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

Chong Liu [1] **Yuqing Zhu** [1] **Kamalika Chaudhuri** [2] **Yu-Xiang Wang** [1]

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 the more general cases where the error rate of the optimal classifier is bounded away from zero. We fill in this gap by introducing Tsybakov Noise condition 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 proof 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

Increasing public concerns on data privacy have accelerated the shaping of public policies regarding the use of personal data for machine learning. For example, the General Data Protection Regulation (GDPR) (European Parliament & Council of the European Union, 2016) that went into effect in 2018 requires all organizations operating in the European Union to provide their customers transparency, access, and *the right to erase* all traces of their data. GDPR and similar new legislations in the US (e.g., the California Consumer Privacy Act (Mathews & Bowman, 2018)) have placed businesses that rely on machine learning models to serve their customers in an awkward position, as there might not be a way to "unlearn" an already-deployed model even if a user requested their information to be erased.

There is a growing body of research on differentially private machine learning (see, e.g., Kasiviswanathan et al., 2011; Chaudhuri et al., 2011; Bassily et al., 2014; Wang et al., 2015; Abadi et al., 2016), which aims at providing formal privacy guarantees that provably nullifies 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. Despite some promising progress, it remains a fundamental challenge how to circumvent the pessimistic dimension dependence (Hardt & Talwar, 2010; Bassily et al., 2014) and develop practical methods in privately releasing deep learning models.

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 splitting the private dataset into K parts.
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 is appealing in practice due to its modular reduction to non-private learning, which allows it to work with any deep learning models in a model-agnostic fashion. The competing alternative, NoisySGD (Abadi et al., 2016), requires significantly more tweaking and modifications to achieve a comparable performance (e.g., on MNIST), if achievable.

This paper builds upon the pioneering work of Bassily et al. (2018), which instantiates the PATE framework with a more data-adaptive scheme of private aggregation that 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. (2018) provides the first

[1]Department of Computer Science, University of California, Santa Barbara, Santa Barbara, California, USA [2]Department of Computer Science and Engineering, University of California, San Diego, La Jolla, California, USA. Correspondence to: Chong Liu <chongliu@cs.ucsb.edu>.

International Workshop on Federated Learning for User Privacy and Data Confidentiality in Conjunction with ICML 2020 (FL-ICML'20).

theoretical analysis of PATE which shows that it is able to PAC-learn any hypothesis classes with finite VC-dimension in the realizable setting (Bassily et al., 2018). 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. (2018) 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 particular, these error bounds become vacuous when the optimal classifier has an error rate $\geq 1/6$.

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: Tsybakov noise condition (Mammen & Tsybakov, 1999), active learning (Hanneke, 2014), as well as the properties of voting classifiers. Our contributions are summarized as follows.

1. We show that PATE consistently learns any VC-classes under the Tsybakov noise condition. 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 the Tsybakov noise condition, we show that the active learning approach allows us to save the privacy budget exponentially when we use the standard privacy aggregation (Algorithm 1).

To the best of our knowledge, these results are new and we are the first that consider noise models and active learning for PATE.

2. Preliminaries

In this section, we introduce the notations , definitions, and discuss specific technical tools that we will build upon.

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 $\varepsilon(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathbb{1}(h(x) \neq y)]$, where $\mathbb{1}(x)$ 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 $\hat{\varepsilon}(h) = \frac{1}{n}\sum_{i=1}^{n}[\mathbb{1}(h(x_i) \neq y_i)]$. 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 $\hat{h} = \operatorname{argmin}_{h\in\mathcal{H}} \hat{\varepsilon}(h)$. $\hat{h}^{\texttt{agg}}$ is used to denote the aggregated classifier in the PATE framework. $\hat{h}^{\texttt{priv}}$ 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 $\texttt{Dis}(h_1, h_2) = \mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}[\mathbb{1}(h_1(x) \neq h_2(x)]$. 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 $\widehat{\texttt{Dis}}(h_1, h_2) = \frac{1}{n}\sum_{i=1}^{n}[\mathbb{1}(h_1(x_i) \neq h_2(x_i)]$. In this paper, we use standard big O notations; and to improve the readability, we use $\lesssim$ and $\tilde{O}$ to hide poly-logarithmic factors.

Differential Privacy and Private Learning. Now we formally define differential privacy.

Definition 1 (Differential Privacy (Dwork & 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 $\mathcal{S} \subseteq \mathcal{R}$:*

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

Remark. 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 DP guarantee. The closer they are to 0, the stronger the guarantee is.

DP has many desirable properties including closure under postprocessing, adaptive composition and so on. For the interest of this paper, it suffices to know that this is a definition and there are various algorithms that can be used to achieve DP. The problem of differentially private learning aims at designing a randomized training algorithm that satisfies Definition 1. More often then not, the research question is about understanding the privacy-utility trade-offs and characterizing the Pareto optimal frontiers.

PATE and Model-Agnostic Private Learning. There are different ways we can instantiate the PATE framework

Algorithm 1 Standard PATE (Papernot et al., 2017)

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 $\hat{y}_j \leftarrow \mathbb{1}(\sum_{k=1}^K \hat{h}_k(x_j) + \mathcal{N}(0, \sigma^2) \geq K/2)$.
5: Nature chooses x_{j+1} adaptively (possibly as a function of the output vector till time j).
6: **end for**

in the introduction to privately aggregate the labels. We show two of them in Algorithm 1 and Algorithm 2. Here, $\mathcal{N}(\mu, \sigma^2)$ denotes the Gaussian distribution with parameters μ, σ, $\texttt{Lap}(\lambda)$ denotes the Laplace distribution with parameter λ, and the margin function is defined as $\text{margin}(x) := |2\sum_{k=1}^K \hat{h}_k(x) - K|$, which measures the absolute value of the difference between the number of votes.

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

Remark. The key difference between the two algorithms is that the standard PATE pays for a unit privacy loss for every public data point labeled, while the stability-based PATE (essentially) pays for only those with a small margin in the votes. This idea is formalized in the following lemma.

Lemma 3 (Adapted from Theorem 3.11 of (Bassily et al., 2018))**.** *If the classifiers $\hat{h}_1, ..., \hat{h}_K$ and the sequence $x_1, ..., x_\ell$ obey that there are at most T of them such that $\text{margin}(x_k) < K/3$ for $K = 136 \log(4\ell T/\min(\delta, \beta)) \cdot \sqrt{T \log(2/\delta)}/\epsilon$. Then with probability at least $1 - \beta$, Algorithm 2 finishes all ℓ queries and for all $i \in [\ell]$ such that $\text{margin}(x_i) \geq K/3$, the output of Algorithm 2 is $\hat{h}^{\texttt{agg}}(x_i)$.*

Lemma 4 (Lemma 4.2 of (Bassily et al., 2018))**.** *If the classifiers $\hat{h}_1, ..., \hat{h}_K$ obeys that each of them makes at most B mistakes on data $(x_1, y_1), ..., (x_\ell, y_\ell)$, then*

$$\left|\{i \in [\ell] \mid \sum_{k=1}^K \mathbb{1}(\hat{h}_k(x_i) \neq y_i) \geq K/3\}\right| \leq 3B.$$

Remark. Lemma 4 implies that if the individual classifiers are accurate — by the statistical learning theory, they are — the corresponding majority voting classifier is not only nearly as accurate, but also has sufficiently large margin that satisfies the conditions in Lemma 3.

Our results provide new theoretical insight into Algorithm 2 and the active learning method we propose shows that we can essentially obtain better bounds.

Algorithm 2 Stability-Based PATE (Bassily et al., 2018)

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 $\epsilon, \delta > 0$.

1: Nature chooses x_1.
2: $\lambda \leftarrow \frac{1}{\epsilon}\left(\sqrt{2T(\epsilon + \log(2/\delta))} + \sqrt{2T \log(2/\delta)}\right)$.
3: $w \leftarrow 3\lambda \log(2(\ell + T)/\delta)$, $\hat{w} \leftarrow w + \texttt{Lap}(\lambda)$.
4: $c = 0$.
5: **for** $j \in [\ell]$ **do**
6: $\text{dist}_j \leftarrow \max\{0, \lceil \text{margin}(x_j)/2 \rceil - 1\}$.
7: $\widehat{\text{dist}}_j \leftarrow \text{dist}_j + \texttt{Lap}(2\lambda)$.
8: **if** $\widehat{\text{dist}}_j > \hat{w}$ **then**
9: Output $\hat{y}_j \leftarrow \mathbb{1}(\sum_{k=1}^K \hat{h}_k(x_j) \geq K/2)$.
10: **else**
11: Output $\hat{y}_j \leftarrow \perp$.
12: $c \leftarrow c + 1$, break if $c \geq T$.
13: $\hat{w} \leftarrow w + \texttt{Lap}(\lambda)$.
14: **end if**
15: Nature chooses x_{j+1} adaptively (based on $\hat{y}_1, ..., \hat{y}_{j-1}$).
16: **end for**

3. Main Results

In this section, first we revisit the PATE framework and point out the gap in existing utility guarantee results. To seal the gap, then we prove improved learning bounds under the Tsybakov noise condtion and in the agnostic setting, respectively. Further, we propose the PATE with disagreement-based active learning method and show its theoretical guarantees are even better.

3.1. PATE Framework Revisited

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.

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: 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 $(\hat{h}_1, ..., \hat{h}_K), D^S, \epsilon, \delta$)
4: For those pseudo labels that are $\perp$, assign them arbitrarily to $\{0, 1\}$.

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

Algorithm 3 shows the PATE with Passive Student Queries (PATE-PSQ) algorithm. For its utility guarantee, we have the following theorem based on Lemma 3 and 4.

Theorem 5 (Utility guarantee of Algorithm 3)**.** *Set* $T = 3(\mathbb{E}[\varepsilon(\hat{h}_1)]m + \sqrt{m\log(m/\beta)/2})$, $K = O(\log(mT/\min(\delta,\beta))\sqrt{T\log(2/\delta)}/\epsilon)$. *Let* $\hat{h}$ *be the output of Algorithm 3 that uses Algorithm 2 for privacy aggregation. With probability at least* $1-\beta$ *(over the randomness of the algorithm and the randomness of all data points drawn iid), we have* $\varepsilon(\hat{h}) \le \tilde{O}\big(\frac{d^2 m\log(2/\delta)}{n^2\epsilon^2} + \sqrt{\frac{d}{m}}\big)$ *for the realizable case, and* $\varepsilon(\hat{h}) \le 3\varepsilon(h^*) + \tilde{O}\big(\frac{m^{1/3}d^{2/3}}{n^{2/3}\epsilon^{2/3}} + \sqrt{\frac{d}{n\epsilon}} + \sqrt{\frac{d}{m}}\big)$ *for the agnostic case.*

Remark (Error bounds when m is sufficiently large)**.** Notice that we do not have to label all public data. So when we have a large number of public data, we can afford to choose m to be smaller so as to minimize the bound. That gives is a $\tilde{O}(\frac{d}{n^{2/3}\epsilon^{2/3}})$ error bound for the realizable case and a $\tilde{O}(\frac{d^{3/5}}{n^{2/5}\epsilon^{2/5}})$ error bound for the agnostic case[1].

We emphasize that for the agnostic setting the bound is vacuous if $\varepsilon(h^*) > 1/6$ and $\hat{h}$ does not match the performance of h^* even as $m, n \to \infty$. This does not match the empirical performance of Algorithm 3 reported in (Papernot et al., 2017; 2018).

3.2. Improved Learning Bounds under TNC

To seal the gap, we consider proving learning bounds under the Tsybakov Noise Condition, which is an important assumption about the data distribution and hypothesis class.

Definition 6 (Tsybakov Noise Condition (Mammen & Tsybakov, 1999; Tsybakov, 2004))**.** *For some* $\eta \in [1,\infty)$ *and* $\tau \in [0,1]$, *for every* $h \in \mathcal{H}$,

$$\texttt{Dis}(h, h^*) \le \eta(\varepsilon(h) - \varepsilon(h^*))^\tau.$$

Remark. In the realizable setting, one can obtain the excess risk in order of $O(1/n)$, where n is the number of input samples. However, realizable setting is naturally unrealistic in many real-world problems. While in agnostic setting one can only obtain $O(1/\sqrt{n})$. By contrast, the TNC provides a way where we assume noise can be controlled and we can do better, i.e., $O(1/n^{1/(2-\tau)})$. Note that $\tau \in [0,1]$. In addition, when $\tau = 0$, TNC is void. When $\tau = 1$, it reduces to Massart noise condition.

Next, we give a novel analysis of Algorithm 3 under the Tsybakov noise condition, which is based on the following lemmas. 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.

Lemma 7. *With probability* $1-\gamma$ *over the training data of* $\hat{h}_1, ..., \hat{h}_K$, *assume Tsybakov noise condition with parameters* η, τ, *then for all* k,

$$\texttt{Dis}(\hat{h}_k, h^*) \le \eta^{\frac{2}{2-\tau}}\Big(\frac{dK\log(n/d) + \log(K/\gamma)}{n}\Big)^{\frac{\tau}{2-\tau}}.$$

Lemma 8. *Assuming Tsybakov noise condition with parameters* η, τ, *the total number of mistakes made by one teacher classifier* $\hat{h}_k$ *with respect to* h^* *can be bounded as:*

$$\sum_{j=1}^{m} \mathbb{1}(\hat{h}_k(x_j) \ne h^*(x_j)) \le 2\max\{m\texttt{Dis}(\hat{h}_k, h^*), \log(K/\gamma)\}.$$

Lemma 9. *Assume Tsybakov noise condition with parameters* ϵ, δ, *the input value of* T *of algorithm 3 is chosen as*

$$T = \tilde{O}\Big(\frac{\eta^{\frac{4}{4-3\tau}} m^{\frac{4-2\tau}{4-3\tau}} d^{\frac{2\tau}{4-3\tau}}}{n^{\frac{2\tau}{4-3\tau}} \epsilon^{\frac{2\tau}{4-3\tau}}} \log^{\frac{2\tau}{4-3\tau}}\Big(\frac{\eta^{\frac{4}{4-3\tau}} m^{\frac{8-5\tau}{4-3\tau}} d^{\frac{2\tau}{4-3\tau}}}{n^{\frac{2\tau}{4-3\tau}} \epsilon^{\frac{2\tau}{4-3\tau}}}\Big)\Big).$$

Theorem 10 (Utility guarantee of Algorithm 3 under the Tsybakov noise condition)**.** *Assume the data distribution* $\mathcal{D}$ *and the hypothesis class* $\mathcal{H}$ *obey the Tsybakov Noise condition with parameters* η, τ. *Then Algorithm 3 with* $T = \tilde{O}\Big(\frac{\eta^{\frac{4}{4-3\tau}} m^{\frac{4-2\tau}{4-3\tau}} d^{\frac{2\tau}{4-3\tau}}}{n^{\frac{2\tau}{4-3\tau}} \epsilon^{\frac{2\tau}{4-3\tau}}} \log^{\frac{2\tau}{4-3\tau}}\Big(\frac{\eta^{\frac{4}{4-3\tau}} m^{\frac{8-5\tau}{4-3\tau}} d^{\frac{2\tau}{4-3\tau}}}{n^{\frac{2\tau}{4-3\tau}} \epsilon^{\frac{2\tau}{4-3\tau}}}\Big)\Big)$ *and* $K = O(\log(mT/\min(\delta,\beta))\sqrt{T\log(1/\delta)}/\epsilon)$ *obeys that with probability at least* $1-\beta$*:*

$$\varepsilon(\hat{h}) \le \varepsilon(h^*) + \tilde{O}\Big(\frac{d}{m} + \frac{\eta^{\frac{4}{4-3\tau}} m^{\frac{\tau}{4-3\tau}} d^{\frac{2\tau}{4-3\tau}}}{n^{\frac{2\tau}{4-3\tau}} \epsilon^{\frac{2\tau}{4-3\tau}}} \cdot \log^{\frac{2\tau}{4-3\tau}}\Big(\frac{\eta^{\frac{4}{4-3\tau}} m^{\frac{8-5\tau}{4-3\tau}} d^{\frac{2\tau}{4-3\tau}}}{n^{\frac{2\tau}{4-3\tau}} \epsilon^{\frac{2\tau}{4-3\tau}}}\Big)\Big).$$

Remark (Bounded noise case)**.** When $\tau = 1$ and $\eta = O(1)$, TNC 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 < 0.5. Theorem 10 implies that the excess risk is bounded by $\tilde{O}(\frac{d^2 m}{n^2\epsilon^2} + \frac{d}{m})$, 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., 2018) in the stronger realizable setting from $\tilde{O}(d^{3/2}\alpha^{-3/2}\epsilon^{-1})$ and $\tilde{O}(d\alpha^{-2})$ to $\tilde{O}(d^{3/2}\alpha^{-1}\epsilon^{-1})$ and $\tilde{O}(d\alpha^{-1})$ 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

[1] These correspond to the $\tilde{O}((d/\alpha)^{3/2})$ sample complexity bound in Theorem 4.6 of (Bassily et al., 2018) for realizable PAC learning for error α; and the $\tilde{O}(d^{3/2}/\alpha^{5/2})$ sample complexity bound in Theorem 4.7 of (Bassily et al., 2018) for agnostic PAC learning with error $O(\alpha + \varepsilon(h^*))$. The privacy parameter ϵ is taken as a constant in these results.

labels y as in Lemma 4; all they have to do is to agree on something for the majority of the data points. Conveniently, TNC implies that the teacher classifiers agree on the optimal classifier h^*. Secondly, when the teachers agree on the optimal classifiers, the synthetic learning problem with the privately released pseudo-labels is nearly realizable.

3.3. Improved Learning Bounds in Agnostic Setting

In this subsection, 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 11 (Voting fail)**.** Consider a uniform distribution on $\mathcal{X} = \{x_1, x_2, x_3, x_4\}$ and that the corresponding label $\mathbb{P}(y = 1) = 1$. 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 $K \to \infty$.

	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
$\hat{h}^{\mathtt{agg}}$	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 12 (Voting win)**.** If $\mathbb{P}[\hat{h}(x) \neq y|x] \leq 0.5 - \Delta$ for all $x \in \mathcal{X}$, then by Hoeffding's inequality $\mathbb{P}[\hat{h}^{\mathtt{agg}}(x) \neq y|x] = \mathbb{P}[\sum_{k=1}^{K} \mathbb{1}(\hat{h}_k(x) \neq y) \geq k/2|x] \leq e^{-2K\Delta^2}$. 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 paradigms. First, we need to go beyond $\mathcal{H}$ and compare with the following classifier

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

Note that this classifier also changes as n gets larger. This classifier can be better or worse than $\hat{h}_1$ that takes n/K data points, $\hat{h}$ that trains on all n data points and h^* that is the optimal classifier in $\mathcal{H}$.

Second, we define the expected margin $\Delta_n(x) := |\mathbb{E}[\hat{h}_1(x)|x] - 0.5|$ to capture for each $x \in \mathcal{X}$, 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 $\Delta(x)$ 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 12, we do not require the teachers to agree on y. Instead, it measures the extent to which they agree on $h^{\mathtt{agg}}_{(\infty)}(x)$, which could be any label.

On the technical level, this definition allows us decouple the stability analysis of PATE and the accuracy analysis that talks about how good $h^{\mathtt{agg}}_{(\infty)}(x)$ is.

Theorem 13. *Suppose the voting classifiers agree w.r.t. $h^{\mathtt{agg}}_{(\infty)}$ on at least $m - T$ examples, and $K \geq \tilde{O}(\sqrt{T}/\epsilon)$, then the output classifier $\hat{h}^S$ of Algorithm 3 in the agnostic setting satisfies,*

$$\varepsilon(\hat{h}^S) - \varepsilon(h^{\mathtt{agg}}_{(\infty)}) \leq \min_{h \in \mathcal{H}} \mathtt{Dis}(h, h^{\mathtt{agg}}_{(\infty)}) + 2T/m + 4\sqrt{d/m}.$$

The voting classifier $\hat{h}^{\mathtt{agg}}$ 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 $\mathrm{Vote}(\mathcal{H})$ where the learning problem becomes realizable.

Theorem 14. *Choose T,K according to Theorem 13. Suppose we train an ensemble classifier within the voting hypothesis space $\mathrm{Vote}_K(\mathcal{H})$ in the student domain, then the output classifier $\hat{h}^S$ of Algorithm 3 in the agnostic setting satisfies,*

$$\varepsilon(\hat{h}^S) - \varepsilon(h^{\mathtt{agg}}_{(\infty)}) \leq 4T/m + 5(Kd + \log(4/\gamma))/m.$$

Therefore, the rest of the problem would be finding the bound of T.

Lemma 15. *From Hoeffding's inequality, for a fixed data*

point x, we learn that w.p. $1-\gamma$,

$$\Big|\frac{1}{K}\sum_{k=1}^{K}\hat{h}_k(x) - \mathbb{E}[\hat{h}_1(x)|x]\Big| < \sqrt{\frac{\log(m/\gamma)}{K}}.$$

From this lemma, we learn that $\forall j \in [m]$, if $\Delta(x_j) \geq \sqrt{\log(m/\gamma)/K}$, then $\hat{h}^{\texttt{agg}} = h^{\texttt{agg}}_{(\infty)}$, and if $\Delta(x_j) \geq 1/6 + \sqrt{\log(m/\gamma)/K}$, then $\hat{h}^{\texttt{agg}} = h^{\texttt{agg}}_{(\infty)} = \hat{h}^{\texttt{priv}}$. It leads to our choice of T, which is

$$\begin{aligned} T &= m\mathbb{E}[\mathbb{1}(\Delta_{n/K}(x) < 1/6 + \sqrt{\log(m/\gamma)/K})] \\ &\quad + \sqrt{m\log(1/\gamma)} \\ &= m\mathbb{P}[\Delta_{n/K}(x) < 1/6 + \sqrt{\log(m/\gamma)/K}] \\ &\quad + \sqrt{m\log(1/\gamma)}. \end{aligned}$$

Remark. Whether the bounds in Theorem 13 and 14 will vanish as $m, n \to \infty$ depends strongly on whether $\mathbb{P}[\Delta_{n/K}(x) < 1/6 + \sqrt{\log(m/\gamma)/K}]$ vanishes as $n \to \infty$. This is a reasonable condition that says the "teachers" will get *more confident* in their individual prediction for all data points as $n \to \infty$. We argue this is a much more modest requirement than requiring the teachers to get *more accurate*.

3.4. PATE with Active Student Queries under TNC

In previous subsections, we've proved stronger learning bounds for PATE framework under the Tsybakov noise condition and in agnostic setting. However, all these results are with passive student queries. Can we do even better? In Algorithm 4, we propose a new algorithm called PATE with Active Student Queries (PATE-ASQ).

Theorem 16 (Utility guarantee of PATE-ASQ)**.** *With probability at least $1-\gamma$, there exists universal constants C_1, C_2 such that for all α such that*

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

the output $\hat{h}_S$ of Algorithm 4 with parameters 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

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

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 $D^T_k \subseteq D^T$ where $k \in [K]$
2: Train K classifiers $\hat{h}_k \in \mathcal{H}$, one from each part D^T_k.
3: Declare "Labeling Service" $\leftarrow$ Algorithm 1 with $\hat{h}_1, ..., \hat{h}_K, \ell, \epsilon, \delta$, with an unspecified "nature".
4: Initiate an active learning oracle (e.g., disagreement-based active learning algorithm (Hanneke, 2014)) 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 $\hat{h}^S$.

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

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

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

Remark. The bound above resembles the same bound we obtain using the positive 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.

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.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the ACM Conference on Computer and Communications Security*, pp. 308–318, 2016.

Bassily, R., Smith, A., and Thakurta, A. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Proceedings of the 54th Annual IEEE Symposium on Foundations of Computer Science*, pp. 464–473, 2014.

Bassily, R., Thakkar, O., and Thakurta, A. Model-agnostic private learning via stability. *arXiv preprint arXiv:1803.05101*, 2018.

Bassily, R., Moran, S., and Alon, N. Limits of private learning with access to public data. In *Advances in Neural Information Processing Systems 32*, 2019.

Beimel, A., Nissim, K., and Stemmer, U. Characterizing the sample complexity of private learners. In *Proceedings of the 4th Innovations in Theoretical Computer Science Conference*, pp. 97–110, 2013.

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

Bousquet, O., Boucheron, S., and Lugosi, G. Introduction to statistical learning theory. *Advanced Lectures on Machine Learning: ML Summer Schools*, pp. 169–207, 2004.

Bun, M. and Steinke, T. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Proceedings of the Theory of Cryptography Conference*, pp. 635–658, 2016.

Bun, M., Nissim, K., Stemmer, U., and Vadhan, S. Differentially private release and learning of threshold functions. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pp. 634–649. IEEE, 2015.

Chaudhuri, K. and Hsu, D. Sample complexity bounds for differentially private learning. In *Proceedings of the 24th Annual Conference on Learning Theory*, pp. 155–186, 2011.

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

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

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Theory of cryptography conference*, pp. 265–284, 2006.

Dwork, C., Rothblum, G. N., and Vadhan, S. Boosting and differential privacy. In *Proceedings of the 51st Annual IEEE Symposium on Foundations of Computer Science*, pp. 51–60, 2010.

European Parliament and Council of the European Union. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation). *Official Journal of the European Union*, 2016.

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

Hardt, M. and Talwar, K. On the geometry of differential privacy. In *Proceedings of the 42nd ACM Symposium on Theory of Computing*, pp. 705–714. ACM, 2010.

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

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

Mathews, K. and Bowman, C. The california consumer privacy act of 2018, 2018.

Papernot, N., Abadi, M., Erlingsson, Ú., Goodfellow, I., and Talwar, K. Semi-supervised knowledge transfer for deep learning from private training data. In *Proceedings of the 5th International Conference on Learning Representations*, 2017.

Papernot, N., Song, S., Mironov, I., Raghunathan, A., Talwar, K., and Erlingsson, Ú. Scalable private learning with pate. In *Proceedings of the 6th International Conference on Learning Representations*, 2018.

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

Thakurta, A. G. and Smith, A. Differentially private feature selection via stability arguments, and the robustness of the lasso. In *Proceedings of the 26th Conference on Learning Theory*, pp. 819–850, 2013.

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

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

Wang, Y.-X., Fienberg, S., and Smola, A. Privacy for free: Posterior sampling and stochastic gradient monte carlo. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 2493–2502, 2015.

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

Zhang, C. and Chaudhuri, K. Beyond disagreement-based agnostic active learning. In *Advances in Neural Information Processing Systems 27*, pp. 442–450, 2014.

Zhao, Z., Papernot, N., Singh, S., Polyzotis, N., and Odena, A. Improving differentially private models with active learning. *arXiv preprint arXiv:1910.01177*, 2019.