The Price of Privacy in Untrusted Recommendation Engines

Siddhartha Banerjee
The University of Texas at Austin
SBANERJEE@MAIL.UTEXAS.EDU

Nidhi Hegde
Technicolor, Paris Research Lab
NIDHI.HEGDE@TECHNICOLOR.COM

Laurent Massoulié
Technicolor, Paris Research Lab
LAURENT.MASSOULIE@TECHNICOLOR.COM

Abstract
Recent increase in online privacy concerns prompts the following question: can a recommendation engine be accurate if end-users do not entrust it with their private data? To provide an answer, we study the problem of learning user ratings for items under local or ‘user-end’ differential privacy, a powerful, formal notion of data privacy.

We develop a systematic approach for lower bounds on the complexity of learning item structure from privatized user inputs, based on mutual information. Our results identify a sample complexity separation between learning in the scarce information regime and the rich information regime, thereby highlighting the role of the amount of ratings (information) available to each user.

In the information-rich regime (where each user rates at least a constant fraction of items), a spectral clustering approach is shown to achieve optimal sample complexity. However, the information-scarce regime (where each user rates only a vanishing fraction of the total item set) is found to require a fundamentally different approach. We propose a new algorithm, MaxSense, and show that it achieves optimal sample complexity in this setting.

The techniques we develop for bounding mutual information may be of broader interest. To illustrate this, we show their applicability to (i) learning based on 1-bit sketches (in contrast to differentially private sketches), and (ii) adaptive learning, where queries can be adapted based on answers to past queries.

Keywords: Differential privacy, recommender systems, lower bounds, partial information

1. Introduction
Recommender systems are fast becoming one of the cornerstones of the Internet; in a world with ever increasing choices, they are one of the most effective ways of matching users with items. Today, many websites (Amazon, Netflix, Yahoo, etc.) use some form of such systems, and research into these algorithms received a fillip from the recently concluded Netflix prize competition. Ironically, the contest also exposed the Achilles heel of such systems, when Narayanan and Shmatikov (2006) demonstrated that the Netflix data could be de-anonymized. Subsequent works such as Calandrino et al. (2011) have reinforced belief in the frailty of these algorithms in the face of privacy attacks.

To design recommender systems in such scenarios, we first need to define what it means for a data-release mechanism to be private. The popular perception has coalesced around the following notion: a person can either participate in a collaborative filtering system and waive all claims to privacy, or avoid such systems entirely. The response of the research community to these concerns has been the development of a third paradigm, between complete exposure and complete silence.
This new approach has been captured in the formal notion of differential privacy (refer Dwork (2006)); essentially it suggests that although perfect privacy is impossible, one can control the leakage of information by deliberately corrupting sensitive data before release. The original definition in Dwork (2006) provides a statistical test that must be satisfied by a data-release mechanism to be private. Accepting this paradigm shifts the focus to designing algorithms that obey this constraint while maximizing relevant notions of utility. This trade-off between utility and privacy has been explored for several problems in database management (refer to Blum et al. (2005); Dwork (2006); Dwork et al. (2006, 2010a,b)) and learning (refer to Blum et al. (2008); Chaudhuri et al. (2011); Gupta et al. (2011); Kasiviswanathan et al. (2008); McSherry and Mironov (2009); Smith (2011)).

In the context of recommender systems, there are two models for ensuring privacy: centralized and local. Under the centralized model the recommender system is trusted to collect data from users; it then responds to queries by publishing results that have been corrupted via any mechanism that obeys the differential privacy constraint. However, users increasingly desire control over their private data given the mistrust in systems with centrally stored data—misgivings that are supported by examples such as the Netflix privacy breach. In cases where the database cannot be trusted to keep data confidential, users can store their data locally, and differential privacy is ensured through suitable randomization at the ‘user-end’ before releasing data to the recommender system. This is precisely the context of the present paper: the design of differentially private algorithms within the setting of untrusted recommender systems.

The latter model is variously known in privacy literature as local differential privacy (see Kasiviswanathan et al. (2008); we henceforth refer to it as local-DP), and also in statistics as the ‘randomized response technique’ (see Warner (1965)). However, there are two unique challenges to local-DP posed by recommender systems which have not been dealt with before:

1. The underlying space (here, the set of ratings over all items) has very high dimensionality.
2. The users have limited information: they rate only a (vanishingly small) fraction of items.

In this work we address both these issues. Assuming an unknown cluster structure for the items, we demonstrate a surprising change in the sample complexity of private learning algorithms when shifting from information-rich to information-scarce settings. No similar phenomenon is known for non-private learning. With the aid of new information-theoretic arguments, we provide lower bounds on the sample complexity in various regimes. On the other hand, these arguments also guide us in developing novel algorithms, particularly in the information-scarce setting, which match the lower bounds upto logarithmic factors. Thus although we pay a ‘price of privacy’ when ensuring local-DP in untrusted recommender systems with information-scarcity, we can design optimal algorithms under such regimes.

1.1. Our Results

We now present a high level view of our technical results, and discuss their relevance to the problem of designing algorithms for untrusted recommender systems. As mentioned before, we focus on learning a stochastic generative model for the data, under user-end, or local differential privacy constraints. This entails a subtle difference in the definition of utility as compared to centralized differential privacy. In the latter, the true model may be known to the database curator, but privacy constraints require the output to be perturbed; the performance measure is the size of database required to output a hypothesis that is private and close to the truth. In contrast, local differential privacy ensures privacy at the user-end; the aim of the system is to learn the model from privatized
responses to appropriately designed queries, and the performance is in terms of the number of users needed for learning.

More precisely, we aim at learning a partition of the items into clusters within which items are statistically identical. The hypothesis class (i.e., set of models) is the set of functions from items \([N]\) to cluster labels \([L]\) (where typically \(L \ll N\), and thus has size \(L^N\)). Further, we assume that each user has rated only \(w\) items out of the possible \(N\). For a learner to be successful, we require that it identify the correct cluster label for all items\(^1\). Our starting point is then given by the following basic lower bound (for exact definitions, see Section 2)

**Informal Theorem 1 (Theorem 5)** For any (finite) hypothesis class \(\mathcal{H}\) to be ‘successfully’ learnt under \(\epsilon\)-differential privacy, the number of users must satisfy:

\[
U_{LB} = \Omega\left( \frac{\log |\mathcal{H}|}{\epsilon} \right).
\]

The above theorem is based on a standard use of Fano’s inequality in statistical learning. Returning to the recommender system problem, note that \(\log |\mathcal{H}| = \Theta(N)\). In the information-rich setting (i.e., where \(w = \Omega(N)\)), we show the above bound is matched (up to logarithmic factors) by a local-DP algorithm based on a novel ‘pairwise-preference’ sketch and spectral clustering techniques:

**Informal Theorem 2 (Theorem 6)** In the information-rich (IR) regime, clustering via the Pairwise-Preference Algorithm succeeds if the number of users exceeds:

\[
U_{IR}^{PP} = O\left( \frac{N \log N}{\epsilon} \right).
\]

In practical scenarios \(w\) is quite small; for example, in a movie ratings system, users usually have seen and rated only a small fraction of the set of movies. Our main results in the paper concern non-adaptive, local-DP learning in the information-scarce regime (where \(w = o(N)\)). Herein, we observe an interesting phase-change in the sample complexity of private learning:

**Informal Theorem 3** In the information-scarce (IS) regime, the sample complexity of non-adaptive, local-DP cluster learning is lower bounded by (Theorem 9): \(U_{LB}^{IS} = \Omega\left( \frac{N^2 w^2}{\epsilon} \right)\). Furthermore, for small \(w\) (in particular, \(w = o(N^{1/3})\)), we have (Theorem 10): \(U_{LB}^{IS} = \Omega\left( \frac{N^2}{w^2} \right)\).

Finally for the IS regime, we develop a new class of algorithms based on a novel sketch, that, under certain separation conditions, matches the above lower bound up to logarithmic factors:

**Informal Theorem 4 (Theorem 11)** For a given \(w\), clustering under the MaxSense Algorithm (Section 5) is successful if the number of users exceeds a threshold given by:

\[
U_{MS} = O\left( \frac{N^2 \log N}{w \epsilon} \right).
\]

**Techniques:** Our main technical contribution lies in the tools we use for the lower bounds. By viewing the privacy mechanism as a noisy channel with certain constraints, we are able to use information theoretic methods to obtain bounds on private learning. Although these connections between privacy and mutual information have been considered in previous works (refer McGregor et al. (2010); Alvim et al. (2011)), our work is novel in that: a) it illustrates its application to problems in private learning (via Fano’s inequality), and b) it shows how non-trivial bounds can be obtained via careful analysis of the information leakage in private mechanisms. Towards the latter, we formalize a notion of ‘channel mis-alignment’ between the ‘sampling channel’ (the partial ratings submitted by users) and the privatization channel. In Section 4 we provide a structural lemma (Lemma 7) that quantifies this mismatch under general conditions, and demonstrate its use by obtaining tight lower

---

1. in Appendix A we also treat the case where we allow a fraction of item misclassifications.
bounds under 1-bit (non-private) sketches. In Section 4.2 we use it to obtain tight lower bounds under local-DP. In Section 6 we discuss its application to adaptive local-DP algorithms, establishing a lower bound of order $\Omega(N \log N)$, which also refines Informal Theorem 1. Though we focus on the item clustering problem, the lower bounds thus obtained apply to learning any finite hypothesis class under privacy constraints, and offer scope for further extensions.

The information theoretic results also suggest that 1-bit privatized sketches are sufficient for learning in such scenarios. Based on this intuition, we show how existing spectral-clustering techniques can be extended to private learning in some regimes. More significantly, in the information-scarce regime, where spectral learning fails, we develop a novel algorithm based on blind probing of a large set of items. This algorithm, in addition to being private and having optimal sample complexity in many regimes, triggers several interesting open questions, which we discuss in Section 6.

1.2. Related Work

**Privacy preserving recommender systems:** The design of recommender systems with differential privacy was studied by McSherry and Mironov (2009) under the centralized model. Like us, they separate the recommender system into two components, a learning phase (based on a database appropriately perturbed to ensure privacy) and a recommendation phase (performed by the users ‘at home’, without interacting with the system). They numerically compare the performance of the algorithm against non-private algorithms. In contrast, we consider a stronger notion of privacy (local-DP), and for our generative model, are able to provide tight analytical guarantees and further, quantify the impact of limited information on privacy.

**Private PAC Learning and Query Release:** Several works have considered private algorithms for PAC-learning. Blum et al. (2008); Gupta et al. (2011) consider the private query release problem (i.e., releasing approximate values for all queries in a given class) in the centralized model. Kasiviswanathan et al. (2008) show equivalences between: a) centralized private learning and agnostic PAC learning, b) local-DP and the statistical query (SQ) model of learning; this line of work is further extended by Beimel et al. (2010). Although some of our results (in particular, Theorem 5) are similar in spirit to lower bounds for PAC (see Kasiviswanathan et al. (2008); Beimel et al. (2010) there are significant differences both in scope and technique. Furthermore:

1. We emphasize the importance of limited information, and characterize its impact on learning with local-DP. Hitherto unconsidered, information scarcity is prevalent in practical scenarios, and as our results shows, it has strong implications on learning performance under local-DP.
2. Via lower bounds, we provide a tight characterization of sample complexity, unlike Kasiviswanathan et al. (2008); Blum et al. (2008); Gupta et al. (2011), which are concerned with showing polynomial bounds. This is important for high dimensional data sets.

**Privacy in Statistical Learning:** Chaudhuri et al. (2011) consider privacy in the context of empirical risk minimization; they analyze the release of classifiers, obtained via algorithms such as SVMs, with (centralized) privacy constraints on the training data. Though they provide performance guarantees, they do not provide related lower bounds. Dwork and Lei (2009) study algorithms for privacy-preserving regression under the centralized model; these however require running time which is exponential in the data dimension. Smith (2011) obtains private, asymptotically-optimal algorithms for statistical estimation, again though, in the centralized model.

**Other Notions of Privacy:** The local-DP model which we consider has been studied before in privacy literature (Kasiviswanathan et al. (2008); Dwork et al. (2006)) and statistics (Warner (1965)).
It is a stronger notion than central differential privacy, and also stronger than two other related not-
tions: pan-privacy (Dwork et al. (2010b)) where the database has to also deal with occasional release
of its state, and privacy under continual observations (Dwork et al. (2010a)), where the database
must deal with additions and deletions, while maintaining privacy.

**Recommendation algorithms based on incoherence properties:** Apart from privacy-preserving
algorithms, there is a large body of work on designing recommender systems under various con-
straints (usually low-rank) on the ratings matrix (for example, Wainwright (2009); Keshavan et al.
(2010)). These methods, though robust, fail in the presence of privacy constraints, as the noise
added as a result of privatization is much more than their noise-tolerance. This is intuitive, as suc-
cessful matrix completion would constitute a breach of privacy; our work builds the case for using
simpler lower dimensional representations of the data, and simpler algorithms based on extracting
limited information (in our case, 1-bit sketches) from each user.

2. Preliminaries

We now define our system model, the notion of differential privacy, and tools from information
theory that form the basis of our techniques. We use \([N]\) to denote the set \(\{1, 2, \ldots, N\}\).

2.1. Recommender Systems

In this paper we consider a specific statistical model wherein items are assumed to have an underly-
ing cluster structure, and user affinities for items depend only on the clusters they belong to. In this
setting, the primary objective of the recommender engine is to learn these clusters (and then reveal
them to the users, who can then compute their own recommendations privately). Our model, though
simpler than the state of the art in recommender engines, is still rich enough to account for many of
the features seen empirically in recommender systems. In addition it yields reasonable accuracy in
non-private settings on meaningful datasets (see Tomozei and Massoulié (2011)).

We thus assume that there is an underlying clustering of users and items into several classes,
such that the affinity of a user for an item is only a function of the user’s class and the item’s class
(this is akin to a bipartite version of the Stochastic Blockmodel of Holland et al. (1983), widely used
in model selection literature). Let \([U]\) be the set of \(U\) users and \([N]\) the set of \(N\) items. The set of
users is divided into \(K\) clusters labelled as \(C_u = \{1, 2, \ldots, K\}\), where cluster \(i\) contains \(\alpha_i U\) users.
Similarly, the set of items is divided into \(L\) clusters \(C_n = \{1, 2, \ldots, L\}\), where cluster \(\ell\) contains
\(\beta\ell N\) items. We use \(A\) to denote the matrix of user/item ratings, where each row corresponds to a
user, and each column an item. For simplicity, we assume \(A_{ij} \in \{0, 1\}\); for example, this could
correspond to ‘like/dislike’ ratings. Finally we have the following statistical model for the ratings:
for user \(u \in U\) with user class \(k\), and item \(n \in [N]\) with item class \(\ell\), the rating \(A_{un}\) is given by a
Bernoulli random variable \(A_{un} \sim \text{Bernoulli}(b_{k\ell})\), where the ratings by users in the same class, and
for items in the same class, are i.i.d.

In order to model limited information, i.e., the fact that users rate only a fraction of all items,
we define a parameter \(w\) to be the number of items a user has rated (more generally, we only need
bounds for \(w\)--for example, we could have \(w = \Theta(f(N))\) for some function \(f\)). We assume that the
rated items are picked uniformly at random. We characterize \(w = \Omega(N)\) as the information-rich regime
and \(w = o(N)\) as the information-scarce regime.

When considering lower bounds, we will specialize this model to the situation where there is
only one user class (\(K = 1\)) and where users have perfect knowledge of the type of the items they rate.
2.2. Differential Privacy

Differential privacy is a framework that, in its most general form, defines conditions under which an algorithm can be said to be privacy preserving with respect to the input. Formally we have:

**Definition 1 (ε-Differential Privacy)** A randomized function $\Psi : \mathcal{X} \rightarrow \mathcal{Y}$ that maps data $X \in \mathcal{X}$ to $Y \in \mathcal{Y}$ is said to be $\epsilon$-differentially private (or $\epsilon$-DP) if, for all values $y \in \mathcal{Y}$ in the range space of $\Psi$, and for all ‘neighboring’ data $x, x'$, we have that:

$$\frac{P[Y = y|X = x]}{P[Y = y|X = x']} \leq e^\epsilon$$

We also assume that $Y$ conditioned on $X$ is independent of any external side information $Z$ (in other words, the output of mechanism $\Psi$ depends only on $X$ and its internal randomness). Furthermore, the definition of ‘neighboring’ is chosen according to the situation, and determines the data that remain private. In the context of ratings matrices, two matrices can be neighbors if: i) they differ in a single row (per-user privacy), or ii) if they differ in a single rating (per-rating privacy).

We consider the local model of differential privacy, where privacy is ensured at the user-database boundary before the data is stored in the system. This paradigm is known in statistics as the ‘Randomized Response’ technique [Warner (1965)] (where it is used for collecting statistics for sensitive questions). For each user $u$, let $X$ be its private data—in the recommendation context, the rated-item labels and corresponding ratings—and let $Y$ be the data that the user makes publicly available to the untrusted engine. Then local-DP requires that the above condition holds, where any two private data (ratings vectors in our case) $x$ and $x'$ are deemed neighboring. It is thus the natural notion of privacy in the case of untrusted databases, as the data is privatized at the user-end before storage in the database; to emphasize this, we alternately refer to it as User-end Differential Privacy.

We conclude this section with a mechanism for releasing a single bit under $\epsilon$-differential privacy. The proof of differential privacy for this mechanism is easy to check using equation (1).

**Proposition 2 (ε-DP bit release):** Given a single bit $S^0$, let output bit $S$ be equal to $S^0$ with probability $\frac{e^\epsilon}{1 + e^\epsilon}$, else equal to $S^0 = 1 - S^0$. Then the map $S^0 \rightarrow S$ is (locally) $\epsilon$-differentially private.

2.3. Preliminaries from Information Theory

For a random variable $X$ taking values in some discrete space $\mathcal{X}$, its entropy is defined as $H(X) \triangleq \sum_{x \in \mathcal{X}} -P[X = x] \log P[X = x]$ 2. For two random variables $X, Y$, the mutual information between them is given by:

$$I(X; Y) \triangleq \sum_{(x,y)} P[X = x, Y = y] \log \left( \frac{P[X = x, Y = y]}{P[X = x]P[Y = y]} \right)$$

Our main tools for constructing lower bounds are variants of Fano’s Inequality, which are commonly used in non-parametric statistics literature (refer Santhanam and Wainwright (2009); Wainwright (2009)). Consider a finite hypothesis class $\mathcal{H}$, $|\mathcal{H}| = M$, indexed by $[M]$. Suppose that we choose a hypothesis $H$ uniformly at random from $\{1, 2, \ldots, M\}$, sample a data set $X_1^U$ of $U$ samples drawn in an i.i.d. manner according to a distribution $P_{\mathcal{H}}(H)$ (in our case, $u \in [U]$ corresponds to a user, and $X_u$ the ratings drawn according to the statistical model in Section 2.1),

2. For notational convenience, we use $\log(\cdot)$ as the logarithm to the base 2 throughout; hence, the entropy is in ‘bits’
and then provide a private version of this data $\hat{X}_1^U$ to the learning algorithm. We can represent this as the Markov chain:

\[
H \in \mathcal{H} \xrightarrow{\text{Sampling}} X_1^U \xrightarrow{\text{Privatization}} \hat{X}_1^U \xrightarrow{\text{Model Selection}} \hat{H}
\]

Further, we define a given learning algorithm to be unreliable for the hypothesis class $\mathcal{H}$ (and a hypothesis drawn uniformly at random) if

\[
\max_{h \in [M]} \mathbb{P}[\hat{H} \neq H | H = h] > \frac{1}{2}.
\]

Fano’s inequality provides a lower bound on the probability of error under any learning algorithm in terms of the mutual information between the underlying hypotheses and the samples. A basic version of the inequality is as follows (see Appendix A for a more general version with discussions):

**Lemma 3 (Fano’s Inequality)** Given a hypothesis $H$ drawn uniformly from $\mathcal{H}$, and $U$ samples $X_1^U$ drawn according to $H$, for any learning algorithm, the average probability of error $P_e \triangleq \mathbb{P}[\hat{H} \neq H]$ satisfies:

\[
P_e \geq 1 - \frac{I(H; X_1^U)}{\log(M)} + 1.
\]

(2)

As a direct consequence of this result, if the samples are such that $I(H; X_1^U) = o(\log M)$, then any algorithm fails to correctly identify almost all of the possible underlying models. Though this is a weak bound, equation 2 turns out to be sufficient to study sample complexity scaling in the cases we consider. In Appendix A, we consider stronger versions, as well as more general criterion for approximate model selection (i.e., with distortion).

### 3. Clustering under Local-DP: The Information-Rich Regime

In this section, we derive a lower bound on the number of users needed for accurate learning under local differential privacy. This relies on a simple bound on the mutual information between any database and its privatized output, and hence is applicable in very general settings. Returning to the clustering problem, we give an algorithm that matches the optimal scaling in $N$ (up to some logarithmic factor) under one of the following two conditions: i) $w = \Omega(N)$, i.e., each user has rated a constant fraction of items (the information-rich regime), or ii) only the ratings are private, not the identity of the rated items.

We obtain a simple lower bound on the scaling required using the following lemma that characterizes a lower bound on the mutual information leakage across any differentially private channel. Equivalent statements of this lemma are given in Alvim et al. (2011); McGregor et al. (2010):

**Lemma 4** Given (private data) r.v. $X \in \mathcal{X}$, a privatized output $Y \in \mathcal{Y}$ obtained by any locally $\epsilon-\text{DP}$ mechanism $\Phi : \mathcal{X} \rightarrow \mathcal{Y}$, and any side information $Z$, we have: $I(X; Y|Z) \leq \epsilon \log e$.

Lemma 4 follows directly from the definitions of mutual information and differential privacy (note that for any such mechanism, the output $Y$ given the input $X$ is conditionally independent of any side-information). It suggests that any mechanism obeying DP results in an output which has at most $\epsilon$ bits of information vis-a-vis the data. Returning to the private learning of item classes, we obtain a lower bound on the number of users needed by considering the following reduction: let $C_N \in \mathcal{H} = \{0,1\}^N$ be the mapping of the item set $[N]$ to two classes represented as $\{0,1\}$; hence the size of the hypothesis class is now $2^N$. Recall that we defined a learning algorithm to be unreliable for $\mathcal{H}$ if

\[
\max_{h \in \mathcal{H}} \mathbb{P}[\hat{C}_N \neq C_N | C_N = h] > \frac{1}{2}.
\]

Using Lemma 4 and Fano’s inequality (Lemma 3), we get the following lower bound on the sample complexity.
Theorem 5 Suppose the underlying clustering $C_N$ is drawn uniformly at random from $\{0, 1\}^N$. Then any learning algorithm obeying $\epsilon$-local-DP is unreliable if the number of queries $U < \left(\frac{N}{\epsilon \log e}\right)$.

Proof (Sketch) Similar to the assumptions in Section 2.3, we have the following model for each user (under local-DP): $C_N \xrightarrow{\text{Sampling}} X_u \xrightarrow{\text{Privatization}} \hat{X}_u$, for each $u \in [U]$. Here sampling refers to each user rating a subset of $w$ items. Now by the Data-Processing Inequality (Theorem 2.8.1 from Cover and Thomas (2006), see Appendix A), we have that

$$I(C_N; \hat{X}^U) \leq \sum_{u=1}^U I(X_u; \hat{X}_u | \hat{X}_u - 1) < U\epsilon \log e,$$

by Lemma 4. Fano’s inequality (Lemma 3) then implies that a learning algorithm is unreliable if the number of queries satisfies: $U < \left(\frac{N}{\epsilon \log e}\right)$.

We note that this is a simplified form of a more general theorem presented in Appendix A. Further, though this bound is not the strongest possible, it turns out to be achievable (up to logarithmic factors) in the information-rich regime, as we show below. A similar bound was given by Beimel et al. (2010) for PAC-learning in the centralized setting using more explicit counting techniques. Such bounds fail to exhibit the correct scaling in the information-scarce case ($w = o(N)$) setting. The reason for this is that we use the Data-Processing inequality in the proof of Theorem 5, instead of jointly analyzing the interaction between the two channels (sampling and privatization). However, unlike proofs based on simple counting arguments, our method allows us to leverage more sophisticated information theoretic tools for other variants of the problem, like those we consider subsequently in Section 4.

To conclude this section, we outline an algorithm for clustering in the information-rich regime. The algorithm proceeds as follows: i) provide each user $u$ with two items $(i_u, j_u)$ picked at random whereupon the user generates a private bit $S^0_u$ equal to 1 if it rated the two items positively, and else 0, ii) let users release as a public sketch a privatized version $S_u$ of their private bit using the $\epsilon$-DP bit release mechanism, iii) construct matrix $\hat{A}$ whose $(i, j)$ entry is obtained by adding the sketches $S_u$ of each user $u$ queried with item pair $(i, j)$, and iv) perform spectral clustering of items based on matrix $\hat{A}$ and return the item classes. We refer to this as the Pairwise-Preference algorithm. The algorithm is formally specified in Appendix B. Its privacy is guaranteed by the use of $\epsilon$-DP bit release, while its performance analysis, given in Theorem 6, is based on a related result on spectral clustering by Tomozei and Massoulié (2011); details are given in Appendix B (in particular, the detailed separability conditions are given in Theorem 21).

Theorem 6 The Pairwise-Preference algorithm is $\epsilon$-differentially private. Further, in the information-rich regime, under the separability assumptions on the model parameters $(\alpha_k), (\beta_\ell)$ and $(b_{k\ell})$ stated in Appendix B, there exists $c > 0$ such that the item clustering is successful with high probability if the number of users satisfies: $U \geq c(N \log N)$.

4. The Information-scarce Setting: Lower Bounds

To get tighter lower bounds on the number of users needed to obtain an accurate item clustering, we need more accurate bounds on the mutual information between the underlying model expressed in terms of item clusters and the available privatized data. In Section 3 we developed a basic lower

---

8
bound by characterizing a constraint on the mutual information across any differentially private channel. We now develop some more refined techniques to study the impact of privatization in the presence of incomplete information.

As in the previous lower bound, we consider a simplified version of the problem, where there is a single class of users, and each item is ranked either 0 or 1 deterministically by each user (i.e., \( b_{ui} = b_i \in \{0, 1\} \) for all items). Let \( C_N(\cdot) : [N] \rightarrow \{0, 1\} \) be the underlying clustering function; in general we can think of this as an \( N \)-bit vector \( \mathbf{Z} \in \{0, 1\}^N \). We assume that the user-data for user \( u \) is given by \( X_u = (I_u, Z_u) \), where \( I_u \) is a size \( w \) subset of \([N]\) representing items rated by user \( u \), and \( Z_u \) are the ratings for the corresponding items; in this case, \( Z_u = \{Z(i)\}_{i \in I_u} \). The set \( I_u \) is assumed to be chosen uniformly at random from amongst all size-\( w \) subsets of \([N]\). We also denote the privatized sketch from user \( u \) as \( S_u \in \mathcal{S} \). Here the space \( \mathcal{S} \) to which sketches belong is assumed to be an arbitrary finite or countably infinite space. The sketch is assumed \( \epsilon \)-differentially private. Finally, as before, we assume that \( \mathbf{Z} \) is chosen uniformly over \( \{0, 1\}^N \).

### 4.1. Local Differential Privacy and Mutual Information

In this section, we establish the main lemma we use for bounding the mutual information under local-DP, and derive a result for the sample complexity of learning with 1-bit sketches, which builds intuition regarding the bounds in the next section.

We define \( \binom{N}{w} \) to be the collection of all size-\( w \) subsets of \([N] = \{1, 2, \ldots, N\} \), \( \mathcal{D} \triangleq \binom{[N]}{w} \times \{0, 1\}^w \) to be the set from which user information (i.e., \((I, Z)\)) is drawn, and define \( D = |\mathcal{D}| = \binom{N}{w} 2^w \). Finally \( \mathbb{E}_{X}[\cdot] \) indicates that the expectation is over the random variable \( X \). We now establish the following bound for the mutual information between the model and the sketch. This is a special case (for \( \mathbf{Z} \) taking the uniform measure over \( \{0, 1\}^N \)) of a more general lemma which we state and prove in Appendix C.

**Lemma 7** Given the Markov Chain \( \mathbf{Z} \rightarrow (I, Z) \rightarrow S \), let \((I_1, Z_1), (I_2, Z_2) \in \mathcal{D} \) be two pairs of ‘user-data’ sets which are independent and identically distributed according to the conditional distribution of the pair \((I, Z)\) given \( S = s \). Then, the mutual information \( \mathcal{I}(\mathbf{Z}; S) \) satisfies:

\[
\mathcal{I}(\mathbf{Z}; S) \leq \mathbb{E}_S \left[ \mathbb{E}_{(I_1, Z_1) \mid S \cup (I_2, Z_2) \mid S} \left[ 2^{|I_1 \cap I_2|} \mathbb{I}_{\{Z_1 \equiv Z_2\}} - 1 \right] \right],
\]

where we use the notation \( \mathbb{I}_{\{Z_1 \equiv Z_2\}} \) to denote that the two user-data sets are consistent on the index set on which they overlap, i.e., \( \mathbb{I}_{\{Z_1 \equiv Z_2\}} \triangleq \mathbb{I}_{\{Z_1(\ell) = Z_2(\ell) \forall \ell \in I_1 \cap I_2\}} \).

Before deriving tighter lower bounds under local-DP, we first consider a related problem that demonstrates the effect of per-user constraints (as opposed to average constraints) on the mutual information. We consider the same item-class learning problem as before with \( w = 1 \) (i.e., each user has access to one rating), but instead of a privacy constraint, we consider a ‘per-user bandwidth’ constraint, wherein each user can communicate only a single bit to the learning algorithm.

This demonstrates an interesting change in the sample complexity of learning with per-user communications constraints (maximum bandwidth in this section, and privacy in next section) versus average-user constraints (mutual information bound or average bandwidth). In the former case as we will show, the sample complexity is \( \Theta(N^2) \). In the latter case, the sample complexity with 1-bit average bandwidth constraint is \( O\left(N \log^2 N\right) \). Indeed, assume \( w = 1 \), and let users reveal their private data \((I, Z)\) with probability \( 1/\log(N) \) and otherwise return a blank symbol. Then the average information released per user is \( O(1) \), and by a coupon collector argument the original sequence \( Z_1^N \) is indeed retrieved after \( O(N \log^2 N) \) queries.
Theorem 8 Suppose \( w = 1 \), with \((I, Z)\) drawn i.i.d uniformly over \([N] \times \{0, 1\}\). Then for any 1-bit sketch derived from \((I, Z)\), it holds that: \( \mathcal{I}(Z, S) = O\left(\frac{w^2}{N}\right) \), and consequently, there exists a constant \( c > 0 \) such that any cluster learning algorithm using queries with 1-bit responses is unreliable if the number of users satisfies \( U < cN^2 \).

Proof (Sketch.) We first note that \( \mathcal{I}(Z, S) \) is a convex function of \( \mathbb{P}[S = s | Z = z] \) for fixed \( \mathbb{P}[Z = z] \) (Theorem 2.7.4, Cover and Thomas (2006)). Thus, the mutual information is maximized at the extremal points of the kernel \( \mathbb{P}[S = s | Z = z] \) which correspond to \( \mathbb{P}[S = s](i, z) \in \{0, 1\} \), implying that the class of deterministic queries with 1-bit response that maximizes mutual information has the following structure: given user-data \((I_u, Z_u)\), user \( u \)'s response \( S_u \in \{0, 1\} \) is of the form \( S_u = 1_A(I_u, Z_u) \), where \( A \subseteq \{(i, z) | i \in [N], z \in \{0, 1\}\} \). In other words, the algorithm provides user \( u \) with an arbitrary set \( A \) of \((\text{items}, \text{ratings})\), and the user identifies if \((I_u, Z_u)\) is contained in \( A \). The mutual information lower bound follows from elementary manipulations. We then get the result from Lemma 7 and Fano’s inequality (Lemma 3).

We note that this is a tight bound—a simple (adaptive) scheme is to ask random queries of the form “Is \((I, Z) = (i, b)\)”? (where \( i \in [N] \) and \( b = \{0, 1\} \)). The average time between two successful queries is \( 2N \), and one needs \( N \) successful queries to learn all the bits.

4.2. Query Complexity Lower Bounds for Clustering under Local-DP

We now exploit the above techniques to obtain lower bounds on the scaling required for accurate clustering with DP in an information-scarce regime, i.e., when \( w = o(N) \). We first obtain a weak lower bound in Theorem 9, valid for all \( w \), and then refine it in Theorem 10 under some additional conditions. Refer to Appendix C for the complete proofs.

Theorem 9 In the information-scarce regime, i.e., when \( w = o(N) \), under \( \epsilon\)-local-DP we have:

\[
\mathcal{I}(Z, S) = O\left(\frac{w^2}{N}\right)
\]

and consequently, there exists a constant \( c > 0 \) such that any cluster learning algorithm with \( \epsilon\)-local-DP is unreliable if the number of users satisfies \( U < c\left(\frac{N^2}{w^2}\right) \).

The above result shows how Lemma 7 can be used to obtain sharper bounds on the mutual information contained in a differentially private sketch in the information-scarce setting in comparison to Lemma 4. Using this result, we get a lower bound of \( \Omega\left(\frac{N^2}{w^2}\right) \) on the number of samples needed to learn the underlying clustering. We now present a tighter bound on the mutual information under some conditions; it relies on a more careful evaluation of the bound in Lemma 7, but matches the performance of the algorithm we present in Section 5, thereby displaying its optimality.

Theorem 10 Under the scaling assumption \( w = o(N^{1/3}) \), and for \( \epsilon < \ln(2) \), it holds that

\[
\mathcal{I}(Z, S) = O\left(\frac{w}{N}\right).
\]

and thus there exists a constant \( c > 0 \) such that any cluster learning algorithm with \( \epsilon\)-local-DP is unreliable if the number of users satisfies \( U < c\left(\frac{N^2}{w}\right) \).
5. The Information-scarce Setting: Cluster Learning

The sample complexity of the pairwise-preference algorithm in Section 3 does not match our lower bounds in an information-scarce setting. Indeed, the probability that two randomly probed items belong to the rated set of size \( w \) is \( O(w^2/N^2) \). The sample complexity is thus magnified from \( \Omega(N \log(N)) \) in the information-rich regime to \( \Omega(N^3 \log(N)/w^2) \), which is polynomially larger than our lower bound for \( w = o(\sqrt{N}) \). We thus turn to the design of a new algorithm that achieves the sample complexity bound from Section 4.

**The MaxSense algorithm:** As in Pairwise Preference, we use a (privatized) 1-bit sketch for learning. A query to user \( u \) is formed by first constructing a random sensing vector \( H_u = (H_{un})_{n \in [N]} \), whose entries \( H_{un} = 1 \) if item \( n \) is being sensed, and 0 otherwise; each entry is set to 1 in an i.i.d. manner with probability \( \theta/w \) for some design parameter \( \theta \). User \( u \) then constructs a private sketch \( S^0_u \), which is the disjunction of its ratings for all items \( n \) that are being sensed (with unrated items given rating of 0): \( S^0_u = \max_{n \in [N]} H_{un}Z_{un} \), where \( Z_{un} \in \{0, 1\} \) equals 1 if user \( u \) rated positively item \( n \). Finally, user \( u \) outputs a privatized version \( S'_u \) of its private sketch \( S^0_u \). The sensing vector \( H_u \) is known publicly, hence can be generated either by the user or by the engine querying the user.

Based on the sketches \( S_u \) and sensing vectors \( H_u \), the algorithm then determines per-item scores \( X_n \) according to \( X_n := \sum_{u \in [U]} H_{un}S_u, n \in [N] \), and performs \( k \)-means clustering of these scores in \( \mathbb{R} \). A formal description of the algorithm is provided in Appendix D. Now we have the following:

**Theorem 11** The MaxSense algorithm is \( \epsilon \)-differentially private. Further, define

\[
\tilde{\epsilon} = \frac{2(\epsilon^e - 1)}{(\epsilon^e + 1)} \quad \text{and} \quad \delta_{\min} = \min_{1 \leq \ell < \ell' \leq L} \left| \sum_{k=1}^{K} \alpha_k e^{-\theta \sum_{\ell=1}^{L} \beta_k \ell (b_{k\ell} - b_{k\ell'})} \right|
\]

where \( \theta \) is the parameter of the item sensing probability \( \theta/w \). Then for any \( d > 0 \), there exists a constant \( C > 0 \) such that the clustering is successful with probability \( 1 - N^{-d} \) if the number of users satisfies:

\[
U \geq C \left( \frac{N^2 \log N}{\epsilon^2 \delta_{\min}^2 w} \right).
\]

\( \delta_{\min} \) here determines separability conditions on the problem: for example, using the notation \( v_k := \sum_{\ell} \beta_{k\ell} \), it can be checked that \( \delta_{\min} \) is strictly positive for all \( \theta \) (except on a set of measure 0) provided the following condition holds:

\[
\forall \ell \neq \ell' \in [L], \exists k \in [K] \text{ such that } \sum_{j:v_{jk} = v_{k\ell'}} \alpha_j (b_{j\ell} - b_{j\ell'}) \neq 0. \tag{4}
\]

Determining whether alternative schemes could achieve similar complexity under weaker separability conditions is, for now, an open problem.

6. Extensions and Conclusion

Theorem 11 demonstrates that despite its simplicity, MaxSense is sufficient to achieve optimal scaling in \( N \) (up to logarithmic terms) under suitable separability condition. More generally, the algorithm suggests a general approach to dealing with partial information under local differential privacy. We now briefly discuss an extension to achieve a better ‘privacy trade-off’, namely, a \( \frac{1}{\ell} \) factor in the scaling required for accurate clustering. Under this extension, each user is asked
\( Q = \lceil \epsilon^{-1} \rceil \) MaxSense questions, each with a privacy parameter of \( \frac{\epsilon}{Q} \) in a way that ensures independence between answers. The user calculates \( Q \) sketches using the \( Q \) sensing vectors and reveals the privatized set of sketches (with each sketch being revealed via a \( \frac{\epsilon}{Q} \)-DP bit release mechanism). Finally, we calculate the item counts and perform clustering as before. The algorithm, which we call the Multi-MaxSense algorithm, is formally presented and analyzed in the Appendix.

Adaptive queries: The lower bounds of Section 4 applied to non-adaptive learning, where queries to users are performed in parallel, without leveraging answers of users 1, ..., \( u - 1 \) when querying user \( u \). One can in fact extend these bounds to the adaptive setting where query to user \( u \) is allowed to depend on the previous queries and answers of users 1, ..., \( u - 1 \). Specifically the following, shown in the Appendix E, holds.

**Theorem 12** Assume \( w = 1 \). If users’ answers are \( \epsilon \)-DP, the number of adaptive queries needed to learn unknown content clustering into two types drawn uniformly at random from \( \{0, 1\}^N \) is \( \Omega(N \log N) \).

The proof again relies on bounding the mutual information between the unknown clusters and a user’s sketch, although now the mutual information conditional on the previous queries and their answers (i.e., of the form \( I(Z, S_u | S_1^{u-1} = s_1^{u-1}) \)) has to be considered. The first step applies an extension of Lemma 7 to bound this mutual information by the variance of a certain empirical sum \( N^{-1} \sum_{n=1}^{N} f_n(Z_n) \) for bounded functions \( f_n \), under the distribution of \( Z \) conditional on \( S_1^{u-1} = s_1^{u-1} \). The crux of the proof then consists in showing that, provided this conditional distribution is close to uniform (i.e., its entropy is \( \geq N - \delta \) for some \( \delta > 0 \)), then the variance of this empirical sum under the conditional distribution is no larger than \( N^{-1} g(\delta) \) for some constant \( g(\delta) \). This intermediate result is of independent interest, and could enable extensions of the latter theorem, e.g. relaxing the assumption that \( w = 1 \).

We leave it as a topic for further research to establish how sharp this lower bound is. In particular, if it can be tightened to a lower bound of \( \Omega(N^2) \) and further extended to \( \Omega(N^2/w) \) for \( w \neq 1 \), this would imply that MaxSense is optimal even when one can use adaptive queries. If on the other hand there is a gap between non-adaptive and adaptive complexities, then this implies that schemes superior to MaxSense in the adaptive case have yet to be identified.

In conclusion, we have initiated a study in the design of recommender systems under local-DP constraints. We have provided lower bounds on the sample complexity in both information-rich and information-scarce regime, quantifying the effect of limited information on private learning. Further, we showed tightness of these results by designing the MaxSense algorithm, which recovers the item clustering under privacy constraints with optimal sample complexity. The lower bound techniques naturally extend to cover model selection for more general (finite) hypothesis classes, while 1-bit sketches appear appropriate for designing efficient algorithms for the same. Development of such algorithms and analysis of matching lower bounds by leveraging and extending the techniques we introduced seem promising future research directions.
References
Mário S. Alvim, Miguel E. Andrés, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. Quantitative Information Flow and Applications to Differential Privacy. In FOSAD, pages 211–230, 2011.

Amos Beimel, Shiva Prasad Kasiviswanathan, and Kobbi Nissim. Bounds on the Sample Complexity for Private Learning and Private Data Release. In TCC 2010, pages 437–454, 2010.

Avrim Blum, Cynthia Dwork, Frank McSherry, and Kobbi Nissim. Practical privacy: the SuLQ framework. In Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, PODS ’05, pages 128–138, 2005.

Avrim Blum, Katrina Ligett, and Aaron Roth. A Learning Theory Approach to Non-interactive Database Privacy. In STOC, pages 609–618, 2008.

Joseph A. Calandrino, Ann Kilzer, Arvind Narayanan, Edward W. Felten, and Vitaly Shmatikov. "You Might Also Like:" Privacy Risks of Collaborative Filtering. In IEEE Symposium on Security and Privacy, pages 231–246, 2011.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate. Differentially Private Empirical Risk Minimization. Journal of Machine Learning Research, 12, 2011.

Thomas M. Cover and Joy A. Thomas. Elements of information theory (2. ed.). Wiley, 2006.

Cynthia Dwork. Differential Privacy. In ICALP, pages 1–12, 2006.

Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In STOC, pages 371–380, 2009.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating Noise to Sensitivity in Private Data Analysis. In TCC, pages 265–284, 2006.

Cynthia Dwork, Moni Naor, Toniann Pitassi, and Guy N. Rothblum. Differential Privacy under Continual Observation. In STOC, pages 715–724, 2010a.

Cynthia Dwork, Moni Naor, Toniann Pitassi, Guy N. Rothblum, and Sergey Yekhanin. Pan-Private Streaming Algorithms. In ICS, pages 66–80, 2010b.

Anupam Gupta, Moritz Hardt, Aaron Roth, and Jonathan Ullman. Privately Releasing Conjunctions and the Statistical Query Barrier. In STOC, pages 803–812, 2011.

Paul W. Holland, Kathryn B. Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. Social Networks, 5(2):109–137, 1983.

Shiva Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What Can We Learn Privately? In Proc. IEEE FOCS, 2008.

Raghunandan H. Keshavan, Andrea Montanari, and Sewoong Oh. Matrix Completion from Noisy Entries. Journal of Machine Learning Research, 11:2057–2078, 2010.
Andrew McGregor, Ilya Mironov, Toniann Pitassi, Omer Reingold, Kunal Talwar, and Salil Vadhan. The Limits of Two-Party Differential Privacy. In *FOCS*, pages 81 – 90, oct. 2010.

Frank McSherry and Ilya Mironov. Differentially Private Recommender Systems: Building Privacy into the Netflix Prize Contenders. In *KDD*, pages 627 – 636, 2009.

Arvind Narayanan and Vitaly Shmatikov. How To Break Anonymity of the Netflix Prize Dataset. *CoRR*, abs/cs/0610105, 2006.

Narayana P. Santhanam and Martin J. Wainwright. Information-theoretic Limits of Selecting Binary Graphical Models in High Dimensions. *CoRR*, abs/0905.2639, 2009.

Adam Smith. Privacy-preserving Statistical Estimation with Optimal Convergence Rates. In *Proceedings of the 43rd annual ACM symposium on Theory of computing*, STOC ’11, pages 813 – 822, 2011.

Dan-Cristian. Tomozei and Laurent Massoulié. Distributed User Profiling via Spectral Methods. *ArXiv e-prints*, 1109.3318, September 2011.

Martin J. Wainwright. Information-theoretic Limits on Sparsity Recovery in the High-dimensional and Noisy Setting. *IEEE Trans. Inf. Theor.*, 55:5728 – 5741, December 2009.

Stanley L. Warner. Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. *Journal of the American Statistical Association*, 60(309):63 – 69, March 1965.

**Appendix A. Differential Privacy and Mutual Information**

In this section we state and prove some basic results instrumental for our lower bounds.

**A.1. Differential Privacy and Mutual Information**

For the sake of convenience, we restate the definition of Differential Privacy:

**Definition 13 (ε-Differential Privacy)** A randomized function \( \Psi : \mathcal{X} \rightarrow \mathcal{Y} \) that maps data \( X \in \mathcal{X} \) to \( Y \in \mathcal{Y} \) is said to be \( \varepsilon \)-differentially private (or \( \varepsilon \)-DP) if, for all values \( y \in \mathcal{Y} \) in the range space of \( \Psi \), and for all ‘neighboring’ data \( x, x' \), we have that:

\[
\frac{\mathbb{P}[Y = y|X = x]}{\mathbb{P}[Y = y|X = x']} \leq e^{\varepsilon}
\]

In this work, we focus on the local model of differential privacy. The local model is formally defined in Kasiviswanathan et al. (2008). Informally, in the context of recommender systems, this means that the ratings of each user are assumed to be private data, and hence any information given by any user to the system is required to obey the above definition vis-a-vis the user’s private data. The definition of neighboring databases for checking differential privacy depends on the exact information that needs privatization. In particular, we can consider the following two cases in the context of recommender systems:

1. The items rated by a user are not considered private, but the ratings are. Now, given a set of \( w \) rated items with ratings (with, say, \( \{0, 1\} \) ratings), the neighboring databases are all possible ratings vectors for these \( w \) items (hence, all vectors in \( \{0, 1\}^w \)).
2. Both ratings as well as rated items are considered private. Now, the neighbors of any set of
rated items and ratings consists of all possible subsets of items, and all possible sets of ratings
for this subset.

We have considered the latter case throughout the paper. However, as mentioned in Section 3, the
first case (where only ratings are private) can be handled by the Pairwise Preference algorithm.
Furthermore, the basic lower bound of Section 3, Lemma 4 is also applicable, thereby giving a
complete characterization of that case.

Two crucial properties of differential privacy, which we use later in our proofs, are ‘composition’
and ‘post-processing’ (refer to Dwork et al. (2006) for details). Composition defines how the
privacy of the data scales upon the application of multiple differentially-private release mechanisms.
Formally we have:

**Proposition 14 (Composition)** If \( k \) outputs, \( \{Y_1, Y_2, \ldots, Y_k\} \) are obtained from data \( X \in \mathcal{X} \) by
\( k \) different randomized functions, \( \{\Psi_1, \Psi_2, \ldots, \Psi_k\} \), where \( \Psi_i \) is \( \epsilon_i \)-differentially private, then the
resultant function is \( \sum_{i=1}^{k} \epsilon_i \) differentially private.

The post-processing property implies that processing the output of a differentially private release
mechanism can only make it more differentially private (i.e., with a smaller \( \epsilon \)) vis-a-vis the private
input. Formally:

**Proposition 15 (Post-processing)** If a function \( \Psi_1 : \mathcal{X} \rightarrow \mathcal{Y} \) is \( \epsilon \)-differentially private, then any
composition function \( \Psi_2 \circ \Psi_1 : \mathcal{X} \rightarrow \mathcal{Z} \) is \( \epsilon' \)-differentially private for some \( \epsilon' \leq \epsilon \).

Before we derive a basic bound on the mutual information leaked across a differentially private
channel, we need to state one important property of mutual information that we use repeatedly in
our proofs. The Data-Processing inequality (see Cover and Thomas (2006) for details) states that
mutual information decreases upon further processing. Formally we have:

**Proposition 16 (Data-Processing Inequality)** For random variables \( X, Y, Z \) forming a Markov
chain \( X \rightarrow Y \rightarrow Z \), we have that:

\[ I(X; Z) \leq I(Y; Z) \]

Finally we give a proof for Lemma 4. Similar results have been presented by McGregor et al.
(2010); Alvim et al. (2011).

**Lemma** (Lemma 4 in the paper) Given a random variable \( X \in \mathcal{X} \) of user’s private data, a priva-
tized output \( Y \in \mathcal{Y} \) obtained by any \( \epsilon \)-local-DP mechanism \( \Phi : \mathcal{X} \rightarrow \mathcal{Y} \), and any side information
\( Z \), we have:

\[ I(X; Y | Z) \leq \epsilon \log e . \]
### Proof

\[
I(X; Y | Z) = \sum_{(x, y) \in \mathcal{X} \times \mathcal{Y}} p(x, y | Z) \log \left( \frac{p(x, y | Z)}{p(x | Z) p(y | Z)} \right)
\]

\[
= \sum_{(x, y) \in \mathcal{X} \times \mathcal{Y}} p(x, y | Z) \log \left( \frac{p(y | x, Z)}{\sum_{x' \in \mathcal{X}} p(x' | Z) p(y | x', Z)} \right)
\]

\[
= \sum_{(x, y) \in \mathcal{X} \times \mathcal{Y}} p(x, y | Z) \log \left( \frac{1}{\sum_{x' \in \mathcal{X}} p(x' | Z) (p(y | x', Z)/p(y | x, Z))} \right)
\]

\[
\leq \sum_{(x, y) \in \mathcal{X} \times \mathcal{Y}} p(x, y | Z) \log \left( \frac{1}{\sum_{x' \in \mathcal{X}} p(x' | Z) e^{-\epsilon}} \right)
\]

\[
\leq \epsilon \log e.
\]

Where inequality \((a)\) is a direct application of the definition of differential privacy (Equation 1), and in particular, the fact that it holds for any side information. \[\square\]

### A.2. Sample-Complexity Lower Bounds for Private Learning

The main tool we use for deriving lower bounds is Fano’s inequality (Lemma 3); in this section we state and derive some stronger forms of the same. The item clustering problem fits in a more general framework of model selection from finite hypothesis classes, with local-DP constraints: we consider a hypothesis class \(\mathcal{H}, |\mathcal{H}| = M\), indexed by \([M]\). Given a hypothesis \(Z\), samples \(X_1^U\) are drawn in an i.i.d. manner according to some distribution \(P_\mathcal{H}(Z)\) (in our case, \(u \in [U]\) corresponds to a user, and \(X_u\) the ratings drawn according to the statistical model in Section 2.1. \(P_\mathcal{H}(Z)\) thus includes both the sampling of items by a user, as well as the ratings given for the sampled items). Let \(\widehat{X}_1^U\) be a privatized version of this data, where for each \(u \in [U]\), the output \(\widehat{X}_u\) is \(\epsilon\)-differentially private with respect to the data \(X_u\) (by local-DP). Note here that \(X_u\) and \(\widehat{X}_u\) need not belong to the same space (for example, in the case of the Multi-MaxSense algorithm, \(X_u\) is a subset of items and their ratings, while \(\widehat{X}_u\) is the collection of privatized responses to the multiple MaxSense queries). Note also that the probability transition kernel \(P_\mathcal{H}\) can be known to the algorithm (although the exact model \(Z\) is unknown). Finally the learning algorithm infers the underlying model from the privatized samples. We can represent this as the Markov chain:

\[
Z \in \mathcal{H} \xrightarrow{\text{Sampling}} X_1^U \xrightarrow{\text{Privatization}} \widehat{X}_1^U \xrightarrow{\text{Model Selection}} \widehat{Z}
\]

In the paper, we considered an algorithm successful only if \(\widehat{Z} = Z\), i.e., the model is identified perfectly. A natural relaxation of this is in terms of a distortion metric, as follows: given a distance function \(d : \mathcal{Z} \times \mathcal{Z} \rightarrow \mathcal{R}_+\), we say the learner is successful if, for a given \(d > 0\), we have:

\[
d(Z, \widehat{Z}) \leq d.
\]

For any \(h \in \mathcal{H}\), we define the set \(B_d(h) \triangleq \{h' \in \mathcal{H} | d(h, h') \leq d\}\). Further, we define \(M_d = \max_{h \in \mathcal{H}} |B_d(h)|\) to be the largest size of such a set. Finally, given a distribution for \(Z\), we define the average error probability \(P_\epsilon\) for a learning algorithm for the hypothesis class \(\mathcal{H}\) as:

\[
P_\epsilon = \mathbb{P} \left[ d(\widehat{Z}, Z) > d \right].
\]
Then we have the following bound on $P_e$:

**Lemma 17 (Generalized Fano’s Inequality)** Given a hypothesis $Z$ drawn uniformly from $\mathcal{H}$, for any learning algorithm, the average error probability satisfies:

$$P_e \geq 1 - \frac{I(Z; \hat{X}_U^U) + 1}{\log M - \log M_d}.$$  

**Proof** First, we define an error indicator $E$ as:

$$E = \begin{cases} 
1 & : d(Z, \hat{Z}) > d \\
0 & : \text{otherwise}
\end{cases},$$

and hence $P_e = \mathbb{P}[E = 1]$. Define $H(x) = -x \log(x) - (1 - x) \log(1 - x)$. Now we have:

$$I(Z; \hat{X}_1^U) \geq I(Z; \hat{Z})$$

(By the Data Processing Inequality)

$$= H(Z) - H(Z|\hat{Z})$$

$$\geq \log M - H(Z|\hat{Z}, E) - H(E|\hat{Z})$$

(Since $Z$ is uniform over $\mathcal{H} \equiv [M]$, and via basic information inequalities)

$$\geq \log M - P_e H(Z|\hat{Z}, E = 1) - (1 - P_e) H(Z|\hat{Z}, E = 0) - 1$$

(Since $H(P_e) \geq H(E|\hat{Z})$ and $H(P_e) \leq 1$)

$$\geq (1 - P_e)(\log M - H(Z|\hat{Z}, E = 1)) - 1$$

(Since $H(Z|\hat{Z}, E = 1) \leq \log M$)

$$\geq (1 - P_e)(\log M - \log M_d) - 1$$

(Since $H(Z|\hat{Z}, E = 0) \leq \log |B_d(\hat{Z})| \leq \log M_d$)

Rearranging, we have:

$$P_e \geq 1 - \frac{I(Z; \hat{X}_1^U) + 1}{\log M - \log M_d}.$$  

We now have two immediate corollaries of this lemma. First, we consider the non-adaptive learning case, i.e., where the data of each user $\hat{X}_u$ is obtained in an i.i.d manner. Then we have:

**Corollary 18** Given a hypothesis $Z$ drawn uniformly from $\mathcal{H}$, for any non-adaptive learning algorithm, the number of users satisfies:

$$P_e \geq 1 - \left(\frac{UI(Z; \hat{X}_u) + 1}{\log M - \log M_d}\right).$$
Before moving on, we note that these results do not imply that we are assuming a prior on the hypothesis class for our algorithms; rather, the lower bound can be viewed as a probabilistic argument that shows that below a certain sample complexity, any learner fails to learn a large fraction of models. In this light, a stronger restatement of the above result is: if \( U = o \left( \frac{\log M - \log M_d}{Z(Z; \hat{X}_u)} \right) \), then \( P_e \to 0 \) and hence any learning algorithm fails to learn the underlying hypothesis even up to a distortion of \( d \) for almost all models in the hypothesis class.

Next, using Lemma 4, we get a bound on the sample complexity of learning under local-DP.

**Corollary 19** Given a hypothesis \( Z \) drawn uniformly from \( \mathcal{H} \), for any learning algorithm on \( U \) privatized samples, each obtained via \( \epsilon \)-local-DP, the average error probability satisfies:

\[
P_e \geq 1 - \frac{1}{\ln 2} \left( \frac{U \epsilon + 1}{\log M - \log M_d} \right).
\]

Returning to our problem of learning item clusters, we note that \( M = \frac{K^N}{K!} \) in that case. Further, by choosing \( d \) as the edit distance (Hamming distance) between two clusterings of items (i.e., for two clusterings \( C_N \) and \( C'_N \), \( d(C_N, C'_N) \) is the the number of items that are mapped to different clusters in the two clusterings), we get that:

\[
M_d = \frac{1}{K!} \sum_{i=0}^{d} \binom{N}{i} (K - 1)^i \\
= \frac{K^N}{K!} \mathbb{P} \left[ \text{Binomial}(N, 1/K) \geq N - d \right] \\
\leq \frac{K^N}{K!} \exp \left( \frac{-NK(1 - \frac{d}{N} - \frac{1}{K})^2}{3} \right)
\]

Now, using the above results, we can derive a more general version of Theorem 5.

**Theorem 20** Suppose the underlying clustering \( C_N(\cdot) : [M] \to [K] \) is drawn uniformly at random from \( \{0, 1\}^N \). Further, for a given tolerance \( d > 0 \) and error threshold \( p_{\text{max}} \), we define a learning algorithm to be unreliable for the hypothesis class \( \mathcal{H} \) if:

\[
\max_{h \in [M]} \mathbb{P} \left[ d(\hat{Z}, Z) > d \right] > p_{\text{max}}.
\]

Then any learning algorithm that obeys \( \epsilon \)-local-DP is unreliable if the number of queries \( U \) satisfies:

\[
U < \left( 1 - p_{\text{max}} \right) \left( \frac{NK(1 - \frac{d}{N} - \frac{1}{K})^2}{3\epsilon} \right).
\]

**Appendix B. The Pairwise Preference Algorithm**

In this appendix, we define and analyze the Pairwise Preference Algorithm for identifying the item classes in the information-rich regime. The algorithm is formally specified in Algorithm 1.

**Theorem 21** (Theorem 8 in the paper) The Pairwise-Preference algorithm is \( \epsilon \)-differentially private. Further, in the information-rich regime, suppose we have the following non-degeneracy conditions on the eigenvalues and eigenvectors of \( \hat{A} \):
• The $L$ largest magnitude eigenvalues of $A$ have distinct absolute values.

• The corresponding eigenvectors $y_1, y_2, \ldots, y_L$ (normalized under the $\alpha$-norm, which is defined as $||y||_\alpha^2 = \sum_k \alpha_k y_k^2$) satisfy:

\[ t_k \neq t_l , 1 \leq k < l \leq L \]

where $t_k \triangleq (y_1(k), \ldots, y_L(k))$.

Then there exists $c > 0$ such that the item clustering is successful with high probability if the number of users satisfies:

\[ U \geq c (N \log N) . \]

**Proof** As mentioned before, privacy for the algorithm is guaranteed by the use of $\epsilon$-DP bit release (Proposition 2), and the composition property of DP (Proposition 14).

---

**Algorithm 1 The Pairwise-Preference Algorithm**

**Setting:** $N$ items $[N]$, $U$ users $[U]$. Each user has a set of $w$ ratings $(W_u, R_u), W_u \in \binom{[N]}{w}, R_u \in \{0, 1\}^w$. Each item $i$ is associated with a cluster $C_N(i)$ from a set of $L$ clusters, $\{1, 2, \ldots, L\}$

**Output:** The cluster labels of each item, i.e., $\{C_N(i)\}_{i \in [N]}$

**Stage 1 (User sketch generation):**

- For each user $u \in [U]$, the algorithm picks a pair of items $P_u = \{i_u, j_u\}$:
  - At random if $w = \Omega(N)$
  - If $W_u$ is known, it picks a random set of two rated items.

- User $u$ generates a private sketch $S^0_u$ given by:

\[ S^0_u(P_u, R_u) = \begin{cases} 
1 & : R_u(i_u) = R_u(j_u) = 1 \\
0 & : \text{otherwise}
\end{cases}, \]

Where $R_u(i) = 1$ if $i \in W_u$ and item $i$ is rated positively, and 0 otherwise.

**Stage 2 (User sketch privatization):** Each user $u \in [U]$ releases a privatized sketch $S_u$ from $S^0_u$ using the $\epsilon$-DP bit release mechanism (Proposition 2).

**Stage 3 (Spectral Clustering):**

- Generate a pairwise-preference matrix $\hat{A}$, where:

\[ \hat{A}_{ij} = \sum_{u \in U(P_u = \{i, j\})} S_u \]

- Extract the top $L$ normalised eigenvectors $x_1, x_2, \ldots, x_L$ corresponding to the $L$ largest magnitude eigenvalues of matrix $\hat{A}$. Embed each row (node) into $L$-dimensional Euclidean space by assigning as coordinates the corresponding entries of the $L$ eigenvectors.

- Perform k-means clustering in the profile space to get the item clusters
We will prove the sample complexity bound for the case where \( w = \Omega(N) \), as the other case (where the rated items are not private) follows similarly. From the definition of the \( \epsilon \)-DP bit release mechanism, we have that:

\[
\mathbb{P}[S_u = 1] = \frac{1 + (e^{\epsilon} - 1)\mathbb{P}[S_u^0 = 1]}{e^{\epsilon} + 1},
\]

and thus for any pair of items \( \{i, j\} \), defining \( b_{ij} \triangleq \sum_{k=1}^K \alpha_k (b_{ki} b_{kj} + (1 - b_{ki})(1 - b_{kj})) \) (i.e., the probability that a random user has identical preference for items \( i \) and \( j \)) and \( \overline{b_{ij}} = 1 - b_{ij} \), we have:

\[
\mathbb{P}[S_u = 1, P_u = \{i, j\}] = \frac{1}{N(N-1)} \left( \frac{1}{e^{\epsilon} + 1} + \left( \frac{e^{\epsilon} - 1}{e^{\epsilon} + 1} \right) \frac{w(w-1)}{N(N-1)} b_{ij} \right) \triangleq \frac{b_{ij}'}{N(N-1)},
\]

and similarly:

\[
\mathbb{P}[S_u = 0, P_u = \{i, j\}] = \frac{1}{N(N-1)} \left( \frac{e^{\epsilon}}{e^{\epsilon} + 1} + \left( \frac{e^{\epsilon} - 1}{e^{\epsilon} + 1} \right) \frac{w(w-1)}{N(N-1)} (b_{ij} - 1) \right) \triangleq \frac{\overline{b_{ij}}'}{N(N-1)},
\]

where, under the assumptions that \( w = \Omega(N) \) and \( \epsilon = \Theta(1) \), we have that \( b_{ij}', \overline{b_{ij}}' \) are both \( \Theta(1) \).

Now, since \( \hat{A}_{ij} = 1_{\{S_u = 1, P_u = \{i, j\}\}} \), we have that:

\[
\hat{A}_{ij} \sim \text{Binomial} \left( U, \frac{b_{ij}}{N(N-1)} \right)
\]

And setting \( U = cN \log N \), we have that:

\[
\mathbb{P}[\hat{A}_{ij} > 0] = 1 - \left( 1 - \frac{b_{ij}}{N(N-1)} \right)^U = \frac{Ub_{ij}}{N(N-1)} + \Theta \left( \frac{U^2}{N^4} \right) = c'b_{ij} \frac{\log N}{N} + \Theta \left( \frac{(\log N)^2}{N^4} \right)
\]

Thus we can interpret \( \hat{A} \) as representing the edges of a random graph over the item set, with an edge between an item in class \( i \) and another in class \( j \) if \( \hat{A}_{ij} > 0 \); the probability of such an edge is \( \Theta \left( \frac{b_{ij} \log N}{N} \right) \). We can now use clustering results from Tomozei and Massoulié (2011) to complete the proof.

We note that the above analysis does not give us exact scaling behavior with respect to \( \epsilon \); this would require more detailed analysis. However, the MaxSense algorithm analyzed in Appendix D allows us to determine the trade-off between privacy and performance more accurately. Furthermore, the MaxSense algorithm, under stronger separation assumptions than the Pairwise-Preference algorithm, achieves the same scaling with respect to \( N \) in the information-rich regime and also optimal scaling for many other regimes for \( w \).
Appendix C. Lower bounds for the Information-scarce Setting

In this appendix, we prove the results stated in Section 4. Recall that we consider a scenario where there is a single class of users, and each item is ranked either 0 or 1 deterministically by each user. 

Let \( C_N(\cdot) : [N] \to \{0, 1\} \) be the underlying clustering function. We assume that the user-data for user \( u \) is given by \( X_u = (I_u, Z_u) \), where \( I_u \) is a size \( w \) subset of \([N]\) representing items rated by user \( u \), and \( Z_u \) are the ratings for the corresponding items; in this case, \( Z_u = \{Z(i)\}_{i \in I_u} \). We also denote the privatized sketch from user \( u \) as \( S_u \in \mathcal{S} \). Here the space \( \mathcal{S} \) from which sketches are drawn is assumed to be an arbitrary finite or countably infinite space. The sketch is assumed to obey \( \epsilon \)-DP. Finally, we assume that \( Z \) is chosen uniformly over \( \{0, 1\}^N \), and the set of items \( I_u \) rated by user \( u \) is also assumed to be chosen uniformly at random from amongst all size-\( w \) subsets of \([N]\).

We first develop two general lemmas in Section 4.1 which we use in our proof, but which can potentially be used for other similar situations. Then, in Section 4.2, we use these to derive tighter bounds on the scaling required for accurate cluster learning.

C.1. Local Differential Privacy and Mutual Information

We first establish two lemmas that we need in order to obtain the lower bound for learning in the information-scarce regime. The first lemma is a simple consequence of differential privacy and establishes a relation between the distribution of a random variable with and without conditioning on a differentially private sketch:

**Lemma 22** Given a discrete random variable \( A \in \mathcal{A} \) and some \( \epsilon \)-differentially private ‘sketch’ variable \( S \in \mathcal{S} \) generated from \( A \), there exists a function \( \lambda : \mathcal{A} \times \mathcal{S} \to [e^{-\epsilon}, e^\epsilon] \) such that for any \( a \in \mathcal{A} \) and \( s \in \mathcal{S} \):

\[
\mathbb{P}(A = a | S = s) = \mathbb{P}(A = a) \lambda(a, s)
\]

**Proof**

\[
\mathbb{P}(A = a | S = s) = \frac{\mathbb{P}(A = a) \mathbb{P}(S = s | A = a)}{\sum_{a' \in \mathcal{A}} \mathbb{P}(A = a') \mathbb{P}(S = s | A = a')}
\]

(From Bayes’ Theorem)

\[
= \mathbb{P}(A = a) \left( \sum_{a' \in \mathcal{A}} \frac{\mathbb{P}(A = a') \mathbb{P}(S = s | A = a')}{\mathbb{P}(S = s | A = a)} \right)^{-1}
\]

\[
= \mathbb{P}(A = a) \left( \sum_{a' \in \mathcal{A}} \frac{\mathbb{P}(S = s | A = a')}{\mathbb{P}(S = s | A = a)} \right)^{-1}
\]

\[
\Delta \mathbb{P}(A = a) \lambda(a, s)
\]

Further, from the definition of \( \epsilon \)-differential privacy, we have that:

\[
e^{-\epsilon} \leq \frac{\mathbb{P}(S = s | A = a')}{\mathbb{P}(S = s | A = a)} \leq e^\epsilon,
\]

and hence we have \( \lambda(a, s) \in [e^{-\epsilon}, e^\epsilon] \), \( \forall a \in \mathcal{A}, s \in \mathcal{S} \). \( \blacksquare \)

Recall that we define \( \binom{N}{w} \) to be the collection of all size-\( w \) subsets of \([N] = \{1, 2, \ldots, N\} \), and \( \mathcal{D} \triangleq \binom{N}{w} \times \{0, 1\}^w \) to be the set from which user information (i.e., \((I, Z)\)) is drawn (and define \( D = |\mathcal{D}| = \binom{N}{w} 2^w \)). Finally \( \mathbb{E}_X[\cdot] \) indicates that the expectation is over the random variable \( X \).
We now establish the following general lemma, which we will use to bound the mutual information between the model and the sketch.

**Lemma 23** Assume that under probability distribution $\mathbb{P}$, the set $I$ of items whose type is available to a given user is independent of the type vector $Z$. Denote $p_s(i, z) := \mathbb{P}(\{I, Z\} = (i, z) | S = s)$. Let also for subsets $j \subset [N]$ denote $p_j(z) := \mathbb{P}(Z_j = z)$. Then the following holds:

$$I(Z; S = s) \leq \sum_{i,z} \sum_{i',z'} p_s(i, z)p_s(i', z') \left[ I_{z=z'} \frac{p_{i,j}(z \cup z')}{p_i(z)p_{i'}(z')} - 1 \right]. \quad (6) \tag{6}$$

**Proof** From the definition of mutual information, we have:

$$I(Z; S) = \sum_{z,s} \mathbb{P}([Z, S] = (z, s)) \log \left( \frac{\mathbb{P}([Z, S] = (z, s))}{\mathbb{P}[Z = z]\mathbb{P}[S = s]} \right)$$

$$= \mathbb{E}_S \left[ I(Z; S = s) \right],$$

where we use the notation:

$$I(Z; S = s) := \sum_z \mathbb{P}[Z = z|S = s] \log \left( \frac{\mathbb{P}[Z = z|S = s]}{\mathbb{P}[Z = z]} \right)$$

Now note that

$$\mathbb{P}[Z = z|S = s] = \sum_{(i_1, z_1)} \mathbb{P}[Z = z, (I_1, Z_1) = (i_1, z_1)|S = s]$$

$$= \sum_{(i_1, z_1)} \mathbb{P}[Z = z|i_1, z_1] \mathbb{P}[(I_1, Z_1) = (i_1, z_1)|s]$$

$$= \sum_{(i_1, z_1)} p_s(i, z_i) \mathbb{P}[Z = z] \frac{\mathbb{P}[Z = z]}{p_i(z_i)} I_{\{z=z_1\}}.$$

Combining the equations, we get

$$I(Z; S = s) = \sum_z \sum_{i_1, z_1} I_{z=z_1} \mathbb{P}(Z = z) \frac{p_s(i_1, z_1)}{p_i(z_1)} \log \left( \sum_{i_2, z_2} I_{z=z_2} \frac{p_s(i_2, z_2)}{p_i(z_2)} \right).$$

Using Jensen’s inequality, the R.H.S. is upper bounded by the corresponding expression where averaging over $z$ conditionally on $Z_{i_1} = z_1$ is taken inside the logarithm, yielding

$$I(Z; S = s) \leq \sum_{i_1, z_1} p_s(i_1, z_1) \log \left( \sum_z I_{z=z_1} \frac{\mathbb{P}(Z = z)}{p_i(z_1)} \sum_{i_2, z_2} p_s(i_2, z_2) \frac{I_{z=z_2}}{p_i(z_2)} \right)$$

$$= \sum_{i_1, z_1} p_s(i_1, z_1) \log \left( \sum_{i_2, z_2} p_s(i_2, z_2) \frac{p_{i_1,j_2}(z_1 \cup z_2)}{p_i(z_1)p_{i_2}(z_2)} \right).$$

The result now follows by using the inequality $\log(x) \leq x - 1$.  

Note that in the above lemma we do not make any assumption regarding: i) the distribution of $Z$, ii) the distribution of the user-data $(I, Z)$. Now assuming that $Z$ is uniformly distributed on $\{0, 1\}^N$, we get the following corollary, which was stated before as Lemma 7:
Corollary 24  Given the Markov Chain $\mathbf{Z} \rightarrow (I, Z) \rightarrow S$, where $\mathbf{Z}$ is drawn uniformly from $\{0,1\}^N$, let $(I_1, Z_1), (I_2, Z_2) \in \mathcal{D}$ be two pairs of ’user-data’ sets which are drawn i.i.d according to the conditional distribution of $(I, Z)$ given $S = s$. Then, the mutual information $\mathcal{I}(\mathbf{Z}; S)$ satisfies:

$$\mathcal{I}(\mathbf{Z}; S) \leq \mathbb{E}_S \left[ \mathbb{E}((I_1, Z_1)|S) \mathbb{E}((I_2, Z_2)|S) \left[ 2^{|I_1 \cap I_2|} \mathbbm{1}_{Z_1 = Z_2} - 1 \right] \right].$$

Unless we specifically mention otherwise, when we refer to Lemma 23, we will mean this corollary.

C.2. Learning with 1-bit sketches

We next provide a proof of Theorem 8 from the paper. We restate the theorem for convenience.

Theorem (Theorem 8 in the paper) Suppose $w = 1$, with $(I, Z)$ drawn i.i.d uniformly over $[N] \times \{0, 1\}$. Then for any 1-bit sketch derived from $(I, Z)$, it holds that:

$$\mathcal{I}(\mathbf{Z}, S) = O \left( \frac{1}{N} \right),$$

and consequently, there exists a constant $c > 0$ such that any cluster learning algorithm using queries with 1-bit responses is unreliable if the number of users satisfies:

$$U < cN^2.$$

Proof  In order to use Lemma 23, we first note that $\mathcal{I}(\mathbf{Z}, S)$ is a convex function of $\mathbb{P}[S = s|\mathbf{Z} = z]$ for fixed $\mathbb{P}[\mathbf{Z} = z]$ (Theorem 2.7.4, Cover and Thomas (2006)). Writing $\mathbb{P}[S = s|\mathbf{Z} = z]$ as $\sum_{(i,z)} \mathbb{P}[S = s|(I, Z) = (i, z)]\mathbb{P}[(I, Z) = (i, z)|\mathbf{Z} = z]$, we observe that the extremal points of the kernel $\mathbb{P}[S = s|(i, z)]$ correspond to $\mathbb{P}[S = s|(i, z)] \in \{0, 1\}$, where the mutual information is maximized. This implies that the class of deterministic queries with 1-bit response that maximizes mutual information has the following structure: given user-data $(I_u, Z_u)$, user $u$’s response $S_u \in \{0, 1\}$ is of the form $S_u = \mathbbm{1}_A(I_u, Z_u)$, where $A \subseteq \{(i, z)|i \in [N], z \in \{0, 1\}\}$. In other words, the algorithm provides user $u$ with an arbitrary set $A$ of items, ratings, and the user identifies if $(I_u, Z_u)$ is contained in $A$.

Defining $p^i_{s,z} \triangleq \mathbb{P}[(I, Z) = (i, z)|S = s]$, for a query response $S = \mathbbm{1}_A(I_u, Z_u)$, we have the following:

$$p^i_{s,z} = \frac{\mathbb{P}[(I, Z) = (i, z)]\mathbb{P}[S = 1|(i, z)]}{\sum_{(j,z')} \mathbb{P}[(I, Z) = (j, z')]|S = 1|(j, z')]} = \frac{\mathbbm{1}_A(i, z)}{\sum_{j=1}^N \{\mathbbm{1}_A(j, 0) + \mathbbm{1}_A(j, 1)\}} = \frac{\mathbbm{1}_A(i, z)}{|A|},$$

and similarly $p^0_{i,z} = \frac{\mathbbm{1}_{\bar{A}}(i, z)}{|A|}$, where $\bar{A}$ is the complement of set $A$. From Lemma 7, for $(I_1, Z_1)|S \perp \perp (I_2, Z_2)|S$ we have:

$$\mathcal{I}(\mathbf{Z}, S) \leq \mathbb{E}_S \left[ \mathbb{E} \left[ 2^{|I_1 \cap I_2|} \mathbbm{1}_{Z_1 = Z_2} - 1 \right] \right] = \sum_{s \in \{0,1\}} \mathbb{P}[S = s] \mathbb{E} \left[ \mathbbm{1}_{I_1 = I_2} \left( 2\mathbbm{1}_{Z_1 = Z_2} - 1 \right) |S = s \right].$$
Introducing the notation \(\mathbb{P}(I = \ell, Z(\ell) = \sigma | S = s) = \pi_{\ell, \sigma}^s\), the following identity is easily established:

\[
N \sum_{\ell=1}^{N} \mathbb{E} \left[ 1_{\{I_1 = I_2 = \ell\}} \left( 21_{\{Z_1(\ell) = Z_2(\ell)\}} - 1 \right) | S = s \right] = \sum_{\ell=1}^{N} \left( \pi_{\ell,0}^s - \pi_{\ell,1}^s \right)^2.
\]  

(7)

The left-hand side of (7) is thus a non-negative definite quadratic form of the variables \(p_{i,z}^s\) (since \(\pi_{\ell,\sigma}^s = \sum_{i,\sigma} |\ell \in i, z(\ell) = \sigma| p_{i,z}^s\)). Thus we have:

\[
I(Z, S) \leq \sum_{s \in \{0, 1\}} \mathbb{P}[S = s] \sum_{\ell=1}^{N} \left( \pi_{\ell,0}^s - \pi_{\ell,1}^s \right)^2
\]

where \(A_s = A\) if \(s = 1\) and \(\bar{A}\) if \(s = 0\). Now for a given \(A\), consider the partitioning of the set \([N]\) into \(C_0 \cup C_1 \cup C_2\), where for \(k = 1, 2, 3, \forall i \in C_k, |A \cap \{(i, 0), (i, 1)\}| = k\). We then have the following:

\[
I(Z, S) \leq \mathbb{P}[S = 1] \frac{|C_1|}{|A|^2} + \mathbb{P}[S = 0] \frac{|C_1|}{|A|^2}
\]

\[
= \frac{|A|}{2N} \frac{|C_1|}{|A|^2} + \frac{\bar{A}}{2N} \frac{|C_1|}{|A|^2} \quad \text{(Since } S = 1_A(I, Z))
\]

\[
= \frac{|C_1|}{2N} \left( \frac{1}{|A|} + \frac{1}{2N - |A|} \right)
\]

\[
\leq \frac{1}{N}.
\]

Now using Fano’s inequality (Lemma 17), we get the theorem. 

---

**C.3. Lower Bound on Scaling for Clustering with Local-DP**

Recall that we defined \(D \triangleq \binom{[N]}{w} \times \{0, 1\}^w\) to be the set from which user information (i.e., \((I, Z)\)) is drawn. We write \(\mathbb{P}^0\) for the base probability distribution on \((I_1, Z_1)\) and \((I_2, Z_2)\), i.e., the two are independent and uniformly distributed over \(D\), and denote by \(\mathbb{E}^0\) mathematical expectation under \(\mathbb{P}^0\). For completeness, we state and prove the following basic asymptotic estimate which we use several times in the subsequent proofs.

**Lemma 25** If \(w = o(N)\), then:

\[
\left| \frac{\binom{N-w}{w}}{\binom{N}{w}} - \left( 1 - \frac{w^2}{N} \right) \right| = \Theta \left( \frac{w^4}{N^2} \right)
\]

**Proof**

\[
\frac{\binom{N-w}{w}}{\binom{N}{w}} = \prod_{k=0}^{w-1} \left( 1 - \frac{w}{N-k} \right)
\]
⇒ \left(1 - \frac{w}{N - w + 1}\right)^w \leq \frac{(N-w)}{w(N)} \leq \left(1 - \frac{w}{N}\right)^w

Now for the upper bound, using the binomial expansion, we have:

\left(1 - \frac{w}{N}\right)^w = 1 - \frac{w^2}{N} + \frac{w^4}{2N^2} - \ldots
\quad = 1 - \frac{w^2}{N} + \Theta\left(\frac{w^4}{N^2}\right)

Similarly for the lower bound, we have:

\left(1 - \frac{w}{N - w + 1}\right)^w = 1 - \frac{w^2}{N - w + 1} + \frac{w^4}{2(N - w + 1)^2} - \ldots
\geq 1 - \frac{w^2}{N} - \frac{w^3}{N(N - w + 1)} + \frac{w^4}{2(N - w + 1)^2} - \ldots
\quad = 1 - \frac{w^2}{N} - \Theta\left(\frac{w^4}{N^2}\right)

Now we can prove Theorems 9 and 10:

**Theorem** (Theorem 9 in the paper) *In the information-scarce regime, i.e., when \( w = o(N) \), we have that:

\[ \mathcal{I}(Z, S) = O\left(\frac{w^2}{N}\right). \]

and consequently, there exists a constant \( c > 0 \) such that any cluster learning algorithm with \( \epsilon \)-local-DP is unreliable if the number of users satisfies:

\[ U < c\left(\frac{N^2}{w^2}\right). \]

**Proof** To bound the mutual information between the underlying model and each private sketch, we use Lemma 7. In particular, we show that the mutual information is bounded by \( \left(\frac{w^2}{N}\right) \) for any given value \( s \) of the private sketch. Below we denote by \( \mathbb{E}_s[\cdot] \) expectations conditionally on \( S = s \).

Consider any sketch realization \( S = s \). Now, we have:

\[ \mathbb{E}_s\left[2^{\left|I_1 \cap I_2\right|} \mathds{1}_{\{Z_1 \equiv Z_2\}} - 1\right] \leq \mathbb{E}_s\left[\mathds{1}_{\{Z_1 \equiv Z_2\}} \left(2^{\left|I_1 \cap I_2\right|} - 1\right)\right] \]
The RHS of the above equation is a non-negative quadratic function of the variables \( \{p_{i,z}\}_{(i,z) \in \mathcal{D}} \), where \( p_{i,z} \triangleq P[(I, Z) = (i, z) | S = s] \). Now, using Lemma 22, we get:

\[
E_0 \left[ 2^{|I_1 \cap I_2|} \mathbb{1}_{\{Z_1 \equiv Z_2\}} - 1 \right] \leq e^{2\epsilon} E^0 \left[ \mathbb{1}_{\{Z_1 \equiv Z_2\}} \left(2^{|I_1 \cap I_2|} - 1\right)\right]
\]

\[
= e^{2\epsilon} \sum_{k=0}^w E^0 \left[ \mathbb{1}_{\{|I_1 \cap I_2| = k\}} \mathbb{1}_{\{Z_1 \equiv Z_2\}} \left(2^{|I_1 \cap I_2|} - 1\right)\right]
\]

\[
= e^{2\epsilon} \sum_{k=0}^w E^0 \left[ \mathbb{1}_{\{|I_1 \cap I_2| = k\}} 2^{-k} \left(2^k - 1\right)\right]
\]

\[
= e^{2\epsilon} (\Delta_1 + \Delta_2)
\]

where

\[
\Delta_1 = \frac{1}{2} E^0 \left[ \mathbb{1}_{\{|I_1 \cap I_2| = 1\}} \right], \Delta_2 = E^0 \left[ \mathbb{1}_{|I_1 \cap I_2| > 1} \left(1 - 2^{-|I_1 \cap I_2|}\right)\right]
\]

We bound each of these terms separately. For \( \Delta_1 \), we have:

\[
\Delta_1 = \frac{1}{2} E^0 \left[ \mathbb{1}_{\{|I_1 \cap I_2| = 1\}} \right] = \frac{1}{2} \sum_{\ell=1}^N E^0 \left[ \mathbb{1}_{|I_1 \cap I_2| = \ell} \right]
\]

\[
= \frac{w^2 (N-w)}{2(N-1)} \left(1 - \frac{w^2}{N} + O\left(\frac{w^4}{N^2}\right)\right) = O\left(\frac{w^4}{N^2}\right) \tag{8}
\]

Similarly for \( \Delta_2 \), we have:

\[
\Delta_2 \leq E^0 \left[ \mathbb{1}_{\{|I_1 \cap I_2| > 1\}} \right] = 1 - E^0 \left[ |I_1 \cap I_2| < 2 \right]
\]

\[
= 1 - \left(1 + \frac{w^2}{N-2w+1}\right) \frac{N-w}{w} \left(1 - \frac{w^2}{N} - O\left(\frac{w^4}{N^2}\right)\right)
\]

\[
= O\left(\frac{w^4}{N^2}\right) \tag{9}
\]

Combining equations (8) and (9), we get the result.
Note that the dominant term in the above proof is the bound on $\Delta_1$, which is closely connected to the case considered in Theorem 8, where we assumed $w = 1$. Now we have:

**Theorem (Theorem 10 in the paper)** Under the scaling assumption $w = o(N^{1/3})$, and for $\epsilon < \ln(2)$, it holds that

$$I(Z, S) = O \left( \frac{w}{N} \right).$$

and thus there exists a constant $c > 0$ such that any cluster learning algorithm with local-DP is unreliable if the number of users satisfies:

$$U < c \left( \frac{N^2}{w} \right).$$

**Proof** In the proof of Theorem 9, the two steps which are weak are the conversion to the base measure $P_0$ using Lemma 22, and the evaluation of the bound for $\Delta_1$. We start off by performing a similar decomposition of the bound, but without first converting to the base measure. For any $S = s$, we have:

$$E \left[ 2^{I_1 \cap I_2} I_{\{Z_1 = Z_2\}} - 1 \right] = \sum_{\ell=1}^{N} E \left[ I_{\{I_1 \cap I_2 = \{\ell\}\}} (2 \ast I_{\{Z_1 = Z_2\}} - 1) \right] + E \left[ I_{\{|I_1 \cap I_2| > 1\}} (2^{I_1 \cap I_2} I_{\{Z_1 = Z_2\}} - 1) \right] = \Delta'_1 + \Delta''_1 + \Delta'_2$$

where

$$\Delta'_1 = \sum_{\ell=1}^{N} E \left[ I_{\{\ell \in I_1 \cap I_2\}} (2 \ast I_{Z_1(\ell) = Z_2(\ell)} - 1) \right]$$

$$\Delta''_1 = - \sum_{\ell=1}^{N} E \left[ I_{\{\ell \in I_1 \cap I_2\}; |I_1 \cap I_2| > 1} (2 \ast I_{Z_1(\ell) = Z_2(\ell)} - 1) \right]$$

$$\Delta'_2 = E \left[ I_{|I_1 \cap I_2| > 1} (2^{I_1 \cap I_2} I_{\{Z_1 = Z_2\}} - 1) \right]$$

Note that $\Delta'_1 + \Delta''_1$ are similar to $\Delta_1$ and $\Delta'_2$ similar to $\Delta_2$ in Theorem 9 (albeit without first converting to the base measure). Unlike before, however, we first bound $\Delta''_1 + \Delta'_2$, establishing that $\Delta''_1 + \Delta'_2 = O(w^4/N^2) = o(w/N)$ whenever $w = o(N^{1/3})$. For $\Delta'_1$, we need to employ a more sophisticated technique for bounding. As before, we write $P^0$ for the base probability distribution under which $(I_1, Z_1)$ and $(I_2, Z_2)$ are independent and uniformly distributed over $D$, and denote by $E^0$ mathematical expectation under $P^0$. For $\Delta''_1$, we have:

$$\Delta''_1 \leq \sum_{\ell=1}^{N} E \left[ I_{\{\ell \in I_1 \cap I_2\}; |I_1 \cap I_2| > 1} \right]$$

$$= E \left[ |I_1 \cap I_2| I_{\{|I_1 \cap I_2| > 1\}} \right]$$
Since the RHS is non-negative, we use Lemma 22 to convert the expectation to the base measure. Thus, we get:

\[
\Delta_1'' \leq e^{2\epsilon} \left[ \mathbb{E}^0 \left[ |I_1 \cap I_2| \right] - \mathbb{P}^0 \left[ |I_1 \cap I_2| = 1 \right] \right]
\]

\[
= e^{2\epsilon} w \left( \frac{N-1}{w-1} \right) - w \left( \frac{N-w}{w-1} \right)
\]

\[
= e^{2\epsilon} \left( \frac{w^2}{N} - \left( \frac{w^2}{N-2w+1} \right) \left( \frac{N-w}{w} \right) \right)
\]

(10)

Similarly for \( \Delta_2' \), we have:

\[
\Delta_2' \leq e^{2\epsilon} \mathbb{E} \left[ \mathbb{I}_{\{|I_1 \cap I_2| > 1\}} \mathbb{I}_{|I_1 \cap I_2^{2} > 2}} \mathbb{I}_{\{Z_1 \equiv Z_2\}} \right]
\]

\[
\leq e^{2\epsilon} \mathbb{E}^0 \left[ \mathbb{I}_{\{|I_1 \cap I_2| > 1\}} \mathbb{I}_{|I_1 \cap I_2^{2} > 2}} \mathbb{I}_{\{Z_1 \equiv Z_2\}} \right]
\]

\[
\leq e^{2\epsilon} \mathbb{P}^0 \left[ |I_1 \cap I_2| > 1 \right],
\]

as \( \mathbb{P}^0 \left[ Z_1 \equiv Z_2 \right] = 2^{-|I_1 \cap I_2|}. \) Now since \( I_1 \) and \( I_2 \) are picked independently and uniformly over all size \( w \) subsets of \( [N] \) (under \( \mathbb{P}^0 \)), we have:

\[
\Delta_2' \leq e^{2\epsilon} \left( 1 - \frac{(N-w) + w(N-w)}{w} \right)
\]

\[
= e^{2\epsilon} \left( 1 - \left( 1 + \frac{w^2}{N-2w+1} \right) \left( \frac{N-w}{w} \right) \right)
\]

(11)

Finally combining equations (10) and (11), we get:

\[
\Delta_1'' + \Delta_2' \leq e^{2\epsilon} \left( 1 + \frac{w^2}{N} - \left( 1 + \frac{2w^2}{N-2w+1} \right) \left( \frac{N-w}{w} \right) \right),
\]

and using Lemma 25, we get:

\[
\Delta_1'' + \Delta_2' \leq e^{2\epsilon} \left( 1 + \frac{w^2}{N} - \left( 1 + \frac{2w^2}{N-2w+1} \right) \left( \frac{N-w}{w} \right) \right)
\]

\[
= O \left( \frac{w^4}{N^2} \right)
\]

Thus, we now have:

\[
\mathbb{E} \left[ 2^{1 \cap J} \mathbb{I}_{\{Z(\ell) = Z'(\ell) \forall \ell \in I \cap J\}} - 1 \right] \leq \sum_{\ell=1}^{N} \mathbb{E} \left[ \mathbb{I}_{\{\ell \in I \cap J\}} \left( 2 + \mathbb{I}_{Z(\ell) = Z'(\ell)} - 1 \right) \right] + O \left( \frac{w^4}{N^2} \right)
\]

Under the scaling assumption \( w = o(N^{1/3}) \), the second term in the right-hand side of the above equation is \( o(w/N) \), and we only need to establish that the first term in the right-hand side is \( O(w/N) \).
As in Theorem 8, we introduce the notation \( P(\ell \in I, Z(\ell) = \sigma | S = s) = \pi_{\ell,\sigma} \) (here we can omit indexing with respect to \( s \) for notational convenience). The following identity is then easily established:

\[
\sum_{\ell=1}^{N} E_1 \left\{ \mathbf{1}_{I_1 \cap I_2} (2 \mathbf{1}_{Z(\ell) = Z_2(\ell)} - 1) \right\} = \sum_{\ell=1}^{N} (\pi_{\ell,0} - \pi_{\ell,1})^2. \tag{12}
\]

The left-hand side of (12) is thus a non-negative definite quadratic form of the variables

\[
p_{i,z} := P(I = i, Z = z | S = s),
\]

where we have that \( \pi_{\ell,\sigma} = \sum_{i,\sigma} P(I = i, Z = z | S = s) \) in (12). We know however by Lemma 22 that these variables are constrained to lie in the convex set defined by the following inequalities:

\[
\sum_{(i,z)\in D} p_{i,z} = 1,
\]

\[
e^{-\epsilon} D \leq p_{i,z} \leq e^{\epsilon} D.
\]

Defining \( \epsilon' := e^{\epsilon} - 1 = \max(e^\epsilon - 1, 1 - e^{-\epsilon}) \), we can relax the last constraint to

\[
1 - \epsilon' \leq p_{i,z} D \leq 1 + \epsilon'.
\]

Provided \( \epsilon \) is small enough (precisely, provided \( \epsilon < \ln(2) \), which we have assumed), it holds that \( \epsilon' < 1 \).

Given this setup, we can now formulate the problem of upper bounding \( \Delta_1' \) as the following optimization problem:

\[
\text{maximize } \sum_{(i,z)\in D} \left( \pi_{\ell,0} - \pi_{\ell,1} \right)^2 \text{ subject to } \sum_{(i,z)\in D} p_{i,z} = 1, \quad p_{i,z} D \in [1 - \epsilon', 1 + \epsilon'] . \tag{13}
\]

In order to evaluate this bound, we need to first characterize the extremal points of the above convex set. We do this in the following lemma.

**Lemma 26** The extremal points of the convex set of distributions \( \{ p_{i,z} \} \) defined by (13) consists precisely of the distributions \( p_{i,z}^A \) indexed by the sets \( A \subset D \) of cardinality

\[
|A| = \binom{N}{w} 2^{w-1} = \frac{D}{2},
\]

defined by

\[
p_{i,z}^A = \begin{cases} \frac{1}{D} & \text{if } (i, z) \in A, \\ \frac{1 - \epsilon'}{D} & \text{if } (i, z) \notin A. \end{cases} \tag{14}
\]
Proof Let \( \{p_{i,z}\} \) be a probability distribution satisfying constraints (13). The aim is to establish the existence of non-negative weights \( \gamma_S \) for each subset \( S \subset D \) of size \( D/2 \), summing to 1, and such that for all \((i, z) \in D\), one has

\[
p_{i,z} = \sum_{S \subseteq D, |S| = D/2} \gamma_S (1 + \epsilon' \mathbbm{1}_{(i,z) \in S} - \epsilon' \mathbbm{1}_{(i,z) \notin S})/D.
\]  

(15)

Let us now express the existence of such weights \( \gamma_S \) as a property of a network flow problem. For each \( n \in [D] \), define

\[
\alpha_n := \left( p_n - \frac{1 - \epsilon'}{D} \right) \frac{D}{2\epsilon'}.
\]

The constraint \( p_n \in [(1 - \epsilon')/D, (1 + \epsilon')/D] \) entails that \( \alpha_n \in [0, 1] \). Construct now a network with for each \( n \in [D] \) two links, labelled \((n \in)\) and \((n \notin)\), and with respective capacities \( \alpha_n \) and \( 1 - \alpha_n \). In addition, for each set \( S \subset [D], |S| = D/2 \), create a route \( r_S \) through this network, which for each \( n \in D \) crosses link \((n \in)\) if \( n \in S \), and crosses link \((n \notin)\) if \( n \notin S \). All such routes are connected to a source and a sink node.

We now claim that the existence of probability weights \( \gamma_S \) satisfying (15) is equivalent to the fact that the maximum flow through this network is equal to 1. Indeed, the existence of a flow of total weight 1 is equivalent to the existence of a probability distribution \( \gamma_S \) on the routes \( r_S \) through this network which match the link capacity constraints, that is to say such that for all \( n \in [D] \), one has

\[
\sum_{S: n \in S} \gamma_S = \alpha_n, \\
\sum_{S: n \notin S} \gamma_S = 1 - \alpha_n.
\]

It is readily seen that this condition implies (15). Conversely, if the probability weights \( \gamma_S \) satisfy (15), using the definition of \( \alpha_n \), it is easily seen that the two previous equations hold.

Let us now establish the existence of such a flow. To this end, we use the max flow-min cut theorem. Any set of links that contains, for some \( n \in [D] \), both links \((n \in)\) and \((n \notin)\), is a cut, and its capacity is at least \( \alpha_n + 1 - \alpha_n \), hence larger than 1. Any cut \( C \) which for each \( n \) either does not contain \((n \in)\) or does not contain \((n \notin)\) must be such that either

\[
|C \cap \{\cup_{n \in [D]} (n \in)\}| > D/2
\]

(16)
or

\[
|C \cap \{\cup_{n \in [D]} (n \notin)\}| > D/2,
\]

(17)

for otherwise we can identify \( S \subset [D], |S| = D/2 \) which crosses this cut \( C \). Assume thus that (16) holds. Assume without loss of generality that \( C \) contains the links \((n \in)\) for all \( n = 1, \ldots, D/2+1 \). The weight of this cut is thus at least \( \sum_{n=1}^{D/2+1} \alpha_n \). We now argue that this must be at least 1. Indeed, it holds that

\[
\sum_{n=1}^{D/2+1} \alpha_n = D/2.
\]

However, if \( \sum_{n=1}^{D/2+1} \alpha_n < 1 \), using the fact that each \( \alpha_n \) is at most 1, it follows that \( \sum_{n=1}^{D} \alpha_n \) is strictly less than \( 1 + D/2 - 1 = D/2 \), a contradiction. The case when cut \( C \) verifies Equation (17) is similar. \(\blacksquare\)
We can now complete the proof of Theorem 10. Since as argued the second term in the right-hand side of (12) is a non-negative definite quadratic form of the $p_{i,z}$, it is in particular a convex function of the $p_{i,z}$, and as such is maximized over the convex set described by (13) at one of its extremal points, which are precisely identified by Lemma 26. It will thus suffice to establish the following inequality for all $A \subset D$ of size half the cardinality of the full set:

$$\sum_{\ell=1}^{N} \left( \pi_{\ell,0}^A - \pi_{\ell,1}^A \right)^2 \leq O(w/N),$$

(18)

where we introduced the notation for all $\ell \in [N]$ and $\sigma \in \{0, 1\}$:

$$\pi_{\ell,\sigma}^A = \sum_{i: \ell \in i} \sum_{z: z(\ell) = \sigma} p_{i,z}^A,$$

and $p_{i,z}^A$ is as defined in (14). Introducing also the sets

$$A_{\ell,\sigma} = \{(i, z) : \ell \in i \text{ and } z(\ell) = \sigma\},$$

we have

$$\pi_{\ell,0}^A - \pi_{\ell,1}^A = \frac{2^r}{N^2 w} |A_{\ell,0} \cap A| - |A_{\ell,1} \cap A| = \frac{2^r}{N^2 w} \langle 1_A, v_\ell \rangle,$$

(19)

where in the last display we used the following notations. $\langle \cdot, \cdot \rangle$ stands for the scalar product in $\mathbb{R}^D$, $1_A$ is the characteristic vector of the set $A$, and $v_\ell$ is defined as

$$v_\ell(i, z) = 1_{\{\ell \in i\}} (1 - 2z(\ell)).$$

Equation (19) entails that the left-hand side of Equation (18) also equals

$$\sum_{\ell=1}^{N} \left( \frac{2^r}{D} \right)^2 \langle 1_A, v_\ell \rangle^2.$$

(20)

The scalar product $\langle v_\ell, v_{\ell'} \rangle$ reads, for $\ell \neq \ell'$:

$$\langle v_\ell, v_{\ell'} \rangle = \sum_{i: \ell, \ell' \in i} \sum_{z} (1 - 2z(\ell))(1 - 2z(\ell')) = \sum_{i: \ell, \ell' \in i} 2^{w-2}2 [(1) * (1) + (1) * (-1)] = 0.$$

Note further that for all $\ell \in [N]$, one has

$$||v_\ell||^2 = \left( \frac{N - 1}{w - 1} \right)^2 \frac{wD}{N}.$$

Orthogonality and equality of norms among the $v_\ell$ readily implies that the expression in (20) is upper-bounded by

$$\left( \frac{2^r}{D} \right)^2 \frac{wD}{N} \langle 1_A, 1_A \rangle^2.$$
Recalling that the vector $1_A$ has $\frac{D}{2}$ entries equal to 1, and all other entries equal to zero, the square of its Euclidean norm $||1_A||^2$ equals precisely $\frac{D}{2}$. Plugging this value in the last display, after cancellation, one obtains that the expression in (20) is bounded by

$$2\epsilon^2 \frac{w}{N}.$$  

This completes the proof. ■

Appendix D. The MaxSense algorithm

In this appendix, we state and prove the privacy and performance guarantees of the MaxSense algorithm, and its variants. The algorithm is formally specified in Algorithm 2.

Algorithm 2 The MaxSense Algorithm

**Setting:** $N$ items $[N]$, $U$ users $[U]$. Each user has a set of $w$ ratings $(W_u, R_u), W_u \in \binom{[N]}{w}, R_u \in \{0,1\}^w$. Each item $i$ is associated with a cluster $C_N(i)$ from a set of $L$ clusters, $\{1,2,\ldots,L\}$

**Output:** The cluster labels of each item, i.e., $\{C_N(i)\}_{i \in [N]}$

**Stage 1 (User sketch generation):**
- For each user $u \in [U]$, generate sensing vector $H_u \in \{0,1\}^{N}$, where $H_{ui}$ is a ‘probe’ for item $i$ given by:

$$H_{ui} \sim \text{Bernoulli} (p), \text{ i.i.d.,}$$

with $p = \frac{\theta}{w}$, where $\theta$ is a chosen constant.
- User $u$ generates a private sketch $S_u^0$ given by:

$$S_u^0(W_u, R_u, H_u) = \max_{i \in [N]} H_{ui} \hat{R}_{ui}$$

Where $\hat{R}_{ui} = R_{ui}$ if $i \in W_u$, and 0 otherwise.

**Stage 2 (User sketch privatization):** Each user $u \in [U]$ releases a privatized sketch $S_u$ from $S_u^0$ using the $\epsilon$-DP bit release mechanism (Proposition 2).

**Stage 3 (Item Clustering):**
- For each item $i \in [N]$, compute a count $B_i$ as:

$$B_i = \sum_{u \in U} H_{ui}S_u$$

- Perform k-means clustering using the counts $\{B_i\}_{i \in [N]}$ with $k = L$. 

32
Theorem (Theorem 11 in the paper) The MaxSense algorithm (Algorithm 2) is ε-differentially private. Further, suppose we define:
\[
\hat{\epsilon} = \frac{2(e^{\epsilon} - 1)}{(e^{\epsilon} + 1)},
\]
\[
\delta_{\min} = \min_{1 \leq i < j \leq L} \left| \sum_{k=1}^{K} \alpha_k e^{-\theta} \sum_{l=1}^{L} \beta_l b_{kl} (b_{kl} - b_{kl'}) \right|,
\]
then for any \( d > 0 \), there exists a constant \( C > 0 \) such that if the number of users satisfies:
\[
U \geq C \left( \frac{N^2 \log N}{\epsilon^2 \delta_{\min}^2 w} \right),
\]
then the clustering is successful with probability \( 1 - N^{-d} \).

Proof
Privacy: For each user \( u \), observe that \( H_u \) is independent of the data \( (W_u, R_u) \), and hence preserves privacy. Next, given \( H_u \), we have that \( (W_u, R_u) \to S_0^u \to S_u \) form a Markov chain, and hence it is sufficient via the post-processing property to prove that \( S_0^u \to S_u \) satisfy ε-differential privacy. This is a direct consequence of using the ε-DP bit release mechanism (Proposition 2). Now, using the post-processing property of differential privacy (Proposition 15), we get our result.

Performance: The intuition behind the correctness of the clustering of items is as follows: First, we show that for any item \( j \), its count \( B_j \) will concentrate around \( \hat{\mu}_{l(j)} \), the expected count for its corresponding cluster. Next, we calculate the minimum separation between the expected counts for any two item-clusters (denoted as \( \Delta_{\min} \)). Finally, we show that for the given scaling of users, with high probability we have that each item count \( B_j \) is within a distance of \( \Delta_{\min}/5 \) from its corresponding \( \hat{\mu}_{l(j)} \). This will then ensure that any two items belonging to the same cluster are within a distance of \( 2\Delta_{\min}/5 \), while two items of different clusters have a separation of at least \( 3\Delta_{\min}/5 \), thereby ensuring successful clustering.

First, for any item \( i \in [N] \), consider the random variable \( B_i = \sum_{u \in U} H_{ui} S_i \). We have that:
\[
\mathbb{E}[B_i] = \sum_{u \in U} \mathbb{E}[H_{ui} S_i] = \sum_u p \left[ \frac{1}{e^{\epsilon} + 1} \mathbb{E}[1 - S_0^u | H_{ui} = 1] + \frac{e^{\epsilon}}{(e^{\epsilon} + 1)} \mathbb{E}[S_0^u | H_{ui} = 1] \right],
\]
using the i.i.d sensing property and the definition of the privacy mechanism. Now, we substitute \( \hat{\epsilon} = \frac{2(e^{\epsilon} - 1)}{(e^{\epsilon} + 1)} \) (Note: for small \( \epsilon \), we have \( \hat{\epsilon} \approx \epsilon \)) to get:
\[
\mathbb{E}[B_i] = \sum_{u=1}^{U} p \left[ \frac{1}{2} - \frac{\hat{\epsilon}}{4} + \frac{\hat{\epsilon}}{2} \mathbb{E}[S_0^u | H_{ui} = 1] \right] \left[ 1 - \prod_{j \in [N]} \left( 1 - \frac{w_{uj} H_{uj}}{N} \right) \right]
\]
\[
= \sum_{u=1}^{U} p \left[ \frac{1}{2} - \frac{\hat{\epsilon}}{4} + \frac{\hat{\epsilon}}{2} \mathbb{E}[S_0^u | H_{ui} = 1] \right] \left( 1 - \frac{w_{ui}}{N} \right) \mathbb{E}\left[ \prod_{j \neq i} \left( 1 - \frac{w_{uj} H_{uj}}{N} \right) \right]
\]

33
We use the shorthand notation $k(u)$ to denote the user-cluster of user $u$ and similarly $l(j)$ to denote the item-cluster of item $j$ to get:

\[
E[B_i] = \sum_{u=1}^{U} p \left[ \frac{1}{2} + \frac{\hat{\epsilon}}{4} - \frac{\hat{\epsilon}}{2} \left( 1 - \frac{w b_{k(u)} l(i)}{N} \right) \prod_{j \neq i} \left( 1 - p + p \left( 1 - \frac{w b_{k(u)} l(j)}{N} \right) \right) \right]
\]

(Using the i.i.d sensing properties of $H_{ui}$)

\[
= \sum_{k=1}^{K} \alpha_k U p \left[ \frac{1}{2} + \frac{\hat{\epsilon}}{4} - \frac{\hat{\epsilon}}{2} \left( 1 - \frac{w b_{k(i)} l(j)}{N} \right) \left( 1 - \frac{p w b_{k(i)}}{N} \right)^{-1} \prod_{j \in [N]} \left( 1 - \frac{p w b_{k(j)}}{N} \right) \right]
\]

(Grouping terms by user and item classes.)

Note that we have dropped the explicit dependence on the user index and retained only the user-cluster label. Similarly, we henceforth write $k$ and $l$ for $k(i), l(j)$ respectively, whenever it does not cause confusion in the notation. Also from the algorithm specification, we have $p w = \theta$. Furthermore, we define:

\[
q_k^0 \triangleq \mathbb{P}[S_u^0 = 0|k(u) = k] = \prod_{j \in [N]} \left( 1 - \frac{p w b_{k(j)}}{N} \right),
\]

i.e., $q_k^0$ is the probability that a user $u$ of cluster $k$ will have a (private) sketch $S_u^0$ equal to 0. Thus we have:

\[
E[B_i] = U p \left[ \frac{1}{2} + \frac{\hat{\epsilon}}{4} - \frac{\hat{\epsilon}}{2} \sum_{k=1}^{K} \alpha_k q_k^0 \left( 1 - \frac{w b_{k(i)}}{N} \right) \left( 1 - \frac{p w b_{k(i)}}{N} \right)^{-1} \right]
\]

\[
= U p \left[ \frac{1}{2} + \frac{\hat{\epsilon}}{4} - \frac{\hat{\epsilon}}{2} \sum_{k=1}^{K} \alpha_k q_k^0 \left( 1 - \frac{w (1-p)}{N} b_{k(i)} \right) \right]
\]

\[
= U p \left[ \frac{1}{2} + \frac{\hat{\epsilon}}{4} - \frac{\hat{\epsilon}}{2} \sum_{k=1}^{K} \alpha_k q_k^0 \right] + U p \frac{(w - \theta)}{N} \frac{\hat{\epsilon}}{2} \sum_{k=1}^{K} \frac{\alpha_k q_k^0 b_{k(i)}}{1 - \frac{\theta}{N} b_{k(i)}}
\]
Before continuing, we need to analyze the term $q^0_k$. We have:

$$
\log q^0_k = \sum_{j=1}^{N} \log \left(1 - \frac{\theta b_{kj}}{N}\right) \\
= \sum_{l=1}^{L} \beta_l N \log \left(1 - \frac{\theta b_{kl}}{N}\right) \\
= \sum_{l=1}^{L} \beta_l N \left(- \frac{\theta b_{kl}}{N} + \Theta \left(\frac{1}{N^2}\right)\right) \\
= -\theta \sum_{l=1}^{L} \beta_l b_{kl} + \Theta \left(\frac{1}{N}\right) \\
\geq -\theta + \Theta \left(\frac{1}{N}\right)
$$

From this we can see that:

$$
\frac{1}{e^\theta} \left(1 + \Theta \left(\frac{1}{N}\right)\right) \leq q^0_k \leq 1,
$$

Thus we see that for any user-cluster, the probability of the MaxSense sketch being 0 is $\Theta(1)$. Intuitively, this means that each sketch has close to 1 bit of information. We define $q^0 = \sum_{k=1}^{K} q^0_k$ (which is the probability that a random user’s sketch is 0). Now, noting that the expectation of $B_i$ only depends on the class $l(i)$ of item $i$, we define $\overline{B}_l = \mathbb{E}[B_i | l(i) = l]$. Then we have:

$$
\overline{B}_l = U_p \left[ \frac{1}{2} + \frac{\tilde{c}}{4} - \frac{\tilde{c}}{2} q^0 \right] + U_p \frac{(w - \theta) \tilde{c}}{N} \sum_{k=1}^{K} \frac{\alpha_k q^0_k b_{kl}}{1 - \frac{\theta b_{kl}}{N}}
$$

Suppose $w = o(N)$. Since $\tilde{c} < 1$, then for sufficiently large $N$, we have that for all item classes $l \in \{1, 2, \ldots, L\}$:

$$
\overline{B}_l \leq U_p
$$

Next, given any two distinct item classes $l, m$, we define $\Delta_{lm} \triangleq \mathbb{E}[|B_l - B_m|]$. Then we have:

$$
\Delta_{lm} \geq |\mathbb{E}[B_l - B_m]| \quad \text{(By Jensen’s Inequality)}
$$

$$
= U_p \frac{(w - \theta) \tilde{c}}{N} \left| \sum_{k=1}^{K} \frac{\alpha_k q^0_k b_{kl}}{1 - \frac{\theta b_{kl}}{N}} - \sum_{k=1}^{K} \frac{\alpha_k q^0_k b_{km}}{1 - \frac{\theta b_{km}}{N}} \right| \\
\geq \frac{U \tilde{c} (c - c^2 w^{-1})}{N} \delta_{lm},
$$

35
where we define:

\[ \delta_{lm} \triangleq \sum_{k=1}^{K} \frac{\alpha_k q_k^0 b_{kl}}{1 - \theta b_{kl}} - \sum_{k=1}^{K} \frac{\alpha_k q_k^0 b_{km}}{1 - \theta b_{km}} \]

\[ = \sum_{k=1}^{K} \frac{\alpha_k q_k^0 (b_{kl} - b_{km})}{1 - \theta b_{kl}} \left( 1 - \frac{\theta b_{km}}{N} \right) \]

\[ \geq \sum_{k=1}^{K} \alpha_k e^{-\theta \sum_{l=1}^{L} \beta_l b_{kl}} (b_{kl} - b_{km}) \]

\[ \geq \delta_{\min} \triangleq \min_{1 \leq l < l' \leq L} \left| \sum_{k=1}^{K} \alpha_k e^{-\theta \sum_{l=1}^{L} \beta_l b_{kl}} (b_{kl} - b_{kl'}) \right|, \]

Let \( \Delta_{\min} \triangleq \min_{l,m \in [L]^2, l \neq m} \Delta_{lm} \). Now, for a given item \( j \), a standard Chernoff bound (applicable since the sketches are independent and bounded) gives us that for any \( a > 0 \):

\[ \mathbb{P}[|B_j - B_l(j)| \geq a B_l(j)] \leq 2 \exp \left( -\frac{a^2}{3} B_l(j) \right) \]

Following the above discussion, we choose \( a = \frac{\Delta_{\min}}{5 B_l(j)} \). Then we have:

\[ \mathbb{P} \left[ |B_j - B_l(j)| \geq \frac{\Delta_{\min}}{5} \right] \leq 2 \exp \left( -\frac{\Delta_{\min}^2}{75 B_l(j)} \right) \]

\[ \leq 2 \exp \left( -\frac{\Delta_{\min}^2}{75 U w} \right), \]

and by taking union bound over all items, we have:

\[ \mathbb{P} \left[ \sup_{j \in [N]} |B_j - B_l(j)| \geq \frac{\Delta_{\min}}{5} \right] \leq \exp \left( \log 2 N - \frac{\Delta_{\min}^2}{75 U w} \right) \]

\[ \leq \exp \left( \log 2 N - \frac{U w^2 \delta_{\min}^2}{75 N^2} \right), \]

where we have used \( p = \frac{c}{w} \). Now if we choose \( U \) as:

\[ U = \left( \frac{75 N^2 (\log 2 + (1 + d) \log N)}{c^2 \delta_{\min}^2 w} \right) = \Theta \left( \frac{N^2 \log N}{c^2 \delta_{\min}^2 w} \right), \]

then we have:

\[ \mathbb{P} \left[ \sup_{j \in [N]} |B_j - B_l(j)| \geq \frac{\Delta_{\min}}{5} \right] \leq \frac{1}{N^d}, \]

Thus, if the number of users scale according to (21), then the clustering is successful with probability \( 1 - N^{-d} \). \( \blacksquare \)
Finally, we define Multi-MaxSense, a generalization of Algorithm 2, wherein we ask multiple MaxSense questions to each user. Each question now has a privacy parameter of $\epsilon_Q$, where $Q$ is the number of questions asked to a user—thus we obtain $\epsilon$-DP via the composition property (Proposition 14). Independence between the answers is ensured as follows: first, for each user, we choose a random partition of $[N]$ into $\frac{1}{p}$ sets, each of size $Np$; we pick $Q$ of these and present them to the user. Next, each user calculates $Q$ sketches using these $Q$ sensing vectors, and reveals the privatized set of sketches (with each sketch revelation obeying $\frac{\epsilon}{Q}$-differential privacy. Finally, we calculate the item counts as before. More formally, the algorithm is given in Algorithm 3.

Algorithm 3 The Multi-MaxSense Algorithm

**Setting:** $N$ items $[N]$, $U$ users $[U]$, each with data $(W_u, R_u) \in ([N] \times \{0, 1\}^w$. Parameter $Q$.

**Output:** The cluster labels of each item, $\{C_N(i)\}_{i \in [N]}$

**Stage 1 (User sketch generation):**
- For each user $u \in [U]$, generate $Q$ sensing vectors $H_{(u,q)} \in \{0, 1\}^N$, where each vector is generated by choosing $Np$ items uniformly and without replacement. As before, $p = \frac{1}{w}$.
- User $u$ generates $Q$ private sketches $S^0_{(u,q)}$ as in Algorithm 2.

**Stage 2 (User sketch privatization):** Each user $u \in [U]$ releases $Q$ privatized sketches, where each sketch is generated using a $\frac{\epsilon}{Q}$-private bit release mechanism (Proposition 2).

**Stage 3 (Item Clustering):**
- For each item $i \in [N]$, compute a count $B_i = \sum_{u \in [U]} \sum_{q \in [Q]} H_{(u,q)i} S_{(u,q)}$
- Perform k-means clustering using the counts $\{B_i\}_{i \in [N]}$ with $k = L$.

Now we have the following theorem.

**Theorem 27** The Multi-MaxSense algorithm (Algorithm 3) is $\epsilon$-differentially private. Further, suppose $Q = \lceil \frac{\epsilon}{c} \rceil$. Then for any $d > 0$, there exists a constant $c$ such that if the number of users satisfies:

$$U \geq c \left( \frac{N^2 \log N}{\epsilon \delta_{\text{min}}^2 w} \right),$$

then the clustering is successful with probability $1 - N^{-d}$.

**Proof**

**Privacy:** Since each user reveals $Q$ bits, and each bit is privatized using a $\frac{\epsilon}{Q}$-differential private mechanism, therefore for any user $u$, the user sketch $\{S_{u,q}\}_{q=1}^Q$ and user data $(W_u, R_u)$ are private using the composition property (Proposition 14). The remaining proof for the privacy of the learning algorithm is as before, using the post-processing property.

**Performance:** To show the improved scaling, we need to observe the following:

1. Due to the way in which the sensing vectors are chosen, the probability of any probe for a item in any sensing vector (i.e., $H_{(u,q)i}$ for some $u \in [U], q \in [Q], i \in [N]$) being set to 1 is $p$, i.i.d.
2. Further, since the multiple sensing vectors given to a single user do not overlap, therefore the 
sketches \( \{S_{u,q}\}_{u,q} \) are also independent.

Hence, the analysis in Algorithm 2 can be repeated with \( U \) being replaced with \( QU \) and \( \epsilon \) being 
replaced with \( \frac{\epsilon}{Q} \). Choosing \( Q = \lceil \epsilon \rceil \) implies that we now have:

\[
\tilde{\epsilon} = \frac{2 \left( \exp \left( \frac{\epsilon}{Q} \right) - 1 \right)}{\exp \left( \frac{\epsilon}{Q} \right) + 1} \geq \frac{2(e - 1)}{e + 2}
\]

Substituting these in equation 21, we get the condition for correct clustering with high probability as:

\[
U \geq e^\prime \left( \frac{N^2 \log N}{\epsilon \delta_{\min}^2 w} \right).
\]

\[\square\]

Appendix E. Lower Bound for Adaptive Queries

The lower bounds of Section 4 applied to non-adaptive learning, where queries to user \( u \) are designed 
without leveraging answers of users \( 1, \ldots, u - 1 \). One can extend these bounds to the adaptive setting 
where query to user \( u \) is allowed to depend on the previous queries and answers of users \( 1, \ldots, u - 1 \). Specifically we now assume that questions are asked to users sequentially, and the 
question to which the \( t \)-th user answers can be affected by the previous sketch releases \( S_1, \ldots, S_{t-1} \) of the \( t - 1 \) previous users. We shall now prove the following:

**Theorem 28** Assume \( w = 1 \). If users’ answers are \( \epsilon \)-DP, the number of adaptive queries needed 
to learn unknown content clustering into two types drawn uniformly at random from \( \{0, 1\}^N \) is 
\( \Omega(N \log N) \).

**Proof**

In the sequel we assume that \( T - 1 \) sketches have been released, and denote by \( P^T \) the probability 
distribution conditionally on the previously observed sketch values. We shall develop bounds of the form:

\[
\mathcal{I}(Z; S_1^T) \leq \delta_T
\]

for a suitable function \( \delta_T \).

These bounds are obtained inductively as follows. First, we expand the mutual information as:

\[
\mathcal{I}(Z; S_1^T) = \sum_{t=1}^{T} \mathcal{I}(Z; S_t|S_1^{t-1}).
\]

Now recall from before that we define:

\[
\mathcal{I}(U; V = v|W = w) := \sum_u \mathbb{P}(U = u|V = v, W = w) \log \left( \frac{\mathbb{P}(U = u|V = v, W = w)}{\mathbb{P}(U = u|W = w)} \right),
\]

38
i.e., the mutual information between $U$ and $V = v$ conditioned on $W = w$. Hence, we have

$$\sum_v \mathbb{P}(V = v|W = w)\mathcal{I}(U; V = v|W = w) = \mathcal{I}(U, V|W = w)$$

and

$$\sum_w \mathbb{P}(W = w)\mathcal{I}(U; V|W = w) = \mathcal{I}(U, V|W).$$

Further, using this definition, we can bound the mutual information as:

$$\mathcal{I}(Z; S^T) \leq \mathcal{I}(Z; S^{T-1}) + \sup_{s, s'_{T-1}} \mathcal{I}(Z; S_T = s|S^{T-1} = s'_{T-1}).$$

Now consider any sequence $\{s^{T-1}_1, s\}$. Defining $\mathbb{P}^T$ to be the probability measure conditional on $S^{T-1}_1 = s^{T-1}_1$, we use Lemma 23 to bound the term $\mathcal{I}(Z; S_T = s|S^{T-1} = s^{T-1}_1)$, to get:

$$\mathcal{I}(Z; S_T = s|S^{T-1}_1 = s^{T-1}_1) \leq \sum_{i_1, z_1} \sum_{i_2,z_2} p^{T+1}(i_1, z_1)p^{T+1}(i_2, z_2) \left[1_{z_1 \equiv z_2} \frac{p_{i_1, i_2}(z_1 \cup z_2)}{p_{i_1}(z_1)p_{i_2}(z_2)} - 1\right],$$

where $p^{T+1}(i, z) = \mathbb{P}^{T+1}[I, Z] = (i, z)$ and $p_{i}^{T+1}(z) = \mathbb{P}^{T+1}[Z = z]$ (in other words, all quantities are defined w.r.t. the probability measure conditional on $S^{T-1}_1 = s^{T-1}_1$ and $S_T = s$).

Using Lemma 22, we have $p^{T+1}(i, z) = f_i(z_i) \frac{1}{N} p_i^T(z_i)$ where the likelihood ratio $f_i(z_i)$ belongs to $[1 - \epsilon', 1 + \epsilon']$ where $\epsilon' = \epsilon - 1$. Using this expression, the RHS of the previous inequality can be rewritten as:

$$\mathcal{I}(Z; S_T = s|S^{T-1}_1 = s^{T-1}_1) \leq \frac{1}{N^2} \text{Var}^T \left[\sum_{i=1}^N f_i(Z_i)\right],$$

where $\text{Var}^T$ is defined w.r.t. the $\mathbb{P}^T$ measure. Let $\mathbb{P}^0$ be the unconditional probability, under which the $Z_i$ are i.i.d. uniform on $\{0, 1\}$. We define $F := \sum_{i=1}^N f_i(Z_i)$; note that under $\mathbb{P}^0$, the random variable $F$ has a variance that is at most $2\epsilon'^2N$. If we could have a similar bound for the variance of $F$ under $\mathbb{P}^T$ rather than under $\mathbb{P}^0$, this would yield an upper bound of order $1/N$ on the mutual information of interest.

We now proceed to show that, provided the two distributions $\mathbb{P}^0$ and $\mathbb{P}^T$ have small Kullback-Leibler divergence, then the variance of $F$ under $\mathbb{P}^T$ is indeed of order at most $N$. The argument proceeds in several steps.

**Step 1: Relating divergence between $\mathbb{P}^T$ and $\mathbb{P}^0$ to the divergence between the law of $F$ under $\mathbb{P}^T$ and under $\mathbb{P}^0$:**

**Lemma 29** For each $f$ in the support of the discrete random variable $F$, let $p_f$ and $p_f^0$ denote the probabilities that $F = f$ under $\mathbb{P}^T$ and $\mathbb{P}^0$ respectively. Then we have:

$$H(\mathbb{P}^0) - H(\mathbb{P}^T) = D(\mathbb{P}^T||\mathbb{P}^0) \geq D(p||p^0) = \sum_f p_f \log \left(\frac{p_f}{p_f^0}\right). \quad (22)$$
Proof. For each $f$, let $N_f$ denote the number of vectors $z \in \{0, 1\}^N$ for which $F = f$, so that $p_f^0 = N_f 2^{-N}$. Write
\[ H(\mathbb{P}^T) = \sum_f p_f \sum_{z:F(z) = f} \frac{\mathbb{P}^T(z)}{p_f} \left[ \log \left( \frac{1}{p_f} \right) + \log \left( \frac{p_f}{\mathbb{P}^T(z)} \right) \right] \]
\[ \leq \sum_f p_f \left[ \log \left( \frac{1}{p_f} \right) + \log(n_f) \right] \]
\[ = \sum_f p_f \left[ \log \left( \frac{1}{p_f} \right) + N \log(2) + \log(p_f^0) \right] \]
\[ = H(\mathbb{P}^0) - D(p||p^0), \]
where the inequality follows by upper-bounding the entropy of a probability distribution on a set of size $N_f$ by $\log(N_f)$.

Step 2: Bounding variance of $F$ under $\mathbb{P}^T$ given divergence constraints:
Let $\bar{F}$ denote the expectation of $F$ under $\mathbb{P}^0$, i.e., $\bar{F} = \sum_f p_f^0 f$. Also let $\sigma^2$ denote the variance of $F$ under $\mathbb{P}^0$. Note that
\[ \text{Var}^T(F) = \inf_{x \in \mathbb{R}} \mathbb{E}^T((F-x)^2) \leq \mathbb{E}^T((F-\bar{F})^2) = \sum_f p_f (f - \bar{F})^2. \]
Assume that the entropy $H(\mathbb{P}^T)$ verifies $H(\mathbb{P}^T) \geq H(\mathbb{P}^0) - \delta$, for some $\delta \geq 0$. Then in view of (22) and the previous display, an upper bound on the variance of $F$ under $\mathbb{P}^T$ is provided by the solution of the following optimization problem:

Maximize \[ \sum_f p_f (f - \bar{F})^2 \]
over \[ p_f \geq 0 \]
such that \[ \sum_f p_f = 1 \]
and \[ \sum_f p_f \log \left( \frac{p_f}{p_f^0} \right) \leq \delta. \] (23)

It is readily seen by introducing the Lagrangian of this optimization problem, and a dual variable $\nu^{-1} > 0$ for the constraint (23) that the optimal of this convex optimization problem is achieved by
\[ p_f := \frac{1}{Z(\nu)} p_f^0 e^{\nu(f - \bar{F})^2}, \]
for a suitable positive constant $\nu$, where the normalization constant $Z(\nu)$ is given by:
\[ Z(\nu) := \sum_f p_f^0 e^{\nu(f - \bar{F})^2} = \mathbb{E}^0 e^{\nu(F - \bar{F})^2}. \]
For this particular distribution, the divergence $D(p||p^0)$ reads:
\[ \sum_f \frac{1}{Z(\nu)} p_f^0 e^{\nu(f - \bar{F})^2} \left[ \nu(f - \bar{F})^2 - \log Z(\nu) \right] = -\log(Z(\nu)) + \frac{\nu}{Z(\nu)} \mathbb{E}^0 (F - \bar{F})^2 e^{\nu(F - \bar{F})^2}, \]
so that constraint (23) reads
\[- \log(Z(\nu)) + \frac{\nu}{Z(\nu)} \mathbb{E}_0((F - \bar{F})^2 e^{\nu(F - \bar{F})^2}) \leq \delta.\] (24)

This characterization in turn allows to establish the following

**Lemma 30** Let \( \psi(\nu) := \log Z(\nu). \) Assume there exist \( a, \nu > 0 \) such that
\[ \nu a - \psi(\nu) \geq \delta. \] (25)

Then the solution to the value of the optimization problem 23 is less than or equal to \( a. \)

**Proof** Note that by Hölder’s inequality, function \( \psi \) is convex, so that its derivative \( \psi' = \frac{Z^{-1}(\nu)}{\nu} \mathbb{E}_0((F - \bar{F})^2 e^{\nu(F - \bar{F})^2}) \) is non-decreasing. Note further that the function \( \nu \psi' - \psi \) appearing in the left-hand side of (24) is non-decreasing for non-negative \( \nu \), as its derivative reads \( \nu \psi'' \). Thus the value \( \nu^* \) which achieves the optimum is such that
\[ \nu^* \psi'(\nu^*) - \psi(\nu^*) = \delta \]
and the sought bound is \( \psi'(\nu^*) \). Now for a given \( a \in \mathcal{R} \), the supremum of \( \nu a - \psi(\nu) \) is achieved precisely at \( \nu \) such that \( a = \psi'(\nu) \). Thus if for some \( \nu \) and some \( a \) Condition (25) holds, it follows that
\[ \sup_{\nu} (\nu a - \psi(\nu)) \geq \delta = \sup_{\nu} (\nu a^* - \psi(\nu)), \]
where \( a^* := \psi'(\nu^*) \). It follows from monotonicity of \( \nu \rightarrow \nu \psi'(\nu) - \psi(\nu) \) that the value \( \nu' \) where the supremum is achieved in the left-hand side, and such that \( a = \psi'(\nu') \), verifies \( \nu' \geq \nu^* \). Monotonicity of \( \psi' \) then implies that \( a \geq a^* \) as announced.

**Step 3: Deriving explicit bounds, using concentration results under \( \mathbb{P}^0 \).**

Define the centered and scaled random variable
\[ G := \frac{F - \bar{F}}{\sigma}. \]

Recall that after centering, each variable \( f_i(Z_i) \) is bounded in absolute value by \( \epsilon' \). Thus, using the Azuma-Hoeffding inequality yields the following bound:
\[ \mathbb{P}^0(G > A) \leq e^{-A^2/2}, \ A > 0, \] (26)
and the same bound holds for \( \mathbb{P}^0(G < -A) \). To obtain the above, we used the fact that after centering, \( f_i(Z_i) \) is of the form \( \sigma_i(2Z_i - 1) \) where \( \sigma_i \) is the standard deviation of \( f_i(Z_i) \).

We now apply these to bound the value of the so-called partition function \( Z(\nu) \) as follows:

**Lemma 31** Let \( \nu > 0 \) be such that \( a := \nu \sigma^2 < 1/2. \) Denoting \( s := 1/(1 - 2a) \), the partition function \( Z(\nu) \) verifies, for all \( A > 0, \)
\[ Z(\nu) \leq 1 + \frac{4a}{1 - 2a}. \] (27)
Proof. Write
\[
Z(\nu) = \int_0^\infty P^0(e^{\nu(F-F)^2} \geq t)dt
= 1 + \int_1^\infty P^0(\nu(F-\overline{F})^2 \geq \log t)dt
= 1 + \int_0^\infty P^0(|G| \geq \frac{1}{a})e^t dt
= 1 + \int_0^\infty P^0(|G| \geq t)2ae^{at^2} dt
= 1 + \int_0^\infty \left[P^0(G \geq t) + P^0(G \leq -t)\right]2ae^{at^2} dt.
\]
Using Hoeffding’s bound (26), the last term is upper-bounded by
\[
1 + 2\int_0^\infty e^{-t^2/2}2ae^{at^2} dt = 1 + 2\left[\frac{-2a}{1-2a} e^{-(t^2/2)(1-2a)}\right]_0^\infty = 1 + \frac{4a}{1-2a}
\]
as announced in (27).

Fix now \(\delta > 0\) and let as before \(\sigma^2\) denote the variance of \(F\) under \(P^0\). We set out to find an \(a > 0\) that is an upper bound of its variance under \(P\) by using the previous two lemmas. In view of Lemma 30, it suffices to verify that for some \(\nu > 0\), Condition \(\nu a - \psi(\nu) \geq \delta\) holds. In view of Lemma 31, denoting the corresponding upper bound to \(\psi(\nu)\) by
\[
\phi(\nu) := \begin{cases} 
\log \left(1 + \frac{4\nu\sigma^2}{1-2\nu\sigma^2}\right) & \text{if } \nu\sigma^2 < 1/2, \\
+\infty & \text{otherwise},
\end{cases}
\]
it suffices to find \(a\) such that for some \(\nu, a\nu - \phi(\nu) \geq \delta\). Maximizing \(\nu a - \phi(\nu)\) over \(\nu\) for fixed \(a\), one finds that the optimal value for \(\nu\) is given by
\[
\nu = \frac{1}{2\sigma^2} \sqrt{\frac{b-4}{b}},
\]
where we introduced the notation \(b := a/\sigma^2\). Plugging this expression for \(\nu\) in \(\nu a - \phi(\nu)\), we have that \(a\) upper-bounds the variance of interest if
\[
2b \sqrt{\frac{b-4}{b}} - \log \left(1 + \frac{2(1-4/b)^{1/2}}{1 - (1-4/b)^{1/2}}\right) \geq \delta.
\]
For \(b \geq 16/3\), it holds that \(1/2 \leq (1-4/b)^{1/2} \leq 1\). Thus under this condition on \(b\), the left-hand side of the above is at least as large as
\[
2b \ast (1/2) - \log \left(\frac{[1 + (1-4/b)^{1/2}]^2}{1 - 1 + 4/b}\right) \geq b - \log(b) \geq (1 - 1/e)b.
\]
We have thus established the following:

Lemma 32 Under the Kullback-Leibler bound of \(\delta\), then the variance of \(F\) under \(P^T\) is upper-bounded by
\[
a = \sigma^2 \max \left(\frac{16}{3}, \frac{\delta}{1 - 1/e}\right).
\]
We can now complete the proof of the Theorem. An upper bound $\delta_T$ on the conditional mutual information obtained after $T$ steps, uniformly over the sketch values observed, is evaluated recursively as

$$\delta_T \leq \delta_{T-1} + \frac{1}{N^2} \sigma^2 \max \left( \frac{16}{3}, \frac{\delta_{T-1}}{1 - 1/e} \right).$$

Recalling that $\sigma^2 \leq 2N\epsilon^2$, we rewrite this for convenience as

$$\delta_T \leq \delta_{T-1} + \frac{C}{N} \max(1, \delta_{T-1}),$$

for some suitable constant $C$.

It then follows that $\delta_T \leq CT/N$ for $T \leq N/C$, and for $T > N/C$ one has

$$\delta_T \leq \left(1 + \frac{C}{N}\right)^T.$$

Thus for any fixed exponent $\alpha > 0$, in order to learn $N^\alpha$ bits of information about the unknown labels $Z_i^N$, one needs at least $T = \alpha \log(N)/\log(1 + C/N) = \Omega(N \log N)$ samples. 

