A Related Work

The literature on differentially private machine learning is enormous and it is impossible for us to provide a comprehensive summary. Thus, our discussion focuses on only a few theoretical aspects.

A.1 Theory of Private Learning

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 "dimension" that measures the extent to which we can construct a specific discretization of the hypothesis space that works for "all distributions" on data. This dimension is often not possible 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).

A.2 Weaker Private Learning Models

This setting of private learning was relaxed in various ways to circumvent this 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 finally 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.

A.3 Private Learning with an Auxiliary Public Dataset

The problem of private learning with access to an additional public data was formally studied in (Alon et al., 2019), which 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. They also provide an upper bound that says $\tilde{O}(d/\alpha^2)$ private data and $\tilde{O}(d/\alpha)$ public data are sufficient to agnostically learn any classes with VC-dimension d to α-excess risk. The algorithm however uses an explicit α-net covering argument due to (Beimel et al., 2016), thus is not an efficient algorithm that is implementable. Our results are complementary in that we focuses on advancing the theoretical understanding of PATE towards more-refined data-dependent analysis, and to investigate the use of active learning. That said, Example 8 implies that PATE does not enjoy the same agnostic learning bounds.

A.4 Privacy-Preserving Prediction

There is another line of work (Dwork and Feldman, 2018) that focuses on the related problem of "privacy-preserving predictions" which does not release the learned model (which we do), but instead privately answer one 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 significantly worse rates. More recent work under this model (see e.g., Dagan and Feldman, 2020; Nandi and Bassily, 2020) relies on the explicit α-net construction due to (Beimel et al., 2016) (just like that in (Alon et al., 2019)) which cannot be implemented efficiently, instead we reduce to the learning bounds of empirical risk minimization (in the passive learning case) or active learning oracles. So if we believe supervised learning is practically easy, the algorithm can be implemented (and has been) in practice (Papernot et al., 2017, 2018).

A.5 Statistical Learning Theory

The Tsybakov noise condition (TNC) (Mammen and Tsybakov, 1999; Tsybakov, 2004) is a natural and well-established condition in generalization 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 the Tsybakov noise condition, 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. To the best of our knowledge, we are the first that formally study these quantities 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.

B Deferred Proofs of Some Technical Results

Theorem B.1 (Restatement of Theorem 5)**.** *Assume the data distribution $\mathcal{D}$ and the hypothesis class $\mathcal{H}$ obey the Tsybakov noise condition with parameter τ. Then Algorithm 3 with*

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

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

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

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

First we will formalize the discussion in the main paper to be a few lemmas, which will be used in the proof afterwards.

Lemma B.2. *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 $k = 1, 2, 3, ..., K$*

$$\mathrm{Dis}(\hat{h}_k, h^*) \le C\left(\frac{dK\log(n/d) + \log(K/\gamma)}{n} \right)^{\frac{\tau}{2-\tau}}.$$

Proof. By the equivalent definition of the Tsybakov noise condition and then the learning bounds under the Tsybakov noise condition (Lemma D.6),

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

□

Lemma B.3. *Under the condition of Lemma B.2, 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:*

$$\sum_{j=1}^{m} \mathbf{1}(\hat{h}_k(x_j) \neq h^*(x_j)) \le O\left(\max\left\{ m\mathrm{Dis}(\hat{h}_k, h^*), \log\left(\frac{K}{\gamma}\right) \right\} \right).$$

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 D.3)

$$\begin{aligned}\sum_{j=1}^{m} \mathbf{1}(\hat{h}_k(x_j) \neq h^*(x_j)) &\le O\left(m\mathrm{Dis}(\hat{h}_k, h^*) + \sqrt{m\mathrm{Dis}(\hat{h}_k, h^*)\log\left(\frac{K}{\gamma}\right)} + \log\left(\frac{K}{\gamma}\right) \right) \\ &\le O\left(\max\left\{ m\mathrm{Dis}(\hat{h}_k, h^*), \log\left(\frac{K}{\gamma}\right) \right\} \right).\end{aligned}$$

□

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 B.4. *Let Algorithm 3 be run with the number of teachers K and the cut-off parameter T chosen according to Theorem 5. Assume the conditions of Lemma B.2. Then with high probability ($\geq 1-\beta$ over the random coins of Algorithm 3 alone and conditioning on the high probability events of Lemma B.2 and Lemma B.3), Algorithm 3 finishes all m queries without exhausting the cut-off budget and that*

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

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 B.2 and B.3.

Proof. Denote the bound from Lemma B.3 by B. By the same Pigeon hole principle argument as in Lemma C.2 (but with y replaced by h^*), we have that the number of queries that have margin smaller than $K/6$ is at most $3B = O(\max\{m\text{Dis}(\hat{h}_k, h^*), \log(K/\gamma)\})$. 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.,

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

□

Now we are ready to put everything together and prove Theorem 5.

Proof of Theorem 5. Denote $\tilde{h} = \text{argmin}_{h\in\mathcal{H}} \widehat{\text{Dis}}(h, h^*)$ where $\widehat{\text{Dis}}$ is the empirical average of the disagreements over the data points that students have [3]. By the triangular inequality of the $0-1$ error,

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

where the second line follows from the uniform Bernstein's inequality — apply the first statement Lemma D.4 in the Appendix with $z = h^*(x)$ 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 $\widehat{\text{Dis}}(\hat{h}^S, h^*) \leq \widehat{\text{Dis}}(\hat{h}^S, \hat{h}^{\text{priv}}) + \widehat{\text{Dis}}(\hat{h}^{\text{priv}}, h^*)$, therefore

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

In the second line, we applied the fact that $\hat{h}^S$ is the minimizer of $\widehat{\text{Dis}}(h, \hat{h}^{\text{priv}})$; in the third line, we applied triangular inequality again and the last line is true because $\widehat{\text{Dis}}(\tilde{h}, h^*) = 0$ 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 B.4 which shows that the choice of T is appropriate such that $\widehat{\text{Dis}}(\hat{h}^{\text{priv}}, h^*) \leq T/m$ with high probability. □

[3] 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 $\mathcal{H}$.

In the light of the above analysis, it is clear that the improvement from our analysis under the Tsybakov noise condition are twofolds: (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.

Proposition B.5 (Restatement of Proposition 11)**.** *Assume the learning problem with n/K i.i.d. data points satisfies (ν, Δ)-approximate high-margin condition. Let Algorithm 2 be instantiated with parameter $T \geq \nu m + \sqrt{2\nu m \log(3/\gamma)} + \frac{2}{3}\log(3/\gamma)$ and $K \geq \max\{\frac{2\log(3m/\gamma)}{\Delta^2}, \frac{3\lambda\left(\log(4m/\delta)+\log(3m/\gamma)\right)}{\Delta}\}$[4]. 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 $h_\infty^{\mathrm{agg}}(x_i)$ for all but T of the $i \in [m]$.*

Proof. By the Bernstein's inequality, with probability $\geq 1-\gamma_2$ over the iid samples of the public data, the number of queries $j \in [m]$ with $\Delta_{n/k}(x_j) \leq \Delta$ 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 iid samples of the private data (hence the K iid teacher classifiers), for all $m-T$ queries with $\Delta_{n/k}(x_i)$ larger than Δ, the realized margin (defined in (2.3)) obeys that

$$\begin{aligned}\mathrm{margin}(x_j) \geq& \mathbf{E}[\mathrm{margin}(x_j)|x_j] - \sqrt{\frac{K\log(m/\gamma_3)}{2}}\\ =& 2K\Delta_{n/K}(x_i) - \sqrt{2K\log(m/\gamma_3)}\\ \geq& 2K\Delta - \sqrt{2K\log(m/\gamma_3)}\end{aligned}$$

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 6 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

$$2K\Delta - \sqrt{2K\log(m/\gamma_3)} - 2\lambda\log(m/\gamma_1) > w + \lambda\log(m/\gamma_1).$$

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

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

Theorem B.6 (Restatement of Theorem 12)**.** *Assume the learning problem with n/K i.i.d. data points satisfies (ν, Δ)-approximate high-margin condition and let K, T be chosen according to Proposition 11, furthermore assume that the privacy parameter of choice $\epsilon \leq \log(2/\delta)$, then the output classifier $\hat{h}^S$ of Algorithm 3 in the agnostic setting satisfies that with probability $\geq 1-2\gamma$*

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

Proof. We follow a similar argument as in the proof of Theorem 5, but replace h^* with h_∞^{agg}. Define $\tilde{h} = \mathrm{argmin}_{h\in\mathcal{H}} \widehat{\mathrm{Dis}}(h, h_\infty^{\mathrm{agg}})$. By the triangular inequality of the $0-1$ error and Lemma D.4 in the Appendix,

$$\varepsilon(\hat{h}^S) - \varepsilon(h_\infty^{\mathrm{agg}}) \leq \mathrm{Dis}(\hat{h}^S, h_\infty^{\mathrm{agg}}) \leq \widehat{\mathrm{Dis}}(\hat{h}^S, h_\infty^{\mathrm{agg}}) + \tilde{O}\Big(\sqrt{\frac{d}{m}}\Big). \tag{3}$$

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

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

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

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

Recall that T is the unstable cutoff in Algorithm 3. The proof completes by invoking Proposition 11 which implies that $\widehat{\mathrm{Dis}}(\hat{h}^{\mathrm{priv}}, h_\infty^{\mathrm{agg}}) \le T/m$ with high probability. □

Theorem B.7 (Restatement of Theorem 13)**.** *Under the same assumption of Theorem 12. 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* $\ge 1-2\gamma$

$$\varepsilon(\hat{h}^S) - \varepsilon(h_\infty^{\mathrm{agg}}) \le \frac{4T}{m} + \frac{5(Kd+\log(4/\gamma))}{m} = \tilde{O}\left(\nu + \frac{\log(4/\gamma)}{m} + \frac{d\sqrt{\nu}}{\Delta\sqrt{m}}\right)$$

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

$$\begin{aligned}
\varepsilon(\hat{h}^S) - \varepsilon(h_\infty^{\mathrm{agg}}) &\le \mathrm{Dis}(\hat{h}^S, h_\infty^{\mathrm{agg}}) \\
&\le \widehat{\mathrm{Dis}}(\hat{h}^S, h_\infty^{\mathrm{agg}}) + 2\sqrt{\frac{(Kd+\log(4/\gamma))\widehat{\mathrm{Dis}}(\hat{h}^S, h_\infty^{\mathrm{agg}})}{m}} + \frac{4(Kd+\log(4/\gamma))}{m} \\
&\le 2\widehat{\mathrm{Dis}}(\hat{h}^S, h_\infty^{\mathrm{agg}}) + \frac{5(Kd+\log(4/\gamma))}{m}, \qquad (4)
\end{aligned}$$

where the second line follows from the first statement of Lemma D.4 in the Appendix with $z = h_\infty^{\mathrm{agg}}(x)$ and the third line is due to $a + 2\sqrt{ab} + b \le 2a + 2b$ for non-negative a, b.

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

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

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

Recall that T is the unstable cutoff in Algorithm 3. The proof completes by using that $\widehat{\mathrm{Dis}}(\hat{h}^{\mathrm{priv}}, h_\infty^{\mathrm{agg}}) \le T/m$ with probability $1-\gamma$ according to Proposition 11 and substitute the choices of T and K accordingly. □

Lemma B.8. *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,*

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

Proof. From (Hanneke, 2014), we learn that for any hypothesis h survive in V must satisfy

$$\varepsilon(h) - \varepsilon(h^*) \le 2U(m, \gamma).$$

Then by the definition of $U(m,\gamma)$ shown in Algorithm 7, we have

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

□

Theorem B.9 (Restatement of Theorem 15)**.** *With probability at least $1-\gamma$, there exists universal constants C_1, C_2 such that for all*

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

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

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

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

obeys that

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

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 $\epsilon \le \log(1/\delta)$, then it follows that

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

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

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

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

Step 2: PATE is just as good. Let $\tilde{h}^{\mathrm{priv}}$ 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 $\hat{h}^{\mathrm{priv}}$ 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,

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

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 $\tilde{h}^{\text{priv}}$ is the labeling function in the student learning problem. So the above implies that the student learning problem is close to realizable:

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

By the above, and the agnostic active learning bounds in Lemma D.7, to achieve an excess risk bound of $\alpha \geq \text{Dis}(\tilde{h}^{\text{priv}}, h^*) := \varepsilon^*$ in the student learning problem with probability at least $1-\gamma/2$, with unbounded m, it suffices to choose ℓ to be

$$\begin{aligned}
&C\theta(\varepsilon^* + \alpha)\left(\frac{(\varepsilon^*)^2}{\alpha^2} + \log(\frac{1}{\alpha})\right)\left(d\log(\theta(\varepsilon^* + \alpha)) + \log(\frac{\log(1/\alpha)}{\gamma})\right) \\
&\leq C\theta(\alpha)(1+\log(1/\alpha))\left(d\log(\theta(\alpha)) + \log(\frac{\log(1/\alpha)}{\gamma})\right).
\end{aligned}$$

This implies an error bound of

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

When m is small, we might not have enough data points to obtain $\alpha = O(\text{Dis}(\tilde{h}^{\text{priv}}, h^*))$ in this case the error is dominated by our bounds in Lemma B.8, which says that we can take $\alpha = C\max\{\varepsilon^*, \frac{d\log(m/d)+\log(2/\gamma)}{m}\}$.

Step 4 Putting everything together.

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

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

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

□

C Results and Analysis from (Bassily et al., 2018a) and the Analysis of Standard PATE

Lemma C.1 (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* $\text{margin}(x_k) < K/3$ *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* $\text{margin}(x_i) \geq K/3$, *the output of Algorithm 2 is* $\hat{h}^{\text{agg}}(x_i)$.

Lemma C.2 (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*

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

Lemma C.2 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 C.1.

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.

Theorem C.3 (Adapted from Theorems 4.6 and 4.7 of (Bassily et al., 2018a)). *Set*

$$T = 3(\mathbf{E}[\varepsilon(\hat{h}_1)]m + \sqrt{m\log(m/\beta)/2}),$$

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

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 iid), we have

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

for the realizable case, and

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

for the agnostic case.

Remark C.4 (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 $\tilde{O}(\frac{d^{3/5}}{n^{2/5}\epsilon^{2/5}})$ error bound for the agnostic case*[5].

Proof. The analysis essentially follows the proof of Theorem 5 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 $m\mathbf{E}[\varepsilon(\hat{h}_1)] + \sqrt{m\log(m/\beta)/2}$, which is B in Lemma C.2. Then following Lemma C.2, by choose $T = 3B = 3(m\mathbf{E}[\varepsilon(\hat{h}_1)] + \sqrt{m\log(m/\beta)/2})$, we ensure that the majority voting classifiers are correct and have high margin in at least $m - T$ examples.

In the realizable setting. Since $\varepsilon(h^*) = 0$ and by standard statistical learning theory in the realizable case (Lemma D.5), for each teacher classifier $\hat{h}_k$ we have

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

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

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

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

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

[5] 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 $O(\alpha + \varepsilon(h^*))$. The privacy parameter ϵ is taken as a constant in these results.

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 dependences 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 D.4),

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

where we applied the triangular inequality in the second line, used that $\hat{h}^S$ is the minimizer of $\widehat{\text{Dis}}(\hat{h}^{\text{priv}}, \cdot)$ in the third line, and then combined Lemma C.1 and Lemma C.2 to show that under the appropriate choice of T and K with high probability, $\hat{h}^{\text{priv}}(x_j)$ correctly returns y_j except for up to T example. Finally, the choice of T is substituted.

In agnostic setting. By Lemma D.5, with high probability, for all teacher classifier $\hat{h}_k$ for $k = 1, ..., K$, we have

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

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

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

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

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

Further, we can write

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

where the first line talks about two cases of Inequality (6): (1) $T/2 \le T - \tilde{O}(\frac{md^{1/2}T^{1/4}}{n^{1/2}\epsilon^{1/2}}) \le 3m\varepsilon(h^*) + \tilde{O}(\sqrt{m})$ 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 D.4)

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

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 $\hat{h}^{\text{priv}}$ and the fact that $h_1 \in \mathcal{H}$. The second last line follows from the fact that in those stable queries $\hat{h}^{\text{priv}}(x_j)$ 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).

□

Theorem C.5 (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

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

Specifically, in the realizable setting, then it follows that

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

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

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

where the third line follows from the learning bound (Lemma D.5) 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 $\text{Dis}(\tilde{h}^{\text{priv}}, h^*)$. First by Lemma B.2, with probability at least $1 - \gamma/2$, $\forall k \in [K]$ we have

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

Next, conditioning on the teachers, $\tilde{h}^{\text{priv}}$ 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,

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

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

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

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

Put everything together, and the excess risk bound is

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

□

Theorem C.6 (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 E.2) of the vote count is 1. Algorithm 1 is a straightforward adaptive composition of ℓ Gaussian mechanisms (Lemma E.5), which satisfies $\frac{\ell}{2\sigma^2}$-zCDP. By Lemma D.1, we get that the choice of σ gives us (ϵ, δ)-differential privacy.

Let us now address Algorithm 2. First note that $\text{margin}(x_j)$ 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 5), 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 6) which outputs $\{\hat{h}^{\text{agg}}(x_j)\}$ 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 $\{\hat{h}^{\text{agg}}(x_j)\}$ 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 E.9) of the distance-to-stability is to discuss the two cases. In the first case, assume for all j such that the output is $\perp$, $\hat{h}^{\text{agg}}(x_j)$ remains the same over D, D', then adding $\hat{h}^{\text{agg}}(x_j)$ to the output obeys 0-DP; in the second case, assume that there exists some j where we output $\perp$ such that, $\hat{h}^{\text{agg}}(x_j)$ is different under D and D', then for all these j we know that $\text{dist}_j = 0$ 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 do 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 D.1), 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. □

D Technical Lemmas

Lemma D.1 (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 D.2 (Proposition 1.4 of (Bun and Steinke, 2016)). *If $\mathcal{M}$ obeys ϵ-DP, then $\mathcal{M}$ obeys $\frac{\epsilon^2}{2}$-CDP.*

Lemma D.3 (Pointwise convergence (Bousquet et al., 2004)). *Let (x, z) be drawn from any distribution $\mathcal{D}$ supported on $\mathcal{X} \times \mathcal{Y}$. Let* Dis *and* $\widehat{\text{Dis}}$ *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 $1 - \gamma$,*

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

where n is the number of data points.

Lemma D.4 (Uniform convergence (Bousquet et al., 2004)). *Under the same conditions of Lemma D.3, 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,*

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

and

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

We will be taking z to be $h^*(x)$ in the cases when we work with noise conditions and $h^{\text{agg}}_\infty(x)$ in the agnostic case.

Lemma D.5 (Learning bound (Bousquet et al., 2004)). *Let d be the VC-dimension of $\mathcal{H}$, the excess risk is bounded with probability $1 - \gamma$,*

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

where n is the number of data points.

Lemma D.6 (Learning bound under the Tsybakov noise condition (Hanneke, 2014)). *Let d be the VC-dimension of the class $\mathcal{H}$. Assume Tsybakov noise condition with parameters η, τ, the excess risk is bounded with probability*

$1-\gamma$,

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

where n is the number of data points.

Lemma D.7 (Agnostic active learning bound (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 7) outputs a classifier with an access risk of α with*

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

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

E Additional Information

E.1 Differential Privacy

Lemma E.1 (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 E.2 (Global sensitivity (Dwork and Roth, 2014)). *A function $f: \mathcal{Z}^* \to \mathcal{R}$ has global sensitivity ϑ if*

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

Lemma E.3 (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 $b \sim \mathrm{Lap}(\vartheta/\epsilon)^p$, satisfies ϵ-DP. The $\mathrm{Lap}(\lambda)^p$ denotes a vector of p i.i.d. samples from the Laplace distribution $\mathrm{Lap}(\lambda)$.*

Definition E.4 (ℓ_2-sensitivity (Dwork and Roth, 2014)). *A function $f: \mathcal{Z} \to \mathcal{R}$ has ℓ_2 sensitivity ϑ_2 if*

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

Lemma E.5 (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 5 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.

$c \leftarrow 0, \lambda \leftarrow \sqrt{32T\log(1/\delta)/\epsilon}, \hat{w} \leftarrow w + \mathrm{Lap}(\lambda)$.
for $q \in \mathcal{Q}$ and $c \leq T$ **do**
  $\hat{q} \leftarrow q + \mathrm{Lap}(2\lambda)$.
  if $\hat{q} > \hat{w}$ **then**
    Output $\top$.
  else
    Output $\perp$. $\hat{w} \leftarrow w + 1, c \leftarrow c + 1$.
  end if
end for

Lemma E.6 (Privacy guarantee of Algorithm 5 (Dwork et al., 2010; Dwork and Roth, 2014)). *Algorithm 5 is (ϵ, δ)-DP.*

Lemma E.7 (Utility guarantee of Algorithm 5 (Dwork et al., 2010; Dwork and Roth, 2014)). *For $\alpha = \log(2mT/\beta)\sqrt{512T\log(1/\delta)}/\epsilon$, and any set of m queries, define the set $L(\alpha) = \{i : q_i(D) \leq w + \alpha\}$. If $|L(\alpha)| \leq T$, then w.p. at least $1-\beta$: $\forall i \notin L(\alpha)$ Algorithm 5 outputs $\top$.*

Definition E.8 (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 $|D \Delta D'| \leq k$. We say f is stable on D if it is (at least) 1-stable on D, and unstable otherwise.*

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

Lemma E.9 (Privacy guarantee of Algorithm 6 (Bassily et al., 2018b)). *If the threshold $\Gamma = \log(1/\delta)/\epsilon$, and the distance to instability function $\text{dist}_f(D) = \text{argmax}_k(f(D)$ is k-stable), then Algorithm 6 is (ϵ, δ)-DP.*

Lemma E.10 (Utility guarantee of Algorithm 6 (Thakurta and Smith, 2013)). *If the threshold $\Gamma = \log(1/\delta)/\epsilon$, and the distance to instability function $\text{dist}_f(D) = \text{argmax}_k(f(D)$ is k-stable), and $f(D)$ is $((\log(1/\delta) + \log(1/\beta))/\epsilon)$-stable, then Algorithm 6 outputs $f(D)$ w.p. at least $1 - \beta$.*

Definition E.11 (Definition 1.1 of (Bun and Steinke, 2016)). *$\mathcal{M}$ obeys (ξ, ρ)-zCDP if for two adjacent dataset D, D', for all $\alpha \in (1, \infty)$, the Renyi-divergence of order α below obeys that*

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

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).

Algorithm 7 Disagreement-Based Active Learning (Hanneke, 2014)

Input: A "data stream" $x_1, x_2, ...$ 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 $V \leftarrow \mathcal{H}$.
Initialize the selected dataset $Q \leftarrow \varnothing$.
Initialize "Current Output" to be any $h \in \mathcal{H}$.
Initialize "Counter" $c \leftarrow 0$
for $j \in [m]$ **do**
  if $x_j \in \text{DIS}(V)$ **then**
    "Request for label" for x_j and get back y_j from the "labeling service"
    Update $Q \leftarrow Q \cup \{(x_j, y_j)\}$
    $c \leftarrow c + 1$.
  end if
  if $\log_2(j) \in \mathbb{N}$ **then**
    Update $V \leftarrow \{h \in V : (\varepsilon_Q(h) - \min_{g \in V} \varepsilon_Q(g))|Q| \leq U(j, \gamma_j)j\}$ where $U(j, \gamma_j) = c'(d \log(\theta(d/j)) + \log(1/\gamma_j))/j + c'\sqrt{\varepsilon(h^*)(d \log(\theta(\varepsilon(h^*)) + \log(1/\gamma_j))/j}$ where c' is a constant and $\gamma_j = \gamma/(\log_2(2j))^2$
    Set "Current Output" to be any $h \in V$.
  end if
  if $c \geq \ell$ **then**
    Break.
  end if
end for

Output: Return "Current Output".

E.2 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 7, 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 E.12 (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,*

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

Remark E.13. *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.

F Simulation with Adult Dataset

In this section, we empirically estimate the expected margin $\Delta_n(x) = |\mathbf{E}[\hat{h}_1(x)|x] - 0.5|$ and $|\mathbf{E}[\mathbf{1}(h(x) \neq h^*(x))|x] - 0.5|$ on the Adult dataset. Note that in $\Delta_n(x)$, we do not require the teachers to agree on y or h^* but measure the extent to which they agree with $\hat{h}^{\mathrm{agg}}$. 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 2000 queries from the public domain. Note that under this setting, the private domain contains 36631 records and the public domain has 10211 unlabelled records. We train h^* with the entire private dataset using the logistic regression model.

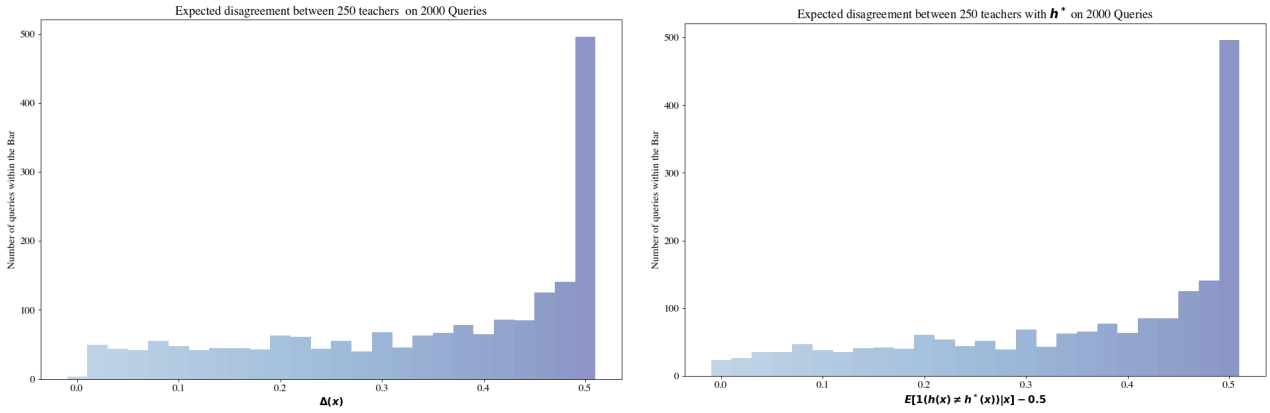

Figure 2: Empirical distribution of the margins on the Adult dataset.

Results shown in Figure 2 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 3, we plot the correlations between $\Delta(x)$ and $\mathbf{E}[\mathbf{1}[h(x) \neq h^*(x)]|x] - 0.5$ over 200 queries. The x-axis is the cumulation of $\Delta(x)$ and the y-axis is the cumulation of $\mathbf{E}[\mathbf{1}[h(x) \neq h^*(x)]|x] - 0.5$. 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.

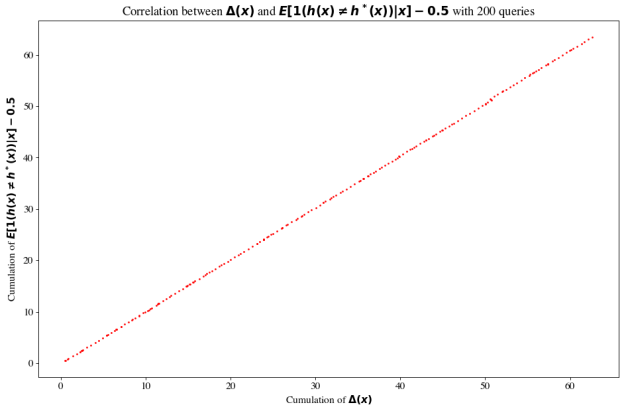

Figure 3: Correlations between margins on the Adult dataset.