SCALABLE PRIVATE LEARNING WITH PATE

Anonymous authors
Paper under double-blind review

ABSTRACT

Recently, there has been increased attention to the privacy concerns raised by machine learning (ML) models trained on highly sensitive data, such as medical records or other personal information. To address those concerns, one attractive approach is the *Private Aggregation of Teacher Ensembles* (PATE), which has shown that knowledge from an ensemble of teacher models trained on disjoint sets of the training data can be aggregated and transferred to a student model with strong privacy guarantees. While promising, PATE applications have so far been limited to simple classification tasks like MNIST; its scalability remained unclear notably because of inherent limitations of the random noise distributions used to provide privacy when transferring knowledge to the student.

In this work, we leverage two key insights: aggregation mechanisms with concentrated noise are more likely to output the correct label and an ensemble of teachers may choose to only answer questions on which they generally agree. Intuitively, such consensus answers are more likely to be correct and also ought to incur lower privacy costs. With these new aggregation mechanisms and tighter privacy analyses, we greatly improve PATE's tradeoffs, thereby leading to better scalability.

In experiments, we improve the state-of-the-art on privacy-preserving ML benchmarks. We also successfully apply PATE with our new ideas to an uncurated task with hundreds of classes and whose data is *imbalanced* and partly *mislabeled*.

1 INTRODUCTION

Many applications of ML often operate (in)directly with data that is considered sensitive—because it contains private information contributed by individuals. This sensitive data is for instance common in healthcare applications. For instance, algorithms designed to solve pattern recognition tasks are now also used to identify medical conditions indicative of cancer (Liu et al., 2017) and diabetic retinopathy (Gulshan et al., 2016). Hence, protecting the privacy of individuals whose data is analyzed to build these systems is important.

In the following, we revisit an approach for learning with privacy named PATE, which stands for the *private aggregation of teacher ensembles* (Papernot et al., 2017). We chose to revisit this particular approach among the many available (see Appendix B) because it is *agnostic* to any underlying ML technique. It is thus applicable to ML tasks where modifications of the learning procedure, e.g., optimizing with the differentially-private SGD of Abadi et al. (2016), is challenging.

The PATE approach first learns an ensemble of teacher models from the sensitive training data. Each teacher is independently trained on a data partition. Their predictions are aggregated through a majority vote to supervise the training of a student model. Intuitively, the partitioning and aggregation provide some form of privacy: aggregated predictions cannot be influenced by any given point. Teachers precisely control sensitive information the student has access to by adding carefully calibrated noise to their predictions during the aggregation step. This noise, sampled from a Laplace distribution, allows one to prove rigorous guarantees of differential privacy (Dwork et al., 2006)—a gold standard of privacy. In this work, we improve PATE to scale its application to uncurated data.

First, we leverage the synergy between privacy and utility, which contrary to the popular belief are not necessarily conflicting, to make better use of the privacy budget spent when transfering knowledge from teachers to the student. We introduce two aggregation mechanisms that explicitly leverage this alignment between accurate and private predictions. They are designed to choose not to label inputs queried by the student when there is either (a) a lack of consensus among teachers, or (b) an already overwhelming agreement between the student's prediction and the teacher votes.

Second, we demonstrate that sampling from a Gaussian distribution rather than the Laplace distribution used in the original PATE proposal facilitates better aggregation of teacher votes. This is particularly true for tasks with large number of output classes, as well as ensembles with many teachers and follows from the faster diminishing tails of the Gaussian distribution compared to the Laplace distribution. However, it requires a privacy analysis from scratch which we detail in Appendix C.

Third, and of possible independent interest, a by-product of our evaluation is that adversarial training techniques support semi-supervised learning on tasks with many classes. Although we confirm the performance of the semi-supervised application of generative adversarial networks (GANs) found in Salimans et al. (2016) and used by the original PATE proposal for SVHN (10 classes); we find that the virtual adversarial training (VAT) of Miyato et al. (2017) outperforms GANs on the Glyph dataset (150 classes, see Section 4.1). With VAT, unlabeled training points are used to regularize the student by ensuring that its predictions remain constant in *adversarial*[1] directions.

Our proposed aggregation mechanisms are complementary to the design choices made by the original PATE proposal, as demonstrated by our significant improvement on the state-of-the-art established by the first instantiation of PATE on all datasets evaluated. We also successfuly apply PATE to a large-scale character recognition problem, the Glyph dataset, which we chose because it involves training data that is *class imbalanced* and contains *erroneously labeled* points.

In short, our contributions are as follows:

- We refine the aggregation step of the PATE framework—as well as the corresponding privacy analysis—to use the Gaussian mechanism in Section 3.1. At no cost in privacy, the accuracy of the aggregation is improved for tasks with a large number of classes.
- We enable aggregation mechanisms to only answer a subset of queries; when teachers agree on the prediction or the student model already correctly predicts the label with confidence.
- We improve the state-of-the-art tradeoff between privacy and utility on all datasets used to evaluate the original PATE proposal: MNIST, SVHN, UCI Adult, and UCI Diabetes. For instance, we achieve on SVHN a test accuracy of 91.7% with an (ε, δ)-differential privacy guarantee of $(5.34, 10^{-6})$. Papernot et al. (2017) achieved 82.72% (90.7%, resp.) accuracy with $\varepsilon = 5.04$ (8.19, resp) for $\delta = 10^{-6}$. Given the (baseline) non-private accuracy of 92.8%, we increase privacy at a fixed utility; and utility at a fixed privacy.
- We then demonstrate our ideas' performance on *uncurated data*: a pattern-recognition task with hundreds of classes, mislabeled data, and class imbalance. In Section 4.3, we highlight the very strong privacy ($\varepsilon \approx 1.0$) provided in conjunction with high student utility.

2 Overview

We introduce essential components of our approach towards a generic and flexible framework for machine learning with provable privacy guarantees for training data.

2.1 The PATE Framework

The PATE framework (see Figure 7 in Appendix A) due to Papernot et al. (2017) comprises three principal components: (1) an ensemble of *teachers*, (2) an *aggregation mechanism*, and (3) a *student*.

Teachers. Each teacher is a model trained independently on a subset of the data whose privacy one wishes to protect. The data is partitioned to ensure no pair of teachers will have trained on overlapping data. Any learning technique suitable for the data can be used for any teacher.

Aggregation Mechanism. The role of this mechanism is to be the intermediary between the student, who is never directly exposed to sensitive data, and the teachers, each of which has unfettered access to a data partition. This step enforces privacy by tracking the privacy budget spent (see below).

Student. The student is another model with access to public unlabeled data. During learning, the student queries the aggregator with public data and receives private labels. Yet, each of these labels

[1] In this context, the adversarial component refers to the phenomenon commonly referred to as adversarial examples (Biggio et al., 2013; Szegedy et al., 2014) and not to the adversarial training approach taken in GANs.

increases the exposure of the teachers' sensitive training data. Hence, the student and aggregator are designed to minimize the querying needed to transfer knowledge from the teachers to the student.

2.2 Differential Privacy and Rényi Differential Privacy

Differential privacy (Dwork et al., 2006) is a gold standard of privacy for statistical algorithms. It requires that the sensitivity of the distribution of an algorithm's output to small perturbations of its input be limited. The following variant of the definition captures this intuition formally:

Definition 1. *A randomized mechanism $\mathcal{M}$ with domain $\mathcal{D}$ and range $\mathcal{R}$ satisfies (ε, δ)-differential privacy if for any two adjacent inputs $D, D' \in \mathcal{D}$ and for any subset of outputs $S \subseteq \mathcal{R}$ it holds that:*

$$\mathbf{Pr}[\mathcal{M}(D) \in S] \leq e^{\varepsilon} \cdot \mathbf{Pr}[\mathcal{M}(D') \in S] + \delta. \tag{1}$$

For our application of differential privacy to ML, adjacent inputs are defined as two datasets that only differ by one training example and the randomized mechanism $\mathcal{M}$ would be the model training algorithm. Composition theorems for differential privacy (Dwork & Roth, 2014) lead to the following natural interpretation of the parameters from its definition: ε is an upper bound on the loss of privacy that accumulates over time and with each query to the dataset ("the privacy budget"), and δ is the probability with which this guarantee may not hold.

To better facilitate tracking privacy budget across multiple queries, we use Rényi Differential Privacy (RDP) recently introduced by Mironov (2017). RDP is closely related to (zero)-Concentrated Differential Privacy (Dwork & Rothblum, 2016; Bun & Steinke, 2016) and the moments accountant (Abadi et al., 2016)—also used by Papernot et al. (2017) in the original PATE proposal.

2.3 Aggregation Mechanisms

The aggregation step is a crucial component of PATE. It enables knowledge transfer from the teachers to the student while enforcing privacy. We improve the LNMax mechanism used by Papernot et al. (2017) which adds Laplace noise to teacher votes and outputs the class with the highest votes.

First, we add Gaussian noise with an accompanying privacy analysis in the RDP framework. This modification effectively reduces the noise needed to achieve the same privacy cost per student query.

Second, the aggregation mechanism is now *selective*: teacher votes are analyzed to decide which student queries are *worth* answering. This takes into account both the privacy cost of each query and its payout in improving the student's utility. Surprisingly, our analysis shows that these two metrics are not at odds and in fact align with each other: the privacy cost is the smallest when teachers agree, and when teachers agree, the label is more likely to be correct thus being more useful to the student.

Third, we propose and study an *interactive* mechanism that takes into account not only teacher votes on a queried example but possible student predictions on that query. Now, queries worth answering are those where the teachers agree on a class but the student is not confident in its prediction on that class. This third modification aligns the two metrics discussed above even further: queries where the student already agrees with the consensus of teachers are not worth expending our privacy budget on, but queries where the student is less confident are useful and answered at a small privacy cost.

2.4 Data-dependent Privacy

A direct privacy analysis of the aggregation mechanism, for reasonable values of the noise parameter, allows answering only few queries before the privacy cost becomes prohibitive. The original PATE proposal used a data-dependent analysis, exploiting the fact that when the teachers have large agreement, the privacy cost is usually much smaller than the data-dependent bound would suggest.

In our work, we perform a data-dependent privacy analysis of the aggregation mechanism with Gaussian noise. This change of noise distribution turns out be technically much more challenging than the Laplace noise case and we defer the details to Appendix C. This increased complexity of the analysis however does not make the algorithm any more complicated and thus allows us to improve the privacy-utility tradeoff.

An additional challenge with data-dependent privacy analyses arises from the fact that the privacy cost itself is now a function of the private data. Further, the data-dependent bound on the privacy cost has large global sensitivity (a metric used in differential privacy to calibrate the noise injected) and is therefore difficult to sanitize. To remedy this, we use the smoothed-sensitivity framework. This consideration also affects the choice of bound in our data-dependent privacy analysis. We chose a bound that tradeoffs utility (i.e., it offers a tight analysis of the aggregators built on the Gaussian mechanism) while still allowing us to prove a small smoothed sensitivity bound.

3 Improved Aggregation Mechanisms

Guarantees provided by PATE stem from the design and analysis of the aggregation step. Here, we detail our improvements to the mechanism used by Papernot et al. (2017). As outlined in Section 2.3, we first replace the Laplace noise added to teacher votes with Gaussian noise, adapting the data-dependent privacy analysis. Next, we describe the Confident and Interactive Aggregators that select queries worth answering in a privacy-preserving way: the privacy budget is shared between the query selection and answer computation. The aggregators use different heuristics to select queries: the former does not take into account student predictions, while the latter does.

3.1 The GNMax Aggregator and Its Privacy Guarantee

This section uses the following notation. For a sample x and classes 1 to m, let $f_j(x) \in [m]$ denote the j-th teacher model's prediction on x and $n_i(x)$ denote the vote count for the i-th class (i.e., $n_i(x) = |\{j : f_j(x) = i\}|$). We define a Gaussian NoisyMax (GNMax) aggregation mechanism as:

$$\mathcal{A}_\sigma(x) := \operatorname*{argmax}_i \left\{ n_i(x) + \mathcal{N}(0, \sigma^2) \right\},$$

where $\mathcal{N}(\mu, \sigma^2)$ denotes the Gaussian distribution with mean μ and variance σ^2. The aggregator outputs the class with noisy plurality after adding Gaussian noise to each vote count. In what follow, *plurality* more generally refers to the highest number of teacher votes assigned among the classes.

The Gaussian distribution is more concentrated than the Laplace distribution used by Papernot et al. (2017). This concentration directly improves the aggregation's utility when the number of classes m is large. The GNMax mechanism satisfies $(\lambda, \lambda/\sigma^2)$-RDP (see Lemma 8 in Appendix C) which holds for all inputs and all $\lambda \geq 1$. A straightforward application of composition theorems leads to loose privacy bounds. As an example, the standard advanced composition theorem applied to experiments in rows 5 and 6 in Fig. 3 would give us $\varepsilon = 8.42$ and $\varepsilon = 10.14$ resp. at $\delta = 10^{-8}$ for the Glyph dataset.

To refine these, we work out a careful *data-dependent* analysis that yields ε's close to 1 for the same δ. The following theorem translates standard RDP guarantees for higher orders into a data-dependent RDP guarantee for a smaller order λ. Its precise statement and proof appear in Appendix C. We use it in conjunction with Lemma 3 to bound the privacy cost of each query to the GNMax algorithm as a function of $\tilde{q}$, the probability that the most common answer will not be output by the mechanism.

Theorem 2 (informal)**.** *Let $\mathcal{M}$ be a randomized algorithm with (μ_1, ε_1)-RDP and (μ_2, ε_2)-RDP guarantees and suppose that given a dataset D, there exists* a likely *outcome i^* such that $\mathbf{Pr}[\mathcal{M}(D) \neq i^*] \leq \tilde{q}$. Then the* data-dependent *Rényi differential privacy for $\mathcal{M}$ of order $\lambda \leq \mu_1, \mu_2$ at D is bounded by a function of $\tilde{q}$, μ_1, ε_1, μ_2, ε_2, which approaches 0 as $\tilde{q} \to 0$.*

The new bound improves on the data-independent privacy for λ as long as the distribution of the algorithm's output *on that input* has a strong peak (i.e., $\tilde{q} \ll 1$). Values of $\tilde{q}$ close to 1 could result in a looser bound. Therefore, in practice we take the minimum between this bound and λ/σ^2 (the data-independent one). The theorem generalizes Theorem 3 from Papernot et al. (2017), where it was shown for a mechanism satisfying ε-differential privacy (i.e., $\mu_1 = \mu_2 = \infty$ and $\varepsilon_1 = \varepsilon_2$).

The final step in our analysis uses the following lemma to bound the probability $\tilde{q}$ when i^* corresponds to the class with the true plurality of teacher votes.

Lemma 3. *For any $i^* \in [m]$, we have* $\mathbf{Pr}[\mathcal{A}_\sigma(D) \neq i^*] \leq \frac{1}{2} \sum_{i \neq i^*} \operatorname{erfc}\left(\frac{n_{i^*} - n_i}{2\sigma}\right)$.

Algorithm 1 - Confident GNMax Aggregator: given a query, consensus among teachers is first estimated in a privacy-preserving way to then only reveal confident teacher predictions.

Input: input x, threshold T, noise parameters σ_1 and σ_2
1: **if** $\max_i\{n_j(x)\} + \mathcal{N}(0, \sigma_1^2) \geq T$ **then** ▷ Privately check for consensus
2: **return** $\operatorname{argmax}_j \left\{n_j(x) + \mathcal{N}(0, \sigma_2^2)\right\}$ ▷ Run the usual max-of-Gaussian
3: **else**
4: **return** $\perp$
5: **end if**

In Appendix C, we detail how these results translate to privacy bounds. In short, for each query to the GNMax aggregator, given teacher votes n_i and the class i^* with maximal support, Lemma 3 gives us the value of $\tilde{q}$ to use in Theorem 2. We optimize over μ_1 and μ_2 to get a data-dependent RDP guarantee for any order λ. Finally, we use composition properties of RDP to analyze a sequence of queries, and translate the RDP bound back to an (ε, δ)-DP bound.

Expensive queries. This data-dependent privacy analysis leads us to the concept of an *expensive* query in terms of its privacy cost. When teacher votes largely disagree, some $n_{i^*} - n_i$ values may be small leading to a large value for $\tilde{q}$: i.e., the lack of consensus amongst teachers indicates that the aggregator is likely to output the wrong label. Thus expensive queries from a privacy perspective are often bad for training too. Conversely, queries with strong consensus enable tight privacy bounds. This synergy motivates the aggregation mechanisms discussed in the following sections: they evaluate the strength of the consensus before answering a query.

3.2 The Confident Aggregator

In this section, we propose a refinement of the GNMax aggregator that enables us to filter out queries for which teachers do not have a sufficiently strong consensus. This filtering enables the teachers to avoid answering expensive queries. We also take note to do this selection step itself in a private manner.

The proposed *Confident Aggregator* is described in Algorithm 1. To select queries with overwhelming consensus, the algorithm checks if the plurality vote crosses a threshold T. To enforce privacy in this step, the comparison is done after adding Gaussian noise with a large deviation σ_1. Then, for queries that pass this noisy threshold check, the aggregator proceeds with the usual GNMax mechanism with a different deviation σ_2. For queries that do not pass the noisy threshold check, the aggregator simply returns $\perp$ and the student discards this example in its training.

In practice, we often choose significantly higher values for σ_1 compared to σ_2. We pick T so that queries where the plurality gets less than half the votes (often very expensive) are unlikely to pass the threshold after adding noise, but we still have a high enough yield amongst the queries with a strong consensus. This tradeoff leads us to look for T's between $0.6\times$ to $0.8\times$ the number of teachers.

The privacy cost of this aggregator is intuitive: we pay for the threshold check for every query, and for the GNMax step only for queries that pass the check. In the work of Papernot et al. (2017), the mechanism paid a privacy cost for every query, expensive or otherwise. In comparison, the Confident Aggregator expends a much smaller privacy cost to check against the threshold, and by answering a significantly smaller fraction of expensive queries, it expends a lower privacy cost overall.

3.3 The Interactive Aggregator

While the Confident Aggregator exclude expensive queries, it ignores the possibility that the student might receive labels that contribute little to learning, and in turn to its utility. By incorporating the student's current predictions for its public training data, we design an *Interactive Aggregator* that discards queries where the student already confidently predicts the same label as the teachers.

Given a set of queries, the Interactive Aggregator (Algorithm 2) selects those answered by comparing student predictions to teacher votes for each class. Similar to Step 1 in the Confident Aggregator, queries where the plurality of these noised differences crosses a threshold are answered with GN-

Algorithm 2 - Interactive GNMax Aggregator: the protocol first compares student predictions to the teacher votes in a privacy-preserving way to then either (a) reinforce the student prediction for the given query or (b) provide the student with a new label predicted by the teachers.

Input: input x, confidence γ, threshold T, noise parameters σ_1 and σ_2, total number of teachers M

Ask the student to provide prediction scores $\mathbf{p}(x)$
if $\max_j\{n_j(x) - Mp_j(x)\} + \mathcal{N}(0, \sigma_1^2) \geq T$ **then** ▷ Student does not agree with teachers
  return $\operatorname{argmax}_j\{n_j(x) + \mathcal{N}(0, \sigma_2^2)\}$ ▷ Teachers provide new label
else if $\max\{p_i(x)\} > \gamma$ **then** ▷ Student agrees with teachers and is confident
  return $\arg\max_j p_j(x)$ ▷ Reinforce student's prediction
else
  return $\perp$ ▷ No output given for this label
end if

Max. This noisy threshold suffices to enforce privacy of the first step because student predictions can be considered public information (the student is trained in a differentially private manner).

For queries that fail this check, the mechanism reinforces the predicted student label if the student is confident enough and does this without looking at teacher votes again. This limited form of supervision comes at a small privacy cost. Moreover, the order of the checks ensures that a student falsely confident in its predictions on a query is not accidentlly reinforced if it disagrees with the teacher consensus. The privacy accounting is identical to the Confident Aggregator except in considering the difference between teachers and the student instead of only the teachers votes.

In practice, the Confident Aggregator can be used to start training a student when it can make no meaningful predictions and training can be finished off with the Interactive Aggregator after the student gains some proficiency.

4 Experimental Evaluation

Our goal is first to show that the improved aggregators introduced in Section 3 enable the application of PATE to uncurated data, thus departing from previous results on tasks with balanced and well-separated classes. We experiment with the Glyph dataset described below to address two aspects left open by Papernot et al. (2017): (a) the performance of PATE on a task with a larger number of classes (the framework was only evaluated on datasets with at most 10 classes) and (b) the privacy-utility trade-offs offered by PATE on data that is class imbalanced and partly mislabeled. In Section 4.2, we evaluate the improvements given by the GNMax aggregator over its Laplace counterpart (LNMax) and demonstrate the necessity of the Gaussian mechanism for uncurated tasks.

In Section 4.3, we then evaluate the performance of PATE with both the Confident and Interactive Aggregators on all datasets used to benchmark the original PATE framework, in addition to Glyph. With the right teacher and student training, the two mechanisms from Section 3 achieve high accuracy with very tight privacy bounds. Not answering queries for which teacher consensus is too low (Confident-GNMax) or the student's predictions already agree with teacher votes (Interactive-GNMax) better aligns utility and privacy: queries are answered at a significantly reduced cost.

4.1 Experimental Setup

MNIST, SVHN, UCI Adult, UCI Diabetes. We evaluate with two computer vision tasks (MNIST and Street View House Numbers (Netzer et al., 2011)), census data from the UCI Adult dataset Kohavi (1996), and medical records from the UCI Diabetes Strack et al. (2014) dataset. This enables a comparative analysis of the utility-privacy trade-off achieved with our Confident-GNMax aggregator and the LNMax originally used in PATE. We replicate the experimental setup and results foud in Papernot et al. (2017) thanks in part to the code and teacher votes made available online.

A detailed description of the experimental setup can be found in Papernot et al. (2017); we provide here only a brief overview. For MNIST and SVHN, teachers are convolutional networks trained on partitions of the training set. For Adult and Diabetes, each teacher is a random forest. The test set is split in two halves: the first is used as unlabeled inputs to simulate the student's public data

and the second is used as a hold out to evaluate test performance. The MNIST and SVHN students are convolutional networks trained using semi-supervised learning with GANs à la Salimans et al. (2016). The students for the Adult and Diabetes datasets are fully supervised random forests.

Glyph. This optical character recognition task has *an order of magnitude more classes* than all previous applications of PATE. The Glyph dataset also posesses many characteristics shared by real-world tasks: e.g., it is imbalanced and some inputs are mislabeled. Each input is a 28×28 grayscale image containing a single glyph generated synthetically from a collection of over 500K computer fonts.[2] Samples representative of the difficulties raised by the data are depicted in Figure 1. The task is to classify inputs as one of the 150 Unicode symbols used to generate them.

This set of 150 classes results from pre-processing efforts. We discarded additional classes that had few samples; some classes had at least 50 times fewer inputs than the most popular classes, and these were almost exclusively incorrectly labeled inputs. We also merged classes that were too ambiguous for even a human to differentiate them. Nevertheless, a manual inspection of samples grouped by classes—favorably to the human observer—led to the conservative estimate that some classes remain 5 times more frequent, and mislabeled inputs represent at least 10% of the data.

To simulate the availability of private and public data (see Appendix A), we split data originally marked as the training set (about 65M points) into partitions given to the teachers. Each teacher is a ResNet (He et al., 2016) made of 32 leaky ReLU layers.[3] We train on batches of 100 inputs for 40K steps with the Momentum algorithm. The learning rate, initially set to 0.1, is decayed after 10K steps to 0.01 and again after 20K steps to 0.001. These parameters were found with a grid search.

We split holdout data in two subsets of 100K and 400K samples: the first acts as public data to train the student and the second as its testing data. The student architecture is the convolutional network found in Miyato et al. (2017), and it is learned in a semi-supervised fashion with virtual adversarial training (VAT): it minimizes an unsupervised term to encourage predictions on training points and their adversarial examples to remain identical. Indeed, we found that GANs did not yield as much utility for Glyph as for SVHN. We train with Adam for 400 epochs and a learning rate of $6 \cdot 10^{-5}$.

4.2 Comparing the Laplace and Gaussian mechanisms

Section 3.1 introduces the GNMax mechanism and the accompanying privacy analysis. With a Gaussian distribution, whose tail diminishes more rapidly than the Laplace distribution, we expect better utility when using the new mechanism (albeit with a more involved privacy analysis).

To study the trade-off between privacy and accuracy with the two mechanisms, we run experiments training several ensembles of M teachers for $M \in \{100, 500, 1000, 5000\}$ on the Glyph data. Recall that 65 million training inputs are partitioned and distributed among the M teachers with each teacher receiving between 650K and 13K inputs for the values of M above. The test data is used to query the teacher ensemble and the resulting labels (after the LNMax and GNMax mechanisms) are compared with the ground truth labels provided in the dataset. This predictive performance of the teachers is essential to good student training with accurate labels and is a useful proxy for utility.

For each mechanism, we compute (ε, δ)-differential privacy guarantees. As is common in literature, for a dataset on the order of 10^8 samples, we choose $\delta = 10^{-8}$ and denote the corresponding ε as the privacy cost. The total ε is calculated on a subset of 4,000 queries, which is representative of the number of labels needed by a student for accurate training (see Section 4.3). We visualize in Figure 2 the effect of the noise distribution (left) and the number of teachers (right) on the tradeoff between privacy costs and label accuracy.

Observations. On the left of Figure 2,we compare our GNMax aggregator to the LNMax aggregator used by the original PATE proposal, on an ensemble of 1000 teachers and for varying noise scales σ. At fixed test accuracy, the GNMax algorithm consistently outperforms the LNMax mechanism in terms of privacy cost. To explain this improved performance, recall notation from Section 3.1. For both mechanisms, the data dependent privacy cost scales linearly with $\tilde{q}$—the likelihood of an answer other than the true plurality. The value of $\tilde{q}$ falls of as $\exp\left(-x^2\right)$ for GNMax and $\exp\left(-x\right)$

[2]Glyph data is not public but similar data is available publicly as part of the notMNIST dataset.

[3]It is available in the TensorFlow models library at `https://github.com/tensorflow/models`.

Figure 1: **Some example inputs from the Glyph dataset along with the class they are labeled as.** Note the ambiguity (between the comma and apostrophe) and the mislabeled input.

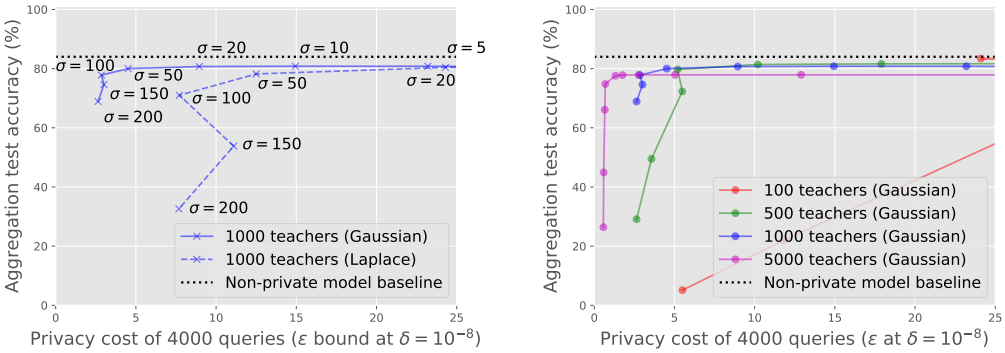

Figure 2: **Trade-off between utility and privacy for the LNMax and GNMax aggregators on Glyph:** effect of the noise distribution (left) and size of the teacher ensemble (right). The LNMax aggregator uses a Laplace distribution and GNMax a Gaussian. Smaller values of the privacy cost ε (often obtained by increasing the noise scale σ—see Section 3) and higher accuracy are better.

Dataset	Aggregator	Queries answered	Privacy bound ε	Accuracy Student	Accuracy Baseline
MNIST	LNMax (Papernot et al., 2017)	100	**2.04**	98.0%	99.2%
	LNMax (Papernot et al., 2017)	1,000	8.03	98.1%	
	Confident GNMax (T=200, σ_1=100, σ_2=40)	317	2.30	**98.5%**	
SVHN	LNMax (Papernot et al., 2017)	500	**5.04**	82.7%	92.8%
	LNMax (Papernot et al., 2017)	1,000	8.19	90.7%	
	Confident GNMax (T=300, σ_1=200, σ_2=40)	3,464	5.34	**91.7%**	
Adult	LNMax (Papernot et al., 2017)	500	2.66	83.0%	85.0%
	Confident GNMax (T=300, σ_1=200, σ_2=40)	524	**1.90**	**83.7%**	
Diabetes	LNMax (Papernot et al., 2017)	500	1.44	**93.9%**	93.8%
	Confident GNMax (T=200, σ_1=100, σ_2=15)	494	**0.7**	**93.9%**	
Glyph	LNMax (Papernot et al., 2017)	4,000	4.3	72.4%	82.2%
	Confident GNMax (T=1000, σ_1=500, σ_2=100)	10,762	2.03	**75.5%**	
	Interactive GNMax (T=3500, σ_1=2000, σ_2=200)	10,422	**0.99**	73.2%	

Figure 3: **Utility and privacy of the students.** The Confident and Interactive GNMax aggregators introduced in Section 3 offer better trade-offs between privacy (characterized by the value of the bound ε) and utility (the accuracy of the student compared to a non-private baseline) than the LNMax aggregator used by the original PATE proposal on all datasets we evaluated with. For MNIST, Adult, Diabetes, and SVHN, we use the labels of ensembles of 250 teachers published by Papernot et al. (2017) and set $\delta = 10^{-5}$ to compute values of ε (to the exception of SVHN where $\delta = 10^{-6}$). All Glyph results use an ensemble of 5000 teachers and ε is computed for $\delta = 10^{-8}$.

for LNMax, where x is the ratio $(n_{i^*} - n_i)/\sigma$. Thus, when $n_{i^*} - n_i$ is (say) 4σ, LNMax would have $\tilde{q} \approx e^{-4} = 0.018...$, whereas GNMax would have $\tilde{q} \approx e^{-16} \approx 10^{-7}$, thereby leading to a much higher likelihood of returning the true plurality. Moreover, this reduced $\tilde{q}$ translates to a smaller privacy cost for a given σ leading to a better utility-privacy trade-off.

As long as each teacher has sufficient data to learn a good-enough model, increasing the number M of teachers improves the trade-off—as illustrated on the right of Figure 2 with GNMax. The larger ensembles lower the privacy cost of answering queries by tolerating larger σ's. Combining the two observations made in this Figure, for a fixed label accuracy, we lower privacy costs by switching to the GNMax aggregator and training a larger number M of teachers.

4.3 Student Training with New Aggregation Mechanisms

As outlined in Section 2, we train a student on public data labeled by the aggregation mechanisms. We take advantage of PATE's flexibility and apply the technique that performs best on each dataset: semi-supervised learning with Generative Adversarial Networks (Salimans et al., 2016) for MNIST and SVHN, Virtual Adversarial Training (Miyato et al., 2017) for Glyph, and fully-supervised random forests for Adult and Diabetes. In addition to evaluating the total privacy cost associated with training the student model, we compare its utility to a non-private baseline obtained by training on the sensitive data (used to train teachers in PATE): we use the baselines of 99.2%, 92.8%, 85.0%, and 93.8% reported by Papernot et al. (2017) respectively for MNIST, SVHN, Adult, and Diabetes, and we measure a baseline of 82.2% for Glyph. We compute (ε, δ)-privacy bounds and denote the privacy cost as the ε value at a value of δ set accordingly to number of training samples.

Confident Aggregator. Given a pool of 500 to 12,000 samples to learn from (depending on the dataset), the student submits queries to the teacher ensemble running the Confident-GNMax aggregator from Section 3.2. A grid search over a range of plausible values for parameters T, σ_1 and σ_2 yielded the values reported in Figure 3, illustrating the tradeoff between utility and privacy achieved. We additionally measure the number of queries selected by the teachers to be answered and compare student utility to a non-private baseline.

The Confident-GNMax aggregator outperforms LNMax for the four datasets considered in the original PATE proposal: it reduces the privacy cost ε, increases student accuracy, or both simultaneously. On the uncurated Glyph data, despite the imbalance of classes and mislabeled data (as evidenced by the 82.2% baseline), the Confident Aggregator achieves 73.5% accuracy with a privacy cost of just $\varepsilon = 1.02$. Roughly 1,300 out of 12,000 queries made are not answered, indicating that several costly queries were successfully avoided. This selectivity is analyzed in more details in Section 4.4.2.

Interactive Aggregator. On Glyph, we evaluate the utility and privacy of an interactive training routine that proceeds in *two rounds*. Round one runs student training with a Confident Aggregator. A grid search targeting the best privacy for roughly 3,400 answered queries (out of 6,000)—sufficient to bootstrap a student—led us to setting $(T{=}3500, \sigma_1{=}1500, \sigma_2{=}100)$ and a privacy cost of $\varepsilon \approx 0.65$.

In round two, this student was then trained with 10,000 more queries made with the Interactive-GNMax Aggregator. We computed the resulting (total) privacy cost and utility at an *exemplar* data point through another grid search of plausible parameter values. The result appears in the last row of Figure 3. With just over 10,422 answered queries in total at a privacy cost of $\varepsilon = 0.99$, the trained student was able to achieve 73.2% accuracy. Note that this students required fewer answered queries compared to the Confident Aggregator. The best overall cost of student training occurred when the privacy costs for the first and second rounds of training were roughly the same.[4]

Baseline. Note that the Glyph student's accuracy remains seven percentage points below the non-private model's accuracy achieved by training on the 65M sensitive inputs. We hypothesize that is due to the uncurated nature of the data considered. Indeed, the class imbalance naturally requires more queries to return labels from the less represented classes. For instance, a student trained on 200K queries is only 77% accurate on test data. In addition, the large fraction of mislabeled inputs are likely to have a large privacy cost: these inputs are sensitive because they are outliers of the distribution, which is reflected by the weak consensus among teachers on these inputs.

[4] The total ε is less than $0.65 \times 2 = 1.3$ due to better composition—via Theorems 6 and 7.

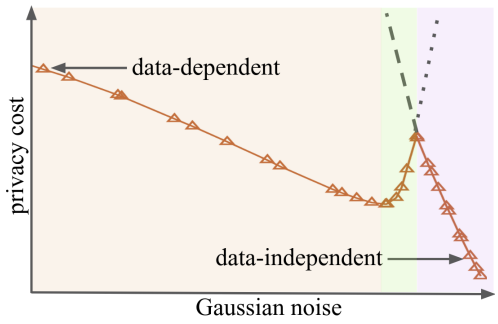

Figure 4: **Interaction between the data-independent and data-dependent bounds.** Shaded areas identify areas of different trade-offs between privacy and accuracy.

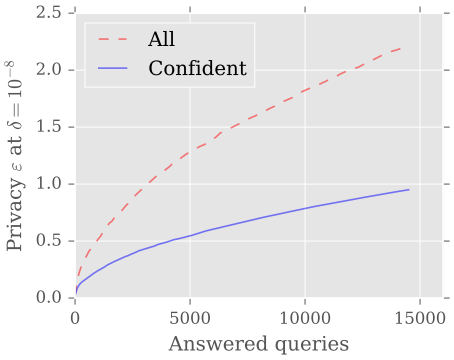

Figure 5: **Privacy cost of answering all vs only inexpensive queries**, keeping the number of answered queries fixed.

4.4 Empirical Considerations for the Privacy Analysis

Here, we discuss in more details two aspects of our privacy analysis that were briefly touched upon in our evaluation. First, the interaction between data dependent and independent analyses in the empirical computation of privacy cost ε. Second, the parameterization of the noisy threshold check.

4.4.1 Data Dependent vs. Independent Analysis

The data-dependent analysis introduces an interesting twist into the standard privacy-accuracy trade-off, which is usually monotone. Consider Figure 4, which exhibits a characteristic elbow: there is a point where accuracy lowers and the privacy cost increases. This elbow point is due to the nature of the differential privacy bound provided by our analysis: it is defined as the minimum of the data-independent and data-dependent bounds. There are three regimes illustrated in Figure 4: (1) the data-dependent bound decreases as more noise is injected and it is lower than the data-independent bound, (2) the data-dependent bound starts trending up, and (3) the data-dependent bound is overtaken by the data-independent analysis.

4.4.2 Noisy Threshold Check and Privacy Costs

Sections 3.1 and 3.2 motivated the need for a noisy threshold checking step before having the teachers answer queries: it prevents most of the privacy budget being consumed by few queries that are expensive and also likely to be incorrectly answered. In Figure 5, we compare the privacy cost ε of answering all queries to only answering confident queries for a fixed number of queries.

We run additional experiments to support the evaluation from Section 4.3. With the votes of 5000 teachers on the Glyph dataset, we plot in Figure 6 the histogram of the plurality vote counts (n_{i^*} in the notation of Section 3.1) across 25,000 student queries. We compare these values to the vote counts of queries that passed the noisy threshold check for two sets of parameters T and σ_1 in Algorithm 1. Smaller values imply weaker teacher agreements and consequently more expensive queries.

When (T=3500, σ_1=1500) we capture a significant fraction of queries where teachers have a strong consensus (roughly > 4000 votes) while managing to filter out many queries with poor consensus. This *moderate check* ensures that although many queries with plurality votes between 2,500 and 3,500 are answered (i.e., only 50–70% of teachers agree on a label) the expensive ones are most likely discarded. For (T=5000, σ_1=1500), queries with poor consensus are completely culled out. This selectivity comes at the expense of a noticeable drop for queries that might have had a strong consensus and little-to-no privacy cost. Thus, this *aggressive check* answer fewer queries with very strong privacy guarantees. We reiterate that this threshold checking step itself is done in a private manner. Emperically, in our Interactive Aggregator experiments, we expend about a third to a half of our privacy budget on this step, which still yields a very small cost *per query* across 6,000 queries.

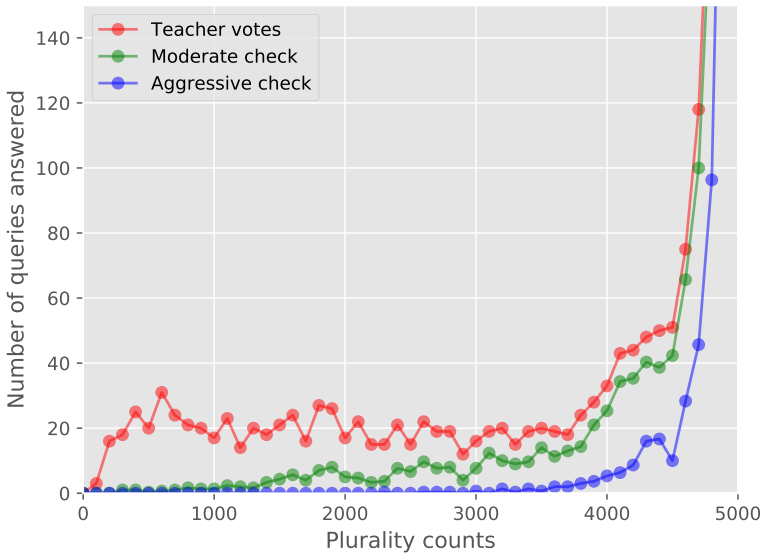

Figure 6: **Performance of the noisy threshold checking:** Teacher votes, the highest line represents the plurality values of teacher votes. Moderate check, the middle line shows the histogram for $(T{=}3500, \sigma_1{=}1500)$. Aggressive check, the bottom line is the histogram for $(T{=}5000, \sigma_1{=}1500)$.

5 Conclusions

The key insight motivating the addition of a noisy thresholding step to the two aggregation mechanisms proposed in our work is that there is a form of synergy between the privacy and accuracy of labels output by the aggregation: *labels that come at a small privacy cost also happen to be more likely to be correct*. As a consequence, we are able to provide more quality supervision to the student by choosing not to output labels when the consensus among teachers is too low to provide an aggregated prediction at a small cost in privacy. This observation was further confirmed in some of our experiments where we observed that if we trained the student on either private or non-private labels, the former almost always gave better performance than the latter—for a fixed number of labels.

Complimentary with these aggregation mechanisms is the use of a Gaussian (rather than Laplace) distribution to perturb teacher votes. In our experiments with Glyph data, these changes proved essential to preserve the accuracy of the aggregated labels—because of the large number of classes. The analysis presented in Appendix 3 details the delicate but necessary adaptation of analog results for the Laplace NoisyMax.

As was the case for the original PATE proposal, semi-supervised learning was instrumental to ensure the student achieves strong utility given a limited set of labels from the aggregation mechanism. However, we found that virtual adversarial training outperform the approach from Salimans et al. (2016) for the Glyph dataset. Thus, future work may continue to investigate other techniques for learning the student in these particular settings with restricted supervision.

REFERENCES

Martín Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 308–318. ACM, 2016.

Charu C Aggarwal. On k-anonymity and the curse of dimensionality. In *Proceedings of the 31st International Conference on Very large Data Bases*, pp. 901–909. VLDB Endowment, 2005.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Proceedings of the 2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, FOCS '14, pp. 464–473, 2014. ISBN 978-1-4799-6517-5.

Raghav Bhaskar, Srivatsan Laxman, Adam Smith, and Abhradeep Thakurta. Discovering frequent patterns in sensitive data. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 503–512. ACM, 2010.

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 387–402. Springer, 2013.

Vincent Bindschaedler, Reza Shokri, and Carl A Gunter. Plausible deniability for privacy-preserving data synthesis. *Proceedings of the VLDB Endowment*, 10(5), 2017.

Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pp. 635–658. Springer, 2016.

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

David Cohn, Les Atlas, and Richard Ladner. Improving generalization with active learning. *Machine learning*, 15(2):201–221, 1994.

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 and Guy N Rothblum. Concentrated differential privacy. *arXiv preprint arXiv:1603.01887*, 2016.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, volume 3876, pp. 265–284. Springer, 2006.

Cynthia Dwork, Guy N Rothblum, and Salil Vadhan. Boosting and differential privacy. In *The 51st Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pp. 51–60, 2010.

Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22):2402–2410, 2016.

Jihun Hamm, Yingjun Cao, and Mikhail Belkin. Learning privately from multiparty data. In *International Conference on Machine Learning*, pp. 555–563, 2016.

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Ron Kohavi. Scaling up the accuracy of Naive-Bayes classifiers: A decision-tree hybrid. In *KDD*, volume 96, pp. 202–207, 1996.

Yun Liu, Krishna Gadepalli, Mohammad Norouzi, George E Dahl, Timo Kohlberger, Aleksey Boyko, Subhashini Venugopalan, Aleksei Timofeev, Philip Q Nelson, Greg S Corrado, et al. Detecting cancer metastases on gigapixel pathology images. *arXiv preprint arXiv:1703.02442*, 2017.

H Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private language models without losing accuracy. *arXiv preprint arXiv:1710.06963*, 2017.

Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *Foundations of Computer Science, 2007. FOCS'07. 48th Annual IEEE Symposium on*, pp. 94–103. IEEE, 2007.

Ilya Mironov. Rényi differential privacy. In *Proceedings of 30th IEEE Computer Security Foundations Symposium, 21-25 August 2017, Santa Barbara, CA*, pp. 263–275, 2017.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *arXiv preprint arXiv:1704.03976*, 2017.

Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *Proceedings of the 2008 IEEE Symposium on Security and Privacy*, pp. 111–125. IEEE, 2008.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, pp. 5, 2011.

Nicolas Papernot, Martín Abadi, Úlfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. In *Proceedings of the 5th International Conference on Learning Representations*, 2017.

Manas Pathak, Shantanu Rane, and Bhiksha Raj. Multiparty differential privacy via aggregation of locally trained classifiers. In *Advances in Neural Information Processing Systems*, pp. 1876–1884, 2010.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems*, pp. 2234–2242, 2016.

Shuang Song, Kamalika Chaudhuri, and Anand D Sarwate. Stochastic gradient descent with differentially private updates. In *Global Conference on Signal and Information Processing (GlobalSIP), 2013 IEEE*, pp. 245–248. IEEE, 2013.

Thomas Steinke and Jonathan Ullman. Tight lower bounds for differentially private selection. In *FOCS 2017*, 2017.

Beata Strack, Jonathan P DeShazo, Chris Gennings, Juan L Olmo, Sebastian Ventura, Krzysztof J Cios, and John N Clore. Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed research international*, 2014, 2014.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *Proceedings of the 2nd International Conference on Learning Representations*, 2014.

Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction APIs. In *USENIX Security Symposium*, pp. 601–618, 2016.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *Proceedings of the 5th International Conference on Learning Representations*, 2017.

A APPENDIX: BACKGROUND ON PATE

Here, we provide an overview of the PATE framework. Readers familiar with the approach may skip to Section 3. To protect the privacy of training data during learning, PATE transfers knowledge from an ensemble of teacher models trained on partitions of the data to a student model. Privacy guarantees may be understood intuitively and expressed rigorously in terms of differential privacy.

Illustrated in Figure 7, the PATE framework consists of three key parts: (1) an ensemble of n teacher models, (2) an aggregation mechanism and (3) a student model.

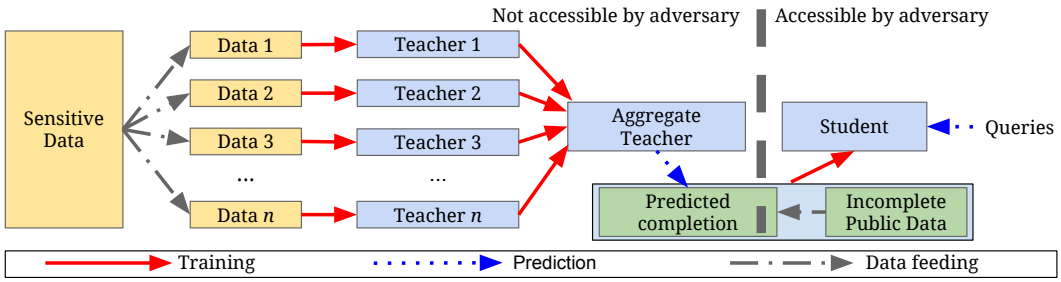

Figure 7: Overview of the approach: (1) an ensemble of teachers is trained on disjoint subsets of the sensitive data, (2) a student model is trained on public data labeled using the ensemble. Figure reproduced from Papernot et al. (2017) with permission.

Teacher models: Training each teacher on a *partition* of the sensitive data produces n different models solving the same task. At inference, teachers independently predict labels. When there is a strong consensus among teachers, the label they almost all agree on does not depend on the model learned by any given teacher. Hence, this collective decision is intuitively private with respect to any given training point—because such a point could have been included only in one of the teachers' training set.

Aggregation mechanism: The aggregation mechanism of the original PATE framework counts votes assigned to each class, adds carefully calibrated Laplacian noise to the resulting vote histogram, and outputs the class with the most noisy votes as the ensemble's prediction. This mechanism is referred to as the max-of-Laplacian mechanism going forward.

For samples x and classes $1, \ldots, m$, let $f_j(x) \in [m]$ denote the j-th teacher model's prediction and n_i denote the vote count for the i-th class (i.e., $n_i = |f_j(x) = i|$). The output of the mechanism is $\mathcal{A}(x) := \text{argmax}_i \left(n_i(x) + \text{Lap}\left(1/\gamma\right)\right)$. Through a rigorous analysis of this mechanism, the PATE framework provides a differentially private API: the privacy cost of each aggregated prediction made by the teacher ensemble is known.

Student model: PATE's final step involves the training of a student model by knowledge transfer from the teacher ensemble using access to public—but *unlabeled*—data. To limit the privacy cost of labeling them, queries are only made to the aggregation mechanism for a subset of public data to train the student in a semi-supervised way using a fixed number of queries. The authors note that every additional ensemble prediction increases the privacy cost spent and thus cannot work with unbounded queries. Fixed queries fixes privacy costs as well as diminishes the avlue of attacks analyzing model parameters to recover training data (Zhang et al., 2017). The student only sees public data and privacy-preserving labels.

B APPENDIX: RELATED WORK

Differential privacy is by now the gold standard of privacy. It offers a rigorous framework whose threat model makes few assumptions about the adversary's capabilities, allowing differentially private algorithms to effectively cope against strong adversaries. This is not the case of all privacy definitions, as demonstrated by successful attacks against anonymization techniques (Aggarwal, 2005; Narayanan & Shmatikov, 2008; Bindschaedler et al., 2017).

The first learning algorithms adapted to provide differential privacy with respect to their training data were often linear and convex (Pathak et al., 2010; Chaudhuri et al., 2011; Song et al., 2013; Bassily et al., 2014; Hamm et al., 2016). More recently, successful developments in deep learning called for differentially private stochastic gradient descent algorithms (Abadi et al., 2016), some of which have been tailored to learn in federated (McMahan et al., 2017) settings.

Differentially private selection mechanisms like GNMax are commonly used in hypothesis testing, frequent itemset mining, and as building blocks of more complicated private mechanisms. The most commonly used differentially private selection mechanisms are exponential mechanism(McSherry & Talwar, 2007) and LNMax(Bhaskar et al., 2010). There are also works on lower bound for such problem (Steinke & Ullman, 2017).

The Confident and Interactive Aggregator proposed in our work use the intuition that selecting samples under certain constraints could result in better training than using samples uniformly at random. In Machine Learning Theory, active learning (Cohn et al., 1994) has been shown to allow learning from fewer labeled examples than the passive case (see e.g. Hanneke (2014)). Similarly, in model stealing (Tramèr et al., 2016), a goal is to learn a model from limited access to a teacher network.

C Appendix: Rényi Differential Privacy and Privacy Proofs

C.1 Rényi Differential Privacy

Papernot et al. (2017) note that the natural approach to bound PATE's privacy loss—by first having a bound for the privacy cost of each label queried and then using strong composition (Dwork et al., 2010) to derive the total cost—yields loose privacy guarantees. Instead, their approach uses *data-dependent* privacy analysis. This takes advantage of the fact that when the consensus among the teachers is very strong, the plurality outcome has overwhelming likelihood leading to a very small privacy cost whenever the consensus occurs. To capture this effect quantitatively, Papernot et al. (2017) rely on the *moments accountant*, introduced by Abadi et al. (2016) and building on previous work (Bun & Steinke, 2016; Dwork & Rothblum, 2016).

In this section, we restate the privacy analysis framework of PATE in the language of Rényi Differential Privacy or RDP (Mironov, 2017). RDP generalizes pure differential privacy ($\delta = 0$) and is closely related to the moments accountant. We choose to use RDP as a more natural analysis framework when dealing with our mechanisms that use Gaussian noise. Defined below, the RDP of a mechanism is stated in terms of the Rényi divergence.

Definition 4 (Rényi Divergence)**.** *The Rényi divergence of order λ between two distributions P and Q is defined as:*

$$D_\lambda(P\|Q) \triangleq \frac{1}{\lambda-1}\log \mathbb{E}_{x\sim Q}\left[\left(P(x)/Q(x)\right)^\lambda\right] = \frac{1}{\lambda-1}\log \mathbb{E}_{x\sim P}\left[\left(P(x)/Q(x)\right)^{\lambda-1}\right].$$

Definition 5 (Rényi Differential Privacy (RDP))**.** *A randomized mechanism $\mathcal{M}$ is said to guarantee (λ,ε)-RDP with $\lambda \geq 1$ if for any neighboring datasets D and D' and any auxiliary information aux,*

$$D_\lambda(\mathcal{M}(D,aux)\|\mathcal{M}(D',aux)) = \frac{1}{\lambda-1}\log \mathbb{E}_{x\sim\mathcal{M}(D,aux)}\left[\left(\frac{\mathbf{Pr}\left[\mathcal{M}(D,aux)=x\right]}{\mathbf{Pr}\left[\mathcal{M}(D',aux)=x\right]}\right)^{\lambda-1}\right] \leq \varepsilon.$$

RDP generalizes pure differential privacy in the sense that ε-differential is equivalent to (∞,ε)-RDP. Mironov (2017) proves the following key facts that allow easy composition of RDP guarantees and their conversion to (ε,δ)-differential privacy bounds.

Theorem 6 (Composition)**.** *If a mechanism $\mathcal{M}$ consists of a sequence of adaptive mechanisms $\mathcal{M}_1,\ldots,\mathcal{M}_k$ such that for any $i \in [k]$, $\mathcal{M}_i$ guarantees (λ,ε_i)-RDP, then $\mathcal{M}$ guarantees $(\lambda,\sum_{i=1}^k \varepsilon_i)$-RDP.*

Theorem 7 (From RDP to DP)**.** *If a mechanism $\mathcal{M}$ guarantees (λ,ε)-RDP, then $\mathcal{M}$ guarantees $(\varepsilon+\frac{\log 1/\delta}{\lambda-1},\delta)$-differential privacy for any $\delta\in(0,1)$.*

While both (ε, δ)-differential privacy and RDP are relaxations of pure ε-differential privacy, the two main advantages of RDP are as follows. First, it composes nicely; second, it captures the privacy guarantee of Gaussian noise in a much cleaner manner compared to (ε, δ)-differential privacy. This lets us do a careful privacy analysis of the GNMax mechanism as stated in Theorem 2. While the analysis of Papernot et al. (2017) leverages the first aspect of such frameworks with the Laplace noise (LNMax mechanism), our analysis of the GNMax mechanism relies on both.

C.2 Privacy Proofs

In this section, we provide the proofs of Theorem 2 and Lemma 3. Moreover, we present Lemma 9, which provides optimal values of μ_1 and μ_2 to apply towards Theorem 2 for the GNMax mechanism. We start off with a lemma about the Rényi differential privacy guarantee of the GNMax.

Lemma 8. *The GNMax aggregator $\mathcal{A}_\sigma$ guarantees $\left(\lambda, \lambda/\sigma^2\right)$-RDP for all $\lambda \geq 1$.*

Proof. The result follows from observing that $\mathcal{A}_\sigma$ can be decomposed into applying the argmax operator to a noisy histogram resulted from adding Gaussian noise to each dimension of the original histogram. The Gaussian mechanism satisfies $(\lambda, \lambda/2\sigma^2)$-RDP (Mironov, 2017), and since each teacher may change two counts (incrementing one and decrementing the other), the overall RDP guarantee is as claimed. ∎

Lemma 3. *For any $i^* \in [m]$, we have $\mathbf{Pr}\left[\mathcal{A}_\sigma(D) \neq i^*\right] \leq \frac{1}{2}\sum_{i \neq i^*} \mathrm{erfc}\left(\frac{n_{i^*} - n_i}{2\sigma}\right)$.*

Proof. For any $i^* \in [n]$, we have

$$\begin{aligned}
\mathbf{Pr}(\mathcal{M}(D) \neq i^*) = \mathbf{Pr}\left[\exists i, n_i + Z_i > n_{i^*} + Z_{i^*}\right] &\leq \sum_{i \neq i^*} \mathbf{Pr}\left[n_i + Z_i > n_{i^*} + Z_{i^*}\right] \\
&= \sum_{i \neq i^*} \mathbf{Pr}\left[Z_i - Z_{i^*} > n_{i^*} - n_i\right] \\
&= \sum_{i \neq i^*} \frac{1}{2}\left(1 - \mathrm{erf}\left(\frac{n_{i^*} - n_i}{2\sigma}\right)\right).
\end{aligned}$$

where the last equality follows from the fact that $Z_i - Z_j$ is a Gaussian random variable with mean zero and variance $2\sigma^2$. ∎

We now present a precise statement of Theorem 2.

Theorem 2. *Let $\mathcal{M}$ be a randomized algorithm with (μ_1, ε_1)-RDP and (μ_2, ε_2)-RDP guarantees and suppose that there exists* a likely *outcome i^* given a dataset D such that $\tilde{q} \geq \mathbf{Pr}\left[\mathcal{M}(D) \neq i^*\right]$. Let $\lambda \leq \mu_1$ and $\tilde{q} \leq e^{(\mu_2 - 1)\varepsilon_2} / \left(\frac{\mu_1}{\mu_1 - 1} \cdot \frac{\mu_2}{\mu_2 - 1}\right)^{\mu_2}$. Then, for any auxiliary information aux and any neighboring dataset D' of D, we have:*

$$D_\lambda(\mathcal{M}(D, aux) \| \mathcal{M}(D', aux)) \leq \frac{1}{\lambda - 1} \log\Big((1 - \tilde{q}) \cdot \boldsymbol{A}(\tilde{q}, \mu_2, \varepsilon_2)^{\lambda - 1} + \tilde{q} \cdot \boldsymbol{B}(\tilde{q}, \mu_1, \varepsilon_1)^{\lambda - 1} \Big) \tag{2}$$

where $\boldsymbol{A}(\tilde{q}, \mu_2, \varepsilon_2) \triangleq (1 - \tilde{q}) / \left(1 - (\tilde{q} e^{\varepsilon_2})^{\frac{\mu_2 - 1}{\mu_2}}\right)$ and $\boldsymbol{B}(\tilde{q}, \mu_1, \varepsilon_1) \triangleq e^{\varepsilon_1} / \tilde{q}^{\frac{1}{\mu_1 - 1}}$.

Proof. Before we proceed to the proof, we introduce some simplifying notation. For a randomized mechanism $\mathcal{M}$, neighboring datasets D and D', and auxiliary information aux, we define

$$\begin{aligned}
\boldsymbol{\beta}_{\mathcal{M}}(\lambda; aux, D, D') &\triangleq D_\lambda(\mathcal{M}(D, aux) \| \mathcal{M}(D', aux)) \\
&= \frac{1}{\lambda - 1} \log \mathbb{E}_{x \sim \mathcal{M}(D, aux)} \left[\left(\frac{\mathbf{Pr}\left[\mathcal{M}(D, aux) = x\right]}{\mathbf{Pr}\left[\mathcal{M}(D', aux) = x\right]} \right)^{\lambda - 1} \right].
\end{aligned}$$

As the proof involves working with the RDP bounds in the exponent, we set $\zeta_1 \triangleq e^{\varepsilon_1(\mu_1 - 1)}$ and $\zeta_2 \triangleq e^{\varepsilon_2(\mu_2 - 1)}$.

Finally, we define the following shortcuts:

$$q_i \triangleq \mathbf{Pr}\left[\mathcal{M}(D) = i\right] \text{ and } q \triangleq \sum_{i \neq i^*} q_i = \mathbf{Pr}\left[\mathcal{M}(D) \neq i^*\right],$$

$$p_i \triangleq \mathbf{Pr}\left[\mathcal{M}(D') = i\right] \text{ and } p \triangleq \sum_{i \neq i^*} p_i = \mathbf{Pr}\left[\mathcal{M}(D') \neq i^*\right],$$

and note that $q \leq \tilde{q}$.

From the definition of Rényi differential privacy of order μ_1, we have:

$$\exp\left(\boldsymbol{\beta}_{\mathcal{M}}(\mu_1; aux, D, D')\right) = \left(\frac{(1-q)^{\mu_1}}{(1-p)^{\mu_1-1}} + \sum_{i \neq i*} \frac{q_i^{\mu_1}}{p_i^{\mu_1-1}}\right)^{1/(\mu_1-1)} \leq \exp\left(\varepsilon_1\right)$$

$$\implies \sum_{i>1} \frac{q_i^{\mu_1}}{p_i^{\mu_1-1}} = \sum_{i>1} q_i \left(\frac{q_i}{p_i}\right)^{\mu_1-1} \leq \zeta_1. \tag{3}$$

Since $\mu_1 > \lambda$, $f(x) \triangleq x^{\frac{\mu_1-1}{\lambda-1}}$ is convex. Applying Jensen's Inequality we have the following:

$$\left(\frac{\sum_{i \neq i^*} q_i \left(\frac{q_i}{p_i}\right)^{\lambda-1}}{q}\right)^{\frac{\mu_1-1}{\lambda-1}} \leq \frac{\sum_{i \neq i^*} q_i \left(\frac{q_i}{p_i}\right)^{\mu_1-1}}{q}$$

$$\implies \sum_{i \neq i^*} q_i \left(\frac{q_i}{p_i}\right)^{\lambda-1} \leq q \left(\frac{\sum_{i \neq i^*} q_i \left(\frac{q_i}{p_i}\right)^{\mu_1-1}}{q}\right)^{\frac{\lambda-1}{\mu_1-1}}$$

$$\overset{(3)}{\implies} \sum_{i \neq i^*} q_i \left(\frac{q_i}{p_i}\right)^{\lambda-1} \leq {\zeta_1}^{\frac{\lambda-1}{\mu_1-1}} \cdot q^{1-\frac{\lambda-1}{\mu_1-1}}. \tag{4}$$

Next, by the bound at order μ_2, we have:

$$\exp\left(\boldsymbol{\beta}_{\mathcal{M}}(\mu_2; aux, D', D)\right) = \left(\frac{(1-p)^{\mu_2}}{(1-q)^{\mu_2-1}} + \sum_{i \neq i^*} \frac{p_i^{\mu_2}}{q_i^{\mu_2-1}}\right)^{1/(\mu_2-1)} \leq \exp\left(\varepsilon_2\right)$$

$$\implies \frac{(1-p)^{\mu_2}}{(1-q)^{\mu_2-1}} + \sum_{i \neq i^*} \frac{p_i^{\mu_2}}{q_i^{\mu_2-1}} \leq \zeta_2.$$

By the data processing inequality of Rényi divergence, we have

$$\frac{(1-p)^{\mu_2}}{(1-q)^{\mu_2-1}} + \frac{p^{\mu_2}}{q^{\mu_2-1}} \leq \zeta_2,$$

which implies $\frac{p^{\mu_2}}{q^{\mu_2-1}} \leq \zeta_2$ and thus

$$p \leq \left(q^{\mu_2-1}\zeta_2\right)^{\frac{1}{\mu_2}}. \tag{5}$$

Combining (4) and (5), we can derive a bound at λ.

$$\exp\left(\boldsymbol{\beta}_{\mathcal{M}}(\lambda, aux, D, D')\right) = \left(\frac{(1-q)^{\lambda}}{(1-p)^{\lambda-1}} + \sum_{i \neq i^*} \frac{q_i^{\lambda}}{p_i^{\lambda-1}}\right)^{1/(\lambda-1)}$$

$$\leq \left(\frac{(1-q)^{\lambda}}{\left(1-\left(q^{\mu_2-1}\zeta_2\right)^{\frac{1}{\mu_2}}\right)^{\lambda-1}} + {\zeta_1}^{\frac{\lambda-1}{\mu_1-1}} \cdot q^{1-\frac{\lambda-1}{\mu_1-1}}\right)^{1/(\lambda-1)}. \tag{6}$$

Although Equation (6) is very close to the corresponding statement in the theorem's claim, one subtlety remains. The bound (6) applies to the exact probability $q = \mathbf{Pr}\left[\mathcal{M}(D) \neq i^*\right]$. In the theorem statement, and in practice, we can only derive an upper bound $\tilde{q}$ on $\mathbf{Pr}\left[\mathcal{M}(D) \neq i^*\right]$. The last step of the proof requires showing that the expression in Equation (6) is monotone in the range of values of q that we care about.

Monotonicity of the bound. Below we demonstrate that for the functions

$$f_1(x) \triangleq \frac{(1-x)^\lambda}{\left(1-(x^{\mu_2-1}\zeta_2)^{\frac{1}{\mu_2}}\right)^{\lambda-1}} \qquad \text{and} \qquad f_2(x) \triangleq {\zeta_1}^{\frac{\lambda-1}{\mu_1-1}} \cdot x^{1-\frac{\lambda-1}{\mu_1-1}},$$

defined on $x \in [0,1]$ as above, $f_1(x) + f_2(x)$ is increasing in $\left[0, \zeta_2 / \left(\frac{\mu_1}{\mu_1-1} \cdot \frac{\mu_2}{\mu_2-1}\right)^{\mu_2}\right]$.

Taking the derivative of $f_1(x)$, we have:

$$\begin{aligned} f_1'(x) &= \frac{-\lambda(1-x)^{\lambda-1}(1-(x^{\mu_2-1}\zeta_2)^{\frac{1}{\mu_2}})^{\lambda-1}}{(1-(x^{\mu_2-1}\zeta_2)^{\frac{1}{\mu_2}})^{2\lambda-2}} \\ &\quad + \frac{(1-x)^\lambda(\lambda-1)(1-(x^{\mu_2-1}\zeta_2)^{\frac{1}{\mu_2}})^{\lambda-2}{\zeta_2}^{\frac{1}{\mu_2}} \cdot \frac{\mu_2-1}{\mu_2} \cdot x^{-\frac{1}{\mu_2}}}{(1-(x^{\mu_2-1}\zeta_2)^{\frac{1}{\mu_2}})^{2\lambda-2}} \\ &= \frac{(1-x)^{\lambda-1}}{(1-(x^{\mu_2-1}\zeta_2)^{\frac{1}{\mu_2}})^{\lambda-1}} \left(-\lambda + (\lambda-1)\left(1-\frac{1}{\mu_2}\right)\frac{1-x}{1-(x^{\mu_2-1}\zeta_2)^{\frac{1}{\mu_2}}}\left(\frac{\zeta_2}{x}\right)^{\frac{1}{\mu_2}}\right). \end{aligned}$$

We intend to show that:

$$f_1'(x) \geq -\lambda + (\lambda-1)\left(1-\frac{1}{\mu_2}\right)\left(\frac{\zeta_2}{x}\right)^{\frac{1}{\mu_2}}. \tag{7}$$

For $x \in \left[0, \zeta_2 / \left(\frac{\mu_1}{\mu_1-1} \cdot \frac{\mu_2}{\mu_2-1}\right)^{\mu_2}\right]$ and $y \in [1, \infty)$, define $g(x,y)$ as:

$$g(x,y) \triangleq -\lambda \cdot y^{\lambda-1} + (\lambda-1)\left(1-\frac{1}{\mu_2}\right)\left(\frac{\zeta_2}{x}\right)^{\frac{1}{\mu_2}} y^\lambda.$$

We claim that $g(x,y)$ is increasing in y and therefore $g(x,y) \geq g(x,1)$, and prove it by showing the partial derivative of $g(x,y)$ with respect to y is non-negative. Take a derivative with respect to y as:

$$\begin{aligned} g_y'(x,y) &= -\lambda(\lambda-1)y^{\lambda-2} + \lambda(\lambda-1)\left(1-\frac{1}{\mu_2}\right)\left(\frac{\zeta_2}{x}\right)^{\frac{1}{\mu_2}} y^{\lambda-1} \\ &= \lambda(\lambda-1)y^{\lambda-2}\left(-1 + \left(1-\frac{1}{\mu_2}\right)\left(\frac{\zeta_2}{x}\right)^{\frac{1}{\mu_2}} y\right). \end{aligned}$$

To see why $g_y'(x,y)$ is non-negative in the respective ranges of x and y, note that:

$$\begin{aligned} x \leq \zeta_2 / \left(\frac{\mu_1}{\mu_1-1} \cdot \frac{\mu_2}{\mu_2-1}\right)^{\mu_2} &\implies x \leq \zeta_2 / \left(\frac{\mu_2}{\mu_2-1}\right)^{\mu_2} \\ &\implies 1 \leq \frac{\zeta_2}{x} \cdot \left(\frac{\mu_2-1}{\mu_2}\right)^{\mu_2} \\ &\implies 1 \leq \frac{\mu_2-1}{\mu_2}\left(\frac{\zeta_2}{x}\right)^{\frac{1}{\mu_2}} \\ &\implies 1 \leq \frac{\mu_2-1}{\mu_2}\left(\frac{\zeta_2}{x}\right)^{\frac{1}{\mu_2}} y \qquad \text{(as } y \geq 1\text{)} \\ &\implies 0 \leq -1 + \frac{\mu_2-1}{\mu_2}\left(\frac{\zeta_2}{x}\right)^{\frac{1}{\mu_2}} y \\ &\implies 0 \leq g_y'(x,y). \qquad \text{(in the resp. range of } x \text{ and } y\text{)} \end{aligned}$$

Consider $\frac{1-x}{1-(x^{\mu_2-1}\zeta_2)^{1/\mu_2}}$. Since $\zeta_2 \geq 1$ and $x \leq 1$, we have $x \leq \zeta_2$ and hence

$$\frac{1-x}{1-(x^{\mu_2-1}\zeta_2)^{\frac{1}{\mu_2}}} \geq \frac{1-x}{1-(x^{\mu_2-1}x)^{\frac{1}{\mu_2}}} = 1.$$

Therefore we can set $y = \frac{1-x}{1-(x^{\mu_2-1}\zeta_2)^{1/\mu_2}}$ and apply the fact that $g(x,y) \geq g(x,1)$ for all $y \geq 1$ to get

$$f_1'(x) \geq -\lambda + (\lambda-1)\left(1-\frac{1}{\mu_2}\right)\left(\frac{\zeta_2}{x}\right)^{\frac{1}{\mu_2}},$$

as required by (7).

Taking the derivative of $f_2(x)$, we have:

$$f_2'(x) = \zeta_1{}^{\frac{\lambda-1}{\mu_1-1}} \cdot \left(1-\frac{\lambda-1}{\mu_1-1}\right) x^{-\frac{\lambda-1}{\mu_1-1}} = \left(\frac{\zeta_1}{x}\right)^{\frac{\lambda-1}{\mu_1-1}} \left(1-\frac{\lambda-1}{\mu_1-1}\right) \geq 1-\frac{\lambda-1}{\mu_1-1}.$$

Combining the two terms together, we have:

$$\begin{aligned} f'(x) &\geq -\lambda + (\lambda-1)\left(1-\frac{1}{\mu_2}\right)\left(\frac{\zeta_2}{x}\right)^{\frac{1}{\mu_2}} + 1 - \frac{\lambda-1}{\mu_1-1} \\ &= (\lambda-1)\left(-\frac{\mu_1}{\mu_1-1} + \frac{\mu_2-1}{\mu_2}\left(\frac{\zeta_2}{x}\right)^{\frac{1}{\mu_2}}\right). \end{aligned}$$

For $f'(x)$ to be non-negative we need:

$$\begin{aligned} &-\frac{\mu_1}{\mu_1-1} + \frac{\mu_2-1}{\mu_2}\left(\frac{\zeta_2}{x}\right)^{\frac{1}{\mu_2}} \geq 0 \\ \iff &\left(\frac{\mu_1}{\mu_1-1}\cdot\frac{\mu_2}{\mu_2-1}\right)^{\mu_2} \leq \frac{\zeta_2}{x}. \end{aligned}$$

So $f(x)$ is increasing for $x \in \left[0, \zeta_2/\left(\frac{\mu_1}{\mu_1-1}\cdot\frac{\mu_2}{\mu_2-1}\right)^{\mu_2}\right]$. This means for $q \leq \tilde{q} \leq \zeta_2/\left(\frac{\mu_1}{\mu_1-1}\cdot\frac{\mu_2}{\mu_2-1}\right)^{\mu_2}$, we have $f(q) \leq f(\tilde{q})$. This completes the proof of the theorem. ∎

Theorem 2 yields data-dependent Rényi differential privacy bounds for *any* value of μ_1 and μ_2 larger than λ. The following lemma simplifies this search by calculating optimal higher moments μ_1 and μ_2 for the GNMax mechanism with variance σ^2.

Lemma 9. *When applying Theorem 2 and Lemma 8 for GNMax with Gaussian of variance σ^2, Eq.* (2) *is minimized at*

$$\mu_2 = \sigma \cdot \sqrt{\log(1/\tilde{q})}, \text{ and } \mu_1 = \mu_2 + 1.$$

Proof. We can minimize both terms in (2) independently. To minimize the first term in (2), we minimize $(\tilde{q}e^{\varepsilon_2})^{1-1/\mu_2}$ by considering logarithms:

$$\begin{aligned} \log\left\{(\tilde{q}e^{\varepsilon_2})^{1-1/\mu_2}\right\} &= \log\left\{\tilde{q}^{1-\frac{1}{\mu_2}}\exp\left(\frac{\mu_2-1}{\sigma^2}\right)\right\} \\ &= \left(1-\frac{1}{\mu_2}\right)\cdot\log\tilde{q} + \frac{\mu_2-1}{\sigma^2} \\ &= \frac{1}{\mu_2}\log\frac{1}{\tilde{q}} + \frac{\mu_2}{\sigma^2} - \frac{1}{\sigma^2} - \log\frac{1}{\tilde{q}}, \end{aligned}$$

which is minimized at $\mu_2 = \sigma \cdot \sqrt{\log(1/\tilde{q})}$.

To minimize the second term in (2), we minimize $e^{\varepsilon_1}/\tilde{q}^{1/(\mu_1-1)}$ as follows:

$$\begin{aligned}\log\left\{\frac{e^{\varepsilon_1}}{\tilde{q}^{1/(\mu_1-1)}}\right\} &= \log\left\{\tilde{q}^{-1/(\mu_1-1)}\exp\left(\frac{\mu_1}{\sigma^2}\right)\right\} \\ &= \frac{\mu_1}{\sigma^2} + \frac{1}{\mu_1-1}\log\frac{1}{\tilde{q}} \\ &= \frac{1}{\sigma^2} + \frac{\mu_1-1}{\sigma^2} + \frac{1}{\mu_1-1}\log\frac{1}{\tilde{q}},\end{aligned}$$

which is minimized at $\mu_1 = 1 + \sigma \cdot \sqrt{\log(1/\tilde{q})}$ completing the proof. ∎