Machine learning models are known to memorize the unique properties of individual data points in a training set. This memorization capability can be exploited by several types of attacks to infer information about the training data, most notably, membership inference attacks. In this paper, we propose an approach based on information leakage for guaranteeing membership privacy. Specifically, we propose to use a conditional form of the notion of maximal leakage to quantify the information leaking about individual data entries in a dataset, i.e., the entrywise information leakage. We apply our privacy analysis to the Private Aggregation of Teacher Ensembles (PATE) framework for privacy-preserving classification of sensitive data and prove that the entrywise information leakage of its aggregation mechanism is Schur-concave when the injected noise has a log-concave probability density. The Schur-concavity of this leakage implies that increased consensus among teachers in labeling a query reduces its associated privacy cost. Finally, we derive upper bounds on the entrywise information leakage when the aggregation mechanism uses Laplace distributed noise.

1 Introduction

In recent years, many useful machine learning applications have emerged that require training on sensitive data. Such applications span across a diverse range of fields such as medical imaging [1], rumor identification in social media [2], or financial fraud detection [3]. While all machine learning applications by definition reveal some information about the training data, privacy concerns arise when machine learning models memorize properties that are unique to individual data entries. In fact, a variety of privacy attacks have demonstrated that it is indeed possible to exploit this “memorization” capability of models to infer information about data entries in the training set [4].

Arguably, the simplest type of privacy attacks against machine learning models is membership inference attacks in which an adversary infers whether or not a certain data point was used in the training [5][6]. In response to such attacks, a number of mitigation techniques have been proposed in the literature, with differential privacy-based methods being the most commonly studied. Differential privacy [7] provides provable and operationally meaningful privacy guarantees, and by definition neutralizes membership inference attacks. Roughly speaking, differential privacy ensures that all datasets differing in only one entry (i.e., adjacent datasets) produce an output with similar probabilities. Moreover, it has several useful properties, such as satisfying data-processing inequalities and composition theorems [7].

The standard definition of differential privacy (i.e., pure differential privacy) uses a parameter $\epsilon$ to define a multiplicative upper bound on the changes in the probability of an output for all adjacent datasets in the input [8]. However, this definition is known to be very strict, and has limited applicability. As such, several relaxations of differential privacy have been proposed, the most notable of which is $(\epsilon, \delta)$-differential privacy [9]. A common interpretation of $(\epsilon, \delta)$-differential privacy is that the guarantees of $\epsilon$-differential privacy hold except with probability $\delta$. Thus, it provides the necessary flexibility for studying a larger class of privacy-preserving mechanisms such as the Gaussian mechanism [9].
Despite the advantages of $(\epsilon, \delta)$-differential privacy, one should note that its privacy guarantees are qualitatively different from those of pure differential privacy (see [10] for illustrative examples). On this account, recently Rényi differential privacy [10] was proposed as an alternative relaxation of pure differential privacy. While Rényi differential privacy satisfies the same useful properties as pure differential privacy, it does not offer any intuitive operational meaning, and its privacy guarantees are usually translated into $(\epsilon, \delta)$-differential privacy for interpretation.

In this paper, we propose to use (a conditional form of) the notion of maximal leakage [11] to measure the amount of information leaking about any single data entry in a dataset, i.e., the entrywise information leakage. Maximal leakage [11] is an operationally meaningful privacy metric that captures the inference capabilities of an adversary trying to deduce some information about the input data by observing the output. Specifically, maximal leakage quantifies the maximal gain in an adversary’s ability to correctly guess any arbitrary discrete function of the input data by observing the output (as opposed to making a guess with no observations). Note that the original definition of maximal leakage quantifies the information leaking about the whole dataset, whereas we are interested in measuring the information leaking about single data entries in the dataset. As such, similar to [12], we consider an adversary who knows the values of all the entries in the dataset, except for a single data entry of interest. Intuitively, in this setup, observations only convey the unique information contributed by the unknown data entry since all other entries are already known to the adversary. To quantify this entrywise information leakage, we propose a conditional form of maximal leakage, namely the pointwise conditional maximal leakage. Then, by allowing the unknown entry to be any of the entries in the dataset, we can derive upper bounds on the entrywise information leakage, and provide meaningful worst-case privacy guarantees.

Maximal leakage satisfies several useful properties, most notably a data-processing inequality and a composition lemma [11]. The data-processing inequality ensures that no manipulation of the output can increase the information leakage, while the composition property characterizes the information leaked through multiple independent observations. Here, we show that the same properties hold for pointwise conditional maximal leakage, rendering it suitable for privacy analysis of more complex information systems.

We apply our privacy analysis to the Private Aggregation of Teacher Ensembles (PATE) framework [13][14]. PATE is a general framework for privacy-preserving classification of sensitive data, and operates by transferring the knowledge of an ensemble of models (called teachers) trained on disjoint partitions of the sensitive data to a student classifier. Specifically, the student is trained using a public unlabelled dataset which will be labelled by the teachers through an aggregation mechanism. The aggregation mechanism is essentially the Report-Noisy-Max mechanism [7] which adds noise to the teachers’ predictions to enable derivation of privacy guarantees.

PATE has several advantages as a privacy-preserving machine learning framework. First, the privacy guarantees result solely from the aggregation mechanism and are agnostic to the specific machine learning techniques used by each teacher. This is because the modular structure of PATE enables us to invoke the data-processing inequality to uncouple the information leaked through the training and aggregation, and guarantee that the overall leakage is less than both. Second, PATE lends itself well to distributed learning by allowing data owners to separately train their own predictors, hence mitigating the need for centralized storage of the sensitive data. Finally, the aggregation mechanism induces a favorable synergy between privacy and accuracy such that increased agreement among the teachers in labelling a query lowers its associated privacy cost. This synergy is one of the main focuses of this paper, and will be extensively studied.

The privacy guarantees established by PATE are characterized in [13][14] in terms of differential privacy, and results from experiments are reported. Here, we will analyze the privacy of the framework in terms of the entrywise information leakage while our goal is to gain a deeper understanding of the operation of the framework, especially the Report-Noisy-Max mechanism used for aggregating teachers’ predictions. As [13][14] present a thorough experimental study, here we refrain from repeating the experiments but focus on giving a rigorous analysis of the framework to explain properties such as the privacy-accuracy synergy.

1.1 Contributions

Our contributions can be summarized as follows:

i) **Introducing pointwise conditional maximal leakage.** We approach membership privacy from a novel angle by studying the information leakage of individual data entries in a database. We begin by deriving a data-processing inequality and a composition lemma for the pointwise conditional maximal leakage, and then apply them to the problem of studying the entrywise information leakage in PATE.

ii) **Proving the privacy-accuracy synergy in PATE.** We show that the entrywise information leakage of the aggregation mechanism in PATE (i.e., the Report-Noisy-Max mechanism) is Schur-concave [15][16] when the injected noise has a log-concave [17][18] probability density. As we will see, this implies that increased consensus among
teachers lowers the privacy cost of labelling a query. Note that many commonly used probability distributions including the Laplace and Gaussian distributions are log-concave rendering this result fairly general.

iii) **Deriving membership privacy guarantees for PATE with Laplace noise.** We derive upper bounds on the entrywise information leakage when the noise injected in the aggregation mechanism has Laplace distribution.

1.2 Other Related Work

**Information leakage metrics.** In recent years, a large body of work has been dedicated to studying various information-theoretic privacy metrics. Most notably, mutual information has been frequently proposed and studied as such a metric (see e.g., [19–21]) by appealing to its operational meaning in communication theory. Similarly, in [22] another information-theoretic quantity namely the total variation distance is studied as a privacy metric in an information disclosure scenario. More closely related to our approach, several information leakage metrics have recently emerged that aim to capture the inference abilities of an adversary trying to guess a secret. For instance, [23] proposes to use the probability of correctly guessing the secret as a privacy metric. In [24] a class of tunable loss functions are introduced to capture a range of adversarial objectives, e.g., refining a belief or guessing the most likely value for the secret. Other methods include posing the privacy problem as a hypothesis test, e.g., in [25]. It is worth mentioning that the majority of the proposed privacy metrics have no clear operational meaning which limits their applicability. A systematic survey of privacy metrics is provided in [26].

**Privacy-preserving machine learning.** Several centralized and decentralized solutions have been proposed in the literature that provide privacy guarantees in terms of differential privacy. To give a few examples, [27] proposes a collaborative framework for privacy-preserving deep learning where the guarantees of differential privacy are obtained by perturbing the gradients. Another example is [28] where the privacy analysis of gradient perturbations are improved by introducing the *moments accountant* framework. Other methods include privacy-preserving logistic regression [29, 30], support vector machines [31] and empirical risk minimization [32, 33].

1.3 Outline of the Paper

The rest of the paper is organized as follows: in Section 2 we will review the definition of maximal leakage and give a short summary of the operation of the PATE framework. In Section 3 we will present the definition of pointwise conditional maximal leakage, and state a few of its key properties. In Section 4 we will present our privacy analysis of the framework and state our results. Section 5 concludes the paper.

2 Background

Throughout this work, upper-case letters are used to represent discrete random variables, upper-case calligraphic letters represent their corresponding alphabets and lower-case letters represent the elements of the alphabets. Furthermore, \([1, n] = \{1, \ldots, n\}\) denotes the set of integers between one and \(n\), \(|\cdot|\) denotes the cardinality of a set, and \(\log(\cdot)\) denotes the natural logarithm. Finally, all sets considered in this paper are assumed to be finite.

We begin by reviewing a few key concepts.

2.1 Maximal Leakage

Let \(X\) be a random variable representing the data containing sensitive information, and \(Y\) be the publicly observed output of a probability kernel \(P_{Y|X}\) with input \(X\). Suppose that an adversary observes \(Y\) and wishes to guess an arbitrary discrete function of \(X\), denoted by \(U\).

**Definition 1 (Maximal leakage [11])** Suppose \(P_{XY}\) is a joint distribution defined on the alphabets \(X\) and \(Y\). The maximal leakage from \(X\) to \(Y\) is defined as

\[
\mathcal{L}(X \rightarrow Y) := \sup_{U: U \subset X \times Y} \log \frac{\mathbb{P}(U = \hat{U}(Y))}{\max_{u \in U} P_U(u)},
\]

(1)

where \(\hat{U}\) is the optimal estimator (i.e., MAP estimator) taking values from the same alphabet as \(U\).

Maximal leakage quantifies the maximal gain in the adversary’s ability to correctly guess \(U\) after observing \(Y\) (compared to correctly guessing \(U\) with no observations). It is shown in [11] Theorem 1] that for finite alphabets \(X\) and \(Y\), \(\mathcal{L}(X \rightarrow Y)\)
simplifies to
\[
L(X \rightarrow Y) = \log \sum_{y \in Y} \max_{x \in X_x: P_X(x) > 0} P_{Y|X}(y | x).
\] (2)

2.2 The PATE Framework

PATE [13][14] is a general framework for privacy-preserving classification of sensitive data. It operates by transferring the knowledge of an ensemble of classifiers, called teachers, trained on (disjoint) partitions of the sensitive data to a student classifier. More specifically, the PATE framework consists of the following three main components:

Teacher models. A teacher is a classification model trained on one of the disjoint partitions of the training data, and can use any classification algorithm suited for the task. At inference, each teacher predicts a label independently of others, to which we will refer as that teacher’s vote. Thus, partitioning data into \( L \) sets (and correspondingly training \( L \) teachers) produces \( L \) primary votes for predicting the label of any new data point.

Aggregation mechanism. To predict the label of a new data point, the aggregation mechanism (i.e., the Report-Noisy-Max mechanism [7]) constructs the histogram of teachers’ votes, adds calibrated noise to each of the bins, and outputs the class label with the maximum noisy vote as the final aggregate prediction. Note that the overall privacy guarantees of the framework result from the addition of noise in the aggregation mechanism.

Student model. The student model is trained using a public unlabelled dataset which will be labelled by the teachers’ ensemble through the aggregation mechanism. Note that to limit the privacy cost of the overall system, the student must be trained with as few queries to the teachers as possible.

3 Pointwise conditional maximal leakage

In this section, we introduce the notion of pointwise conditional maximal leakage, and state two of its important properties. Recall that maximal leakage is defined in a setup where an adversary wishes to guess an arbitrary discrete function \( U \) of the private input data \( X \) by observing the output \( Y \). Here, we consider the case where the adversary has some a priori knowledge about \( X \). We model this a priori knowledge as the outcome of a random variable, and accordingly define a conditional form of maximal leakage. Consider an adversary that knows the outcome of a random variable \( Z \).

Definition 2 (Pointwise conditional maximal leakage) Suppose \( P_{XYZ} \) is a joint distribution defined on the alphabets \( X, Y \) and \( Z \), and that the value of the random variable \( Z \) is a priori given as \( z \in Z \). The pointwise conditional maximal leakage from \( X \) to \( Y \) given \( Z = z \) is defined as

\[
L(X \rightarrow Y|Z = z) := \sup_{U: U \rightarrow (X, Z) \rightarrow Y} \log \frac{P(U = \hat{U}(Y, Z = z))}{P(U = \hat{U}(Z = z))},
\] (3)

where \( \hat{U} \) is the optimal estimator of \( U \) given \( Y \) and \( Z = z \), and \( \hat{U} \) is optimal estimator of \( U \) given only \( Z = z \).

Proposition 3 For finite alphabets \( X, Y \) and \( Z \), the pointwise conditional maximal leakage can be expressed as

\[
L(X \rightarrow Y|Z = z) = \log \sum_{y \in Y} \max_{x: P_X(x|z) > 0} P_{Y|XZ}(y|x, z).
\] (4)

The proof is given in Appendix A.1. Our definition of conditional maximal leakage differs slightly from the one proposed in [11]. In [11] the leakage is conditioned on the random variable \( Z \) itself, which translates into a maximization over the outcomes of \( Z \) in (4). We, on the other hand, are conditioning the leakage directly on outcomes of \( Z \) since we are interested in characterizing the leakage for all outcomes, not just the one with the highest leakage.

Remark 4 If the Markov chain \( Z \rightarrow X \rightarrow Y \) holds, (4) becomes

\[
L(X \rightarrow Y|Z = z) = \log \sum_{y \in Y} \max_{x: P_X(x|z) > 0} P_{Y|X}(y|x).
\] (5)

Similarly to [11], we now state two important properties of the pointwise conditional maximal leakage: a data-processing inequality and a composition lemma. These properties will be used in the next section to analyze the entrywise information leakage of the PATE framework.
The aggregation mechanism in PATE is essentially the Report-Noisy-Max mechanism \cite{7} which operates by adding i.i.d. noise samples to the bins of the votes’ histogram, and returning the class label with the highest (noisy) value. Let Lap(b) denote the Laplace distribution with location 0 and scale b. Suppose \( N = (N_1, \ldots, N_m) \) is a sequence of i.i.d.

**Lemma 5 (Composition)** If the Markov chain \( Y_1 - (X, Z) - Y_2 \) holds, then,
\[
\mathcal{L}(X \rightarrow (Y_1, Y_2) \mid Z = z) \leq \mathcal{L}(X \rightarrow Y_1 \mid Z = z) + \mathcal{L}(X \rightarrow Y_2 \mid Z = z).
\]  
(6)

Lemma 5 states that the information leaked to multiple independent observations is upper bounded by the sum of the information leaked through each of the observations.

**Lemma 6 (Data-processing inequality)** If the Markov chain \( (X, Z) - Y_1 - Y_2 \) holds, then,
\[
\mathcal{L}(X \rightarrow Y_2 \mid Z = z) \leq \min\{\mathcal{L}(X \rightarrow Y_1 \mid Z = z), \mathcal{L}(Y_1 \rightarrow Y_2 \mid Z = z)\}.
\]  
(7)

Lemma 6 states that all processing of the output can only decrease the information leakage. Further, it allows us to upper bound the end-to-end leakage of a complex mechanism in terms of the leakages of its smaller intermediate mechanisms. The proofs of Lemma 5 and Lemma 6 are given in Appendix A.2 and A.3 respectively.

### 4 Information leakage analysis of PATE

In this section we we will use the pointwise conditional maximal leakage to measure the information leaking about individual data entries in the PATE framework. We will begin by describing the our system model in Section 4.1. Then, in Section 4.2 we will first prove that increased consensus among teachers in answering queries induces a lower privacy cost (i.e., the privacy-accuracy synergy), and then, state bounds on the entrywise leakage when the noise is Laplace distributed.

#### 4.1 System Model

Suppose \( d = ((x_1, y_1), \ldots, (x_n, y_n)) \in \mathcal{X}^n \times \mathcal{Y}^n \) represents the training data where \( \mathcal{X} \) is the arbitrary but finite domain set, \( \mathcal{Y} = [1, m] \) is the label set. The pairs \( (x_i, y_i) \) are sampled independently according to some distribution \( P \) over \( \mathcal{X} \times \mathcal{Y} \), i.e., \( D \sim P^n \). We use the training data \( d \) to train \( L \) teachers for a classification task with \( m \geq 2 \) classes in the PATE framework. Let \( (d^{(1)}, \ldots, d^{(L)}) \) represent a disjoint partitioning of the training set such that \( d^{(i)} \neq \emptyset \) for all \( i \in [1, L] \). \( \bigcup_{i=1}^{L} d^{(i)} = d \) and \( d^{(i)} \cap d^{(j)} = \emptyset \) for all \( i \neq j \). Each partition \( d^{(i)} \) is used to train a teacher model \( f_i: \mathcal{X} \rightarrow [1, m] \). This results in a total of \( L \) teacher models, classifying queries independently of each other.

The student model is trained using a public and unlabelled dataset, which will be labelled by the teachers ensemble in a privacy-preserving manner. Let \( (x'_1, \ldots, x'_k) \in \mathcal{X}^k \) be the independently sampled unlabelled dataset and suppose that the student queries the ensemble about the label of \( x'_i \) for each \( i \in [1, k] \). Each teacher separately predicts a label for \( x'_i \)

The aggregation mechanism in PATE is essentially the Report-Noisy-Max mechanism \cite{7} which operates by adding i.i.d. noise samples to the bins of the votes’ histogram, and returning the class label with the highest (noisy) value. Let Lap(b) denote the Laplace distribution with location 0 and scale b. Suppose \( N = (N_1, \ldots, N_m) \) is a sequence of i.i.d.
Laplace random variables, where $N_j \sim \text{Lap}(\frac{1}{j})$ for all $j \in [1, m]$ represents the noise added to the $j$th bin. Note that $\gamma$ determines the dispersion of the noise, and thus, affects the privacy guarantees of the system. Roughly speaking, smaller values of $\gamma$ correspond to larger noise, and in turn, stronger privacy guarantees. Finally, let $Y'_i = \max_{j} v_j(x'_i) + N_j$ be the random variable denoting the predicted label for $x'_i$ returned by the aggregation mechanism. Labelling the entire dataset $(x'_1, \ldots, x'_K)$ produces $k$ such predictions, each of which entailing a privacy cost. The system model is depicted in Figure 1.

### 4.2 Measuring the Entrywise Information Leakage

Now, we will measure the information leaking about individual data entries in the training set using the notion of pointwise conditional maximal leakage. In order to be able to quantify the entrywise leakage, let us consider the following scenario: assume an adversary knows the values of all the entries in the training set except for a single entry. Let $V(x'_i)$ be the histogram of the known votes for input $x'_i$. Based on the scenario we described, let $D$ know the vote of the teacher whose training partition includes $x'_i$. Note that we are considering a general setup in which any single data entry can arbitrarily affect the vote of its teacher, resulting in observations which are highly informative of the training data entry of interest. In other words, if the adversary can already predict the last vote there is no information left to be leaked.

Based on the scenario we described, let $D^- = D \setminus D^*$ be the random variable representing the portion of the training set which is known to the adversary, and let $V^-(x'_i) = (V^-_1(x'_i), \ldots, V^-_m(x'_i))$ be the random variable representing the histogram of the known votes for input $x'_i$. Note that $\sum_{j=1}^m V^-_j(x'_i) = L - 1$ for all $x'_i \in X$. For simplicity, let $Y^* = (Y_1^*, \ldots, Y_K^*)$ denote the sequence of random variables representing the predicted labels for the queries $(x'_1, \ldots, x'_K)$. We are interested in quantifying the information leaking about $D^*$ to $Y^*$ assuming that the adversary knows $D^- = d^-$, that is

$$
\mathcal{L}(D^* \rightarrow Y^* \mid D^- = d^-) = \log \sum_{\substack{y' \in \mathcal{Y}^n \setminus \mathcal{Y}^* \\text{d} \in \mathcal{X}^n \times \mathcal{Y}^*\setminus \mathcal{Y}^* \\text{P} \text{d} \text{d}^- > 0}} \mathbb{P}(y' \mid d^*, d^-)
$$

$$
= \log \sum_{\substack{y' \in \mathcal{Y}^n \setminus \mathcal{Y}^* \\text{d} \in \mathcal{X}^n \times \mathcal{Y}^*\setminus \mathcal{Y}^* \\text{P} \text{d} \text{d}^- > 0}} \mathbb{P}(y' \mid d)
$$

(8)

$$
= \mathcal{L}(D \rightarrow Y^* \mid D^- = d^-),
$$

where (a) follows from (5) since the Markov chain $D^- = D - Y'$ clearly holds. Using Lemma 5 we can upper bound the information leaked through multiple queries by writing

$$
\mathcal{L}(D \rightarrow Y^* \mid D^- = d^-) \leq \sum_{i=1}^k \mathcal{L}(D \rightarrow Y'_i \mid D^- = d^-),
$$

(9)

that is, the information leaked to the output of multiple queries is upper bounded by the sum of the information leaked through individual queries. Further, using Lemma 5 we can upper bound the information leaked to the output of a single query as

$$
\mathcal{L}(D \rightarrow Y'_i \mid D^- = d^-) \leq \min\{\mathcal{L}(D \rightarrow V(x'_i) \mid D^- = d^-), \mathcal{L}(V(x'_i) \rightarrow Y'_i \mid D^- = d^-)\},
$$

(10)
i.e., the information leaked to the output of a single query is upper bounded by the smallest of the information leaked through the training and the information leaked through the aggregation mechanism.

As we can make no assumptions about the privacy of the training process, we now turn to evaluating the information leaked through the aggregation mechanism. Let $\delta_j = (0, \ldots, 0, 1, 0, \ldots, 0)$ be a sequence with all components equal to
We will now evaluate (11) using the ideas of majorization theory \([15, 16]\), and assuming that the noise in the aggregation mechanism has a log-concave \([17, 18]\) probability density. Specifically, we will find the \(v^-\) maximizing (or minimizing) the information leakage of the aggregation mechanism for any noise with log-concave probability density.

**Definition 7 (Majorization)** Consider \(p, q \in \mathbb{R}^n\) with non-increasingly ordered elements, i.e., \(p_1 \geq p_2 \geq \ldots \geq p_n\) and \(q_1 \geq q_2 \geq \ldots \geq q_n\). We say that \(p\) majorizes \(q\), and write \(p \succ q\) if

\[
\sum_{i=1}^{m} p_i \geq \sum_{i=1}^{m} q_i, \quad \text{for } m = 1, \ldots, n-1 \quad \text{and} \quad \sum_{i=1}^{n} p_i = \sum_{i=1}^{n} q_i.
\]

Note that majorization only describes a partial ordering. For example, \((4, 4, 1)\) and \((5, 2, 2)\) cannot be compared in terms of majorization. On the other hand, if we define \(Q = \{(q_1, q_2, q_3) \in \mathbb{R}_+^3 : \sum_{i=1}^{3} q_i = 9\}\), then \((3, 3, 3)\) is majorized by all \(q \in Q\) while \((0, 0, 9)\) and \((0, 0, 9)\) majorize all \(q \in Q\).

**Definition 8 (Schur-concave function)** Consider a real-valued function \(\Phi\) defined on \(\mathbb{I}^n \subset \mathbb{R}^n\). \(\Phi\) is said to be Schur-concave on \(\mathbb{I}^n\) if \(p \succ q\) on \(\mathbb{I}^n\) implies \(\Phi(p) \leq \Phi(q)\).

**Definition 9 (Log-concave function)** A non-negative function \(f : \mathbb{R}^n \rightarrow \mathbb{R}_+\) is said to be log-concave if it can be written as \(f(x) = \exp \phi(x)\) for some concave function \(\phi : \mathbb{R}^n \rightarrow (-\infty, \infty)\).

Note that many commonly used probability density functions (and their corresponding CDFs) are log-concave, such as the Laplace and the Gaussian distributions \([17]\).

**Theorem 10** Consider the aggregation mechanism in PATE (i.e., the Report-Noisy-Max mechanism) where the noise has a log-concave probability density. Then, \(\mathcal{L}(V(x'_i) \rightarrow Y'_i \mid V^-(x'_i) = v^-)\) is Schur-concave in \(v^-\). Thus, assuming that \(L - 1\) is divisible by \(m\), \(\mathcal{L}(V(x'_i) \rightarrow Y'_i \mid V^-(x'_i) = v^-)\) is maximized when

\[
v^- = v^\text{max} = \left(\frac{L - 1}{m}, \ldots, \frac{L - 1}{m}\right),
\]

and is minimized when

\[
v^- = v^\text{min} = (0, \ldots, 0, L - 1, 0, \ldots, 0) = (L - 1) \delta_j,
\]

for each \(j \in [1, m]\). The proof is given in Appendix \([B]\).

**Remark 11** The Schur-concavity of the entrywise information leakage of the aggregation mechanism \(\mathcal{L}(V(x'_i) \rightarrow Y'_i \mid V^-(x'_i) = v^-)\) implies that stronger consensus among teachers lowers the amount of information leaked about any individual data entry.

The preceding remark points to one of the main advantages of the PATE framework: increased accuracy of the teacher models results in stronger consensus in predicting the label of a given query, which, in turn, results in stronger privacy guarantees. Note that \([13, 14]\) come to the same conclusions regarding the synergy between privacy and accuracy for the case of Laplace and Gaussian noise distributions, whereas here we have stated a systematic proof of this desired property and have generalized it to the class of log-concave probability densities.

Applying Theorem \([10]\) to \((11)\) yields the following result for the leakage of the aggregation mechanism with Laplace noise.
Therefore, note that due to the Schur-concavity of $L$, this result is a direct consequence of Theorem 13 and Lemma 5, and characterizes the overall information leaked about $\mathcal{D}$. The proof of this result is given in Appendix C.1. Proposition 12 describes a bound that depends on $k$, a single data entry as a result of training a student classifier using $\mathcal{D}$. Consider the setting of Theorem 13. The information leaked about $\mathcal{D}$ is upper bounded by

$$
\mathcal{L}(\mathcal{D} \to Y' | D^{-} = d^{-}) = \mathcal{L}(\mathcal{D} \to Y' | D^{-} = d^{-}) \leq \gamma.
$$

The proof of this result is given in Appendix C.2. Note that while Theorem 13 provides a simple and general upper bound on the entrywise information leakage, it is possible to get tighter bounds by directly calculating the leakage of the aggregation mechanism in (11) using the conditional probabilities. This is demonstrated in the following example.

**Example 14** Suppose the PATE framework has been implemented with $L = 11$ teachers to classify queries into $m = 4$ classes. Further, suppose that for a given query $x'_i$, the histogram of teachers’ votes is (any permutation of) $(5, 3, 2, 1)$, and that Laplace noise with $\gamma = 0.1$ is used in the aggregation mechanism. Depending on which partition of the training set includes $D^*$, the adversary has obtained one of the following values: $v^- \in \{(4, 3, 2, 1), (5, 2, 2, 1), (5, 3, 1, 1), (5, 3, 2, 0)\}$. We can now directly use (11) to calculate the leakage of the aggregation mechanism using the probability density function of the Laplace distribution for each $v^-$. For simplicity of notation we define $\mathcal{L}(v^-) := \mathcal{L}(\mathcal{D} \to Y' | \mathcal{D} \to Y' | V^{-}(x'_i) = v^-)$. Then, we have one of the following cases:

- $v^- = (4, 3, 2, 1) \implies \mathcal{L}(v^-) = 8.50 \times 10^{-2}$.
- $v^- = (5, 2, 2, 1) \implies \mathcal{L}(v^-) = 8.40 \times 10^{-2}$.
- $v^- = (5, 3, 1, 1) \implies \mathcal{L}(v^-) = 8.37 \times 10^{-2}$.
- $v^- = (5, 3, 2, 0) \implies \mathcal{L}(v^-) = 8.35 \times 10^{-2}$.

Therefore, $\mathcal{L}(D^* \to Y' | D^{-} = d^{-}) \leq 8.50 \times 10^{-2}$ which is a tighter bound than the one provided by Theorem 13. Note that due to the Schur-concavity of $\mathcal{L}(\mathcal{D} \to Y' | \mathcal{D} \to Y' | V^{-}(x'_i) = v^-)$ it was already expected that information leakage would be largest for $(4, 3, 2, 1)$, and it would have sufficed to just consider this case.

We conclude the paper by summarizing our results in the following Corollary.

**Corollary 15** Consider the setting of Theorem 13. The information leaked about $D^*$ as the result of training a student model on $k$ samples is upper bounded by

$$
\mathcal{L}(D^* \to Y' | D^{-} = d^{-}) \leq k\gamma.
$$

This result is a direct consequence of Theorem 13 and Lemma 5 and characterizes the overall information leaked about a single data entry as a result of training a student classifier using $k$ queries to the teachers.
5 Conclusions

In this paper, we proposed an approach based on information leakage for quantifying membership privacy. Particularly, we showed that the pointwise conditional maximal leakage, a conditional form of maximal leakage, can be used to measure the information leaking about individual data entries in a dataset. We applied our privacy analysis to PATE and derived novel privacy guarantees for this privacy-preserving classification framework in the form of upper bounds on its entrywise information leakage when the injected noise has Laplace distribution. We also showed that the privacy-accuracy synergy of PATE can be explained by studying the entrywise information leakage of the framework while it was only intuitively justified through the lens of differential privacy.

As our work has taken a step towards gaining a deeper understanding of some underlying privacy principles, our results can be used in the design of machine learning algorithms that preserve both privacy and utility. For example, we envision that based on the precise characterization of the leakage presented in our work one can employ data-dependent perturbation, i.e., noise that scales based on the privacy cost entailed by each query. This reduces the overall amount of noise, which, in turn, will improve the utility of the system. Another potential application is in privacy thresholding schemes where queries which are expensive in terms of privacy will not be answered at all. Once again this method will improve both the privacy and the utility of the system since the expensive queries are precisely those which were not labelled with certainty by the teachers.

References

[1] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[2] Gang Liang, Wenbo He, Chun Xu, Liangyin Chen, and Jinquan Zeng. Rumor identification in microblogging systems based on users’ behavior. *IEEE Transactions on Computational Social Systems*, 2(3):99–108, 2015.

[3] Jarrod West and Maumita Bhattacharya. Intelligent financial fraud detection: a comprehensive review. *Computers & security*, 57:47–66, 2016.

[4] Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael P Wellman. Sok: Security and privacy in machine learning. In *2018 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 399–414. IEEE, 2018.

[5] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE, 2017.

[6] Yunhui Long, Vincent Bindschaedler, Lei Wang, Diyue Bu, Xiaofeng Wang, Haixu Tang, Carl A Gunter, and Kai Chen. Understanding membership inferences on well-generalized learning models. *arXiv preprint arXiv:1802.04889*, 2018.

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

[8] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[9] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 486–503. Springer, 2006.

[10] Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275. IEEE, 2017.

[11] Ibrahim Issa, Aaron B Wagner, and Sudeep Kamath. An operational approach to information leakage. *IEEE Transactions on Information Theory*, 2019.

[12] Mário S Alvim, Miguel E Andréis, Konstantinos Chatzikokolakis, Pierpaolo Degano, and Catuscia Palamidessi. Differential privacy: on the trade-off between utility and information leakage. In *International Workshop on Formal Aspects in Security and Trust*, pages 39–54. Springer, 2011.

[13] Nicolas Papernot, Martin Abadi, Úlfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. *ICLR*, 2017.

[14] Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. Scalable private learning with pate. *ICLR*, 2018.
Appendix A  Proofs of the results in section 3

A.1 Proof of Proposition 3

This result follows readily from [14, Theorem 1] by considering \( L(X' \rightarrow Y) \) such that \( P_{X'} = P_{X|Z=z} \). Nevertheless, we provide an alternative proof.

Upper bound: First, we prove the upper bound on \( L(X \rightarrow Y | Z = z) \). Consider any discrete \( U \) satisfying \( U - (X,Z) - Y \) and define

\[
L_U(X \rightarrow Y | Z = z) := \log \frac{P(U = \hat{U}(Y,Z = z))}{P(U = \tilde{U}(Z = z))},
\]

which is maximized when \( U = Y \), giving rise to the upper bound

\[
L(X \rightarrow Y | Z = z) \leq I(X;Y|Z = z).
\]
where $\hat{U}$ and $\bar{U}$ are MAP estimators of $U$. Then, $\mathcal{L}(X \to Y \mid Z = z) = \sup_{U \colon U - (X, Z) - Y} \mathcal{L}_U(X \to Y \mid Z = z)$.

For each $z \in Z$, define $\mathcal{U}_z := \{u : P_{U|Z}(u \mid z) > 0\}$. The two probabilities in $\mathcal{L}_U(X \to Y \mid Z = z)$ are

$$
P(U = \hat{U}(Z = z)) = \max_{u \in \mathcal{U}_z} P_{U|Z}(u \mid z),$$

and

$$
P(U = \bar{U}(Z = z)) = \sum_{y \in Y} \max_{u \in \mathcal{U}_z} P_{U|Y,Z}(u, y \mid z)$$

$$= \sum_{y \in Y} \max_{u \in \mathcal{U}_z} \sum_{x : P_{X|Z}(x \mid z) > 0} P_{U|X,Z}(u \mid x, z) P_{Y|X,Z}(y \mid x, z) P_{X|Z}(x \mid z)$$

$$\leq \sum_{y \in Y} \max_{x : P_{X|Z}(x \mid z) > 0} P_{Y|X,Z}(y \mid x, z) \max_{u \in \mathcal{U}_z} \sum_{x : P_{X|Z}(x \mid z) > 0} P_{U|X,Z}(u \mid x, z)$$

$$= \max_{u \in \mathcal{U}_z} P_{U|Z}(u \mid z) \sum_{y \in Y} \left( \max_{x : P_{X|Z}(x \mid z) > 0} P_{Y|X,Z}(y \mid x, z) \right).$$

Thus,

$$\mathcal{L}_U(X \to Y \mid Z = z) \leq \log \sum_{y \in Y} \max_{x : P_{X|Z}(x \mid z) > 0} P_{Y|X,Z}(y \mid x, z)$$

for all $U$ such that $U - (X, Z) - Y$ holds. Then,

$$\mathcal{L}(X \to Y \mid Z = z) \leq \log \sum_{y \in Y} \max_{x : P_{X|Z}(x \mid z) > 0} P_{Y|X,Z}(y \mid x, z).$$

**Lower bound:** To prove the lower bound on $\mathcal{L}(X \to Y \mid Z = z)$, we will consider a discrete $U$ for which $\mathcal{L}_U(X \to Y \mid Z = z)$ attains the bound. We fix a $U'$ such that $U' - (X, Z) - Y$ holds and $H(X \mid U') = 0$, that is, the value of $X$ is completely determined by the value of $U'$. Further, we assume that $U' \mid Z = z$ is uniformly distributed, i.e., $P_{U'|Z}(u \mid z) = \frac{1}{|\mathcal{U}|}$ for all $z \in Z$ and $u \in \mathcal{U}_z$. Then,

$$\mathcal{L}_{U'}(X \to Y \mid Z = z) = \log \sum_{y \in Y} \frac{\max_{u \in \mathcal{U}_z} P_{U'|Z}(u \mid z) P_{Y'|U',Z}(y \mid u, z)}{\max_{u \in \mathcal{U}_z} P_{U'|Z}(u \mid z)}$$

$$= \log \sum_{y \in Y} \frac{\max_{u \in \mathcal{U}_z} P_{U'|Z}(u \mid z) P_{Y'|X,Z}(y \mid x, z)}{\max_{u \in \mathcal{U}_z} P_{U'|Z}(u \mid z)}$$

$$= \log \sum_{y \in Y} \frac{\frac{1}{|\mathcal{U}|} \max_{x : P_{X|Z}(x \mid z) > 0} P_{Y'|X,Z}(y \mid x, z)}{\frac{1}{|\mathcal{U}|}}$$

$$= \log \sum_{y \in Y} \max_{x : P_{X|Z}(x \mid z) > 0} P_{Y'|X,Z}(y \mid x, z).$$

Therefore,

$$\mathcal{L}(X \to Y \mid Z = z) = \sup_{U : U - (X, Z) - Y} \mathcal{L}_U(X \to Y \mid Z = z)$$

$$\geq \mathcal{L}_{U'}(X \to Y \mid Z = z)$$

$$= \log \sum_{y \in Y} \max_{x : P_{X|Z}(x \mid z) > 0} P_{Y'|X,Z}(y \mid x, z).$$

Hence, from (23) and (25) it follows that

$$\mathcal{L}(X \to Y \mid Z = z) = \log \sum_{y \in Y} \max_{x : P_{X|Z}(x \mid z) > 0} P_{Y'|X,Z}(y \mid x, z).$$
A.2 Proof of Lemma 5

Consider the Markov chain $Y_1 - (X, Z) - Y_2$. Then,

$$
\mathcal{L}(X \to (Y_1, Y_2) \mid Z = z) - \mathcal{L}(X \to Y_1 \mid Z = z) = \log \frac{\sum_{y_1, y_2} \max_{x : P_{X \mid Z}(x \mid z) > 0} P_{Y_1|XZ}(y_1, y_2 \mid x, z)}{\sum_{y_1} \max_{x : P_{X \mid Z}(x \mid z) > 0} P_{Y_1|XZ}(y_1 \mid x, z)}
$$

$$
= \log \frac{\sum_{y_1} \max_{x \mid XZ} P_{Y_1|XZ}(y_1 \mid x, z) P_{Y_2|XZ}(y_2 \mid x, z)}{\sum_{y_1} \max_{x : P_{X \mid Z}(x \mid z) > 0} P_{Y_1|XZ}(y_1 \mid x, z)}
$$

$$
\leq \log \frac{\sum_{y_1} \max_{x \mid XZ} P_{Y_1|XZ}(y_1 \mid x, z) (\max_{x' \mid XZ} P_{Y_2|XZ}(y_2 \mid x', z))}{\sum_{y_1} \max_{x : P_{X \mid Z}(x \mid z) > 0} P_{Y_1|XZ}(y_1 \mid x, z)}
$$

$$
= \log \sum_{y_2} \max_{x' \mid XZ} P_{Y_2|XZ}(y_2 \mid x', z)
$$

$$
= \mathcal{L}(X \to Y_2 \mid Z = z). \quad (27)
$$

Therefore,

$$
\mathcal{L}(X \to (Y_1, Y_2) \mid Z = z) \leq \mathcal{L}(X \to Y_1 \mid Z = z) + \mathcal{L}(X \to Y_2 \mid Z = z). \quad (28)
$$

A.3 Proof of Lemma 6

Our proof follows the same reasoning as the proof of [11, Lemma 1]. For all discrete $U$ satisfying $U - (X, Z) - Y_1 - Y_2$ it holds that

$$
\mathcal{L}_U(X \to Y_2 \mid Z = z) \leq \mathcal{L}_U(X \to Y_1 \mid Z = z), \quad (29)
$$

where $\mathcal{L}_U$ is defined in [19]. Therefore,

$$
\mathcal{L}(X \to Y_2 \mid Z = z) = \sup_{U: U - (X, Z) - Y_1 - Y_2} \mathcal{L}_U(X \to Y_2 \mid Z = z)
$$

$$
\leq \sup_{U: U - (X, Z) - Y_1} \mathcal{L}_U(X \to Y_1 \mid Z = z)
$$

$$
= \mathcal{L}(X \to Y_1 \mid Z = z). \quad (30)
$$

Similarly,

$$
\mathcal{L}(X \to Y_2 \mid Z = z) = \sup_{U: U - (X, Z) - Y_1 - Y_2} \mathcal{L}_U(X \to Y_2 \mid Z = z)
$$

$$
\leq \sup_{U: U - Z - Y_1 - Y_2} \mathcal{L}_U(Y_1 \to Y_2 \mid Z = z)
$$

$$
= \mathcal{L}(Y_1 \to Y_2 \mid Z = z). \quad (31)
$$

Thus,

$$
\mathcal{L}(X \to Y_2 \mid Z = z) \leq \min\{\mathcal{L}(X \to Y_1 \mid Z = z), \mathcal{L}(Y_1 \to Y_2 \mid Z = z)\}. \quad (32)
$$

Appendix B Proof of Theorem 10

Before stating the proof, let us recall some concepts/results from majorization theory.

**Definition 16 (Symmetric function)** Let $x = (x_1, \ldots, x_n) \in \mathcal{I}^n \subset \mathbb{R}^n$ and consider a real-valued function $\Phi : \mathcal{I}^n \to \mathbb{R}$. The function $\Phi(x)$ is said to be symmetric if $x$ can be arbitrarily permuted without changing the value of $\Phi(x)$. 

Then, the global maximum is achieved by
\[ v_i = \frac{\partial f}{\partial x_i} - \frac{\partial f}{\partial x_j} \leq 0 \quad \text{for all} \quad 1 \leq i, j \leq n. \] (33)
Since \( \Phi(x) \) must be symmetric, it is sufficient to verify the reduced condition
\[ (x_1 - x_2) \left( \frac{\partial f}{\partial x_1} - \frac{\partial f}{\partial x_2} \right) \leq 0. \] (34)

**Proposition 18** ([16, Theorem 2.21]) Let \( x = (x_1, \ldots, x_n) \in \mathbb{R}_+^n \) and let \( f : \mathbb{R}_+^n \to \mathbb{R}_+ \) be a Schur-concave function. Consider the following problems
\[
\max_x f(x) \quad \text{such that} \quad \sum_{i=1}^n x_i = S, \tag{35}
\]
and
\[
\min_x f(x) \quad \text{such that} \quad \sum_{i=1}^n x_i = S. \tag{36}
\]
Then, the global maximum is achieved by
\[ x_{\text{max}} = \frac{S}{n}(1, \ldots, 1), \tag{37} \]
and the global minimum is achieved by
\[ x_{\text{min}} = (0, \ldots, 0, S, 0, \ldots, 0). \tag{38} \]

We now prove that the entrywise information leakage of the aggregation mechanism is Schur-concave when the injected noise has a log-concave probability density. In order to simplify the proof, we will assume that the elements of \( v^- \) (i.e., the histogram of known votes) can take non-negative real values. The results of the proof, however, will be readily applicable to histograms of non-negative integers.

Using [11] we define
\[ f_j(v^-) := \mathbb{P}(Y_i' = j \mid V(x_i') = v^- + \delta_j), \tag{39} \]
where \( \delta_j = (0, \ldots, 0, 1, 0, \ldots, 0) \) represents a single vote for class \( j \). Then,
\[
\mathcal{L}(V(x_i') \to Y_i' \mid V^- (x_i') = v^-) = \log \sum_{j=1}^m f_j(v^-) \tag{40}
\]
where \( f(v^-) = \sum_{j=1}^m f_j(v^-) \). It is clear from \([10]\) that the leakage does not depend on the order of elements in \( v^- \), thus \( \mathcal{L}(V(x_i') \to Y_i' \mid V^- (x_i') = v^-) \) is symmetric. Moreover, according to \([15], 3.B.1]\), the composition of an increasing function and a Schur-concave function remains Schur-concave. Since \( \log(\cdot) \) is an increasing function, to prove the Schur-concavity of the entrywise leakage we only need to verify Schur’s condition for \( f(v^-) \).

Without loss of generality assume that \( v^- = (v_1^- \ldots, v_m^-) \) is non-increasingly ordered, i.e., \( v_1^- \geq \ldots \geq v_m^- \). Let \( N = (N_1, \ldots, N_m) \) denote the tuple of noise, where the elements are independent, identically distributed and have a log-concave probability density. We write
\[
f_j(v^-) = \mathbb{P}(Y_i' = j \mid V(x_i') = v^- + \delta_j)
= \mathbb{P}\{v_j^- + N_j + 1 > v_j^- + N_1, \ldots, v_j^- + N_j + 1 > v_m^- + N_m\}
= \int_{-\infty}^{\infty} \prod_{j=1}^m \mathbb{P}\{N_i < (v_j^- - v_i^- + t + 1)\} \, g(t) \, dt
= \int_{-\infty}^{\infty} \prod_{j=1}^m G(v_j^- - v_i^- + t + 1) \, g(t) \, dt, \tag{41}
\]
\]
\]
\]
where \( g(t) \) is the probability density function of \( N_j \) and \( G(t) = \int_{-\infty}^{t} g(t') \, dt' \) is its corresponding cumulative distribution function. According to [18, Proposition 1], if \( g \) is log-concave, then \( G \) is also log-concave. We now check Schur’s condition by writing

\[
\frac{\partial f(v^-)}{\partial v_i^-} - \frac{\partial f_j(v^-)}{\partial v_i^-} = \sum_{j=1}^{m} \frac{\partial f_j(v^-)}{\partial v_i^-} - \frac{\partial f_j(v^-)}{\partial v_i^-},
\]

where we have one of the following three cases:

if \( j = 1 \), then,

\[
\frac{\partial f_1(v^-)}{\partial v_1^-} - \frac{\partial f_1(v^-)}{\partial v_2^-} = \int_{-\infty}^{\infty} \left[ \sum_{l=2}^{m} g(v_l^- - v_1^- + t) \prod_{k=2}^{m} G(v_k^- - v_1^- + t + 1) \right] g(t) \, dt \\
- \int_{-\infty}^{\infty} \left[ -g(v_1^- - v_2^- + t + 1) \right] \prod_{l=3}^{m} G(v_l^- - v_1^- + t + 1) g(t) \, dt,
\]

if \( j = 2 \), then,

\[
\frac{\partial f_2(v^-)}{\partial v_1^-} - \frac{\partial f_2(v^-)}{\partial v_2^-} = \int_{-\infty}^{\infty} \left[ -g(v_2^- - v_1^- + t + 1) \right] \prod_{j=3}^{m} G(v_j^- - v_1^- + t + 1) g(t) \, dt \\
- \int_{-\infty}^{\infty} \sum_{l \neq 2}^{m} g(v_2^- - v_l^- + t + 1) \prod_{k=1}^{m} G(v_k^- - v_1^- + t + 1) g(t) \, dt,
\]

and if \( j \neq 1, 2 \), then,

\[
\frac{\partial f_j(v^-)}{\partial v_1^-} - \frac{\partial f_j(v^-)}{\partial v_2^-} = \int_{-\infty}^{\infty} \left[ g(v_j^- - v_1^- + t + 1)G(v_j^- - v_2^- + t + 1) \\
+ g(v_j^- - v_1^- + t + 1)G(v_j^- - v_1^- + t + 1) \right] \prod_{l \neq j}^{m} G(v_j^- - v_l^- + t + 1) g(t) \, dt.
\]

Then,

\[
\frac{\partial f(v^-)}{\partial v_1^-} - \frac{\partial f(v^-)}{\partial v_2^-} = A_1 - A_2 + \sum_{j=3}^{m} B_{(1,j)} - B_{(2,j)} + B_{(3,j)} - B_{(4,j)},
\]

where

\[
A_1 = 2 \int_{-\infty}^{\infty} g(v_1^- - v_2^- + t + 1) \prod_{k=3}^{m} G(v_k^- - v_1^- + t + 1) g(t) \, dt,
\]

\[
A_2 = 2 \int_{-\infty}^{\infty} g(v_2^- - v_1^- + t + 1) \prod_{k=3}^{m} G(v_k^- - v_1^- + t + 1) g(t) \, dt,
\]

\[
B_{(1,j)} = \int_{-\infty}^{\infty} g(v_1^- - v_j^- + t + 1)G(v_1^- - v_2^- + t + 1) \prod_{k=3}^{m} G(v_k^- - v_1^- + t + 1) g(t) \, dt,
\]

\[
B_{(2,j)} = \int_{-\infty}^{\infty} g(v_j^- - v_1^- + t + 1)G(v_j^- - v_2^- + t + 1) \prod_{k=3}^{m} G(v_k^- - v_1^- + t + 1) g(t) \, dt,
\]

\[
B_{(3,j)} = \int_{-\infty}^{\infty} g(v_j^- - v_2^- + t + 1)G(v_j^- - v_1^- + t + 1) \prod_{k=3}^{m} G(v_k^- - v_1^- + t + 1) g(t) \, dt,
\]

\[
B_{(4,j)} = \int_{-\infty}^{\infty} g(v_j^- - v_2^- + t + 1)G(v_j^- - v_1^- + t + 1) \prod_{k=3}^{m} G(v_k^- - v_2^- + t + 1) g(t) \, dt.
\]
We now prove that \( A_1 - A_2 \leq 0 \). By a change of variable in \( A_1 \) we let \( v_1^- - v_2^- + t = u \). Then,
\[
A_1 - A_2 = \int_{-\infty}^{\infty} \prod_{k=3}^{m} G(u + v_k^- - v_k^- + 1) \cdot [g(u + 1)g(u + v_2^- - v_1^-) - g(u)g(u + v_2^- - v_1^- + 1)] \, du. 
\]  
(54)

We now apply Proposition 19 to the preceding equation by noting that \( u \geq u + v_2^- - v_1^- \) (due to the non-increasing order of the elements in \( v^- \)), and write
\[
g(u + 1)g(u + v_2^- - v_1^-) - g(u)g(u + v_2^- - v_1^- + 1) \leq 0. 
\]  
(55)

Since \( \prod_{k=3}^{m} G(u + v_2^- - v_k^- + 1) \geq 0 \), we conclude that
\[
A_1 - A_2 \leq 0. 
\]  
(56)

Similarly, Proposition 19 and Proposition 20 can be used to show that \( B_{1,j} - B_{2,j} + B_{3,j} - B_{4,j} \leq 0 \) for all \( j = 3, \ldots, m \). Therefore, we have verified Schur’s condition for \( f(v^-) \), and conclude that \( \mathcal{L}(V(x'_j) \to Y'_j \mid V^-(x'_j) = v^-) \) is Schur-concave. Finally, by Proposition 18, the entrywise leakage is maximized by
\[
v^- = v_{\text{max}}^- = \left( \frac{L-1}{m}, \ldots, \frac{L-1}{m} \right), 
\]  
(57)

and is minimized by
\[
v^- = v_{\text{min}}^- = (0, \ldots, 0, L-1, 0, \ldots, 0) = (L-1) \delta_j, 
\]  
(58)

for each \( j \in \llbracket 1, m \rrbracket \).

### Appendix C  Proofs for the leakage with Laplace noise

#### C.1  Proof of Proposition 12

Let \( N = (N_1, \ldots, N_m) \) be the sequence of i.i.d. Laplace random variables, where \( N_j \sim \text{Lap}(\frac{1}{\gamma}) \) for all \( j \in \llbracket 1, m \rrbracket \). To find an upper bound on the leakage, we will apply Theorem 10 and calculate \( \mathcal{L}(V(x'_i) \to Y'_i \mid V^-(x'_i) = v^-) \) for \( v^- = v_{\text{max}}^- = (\frac{k-1}{m}, \ldots, \frac{k-1}{m}) \). We write
\[
\mathcal{L}(V(x'_i) \to Y'_i \mid V^-(x'_i) = v_{\text{max}}^-) = \log \sum_{j=1}^{m} \mathbb{P}(Y'_i = j \mid V(x'_i) = v_{\text{max}}^- + \delta_j), 
\]  
(59)

where
\[
\mathbb{P}(Y'_i = j \mid V(x'_i) = v_{\text{max}}^- + \delta_j) = \mathbb{P}\{N_j + 1 > N_1, \ldots, N_j + 1 > N_m\} 
\]  
\[
= \int_{-\infty}^{\infty} \prod_{j=1 \atop j \neq i}^{m} \mathbb{P}\{N_l < (t + 1)\} \cdot \frac{\gamma}{2} e^{-\gamma |t|} dt, 
\]  
(60)
and
\[ P(N_t < (t + 1)) = \begin{cases} \frac{1}{2} e^{\gamma(t+1)} & t \leq -1, \\ 1 - \frac{1}{2} e^{-\gamma(t+1)} & t \geq -1. \end{cases} \] (61)

Thus, we have
\[ P(Y' = j \mid V(x'_1) = f_{\text{max}} + \delta_j) = \frac{\gamma}{2} \int_{-\infty}^{1} \left[ \frac{1}{2} e^{\gamma(t+1)} \right]^{m-1} \cdot e^{\gamma t} dt 
+ \frac{\gamma}{2} \int_{-1}^{0} \left[ 1 - \frac{1}{2} e^{-\gamma(t+1)} \right]^{m-1} \cdot e^{\gamma t} dt + \frac{\gamma}{2} \int_{0}^{\infty} \left[ 1 - \frac{1}{2} e^{-\gamma(t+1)} \right]^{m-1} \cdot e^{-\gamma t} dt. \] (62)

It is straightforward to calculate integrals \( A \) and \( C \) as
\[ A = \frac{2^{-m}}{m} e^{-\gamma} \quad \text{and} \quad C = 1 - \left[ 1 - \frac{1}{2} e^{-\gamma} \right]^m \cdot e^{\gamma}. \] (63)

Integral \( B \) can be written as
\[ B = \frac{1}{2} \left( 1 - \frac{1}{2} e^{-\gamma} \right)^{m-1} - 2^{-m} e^{-\gamma} - \frac{\gamma(m-1)}{4} e^{-\gamma} \int_{-1}^{0} \left( 1 - \frac{1}{2} e^{-\gamma(t+1)} \right)^{m-2} dt. \] (64)

We define
\[ H(m) := \gamma \int_{-1}^{0} \left( 1 - \frac{1}{2} e^{-\gamma(t+1)} \right)^m dt = \gamma \sum_{k=0}^{m} \binom{m}{k} \left( -\frac{1}{2} \right)^k e^{-\gamma k} \int_{-1}^{0} e^{-\gamma k t} dt = \sum_{k=0}^{m} \binom{m}{k} \left( -\frac{1}{2} \right)^k \frac{1}{k} (1 - e^{-\gamma k}). \] (65)

Using \( \binom{m}{k} = \binom{m-1}{k} + \binom{m-1}{k-1} \) for \( m \geq 1 \), we get
\[ H(m) = \sum_{k=0}^{m-1} \binom{m-1}{k} \left( -\frac{1}{2} \right)^k \frac{1}{k} (1 - e^{-\gamma k}) + \sum_{k=0}^{m-1} \binom{m-1}{k-1} \left( -\frac{1}{2} \right)^k \frac{1}{k} (1 - e^{-\gamma k}) = H(m-1) + \frac{1}{m} \left( 2^{-m} - (1 - \frac{1}{2} e^{-\gamma})^m \right), \] (66)

and \( H(0) = \gamma \). Thus,
\[ H(m) = \begin{cases} \gamma & m = 0, \\ \gamma + \sum_{k=1}^{m} \frac{2^{-k} - (1 - \frac{1}{2} e^{-\gamma})^k}{k} & m \geq 1, \end{cases} \] (67)

Note that \( H(m) \) is non-negative and monotonically decreasing in \( m \). Since \( \sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{1}{12} \) for \( |t| < 1 \), we have \( \lim_{m \to \infty} H(m) = 0 \). Hence, integral \( B \) can be written as
\[ B = \frac{1}{2} \left( 1 - \frac{1}{2} e^{-\gamma} \right)^{m-1} - 2^{-m} e^{-\gamma} - \frac{m-1}{4} e^{-\gamma} H(m-2). \] (68)

Finally, we have
\[ \mathcal{L}(V(x'_1) \mid V^{-}(x'_1) = v^-) \leq \mathcal{L}(V(x'_1) \mid V^{-}(x'_1) = v_{\text{max}}) \]
\[ = \frac{1}{m} \left[ 1 - \left( 1 - \frac{1}{2} e^{-\gamma} \right)^m \right] e^{\gamma} + \frac{1}{m} \left( 1 - \frac{1}{2} e^{-\gamma} \right)^{m-1} - \frac{m-1}{4} e^{-\gamma} H(m-2). \] (69)
C.2 Proof of Theorem 13

In order to prove the bound, we will show that
\[ k(m) := \exp \left( L(V(x'_i) \rightarrow Y'_i \mid V^-(x'_i) = v_{\max}^-)[m] \right) \]
is concave in \( m \) and that
\[ \lim_{m \to \infty} \exp \left( L(V(x'_i) \rightarrow Y'_i \mid V^-(x'_i) = v_{\max}^-) \right) = e^\gamma. \tag{70} \]

Since \( m \) is an integer, we will check the second-order difference of the leakage with respect to \( m \). The first-order difference is
\[ \Delta k(m) = k(m + 1) - k(m) \tag{71} \]
and the second-order difference is
\[ \Delta^2 k(m) = \Delta k(m + 1) - \Delta k(m) \]
\[ = - \frac{1}{2} e^{-\gamma} H(m) \tag{a} \leq 0, \tag{72} \]
where (a) follows from the fact that \( H(m) \) is non-negative. Thus, we have shown that \( \exp \left( L(V(x'_i) \rightarrow Y'_i \mid V^-(x'_i) = v_{\max}^-) \right) \) is concave in \( m \). Furthermore, it is straightforward to verify that (70) holds. Hence, we have
\[ L(V(x'_i) \rightarrow Y'_i \mid V^-(x'_i) = v_{\max}^-) \leq \gamma. \tag{73} \]

Finally, we get
\[ L(D^* \rightarrow Y'_i \mid D^- = d^-) = L(D \rightarrow Y'_i \mid D^- = d^-) \]
\[ \leq L(V(x'_i) \rightarrow Y'_i \mid V^-(x'_i) = v^-) \]
\[ \leq L(V(x'_i) \rightarrow Y'_i \mid V^-(x'_i) = v_{\max}^-) \]
\[ \leq \gamma. \tag{74} \]