Journal of Machine Learning Research 22 (2021) 1-44 Submitted 11/20; Revised 7/21; Published 11/21

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

Chong Liu† chongliu@cs.ucsb.edu
Yuqing Zhu† yuqingzhu@ucsb.edu
Kamalika Chaudhuri⋆ kamalika@cs.ucsd.edu
Yu-Xiang Wang† yuxiangw@cs.ucsb.edu

† *Department of Computer Science, University of California, Santa Barbara, CA 93106, USA*
⋆ *Department of Computer Science and Engineering, University of California, San Diego, La Jolla, CA 92093, USA*

Editor: Andreas Krause

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, and empirical studies show the effectiveness of this new idea. 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.

Keywords: Model-agnostic private learning, Private Aggregation of Teacher Ensembles, Differential privacy, Tsybakov noise condition, Active learning

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, 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).

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; Shokri and Shmatikov, 2015).

In a nutshell, differentially private machine learning aims at providing formal privacy guarantees that provably reduce 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 satisfying DP guarantees work well in low-dimensional regime where 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 injected 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 *teacher* classifiers, 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 number of DP mechanisms that can be used to instantiate the PATE Framework. Laplace mechanism and Gaussian mechanism are used in Papernot et al. (2017, 2018) respectively. This paper primarily considers the new mechanism 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 PATE 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, i.e, expected risk of best hypothesis equals 0. And in this case, the center of teacher agreement is true label. However, 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, a.k.a., inconsistency.

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

Algorithm	**PATE** (Gaussian Mech.) Papernot et al. (2017)	**PATE** (SVT-based) Bassily et al. (2018b)	**PATE** (SVT-based) This paper	**PATE** (Active Learning) This paper
Realizable	$\tilde{O}\left(\frac{d}{(n\epsilon)^{2/3}} \vee \frac{d}{m}\right)$	$\tilde{O}\left(\frac{d}{(n\epsilon)^{2/3}} \vee \sqrt{\frac{d}{m}}\right)$	$\tilde{O}\left(\frac{d^{3/2}}{n\epsilon} \vee \frac{d}{m}\right)$	$\tilde{O}\left(\frac{d^{3/2}\theta^{1/2}}{n\epsilon} \vee \frac{d}{m}\right)$
τ-TNC	$\tilde{O}\left(\left(\frac{d^{3/2}}{n\epsilon}\right)^{\frac{2\tau}{4-\tau}} \vee \frac{d}{m}\right)$	same as agnostic	$\tilde{O}\left(\left(\frac{d^{3/2}}{n\epsilon}\right)^{\frac{\tau}{2-\tau}} \vee \frac{d}{m}\right)$	$\tilde{O}\left(\left(\frac{d^{3/2}\theta^{1/2}}{n\epsilon}\right)^{\frac{\tau}{2-\tau}} \vee \frac{d}{m}\right)$
Agnostic (vs h^*)	$\Omega(\texttt{Err}(h^*))$ required.	$13\texttt{Err}(h^*)+$ $\tilde{O}\left(\frac{d^{3/5}}{n^{2/5}\epsilon^{2/5}} \vee \sqrt{\frac{d}{m}}\right)$	$\Omega(\texttt{Err}(h^*))$ required.	$\Omega(\texttt{Err}(h^*))$ required.
Agnostic (vs $h_\infty^{\texttt{agg}}$)	-	-	Consistent under weaker conditions.	-

- Results new to this paper are highlighted in blue.
- Teacher number 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. δ is assumed to be in its typical range $\delta < 1/\text{poly}(n)$ and $\epsilon < \log(1/\delta)$. The TNC parameter τ ranges between $(0, 1]$. See Table 2 for a checklist of notations.
- Proofs of utility guarantees of PATE (Gaussian mechanism) can be found in Appendix A.

To fill in the gap, in this paper, we revisit the problem of model-agnostic private learning in PATE framework in two non-realizable settings: under the Tsybakov Noise Condition (TNC) (Mammen and Tsybakov, 1999; Tsybakov, 2004) and in agnostic setting. By making TNC assumption, teachers stay close to the best hypothesis h^* in hypothesis class, thus we consider h^* as the new center for teachers to agree on, instead of considering true label in the realizable setting. We make no assumptions in agnostic setting, and a different center of teacher gravity is considered. In addition, we introduce active learning (Hanneke, 2014) to PATE and propose a new practical algorithm.

Summary of results. Our contributions are summarized as follows.

1. We show that PATE consistently learns any VC-classes under TNC with fast rates and requires very few unlabeled public data points. When specializing to the realizable case, we show that the sample complexity bound of the SVT-based PATE is $\tilde{O}(d^{3/2}/\alpha\epsilon)$ and $\tilde{O}(d/\alpha)$ for the private and public datasets respectively. The best known results (Bassily et al., 2018b) is $\tilde{O}(d^{3/2}/\alpha^{3/2}\epsilon)$ (for private data) and $\tilde{O}(d/\alpha^2)$ (for public data).

2. We analyze standard Gaussian mechanism-based PATE (Papernot et al., 2018) under TNC. In the realizable case, we obtained a sample complexity of $\tilde{O}(d^{3/2}/\alpha\epsilon)$ and $\tilde{O}(d/\alpha)$ for the private and public datasets respectively, which matches the bound of (Bassily et al., 2018b) with a simpler and more practical algorithm that uses fewer public data points.

3. 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.

4. We propose a new active learning-based algorithm, PATE with Active Student Queries (PATE-ASQ), to adaptively select which public data points to release. Under TNC, we show that active learning with standard Gaussian mechanism is able to match the same learning bounds of the SVT-based method for privacy aggregation (Algorithm 1), except some additional dependence.

5. Finally, our experiments on real-life datasets demonstrate that PATE-ASQ achieves significantly better accuracy than standard PATE algorithms while incurring the same or lower privacy loss.

These results (summarized in Table 1) provide strong theoretical insight into how PATE works. Interestingly, our theory suggests that *Gaussian mechanism suffice*s especially if we use active learning and that it is better *not* to label all public data when the number of public data points m is large. The remaining data points can be used for semi-supervised learning. These tricks have been proposed in *empirical* studies of PATE (see, e.g., semi-supervised learning (Papernot et al., 2017, 2018), active learning (Zhao et al., 2019)), thus our *theory* can be viewed as providing formal justifications to these PATE variants that are producing strong empirical results in *deep learning with differential privacy.*

Motivation and applicability. We conclude the introduction by commenting on the applicability of the knowledge transfer model of differentially private learning and PATE. First, while this model applies only to those cases when a (small) public unlabeled dataset is available, it gains a more favorable privacy-utility tradeoff on those applicable cases. Second, public datasets are often readily available (e.g., census microdata) or can be acquired at a low cost (e.g., incentivizing patients to opt-in) especially if we do not need labels (e.g., getting doctor's diagnosis is expensive). Note that this setting is different from label differential privacy (Chaudhuri and Hsu, 2011) where only labels are considered private. In our problem, even if the public data points are labeled, they are scarce and learning directly from them without using the private data will not give the same learning bound. In addition, PATE uses standard off-the-shelf learners / optimizers as blackboxes, thereby retaining their computational efficiency. For these reasons, we argue that the "knowledge transfer" model is widely applicable and could enable practical algorithms with formal DP guarantees in the many applications where the standard private learning model fails to be sufficiently efficient, private and accurate at the same time.

2. Related Work

The literature on differentially private machine learning is enormous and it is impossible for us to provide an exhaustive discussion. Instead we focus on a few closely related work and only briefly discuss other representative results in the broader theory of private learning.

2.1 Private Learning with an Auxiliary Public Dataset

The use of an auxiliary unlabeled public dataset was pioneered in empirical studies (Papernot et al., 2017, 2018) where PATE was proposed and shown to produce stronger results than NoisySGD in many regimes. Our work builds upon Bassily et al. (2018b)'s first analysis of PATE and substantially improves the theoretical underpinning. To the best of our knowledge, our results are new and we are the first that consider *noise models* and *active learning* for PATE.

Alon et al. (2019) also studied the problem of private learning with access to an additional public dataset. Specifically, their result reveals an interesting "theorem of the alternatives"-type result that says either a VC-class is learnable without an auxiliary public dataset, or we need at least $m = \Omega(d/\alpha)$ public data points, which essentially says that our sample complexity on the (unlabeled) public data points are optimal. They also provide an upper bound that says $\tilde{O}(d/\alpha^2)$ private data and $\tilde{O}(d/\alpha)$ public data are sufficient (assuming constant privacy parameter ϵ) to *agnostically learn* any classes with VC-dimension d to α-excess risk. Their algorithm however uses 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). Moreover, we show that under TNC, the inefficient construction is not needed and PATE is indeed consistent and enjoys faster rates. It remains an open problem how to achieve consistent private agnostic learning with only access to ERM oracles.

2.2 Privacy-Preserving Prediction

There is another line of work (Dwork and Feldman, 2018) that focuses on the related problem of "privacy-preserving prediction" which does not release the learned model (which we do), but instead privately answer one randomly drawn query x (which we need to answer many, so as to train a model that can be released). While their technique can be used to obtain bounds in our setting, it often involves weaker parameters. More recent works under this model (see e.g., Dagan and Feldman, 2020; Nandi and Bassily, 2020) notably achieve consistent agnostic learning in this setting with rates comparable to that of Alon et al. (2019). However, they rely on the same explicit α-net construction (Beimel et al., 2016), which renders their algorithm computationally inefficient in practice. In contrast, we analyze an oracle-efficient algorithm via a reduction to supervised learning (which is practically efficient if we believe supervised learning is easy).

2.3 Theory of Private Learning

More broadly, the learnability and sample complexity of private learning were studied under various models in Kasiviswanathan et al. (2011); Beimel et al. (2013, 2016); Chaudhuri and Hsu (2011); Bun et al. (2015); Wang et al. (2016); Alon et al. (2019). The VC-classes were shown to be learnable when the either the hypothesis class or the data-domain is finite (Kasiviswanathan et al., 2011). Beimel et al. (2013) characterizes the sample complexity of private learning in the realizable setting with a particular "dimension" that measures the extent to which we can construct a specific discretization of the hypothesis space that works for "all distributions" on data. Such a discretization does not exist, when $\mathcal{H}$ and $\mathcal{X}$ are both continuous. Specifically, the problem of learning threshold functions on $[0, 1]$ having VC-dimension of 1 is not privately learnable (Chaudhuri and Hsu, 2011; Bun et al., 2015).

2.4 Weaker Private Learning Models

This setting of private learning was relaxed in various ways to circumvent the above artifact. These include protecting only the labels (Chaudhuri and Hsu, 2011; Beimel et al., 2016), leveraging prior knowledge with a prior distribution (Chaudhuri and Hsu, 2011), switching to the general learning setting with Lipschitz losses (Wang et al., 2016), relaxing the distribution-free assumption (Wang et al., 2016), and the setting we consider in this paper — when we assume the availability of an auxiliary public data (Bassily et al., 2018b; Alon et al., 2019). Note that these settings are closely related to each other in that some additional information about the distribution of the data is needed.

2.5 Tsybakov Noise Condition and Statistical Learning Theory

The Tsybakov Noise Condition (TNC) (Mammen and Tsybakov, 1999; Tsybakov, 2004) is a natural and well-established condition in learning theory that has long been used in the analysis of passive as well as active learning (Boucheron et al., 2005). The Tsybakov noise condition is known to yield better convergence rates for passive learning (Hanneke, 2014), and label savings for active learning (Zhang and Chaudhuri, 2014). However, the contexts under which we use these techniques are different. For instance, while we are making the assumption of TNC, the purpose is not for active learning, but rather to establish stability. When we apply active learning, it is for the synthetic learning problem with pseudo-labels that we release privately, which does not actually satisfy TNC. To the best of our knowledge, we are the first that formally study noise models in the theory of private learning. Lastly, active learning was considered for PATE learning in (Zhao et al., 2019), which demonstrates the clear practical benefits of adaptively selecting what to label. We remain the first that provides theoretical analysis with provable learning bounds.

3. Preliminaries

In this section, first we introduce symbols and notations that we will use throughout this paper. Then we formally introduce differential privacy and discuss existing progress on PATE and model-agnostic private learning. Finally we introduce disagreement-based active learning, which is the key tool we will use for our new active learning-based PATE algorithm.

3.1 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 $\texttt{Err}(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathbb{1}(h(x) \neq y)]$, 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 $\widehat{\texttt{Err}}(h) = \frac{1}{n}\sum_{i=1}^n[\mathbb{1}(h(x_i) \neq y_i)]$. The

Table 2: Summary of symbols and notations.

Symbol	Definition	Description
$\mathbb{1}(x)$	$= 1(x = \texttt{T}), = 0(x = \texttt{F})$	indicator function
$\texttt{Err}(h)$	$\mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathbb{1}(h(x) \neq y)]$	expected risk of h w.r.t. $\mathcal{D}$
$\widehat{\texttt{Err}}(h)$	$\frac{1}{n}\sum_{i=1}^{n}[\mathbb{1}(h(x_i) \neq y_i)]$	empirical risk of h w.r.t. dataset $\{(x_i, y_i)\|i \in [n]\}$
$\mathcal{D}$		distribution over $\mathcal{Z}$
d		VC dimension
$\mathcal{D}_{\mathcal{X}}$		marginal distribution over $\mathcal{X}$
D^T	$\{(x_i^T, y_i^T)\|i \in [n]\} \sim \mathcal{D}$	labeled private teacher dataset
D^S	$\{(x_j^S)\|j \in [m]\} \sim \mathcal{D}_{\mathcal{X}}$	unlabeled public student dataset
$\texttt{DIS}$		region of disagreement in active learning
$\texttt{Dis}(h_1, h_2)$	$\mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}[\mathbb{1}(h_1(x) \neq h_2(x)]$	expected disagreement of h_1 and h_2 w.r.t $\mathcal{D}$
$\widehat{\texttt{Dis}}(h_1, h_2)$	$\frac{1}{n}\sum_{i=1}^{n}[\mathbb{1}(h_1(x_i) \neq h_2(x_i))]$	empirical disagreement of h_1 and h_2 w.r.t. $\{(x_i, y_i)\|i \in [n]\}$
$\mathcal{H}$	$\mathcal{H} \subseteq \{0, 1\}^{\mathcal{X}}$	hypothesis class
h		hypothesis, a function mapping from $\mathcal{X}$ to $\mathcal{Y}$
h^*	$\mathrm{argmin}_{h\in\mathcal{H}}\, \texttt{Err}(h)$	best hypothesis
$\hat{h}$	$\mathrm{argmin}_{h\in\mathcal{H}}\, \widehat{\texttt{Err}}(h)$	Empirical Risk Minimizer (ERM)
$\hat{h}^{\texttt{agg}}$		aggregated classifier in PATE
$\hat{h}^{\texttt{priv}}$		privately aggregated classifier in PATE
$h_{\infty}^{\texttt{agg}}$		infinite ensemble classifier
K		number of teachers
ℓ		labeling budget
m		number of unlabeled student points
n		number of labeled teacher points
$[n]$	$\{1, 2, ..., n\}$	integer set
$\tilde{O}$		big O notation hiding poly-logarithmic factors
$r(x)$	$\mathbb{E}[y\|x]$	regression function from x to y
T		cut-off threshold
$\mathcal{X}$		feature space
$\mathcal{Y}$	$\{0, 1\}$	label space
$\mathcal{Z}$	$\mathcal{X} \times \mathcal{Y}$	sample space
$\mathcal{Z}^*$	$\bigcup_{n\in\mathbb{N}} \mathcal{Z}^n$	space of a dataset of unspecified size
α		excess risk
β, γ		failure probabilities
ϵ, δ	Definition 1	parameters of differential privacy
ν, ξ	Definition 17	parameters of high margin condition
τ	Definition 8	parameter of the Tsybakov noise condition
θ	Hanneke (2014)	disagreement coefficient of active learning
$\widehat{\Delta}$	Eq. 1	realized margin
Δ	Eq. 4	expected margin
$\perp$		randomly assigned label
$\vee$	$X \vee Y = \max\{X, Y\}$	max operation
$\lesssim, \gtrsim$		inequalities hiding logarithmic factors
c, c', C		constants

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

best hypothesis h^* is defined as $h^* = \text{argmin}_{h \in \mathcal{H}} \texttt{Err}(h)$, and the Empirical Risk Minimizer (ERM) $\hat{h}$ is defined as $\hat{h} = \text{argmin}_{h \in \mathcal{H}} \widehat{\texttt{Err}}(h)$. $\hat{h}^{\texttt{agg}}$ 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 $\widehat{\texttt{Dis}}(h_1, h_2) = \frac{1}{n} \sum_{i=1}^n [\mathbb{1}(h_1(x_i) \neq h_2(x_i))]$. 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. For reader's easy reference, we summarize the symbol and notations above in Table 2.

3.2 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 $\|D - D'\|_1 = 1$) for all $\mathcal{S} \subseteq \mathcal{R}$:*

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

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 quantify 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.

3.3 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.

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

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 cut-off T, privacy parameters $\epsilon, \delta > 0$.

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

teacher ensemble's predictions are *stable* for most input data. The stability is quantified in terms of the margin function, defined as

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

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

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 A.

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 A.

Lemma 3 (Adapted from Theorem 3.11 of Bassily et al. (2018a)) *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 $\widehat{\Delta}(x_k) < K/3$ for $K = 136\log(4\ell T/\min(\delta, \beta/2)) \cdot \sqrt{T\log(2/\delta)}/\epsilon$. Then with probability at least $1-\beta$, Algorithm 2 finishes all ℓ queries and for all $i \in [\ell]$ such that $\widehat{\Delta}(x_i) \geq K/3$, the output of Algorithm 2 is $\hat{h}^{\mathsf{agg}}(x_i)$.*

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

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

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.

Next, we state and provide a straightforward proof of the following results due to (Bassily et al., 2018a). The results are already stated in the referenced work in the form of sample complexities, but we include a more direct analysis of the error bound and clarify a few technical subtleties.

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

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

Theorem 5 (Adapted from Theorems 4.6 and 4.7 of (Bassily et al., 2018a)) *Set*

$$T = 3\Big(\mathbb{E}[\mathtt{Err}(\hat{h}_1)]m + \sqrt{\frac{m\log(m/\beta)}{2}}\Big),$$

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

Let $\hat{h}^S$ 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 i.i.d.), we have

$$\texttt{Err}(\hat{h}^S) \leq \tilde{O}\Big(\frac{d^2 m \log(1/\delta)}{n^2\epsilon^2} + \sqrt{\frac{d}{m}}\Big)$$

for the realizable case, and

$$\texttt{Err}(\hat{h}^S) \leq 13\texttt{Err}(h^*) + \tilde{O}\Big(\frac{m^{1/3}d^{2/3}}{n^{2/3}\epsilon^{2/3}} + \sqrt{\frac{d}{m}}\Big)$$

for the agnostic case [1].

We provide a self-contained proof of this result in Appendix A.

Remark 6 (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 us a $\tilde{O}(\frac{d}{n^{2/3}\epsilon^{2/3}})$ error bound for the realizable case and a $O(\texttt{Err}(h^*)) + \tilde{O}(\frac{d^{3/5}}{n^{2/5}\epsilon^{2/5}})$ error bound for the agnostic case* [2].

3.4 Disagreement-Based Active Learning

We adopt the disagreement-based active learning algorithm that comes with strong learning bounds (see, e.g., an excellent treatment of the subject in (Hanneke, 2014)). The exact algorithm, described in Algorithm 4, keeps updating a subset of the hypothesis class $\mathcal{H}$ called a *version space* by collecting labels only from those data points from a certain *region of disagreement* and eliminates candidate hypothesis that are certifiably suboptimal.

Definition 7 (Region of disagreement (Hanneke, 2014)) *For a given hypothesis class $\mathcal{H}$, its region of disagreement is defined as a set of data points over which there exists two hypotheses disagreeing with each other,*

$$\texttt{DIS}(\mathcal{H}) = \{x \in \mathcal{X} : \exists h, g \in \mathcal{H} \text{ s.t. } h(x) \neq g(x)\}.$$

Region of disagreement is the key concept of the disagreement-based active learning algorithm. It captures the uncertainty region of data points for the current version space. The algorithm is fed a sequence of data points and runs in the online fashion, whenever there exists a data point in this region, its label will be queried. Then any *bad* hypotheses will be removed from the version space.

The algorithm, as it is written is not directly implementable, as it represents the version spaces explicitly, but there are practical implementations that avoids explicitly representing the versions spaces by a reduction to supervised learning oracles. In our experiments, we implement the PATE-ASQ algorithm and show it works well in practice while no explicit region of disagreement is maintained.

1. The numerical constant 13 might be improvable (and it is indeed worse than the result stated in Bassily et al. (2018b)), though we decide to present this for the simplicity of the proof.
2. These correspond to the $\tilde{O}((d/\alpha)^{3/2})$ sample complexity bound in Theorem 4.6 of (Bassily et al., 2018a) 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., 2018a) for agnostic PAC learning with error $O(\alpha + \texttt{Err}(h^*))$. The privacy parameter ϵ is taken as a constant in these results.

Algorithm 4 Disagreement-Based Active Learning (Hanneke, 2014)

Input: A "data stream" $x_1, x_2, \ldots$ sampled i.i.d. from distribution $\mathcal{D}$. A hypothesis class $\mathcal{H}$. An on-demand "labeling service" that outputs label $y_i \sim P(y|x = x_i)$ when requested at time i. Parameter ℓ, m, γ.

```
Initialize the version space V ← H.
Initialize the selected dataset Q ← ∅.
Initialize "Current Output" to be any h ∈ H.
Initialize "Counter" c ← 0.
for j ∈ [m] do
    if x_j ∈ DIS(V) then
        "Request for label" for x_j and get back y_j from the "labeling service".
        Update Q ← Q ∪ {(x_j, y_j)}.
        c ← c + 1.
    end if
    if log_2(j) ∈ ℕ then
        Update V ← {h ∈ V : (Err_Q(h) − min_{g∈V} Err_Q(g))|Q| ≤ U(j, γ_j)j},
        where
        U(j, γ_j) = c′(d log(θ(d/j))+log(1/γ_j))/j + c′ sqrt(Err(h*)(d log(θ(Err(h*)) + log(1/γ_j))/j),
        c′ is a constant, and γ_j = γ/(log_2(2j))^2.
        Set "Current Output" to be any h ∈ V.
    end if
    if c ≥ ℓ then
        Break.
    end if
end for
```

Output: Return "Current Output".

4. Main Results

In Section 4.1 and 4.2, we present a more refined theoretical analysis of PATE-PSQ (Algorithm 3) that uses SVT-based PATE (Algorithm 2) as the subroutine. Our results provide stronger learning bounds and new theoretical insights under various settings. In Section 4.3, 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 PATE (Algorithm 1) as the subroutine. For comparison, we also include an analysis of standard PATE (with Gaussian mechanism) in Appendix A. Table 1 summarizes these technical results.

4.1 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 existing bound presented in Theorem 5 is vacuous if $\texttt{Err}(h^*) > 1/26$. 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 $\texttt{Err}(h^*) > 0$, we introduce a large family of learning problems that satisfy the so-called 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 $r(x) = \mathbb{E}[y|x]$. The Tsybakov noise condition is defined in terms of the distribution of $r(x)$.

Definition 8 (Tsybakov noise condition) *The joint distribution of the data (x, y) satisfies the Tsybakov noise condition with parameter τ if there exists a universal constant $C > 0$ such that for all $t \geq 0$*

$$\mathbb{P}(|r(x)| \leq t) \leq C t^{\frac{\tau}{1-\tau}}.$$

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 $|r(x)|$ 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 8 (see a proof from Bousquet et al. (2004, Definition 7)).

Lemma 9 (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,

$$\texttt{Dis}(h, h_{\text{Bayes}}) \leq \eta(\texttt{Err}(h) - \texttt{Err}(h_{\text{Bayes}}))^{\tau}. \tag{2}$$

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

In the remainder of this section, we make the assumption that the Bayes optimal classifier $h_{\text{Bayes}} \in \mathcal{H}$ and works with the slightly weaker condition that requires (2) to hold only for $h \in \mathcal{H}$ and that we replace h_{Bayes} by the optimal classifier $h^* \in \mathcal{H}$ [3].

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 $\texttt{Err}(h^*)$ 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 10 (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*

$$T = \tilde{O}\Big(\Big(\frac{m^{2-\tau} d^{\tau}}{n^{\tau}\epsilon^{\tau}}\Big)^{\frac{2}{4-3\tau}}\Big),$$

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

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

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

Remark 11 (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 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. (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.*

3. This slightly different condition, that requires (2) to hold only for $h \in \mathcal{H}$ but with h_{Bayes} replaced by the optimal classifier h^* (without assuming that $h^* = h_{\text{Bayes}}$) 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.

Remark 12 (Optimal choice of m) *The upper bound above can be minimized by choosing $m^* = (d^{4-5\tau} n^{2\tau} \epsilon^{2\tau})^{\frac{1}{4-2\tau}}$. When number of available public data points $m \geq m^*$, then m is not a limiting factor and we should subsample these data points. When $m < m^*$, then d/m is the leading factor, we should use all m data points.*

There are two key observations behind the improvement. First, the teacher classifiers do not have to agree on the labels y as in Lemma 4; 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 10.

Lemma 13 (Performance of teacher classifer w.r.t. h^*) *With probability $1-\gamma$ over the training data of $\hat{h}_1, ..., \hat{h}_K$, assume $h^* \in \mathcal{H}$ is the Bayes optimal classifier and Tsybakov noise condition with parameter τ, then there is a universal constant C such that for all $k = 1, 2, 3, ..., K$*

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

Proof By the equivalent definition of the Tsybakov noise condition and then the learning bounds under TNC (Lemma 38),

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

■

Lemma 14 (Total number of mistakes made by one teacher) *Under the condition of Lemma 13, with probability $1-\gamma$, for all $k = 1, 2, ..., K$ the total number of mistakes made by one teacher classifier $\hat{h}_k$ with respect to h^* can be bounded as:*

$$\sum_{j=1}^{m} \mathbb{1}(\hat{h}_k(x_j) \neq h^*(x_j)) \leq O\bigg(\max\Big\{m\texttt{Dis}(\hat{h}_k, h^*), \log\Big(\frac{K}{\gamma}\Big)\Big\}\bigg).$$

Proof Number of mistakes made by $\hat{h}_k$ with respect to h^* is the empirical disagreement between $\hat{h}_k$ and h^* on m data points, therefore, by Bernstein's inequality (Lemma 35),

$$\begin{aligned}\sum_{j=1}^{m} \mathbb{1}(\hat{h}_k(x_j) \neq h^*(x_j)) &\leq O\bigg(m\texttt{Dis}(\hat{h}_k, h^*) + \sqrt{m\texttt{Dis}(\hat{h}_k, h^*)\log\Big(\frac{K}{\gamma}\Big)} + \log\Big(\frac{K}{\gamma}\Big)\bigg)\\ &\leq O\bigg(\max\Big\{m\texttt{Dis}(\hat{h}_k, h^*), \log\Big(\frac{K}{\gamma}\Big)\Big\}\bigg).\end{aligned}$$

■

Using the above two lemmas we establish a bound on the number of examples where the differentially privately released labels differ from the prediction of h^*.

Lemma 15 (Total queries and cut-off budget) *Let Algorithm 3 be run with the number of teachers K and the cut-off parameter T chosen according to Theorem 10. Assume the conditions of Lemma 13. Then with high probability ($\geq 1-\beta$ over the random coins of Algorithm 3 alone and conditioning on the high probability events of Lemma 13 and Lemma 14), Algorithm 3 finishes all m queries without exhausting the cut-off budget and that*

$$\sum_{j=1}^{m} \mathbb{1}(\hat{h}_j^{\texttt{priv}} \neq h^*(x_j)) \leq T.$$

The $\tilde{O}$ notation in the choice of K and T hides polynomial factors of $\log(K/\gamma), \log(m/\beta)$ *where γ is from Lemma 13 and 14.*

Proof Denote the bound from Lemma 14 by B. By the same Pigeon hole principle argument as in Lemma 4 (but with y replaced by h^*), we have that the number of queries that have margin smaller than $K/6$ is at most $3B = O(\max\{m\texttt{Dis}(\hat{h}_k, h^*), \log(K/\gamma)\})$. The choice of K ensures that with high probability, over the Laplace random variables in Algorithm 2, in at least $m - 3B$ queries where the answer $\hat{y}_j = h^*(x_j)$, i.e.,

$$\sum_{j=1}^{m} \mathbb{1}(\hat{h}_j^{\texttt{priv}} \neq h^*(x_j)) \leq 3B := T.$$

■

Now we are ready to put everything together and prove Theorem 10.

Proof [Proof of Theorem 10] Denote $\tilde{h} = \operatorname{argmin}_{h\in\mathcal{H}} \widehat{\texttt{Dis}}(h, h^*)$ where $\widehat{\texttt{Dis}}$ is the empirical average of the disagreements over the data points that students have[4]. By the triangular inequality of the $0-1$ error,

$$\begin{aligned}\texttt{Err}(\hat{h}^S) - \texttt{Err}(h^*) &\leq \texttt{Dis}(\hat{h}^S, h^*) \\ &\leq \widehat{\texttt{Dis}}(\hat{h}^S, h^*) + 2\sqrt{\frac{(d+\log(4/\gamma))\widehat{\texttt{Dis}}(\hat{h}^S, h^*)}{m}} + \frac{4(d+\log(4/\gamma))}{m} \\ &\leq 2\widehat{\texttt{Dis}}(\hat{h}^S, h^*) + \frac{5(d+\log(4/\gamma))}{m},\end{aligned} \tag{3}$$

where the second line follows from the uniform Bernstein's inequality — apply the first statement Lemma 36 in Appendix C with $z = h^*(x)$ and the third line is due to $a+2\sqrt{ab}+b \leq 2a+2b$ for non-negative a, b.

4. Note that in this case we could take $\tilde{h} = h^*$ since $h^* \in \mathcal{H}$. We are defining this more generally so later we can substitute h^* with other label vector that are not necessarily generated by any hypothesis in $\mathcal{H}$.

By the triangular inequality, we have $\widehat{\texttt{Dis}}(\hat{h}^S, h^*) \le \widehat{\texttt{Dis}}(\hat{h}^S, \hat{h}^{\texttt{priv}}) + \widehat{\texttt{Dis}}(\hat{h}^{\texttt{priv}}, h^*)$, therefore

$$\begin{aligned}(3) &\le 2\widehat{\texttt{Dis}}(\hat{h}^S, \hat{h}^{\texttt{priv}}) + 2\widehat{\texttt{Dis}}(\hat{h}^{\texttt{priv}}, h^*) + \frac{5(d+\log(4/\gamma))}{m} \\ &\le 2\widehat{\texttt{Dis}}(\tilde{h}, \hat{h}^{\texttt{priv}}) + 2\widehat{\texttt{Dis}}(\hat{h}^{\texttt{priv}}, h^*) + \frac{5(d+\log(4/\gamma))}{m} \\ &\le 2\widehat{\texttt{Dis}}(\tilde{h}, h^*) + 4\widehat{\texttt{Dis}}(\hat{h}^{\texttt{priv}}, h^*) + \frac{5(d+\log(4/\gamma))}{m} \\ &= 4\widehat{\texttt{Dis}}(\hat{h}^{\texttt{priv}}, h^*) + \frac{5(d+\log(4/\gamma))}{m}.\end{aligned}$$

In the second line, we applied the fact that $\hat{h}^S$ is the minimizer of $\widehat{\texttt{Dis}}(h, \hat{h}^{\texttt{priv}})$; in the third line, we applied triangular inequality again and the last line is true because $\widehat{\texttt{Dis}}(\tilde{h}, h^*) = 0$ since $\tilde{h}$ is the minimizer and that $h^* \in \mathcal{H}$.

Recall that T is the unstable cutoff in Algorithm 3. The proof completes by invoking Lemma 15 which shows that the choice of T is appropriate such that $\widehat{\texttt{Dis}}(\hat{h}^{\texttt{priv}}, h^*) \le T/m$ with high probability. ■

In the light of the above analysis, it is clear that the improvement from our analysis under TNC are two-folds: (1) We worked with the disagreement with respect to h^* rather than y. (2) We used a uniform Bernstein bound rather than a uniform Hoeffding bound that leads to the faster rate in terms of the number of public data points needed.

Remark 16 (Reduction to ERM) *The main challenge in the proof is to appropriately take care of $\hat{h}^{\texttt{priv}}$. 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 are using the SVT-based Algorithm 2, $\hat{h}^{\texttt{priv}}$ 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 37. 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 36) directly. This is without loss of generality because all learnable problems are learnable by (asymptotic) ERM (Vapnik, 1995; Shalev-Shwartz et al., 2010).*

4.2 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 has taught us that the voting classifier is qualitatively different from the individual voters. In particular, the error rate of 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 with two examples.

Example 1 (Voting fails) *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 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
$\hat{h}^{\mathsf{agg}}$	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 2 (Voting wins) *If $\mathbb{P}[\hat{h}(x) \neq y|x] \leq 0.5 - \xi$, where ξ is a small constant, for all $x \in \mathcal{X}$, then by Hoeffding's inequality,*

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

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

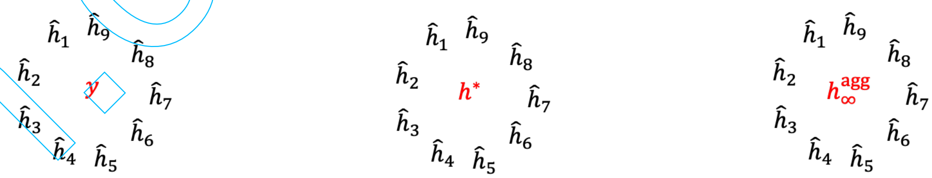
These cases call for an alternative distribution-dependent theory of learning that characterizes the performance of Algorithm 3 more accurately.

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

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

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.

Considering different centers for teacher classifiers to agree on is one of the key ideas of this paper. Figure 2 shows three kinds of centers for teachers $\hat{h}_1, \hat{h}_2, ..., \hat{h}_9$ to agree on.



(a) True label y is the center for in realizable setting. (b) Best hypothesis h^* is the center under TNC. (c) h^{agg}_∞ is our new construction for agnostic setting.

Figure 2: Centers for teachers $\hat{h}_1, \hat{h}_2, ..., \hat{h}_9$ to agree on.

In Bassily et al. (2018b), the center is the true label y in the realizable setting. In Section 4.1 under TNC, we analyze the performance of PATE-PSQ, where the center is the best hypothesis h^*. Now we are interested in the new center h^{agg}_∞ for teachers to agree on.

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

$$\Delta_n(x) := \Big|\mathbb{E}[\hat{h}_1(x)|x] - \frac{1}{2}\Big|. \tag{4}$$

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 2, we do not require the teachers to agree on y. Instead, it measures the extent to which they agree on $h^{\mathsf{agg}}_\infty(x)$, which could be any label.

When the expected margin is bounded away from 0 for x, then the voting classifier outputs $h^{\mathsf{agg}}_\infty(x)$ 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 17 (Approximate high margin) *We say that a learning problem with n i.i.d. samples satisfy* (ν, ξ)*-approximate high-margin condition if* $\mathbb{P}_{x\sim\mathcal{D}}[\Delta_n(x) > \xi] \leq \nu$.

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 ξ. Observe that every learning problem has ξ 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 ξ (see more details in Remark 21). In Appendix E, we demonstrate that for the problem of linear classification on Aadult dataset — clearly an agnostic learning problem — (ν, ξ)-approximate high margin condition is satisfied with a small ν and large ξ.

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 18 *Assume the learning problem with* n/K *i.i.d. data points satisfies* (ν, ξ)*-approximate high-margin condition. Let Algorithm 2 be instantiated with parameters*

$$T \geq \nu m + \sqrt{2\nu m \log\left(\frac{3}{\gamma}\right)} + \frac{2}{3}\log\left(\frac{3}{\gamma}\right),$$

$$K \geq \max\left\{\frac{2\log(3m/\gamma)}{\xi^2}, \frac{3\lambda\big(\log(4m/\delta)+\log(3m/\gamma)\big)}{\xi}\right\},^{5}$$

then with high probability (over the randomness of the n i.i.d. samples of the private dataset, m i.i.d. samples of the public dataset, and that of the randomized algorithm), Algorithm 2 finishes all m rounds and the output is the same as $h_\infty^{\mathtt{agg}}(x_i)$ for all but T of the $i \in [m]$.

This proposition provides the utility guarantee to Algorithm 2 and generalizes Lemma 15 from fixing $\xi = 1/6$ into allowing much smaller ξ at a cost of increasing ν.

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

Theorem 19 *Assume the learning problem with n/K i.i.d. data points satisfies (ν, ξ)-approximate high-margin condition and let K, T be chosen according to Proposition 18, 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 $\geq 1 - 2\gamma$,*

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

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. Note if the VC dimension of $\mathcal{H}$ is d, then the VC dimension of $\mathrm{Vote}_K(\mathcal{H}) \leq Kd$. In practice, this suggests using ensemble methods such as AdaBoost for K iterations.

Theorem 20 *Under the same assumption of Theorem 19, 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 that with probability $\geq 1 - 2\gamma$,*

$$\mathtt{Err}(\hat{h}^S) - \mathtt{Err}(h_\infty^{\mathtt{agg}}) \leq \frac{4T}{m} + \frac{5(Kd + \log(4/\gamma))}{m} = \tilde{O}\Big(\nu + \frac{\log(4/\gamma)}{m} + \frac{d\sqrt{\nu}}{\xi\sqrt{m}}\Big).$$

Remark 21 *Whether the bounds in Theorem 19 and 20 will vanish as $m, n \to \infty$ depends strongly on how parameter ν and ξ 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, ξ)-approximate high-margin condition with $\xi = 1/6$ and $\nu \leq \tilde{O}((dK/n)^{\frac{\tau}{2-\tau}})$. Substituting this quantities into Theorem 19 and using the fact that ν also bounds the disagreement between h^* and $h_\infty^{\mathtt{agg}}$ 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.

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

4.3 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 variants of PATE that aim at *passively* releasing *almost all* student queries. In this section we address the following question:

Can we do even 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 5, we propose a new algorithm called PATE with Active Student Queries (PATE-ASQ) which uses the disagreement-based active learning algorithm (Algorithm 4) as the subroutine. Then we provide its utility guarantee.

Algorithm 5 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 4) 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 $\hat{h}^S$.

Theorem 22 (Utility guarantee of Algorithm 5) *With probability at least $1-\gamma$, there exists universal constants C_1, C_2 such that for all*

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

the output $\hat{h}^S$ of Algorithm 5 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

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

Specifically, when we choose

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

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

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

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

Remark 23 *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 24 (Blackbox reduction, revisited) *In contrary to our discussion in Remark 16, 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.*

Remark 25 (Relationships between SVT and active learning) *There is an intriguing analogy between the Algorithm 2 which simply labels all queries with an advanced DP mechanism and Algorithm 1 which uses active learning with a simple DP mechanism. On a high level, 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 curious question is whether the two sets of selected data points are substantially overlapping. If not, then we might be able to combine the two and achieve even better private-utility tradeoff.

5. Experiments

In this section, we present our empirical studies of PATE-PSQ and PATE-ASQ algorithms. Section 5.1 describes how we set up our experiments, and Section 5.2 show our results.

5.1 Experimental Settings

Algorithms compared. We focus on comparing the classification accuracy of the passive and active learning versions of PATE on a holdout test set ("Utility") when both algorithms are calibrated to the same privacy budget ϵ ("privacy"). To set baselines, we also compare them with non-private versions of them (no noise added to the votes, or $\epsilon = +\infty$), denoted by PATE-PSQ-NP and PATE-ASQ-NP. We remark that the PATE-PSQ we implement is the Gaussian mechanism version (Papernot et al., 2018). While we have shown that it has higher *asymptotic* sample complexity comparing to the more advanced version based on SVT (Bassily et al., 2018b) (Section 4.1), we found that the Gaussian mechanism version performs better for the realistically-sized datasets that we considered. Linear models are used for all of these algorithms for simplicity. For active learning, we follow the practical implementation of the disagreement-based active learning by Yan et al. (2018), which does not require the learner to explicitly maintain the (exponentially large) region of disagreement.

Datasets. We do our experiments on three binary classification datasets, mushroom, a9a, and real-sim. All of them are obtained from LIBSVM dataset website [6]. See Table 3 for the statistics of them. If a dataset had been previously split into training and testing parts, we combine them together and record the total number of all data points. For all datasets, 80% of all data points are randomly selected to be considered private and used to train teacher classifiers. 2% of all data points are randomly selected as public student unlabeled data points. The remaining 18% data points are reserved for testing. We repeat these random selection processes for 30 times.

Table 3: Statistics of datasets.

Dataset	# All	# Train	# Unlabeled	# Budget	# Test	# Dimension
mushroom	8,124	6,499	163	49	1,462	112
a9a	48,842	39,073	977	293	8,792	123
real-sim	72,309	57,847	1,447	434	13,015	20,958

Parameter settings. Number of teachers K is set on all datasets so that each teacher classifier gets trained with approximately 100 data points. 30% of student unlabeled data points are set as the total budget of queries for PATE-ASQ and PATE-ASQ-NP. See Table 3. $\epsilon = 0.5, 1.0, 2.0$ and $\delta = 1/n$ are set as privacy parameters for all datasets, where n is number of private teacher data points. All privacy accounting and calibration are conducted via AutoDP (Wang et al., 2019), and the tight analytical calibration and composition of Gaussian mechanisms are due to (Balle and Wang, 2018).

Privacy loss vs. privacy budget. Besides the privacy budget parameter ϵ that the algorithms receive as an input, it is often the case that the active learning algorithm halts before exhausting the query budget of (30% of the total number of unlabeled data points). Therefore the privacy loss incurred after running PATE-ASQ might be smaller than the prescribed privacy budget. We refer to the privacy loss $\epsilon_{\texttt{ex post}}$, since it is determined by the output.

6. `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/`

Table 4: Utility and privacy results of different PATE models. **# Queries** shows the number of queries actually answered in experiments. **Accuracy** is reported as $\texttt{mean} \pm 1.96 \times \texttt{standard\_error}/\sqrt{30}$, i.e., 98% asymptotic confidence interval of the expected accuracy based on inverting Wald's test. All "PATE-" prefixes of methods are omitted to improve readability.

Dataset	Method	# Queries	ϵ	$\epsilon_{\texttt{ex post}}$	Accuracy
mushroom	PSQ-NP	163	$+\infty$	$+\infty$	$\mathbf{0.9773} \pm 0.0006$
	ASQ-NP	47.3 ± 0.2	$+\infty$	$+\infty$	0.9146 ± 0.0036
	PSQ	163	0.5	0.5	0.6416 ± 0.0036
	ASQ	40.1 ± 0.7	0.5	**0.4461**	$\mathbf{0.6418} \pm 0.0091$
	PSQ	163	1.0	1.0	0.7534 ± 0.0045
	ASQ	42.9 ± 0.5	1.0	**0.9267**	$\mathbf{0.7727} \pm 0.0098$
	PSQ	163	2.0	2.0	$\mathbf{0.8974} \pm 0.0027$
	ASQ	46.5 ± 0.3	2.0	**1.9410**	0.8858 ± 0.0059
a9a	PSQ-NP	977	$+\infty$	$+\infty$	$\mathbf{0.5555} \pm 0.0157$
	ASQ-NP	225.6 ± 5.0	$+\infty$	$+\infty$	0.5461 ± 0.0160
	PSQ	977	0.5	0.5	0.5040 ± 0.0034
	ASQ	293	0.5	0.5	$\mathbf{0.5212} \pm 0.0088$
	PSQ	977	1.0	1.0	0.5171 ± 0.0050
	ASQ	290.8 ± 0.8	1.0	**0.9958**	$\mathbf{0.5369} \pm 0.0103$
	PSQ	977	2.0	2.0	0.5176 ± 0.0070
	ASQ	290.3 ± 0.9	2.0	**1.9896**	$\mathbf{0.5543} \pm 0.0089$
real-sim	PSQ-NP	1,447	$+\infty$	$+\infty$	0.8234 ± 0.0014
	ASQ-NP	434	$+\infty$	$+\infty$	$\mathbf{0.8289} \pm 0.0008$
	PSQ	1,447	0.5	0.5	0.6355 ± 0.0065
	ASQ	434	0.5	0.5	$\mathbf{0.7389} \pm 0.0014$
	PSQ	1,447	1.0	1.0	0.7550 ± 0.0058
	ASQ	434	1.0	1.0	$\mathbf{0.8040} \pm 0.0009$
	PSQ	1,447	2.0	2.0	0.8025 ± 0.0037
	ASQ	434	2.0	2.0	$\mathbf{0.8231} \pm 0.0009$

5.2 Experimental Results

The results are presented in Table 4, where both utility (classification accuracy on the test set) and privacy (privacy budget ϵ and privacy loss $\epsilon_{\text{ex post}}$) metrics are reported. Best results in each category are marked in bold fonts. We make a few observations of the results below.

1. Given the same privacy budget, ASQ performs substantially better than PSQ in most cases. The improvement is sometimes 10% (real-sim / $\epsilon = 0.5$). The only exception is when $\epsilon = 2.0$ on the "mushroom" dataset, in which the active learning performed

substantially worse than the passive-learning counterpart in the non-private baseline as well.

2. ASQ incurs a smaller private loss $\epsilon_{\texttt{ex post}}$ than PSQ, due to possibly fewer queries being selected by the active learning algorithm than the pre-specified query budget.

3. As ϵ increases, less noise is injected by the Gaussian mechanisms, and the performance improves for both PSQ and ASQ. In the regime of small ϵ (stronger privacy), we often see a greater improvement in ASQ.

4. ASQ requires privately releasing a much smaller number of labels while maintaining comparable performances as PSQ. Although ASQ algorithms use up all labeling budget on real-sim datasets, ASQ algorithms do not run out of them on mushroom and a9a datasets in most cases.

5. ASQ-NP does not always perform better than PSQ-NP algorithms, which meets our understanding from active learning literature. It only performs better than PSQ-NP on real-sim datasets.

6. Conclusion

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 (TNC) 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 and showed that simple Gaussian mechanism suffices for obtaining the same fast rates under TNC. In addition, our experiments on PATE-ASQ show it works as an efficient algorithm in practice.

Future work includes understanding different selections made by sparse vector technique and active learning, as well as addressing the open theoretical problem *at large* — developing ERM-oracle efficient algorithm for the private agnostic learning when a public unlabeled dataset is available.

Acknowledgments

The work is partially supported by NSF CAREER Award #2048091 and generous gifts from NEC Labs, Google and Amazon Web Services. The authors would like to thank Songbai Yan for sharing their code with a practical implementation of the disagreement-based active learning algorithms from Yan et al. (2018).

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.

Borja Balle and Yu-Xiang Wang. Improving gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *International Conference in Machine Learning (ICML-18)*, 2018.

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.

Reza Shokri and Vitaly Shmatikov. Privacy-preserving deep learning. In *Conference on Computer and Communications Security (CCS-15)*, pages 1310–1321, 2015.

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.

Yu-Xiang Wang, Borja Balle, and Shiva Kasiviswanathan. Subsampled rényi differential privacy and analytical moments accountant. In *International Conference on Artificial Intelligence and Statistics (AISTATS-19)*, 2019.

Songbai Yan, Kamalika Chaudhuri, and Tara Javidi. Active learning with logged data. In *International Conference on Machine Learning (ICML-18)*, pages 5521–5530, 2018.

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.

Appendix A. Proofs of Stated Technical Results

A.1 Proofs of Existing Results

In this subsection, we provide the privacy analysis as well as reproving the results of Bassily et al. (2018b) in our notation so that it becomes clear where the improvement is coming from.

Theorem 26 (Restatement of Theorem 2) *Algorithm 1 and 2 are both (ϵ, δ)-DP.*

The proof for Algorithm 1 follows straightforwardly from Gaussian mechanism because the number of "teachers" who predict 1 will have a global sensitivity of 1. The proof for Algorithm 2 is more delicate. It follows the arguments in the proof of Theorem 3.6 of Bassily et al. (2018a) for the most part, which combines the *sparse vector technique* (SVT) (Hardt and Rothblum, 2010) with the *distance to stability* approach from Thakurta and Smith (2013). The only difference in the stated result here is that we used the modern CDP approach to handle the composition which provides tighter constants.

Proof First note that the global sensitivity (Definition 41) of the vote count is 1. Algorithm 1 is a straightforward adaptive composition of ℓ Gaussian mechanisms (Lemma 44), which satisfies $\frac{\ell}{2\sigma^2}$-zCDP. By Lemma 51, we get that the choice of σ gives us (ϵ, δ)-differential privacy.

Let us now address Algorithm 2. First note that $\widehat{\Delta}(x_j)$ as a function of the input dataset D has a global sensitivity of 2 for all x_j, thus dist_j has a global sensitivity of 1. Following the proof of Theorem 3.6 of Bassily et al. (2018a), Algorithm 2 can be considered a composition of Sparse Vector Technique (SVT) (Algorithm 6), which outputs a binary vector of $\{\perp, \top\}$ indicating the failures and successes of passing the screening by SVT, and the distance-to-stability mechanism (Algorithm 7) which outputs $\{\hat{h}^{\mathsf{agg}}(x_j)\}$ for all coordinates where the output is $\perp$. Check that the length of this binary vector is random and is between T and ℓ. The number of $\top$ is smaller than T. If $\{\hat{h}^{\mathsf{agg}}(x_j)\}$ is not revealed, then this would be the standard SVT, and the challenge is to add the additional outputs.

The key trick of the proof inspired from the privacy analysis (Lemma 48) of the distance-to-stability is to discuss the two cases. In the first case, assume for all j such that the output is $\perp$, $\hat{h}^{\mathsf{agg}}(x_j)$ remains the same over D, D', then adding $\hat{h}^{\mathsf{agg}}(x_j)$ to the output obeys 0-DP; in the second case, assume that there exists some j where we output $\perp$ such that, $\hat{h}^{\mathsf{agg}}(x_j)$ is different under D and D', then for all these j we know that $\mathrm{dist}_j = 0$ for both D and D'. By the choice of λ, w, we know that the second case happens with probability at most $\delta/2$ using the tail of Laplace distribution and a union bound over all $\ell + T$ independent Laplace random variables. Note that this holds uniformly over all possible adaptive choices of the nature, since this depends only on the added noise.

Conditioning on the event that the second case does not happen, the output of the algorithm is only the binary vector of $\{\perp, \top\}$ from SVT. The SVT with cutoff T is an adaptive composition of T SVTs with cutoff$= 1$. By our choice of parameter λ, each such SVT with cutoff$= 1$ obeys pure-DP with privacy parameter $2/\lambda$, hence also satisfy CDP with parameter $2/\lambda^2$ by Proposition 1.4 of Bun and Steinke (2016). Composing over T SVTs, we get a CDP parameter of $2T/\lambda^2$. By Proposition 1.3 of Bun and Steinke (2016) (Lemma 51), we can convert CDP to DP. The choice of λ is chosen such that the composed mechanism obeys $(\epsilon, \delta/2)$-DP. Combining with the second case above, this establishes the (ϵ, δ)-DP of Algorithm 2. ■

Theorem 27 (Restatement of Theorem 5) *Set*

$$T = 3\Big(\mathbb{E}[\texttt{Err}(\hat{h}_1)]m + \sqrt{\frac{m\log(m/\beta)}{2}}\Big),$$

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

Let $\hat{h}^S$ 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 i.i.d.), we have

$$\texttt{Err}(\hat{h}^S) \le \tilde{O}\Big(\frac{d^2 m\log(1/\delta)}{n^2\epsilon^2} + \sqrt{\frac{d}{m}}\Big)$$

for the realizable case, and

$$\texttt{Err}(\hat{h}^S) \le 13\texttt{Err}(h^*) + \tilde{O}\Big(\frac{m^{1/3}d^{2/3}}{n^{2/3}\epsilon^{2/3}} + \sqrt{\frac{d}{m}}\Big)$$

for the agnostic case.

Proof The analysis essentially follows the proof of Theorem 10 by replacing h^* with y. First, by Hoeffding's inequality, with probability $1-\beta$ over the teacher data points, the total number of mistakes made by each teacher classifier is at most $m\mathbb{E}[\texttt{Err}(\hat{h}_1)] + \sqrt{m\log(m/\beta)/2}$, which is B in Lemma 4. Then following Lemma 4, by choose $T = 3B = 3(m\mathbb{E}[\texttt{Err}(\hat{h}_1)] + \sqrt{m\log(m/\beta)/2})$, we ensure that the majority voting classifiers are correct and have high margin in at least $m - T$ examples.
In the realizable setting. Since $\texttt{Err}(h^*) = 0$ and by standard statistical learning theory in the realizable case (Lemma 37), for each teacher classifier $\hat{h}_k$ we have

$$\texttt{Err}(\hat{h}_k) \le 4\frac{d\log(n/K) + \log(4/\gamma)}{n/K}.$$

Substitute our choice of $K = \tilde{O}(\sqrt{T\log(1/\delta)}/\epsilon)$ as in Lemma 3 we get that w.h.p.

$$\texttt{Err}(\hat{h}_k) \le \tilde{O}\Big(\frac{d\sqrt{T\log(1/\delta)}}{n\epsilon}\Big).$$

Plug in the bound into our choice of $T = 3(m\mathbb{E}[\texttt{Err}(\hat{h}_1)] + \sqrt{m\log(m/\beta)/2})$, we get

$$T \le \tilde{O}\Big(\frac{dm\sqrt{T\log(1/\delta)}}{n\epsilon} + \sqrt{\frac{m\log(m/\beta)}{2}}\Big).$$

By solving the quadratic inequality, we get that T obeys

$$T \le \tilde{O}\Big(\frac{d^2 m^2 \log(1/\delta)}{n^2\epsilon^2} + \sqrt{m}\Big).$$

Recall that this choice of K and T ensures that Algorithm 2 will have at most T unstable queries during the m rounds, which implies that with high probability, the privately released pseudo-labels to those "stable" queries are the same as the corresponding true labels.

Now the next technical subtlety is to deal with the dependence in the student learning problem created by the pseudo-labels via a reduction to an ERM learner. By the standard Hoeffding-style uniform convergence bound (Lemma 36),

$$\begin{aligned}
\texttt{Err}(\hat{h}^S) &\le \widehat{\texttt{Err}}(\hat{h}^S) + \tilde{O}\Big(\sqrt{\frac{d}{m}}\Big) \\
&\le \widehat{\texttt{Err}}(\hat{h}^{\texttt{priv}}) + \widehat{\texttt{Dis}}(\hat{h}^{\texttt{priv}}, \hat{h}^S) + \tilde{O}\Big(\sqrt{\frac{d}{m}}\Big) \\
&\le 2\widehat{\texttt{Err}}(\hat{h}^{\texttt{priv}}) + \tilde{O}\Big(\sqrt{\frac{d}{m}}\Big) \\
&\le \frac{2T}{m} + \tilde{O}\Big(\sqrt{\frac{d}{m}}\Big) \\
&= \tilde{O}\Big(\frac{d^2 m \log(1/\delta)}{n^2\epsilon^2} + \sqrt{\frac{d}{m}}\Big). \qquad (5)
\end{aligned}$$

where we applied the triangular inequality in the second line, used that $\hat{h}^S$ is the minimizer of $\widehat{\texttt{Dis}}(\hat{h}^{\texttt{priv}}, \cdot)$ in the third line, and then combined Lemma 3 and Lemma 4 to show that under the appropriate choice of T and K with high probability, $\hat{h}^{\texttt{priv}}(x_j)$ correctly returns y_j except for up to T example. Finally, the choice of T is substituted.

In agnostic setting. By Lemma 37, with high probability, for all teacher classifier $\hat{h}_k$ for $k = 1, ..., K$, we have

$$\texttt{Err}(\hat{h}_k) - \texttt{Err}(h^*) \le \tilde{O}\Big(\sqrt{\frac{d\log(n/K) + \log(4/\gamma)}{n/K}}\Big).$$

Substitute the choice of $K = \tilde{O}(\sqrt{T\log(1/\delta)}/\epsilon)$ from Lemma 3, we get

$$\texttt{Err}(\hat{h}_k) \le \texttt{Err}(h^*) + \tilde{O}\Big(\frac{d^{1/2}T^{1/4}}{n^{1/2}\epsilon^{1/2}}\Big).$$

Plug in the above bound into our choice $T = 3(m\mathbb{E}[\texttt{Err}(\hat{h}_1)] + \sqrt{m\log(m/\beta)/2})$, we get that

$$T \le 3m\texttt{Err}(h^*) + \tilde{O}(\sqrt{m}) + \tilde{O}\Big(\frac{md^{1/2}T^{1/4}}{n^{1/2}\epsilon^{1/2}}\Big). \qquad (6)$$

Further, we can write

$$\begin{aligned} T &\le 2(3m\texttt{Err}(h^*) + \tilde{O}(\sqrt{m})) \cdot \mathbb{1}\Big(\tilde{O}\Big(\frac{md^{1/2}T^{1/4}}{n^{1/2}\epsilon^{1/2}}\Big) \le \frac{T}{2}\Big) \\ &\quad + \Big(2\tilde{O}\Big(\frac{md^{1/2}}{n^{1/2}\epsilon^{1/2}}\Big)\Big)^{4/3} \cdot \mathbb{1}\Big(\tilde{O}\Big(\frac{md^{1/2}T^{1/4}}{n^{1/2}\epsilon^{1/2}}\Big) > \frac{T}{2}\Big) \\ &\le 6m\texttt{Err}(h^*) + \tilde{O}(\sqrt{m}) + \tilde{O}\Big(\frac{m^{4/3}d^{2/3}}{n^{2/3}\epsilon^{2/3}}\Big), \end{aligned} \tag{7}$$

where the first line talks about two cases of Inequality (6): (1) $T/2 \le T - \tilde{O}(\frac{md^{1/2}T^{1/4}}{n^{1/2}\epsilon^{1/2}}) \le 3m\texttt{Err}(h^*)+\tilde{O}(\sqrt{m})$ if $\tilde{O}(\frac{md^{1/2}T^{1/4}}{n^{1/2}\epsilon^{1/2}}) \le T/2$, and (2) $T^{3/4} \le 2\tilde{O}(\frac{md^{1/2}T^{1/4}}{n^{1/2}\epsilon^{1/2}})$ if $\tilde{O}(\frac{md^{1/2}T^{1/4}}{n^{1/2}\epsilon^{1/2}}) > T/2$; The second line is due to the indicator function is always ≤ 1.

Similar to the realizable case, now we apply a reduction to ERM. By the Hoeffding's style uniform convergence bound (implied by Lemma 36)

$$\begin{aligned} \texttt{Err}(\hat{h}^S) &\le \widehat{\texttt{Err}}(\hat{h}^S) + \tilde{O}\Big(\sqrt{\frac{d}{m}}\Big) \\ &\le \widehat{\texttt{Err}}(\hat{h}^{\texttt{priv}}) + \widehat{\texttt{Dis}}(\hat{h}^{\texttt{priv}}, \hat{h}^S) + \tilde{O}\Big(\sqrt{\frac{d}{m}}\Big) \\ &\le \widehat{\texttt{Err}}(\hat{h}^{\texttt{priv}}) + \widehat{\texttt{Dis}}(\hat{h}^{\texttt{priv}}, \hat{h}_1) + \tilde{O}\Big(\sqrt{\frac{d}{m}}\Big) \\ &\le 2\widehat{\texttt{Err}}(\hat{h}^{\texttt{priv}}) + \widehat{\texttt{Err}}(\hat{h}_1) + \tilde{O}\Big(\sqrt{\frac{d}{m}}\Big) \\ &\le \frac{2T}{m} + \texttt{Err}(h^*) + \tilde{O}\Big(\sqrt{\frac{d}{m}}\Big) \\ &\le 13\texttt{Err}(h^*) + \tilde{O}\Big(\frac{m^{1/3}d^{2/3}}{n^{2/3}\epsilon^{2/3}} + \sqrt{\frac{d}{m}}\Big). \end{aligned}$$

where the second and fourth lines use the triangular inequality of $0-1$ error, the third line uses the fact that $\hat{h}^S$ is the empirical risk minimizer of the student learning problem with labels $\hat{h}^{\texttt{priv}}$ and the fact that $h_1 \in \mathcal{H}$. The second last line follows from the fact that in those stable queries $\hat{h}^{\texttt{priv}}(x_j)$ outputs y_j, and a standard agnostic learning bound. Finally, in the last line, we obtain the stated result by substituting the upper bound of T from (7). ■

The results stated in Table 1 are obtained by minimizing the bound by choosing a random subset of data points to privately release labels.

A.2 Learning bound for PATE with Gaussian Mechanism

In this subsection, we provide a theoretical analysis of the version of PATE from Papernot et al. (2017, 2018) that uses Gaussian mechanism to release the aggregated teacher labels. We will focus on the setting assuming τ-TNC. Though this result is not our main contribution, we note that standard PATE is a practical algorithm and this is the first learning-theoretic guarantees of PATE.

Theorem 28 (Utility guarantee of Algorithm 1) *Assume the data distribution $\mathcal{D}$ and the hypothesis class $\mathcal{H}$ obey the Tsybakov noise condition with parameter τ, then with probability at least $1-\gamma$, there exists universal constant C such that the output $\hat{h}_S$ of Algorithm 1 with parameter K satisfying*

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

obeys that

$$\texttt{Err}(\hat{h}^S) - \texttt{Err}(h^*) \le \tilde{O}\Big(\frac{d}{m} + \Big(\frac{d\sqrt{m}}{n\epsilon}\Big)^{\frac{\tau}{2-\tau}}\Big).$$

Specifically, in the realizable setting, then it follows that

$$\texttt{Err}(\hat{h}^S) - \texttt{Err}(h^*) \le \tilde{O}\Big(\frac{d}{m} + \frac{d\sqrt{m}}{n\epsilon}\Big).$$

Proof By the triangular inequality of the $0-1$ error,

$$\begin{aligned}
\texttt{Err}(\hat{h}^S) - \texttt{Err}(h^*) &\le \texttt{Dis}(\hat{h}^S, h^*)\\
&\le \texttt{Dis}(\hat{h}^S, \tilde{h}^{\texttt{priv}}) + \texttt{Dis}(\tilde{h}^{\texttt{priv}}, h^*)\\
&\le 2\texttt{Dis}(\tilde{h}^{\texttt{priv}}, h^*) + 2\sqrt{\frac{(d+\log(4/\gamma))\texttt{Dis}(\tilde{h}^{\texttt{priv}}, h^*)}{m}} + \frac{4(d+\log(4/\gamma))}{m}\\
&\le 4\texttt{Dis}(\tilde{h}^{\texttt{priv}}, h^*) + \tilde{O}\Big(\frac{d}{m}\Big) \qquad (8)
\end{aligned}$$

where the third line follows from the learning bound (Lemma 37) with $\tilde{h}^{\texttt{priv}}$ being the labeling function for the student dataset. The last line is due to $a + 2\sqrt{ab} + b \le 2a + 2b$ for non-negative a, b.

The remaining problem would be finding the upper bound of $\texttt{Dis}(\tilde{h}^{\texttt{priv}}, h^*)$. First by Lemma 13, with probability at least $1-\gamma/2$, $\forall k \in [K]$ we have

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

Next, conditioning on the teachers, $\tilde{h}^{\texttt{priv}}$ is independent for each input and well-defined for all input. Let $Z \sim \mathcal{N}(0, \sigma^2)$. By Gaussian-tail bound and Markov's inequality,

$$\begin{aligned}
&\texttt{Dis}(\tilde{h}^{\texttt{priv}}, h^*)\\
&\leq \mathbb{P}\Big[|Z| \leq \sigma\sqrt{2\log\big(\tfrac{2}{\beta}\big)}\Big]\mathbb{P}\Big[\sum_{k=1}^{K} \mathbb{1}(\hat{h}_k(x) \neq h^*(x)) \geq \frac{K}{2} - |Z| \Big| |Z| \leq \sigma\sqrt{2\log\big(\tfrac{2}{\beta}\big)}\Big]\\
&\quad + \mathbb{P}\Big[|Z| > \sigma\sqrt{2\log\big(\tfrac{2}{\beta}\big)}\Big]\\
&\leq \frac{1}{K/2 - \sigma\sqrt{2\log(2/\beta)}} \sum_{k=1}^{K} \mathbb{E}[\mathbb{1}(\hat{h}_k(x) \neq h^*(x))] + \beta\\
&\leq \frac{3}{K}\sum_{k=1}^{K} \texttt{Dis}(\hat{h}_k, h^*) + \frac{1}{n}\\
&\lesssim \eta^{\frac{2}{2-\tau}}\Big(\frac{dK\log(n/d) + \log(2K/\gamma)}{n}\Big)^{\frac{\tau}{2-\tau}}.
\end{aligned}$$

In the last line, we choose $\beta = 1/n$ and applied the assumption that $K \geq 6\sigma\sqrt{2\log(2n)}$.

Note that our choice of σ satisfies that

$$\sqrt{\frac{2m\log(1/\delta)}{\sigma^2}} + \frac{m}{2\sigma^2} = \epsilon.$$

Solve the equation and we find that

$$\sigma = \frac{\sqrt{2m\log(1/\delta)} + \sqrt{2m\log(1/\delta) + 2\epsilon m}}{2\epsilon}.$$

Therefore, the choice of K is

$$K = \frac{6\sqrt{\log(2n)}(\sqrt{m\log(1/\delta)} + \sqrt{m\log(1/\delta) + \epsilon m})}{\epsilon} = \tilde{O}\Big(\frac{\sqrt{m}}{\epsilon}\Big),$$

where ϵ is assumed to be small. Put everything together, and the excess risk bound is

$$\texttt{Err}(\hat{h}^S) - \texttt{Err}(h^*) \leq \tilde{O}\bigg(\frac{d}{m} + \Big(\frac{d\sqrt{m}}{n\epsilon}\Big)^{\frac{\tau}{2-\tau}}\bigg).$$

■

Remark 29 *When m is sufficient large ($\frac{d}{m} < \big(\frac{d\sqrt{m}}{n\epsilon}\big)^{\frac{\tau}{2-\tau}}$), it suffices to use a subset of randomly chosen data points to optimize the bound and we obtain an excess risk bound of $\tilde{O}\Big(\big(\frac{d^{3/2}}{n\epsilon}\big)^{\frac{2\tau}{4-\tau}}\Big)$. When $\tau = 1$, this yields the $\frac{d}{(n\epsilon)^{2/3}}$ rate that matches Bassily et al. (2018b)'s analysis of SVT-based PATE. To avoid any confusions, Gaussian mechanism-based PATE is still theoretically inferior comparing to SVT-based PATE as we established in Theorem 10.*

Appendix B. Deferred Proofs of Our Results in Main Paper

In this section, we present full proofs of our results shown in the main paper.

Proposition 30 (Restatement of Proposition 18) *Assume the learning problem with n/K i.i.d. data points satisfies (ν, ξ)-approximate high-margin condition. Let Algorithm 2 be instantiated with parameters*

$$T \geq \nu m + \sqrt{2\nu m \log\left(\frac{3}{\gamma}\right)} + \frac{2}{3}\log\left(\frac{3}{\gamma}\right)$$

$$K \geq \max\left\{\frac{2\log(3m/\gamma)}{\xi^2}, \frac{3\lambda\big(\log(4m/\delta) + \log(3m/\gamma)\big)}{\xi}\right\},^{7}$$

then with high probability (over the randomness of the n i.i.d. samples of the private dataset, m i.i.d. samples of the public dataset, and that of the randomized algorithm), Algorithm 2 finishes all m rounds and the output is the same as $h^{\mathsf{agg}}_{\infty}(x_i)$ for all but T of the $i \in [m]$.

Proof By the Bernstein's inequality, with probability $\geq 1-\gamma_2$ over the i.i.d. samples of the public data, the number of queries $j \in [m]$ with $\Delta_{n/K}(x_j) \leq \xi$ is smaller than $\nu m + \sqrt{2\nu m \log(1/\gamma_2)} + \frac{2}{3}\log(1/\gamma_2)$. T is an upper bound of the above quantity if we choose $\gamma_2 = \gamma/3$.

Conditioning on the above event, by Hoeffding's inequality and a union bound, with probability $\geq 1-\gamma_3$ over the i.i.d. samples of the private data (hence the K i.i.d. teacher classifiers), for all $m-T$ queries with $\Delta_{n/K}(x_i)$ larger than ξ, the realized margin (defined in (1)) obeys that

$$\begin{aligned}\widehat{\Delta}(x_j) &\geq \mathbb{E}[\widehat{\Delta}(x_j)|x_j] - \sqrt{2K\log\left(\frac{m}{\gamma_3}\right)}\\ &= 2K\Delta_{n/K}(x_i) - \sqrt{2K\log\left(\frac{m}{\gamma_3}\right)}\\ &\geq 2K\xi - \sqrt{2K\log\left(\frac{m}{\gamma_3}\right)}.\end{aligned}$$

It remains to check that under our choice of T, K, $\widehat{\mathrm{dist}}_j > \hat{w}$ for all $j \in [m]$ except the (up to) T exceptions.

By the tail of Laplace distribution and a union bound, with probability $\geq 1-\gamma_1$, all m Laplace random variables that perturb the distance to stability $\widehat{\mathrm{dist}}_j$ in Algorithm 7 is larger than $-2\lambda\log((m+T)/(2\gamma_1))$ and all T Laplace random variables that perturb the threshold w is smaller than $\lambda\log((m+T)/(2\gamma_1))$, where λ is chosen according to Algorithm 2. We simplify the above bound by using $T < m$.

It suffices that K is chosen such that

$$2K\xi - \sqrt{2K\log\left(\frac{m}{\gamma_3}\right)} - 2\lambda\log\left(\frac{m}{\gamma_1}\right) > w + \lambda\log\left(\frac{m}{\gamma_1}\right).$$

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

Substitute Algorithm 2's choice $w = 3\lambda \log(2(m+T)/\delta) \le 3\lambda \log(4m/\delta)$. Assume $K \ge 2\log(m/\gamma_3)/\xi^2$, we have $2K\xi - \sqrt{2K\log(m/\gamma_3)} \ge K\xi$, thus it suffices that further $K\xi > 3\lambda\big(\log(4m/\delta) + \log(m/\gamma_1)\big)$.

The proof is complete by taking $\gamma_2 = \gamma_3 = \gamma/3$ and take union bound over all high probability events described above. ■

Theorem 31 (Restatement of Theorem 19) *Assume the learning problem with n/K i.i.d. data points satisfies (ν, ξ)-approximate high-margin condition and let K, T be chosen according to Proposition 18, furthermore assume that the privacy parameter of choice $\epsilon \le \log(2/\delta)$, then the output classifier $\hat{h}^S$ of Algorithm 3 in the agnostic setting satisfies that with probability $\ge 1 - 2\gamma$,*

$$\mathtt{Err}(\hat{h}^S) - \mathtt{Err}(h^{\mathsf{agg}}_\infty) \le \min_{h\in\mathcal{H}} \mathtt{Dis}(h, h^{\mathsf{agg}}_\infty) + \frac{2T}{m} + \tilde{O}\Big(\sqrt{\frac{d}{m}}\Big) \le \min_{h\in\mathcal{H}} \mathtt{Dis}(h, h^{\mathsf{agg}}_\infty) + 2\nu + \tilde{O}\Big(\sqrt{\frac{d}{m}}\Big).$$

Proof We follow a similar argument as in the proof of Theorem 10, but replace h^* with h^{agg}_∞. Define $\tilde{h} = \mathrm{argmin}_{h\in\mathcal{H}} \widehat{\mathtt{Dis}}(h, h^{\mathsf{agg}}_\infty)$. By the triangular inequality of the $0-1$ error and Lemma 36 in Appendix C,

$$\mathtt{Err}(\hat{h}^S) - \mathtt{Err}(h^{\mathsf{agg}}_\infty) \le \mathtt{Dis}(\hat{h}^S, h^{\mathsf{agg}}_\infty) \le \widehat{\mathtt{Dis}}(\hat{h}^S, h^{\mathsf{agg}}_\infty) + \tilde{O}\Big(\sqrt{\frac{d}{m}}\Big). \tag{9}$$

By the triangular inequality, we have $\widehat{\mathtt{Dis}}(\hat{h}^S, h^{\mathsf{agg}}_\infty) \le \widehat{\mathtt{Dis}}(\hat{h}^S, \hat{h}^{\mathtt{priv}}) + \widehat{\mathtt{Dis}}(\hat{h}^{\mathtt{priv}}, h^{\mathsf{agg}}_\infty)$, therefore,

$$\begin{aligned}
(9) &\le \widehat{\mathtt{Dis}}(\hat{h}^S, \hat{h}^{\mathtt{priv}}) + \widehat{\mathtt{Dis}}(\hat{h}^{\mathtt{priv}}, h^{\mathsf{agg}}_\infty) + \tilde{O}\Big(\sqrt{\frac{d}{m}}\Big)\\
&\le \widehat{\mathtt{Dis}}(\tilde{h}, \hat{h}^{\mathtt{priv}}) + \widehat{\mathtt{Dis}}(\hat{h}^{\mathtt{priv}}, h^{\mathsf{agg}}_\infty) + \tilde{O}\Big(\sqrt{\frac{d}{m}}\Big)\\
&\le \widehat{\mathtt{Dis}}(\tilde{h}, h^{\mathsf{agg}}_\infty) + 2\widehat{\mathtt{Dis}}(\hat{h}^{\mathtt{priv}}, h^{\mathsf{agg}}_\infty) + \tilde{O}\Big(\sqrt{\frac{d}{m}}\Big)\\
&\le \min_{h\in\mathcal{H}} \mathtt{Dis}(h, h^{\mathsf{agg}}_\infty) + 2\widehat{\mathtt{Dis}}(\hat{h}^{\mathtt{priv}}, h^{\mathsf{agg}}_\infty) + \tilde{O}\Big(\sqrt{\frac{d}{m}}\Big).
\end{aligned}$$

In the second line, we applied the fact that $\hat{h}^S = \mathrm{argmin}_{h\in\mathcal{H}} \widehat{\mathtt{Dis}}(h, \hat{h}^{\mathtt{priv}})$; in the third line, we applied triangular inequality again and the last line is true because $\tilde{h} = \mathrm{argmin}_{h\in\mathcal{H}} \widehat{\mathtt{Dis}}(h, h^{\mathsf{agg}}_\infty)$.

Recall that T is the unstable cutoff in Algorithm 3. The proof completes by invoking Proposition 18 which implies that $\widehat{\mathtt{Dis}}(\hat{h}^{\mathtt{priv}}, h^{\mathsf{agg}}_\infty) \le T/m$ with high probability.

■

Theorem 32 (Restatement of Theorem 20) *Under the same assumption of Theorem 19, 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 that with probability $\geq 1-2\gamma$,*

$$\texttt{Err}(\hat{h}^S) - \texttt{Err}(h_\infty^{\texttt{agg}}) \leq \frac{4T}{m} + \frac{5(Kd+\log(4/\gamma))}{m} = \tilde{O}\Big(\nu + \frac{\log(4/\gamma)}{m} + \frac{d\sqrt{\nu}}{\xi\sqrt{m}}\Big).$$

Proof Define $\hat{h}^S = \mathrm{argmin}_{h\in \mathrm{Vote}_K(\mathcal{H})} \widehat{\texttt{Dis}}(h, \hat{h}^{\texttt{priv}})$ and $\tilde{h} = \mathrm{argmin}_{h\in \mathrm{Vote}_K(\mathcal{H})} \widehat{\texttt{Dis}}(h, h_\infty^{\texttt{agg}})$. By the triangular inequality of the $0-1$ error,

$$\begin{aligned}\texttt{Err}(\hat{h}^S) - \texttt{Err}(h_\infty^{\texttt{agg}}) &\leq \texttt{Dis}(\hat{h}^S, h_\infty^{\texttt{agg}}) \\ &\leq \widehat{\texttt{Dis}}(\hat{h}^S, h_\infty^{\texttt{agg}}) + 2\sqrt{\frac{(Kd+\log(4/\gamma))\widehat{\texttt{Dis}}(\hat{h}^S, h_\infty^{\texttt{agg}})}{m}} + \frac{4(Kd+\log(4/\gamma))}{m} \\ &\leq 2\widehat{\texttt{Dis}}(\hat{h}^S, h_\infty^{\texttt{agg}}) + \frac{5(Kd+\log(4/\gamma))}{m}, \end{aligned} \tag{10}$$

where the second line follows from the first statement of Lemma 36 in Appendix C with $z = h_\infty^{\texttt{agg}}(x)$ and the third line is due to $a + 2\sqrt{ab} + b \leq 2a + 2b$ for non-negative a, b.

By the triangular inequality, we have $\widehat{\texttt{Dis}}(\hat{h}^S, h_\infty^{\texttt{agg}}) \leq \widehat{\texttt{Dis}}(\hat{h}^S, \hat{h}^{\texttt{priv}}) + \widehat{\texttt{Dis}}(\hat{h}^{\texttt{priv}}, h_\infty^{\texttt{agg}})$, therefore,

$$\begin{aligned}(10) &\leq 2\widehat{\texttt{Dis}}(\hat{h}^S, \hat{h}^{\texttt{priv}}) + 2\widehat{\texttt{Dis}}(\hat{h}^{\texttt{priv}}, h_\infty^{\texttt{agg}}) + \frac{5(Kd+\log(4/\gamma))}{m} \\ &\leq 2\widehat{\texttt{Dis}}(\tilde{h}, \hat{h}^{\texttt{priv}}) + 2\widehat{\texttt{Dis}}(\hat{h}^{\texttt{priv}}, h_\infty^{\texttt{agg}}) + \frac{5(Kd+\log(4/\gamma))}{m} \\ &\leq 2\widehat{\texttt{Dis}}(\tilde{h}, h_\infty^{\texttt{agg}}) + 4\widehat{\texttt{Dis}}(\hat{h}^{\texttt{priv}}, h_\infty^{\texttt{agg}}) + \frac{5(Kd+\log(4/\gamma))}{m} \\ &\leq 4\widehat{\texttt{Dis}}(\hat{h}^{\texttt{priv}}, h_\infty^{\texttt{agg}}) + \frac{5(Kd+\log(4/\gamma))}{m}.\end{aligned}$$

In the second line, we applied the fact that $\hat{h}^S = \mathrm{argmin}_{h\in \mathrm{Vote}_K(\mathcal{H})} \widehat{\texttt{Dis}}(h, \hat{h}^{\texttt{priv}})$; in the third line, we applied triangular inequality again and the last line is true because $\widehat{\texttt{Dis}}(\tilde{h}, h_\infty^{\texttt{agg}}) = 0$ since $\tilde{h}$ is the minimizer and that $h_\infty^{\texttt{agg}} \in \mathrm{Vote}_K(\mathcal{H})$.

Recall that T is the unstable cutoff in Algorithm 3. The proof completes by using that $\widehat{\texttt{Dis}}(\hat{h}^{\texttt{priv}}, h_\infty^{\texttt{agg}}) \leq T/m$ with probability $1-\gamma$ according to Proposition 18 and substitute the choices of T and K accordingly. ■

Lemma 33 *If the disagreement-based agnostic active learning algorithm is given a stream of m unlabeled data points, then with probability at least $1-\gamma$, the algorithm returns a hypothesis h obeying that,*

$$\texttt{Err}(h) - \texttt{Err}(h^*) \lesssim \frac{d\log(\theta(d/m)) + \log(1/\gamma)}{m} + \sqrt{\frac{\texttt{Err}(h^*)(d\log(\theta(\texttt{Err}(h^*)) + \log(1/\gamma))}{m}}.$$

Proof From Lemma 3.1 of Hanneke (2014), we learn that for any hypothesis h survive in version space V must satisfy

$$\texttt{Err}(h) - \texttt{Err}(h^*) \leq 2U(m, \gamma).$$

Then by the definition of $U(m,\gamma)$ shown in Algorithm 4, we have

$$\mathtt{Err}(h) - \mathtt{Err}(h^*) \lesssim \frac{d\log(\theta(d/m)) + \log(1/\gamma)}{m} + \sqrt{\frac{\mathtt{Err}(h^*)(d\log(\theta(\mathtt{Err}(h^*)) + \log(1/\gamma))}{m}}.$$

■

Theorem 34 (Restatement of Theorem 22) *With probability at least* $1-\gamma$*, there exists universal constants* C_1, C_2 *such that for all*

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

the output $\hat{h}^S$ *of Algorithm 5 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

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

Specifically, when we choose

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

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

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

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

Proof **Step 1: Teachers are good.** By Lemma 13, with probability at least $1-\gamma/2$, $\forall k \in [K]$ we have

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

Step 2: PATE is just as good. Let $\tilde{h}^{\mathtt{priv}}$ be a randomized classifier from Line 4 of Algorithm 1. Conditioning on the teachers, this classifier is independent for each input and well-defined for all input. Note that $\hat{h}^{\texttt{priv}}$ that uses Algorithm 2 do not have these properties. Let $Z \sim \mathcal{N}(0, \sigma^2)$. By Gaussian-tail bound and Markov's inequality,

$$\begin{aligned}
&\texttt{Dis}(\tilde{h}^{\texttt{priv}}, h^*)\\
&\leq \mathbb{P}\Big[|Z| \leq \sigma\sqrt{2\log\big(\frac{2}{\beta}\big)}\Big]\mathbb{P}\Big[\sum_{k=1}^{K}\mathbb{1}(\hat{h}_k(x) \neq h^*(x)) \geq \frac{K}{2} - |Z|\Big||Z| \leq \sigma\sqrt{2\log\big(\frac{2}{\beta}\big)}\Big]\\
&\quad + \mathbb{P}\Big[|Z| > \sigma\sqrt{2\log\big(\frac{2}{\beta}\big)}\Big]\\
&\leq \frac{1}{K/2 - \sigma\sqrt{2\log(2/\beta)}}\sum_{k=1}^{K}\mathbb{E}[\mathbb{1}(\hat{h}_k(x) \neq h^*(x))] + \beta\\
&\leq \frac{3}{K}\sum_{k=1}^{K}\texttt{Dis}(\hat{h}_k, h^*) + \frac{1}{n}\\
&\lesssim \eta^{\frac{2}{2-\tau}}\Big(\frac{dK\log(n/d) + \log(2K/\gamma)}{n}\Big)^{\frac{\tau}{2-\tau}}.
\end{aligned}$$

In the last line, we choose $\beta = 1/n$ and applied the assumption that $K \geq 6\sigma\sqrt{2\log(2n)}$.

Step 3: Oracle reduction to active learning bounds. Note that $\tilde{h}^{\texttt{priv}}$ is the labeling function in the student learning problem. So the above implies that the student learning problem is close to realizable:

$$\min_{h\in\mathcal{H}}\texttt{Dis}(\tilde{h}^{\texttt{priv}}, h) \leq \texttt{Dis}(\tilde{h}^{\texttt{priv}}, h^*) \lesssim \eta^{\frac{2}{2-\tau}}\Big(\frac{dK\log(n/d) + \log(2K/\gamma)}{n}\Big)^{\frac{\tau}{2-\tau}}.$$

By the above, and the agnostic active learning bounds in Lemma 39, to achieve an excess risk bound of $\alpha \geq \texttt{Dis}(\tilde{h}^{\texttt{priv}}, h^*) := \texttt{Err}^*$ in the student learning problem with probability at least $1-\gamma/2$, with unbounded m, it suffices to choose ℓ to be

$$\begin{aligned}
&C\theta(\texttt{Err}^* + \alpha)\Big(\frac{(\texttt{Err}^*)^2}{\alpha^2} + \log\big(\frac{1}{\alpha}\big)\Big)\Big(d\log(\theta(\texttt{Err}^* + \alpha)) + \log\big(\frac{\log(1/\alpha)}{\gamma}\big)\Big)\\
&\leq C\theta(\alpha)(1 + \log(1/\alpha))\Big(d\log(\theta(\alpha)) + \log\big(\frac{\log(1/\alpha)}{\gamma}\big)\Big).
\end{aligned}$$

This implies an error bound of

$$\begin{aligned}
\texttt{Dis}(\hat{h}_S, \tilde{h}^{\texttt{priv}}) &\leq \min_{h\in\mathcal{H}}\texttt{Dis}(\tilde{h}^{\texttt{priv}}, h) + \alpha\\
&\leq \texttt{Dis}(\tilde{h}^{\texttt{priv}}, h^*) + \alpha \leq 2\alpha.
\end{aligned}$$

When m is small, we might not have enough data points to obtain $\alpha = O(\texttt{Dis}(\tilde{h}^{\texttt{priv}}, h^*))$ in this case the error is dominated by our bounds in Lemma 33, which says that we can take

$$\alpha = C\max\Big\{\texttt{Err}^*, \frac{d\log(m/d) + \log(2/\gamma)}{m}\Big\}.$$

Step 4 Putting everything together.

$$\begin{aligned}\texttt{Err}(\hat{h}^S) - \texttt{Err}(h^*) &\leq \texttt{Dis}(\hat{h}^S, \tilde{h}^{\texttt{priv}}) + \texttt{Dis}(\tilde{h}^{\texttt{priv}}, h^*) \\ &\lesssim \texttt{Dis}(\tilde{h}^{\texttt{priv}}, h^*) + \alpha \\ &\lesssim \eta^{\frac{2}{2-\tau}} \Big(\frac{dK \log(n/d) + \log(2K/\gamma)}{n}\Big)^{\frac{\tau}{2-\tau}} + \alpha.\end{aligned}$$

The proof is complete by substituting our choice of $K = 6\sigma\sqrt{2\log(2n)}$, and furthermore by the standard privacy calibration of the Gaussian mechanism, our choice of σ satisfies that

$$\sqrt{\frac{2\ell\log(1/\delta)}{\sigma^2}} + \frac{\ell}{2\sigma^2} = \epsilon.$$

following the specification of Algorithm 1. Solve the equation and we find that

$$\sigma = \frac{\sqrt{2\ell\log(1/\delta)} + \sqrt{2\ell\log(1/\delta) + 2\epsilon\ell}}{2\epsilon},$$

where ϵ is assumed to be small. ■

Appendix C. Technical Lemmas from Statistical learning Theory

In this section, we cite a few results from statistical learning theory that we will use as part of our analysis.

Lemma 35 (Pointwise convergence (Bousquet et al., 2004)) *Let (x, z) be drawn from any distribution $\mathcal{D}$ supported on $\mathcal{X} \times \mathcal{Y}$. Let $\texttt{Dis}$ and $\widehat{\texttt{Dis}}$ be the expected and empirical disagreement evaluated on n i.i.d. samples from $\mathcal{D}$. For each fixed $h \in \mathcal{H}$, the following generalization error bound holds with probability $1 - \gamma$,*

$$\texttt{Dis}(h, z) \leq \widehat{\texttt{Dis}}(h, z) + \sqrt{\frac{2\texttt{Dis}(h, z)\log(1/\gamma)}{n}} + \frac{2\log(1/\gamma)}{3n},$$

where n is the number of data points.

This is a standard application of the Bernstein's inequality.

Lemma 36 (Uniform convergence (Bousquet et al., 2004)) *Under the same conditions of Lemma 35, and in addition assume that d is the VC-dimension of $\mathcal{H}$, Then with probability at least $1 - \gamma$, $\forall h \in \mathcal{H}$ simultaneously,*

$$\texttt{Dis}(h, z) - \widehat{\texttt{Dis}}(h, z) \leq 2\sqrt{\frac{(d + \log(4/\gamma))\widehat{\texttt{Dis}}(h, z)}{n}} + \frac{4(d + \log(4/\gamma))}{n}.$$

and

$$\texttt{Dis}(h, z) - \widehat{\texttt{Dis}}(h, z) \leq 2\sqrt{\frac{(d + \log(4/\gamma))\texttt{Dis}(h, z)}{n}} + \frac{4(d + \log(4/\gamma))}{n}.$$

The above lemma is simply the uniform Bernstein's inequality over a hypothesis class with VC-dimension d. We will be taking z to be h^* in the cases when we work with noise conditions and $h_\infty^{\mathsf{agg}}(x)$ in the agnostic case.

Lemma 37 (Learning bound (Bousquet et al., 2004)) *Let d be the VC-dimension of $\mathcal{H}$, the excess risk is bounded with probability $1-\gamma$,*

$$\mathtt{Err}(\hat{h}) \leq \mathtt{Err}(h^*) + 2\sqrt{\mathtt{Err}(h^*)\frac{d\log(n)+\log(4/\gamma)}{n}} + 4\frac{d\log(n)+\log(4/\gamma)}{n},$$

where n is the number of data points we sample.

Lemma 38 (Passive learning bound under TNC (Lemma 3.4 of Hanneke (2014))) *Let d be the VC-dimension of the class $\mathcal{H}$. Assume Tsybakov noise condition with parameters τ, the excess risk is bounded with probability $1-\gamma$,*

$$\mathtt{Err}(\hat{h}) - \mathtt{Err}(h^*) \lesssim \left(\frac{1}{n}\Big(d\log\big(\frac{n}{d}\big) + \log\big(\frac{1}{\gamma}\big)\Big)\right)^{\frac{1}{2-\tau}},$$

where n is the number of data points.

Finally, we need a result from active learning.

Lemma 39 (Agnostic active learning bound (Theorem 5.4 of Hanneke (2014))) *Let $\mathcal{H}$ be a class with VC-dimension d. With probability at least $1-\gamma$, there is a universal constant C, such that the agnostic active learning algorithm (see Algorithm 4) outputs a classifier with an access risk of α with*

$$C\theta(\mathtt{Err}^* + \alpha)\left(\frac{(\mathtt{Err}^*)^2}{\alpha^2} + \log\Big(\frac{1}{\alpha}\Big)\right)\left(d\log(\theta(\mathtt{Err}^*+\alpha)) + \log\Big(\frac{\log(1/\alpha)}{\gamma}\Big)\right),$$

where $\mathtt{Err}^* = \operatorname{argmin}_{h\in\mathcal{H}} \mathtt{Err}(h)$.

Appendix D. Additional Information about Differential Privacy

In this section, we cite a few results from differential privacy that we will use as part of our analysis.

Lemma 40 (Post-processing (Dwork et al., 2006)) *If a randomized algorithm $\mathcal{M} : \mathcal{Z}^* \to \mathcal{R}$ is (ϵ,δ)-DP, then for any function $f : \mathcal{R} \to \mathcal{R}'$, $f \circ \mathcal{M}$ is also (ϵ,δ)-DP.*

Definition 41 (Global sensitivity (Dwork and Roth, 2014)) *A function $f : \mathcal{Z}^* \to \mathcal{R}$ has global sensitivity ϑ if*

$$\max_{|D-D'|=1} \|f(D) - f(D')\|_1 = \vartheta.$$

Lemma 42 (Laplace mechanism (Dwork et al., 2006)) *If a function $f : \mathcal{Z}^n \to \mathcal{R}^p$ has global sensitivity ϑ, then the randomized algorithm $\mathcal{M}$, which on input D outputs $f(D)+b$, where $b \sim \mathtt{Lap}(\vartheta/\epsilon)^p$, satisfies ϵ-DP. The $\mathtt{Lap}(\lambda)^p$ denotes a vector of p i.i.d. samples from the Laplace distribution $\mathtt{Lap}(\lambda)$.*

Definition 43 (ℓ_2-sensitivity (Dwork and Roth, 2014)) *A function $f : \mathcal{Z} \to \mathcal{R}$ has ℓ_2 sensitivity ϑ_2 if*

$$\max_{|D-D'|=1} \|f(D) - f(D')\|_2 = \vartheta_2.$$

Lemma 44 (Gaussian mechanism (Dwork and Roth, 2014)) *If a function $f : \mathcal{Z}^n \to \mathcal{R}^p$ has ℓ_2-sensitivity ϑ_2, then the randomized algorithm $\mathcal{M}$, which on input D outputs $f(D) + b$, where $b \sim \mathcal{N}(0, \sigma^2)^p$, satisfies (ϵ, δ)-DP, where $\sigma \geq c\vartheta_2/\epsilon$ and $c^2 > 2\log(1.25/\delta)$. The $\mathcal{N}(0, \sigma^2)^p$ denotes a vector of p i.i.d. samples from the Gaussian distribution $\mathcal{N}(0, \sigma^2)$.*

Algorithm 6 Sparse Vector Technique (Dwork et al., 2010; Dwork and Roth, 2014)

Input: Dataset D, query set $\mathcal{Q} = \{q_1, ..., q_m\}$, privacy parameters $\epsilon, \delta > 0$, unstable query cutoff T, threshold w.

```
1: c ← 0, λ ← √(32T log(1/δ)/ε), ŵ ← w + Lap(λ).
2: for q ∈ Q and c ≤ T do
3:    q̂ ← q + Lap(2λ).
4:    if q̂ > ŵ then
5:       Output ⊤.
6:    else
7:       Output ⊥. ŵ ← w + 1, c ← c + 1.
8:    end if
9: end for
```

Lemma 45 (Privacy guarantee of Algorithm 6 (Dwork and Roth, 2014)) *Algorithm 6 is (ϵ, δ)-DP.*

Lemma 46 (Utility guarantee of Algorithm 6 (Dwork and Roth, 2014)) *For $\phi = \log(2mT/\beta)\sqrt{512T\log(1/\delta)}/\epsilon$, and any set of m queries, define the set $L(\phi) = \{i : q_i(D) \leq w + \phi\}$. If $|L(\phi)| \leq T$, then w.p. at least $1 - \beta : \forall i \notin L(\phi)$ Algorithm 6 outputs $\top$.*

Definition 47 (k-stability (Thakurta and Smith, 2013)) *A function $f : \mathcal{Z} \to \mathcal{R}$ is k stable on dataset D if adding or removing any k elements from D does not change the value of f, i.e., $f(D) = f(D')$ for all D' such that $|D - D'| \leq k$. We say f is stable on D if it is (at least) 1-stable on D, and unstable otherwise.*

Lemma 48 (Privacy guarantee of Algorithm 7 (Bassily et al., 2018b)) *If the threshold $\Gamma = \log(1/\delta)/\epsilon$, and the distance to instability function $\mathrm{dist}_f(D) = \mathrm{argmax}_k(f(D)$ is k-stable), then Algorithm 7 is (ϵ, δ)-DP.*

Lemma 49 (Utility guarantee of Algorithm 7 (Thakurta and Smith, 2013)) *If the threshold $\Gamma = \log(1/\delta)/\epsilon$, and the distance to instability function $\mathrm{dist}_f(D) = \mathrm{argmax}_k(f(D)$ is k-stable), and $f(D)$ is $((\log(1/\delta) + \log(1/\beta))/\epsilon)$-stable, then Algorithm 7 outputs $f(D)$ w.p. at least $1 - \beta$.*

Algorithm 7 Distance to Instability Framework (Thakurta and Smith, 2013)

Input: Dataset D, function $f : \mathcal{Z} \to \mathcal{R}$, distance to instability $\text{dist}_f : \mathcal{Z} \to \mathcal{R}$, thereshold Γ, privacy parameter $\epsilon > 0$.

1: $\widehat{\text{dist}} \leftarrow \widehat{\text{dist}}_f(D) + \texttt{Lap}(1/\epsilon)$.
2: **if** $\widehat{\text{dist}} > \Gamma$ **then**
3: Output $f(D)$.
4: **else**
5: Output $\perp$.
6: **end if**

Definition 50 (Definition 1.1 of (Bun and Steinke, 2016)) *$\mathcal{M}$ obeys (ξ, ρ)-zCDP if for two adjacent dataset D, D', for all $\phi \in (1, \infty)$, the Renyi-divergence of order ϕ below obeys that*

$$D_\phi(\mathcal{M}(D) \| \mathcal{M}(D')) \leq \xi + \rho\alpha.$$

When $\xi = 0$, we also call it ρ-zCDP (or simply ρ-CDP, since we are not considering other versions of CDPs in this paper).

The following two lemmas will be used in the privacy analysis of the SVT-based PATE.

Lemma 51 (Proposition 1.3 of (Bun and Steinke, 2016)) *If $\mathcal{M}$ obeys ρ-zCDP, then $\mathcal{M}$ is $(\rho + 2\sqrt{\rho \log(1/\delta)}, \delta)$-DP for any $\delta > 0$.*

Lemma 52 (Proposition 1.4 of (Bun and Steinke, 2016)) *If $\mathcal{M}$ obeys ϵ-DP, then $\mathcal{M}$ obeys $\frac{\epsilon^2}{2}$-CDP.*

Appendix E. Simulation with adult dataset

In this section, we empirically estimate the expected margin $\Delta(x) = |\mathbb{E}[\hat{h}_1(x)|x] - 0.5|$ and $|\mathbb{E}[\mathbb{I}(h(x) \neq h^*(x))|x] - 0.5|$ on the Adult dataset. Note that in $\Delta(x)$, we do not require the teachers to agree on y or h^* but measure the extent to which they agree with $\hat{h}^{\mathsf{agg}}$. In the latter one, we measure the degree of agreement between teachers and h^*.

The UCI Adult dataset is also known as "Census Income" dataset, which is used to predict whether an individual's annual income exceeds \$50,000. We partition the original training set as the private dataset and the testing set as the public dataset. To simulate the PATE setting, we train 250 logistic regression models on the private domain and use this ensemble to answer 2,000 queries from the public domain. Note that under this setting, the private domain contains 36,631 records and the public domain has 10,211 unlabelled records. We train h^* with the entire private dataset using the logistic regression model.

Results shown in Figure 3 demonstrate that even though we do not know the distribution of the data at all, the teacher ensemble agrees on the large majority of the examples. Moreover, when they agree, they agree on h^* in most cases and only in very rare cases when they agree on the wrong answers with high-margin. To say it differently, our assumption on the Tsybakov noise condition could be a good approximation to the real-life datasets. In Figure 4, we plot the correlations between $\Delta(x)$ and $\mathbb{E}[\mathbb{1}[h(x) \neq h^*(x)]|x] - 0.5$ over

200 queries. The x-axis is the cumulation of $\Delta(x)$ and the y-axis is the cumulation of $\mathbb{E}[\mathbb{1}[h(x) \neq h^*(x)]|x] - 0.5$. There is a nearly perfect linear line in the figure, which indicates they are highly correlated and majority voting tends to agree on h^* in most cases on the Adult dataset.

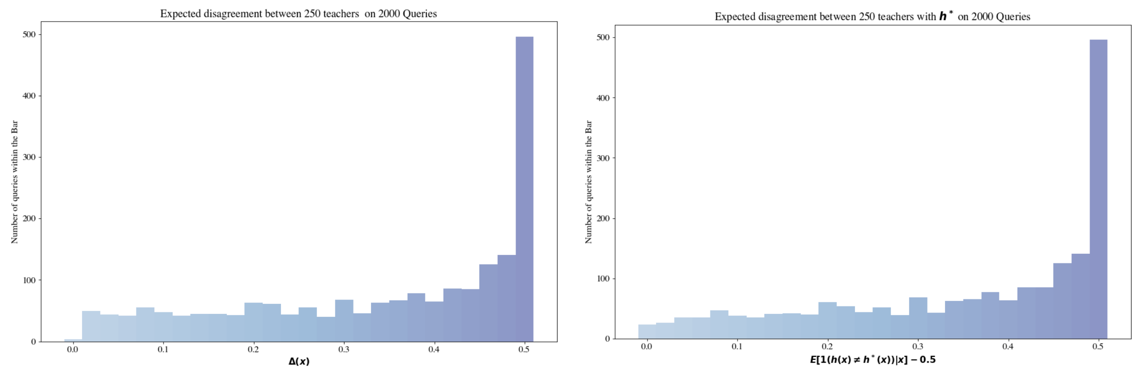


Figure 3: Empirical distribution of the margins on the Adult dataset.

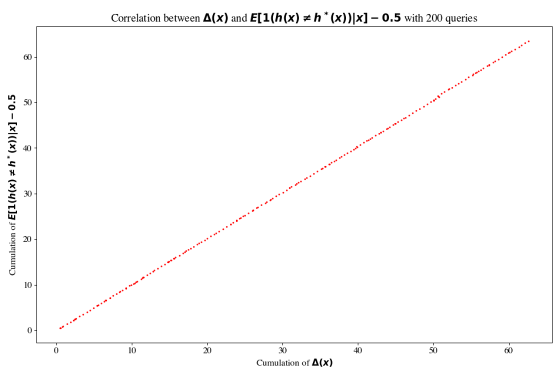


Figure 4: Correlations between margins on the Adult dataset.