^m,\mathcal{A}}[h_{\mathcal{A}} \notin B_{\mathcal{D}}(h',\epsilon)] \geq 1/8$. Therefore $\Pr_{X\sim\mathcal{D}^m,\mathcal{A}}[h_{\mathcal{A}} \notin B_{\mathcal{D}}(h^*,\epsilon)] \geq 1/8$ for at least one $h^* \in \{h, h'\}$.

Case 2: $1/(4\epsilon) \leq m < (CV/\epsilon)\log(C/\epsilon)$.

First, let $r > 0$ be the solution to the equation $(CV/r)\log(C/r) = m$, so $r > \epsilon$. Moreover, the bound on m and ϵ imply

$$m \geq \frac{1}{4\epsilon} > \frac{CV}{\Delta/6}\log\frac{C}{\Delta/6}$$

so $r < \Delta/6$. Finally, using the bound on α, definition of d', and fact $r > \epsilon$, we have

$$\alpha \leq \frac{\log\left(2^{d'-1}-1\right)}{32CV\log\frac{C}{\epsilon}} < \frac{\log\left(2^{d''-1}-1\right)}{32CV\log\frac{C}{r}}$$

where $d'' := \mathrm{ddim}_{12r}(\mathcal{H}, \rho_{\mathcal{D}})$; this implies

$$m = \frac{CV}{r}\log\frac{C}{r} < \frac{\log\left(2^{d''-1}-1\right)}{32\alpha r}.$$

The conditions of Lemma 13 are thus satisfied, so there exists $h^* \in \mathcal{H}$ such that $\Pr_{X\sim\mathcal{D}^m,\mathcal{A}}[\mathcal{A}(S_{X,h^*}) \notin B_{\mathcal{D}}(h^*, r)] \geq 1/8$.

Case 3: $(CV/\epsilon)\log(C/\epsilon) \leq m < \log(2^{d-1}-1)/(32\alpha\epsilon)$.

The conditions of Lemma 13 are satisfied in this case with $r := \epsilon < \Delta/6$, so there exists $h^* \in \mathcal{H}$ such that $\Pr_{X\sim\mathcal{D}^m,\mathcal{A}}[\rho_{\mathcal{D}}(h^*, \mathcal{A}(S_{X,h^*})) > \epsilon] \geq 1/8$. □

D.5 Example

The following lemma shows that if $\mathcal{D}$ is the uniform distribution on $\mathbb{S}^{n-1}$, then $\mathrm{ddim}_r(\mathcal{H}, \rho_{\mathcal{D}}) \geq n-2$ for all scales $r > 0$.

Lemma 14. *Let $\mathcal{H} := \mathcal{H}_{linear}$ be the class of linear separators through the origin in $\mathbb{R}^n$ and $\mathcal{D}$ be the uniform distribution on $\mathbb{S}^{n-1}$. For any $u \in \mathbb{S}^{n-1}$ and any $r > 0$, there exists an $(r/2)$-packing of $(B_{\mathcal{D}}(h_u, r), \rho_{\mathcal{D}})$ of size at least 2^{n-2}.*

Proof. Let μ be the uniform distribution over $\mathcal{H}$; notice that this is also the uniform distribution over $\mathbb{S}^{n-1}$.

We call a pair hypotheses h_v and h_w in $\mathcal{H}$ *close* if $\rho_{\mathcal{D}}(h_v, h_w) \leq r/2$. Observe that if any set of hypotheses has no close pairs, then it is an $(r/2)$-packing.

Using a technique due to Long (1995), we now construct an $(r/2)$-packing of $B_{\mathcal{D}}(h_u, r)$ by first randomly choosing hypotheses in $B_{\mathcal{D}}(h_u, r)$, and then removing hypotheses until no close pairs remain. First, we bound the probability p that two hypotheses h_v and h_w, chosen independently and uniformly at random from

Algorithm $\mathcal{A}_2$.
Input: private labeled dataset $S \subseteq \mathcal{X} \times \{\pm 1\}$, privacy parameter $\alpha \in (0,1)$, accuracy parameter $\epsilon \in (0,1)$, confidence parameter $\delta \in (0,1)$.
Output: $h_{\mathcal{A}} \in \mathcal{H}$.

1. Let $\mathcal{G}$ be a $(\epsilon/4)$-cover for $(\mathcal{H}, \rho_X)$ that is also an $(\epsilon/4)$-packing.
2. Randomly choose $h_{\mathcal{A}} \in \mathcal{G}$ according to the distribution $(p_g : g \in \mathcal{G})$, where $p_g \propto \exp(-\alpha|S|\operatorname{err}(g,S)/2)$ for each $g \in \mathcal{G}$, and return $h_{\mathcal{A}}$.

Figure 3: Learning algorithm for α-label privacy.

$B_{\mathcal{D}}(h_u, r)$, are close:

$$\begin{aligned}
p &= \Pr_{(h_v,h_w)\sim\mu^2}[\rho_{\mathcal{D}}(h_v,h_w) \le r/2 \mid h_v \in B_{\mathcal{D}}(h_u,r) \ \wedge \ h_w \in B_{\mathcal{D}}(h_u,r)] \\
&= \frac{\Pr_{(h_v,h_w)\sim\mu^2}[\rho_{\mathcal{D}}(h_v,h_w) \le r/2 \ \wedge \ h_v \in B_{\mathcal{D}}(h_u,r) \mid h_w \in B_{\mathcal{D}}(h_u,r)]}{\Pr_{h_v\sim\mu}[h_v \in B_{\mathcal{D}}(h_u,r)]} \\
&\le \frac{\Pr_{(h_v,h_w)\sim\mu^2}[\rho_{\mathcal{D}}(h_v,h_w) \le r/2 \mid h_w \in B_{\mathcal{D}}(h_u,r)]}{\Pr_{h_v\sim\mu}[h_v \in B_{\mathcal{D}}(h_u,r)]} \\
&= \frac{\Pr_{(h_v,h_w)\sim\mu^2}[h_v \in B_{\mathcal{D}}(h_w,r/2) \mid h_w \in B_{\mathcal{D}}(h_u,r)]}{\Pr_{h_v\sim\mu}[h_v \in B_{\mathcal{D}}(h_u,r)]} \\
&= \frac{\Pr_{(h_v,h_w)\sim\mu^2}[h_v \in B_{\mathcal{D}}(h_u,r/2)]}{\Pr_{h_v\sim\mu}[h_v \in B_{\mathcal{D}}(h_u,r)]} \\
&= 2^{-(n-1)}.
\end{aligned}$$

where the second-to-last equality follows by symmetry, and the last equality follows by the fact that $B_{\mathcal{D}}(h_u, r)$ corresponds to a $(n-1)$-dimensional spherical cap of $\mathbb{S}^{n-1}$. Now, choose $N := 2^{n-1}$ hypotheses $h_{v_1}, \ldots, h_{v_N}$ independently and uniformly at random from $B_{\mathcal{D}}(h_u, r)$. The expected number of close pairs among these N hypotheses is

$$M := \mathbb{E}\left[\sum_{i<j} \mathbb{1}[h_{v_i} \text{ and } h_{v_j} \text{ are close}]\right] = \sum_{i<j} \Pr\left[h_{v_i} \text{ and } h_{v_j} \text{ are close}\right] = \sum_{i<j} p \le \binom{N}{2} \cdot 2^{-(n-1)}.$$

Therefore, there exists N hypotheses $h_{v_1}, \ldots, h_{v_N}$ in $B_{\mathcal{D}}(h_u, r)$ among which there are at most M close pairs. Removing one hypothesis from each such close pair leaves a set of at least $N - M$ hypotheses with no close pairs—this is our $(r/2)$-packing of $B_{\mathcal{D}}(h_u, r)$. Since $N = 2^{n-1}$, the cardinality of this packing is at least

$$N - M \ge 2^{n-1} - \frac{2^{n-1}\left(2^{n-1}-1\right)}{2} \cdot 2^{-(n-1)} > 2^{n-2}. \qquad \square$$

E Upper bounds for learning with α-label privacy

Algorithm $\mathcal{A}_2$ for learning with α-label privacy, given in Figure 3, differs from the algorithms for learning with α-privacy in that it is able to use the unlabeled data itself to construct a finite set of candidate hypotheses. The algorithm and its analysis are very similar to work due to Chaudhuri et al. (2006); we give the details for completeness.

Theorem 7. *Algorithm $\mathcal{A}_2$ preserves α-label privacy.*

Proof. The algorithm only accesses the labels in S in the final step. It follows from standard arguments in (McSherry and Talwar, 2007) that α-label privacy is guaranteed. $\square$

Theorem 8. *Let $\mathcal{P}$ be any probability distribution over $\mathcal{X} \times \{\pm 1\}$ whose marginal over $\mathcal{X}$ is $\mathcal{D}$. There exists a universal constant $C > 0$ such that for any $\alpha, \epsilon, \delta \in (0,1)$, the following holds. If $S \subseteq \mathcal{X} \times \{\pm 1\}$ is an i.i.d. random sample from $\mathcal{P}$ of size*

$$m \ge C \cdot \left(\frac{\eta}{\epsilon^2} + \frac{1}{\epsilon}\right) \cdot \left(V \log\frac{1}{\epsilon} + \log\frac{1}{\delta}\right) + \frac{C}{\alpha\epsilon} \cdot \log\frac{\mathbb{E}_{X\sim\mathcal{D}^m}[\mathcal{N}_{\epsilon/8}(\mathcal{H}, \rho_X)]}{\delta}$$

where $\eta := \inf_{h' \in \mathcal{H}} \Pr_{(x,y)\sim\mathcal{P}}[h'(x) \neq y]$ *and* V *is the VC dimension of* $\mathcal{H}$*; then with probability at least* $1-\delta$*, the hypothesis* $h_{\mathcal{A}} \in \mathcal{H}$ *returned by* $\mathcal{A}_2(S, \alpha, \epsilon, \delta)$ *satisfies*

$$\Pr_{(x,y)\sim\mathcal{P}}[h_{\mathcal{A}}(x) \neq y] \leq \eta + \epsilon.$$

Remark 1. The first term in the sample size requirement (which depends on VC dimension) can be replaced by distribution-based quantities used for characterizing uniform convergence such as those based on l_1-covering numbers (Pollard, 1984).

Proof. Let $\mathrm{err}(h) := \Pr_{(x,y)\sim\mathcal{P}}[h(x) \neq y]$, and let $h^* \in \mathcal{H}$ minimize $\mathrm{err}(h)$ over $h \in \mathcal{H}$. Let $S := \{(x_1, y_1), \ldots, (x_m, y_m)\}$ be the i.i.d. sample drawn from $\mathcal{P}^m$, and $X := \{x_1, \ldots, x_m\}$ be the unlabeled components of S. Let $g_0 \in \mathcal{G}$ minimize $\mathrm{err}(g, S)$ over $g \in \mathcal{G}$. Since $\mathcal{G}$ is an $(\epsilon/4)$-cover for $(\mathcal{H}, \rho_X)$, we have that $\mathrm{err}(g_0, S) \leq \inf_{h' \in \mathcal{H}} \mathrm{err}(h', S) + \epsilon/4$. Since $\mathcal{G}$ is also an $(\epsilon/4)$-packing for $(\mathcal{H}, \rho_X)$, we have that $|\mathcal{G}| \leq \mathcal{N}_{\epsilon/8}(\mathcal{H}, \rho_X)$ (Pollard, 1984). Let $\mathcal{F} := \{f_h : h \in \mathcal{H}\}$ where $f_h(x, y) := \mathbb{1}[h(x) \neq y]$. We have $\mathbb{E}_{(x,y)\sim\mathcal{P}}[f_h(x, y)] = \mathrm{err}(h)$ and $m^{-1} \sum_{(x,y)\in S} f_h(x, y) = \mathrm{err}(h, S)$. Let E be the event that for all $h \in \mathcal{H}$,

$$\mathrm{err}(h, S) \leq \mathrm{err}(h) + \sqrt{\mathrm{err}(h)\varepsilon_m} + \varepsilon_m \quad \text{and} \quad \mathrm{err}(h) \leq \mathrm{err}(h, S) + \sqrt{\mathrm{err}(h)\varepsilon_m}$$

where $\varepsilon_m := (8V \log(2em/V) + 4\log(16/\delta))/m$. By Lemma 5, the fact $\mathcal{S}(\mathcal{F}, n) = \mathcal{S}(\mathcal{H}, n)$, and union bounds, we have $\Pr_{S\sim\mathcal{P}^m}[E] \geq 1 - \delta/2$. Now let E' be the event that

$$\mathrm{err}(h_{\mathcal{A}}, S) \leq \mathrm{err}(g_0, S) + t_m$$

where $t_m := 2\log(2\mathbb{E}_{X\sim\mathcal{D}^m}[|\mathcal{G}|]/\delta)/(\alpha m)$. The probability of E' can be bounded as

$$\begin{aligned}
\Pr_{S\sim\mathcal{P}^m,\mathcal{A}}[E'] &= 1 - \Pr_{S\sim\mathcal{P}^m,\mathcal{A}}[\mathrm{err}(h_{\mathcal{A}}, S) > \mathrm{err}(g_0, S) + t_m] \\
&= 1 - \mathbb{E}_{S\sim\mathcal{P}^m}\left[\Pr_{\mathcal{A}}\left[\mathrm{err}(h_{\mathcal{A}}, S) > \mathrm{err}(g_0, S) + t_m \mid S\right]\right] \\
&\geq 1 - \mathbb{E}_{S\sim\mathcal{P}^m}\left[|\mathcal{G}| \exp\left(-\frac{\alpha m t_m}{2}\right)\right] \\
&= 1 - \mathbb{E}_{X\sim\mathcal{D}^m}[|\mathcal{G}|] \cdot \exp\left(-\frac{\alpha m t_m}{2}\right) \\
&\geq 1 - \frac{\delta}{2}
\end{aligned}$$

where the first inequality follows from Lemma 7, and the second inequality follows from the definition of t_m. By the union bound, $\Pr_{S\sim\mathcal{P}^m,\mathcal{A}}[E \cap E'] \geq 1 - \delta$. In the event $E \cap E'$, we have

$$\begin{aligned}
\mathrm{err}(h_{\mathcal{A}}) - \mathrm{err}(h^*) &\leq \mathrm{err}(h_{\mathcal{A}}) - \mathrm{err}(h_{\mathcal{A}}, S) + \mathrm{err}(h^*, S) - \mathrm{err}(h^*) + \mathrm{err}(h_{\mathcal{A}}, S) - \mathrm{err}(h^*, S) \\
&\leq \sqrt{\mathrm{err}(h_{\mathcal{A}})\varepsilon_m} + \sqrt{\mathrm{err}(h^*)\varepsilon_m} + \varepsilon_m + \mathrm{err}(g_0, S) - \mathrm{err}(h^*, S) + t_m \\
&\leq \sqrt{\mathrm{err}(h_{\mathcal{A}})\varepsilon_m} + \sqrt{\mathrm{err}(h^*)\varepsilon_m} + \varepsilon_m + \epsilon/4 + t_m
\end{aligned}$$

since $\mathrm{err}(g_0, S) \leq \inf_{h' \in \mathcal{H}} \mathrm{err}(h', S) + \epsilon/4 \leq \mathrm{err}(h^*, S) + \epsilon/4$. By various algebraic manipulations, this in turn implies

$$\mathrm{err}(h_{\mathcal{A}}) \leq \mathrm{err}(h^*) + C' \cdot \left(\sqrt{\mathrm{err}(h^*)\varepsilon_m} + \varepsilon_m + t_m\right) + \epsilon/2$$

for some constant $C' > 0$. The lower bound on m now implies the theorem. □