Revisiting Model-Agnostic Private Learning: Faster Rates and Active Learning

Chong Liu
UC Santa Barbara

Yuqing Zhu
UC Santa Barbara

Kamalika Chaudhuri
UC San Diego

Yu-Xiang Wang
UC Santa Barbara

Abstract

The Private Aggregation of Teacher Ensembles (PATE) framework is one of the most promising recent approaches in differentially private learning. Existing theoretical analysis shows that PATE consistently learns any VC-classes in the realizable setting, but falls short in explaining its success in more general cases where the error rate of the optimal classifier is bounded away from zero. We fill in this gap by introducing the Tsybakov Noise Condition (TNC) and establish stronger and more interpretable learning bounds. These bounds provide new insights into when PATE works and improve over existing results even in the narrower realizable setting. We also investigate the compelling idea of using active learning for saving privacy budget. The novel components in the proofs include a more refined analysis of the majority voting classifier — which could be of independent interest — and an observation that the synthetic "student" learning problem is nearly realizable by construction under the Tsybakov noise condition.

1 INTRODUCTION

Differential privacy (DP) (Dwork et al., 2006) is one of the most popular approaches towards addressing the privacy challenges in the era of artificial intelligence and big data. While differential privacy is certainly not a solution to all privacy-related problems, many would agree that it represents a gold standard and is a key enabler in many applications (Machanavajjhala et al., 2008; Erlingsson et al., 2014; McMahan et al., 2018).

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

Recently, there has been an increasing demand in training machine learning and deep learning models with DP guarantees, which has motivated a growing body of research on this problem (Kasiviswanathan et al., 2011; Chaudhuri et al., 2011; Bassily et al., 2014; Wang et al., 2015; Abadi et al., 2016).

In a nutshell, differentially private machine learning aims at providing formal privacy guarantees that provably nullify the risk of identifying individual data points in the training data, while still allowing the learned model to be deployed and to provide accurate predictions. Many of these methods work well in low-dimensional regime when the model is small and the data is large. It however remains a fundamental challenge how to avoid the *explicit* dependence in the *ambient dimension* of the model and to develop practical methods in privately releasing deep learning models with a large number of parameters.

The "knowledge transfer" model of differentially private learning is a promising recent development (Papernot et al., 2017, 2018) which relaxes the problem by giving the learner access to a public unlabeled dataset. The main workhorse of this model is the Private Aggregation of Teacher Ensembles (PATE) framework:

The *PATE* Framework:

1. Randomly partition the private dataset into K splits.
2. Train one "teacher" classifier on each split.
3. Apply the K "teacher" classifiers on public data and *privately release* their majority votes as pseudo-labels.
4. Output the "student" classifier trained on the pseudo-labeled public data.

PATE achieves DP via the sample-and-aggregate scheme (Nissim et al., 2007) for releasing the pseudo-labels. Since the teachers are trained on disjoint splits of the private dataset, adding or removing one data point could affect only one of the teachers, hence limiting the influence of any single data point. The noise in-

Table 1: Summary of our results: excess risk bounds for PATE algorithms.

Algorithm	**PATE** (Gaussian mechanism) Papernot et al. (2017)	**PATE** (SVT-based) Bassily et al. (2018b)	**PATE** (SVT-based) This paper	**PATE** (Active Learning) This paper
Realizable	$\tilde{O}\Big(\max\Big\{\frac{d}{(n\epsilon)^{2/3}}, \frac{d}{m}\Big\}\Big)$	$\tilde{O}\Big(\max\Big\{\frac{d}{(n\epsilon)^{2/3}}, \sqrt{\frac{d}{m}}\Big\}\Big)$	$\tilde{O}\Big(\max\Big\{\frac{d^{1.5}}{n\epsilon}, \frac{d}{m}\Big\}\Big)$	$\tilde{O}\Big(\max\Big\{\frac{d^{1.5}\theta^{0.5}}{n\epsilon}, \frac{d}{m}\Big\}\Big)$
τ-TNC	$\tilde{O}\Big(\max\Big\{\big(\frac{d^{1.5}}{n\epsilon}\big)^{\frac{2\tau}{4-\tau}}, \frac{d}{m}\Big\}\Big)$	same as agnostic	$\tilde{O}\Big(\max\Big\{\big(\frac{d^{1.5}}{n\epsilon}\big)^{\frac{\tau}{2-\tau}}, \frac{d}{m}\Big\}\Big)$	$\tilde{O}\Big(\max\Big\{\big(\frac{d^{1.5}\theta^{0.5}}{n\epsilon}\big)^{\frac{\tau}{2-\tau}}, \frac{d}{m}\Big\}\Big)$
Agnostic (vs h^*)	$\Omega(\varepsilon(h^*))$ required.	$13\varepsilon(h^*) + \tilde{O}\Big(\max\Big\{\frac{d^{0.6}}{n^{0.4}\epsilon^{0.4}}, \sqrt{\frac{d}{m}}\Big\}\Big)$	$\Omega(\varepsilon(h^*))$ required.	$\Omega(\varepsilon(h^*))$ required.
Agnostic (vs h^{agg}_∞)	-	-	Consistent under weaker conditions.	-

- Results new to this paper are highlighted in blue. Hyperparameter K is chosen optimally. The number of public data points we privately label is chosen optimally (subsampling the available public data to run PATE) to minimize the risk bound.
- n and m denote the number of data points in the private and public dataset respectively. ϵ is the privacy budget for (ϵ, δ)-DP, with δ assumed to be in its typical range $\delta < 1/\text{poly}(n)$ and $\epsilon < \log(1/\delta)$. The TNC parameter τ ranges between $[0, 1]$. d denotes the *VC-dimension* of the hypothesis class, and θ denotes the *disagreement coefficient* (Hanneke, 2014). $\tilde{O}$ hides logarithmic terms in m, n, δ^{-1} and γ^{-1} where γ is the failure probability.
- Precise theorem statements of these results are found in Section 3.1, 3.2 and 3.3. Results about PATE (Gaussian mechanism) can be found in Appendix C.

jected in the aggregation will then be able to "obfuscate" the output and obtain provable privacy guarantees.

This approach is appealing in practice as it does not place any restrictions on the teachers, thus allowing any deep learning models to be used in a *model-agnostic* fashion. The competing alternative for differentially private deep learning, NoisySGD (Abadi et al., 2016), is not *model-agnostic*, and it requires significantly more tweaking and modifications to the model to achieve a comparable performance, (e.g., on MNIST), if achievable.

There are a lot of different DP mechanisms that could be used to instantiate the PATE Framework. Laplace mechanism and Gaussian mechanism are used in (Papernot et al., 2017, 2018) respectively. This paper builds upon the pioneering work of (Bassily et al., 2018b), which instantiates the PATE framework with a more data-adaptive scheme of private aggregation based on the Sparse Vector Technique (SVT). This approach allows the algorithm to privately label many examples while paying a privacy loss for only a small subset of them (see Algorithm 2 for details). Moreover, Bassily et al. (2018b) provides the first theoretical analysis of PATE which shows that it is able to PAC-learn any hypothesis classes with finite VC-dimension in the realizable setting. This is a giant leap from the standard differentially private learning models (without the access to a public unlabeled dataset) because the VC-classes are *not* privately learnable in general (Bun et al., 2015; Wang et al., 2016). Bassily et al. (2018b) also establishes a set of results on the agnostic learning setting, albeit less satisfying, as the excess risk, i.e., the error rate of the learned classifier relative to the optimal classifier, does not vanish as the number of data points increases.

In this paper, we revisit the problem of model-agnostic private learning under the PATE framework with several new analytical and algorithmic tools from the statistical learning theory including: the Tsybakov Noise Condition (TNC) (Mammen and Tsybakov, 1999), active learning (Hanneke, 2014), as well as the properties of voting classifiers. Our results are summarized in Table 1 and our contributions are:

1. We show that PATE consistently learns any VC-classes under TNC. When specializing to the realizable case, the sample complexity bound for achieving α-excess risk improves from $O(d^{1.5}/\alpha^{1.5}\epsilon)$ and $O(d/\alpha^2)$ to $O(d^{1.5}/\alpha\epsilon)$ and $O(d/\alpha)$ respectively on the private and public data.

2. We show that PATE learning is *inconsistent* for agnostic learning in general and derive new learning bounds that compete against a sequence of limiting majority voting classifiers.

3. We adapt the disagreement-based active learning algorithm to actively select which student queries to answer. Under TNC, we show that the active learning approach allows us to save the privacy budget exponentially when we use the standard privacy aggregation (Algorithm 1).

Related work and our novelty. Our work builds upon (Bassily et al., 2018b) and substantially improves the theoretical underpinning of PATE. To the best of our knowledge, these results are new and we are the first that consider *noise models* and *active learning* for PATE. Independent to our work, Alon et al. (2019) also studied the problem of private learning with access to an additional public dataset, and established that all VC-classes can be privately learned in the agnostic setting with optimal rates. Dagan and Feldman (2020); Nandi and Bassily (2020) established an analogous

result on a related but different setting of privacy-preserving predictions (Dwork and Feldman, 2018). The underlying idea of all three papers rely on an explicit (distribution-independent) α-net construction due to (Beimel et al., 2016) and exponential mechanism for producing pseudo-labels, which cannot be efficiently implemented. Our contributions are complementary as we focus on *oracle-efficient* algorithms that reduce to the learning bounds of ERM oracles (for passive learning) and active learning oracles. Our algorithms can therefore be implemented (and has been) in practice (Papernot et al., 2017, 2018). Due to space restrictions, further discussions are deferred to Appendix A.

2 PRELIMINARIES

In this section, we introduce the notations, definitions, and discuss specific technical tools that we will use throughtout this paper.

2.1 Symbols and Notations

We use $[n]$ to denote the set $\{1, 2, ..., n\}$. Let $\mathcal{X}$ denote the feature space, $\mathcal{Y} = \{0, 1\}$ denote the label, $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ to denote the sample space, and $\mathcal{Z}^* = \bigcup_{n\in\mathbb{N}} \mathcal{Z}^n$ to denote the space of a dataset of unspecified size. A hypothesis (classifier) h is a function mapping from $\mathcal{X}$ to $\mathcal{Y}$. A set of hypotheses $\mathcal{H} \subseteq \{0, 1\}^{\mathcal{X}}$ is called the hypothesis class. The VC dimension of $\mathcal{H}$ is denoted by d. Also, let $\mathcal{D}$ denote the distribution over $\mathcal{Z}$, and $\mathcal{D}_{\mathcal{X}}$ denote the marginal distribution over $\mathcal{X}$. $D^T = \{(x_i^T, y_i^T)|i \in [n]\} \sim \mathcal{D}$ is the labeled private teacher dataset, and $D^S = \{(x_j^S)|j \in [m]\} \sim \mathcal{D}_{\mathcal{X}}$ is the unlabeled public student dataset.

The expected risk of a certain hypothesis h with respect to the distribution $\mathcal{D}$ over $\mathcal{Z}$ is defined as $\varepsilon(h) = \mathbf{E}_{(x,y)\sim\mathcal{D}}[\mathbf{1}(h(x) \neq y)]$, where $\mathbf{1}(x)$ is the indicator function which equals to 1 when x is true, 0 otherwise. The empirical risk of a certain hypothesis h with respect to a dataset $\{(x_i, y_i)|i \in [n]\}$ is defined as $\hat{\varepsilon}(h) = \frac{1}{n}\sum_{i=1}^{n}[\mathbf{1}(h(x_i) \neq y_i)]$. The best hypothesis h^* is defined as $h^* = \operatorname{argmin}_{h\in\mathcal{H}} \varepsilon(h)$, and the Empirical Risk Minimizer (ERM) $\hat{h}$ is defined as $\hat{h} = \operatorname{argmin}_{h\in\mathcal{H}} \hat{\varepsilon}(h)$. $\hat{h}^{\text{agg}}$ is used to denote the aggregated classifier in the PATE framework. $\hat{h}^{\text{priv}}$ denotes the privately aggregated one. The expected disagreement between a pair of hypotheses h_1 and h_2 with respect to the distribution $\mathcal{D}_{\mathcal{X}}$ is defined as $\text{Dis}(h_1, h_2) = \mathbf{E}_{x\sim\mathcal{D}_{\mathcal{X}}}[\mathbf{1}(h_1(x) \neq h_2(x)]$. The empirical disagreement between a pair of hypotheses h_1 and h_2 with respect to a dataset $\{(x_i, y_i)|i \in [n]\}$ is defined as $\widehat{\text{Dis}}(h_1, h_2) = \frac{1}{n}\sum_{i=1}^{n}[\mathbf{1}(h_1(x_i) \neq h_2(x_i))]$. Throughout this paper, we use standard big O notations; and to improve the readability, we use $\lesssim$ and $\tilde{O}$ to hide poly-logarithmic factors.

2.2 Differential Privacy and Private Learning

Now we formally introduce differential privacy.

Definition 1 (Differential Privacy (Dwork and Roth, 2014)). *A randomized algorithm $\mathcal{M} : \mathcal{Z}^* \to \mathcal{R}$ is (ϵ, δ)-DP (differentially private) if for every pair of* neighboring *datasets $D, D' \in \mathcal{Z}^*$ (denoted by $\|D\Delta D'\|_1 = 1$) for all $\mathcal{S} \subseteq \mathcal{R}$:*

$$\mathbf{P}(\mathcal{M}(D) \in \mathcal{S}) \leq e^{\epsilon} \cdot \mathbf{P}(\mathcal{M}(D') \in \mathcal{S}) + \delta.$$

The definition says that if an algorithm $\mathcal{M}$ is DP, then no adversary can use the output of $\mathcal{M}$ to distinguish between two parallel worlds where an individual is in the dataset or not. ϵ, δ are privacy loss parameters that quantifies the strength of the DP guarantee. The closer they are to 0, the stronger the guarantee is.

The problem of DP learning aims at designing a randomized training algorithm that satisfies Definition 1. More often than not, the research question is about understanding the privacy-utility trade-offs and characterizing the Pareto optimal frontiers.

2.3 PATE and Model-Agnostic Private Learning

There are different ways we can instantiate the PATE framework to privately aggregate the teachers' predicted labels. The simplest, described in Algorithm 1, uses Gaussian mechanism to perturb the voting score.

Algorithm 1 Standard PATE (Papernot et al., 2018)

Input: "Teachers" $\hat{h}_1, ..., \hat{h}_K$ trained on *disjoint* subsets of the private data. "Nature" chooses an *adaptive* sequence of data points $x_1, ..., x_\ell$. Privacy parameters $\epsilon, \delta > 0$.

1: Find σ such that $\sqrt{\frac{2\ell\log(1/\delta)}{\sigma^2}} + \frac{\ell}{2\sigma^2} = \epsilon$.
2: Nature chooses x_1.
3: **for** $j \in [\ell]$ **do**
4: Output $\hat{y}_j \leftarrow \mathbf{1}(\sum_{k=1}^{K} \hat{h}_k(x_j) + \mathcal{N}(0, \sigma^2) \geq K/2)$.
5: Nature chooses x_{j+1} adaptively (as a function of the output vector till time j).
6: **end for**

An alternative approach due to (Bassily et al., 2018b) uses the Sparse Vector Technique (SVT) in a nontrivial way to privately label substantially more data points in the cases when teacher ensemble's predictions are *stable* for most input data. The stability is quantified in terms of the margin function, defined as

$$\text{margin}(x) := \left|2\sum_{k=1}^{K} \hat{h}_k(x) - K\right|,$$

which measures the absolute value of the difference between the number of votes (see Algorithm 2).

Algorithm 2 SVT-based PATE (Bassily et al., 2018b)

Input: "Teacher" classifiers $\hat{h}_1, ..., \hat{h}_K$ trained on *disjoint* subsets of the private data. "Nature" chooses an *adaptive* sequence of data points $x_1, ..., x_\ell$. Unstable cutoff T, privacy parameters $\epsilon, \delta > 0$.

Nature chooses x_1.
$\lambda \leftarrow (\sqrt{2T(\epsilon + \log(2/\delta))} + \sqrt{2T\log(2/\delta)})/\epsilon$.
$w \leftarrow 3\lambda \log(2(\ell + T)/\delta), \hat{w} \leftarrow w + \text{Lap}(\lambda)$.
$c = 0$.
for $j \in [\ell]$ **do**
  $\text{dist}_j \leftarrow \max\{0, \lceil \text{margin}(x_j)/2 \rceil - 1\}$.
  $\widehat{\text{dist}}_j \leftarrow \text{dist}_j + \text{Lap}(2\lambda)$.
  if $\widehat{\text{dist}}_j > \hat{w}$ **then**
    Output $\hat{y}_j \leftarrow \mathbf{1}(\sum_{k=1}^K \hat{h}_k(x_j) \geq K/2)$.
  else
    Output $\hat{y}_j \leftarrow \perp$.
    $c \leftarrow c + 1$, break if $c \geq T$.
    $\hat{w} \leftarrow w + \text{Lap}(\lambda)$.
  end if
  Nature chooses x_{j+1} adaptively (based on $\hat{y}_1, ..., \hat{y}_j$).
end for

In both algorithms, the privacy budget parameters ϵ, δ are taken as an input and the following privacy guarantee applies to all input datasets.

Theorem 2. *Algorithm 1 and 2 are both (ϵ, δ)-DP.*

Careful readers may note the slightly improved constants in the formula for calibrating privacy than when these methods were first introduced. We include the new proof based on the *concentrated differential privacy* (Bun and Steinke, 2016) approach in the Appendix C.

The key difference between the two private-aggregation mechanisms is that the standard PATE pays for a unit privacy loss for every public data point labeled, while the SVT-based PATE essentially pays only for those queries where the voted answer from the teacher ensemble is close to be unstable (those with a small margin). Combining this intuition with the fact that the individual classifiers are accurate — by the statistical learning theory, they are — the corresponding majority voting classifier can be shown to be accurate with a large margin. These two critical observations of (Bassily et al., 2018b) lead to the first learning theoretic guarantees for SVT-based PATE. For completeness, we include this result with a concise new proof in Appendix C.

Algorithm 3 PATE-PSQ

Input: Labeled private teacher dataset D^T, unlabeled public student dataset D^S, unstable query cutoff T, privacy parameters $\epsilon, \delta > 0$; number of splits K.

Randomly and evenly split the teacher dataset D^T into K parts $D_k^T \subseteq D^T$ where $k \in [K]$.
Train K classifiers $\hat{h}_k \in \mathcal{H}$, one from each part D_k^T.
Call Algorithm 2 with parameters $(\hat{h}_1, ..., \hat{h}_K), D^S, T, \epsilon, \delta$ and $\ell = m$ to obtain pseudo-labels for the public dataset $\hat{y}_1^S, ..., \hat{y}_m^S$. (Alternatively, call Algorithm 1 with parameters $(\hat{h}_1, ..., \hat{h}_K), D^S, \epsilon, \delta$)
For those pseudo labels that are $\perp$, assign them arbitrarily to $\{0, 1\}$.

Output: $\hat{h}^S$ trained on pseudo-labeled student dataset.

3 MAIN RESULTS

In Section 3.1 and 3.2, we present a more refined theoretical analysis of PATE with Passive Student Queries algorithm (PATE-PSQ, Algorithm 3) that uses SVT-based Algorithm 2 as the subroutine. Our results provide stronger learning bounds and new theoretical insights under various settings. In Section 3.3, we propose a new active learning based method and show that we can obtain qualitatively the same theoretical gain while using the simpler (an often more practical) Gaussian mechanism-based Algorithm 1 as the subroutine. For comparison, we also include an analysis of standard PATE (with Gaussian mechanism) in Appendix C. Table 1 summarizes these technical results.

3.1 Improved Learning Bounds under TNC

Recall that our motivation is to analyze PATE in the cases when the best classifier does not achieve 0 error and that the existing bound presented in Theorem C.3 is vacuous if $\varepsilon(h^*) > 1/26$. The error bound of $\hat{h}^S$ does not match the performance of h^* even as $m, n \to \infty$ and even if we output the voted labels without adding noise. This does not explain the empirical performance of Algorithm 3 reported in (Papernot et al., 2017, 2018) which demonstrates that the retrained classifier from PATE could get quite close to the best non-private baselines even if the latter are far from being perfect. For instance, on Adult dataset and SVHN dataset, the non-private baselines have accuracy 85% and 92.8% and PATE achieves 83.7% and 91.6% respectively.

To under stand how PATE works in the regime where the best classifier h^* obeys that $\varepsilon(h^*) > 0$, we introduce a large family of learning problems that satisfy the so-called the Tsybakov Noise Condition (TNC), under

which we show that PATE is consistent with fast rates. To understand TNC, we need to introduce a few more notations. Let label $y \in \{0, 1\}$ and define the regression function $r(x) = \mathbf{E}[y|x]$. The Tsybakov noise condition is defined in terms of the distribution of $r(x)$.

Definition 3 (Tsybakov noise condition). *The joint distribution of the data (x, y) satisfies the Tsybakov noise condition with parameter $\tau \in [0, 1]$ if there exists a universal constant $C > 0$ such that for all $t \geq 0$*

$$\mathbf{P}(|r(x) - 0.5| \leq t) \leq Ct^{\frac{\tau}{1-\tau}}.$$

Note that when $r(x) = 0.5$, the label is purely random and when $r(x) = 0$ or 1, y is a deterministic function of x. The Tsybakov noise condition essentially is reasonable "low noise" condition that does not require a uniform lower bound of $|r(x) - 0.5|$ for all x. When the label-noise is bounded for all x, e.g., when $y = h^*(x)$ with probability 0.6 and $1 - h^*(x)$ with probability 0.4, then the Tsybakov noise condition holds with $\tau = 1$. The case when $\tau = 1$ is also known as the *Massart noise condition* or *bounded noise condition* in the statistical learning literature.

For our purpose, it is more convenient to work with the following equivalent definition of TNC, which is equivalent to Definition 3 (see a proof from (Bousquet et al., 2004, Definition 7)).

Lemma 4 (Equivalent definition of TNC). *We say that a distribution of (x, y) satisfies the Tsybakov noise condition with parameter $\tau \in [0, 1]$* if and only if *there exists $\eta \in [1, \infty)$ such that, for every labeling function h,*

$$\mathrm{Dis}(h, h_{\mathrm{Bayes}}) \leq \eta(\varepsilon(h) - \varepsilon(h_{\mathrm{Bayes}}))^\tau. \qquad (1)$$

where $h_{\mathrm{Bayes}}(x) = \mathbf{1}(r(x) > 0.5)$ is the Bayes optimal classifier.

In the remainder of this section, we make the assumption that the Bayes optimal classifier $h_{\mathrm{Bayes}} \in \mathcal{H}$ and works with the slightly weaker condition that requires (1) to hold only for $h \in \mathcal{H}$ and that we replace h_{Bayes} by the optimal classifier $h^* \in \mathcal{H}$ [1].

We emphasize that the Tsybakov noise condition is not our invention. It has a long history from statistical learning theory to interpolate between the realizable setting and the agnostic setting. Specifically, problems satisfying TNC admit fast rates. For $\tau \in [0, 1]$, the empirical risk minimizer achieves an excess risk of $O(1/n^{1/(2-\tau)})$, which clearly interpolates the realizable case of $O(1/n)$ and the agnostic case of $O(1/\sqrt{n})$.

Next, we give a novel analysis of Algorithm 3 under TNC. The analysis is simple but revealing, as it not only avoids the strong assumption that requires $\varepsilon(h^*)$ to be close to 0, but also achieves a family of fast rates which significantly improves the sample complexity of PATE learning even for the realizable setting.

Theorem 5 (Utility guarantee of Algorithm 3 under TNC). *Assume the data distribution $\mathcal{D}$ and the hypothesis class $\mathcal{H}$ obey the Tsybakov noise condition with parameter τ. Then Algorithm 3 with*

$$T = \tilde{O}\left(\sqrt[4-3\tau]{\frac{m^{4-2\tau}d^{2\tau}}{n^{2\tau}\epsilon^{2\tau}}}\right)$$

$$K = O\left(\frac{\log(mT/\min(\delta, \beta))\sqrt{T\log(1/\delta)}}{\epsilon}\right)$$

obeys that with probability at least $1 - \beta$:

$$\varepsilon(\hat{h}) \leq \varepsilon(h^*) + \tilde{O}\left(\frac{d}{m} + \left(\frac{md^2}{n^2\epsilon^2}\right)^{\frac{\tau}{4-3\tau}}\right).$$

Remark 6 (Bounded noise case). *When $\tau = 1$, the Tsybakov noise condition is implied by the bounded noise assumption, a.k.a., Massart noise condition, where the labels are generated by the Bayes optimal classifier h^* and then toggled with a fixed probability less than 0.5. Theorem 5 implies that the excess risk is bounded by $\tilde{O}(\frac{d^2m}{n^2\epsilon^2} + \frac{d}{m})$, with $K = \tilde{O}(\frac{dm}{n\epsilon^2})$, which implies a sample complexity upper bound of $\tilde{O}(\frac{d^{3/2}}{\alpha\epsilon})$ private data points and $\tilde{O}(d/\alpha)$ public data points. The results improve over the sample complexity bound from (Bassily et al., 2018b) in the stronger realizable setting from $\tilde{O}(\frac{d^{3/2}}{\alpha^{3/2}\epsilon})$ and $\tilde{O}(d/\alpha^2)$ to $\tilde{O}(\frac{d^{3/2}}{\alpha\epsilon})$ and $\tilde{O}(d/\alpha)$ respectively in the private and public data.*

There are two key observations behind the improvement. First, the teacher classifiers do not have to agree on the labels y as in Lemma C.2; all they have to do is to agree on something for the majority of the data points. Conveniently, the Tsybakov noise condition implies that the teacher classifiers agree on the Bayes optimal classifier h^*. Second, when the teachers agree on h^*, the synthetic learning problem with the privately released pseudo-labels is nearly realizable. These intuitions can be formalized with a few lemmas, which will be used in the proof of Theorem 5. See Appendix B.

Remark 7 (Reduction to ERM). *The main challenge in the proof is to appropriately take care of $\hat{h}^{\mathrm{priv}}$. Although we are denoting it as a classifier, it is in fact a vector that is defined only on $x_1, ..., x_m$ rather than a general classifier that can take any input x. Since we*

[1] This slightly different condition, that requires (1) to hold only for $h \in \mathcal{H}$ but with h_{Bayes} replaced by the optimal classifier h^* (without assuming that $h^* = h_{\mathrm{Bayes}}$) is all we need. This is formally referred to as the Bernstein class condition by Hanneke (2014). Very confusingly, when the Tsybakov noise condition is being referred to in more recent literature, it is in fact the Bernstein class condition — a slightly weaker but more opaque definition about both the hypothesis class $\mathcal{H}$ and the data generating distribution.

are using the SVT-based aggregation Algorithm 2, $\hat{h}^{\text{priv}}$ is only well-defined for the student dataset. Moreover, these privately released "pseudo-labels" are not independent, which makes it infeasible to invoke a generic learning bound such as Lemma D.5. Our solution is to work with the empirical risk minimizer (ERM, rather than a generic PAC learner as a blackbox) and use uniform convergence (Lemma D.4) directly. This is without loss of generality because all learnable problems are learnable by (asymptotic) ERM (Vapnik, 1995; Shalev-Shwartz et al., 2010).

3.2 Challenges and New Bounds under Agnostic Setting

In this section, we present a more refined analysis of the agnostic setting. We first argue that agnostic learning with Algorithm 3 will not be consistent in general and competing against the best classifier in $\mathcal{H}$ seems not the right comparator. The form of the pseudo-labels mandate that $\hat{h}^S$ is aiming to fit a labeling function that is inherently a voting classifier. The literature on ensemble methods have taught us that the voting classifier is qualitatively different from the individual voters. In particular, the error rate of the majority voting classifier can be significantly better, about the same, or significantly worse than the average error rate of the individual voters. We illustrate this matter with two examples.

Example 8 (Voting fail)**.** *Consider a uniform distribution on $\mathcal{X} = \{x_1, x_2, x_3, x_4\}$ and that the corresponding label $\mathbf{P}(y = 1) = 1$. Let the hypothesis class be $\mathcal{H} = \{h_1, h_2, h_3\}$ whose evaluation on $\mathcal{X}$ are given in Figure 1. Check that the classification error of all three classifiers is 0.5. Also note that the empirical risk minimizer $\hat{h}$ will be a uniform distribution over h_1, h_2, h_3. The majority voting classifiers, learned with iid data sets, will perform significantly worse and converge to a classification error of 0.75 exponentially quickly as the number of classifiers K goes to ∞.*

	x_1	x_2	x_3	x_4	Error
y	1	1	1	1	0
h_1	1	1	0	0	0.5
h_2	1	0	1	0	0.5
h_3	1	0	0	1	0.5
$\hat{h}^{\text{agg}}$	1	0	0	0	0.75

Figure 1: An example where majority voting classifier is significantly worse than the best classifier in $\mathcal{H}$.

This example illustrates that the PATE framework cannot consistently learn a VC-class in the agnostic setting in general. On a positive note, there are also cases where the majority voting classifier boosts the classification accuracy significantly, such as the following example.

Example 9 (Voting win)**.** *If $\mathbf{P}[\hat{h}(x) \neq y|x] \leq 0.5 - \Delta$, where Δ is a small constant, for all $x \in \mathcal{X}$, then by Hoeffding's inequality,*

$$\mathbf{P}[\hat{h}^{\text{agg}}(x) \neq y|x] = \mathbf{P}\Big[\sum_{k=1}^{K} \mathbf{1}(\hat{h}_k(x) \neq y) \geq \frac{k}{2}\Big|x\Big] \leq e^{-2K\Delta^2}.$$

Thus the error goes to 0 exponentially as $K \to \infty$.

These cases call for an alternative distribution-dependent theory of learning that characterizes the performance of Algorithm 3 more accurately.

Next, we propose two changes to the learning paradigm. First, we need to go beyond $\mathcal{H}$ and compare with the following infinite ensemble classifier

$$h_\infty^{\text{agg}}(x) := \mathbf{1}\Big(\mathbf{E}\Big[\frac{1}{K}\sum_{k=1}^{k} \hat{h}_k(x)\Big|x\Big] \geq \frac{1}{2}\Big) = \mathbf{1}\Big(\mathbf{E}[\hat{h}_1(x)|x] \geq \frac{1}{2}\Big).$$

The classifier outputs the majority voting result of infinitely many independent teachers, each trained on n/K i.i.d. data points. As discussed earlier, this classifier can be better or worse than a single classifier $\hat{h}_1$ that takes n/K data points, $\hat{h}$ that trains on all n data points or h^* that is the optimal classifier in $\mathcal{H}$. Note that this classifier also changes as n/K gets larger.

Second, we define the *expected margin* for a classifier $\hat{h}_1$ trained on $\tilde{n}$ i.i.d. samples to be

$$\Delta_{\tilde{n}}(x) := \Big|\mathbf{E}[\hat{h}_1(x)|x] - \frac{1}{2}\Big|.$$

This quantity captures for a fixed $x \in \mathcal{X}$, how likely the teachers will agree. For a fixed learning problem $\mathcal{H}, \mathcal{D}$ and the number of i.i.d. data points $\hat{h}_1$ is trained upon, the expected margin is a function of x alone. The larger $\Delta_{n/K}(x)$ is, the more likely that the ensemble of K teachers agree on a prediction in $\mathcal{Y}$ with high-confidence. Note that unlike in Example 9, we do not require the teachers to agree on y. Instead, it measures the extent to which they agree on $h_\infty^{\text{agg}}(x)$, which could be any label.

When the expected margin is bounded away from 0 for x, then the voting classifier outputs $h_\infty^{\text{agg}}(x)$ with probability converging exponentially to 1 as K gets larger. On the technical level, this definition allows us to *decouple* the stability analysis and accuracy of PATE as the latter relies on how good h_∞^{agg} is.

Definition 10 (Approximate high margin)**.** *We say that a learning problem with n i.i.d. samples satisfy (ν, Δ)-approximate high-margin condition if $\mathbf{P}_{x\sim\mathcal{D}}[\Delta_n(x) > \Delta] \leq \nu$.*

This definition says that with high probability, except for $O(\nu m)$ data points, all other data points in the public dataset have an expected margin of at least Δ. Observe that every learning problem has Δ that increases from 0 to 0.5 as we vary ν from 0 to 1. The realizability assumption and the Tsybakov noise condition that we considered up to this point imply upper bounds of ν at fixed Δ (see more details in Remark 14). In Appendix F, we demonstrate that for the problem of linear classification on Aadult dataset — clearly an agnostic learning problem — (ν, Δ)-approximate high margin condition is satisfied with a small ν and large Δ.

The following proposition shows that when a problem is approximate high-margin, there are choices T and K under which the SVT-based PATE provably labels almost all data points with the output of h_∞^{agg}.

Proposition 11. *Assume the learning problem with n/K i.i.d. data points satisfies (ν, Δ)-approximate high-margin condition. Let Algorithm 2 be instantiated with parameter $T \geq \nu m + \sqrt{2\nu m \log(3/\gamma)} + \frac{2}{3}\log(3/\gamma)$ and $K \geq \max\{\frac{2\log(3m/\gamma)}{\Delta^2}, \frac{3\lambda\left(\log(4m/\delta)+\log(3m/\gamma)\right)}{\Delta}\}$[2]. Then with high probability (over the randomness of the n iid samples of the private dataset, m iid samples of the public dataset, and that of the randomized algorithm), Algorithm 2 finishes all m rounds and the output is the same as $h_\infty^{\text{agg}}(x_i)$ for all but T of the $i \in [m]$.*

This proposition provides the utility guarantee to Algorithm 2 and generalizes Lemma B.4 from fixing $\Delta = 1/6$ into allowing much smaller Δ at a cost of increasing ν.

Next, we state the learning bounds under the approximate-high margin condition.

Theorem 12. *Assume the learning problem with n/K iid data points satisfies (ν, Δ)-approximate high-margin condition and let K, T be chosen according to Proposition 11, furthermore assume that the privacy parameter of choice $\epsilon \leq \log(2/\delta)$, then the output classifier $\hat{h}^S$ of Algorithm 3 in the agnostic setting satisfies that with probability $\geq 1 - 2\gamma$*

$$\begin{aligned}\varepsilon(\hat{h}^S) - \varepsilon(h_\infty^{\text{agg}}) &\leq \min_{h\in\mathcal{H}} \text{Dis}(h, h_\infty^{\text{agg}}) + \frac{2T}{m} + \tilde{O}\Big(\sqrt{\frac{d}{m}}\Big) \\ &\leq \min_{h\in\mathcal{H}} \text{Dis}(h, h_\infty^{\text{agg}}) + 2\nu + \tilde{O}\Big(\sqrt{\frac{d}{m}}\Big).\end{aligned}$$

The voting classifier $\hat{h}^{\text{agg}}$ is usually not in the original hypothesis class $\mathcal{H}$, so we can take a wider view of the hypothesis class and define the voting hypothesis space $\text{Vote}(\mathcal{H})$ where the learning problem becomes realizable. Note if the VC dimension of $\mathcal{H}$ is d, then the VC dimension of $\text{Vote}_K(\mathcal{H}) \leq Kd$. In practice, this suggests using ensemble methods such as AdaBoost for K iterations.

Theorem 13. *Under the same assumption of Theorem 12. Suppose we train an ensemble classifier within the voting hypothesis space $\text{Vote}_K(\mathcal{H})$ in the student domain, then the output classifier $\hat{h}^S$ of Algorithm 3 in the agnostic setting satisfies that with probability $\geq 1 - 2\gamma$*

$$\begin{aligned}\varepsilon(\hat{h}^S) - \varepsilon(h_\infty^{\text{agg}}) &\leq \frac{4T}{m} + \frac{5(Kd + \log(4/\gamma))}{m} \\ &= \tilde{O}\Big(\nu + \frac{\log(4/\gamma)}{m} + \frac{d\sqrt{\nu}}{\Delta\sqrt{m}}\Big).\end{aligned}$$

Remark 14. *Whether the bounds in Theorem 12 and 13 will vanish as $m, n \to \infty$ depends strongly on how parameter ν and Δ change as n/K gets larger. Intuitively, if the learner converges to a single classifier h^*, as in the realizable case or under TNC, then we can show that the learning problem satisfy (ν, Δ)-approximate high-margin condition with $\Delta = 1/6$ and $\nu \leq \tilde{O}((dK/n)^{\frac{\tau}{2-\tau}})$. Substituting this quantities into Theorem 12 and using the fact that ν also bounds the disagreement between h^* and h_∞^{agg} allows us obtain a bound that vanishes as n gets larger. More generally, in the agnostic case, it is reasonable to assume that the "teachers" will get* more confident *in their individual prediction for most data points as $n/K \to \infty$. We argue this is a more modest requirement than requiring the "teachers" to get* more accurate.

3.3 PATE with Active Student Queries

In previous subsections, we have proved stronger learning bounds for PATE framework under TNC and in agnostic setting. However, all these results are based on the variant of PATE that aims passively releasing *almost all* student queries. In this section we address the following question:

Can we do better if we cherry-pick queries to label?

The hope is that this allows us to spend privacy budget only on those queries that add new information for the interest of training a classifier, hence resulting in a more favorable privacy-utility tradeoff. Without privacy constraints, this problem is known as active learning and it is often possible to save exponentially in the number of labels needed comparing to the passive learning model.

In Algorithm 4, we propose a new algorithm called PATE with Active Student Queries (PATE-ASQ) which uses the disagreement-based active learning algorithm

[2] $\lambda = (\sqrt{2T(\epsilon + \log(2/\delta))} + \sqrt{2T\log(2/\delta)})/\epsilon$ according to Algorithm 2.

Algorithm 4 PATE-ASQ

Input: Labeled private teacher dataset D^T, unlabeled public student dataset D^S, privacy parameters $\epsilon, \delta > 0$, number of splits K, maximum number of queries ℓ, failure probability γ.

1: Randomly and evenly split the teacher dataset D^T into K parts $D_k^T \subseteq D^T$ where $k \in [K]$
2: Train K classifiers $\hat{h}_k \in \mathcal{H}$, one from each part D_k^T.
3: Declare "Labeling Service" $\leftarrow$ Algorithm 1 with $\hat{h}_1, ..., \hat{h}_K$, ℓ, ϵ, δ, with an unspecified "nature".
4: Initiate an active learning oracle (e.g., Algorithm 7) with an iterator over D^S being the "data stream", hypothesis class $\mathcal{H}$, failure probability γ. Set the "labeling service" to be Algorithm 1 with parameter $\hat{h}_1, ..., \hat{h}_K$, ℓ, ϵ, δ, and set the "nature" to be the "request for label" calls in the active learning oracle.
5: Set $\hat{h}^S$ to be the "current output" from active learning oracle.

Output: Return $\hat{h}^S$.

(Algorithm 7 in Appendix E) as the subroutine. Then we provide its utility guarantee.

Theorem 15 (Utility guarantee of Algorithm 4)**.** *With probability at least $1-\gamma$, there exists universal constants C_1, C_2 such that for all*

$$\alpha \geq C_1 \max\left\{\eta^{\frac{2}{2-\tau}}\left(\frac{dK\log(n/d)+\log(2K/\gamma)}{n}\right)^{\frac{\tau}{2-\tau}}, \frac{d\log((m+n)/d)+\log(2/\gamma)}{m}\right\},$$

the output $\hat{h}^S$ of Algorithm 4 with parameter ℓ, K satisfying

$$\ell = C_2\theta(\alpha)\Big(1+\log\big(\frac{1}{\alpha}\big)\Big)\Big(d\log(\theta(\alpha))+\log\Big(\frac{\log(1/\alpha)}{\gamma/2}\Big)\Big)$$

$$K = \frac{6\sqrt{\log(2n)}(\sqrt{\ell\log(1/\delta)}+\sqrt{\ell\log(1/\delta)+\epsilon\ell})}{\epsilon}$$

obeys that

$$\varepsilon(\hat{h}^S) - \varepsilon(h^*) \leq \alpha.$$

Specifically, when we choose

$$\alpha = C_1 \max\left\{\eta^{\frac{2}{2-\tau}}\left(\frac{dK\log(n/d)+\log(2K/\gamma)}{n}\right)^{\frac{\tau}{2-\tau}}, \frac{d\log((m+n)/d)+\log(2/\gamma)}{m}\right\},$$

and also $\epsilon \leq \log(1/\delta)$, then it follows that

$$\varepsilon(\hat{h}^S) - \varepsilon(h^*) = \tilde{O}\Big(\max\Big\{\big(\frac{d^{1.5}\sqrt{\theta(\alpha)\log(1/\delta)}}{n\epsilon}\big)^{\frac{\tau}{2-\tau}}, \frac{d}{m}\Big\}\Big),$$

where $\tilde{O}$ hides logarithmic factors in $m, n, 1/\gamma$.

Remark 16. *The bound above resembles the learning bound we obtain using the passive student queries with Algorithm 2 as the privacy procedure, except for the additional dependence on the disagreement coefficients. Interestingly, active learning achieves this bound without using the sophisticated (and often not practical) algorithmic components from DP, such as sparse sector technique to save privacy losses. Instead, we can get away with using simple Gaussian mechanism as in Algorithm 1.*

Remark 17 (Blackbox reduction, revisited)**.** *In contrary to our discussion in Remark 7, notice that we are using Algorithm 1 instead of Algorithm 2 as the labeling services, which allows us to reduce to any learner as a blackbox. This makes it possible to state formally results even for deep neural networks or other family of methods where obtaining ERM is hard but learning is conjectured to be easy in theory and in practice.*

3.4 Further Discussion

One interesting observation from our analysis is that using passive learning with a more advanced private query-release technique (Algorithm 2) ends up having qualitatively the same learning bound when using active learning with a simple private query-release mechanism (Algorithm 1). Both approaches are doing selection. Active learning selects those queries that are near the decision boundary to be informative for learning; the sparse-vector-technique approach *essentially* selects those queries that are not stable to spend privacy budget on.

One question is that are these data points that are being selected substantially overlapping? If not, then we might be able to combine the two and achieve even better private-utility tradeoff.

4 CONCLUSIONS

Existing theoretical analysis shows that PATE framework consistently learns any VC-classes in the realizable setting, but not in the more general cases. We show that PATE learns any VC-classes under Tsybakov noise condition with fast rates. When specializing to the realizable case, our results improve the best known sample complexity bound for both the public and private data. We show that PATE is incompatible with the agnostic learning setting because it is essentially trying to learn a different class of voting classifiers which could be better, worse, or comparable to the best classifier in the base-class. Lastly, we investigated the PATE framework with active learning for further saving of the privacy budget. Future work includes understanding the geometry of active learning further and to conduct an empirical study on these algorithms.

Acknowledgments

The research is partially supported by a start-up grant from the UCSB Computer Science Department and generous gifts from NEC Labs and Google.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Conference on Computer and Communications Security (CCS-16)*, pages 308–318, 2016.

Noga Alon, Raef Bassily, and Shay Moran. Limits of private learning with access to public data. In *Neural Information Processing Systems (NeurIPS-19)*, pages 10342–10352, 2019.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Symposium on Foundations of Computer Science (FOCS-14)*, pages 464–473, 2014.

Raef Bassily, Om Thakkar, and Abhradeep Thakurta. Model-agnostic private learning via stability. *arXiv preprint arXiv:1803.05101*, 2018a.

Raef Bassily, Om Thakkar, and Abhradeep Guha Thakurta. Model-agnostic private learning. In *Neural Information Processing Systems (NeurIPS-18)*, pages 7102–7112, 2018b.

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

Amos Beimel, Kobbi Nissim, and Uri Stemmer. Private learning and sanitization: Pure vs. approximate differential privacy. *Theory of Computing*, 12(890): 1–61, 2016.

Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9:323–375, 2005.

Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. Introduction to statistical learning theory. *Advanced Lectures on Machine Learning: ML Summer Schools*, pages 169–207, 2004.

Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference (TCC-16)*, pages 635–658, 2016.

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

Kamalika Chaudhuri and Daniel Hsu. Sample complexity bounds for differentially private learning. In *Annual Conference on Learning Theory (COLT-11)*, pages 155–186, 2011.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12 (3):1069–1109, 2011.

Yuval Dagan and Vitaly Feldman. Pac learning with stable and private predictions. In *Annual Conference on Learning Theory (COLT-20)*, pages 1389–1410, 2020.

Cynthia Dwork and Vitaly Feldman. Privacy-preserving prediction. In *Annual Conference on Learning Theory (COLT-18)*, pages 1693–1702, 2018.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference (TCC-06)*, pages 265–284, 2006.

Cynthia Dwork, Guy N Rothblum, and Salil Vadhan. Boosting and differential privacy. In *Symposium on Foundations of Computer Science (FOCS-10)*, pages 51–60, 2010.

Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Conference on Computer and Communications Security (CCS-14)*, pages 1054–1067, 2014.

Steve Hanneke. Theory of disagreement-based active learning. *Foundations and Trends in Machine Learning*, 7(2-3):131–309, 2014.

Moritz Hardt and Guy N Rothblum. A multiplicative weights mechanism for privacy-preserving data analysis. In *Symposium on Foundations of Computer Science (FOCS-20)*, pages 61–70, 2010.

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

Ashwin Machanavajjhala, Daniel Kifer, John Abowd, Johannes Gehrke, and Lars Vilhuber. Privacy: Theory meets practice on the map. In *International Conference on Data Engineering (ICDE-08)*, pages 277–286, 2008.

Enno Mammen and Alexandre B Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27 (6):1808–1829, 1999.

H Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. In *International Conference on Learning Representations (ICLR-18)*, 2018.

Anupama Nandi and Raef Bassily. Privately answering classification queries in the agnostic pac model. In *International Conference on Algorithmic Learning Theory (ALT-20)*, pages 687–703, 2020.

Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In *Symposium on Theory of Computing (STOC-07)*, pages 75–84, 2007.

Nicolas Papernot, Martín Abadi, Úlfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. In *International Conference on Learning Representations (ICLR-17)*, 2017.

Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. Scalable private learning with pate. In *International Conference on Learning Representations (ICLR-18)*, 2018.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11(90):2635–2670, 2010.

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

Alexander B Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

Vladimir N Vapnik. *The nature of statistical learning theory*. Springer, 1995.

Yu-Xiang Wang, Stephen Fienberg, and Alex Smola. Privacy for free: Posterior sampling and stochastic gradient monte carlo. In *International Conference on Machine Learning (ICML-15)*, pages 2493–2502, 2015.

Yu-Xiang Wang, Jing Lei, and Stephen E. Fienberg. Learning with differential privacy: Stability, learnability and the sufficiency and necessity of erm principle. *Journal of Machine Learning Research*, 17(183): 1–40, 2016.

Chicheng Zhang and Kamalika Chaudhuri. Beyond disagreement-based agnostic active learning. In *Neural Information Processing Systems (NeurIPS-14)*, pages 442–450, 2014.

Zhengli Zhao, Nicolas Papernot, Sameer Singh, Neoklis Polyzotis, and Augustus Odena. Improving differentially private models with active learning. *arXiv preprint arXiv:1910.01177*, 2019.