Algorithms for bounding contribution for histogram estimation under user-level privacy

Yuhan Liu [1] **Ananda Theertha Suresh** [2] **Wennan Zhu** [2] **Peter Kairouz** [2] **Marco Gruteser** [2]

Abstract

We study the problem of histogram estimation under user-level differential privacy, where the goal is to preserve the privacy of *all* entries of any single user. We consider the heterogeneous scenario where the quantity of data can be different for each user. In this scenario, the amount of noise injected into the histogram to obtain differential privacy is proportional to the maximum user contribution, which can be amplified by few outliers. One approach to circumvent this would be to bound (or limit) the contribution of each user to the histogram. However, if users are limited to small contributions, a significant amount of data will be discarded. In this work, we propose algorithms to choose the best user contribution bound for histogram estimation under both bounded and unbounded domain settings. When the size of the domain is bounded, we propose a user contribution bounding strategy that almost achieves a two-approximation with respect to the best contribution bound in hindsight. For unbounded domain histogram estimation, we propose an algorithm that is logarithmic-approximation with respect to the best contribution bound in hindsight. This result holds without any distribution assumptions on the data. Experiments on both real and synthetic datasets verify our theoretical findings and demonstrate the effectiveness of our algorithms. We also show that clipping bias introduced by bounding user contribution may be reduced under mild distribution assumptions, which can be of independent interest.

Part of the work was done during an internship at Google. [1]Cornell University, Ithaca, NY [2]Google Research. Correspondence to: Yuhan Liu <yl2976@cornell.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

1. Introduction

Differential privacy (DP) (Dwork et al., 2006) provides a rigorous formulation of privacy and has been applied to many algorithmic and learning tasks that involve the access to private and sensitive information. Notable applications include private data release (Hardt et al., 2012), learning histograms (Dwork et al., 2006), statistical estimation (Diakonikolas et al., 2015; Kamath et al., 2019; Acharya et al., 2021; Kamath et al., 2020; Acharya et al., 2019a;b), and machine learning (Chaudhuri et al., 2011; Bassily et al., 2014; McMahan et al., 2018b; Dwork et al., 2014; Abadi et al., 2016).

In several applications, each user may contribute many data samples to a dataset. For example, one may have multiple health records in a hospital, or may type many words on their phone's virtual keyboard. Naturally, users would hope that *all* of their private data is protected. To achieve this, private algorithms should guarantee *user-level* differential privacy.

However, many existing works assume that each user only contributes one data sample. Thus, an algorithm designed under this assumption can only be used to protect the privacy of each data sample but not the user. In other words, such algorithms achieve *item-level* privacy, but they cannot protect privacy at the user level and may not meet the increasing privacy concerns in most applications where users may contribute a lot of data. Therefore, there has been a growing interest in revisiting differential privacy in the *user-level* setting.

User-level privacy is much more stringent than the item-level counterpart. Under user-level privacy, the amount of noise added is dependent on the *sensitivity* of the differential privacy mechanism, which is typically the maximum number of items contributed by any single user. Hence, even if a majority of users contribute little data and very few outliers contribute a large amount of data, then the amount of noise added will be significantly large. This begs an important question: should we limit the user contribution while providing differential privacy? This question was initiated by Amin et al. (2019) in the context of empirical risk minimization. They showed that not restricting user

contribution results in a large amount of noise injection and restricting user contribution to achieve low sensitivity, may result in loss of a large amount of useful data and suffer from bias. With this observation, they provided an algorithm that determines near-optimal user contribution bound. In this work, we study the problem of bounding user contribution for general differentially private histogram estimation.

Histogram estimation is a fundamental problem that arises in many real-world applications such as demographic data and user preferences. For example, Chen et al. (2019) used histogram estimation to compute unigram language models via *federated learning* (McMahan et al., 2017; Kairouz et al., 2019). Beyond machine learning, *federated analytics* uses histogram estimation to support the Now Playing feature on Google's Pixel phones, a tool that shows users what song is playing in the room around them (Ramage & Mazzocchi, 2020).

Histogram estimation can be broadly divided into two categories: estimation over bounded domains and unbounded domains. In some examples such as estimating unigram language models over finite known vocabulary, the size of the domain is finite and can be counted. We refer to this scenario as bounded domain histogram estimation. In some examples such as finding all possible words used in English, the size of the domain is the set of all strings and hence unbounded or extremely large. We refer to this scenario as unbounded histogram estimation.

There is abundant literature on histogram estimation in the context of item-level privacy, such as (Hay et al., 2010; Suresh, 2019; Xu et al., 2013). Little has been known, however, under user-level privacy. Recently, (Liu et al., 2020; Levy et al., 2021), studied the problem of histogram estimation under user-level privacy when data from users are generated from near-identical distributions. Since users' data may come from diverse distributions in practice, we cannot leverage techniques from these works. Motivated by the need for algorithms that work well with heterogenous user data, we ask the following question:

Can we design private algorithms to find a (nearly) optimal bound of user contributions for histogram estimation?

Somewhat surprisingly, despite many recent works in the area, the above question has not been extensively studied. We take a step towards answering the above question by designing algorithms that perform well in the heterogeneous setting where both the number of samples and data distribution can be unknown and different across users (hence both need to be viewed as private information).

Our contributions are as follows. We first study the problem of bounded domain histogram estimation, where the domain size is finite and can be enumerated efficiently. In this setting, we propose private user contribution bounding algorithms that obtain a factor two approximation compared to the best contribution bound in hindsight. We then study the problem of unbounded domain histogram estimation, where the domain size is very large and cannot be enumerated efficiently. In this setting, we propose a private user contribution bounding algorithm that achieves a logarithmic approximation compared to the best algorithm in hindsight[1]. Finally, we investigate if the bias introduced by these user contribution bounding algorithms can be reduced by post-processing techniques and show that under mild non i.i.d. distribution assumptions, the amount of bias can be reduced by simple post processing techniques. We also provide a complete proof of the gap between the *debiased* and the *non-debiased* algorithms. We evaluate these algorithms on standard federated datasets and demonstrate the practicality of the algorithms.

The paper is organized as follows. In Section 2, we discuss related work. In Section 3, we introduce the definition and problem formulation. In Section 4 we introduce our algorithms for the bounded domain setting with no i.i.d. assumptions. In Section 5 we describe our algorithms for unbounded domain histogram estimation. In Section 6 we briefly introduce our debiasing algorithm and its guarantees. In Section 7 we show the experiment results. In Section 8, we conclude with a discussion about how to extend our methods to federated settings.

2. Related work

Given its importance, user-level privacy has been studied by several works in the last decade. One of the primary motivations for user-level privacy is federated learning, where the goal is to learn a model at the server while keeping the raw data on edge devices such as cell phones (McMahan et al., 2017; Kairouz et al., 2019). Ensuring privacy at the user level is a crucial concern in federated learning. Even though users do not send their original data, various works (Phong et al., 2017; Wang et al., 2019) have shown it is still possible to reconstruct user's data if additional privacy-preserving mechanisms are not used. Therefore, user-level privacy has been studied under various machine learning tasks in the federated learning setup (McMahan et al., 2018b;a; Augenstein et al., 2019). Indeed, understanding the fundamental privacy-utility trade-offs under user-level privacy is one of the main challenges in federated learning (Kairouz et al., 2019, Section 4.3.2).

Several works studied fundamental theoretical problems in user-level private learning. Ghazi et al. (2021) studied PAC learnability under user-level privacy. Levy et al. (2021) stud-

[1]We emphasize that our near-optimality results are relative to a specific family of DP algorithms, i.e. those that follow the clipping and additive noise recipe. We do not claim optimality over all possible DP mechanisms.

ied high-dimensional distribution estimation and optimization and designed efficient algorithms. Both works require i.i.d. data and assume a fixed number of samples across users. There are several recent works closely related to user-level private histogram estimation. Amin et al. (2019) studied the inherent bias and variance trade-off in bounding user contributions under user-level privacy for empirical risk minimization. Their analysis applies to estimating the total count of one symbol in the aggregate histogram. We extend their work to the setting of $d > 1$ symbols.

Liu et al. (2020) and Levy et al. (2021) studied a closely related problem of discrete distribution estimation and designed optimal algorithms in terms of user complexity (the minimum number of users required to learn an unknown distribution with given accuracy) up to logarithmic factors. However, their analysis assumes that all users' data are drawn from nearly identical distributions. Furthermore, the algorithms algorithms in Liu et al. (2020) may be impractical due to time inefficiency and large constants in user complexity.

Narayanan et al. (2022) studied robust high dimensional mean estimation under user-level privacy, assuming i.i.d. and fixed number of samples across users. While their algorithm is robust to at most 49% of the samples being arbitrarily adversarial, their result still requires that the remaining samples are independent and identically distributed. Cummings et al. (2022) studied mean estimation of Bernoulli random variables, allowing different distributions and number of samples for different users. Their setup can be viewed as a special case of histogram estimation when the domain size is 2. However, no theoretical guarantee is provided when the domain size is larger than 2. Wilson et al. (2020) proposed differentially private SQL with bounded user contributions.

Huang et al. (2021) provides an instance optimal algorithm for bounded-domain histogram estimation. However, it is unclear how their algorithm extends to the case of unbounded domains. In addition, we prove optimality against the best contribution bound in hindsight *for any fixed dataset*, which is orthognonal to their formulation of neighborhood instance optimality. We discuss the differences and contributions compared to their work in more detail in subsequent sections.

3. Problem formulation

Differential privacy (DP) is studied in the central and local settings (Dwork et al., 2006; Kasiviswanathan et al., 2011; Duchi et al., 2013). In this paper, we study the problem under the lens of central differential privacy, where the goal is ensure the algorithm's outcomes do not reveal too much information about any user's data. We now define differential privacy, starting with the basic definition of neighboring datasets. Here, we assume the number of users is known and fixed and hence we use the replacement notion of neighboring datasets.

Definition 3.1. Let $D = \{X_1, \ldots, X_n\}$ respresent a dataset of n users. Each X_i consists of m_i samples $\{X_{i,j}\}_{j=1}^{m_i}$. Let $D' = \{X'_i\}_{i=1}^{n}$ be another dataset. We say D and D' are neighboring (or adjacent) datasets if for some $j \in [n]$,

$$X_i = X'_i, \text{for all } i \neq j.$$

Definition 3.2 (Differential privacy)**.** A randomized mechanism $\mathcal{M}$ with range $\mathcal{R}$ satisfies (ε, δ)-differential privacy if for any two adjacent datasets D, D' and for any subset of output $\mathcal{S} \subseteq \mathcal{R}$, it holds that

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

If $\delta = 0$, then the privacy is also referred to as *pure DP*, and for simplicity we say that the algorithm satisfies ε-DP. If $\delta > 0$, we refer to it as *approximate DP*.

We consider the following problem. There are n users and user i has a histogram $N_i = (N_{i,1}, \ldots, N_{i,d}) \in \mathbb{Z}_{\geq 0}^d$ over a discrete domain of size $d \in \mathbb{N}$. Without loss of generality, we can assume the domain to be $[d] := \{1, \ldots, d\}$. Let $m_i = \|N_i\|_1$ be the size of histogram N_i. The dataset D is the collection of users' histograms. The goal is to estimate the population-level histogram, i.e., the sum of the histograms

$$\overline{N}(D) = \sum_{i=1}^{n} N_i.$$

We make no assumptions about the distribution and size of each user's histogram. Given an (ε, δ)-differentially private algorithm whose output histogram is $\widehat{N}$, we characterize its performance by the expected ℓ_1 distance between the algorithm output and the true population-level histogram

$$\mathbb{E}\|\overline{N} - \widehat{N}\|_1 = \sum_{j=1}^{d} \mathbb{E}|\overline{N}_j - \widehat{N}_j|,$$

where the expectation is over the randomization in the differential privacy algorithm.

4. Optimal user contribution for histograms over bounded domains

We first consider the problem of estimating the population-level histogram when the size of the domain d is small enough to be enumerated. Let $\|\cdot\|_q$ denote the ℓ_q norm. For a vector $x \in \mathbb{R}^d$ and $C \in \mathbb{R}^+$, define the ℓ_q clipping function,

$$\text{clip}_q(x, C) = \frac{C \cdot x}{\max(C, \|x\|_q)}.$$

A standard strategy is to *clip* each user contribution either in ℓ_1 (when $\delta = 0$) or ℓ_2 norm (when $\delta > 0$) and add a suitable amount of Laplace or Gaussian noise respectively (Dwork et al., 2006; Balle & Wang, 2018). For completeness, the details are shown in Algorithm 1. In the rest of the paper, we use the term clipping and bounding user contribution interchangeably. When $\delta > 0$, choosing an appropriate noise level $\sigma = \sigma(\varepsilon, \delta)$ guarantees (ε, δ)-differential privacy.

Algorithm 1 Bounded domain histogram estimation

1: Input: histograms $N_1, \ldots, N_n$, clip threshold C, privacy parameter ε, δ, noise level $\sigma = \sigma(\varepsilon, \delta)$.
2: Clipping: for each user i, do

$$\widetilde{N}_i^1 = \text{clip}_1(N_i, C) \quad \text{and} \quad \widetilde{N}_i^2 = \text{clip}_2(N_i, C)$$

3: If $\delta = 0$, return $\widehat{N}_L = \sum_{i=1}^n \widetilde{N}_i^1 + \text{Lap}(C/\varepsilon)$.
4: If $\delta > 0$, return $\widehat{N}_G = \sum_{i=1}^n \widetilde{N}_i^2 + \text{N}(0, C^2\sigma^2\mathbb{I})$.

Lemma 4.1. *[(Dwork et al., 2006; Balle & Wang, 2018)] When $\delta = 0$, Algorithm 1 guarantees ε-DP. When $\delta > 0$. When $\varepsilon \le 1$, choosing $\sigma^2 = 2\log(1.32/\delta)/\varepsilon^2$ guarantees (ε, δ)-DP for Algorithm 1. When $\varepsilon > 1$, choosing $\sigma = \alpha/\sqrt{2\varepsilon}$ where α is defined in Balle & Wang (2018, Algorithm 1) guarantees (ε, δ)-DP.*

For $\delta > 0$, it is noted that the expression for $\varepsilon > 1$ is much more complicated that $\varepsilon \le 1$, thus for simplicity we mainly focus on $\varepsilon \le 1$.

4.1. Selecting the optimal threshold non-privately

There is a bias-variance trade-off in choosing the clipping threshold C. If C is small, then the noise magnitude (or variance) is small, but the clipped histogram would have large error (or bias). On the other hand, if C is large, the clipped histogram would be more accurate (less bias), but the added noise would be large (high variance). For any dataset D, we provide an accurate characterization of the best threshold that balances the bias and variance for both the Laplace and Gaussian estimator. For the Laplace estimator, the proof is similar to that of Amin et al. (2019) and is omitted.

Lemma 4.2. *Let $\mathcal{L}_L(C, D) = \mathbb{E}[\|\widehat{N}_L - \overline{N}(D)\|_1]$. For any dataset D, choosing C^* as the top $\lceil d/\varepsilon \rceil$ element in $\{m_i\}_{i=1}^n$ yields 2-approximation*

$$\mathcal{L}_L(C^*, D) \le 2 \inf_{C \ge 0} \mathcal{L}_L(C, D),$$

where the expectation is over the Laplace mechanism.

We now state the result for the Gaussian mechanism in Theorem 4.3. The complete proof is in Appendix A.1.

Theorem 4.3. *Let $\mathcal{L}_G(C, D) = \mathbb{E}[\|\widehat{N}_G - \overline{N}(D)\|_1]$. Let $\varepsilon \le 1$ and $M = d\sigma\sqrt{\frac{2}{\pi}}$. For any dataset D, choosing C^* such that*

$$C^* = \arg\min_{C \ge 0} \left\{ \sum_{i:\|N_i\|_2 > C} \frac{\|N_i\|_1}{\|N_i\|_2} \le M \right\}$$

yields 2-approximation,

$$\mathcal{L}_G(C^*, D) \le 2 \inf_{C \ge 0} \mathcal{L}_G(C, D).$$

4.2. Choosing the optimal threshold privately

We now discuss how to find C^* privately using an additional privacy budget of (ε', δ'), and further provide complete guarantees in terms of excess error compared to $2\inf_{C\ge 0}\mathcal{L}_G(C, D)$. We emphasize that it only requires a very small extra privacy budget compared to the original (ε, δ) to achieve good performance, as we will later show in the experiments. For $\delta = 0$, one can levarage many existing algorithms to privately find the d/ε quantile in Lemma 4.2 (Dick et al., 2023, Theorem 2). Thus we mainly focus on the $\delta > 0$ case.

Note that computing the optimal C in a differentially private way is possible though difficult because the sensitivity of $\sum_{i:\|N_i\|_2>C} \frac{\|N_i\|_1}{\|N_i\|_2}$ can be very large for some datasets. However, observe that by Cauchy-Schwarz inequality,

$$\|N_i\|_1/\|N_i\|_2 \le \sqrt{\|N_i\|_0}.$$

Hence, if each user's histogram has very few non-zero entries, then the sensitivity would be low.

We observe this to be the case in practice. To illustrate, we plot the unique number of symbols contributed by each user in Sentiment140 (Go et al., 2009) and SNAP datasets in Figure 1. Observe that even though d is large, most users have fewer than 200 samples. Hence, we assume that each user's histogram is at most s sparse. Under this assumption, the sensitivity is upper bounded by $\sqrt{s}$. Note that we can simply set $s = d$ if the bound is not known.

We first note that C^* can be written as a minimizer to a convex function,

$$G(C) = \sum_{i=1}^n f_i(C) + CM,$$

where $f_i(C) = \max\left\{1 - \frac{C}{\|N_i\|_2}, 0\right\}\|N_i\|_1$. Hence we can use techniques from differentially private convex optimization algorithms. We consider two such algorithms and provide their corresponding guarantees.

Estimating C^* with DP-SGD. We first consider the DP-SGD algorithm (Bassily et al., 2014, Algorithm 1) to estimate C by minimizing $G(C)$. Using Bassily et al. (2014, Theorem 2.4), we have the following guarantee.

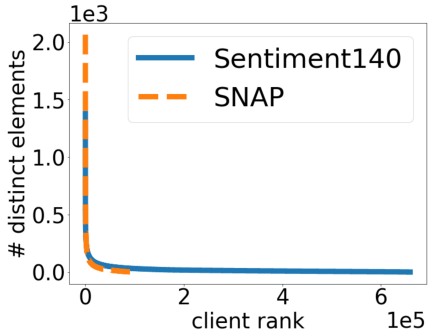

Figure 1. Number of distinct words in each client. The x-axis is the rank of clients ordered from the highest count of distinct words to the lowest.

Corollary 4.4. *Let C_m be an upper bound on C^* and let C_{DPSGD} be the output of Bassily et al. (2014, Algorithm 1). Assume that $n \geq d$. Then, $\mathbb{E}[\mathcal{L}_G(C_{DPSGD}, D)] - 2\inf_{C\geq 0}\mathcal{L}_G(C, D)$ is upper bounded by*

$$O\left(\frac{C_m\left(\sqrt{s}+\frac{\sqrt{\log(1/\delta)}}{\varepsilon}\right)}{\varepsilon'}\log^{3/2}(n/\delta')\sqrt{\log(1/\delta')}\right).$$

Proof. The proof directly follows by noting that $G(C) = \sum_{i=1}^{n}(f_i(C) + \frac{MC}{n})$. Hence the Lipschitz constant for $f_i(C) + MC/n$ w.r.t. C is $L = \sqrt{s} + M/n = \sqrt{s} + O\left(\frac{\sqrt{\log(1/\delta)}}{\varepsilon}\right)$. Since $C_m \geq C^*$, setting the domain diameter $\|\mathcal{C}\|_2 = C_m$ in Bassily et al. (2014, Theorem 2.4) completes the proof. □

Estimating C^* with output perturbation. We consider the second algorithm based on output perturbation (Chaudhuri et al., 2011), which ensures $(\varepsilon', 0)$-DP and is good for small n and ε'. Here, we solve a regularized convex optimization problem and perturb the output to provide differential privacy. The algorithm is outlined in Algorithm 2.

Algorithm 2 Clipping threshold estimation with output perturbation

1: Input: histograms $N_1, \ldots, N_n$, an upper bound of C^* denoted by C_m, sparsity parameter s, privacy parameter ε'.
2: Let $\lambda = \frac{2\sqrt{2s}}{C_m\sqrt{n\varepsilon}}$ and $\Delta = \frac{4\sqrt{s}}{\lambda n}$ Compute C', the minimizer of $F(C) = \frac{1}{n}G(C) + \frac{\lambda}{2}C^2$.
3: Return $C_{\text{output}} = C' + \text{Lap}(\Delta/\varepsilon')$.

With appropriate parameters, the combined algorithm almost achieves a 2-approximation with respect to the best clipping threshold.

Corollary 4.5. *Algorithm 2 is $(\varepsilon', 0)$ differentially private. If C_m is an upper bound on C^*, setting $\lambda = \frac{2\sqrt{2s}}{C_m\sqrt{n\varepsilon'}}$ yields an error*

$$\mathbb{E}[\mathcal{L}_G(C_{output}, D)] \leq 2\inf_{C\geq 0}\mathcal{L}_G(C, D) + 2\sqrt{2}\,C_m\sqrt{\frac{ns}{\varepsilon'}}.$$

Comparing DP-SGD and Algorithm 2, we can see that DP-SGD has a better asymptotic dependence on n, and Algorithm 2 has a better dependence on ε'. Furthermore, DP-SGD provides an approximate DP guarantee and Algorithm 2 gives a pure DP guarantee. Finally, the time complexity of DP-SGD is typically $O(n^2)$, however it has been improved recently to $O(n)$ with similar guarantees (Feldman et al., 2020).

Remark on the sparsity s. We emphasize that Lemma 4.2 and Theorem 4.3 are general results that *do not* require bounded ℓ_0 norms of histograms. Moreover, the convergence results in Corollary 4.4 and 4.5 follow directly by replacing with s with d if the ℓ_0 bound is not satisfied. Hence as long as n is sufficiently larger than d, the excess error is still small. Empirically, we show that even setting $s = d$ yields a performance close to the true 2-approximation threshold, and is much better than choosing the threshold according to (Amin et al., 2019). See Appendix D.2 for details.

Comparison with Huang et al. (2021) (Huang et al., 2021) proposed a clipping-based algorithm to very similar to ours to minimize ℓ_2 estimation error. Their algorithm first applies ℓ_2 clipping and then adds suitable amount of Gaussian noise. They further proved instance optimality over a neighborhood of D. However there are several key differences.

Huang et al. (2021) proved instance optimality against all differentially private algorithms over some *neighborhood* of D, while our results indicate the optimality of all clipping algorithms for any *fixed* dataset. Therefore, the results in Huang et al. (2021) and our work are orthogonal to each other (in that neither result implies the other) and take different perspectives on the same problem.

The technical difference is that they focused on minimizing the ℓ_2 norm instead of the ℓ_1 norm. As a result, their threshold is a quantile of the ℓ_2 norms of all users' histograms, which is very different from the optimal threshold in Theorem 4.3. Not surprisingly, as we will demonstrate in Section 7, their algorithm does not perform well when the error metric is ℓ_1.

Algorithm 3 Unbounded domain histogram estimation

Input: privacy parameters ε, δ, histograms $N_1, \ldots, N_n$, threshold C.
$\overline{N} = \sum_{i=1}^n N_i$.
$t = C + \frac{C}{\varepsilon} \log \frac{C}{2\delta}$.
For each user i, $h_i = \text{rand-clip}(N_i, C)$.
$\widetilde{N} = \sum_{i=1}^n h_i + Z$ where $Z = [Z_j \mathbf{1}_{\overline{N}_j > 0}]_{j=1}^d$ and $Z_j \sim \text{Lap}(C/\varepsilon)$.
Return $\widehat{N}$ where for each item $j \in [d]$ such that $\overline{N}_j > 0$,

$$\widehat{N}_j = \widetilde{N}_j \mathbf{1}_{\widetilde{N}_j > t}.$$

5. Optimal contribution for histograms over unbounded domains

In the unbounded domain setting, the domain size can be prohibitively large or even infinite, so it is not practical to add noise to all items in the domain. We describe an algorithm for unbounded domain histogram estimation in Algorithm 3 based on the sparse vector technique (Dwork et al., 2009). Even though d is very large, the run time of the algorithm depends only on the number of items with non-zero counts. However, in this approach, the privacy guarantee not only depends on the ℓ_1 norm of the user contribution but also on the ℓ_∞ and ℓ_0 norms.

While the standard ℓ_1 clipping defined in the previous section reduces the ℓ_1 norm, it does not reduce the ℓ_0 norm of the histogram. Hence, we use the following randomized clipping strategy: for a histogram N and integer $C > 0$, let rand-clip(N, C) be the histogram obtained by sampling $\min\{\|N\|_1, C\}$ items without replacement. Since all histograms are integer-valued, the ℓ_∞ norm of the clipped histogram is upper bounded by the ℓ_1 norm.

In this technique, each user first uses rand-clip to clip their histogram to ensure ℓ_1 and ℓ_∞ norm to be less than C. Then an appropriate amount of Laplace noise is added to each non-zero count. Finally, we delete all items with counts less than a threshold t and output the histogram of the remaining symbols and their noisy counts. The privacy guarantee is stated in Lemma 5.1.

Lemma 5.1. *Algorithm 3 is (ε, δ)-diffentially private.*

Proof. Note that by random clipping, each $\|\bar{h}_i\|_r \le C$ for $r = 0, 1, 2, \infty$. By the guarantee of the Laplace mechanism, $\widetilde{N}$ is $(\varepsilon, 0)$-DP. Recall that the CDF of Z is given by $\Phi(x) = 1 - \frac{1}{2} e^{-\varepsilon x / C}$ for $x > 0$. Thus,

$$\Phi\left(\frac{C}{\varepsilon} \log \frac{C}{2\delta}\right) = 1 - \frac{\delta}{C} \ge (1-\delta)^{1/C}.$$

The final inequality is due to Bernoulli's inequality $(1 + x)^r \le 1 + rx$ for $x \ge -1$ and $r \in [0, 1]$. Instantiating Google (2020, Theorem 1) proves the differential privacy guarantees. □

Assume that C_m is an upper bound on $\|N_i\|_1, i \in [n]$ which could potentially be very large. Then, we only need to focus on $C \le C_m$ (we are minimizing the error with respect to C, and it is common to assume some bound on optimization variables). In the next theorem we provide a tight characterization of the expected ℓ_1 error of Algorithm 3 up to logarithmic factors.

Theorem 5.2. *Assume that $1 \le C \le C_m \le e^{\varepsilon/(3\delta)}$, $\delta \le 1/8$, where C_m is the maximum contribution of any user before clipping.*

$$\frac{1}{2} \sum_{i=1}^n \max\{\|N_i\|_1 - C, 0\} + \frac{1}{12 \log \frac{C_m}{2\delta}} \sum_{j:\overline{N}_j > 0} \mathbb{E}[\min(\bar{h}_j, t)] \le \mathbb{E}[\|\widehat{N} - \overline{N}\|_1] \le 2 \sum_{i=1}^n \max\{\|N_i\|_1 - C, 0\} + \sum_{j:\overline{N}_j > 0} \mathbb{E}[\min(\bar{h}_j, t)]. \quad (1)$$

Details of the proof is in Appendix B. We argue that the assumption on C, ε, and δ is very mild. δ is set as $O(1/n)$ and ε is chosen to be a constant near 1 (say 0.5 to 5), which implies that the upper bound on C is exponential in n.

If C^* minimizes the upper bound in Theorem 5.2, then C^* yields a logarithmic approximation. However, the upper bound in (1) depends on C directly via the $\sum_{i=1}^n \max\{\|N_i\|_1 - C, 0\}$ and indirectly via randomly clipped histogram $\bar{h}_j$ and the threshold t. Furthermore, the expression is non-convex in C, thus convex optimization approaches may not yield provable guarantees. Hence, to privately estimate C^*, one can obtain the function values for all integers $0 < C \le C_m$ and apply the exponential mechanism with an additional small privacy budget ε'.

Corollary 5.3. *Let $\mathcal{L}_O(C, D)$ be the expected ℓ_1 error of Algorithm 3 on dataset D given threshold C. Let $\hat{C}$ be the output of the exponential mechanism with additional privacy budget ε'. Then, $\mathcal{L}_O(\hat{C}, D)$ is upper bounded by*

$$12 \log \frac{C_m}{2\delta} \inf_{C \ge 1} \mathcal{L}_O(C, D) + \frac{6 C_m \log C_m}{\varepsilon'}.$$

Proof. Let C_{opt} be the threshold that minimizes the ℓ_1 error. Write the lower and upper bounds in Theorem 5.2 as $L(C)$ and $U(C)$ respectively. Using the fact that C^* minimizes (1) and applying Theorem 5.2,

$$U(C^*) \le 12 \log \frac{C_m}{2\delta} L(C^*) \le 12 \log \frac{C_m}{2\delta} L(C_{opt}) \le 12 \log \frac{C_m}{2\delta} \mathcal{L}_O(C_{opt}, D).$$

Note that changing the data of one user changes (1) by at most $3C_m$. By the utility of the exponential mechanism,

$$\mathcal{L}_O(\hat{C}, D) \le U(\hat{C}) \le U(C^*) + \frac{6C_m \log C_m}{\varepsilon'} \le 12 \log \frac{C_m}{2\delta} \mathcal{L}_O(C_{opt}, D) + \frac{6C_m \log C_m}{\varepsilon'}.$$

□

In practice, the expectation in (1) can be hard to compute. By Jensen's inequality,

$$\mathbb{E}[\min(\bar{h}_j, t)] \le \min(\mathbb{E}[\bar{h}_j], t).$$

Hence, we propose to replace the former with $\min(\mathbb{E}[\bar{h}_j], t)$. Observe that

$$\mathbb{E}[\bar{h}_j] = \mathbb{E}\left[\sum_{i=1}^{n} h_{i,j}\right] = \sum_{i=1}^{n} \mathbb{E}[h_{i,j}] = \sum_{i=1}^{n} \frac{CN_{ij}}{\max(C, \|N\|_1)}.$$

Hence,

$$V(C) = 2\sum_{i=1}^{n} \max\{\|N_i\|_1 - C, 0\} + \sum_{j:\overline{N}_j > 0} \min\left(\sum_i \frac{CN_{ij}}{\max(C, \|N\|_1)}, C + \frac{C}{\epsilon} \log \frac{C}{2\delta}\right).$$

Searching over all possible (integer) values of C can also be inefficient. We can instead search over a subset $\mathcal{C} \subseteq [C_m]$. We describe the procedure to privately find the best threshold C in Algorithm 4.

Algorithm 4 Private threshold selection for histograms over unbounded domain

1: Input: privacy parameters ε, δ for Algorithm 3, privacy parameters ε', δ' for estimating the optimal threshold, user histograms $N_1, \ldots, N_n, C_m$.
2: Select a subset $\mathcal{C}$ of $\{1, \ldots, C_m\}$.
3: For each $C \in \mathcal{C}$, compute $V(C)$
4: Return $\hat{C}$, the output of the exponential algorithm with privacy parameter ε' and sensitivity $5C_m/2$ over $\{V(C) : C \in \mathcal{C}\}$.

In Section 7, we empirically demonstrate that Algorithm 4 can also achieve an error very close to the true optimal threshold.

6. Bias reduction

We prove that the bias from clipping can be significantly reduced when N_i's satisfy some mild distribution assumptions and show that the debiasing method provides improvements even on real datasets where these assumptions may not necessarily hold. Consider the special case of $d = 1$, which we refer to as *count estimation*. Let $\mathcal{D}$ be a family of distributions over $\mathbb{Z}_{\ge 0}$. For each user i, N_i is drawn independently from some distribution in $\mathcal{D}$ with mean $\lambda_i > 0$. λ_i's can be arbitrary and do not need to be equal.

In addition to the absolute error of counts $|\widehat{N} - \overline{N}|$, we also want to characterize the accuracy for estimating the mean $\bar{\lambda} = \frac{1}{n}\sum_{i=1}^{n} \lambda_i$. Let $\hat{\lambda} = \widehat{N}/n$ be an estimate of $\bar{\lambda}$. We are interested in the expected square error

$$\mathbb{E}[(\bar{\lambda} - \hat{\lambda})^2],$$

where the expectation is over the randomness of the algorithm and the dataset.

In this work we set $\mathcal{D}$ to be the family of Poisson distributions since they arise in many applications. For example, they can be used to model the occurrences of a memoryless event in a fixed time window. Also, they are good approximations of the binomial distribution $\text{Bin}(m, p)$ when mp is a constant (Le Cam, 1960), and can be very useful when estimating the count of one element in a histogram over a very large domain (e.g. the count of a particular word).

It is easy to see that clipping inevitably induces bias. In many practical situations, it is often reasonable to make mild distribution assumptions on user data. In this section, we ask if the clipping bias can be reduced with such assumptions. We answer affirmatively for bounded domain with $d = 1$ under non-i.i.d. Poisson assumptions on each user's count.

Our algorithm is shown in Algorithm 5. It essentially adds a post-processing procedure on the output of Algorithm 1 to reduce the clipping bias. Since this is a post-processing step, it does not affect the privacy guarantees. We show a detailed analysis of the performance of Algorithm 5 and discuss two possible extensions to high dimensions in the appendix.

Algorithm 5 Debiasing algorithm for Poisson distribution

1: Input: $N_1, \ldots, N_n, C \in \mathbb{N}$.
2: $h(\lambda) = \mathbb{E}_{X \sim \text{Poi}(\lambda)}[\text{clip}(X, C)]$
3: $Y_i = \text{clip}(N_i, C)$
4: Return $\widehat{N} = g\left(\sum_{i=1}^{n} Y_i + Lap(C/\varepsilon)\right)$, where $g(y) = nh^{-1}(y/n)$

7. Experiments

We run experiments on two real-world datasets: Sentiment140 (Go et al., 2009), a twitter dataset that contains user tweets, and SNAP (Cho et al., 2011), a social network dataset that contains the location information of check-ins by users. For Sentiment140, we parse each user's tweets to words, and treat each word as an element. For SNAP,

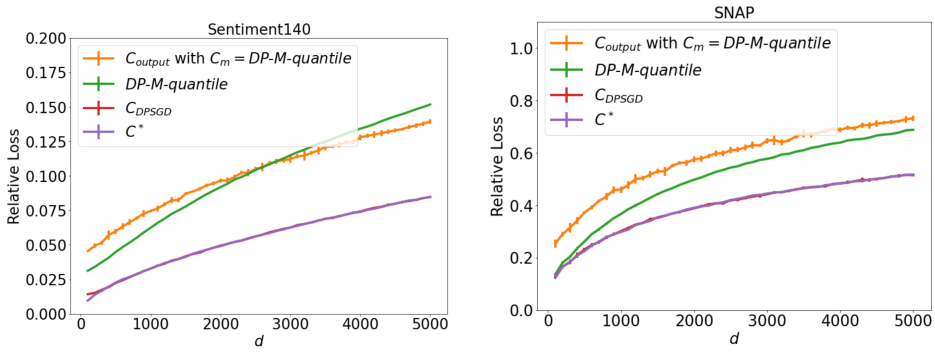

Figure 2. Histogram estimation over bounded domains. **Left:** Sentiment140 **Right:** SNAP.

Table 1. Relative loss of the best threshold and Algorithm 4 for unbounded domains.

Dataset	Support size d (unknown)	Best C	Private C^*	Median	90% quantile
i.i.d.	50	0.001	0.0015	0.0397	0.0049
i.i.d.	100	0.0026	0.0026	0.0394	0.0053
i.i.d.	200	0.0048	0.0048	0.0397	0.0068
non-i.i.d.	50	0.0094	0.0124	0.6957	0.1645
non-i.i.d.	100	0.0175	0.0224	0.6957	0.1685
non-i.i.d.	200	0.0461	0.0462	0.7016	0.1713
Sent. 140	100	0.0076	0.0285	0.6027	0.2742
Sent. 140	1000	0.1250	0.1483	0.5912	0.2699

each element is a location, and each user has check-ins to multiple locations in the dataset.

Since running on all elements (an order of 10^6) is costly and and the error is ususally prohibitively large, we choose the top d elements in the datasets and only run experiments on those. We measure the relative loss of $\widehat{N}$,

$$\frac{\sum_{j=1}^{d} |\overline{N}_j - \widehat{N}_j|}{\|\overline{N}\|_1}. \tag{2}$$

7.1. Bounded domain

In all experiments, the privacy budget for estimating C is $\varepsilon = 0.1, \delta = 1/2n$, and the budget for Algorithm 1 is $\varepsilon = 1, \delta = 1/2n$. For DP-SGD with sparsity assumptions, we set $s = 0.1d$[2] and clip each $\|N_i\|_1/\|N_i\|_2$ to $\sqrt{s}$ when estimating C^*. This introduces bias when the assumption is not satisfied for some users. However if the percentage of such users is small, this effect can be negligible.

We evaluate different algorithms for estimating the clipping threshold C for the Gaussian mechanism given in Algorithm 1. We compare the performance of the following methods: (i) C^*: The non-private clipping threshold given in Theorem 4.3. (ii) *DP-M-quantile*: inspired by the 2-approximation quantile in (Amin et al., 2019), we set C to be the M^{th} largest value of $\|N_i\|_2$, where M is given in Theorem 4.3. We estimate it by gradient descent with differential privacy, e.g. Andrew et al. (2021, Section 2). This corresponds to a slightly different private version of the clipped-mean estimator in Huang et al. (2021, Section 3). (iii) C_{DPSGD}: estimation of C^* with DPSGD algorithm (Corollary 4.4). and (iv) C_{output}: estimation of C^* with output pertubation (Algorithm 2).

In Figure 2, we show the comparison of these threshold estimation algorithms with different choices of d in $[100, 5000]$. The results with both datasets are similar, but SNAP has much higher errors, possibly because of the location information in SNAP is more non-i.i.d compared to the words in Sentiment140. Setting C to *DP-M-quantile* according to (Amin et al., 2019) and (Huang et al., 2021) does not have any theoretical support, and the errors are relatively high. For Algorithm 2, we run experiments with $C_m =$ *DP-M-quantile* and $C_m = 150$ (see Appendix D). Of all the algorithms, C_{DPSGD} has similar performance to the true C^* without differential privacy.

7.2. Unbounded domain

In this section, we run experiments for Algorithm 3 with threshold C chosen by Algorithm 4. We tested on both

[2]The choice $s = 0.1d$ is arbitrary (i.e. not a function of the underlying datasets) and has not been tuned.

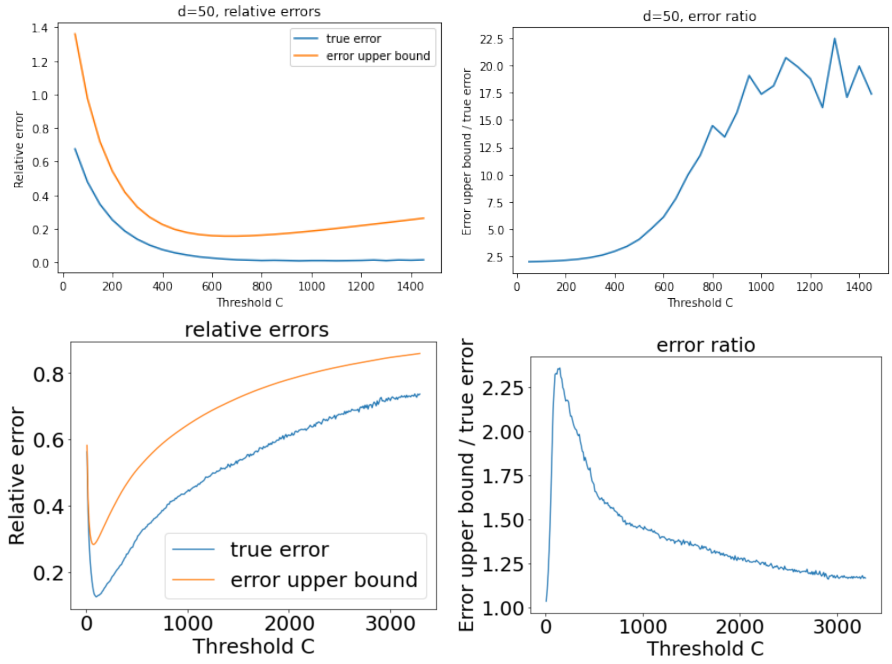

Figure 3. Error plots for different datasets for unbounded domains. **Top**: synthetic non-i.i.d data with $d = 50$ (unknown). **Bottom:** Sentiment 140 with $d = 1000$ (unknown). **Left column**: comparison of true error and predicted error upper bound. **Right column:** the orange/blue ratio in the left plots.

real and synthetic datasets. We set $\varepsilon = 1, \varepsilon' = 0.1$ and $\delta = 1/(2n)$ where n is the number of users in the respective datasets. We compared our algorithms to two non-private baselines where C is the median and the 90% quantile of the ℓ_1 norms. In this section d should be interpreted as the actual support size of the aggregate histogram that is not known beforehand. Our algorithm *does not* require prior knowledge of d.

For Sentiment 140 we choose $d = 100$ and 1000 words and treat those as the support of the histograms. We generated synthetic datasets with $n = 5 \times 10^5$ users with both i.i.d. and non-i.i.d. data over support sizes $d = 50$, 100, or 200. Let $\mathbf{p}$ be a discrete distribution over $[d]$ with probability mass proportional to $1/(j + 50)$ for $j \in [d]$. In the i.i.d. setting, each user draws $\text{Poi}(100)$ samples from $\mathbf{p}$. In the non-i.i.d. setting, let $\lambda_1, \ldots, \lambda_i \sim 100\text{Dir}(2)$, user i draws $\text{Poi}(\lambda_i)$ samples from $\mathbf{p}_i \sim \text{Dir}(\mathbf{p}/2)$.

We search for the best C over $[10, 20, \ldots, 1500]$. Figure 3 compares the average error of Algorithm 3 over 3 independent runs (blue) and the error upper bound in Algorithm 4 (orange) on the non-i.i.d. synthetic dataset with $d = 50$. We can see that (1) indeed upper bounds the expected error, and their ratio is within $O(\log(C_m/2\delta))$.

Table 1 compares the performance of the best C (obtained non-privately) and the private estimate obtained by Algorithm 4. The performance is measured by relative loss defined by (2). We can see that for all datasets, the private estimate is close to the performance achieved by the best threshold, and significantly outperforms the non-privately chosen median and 90% quantiles. This further suggests the need to threshold according to the dataset instead of choosing a fixed threshold or quantile. Furthermore, focusing on the results for Sentiment 140 with $d = 1000$ (i.e. the bottom row of Figure 3 and Table 1), we observe that our algorithm yields good performance even when C_m significantly overshoots the true optimal threshold, which demonstrates robustness against the choice of C_m.

8. Conclusion

We studied histogram estimation under user-level differential privacy in the heterogeneous scenario for bounded and unbounded domains. We proposed algorithms to choose the best user contribution bound that achieve 2-approximation and logarithmic approximation for bounded and unbounded domains respectively. We also showed that clipping bias introduced by bounding user contribution may be reduced under distribution assumptions. Finally, we empirically demonstrated the practicality of the proposed methods.

9. Acknowledgements

The authors thank Alex Kulesza for helpful comments and discussions.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 308–318, 2016.

Acharya, J., Canonne, C. L., and Tyagi, H. Inference under information constraints: Lower bounds from chi-square contraction. *Proceedings of Machine Learning Research vol*, 99:1–15, 2019a.

Acharya, J., Sun, Z., and Zhang, H. Hadamard response: Estimating distributions privately, efficiently, and with little communication. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1120–1129, 2019b.

Acharya, J., Sun, Z., and Zhang, H. Differentially private assouad, fano, and le cam. In *Algorithmic Learning Theory*, pp. 48–78. PMLR, 2021.

Amin, K., Kulesza, A., Munoz, A., and Vassilvtiskii, S. Bounding user contributions: A bias-variance trade-off in differential privacy. In *International Conference on Machine Learning*, pp. 263–271. PMLR, 2019.

Andrew, G., Thakkar, O., McMahan, H. B., and Ramaswamy, S. Differentially private learning with adaptive clipping. In *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=RUQ1zwZR8_.

Augenstein, S., McMahan, H. B., Ramage, D., Ramaswamy, S., Kairouz, P., Chen, M., Mathews, R., and y Arcas, B. A. Generative models for effective ml on private, decentralized datasets. In *International Conference on Learning Representations*, 2019.

Balle, B. and Wang, Y.-X. Improving the Gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 394–403. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/balle18a.html.

Bassily, R., Smith, A., and Thakurta, A. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pp. 464–473. IEEE, 2014.

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

Chen, M., Suresh, A. T., Mathews, R., Wong, A., Allauzen, C., Beaufays, F., and Riley, M. Federated learning of n-gram language models. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pp. 121–130, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/K19-1012. URL https://aclanthology.org/K19-1012.

Cho, E., Myers, S. A., and Leskovec, J. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1082–1090, 2011.

Cummings, R., Feldman, V., McMillan, A., and Talwar, K. Mean estimation with user-level privacy under data heterogeneity. In *Advances in Neural Information Processing Systems*, 2022.

Diakonikolas, I., Hardt, M., and Schmidt, L. Differentially private learning of structured discrete distributions. In *Advances in Neural Information Processing Systems 28*, NIPS '15, pp. 2566–2574. Curran Associates, Inc., 2015.

Dick, T., Kulesza, A., Sun, Z., and Suresh, A. T. Subset-based instance optimality in private estimation. *arXiv preprint arXiv:2303.01262*, 2023.

Duchi, J. C., Jordan, M. I., and Wainwright, M. J. Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pp. 429–438. IEEE, 2013.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pp. 265–284. Springer, 2006.

Dwork, C., Naor, M., Reingold, O., Rothblum, G. N., and Vadhan, S. On the complexity of differentially private data release: efficient algorithms and hardness results. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pp. 381–390, 2009.

Dwork, C., Talwar, K., Thakurta, A., and Zhang, L. Analyze Gauss: Optimal bounds for privacy-preserving principal component analysis. In *Proceedings of the 46th Annual ACM Symposium on the Theory of Computing*, STOC '14, pp. 11–20, New York, NY, USA, 2014. ACM.

Feldman, V., Koren, T., and Talwar, K. Private stochastic convex optimization: Optimal rates in linear time. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2020, pp. 439–449, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369794. doi: 10.1145/3357713.3384335. URL https://doi.org/10.1145/3357713.3384335.

Ghazi, B., Kumar, R., and Manurangsi, P. User-level differentially private learning via correlated sampling. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=PqiCvohYSAx.

Go, A., Bhayani, R., and Huang, L. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1(12), 2009.

Google. Delta calculation for thresholding. https://github.com/google/differential-privacy/blob/main/common_docs/Delta_For_Thresholding.pdf, 2020.

Hardt, M., Ligett, K., and McSherry, F. A simple and algorithm for differentially private data release. In *Proceedings of the 25th International Conference on Neural Information Processing Systems-Volume 2*, pp. 2339–2347, 2012.

Hay, M., Rastogi, V., Miklau, G., and Suciu, D. Boosting the accuracy of differentially private histograms through consistency. *Proceedings of the VLDB Endowment*, 3 (1-2):1021–1032, 2010.

Huang, Z., Liang, Y., and Yi, K. Instance-optimal mean estimation under differential privacy. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 25993–26004. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/da54dd5a0398011cdfa50d559c2c0ef8-Paper.pdf.

Kairouz, P., Oh, S., and Viswanath, P. The composition theorem for differential privacy. *IEEE Transactions on Information Theory*, 63(6):4037–4049, 2017.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.

Kamath, G., Li, J., Singhal, V., and Ullman, J. Privately learning high-dimensional distributions. In *Proceedings of the 32nd Annual Conference on Learning Theory*, 2019.

Kamath, G., Singhal, V., and Ullman, J. Private mean estimation of heavy-tailed distributions. In *Conference on Learning Theory*, pp. 2204–2235. PMLR, 2020.

Kasiviswanathan, S. P., Lee, H. K., Nissim, K., Raskhodnikova, S., and Smith, A. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

Le Cam, L. An approximation theorem for the poisson binomial distribution. *Pacific Journal of Mathematics*, 10 (4):1181–1197, 1960.

Levy, D., Sun, Z., Amin, K., Kale, S., Kulesza, A., Mohri, M., and Suresh, A. T. Learning with user-level privacy. In *Advances in Neural Information Processing Systems*, 2021.

Liu, Y., Suresh, A. T., Yu, F. X. X., Kumar, S., and Riley, M. Learning discrete distributions: user vs item-level privacy. In *Advances in Neural Information Processing Systems*, volume 33, pp. 20965–20976, 2020.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

McMahan, B., Andrew, G., Mironov, I., Papernot, N., Kairouz, P., Chien, S., and Úlfar Erlingsson. A general approach to adding differential privacy to iterative training procedures. 2018a. URL https://arxiv.org/pdf/1812.06210.pdf. Workshop on Privacy Preserving Machine Learning (NeurIPS 2018).

McMahan, H. B., Ramage, D., Talwar, K., and Zhang, L. Learning differentially private recurrent language models. In *International Conference on Learning Representations*, 2018b.

Narayanan, S., Mirrokni, V., and Esfandiari, H. Tight and robust private mean estimation with few users. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16383–16412. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/narayanan22a.html.

Phong, L. T., Aono, Y., Hayashi, T., Wang, L., and Moriai, S. Privacy-preserving deep learning: Revisited and enhanced. In *International Conference on Applications and Techniques in Information Security*, pp. 100–110. Springer, 2017.

Ramage, D. and Mazzocchi, S. Federated analytics: Collaborative data science without data collection. https://ai.googleblog.com/2020/05/federated-analytics-collaborative-data.html, 2020.

Suresh, A. T. Differentially private anonymized histograms. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Wang, Z., Song, M., Zhang, Z., Song, Y., Wang, Q., and Qi, H. Beyond inferring class representatives: User-level privacy leakage from federated learning. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*, pp. 2512–2520. IEEE, 2019.

Wilson, R. J., Zhang, C. Y., Lam, W., Desfontaines, D., Simmons-Marengo, D., and Gipson, B. Differentially private sql with bounded user contribution. *Proceedings on Privacy Enhancing Technologies*, 2:230–250, 2020.

Xu, J., Zhang, Z., Xiao, X., Yang, Y., Yu, G., and Winslett, M. Differentially private histogram publication. *The VLDB Journal*, 22(6):797–822, 2013.

A. Detailed proof for bounded domain algorithms

A.1. Proof of Theorem 4.3

Proof. Recall that σ is a function of ε, δ and $M = d\sigma\sqrt{\frac{2}{\pi}}$. We can upper bound the error as follows.

$$\begin{aligned}
\mathbb{E}[\|\widehat{N} - \overline{N}\|_1] &= \mathbb{E}\left[\left\|\sum_{i=1}^{n} \text{clip}_2(N_i, C) + \mathcal{N}(0, \mathbb{I}C^2\sigma^2) - \sum_{i=1}^{n} N_i\right\|_1\right] \\
&\leq \mathbb{E}\left[\left\|\sum_{i} \text{clip}_2(N_i, C) - \sum_{i=1}^{n} N_i\right\|_1\right] + \mathbb{E}[\|\mathcal{N}(0, \mathbb{I}C^2\sigma^2)\|_1] \\
&= \sum_{i:||N_i||_2 > C} \left(1 - \frac{C}{||N_i||_2}\right) ||N_i||_1 + C \cdot M \\
&= \sum_{i=1}^{n} \max\left\{1 - \frac{C}{||N_i||_2}, 0\right\} ||N_i||_1 + C \cdot M = G(C). \qquad (3)
\end{aligned}$$

Equation 3 is convex with respect to C. To optimize the upper bound on the error, we will take the sub-derivative with respect to C and set it to zero. This gives us the following equation

$$\sum_{i:||N_i||_2 > C} \frac{||N_i||_1}{||N_i||_2} = M. \qquad (4)$$

Roughly we want to choose C that satisfies the above equality. The precise value of C^* is

$$C^* = \arg\min_{C \geq 0} \left\{ \sum_{i:\|N_i\|_2 > C} \frac{\|N_i\|_1}{\|N_i\|_2} \leq M \right\}$$

C^* minimizes the right hand side of (3), and it also makes the expected ℓ_1 loss at most twice the loss of the optimal loss with this algorithm. Formally, suppose Q is the ℓ_2-norm that minimizes $\mathbb{E}[\|\widehat{N} - \overline{N}\|_1]$. Let $\mathcal{Z} = [Z_1, \ldots, Z_d] \sim \mathcal{N}(0, I\sigma^2)$ and $\text{clip}_2(N_i, Q)_j$ be the j the coordinate of $\text{clip}_2(N_i, Q)$, then:

$$\begin{aligned}
&\mathbb{E}\left[\left\|\sum_{i} \text{clip}_2(N_i, Q) + \mathcal{Z} - \sum_{i} N_i\right\|_1\right] = \sum_{j=1}^{d} \mathbb{E}\left[|\text{clip}_2(N_i, Q)_j - N_{i,j} + Z_j|\right] \\
=&\sum_{j=1}^{d} \mathbb{E}\left[\left|\sum_{i} \text{clip}_2(N_i, Q)_j + Z_j - \sum_{i} N_{i,j}\right| \middle| Z_j < 0\right] \cdot \Pr(Z_j < 0) \\
&+ \sum_{j=1}^{d} \mathbb{E}\left[\left|\sum_{i} \text{clip}_2(N_i, Q)_j + Z_j - \sum_{i} N_{i,j}\right| \middle| Z_j \geq 0\right] \cdot \Pr(Z_j \geq 0) \\
\geq&\sum_{j=1}^{d} \mathbb{E}\left[\left|\sum_{i} \text{clip}_2(N_i, Q)_j + Z_j - \sum_{i} N_{i,j}\right| \middle| Z_j < 0\right] \cdot \Pr(Z_j < 0) \\
=&\frac{1}{2}\sum_{j=1}^{d} \mathbb{E}\left[\left|\sum_{i} \text{clip}_2(N_i, Q)_j + Z_j - \sum_{i} N_{i,j}\right| \middle| Z_j < 0\right] \\
=&\frac{1}{2}\left(Q \cdot M + \sum_{i:||N_i||_2 > Q} \left(1 - \frac{Q}{||N_i||_2}\right) ||N_i||_1\right) \\
\geq&\frac{1}{2}\left(C^* \cdot M + \sum_{i:||N_i||_2 > C^*} \left(1 - \frac{C^*}{||N_i||_2}\right) ||N_i||_1\right) = \frac{1}{2}G(C^*).
\end{aligned}$$

This shows that C^* yields a 2-approximation. □

A.2. Proof of Corollary 4.5

To estimate C^* privately, one can use the output perturbation algorithm. For ease of analysis we consider the regularized problem. More precisely, let

$$f_i(C) = \max\{1 - C/\|N_i\|_2, 0\}\|N_i\|_1.$$

Note that f_i is L-Lipschitz where $L = \sqrt{d}$. The goal is to minimize the following function

$$F_1(C) = \frac{1}{n}\sum_{i=1}^{n} f_i(C) + \frac{CM}{n} + \frac{\lambda}{2}C^2. \tag{5}$$

Let $C_1^* = \arg\min_{C\geq 0} F_1(C)$. We first compute the sensitivity of C_1^* as a function of the dataset.

We first compute the sensitivity of C'. Consider a pair of neighboring datasets D and D' which only differ by the nth user.

Lemma A.1. *Let N_n' be a histogram and $f_n'(C)$ defined similarly as $f_n(C)$ with N_n replaced by N_n'. Let $F_1(C) = \frac{1}{n}\sum_{i=1}^{n} f_i(C) + \frac{CM}{n} + \frac{\lambda}{2}C^2$ and $F_2(C) = \frac{1}{n}\sum_{i=1}^{n-1} f_i(C) + \frac{1}{n}f_n'(C) + \frac{CM}{n} + \frac{\lambda}{2}C^2$. Let $C_1^* = \arg\min_{C\geq 0} F_2(C)$ and $C_2^* = \arg\min_{C\geq 0} F_2(C)$. Then,*

$$|C_1^* - C_2^*| \leq \Delta := \frac{4\sqrt{s}}{\lambda n},$$

Proof. Observe that $f_i(C)$ is $\sqrt{s}$ Lipschitz. Let $L = \sqrt{s}$.

$$\begin{aligned}
&n(F_1(C_2^*) - F_1(C_1^*)) \\
&= \sum_{i=1}^{n} f_i(C_2^*) - \sum_{i=1}^{n} f_i(C_1^*) + M(C_2^* - C_1^*) + \frac{n\lambda}{2}((C_2^*)^2 - (C_1^*)^2) \\
&= \sum_{i=1}^{n-1} f_i(C_2^*) - \sum_{i=1}^{n-1} f_i(C_1^*) + M(C_2^* - C_1^*) + \frac{n\lambda}{2}((C_2^*)^2 - (C_1^*)^2) + f_n(C_2^*) - f_n(C_1^*) \\
&= n(F_2(C_2^*) - F_2(C_1^*)) + f_n(C_2^*) - f_n(C_1^*) - (f_n'(C_2^*) - f_n'(C_1^*)) \\
&\leq |f_n(C_2^*) - f_n(C_1^*)| + |f_n'(C_2^*) - f_n'(C_1^*)| \\
&\leq 2L|C_2^* - C_1^*|
\end{aligned}$$

Since F_1 is λ-strongly convex, we have

$$F_1(C_2^*) - F_1(C_1 *) \geq \frac{\lambda}{2}|C_2^* - C_1^*|^2$$

Combining the two parts,

$$|C_2^* - C_1^*| \leq \frac{4L}{\lambda n}$$

□

Now we can characterize performance of the combined algorithm which uses the output of Algorithm 2, $\hat{C}$, as the clipping threshold in Algorithm 1.

Lemma A.2. *Let C_m be an upper bound on C^*. Then*

$$\mathbb{E}[\mathcal{L}_G(\hat{C}, D)] - 2\inf_{C\geq 0}\mathcal{L}_G(C, D) \leq \frac{n\lambda C_m^2}{2} + \frac{4s}{\lambda\varepsilon'}.$$

Proof. Recall the definition of C_1^* from Lemma A.1. First we write the expression,

$$\begin{aligned}
&\mathbb{E}[\mathcal{L}_G(\hat{C}, D)] - 2 \inf_{C>0} \mathcal{L}_G(C, D) \\
&\leq \mathbb{E}[G(\hat{C})] - G(C^*) \\
&\leq \mathbb{E}[M\hat{C}] + \mathbb{E}\left[\sum_{i=1}^{n} f_i(\hat{C})\right] - G(C^*) \\
&= C_1^* M + \mathbb{E}\left[\sum_{i=1}^{n} f_i(\hat{C}) - \sum_{i=1}^{n} f_i(C_1^*)\right] + \sum_{i=1}^{n} f_i(C_1^*) - C^* M - \sum_{i=1}^{n} f_i(C^*).
\end{aligned}$$

The first inequality comes from the proof of Lemma 4.3. We bound the terms separately,

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{n} f_i(\hat{C}) - \sum_{i=1}^{n} f_i(C_1^*)\right] &\leq \mathbb{E}\left|\sum_{i=1}^{n} f_i(\hat{C}) - \sum_{i=1}^{n} f_i(C_1^*)\right| \\
&\leq n\sqrt{s}\mathbb{E}[|\hat{C} - C_1^*|] \\
&\leq n\sqrt{s}\frac{\Delta}{\varepsilon'} = \frac{4\sqrt{s}(\sqrt{s})}{\lambda\varepsilon'}. \qquad (6)
\end{aligned}$$

The remaining terms are bounded using the following fact

$$\sum_{i=1}^{n} f_i(C_1^*) + C_1^* M + n\frac{\lambda}{2}(C_1^*)^2 \leq \sum_{i=1}^{n} f_i(C_*) + C^* M + n\frac{\lambda C_*^2}{2}.$$

Hence,

$$\sum_{i=1}^{n} f_i(C_1^*) + C_1^* M - \sum_{i=1}^{n} f_i(C_*) - C^* M \leq \frac{n\lambda C_*^2}{2} \leq \frac{n\lambda C_m^2}{2}. \qquad (7)$$

Combining equation 6 and 7 yields the desired result. □

The proof of differential privacy follows from Lemma A.1 and the definition of Laplace mechanism. Setting $\lambda = \frac{2\sqrt{2s}}{C_m\sqrt{n\varepsilon'}}$ in Lemma A.2 yields the error.

B. Unbounded domain histograms

Recall that $\widehat{N}$ is the output of Algorithm 3 and the expected error is characterized by

$$\mathbb{E}[\|\widehat{N} - \overline{N}\|_1] = \sum_{j:\overline{N}_j>0} \mathbb{E}\left[\left|\widetilde{N}_j 1_{\widetilde{N}_j>t} - \sum_i N_{ij}\right|\right] \qquad (8)$$

Obviously (8) depends on the choice of the threshold C. The error can be large if C is too small or too large, so the goal of our work is to find the best choice of C.

B.1. Approximation of estimation error

Theorem B.1. *Let $\overline{N}$ be the true aggregate histogram and $\widehat{N}$ be the private estimate obtained by the unbounded-domain algorithm.*

$$\begin{aligned}
\mathbb{E}[\|\widehat{N} - \overline{N}\|_1 \mid \bar{h}_j] = \Theta\Bigg(&\sum_{j:\bar{h}_j>t} \left(\frac{C}{\varepsilon} + e^{-\frac{\varepsilon}{C}(\bar{h}_j - t)}\left(t - \frac{C}{\varepsilon}\right)\right) \\
&+ \sum_{j:\bar{h}_j\leq t} \left(\bar{h}_j + \overline{N}_j + e^{-\frac{\varepsilon}{C}(t-\bar{h}_j)}\left(|\overline{N}_j - t| - \overline{N}_j + \frac{C}{\varepsilon}\right)\right) \\
&+ \sum_{i=1}^{n} \max\{\|N_i\|_1 - C, 0\}\Bigg) \qquad (9)
\end{aligned}$$

Proof. First, we state a fact about exponential distributions.

Lemma B.2. *Let X be an exponential distribution with rate ν (i.e., $X \geq 0$ and $\Pr[X \geq t] = e^{-\nu x}$) and $a \geq 0$, then*

$$\mathbb{E}[X|0 \leq X \leq a] \Pr[0 \leq X \leq a] = \frac{1}{\nu} - e^{-\nu a}\left(a + \frac{1}{\nu}\right).$$

Proof. Recall that $\mathbb{E}[X] = \frac{1}{\nu}$. Due to the memoryless property of X, we have $\mathbb{E}[X|X > a] = a + \frac{1}{\nu}$. Thus,

$$\begin{aligned}
\mathbb{E}[X] &= \mathbb{E}[X|0 \leq X \leq a] \Pr[0 \leq X \leq a] + \mathbb{E}[X|X > a] \Pr[X > a] \\
&= \mathbb{E}[X|0 \leq X \leq a] \Pr[0 \leq X \leq a] + e^{-\nu a}\left(a + \frac{1}{\nu}\right) \\
&= \frac{1}{\nu}.
\end{aligned}$$

Rearranging the terms proves the lemma.

□

We first divide the summation into two parts based on if $Z_j \geq 0$ or $Z_j < 0$. For notational simplicity, all expectations in this section contain an implicit condition on $\bar{h}_j$.

$$\begin{aligned}
&\mathbb{E}[\|\widehat{N} - \overline{N}\|_1 \mid \bar{h}_j] \\
&= \sum_{j:\overline{N}_j > 0} \mathbb{E}\left[\left|\left(\bar{h}_j + Z_j\right) 1_{\bar{h}_j + Z_j > t} - \sum_i N_{ij}\right|\right] \\
&= \frac{1}{2}\left(\sum_{j:\overline{N}_j > 0} \mathbb{E}\left[\left|\left(\bar{h}_j + Z_j\right) 1_{\bar{h}_j + Z_j > t} - \bar{h}_j\right| \middle| Z_j < 0\right] + \sum_{i:||N_i||_1 > C} (||N_i||_1 - C)\right) (**) \\
&+ \frac{1}{2}\left(\sum_{j:\overline{N}_j > 0} \mathbb{E}\left[\left|\left(\bar{h}_j + Z_j\right) 1_{\bar{h}_j + Z_j > t} - \overline{N}_j\right| \middle| Z_j \geq 0\right]\right), (*)
\end{aligned}$$

where to prove $(**)$, we use the fact that if $Z_j < 0$, then $\bar{h}_j + Z_j < \bar{h}_j < \overline{N}_j$ and furthermore

$$\begin{aligned}
\sum_j \overline{N}_j - \bar{h}_j &= \sum_j \sum_i N_{i,j} - h_{i,j} \\
&= \sum_i \sum_j N_{i,j} - h_{i,j} \\
&= \sum_i ||N_i||_1 - ||h_i||_1 \\
&= \sum_{i:||N_i||_1 > C} (||N_i||_1 - C).
\end{aligned}$$

We now bound the middle term. We use the fact that conditioned on $Z_j < 0$, $|Z_j| = -Z_j$ is an exponential distribution with

mean C/ε.

$$
\begin{aligned}
&\sum_{j:\overline{N}_j>0} \mathbb{E}\left[\left|\left(\bar{h}_j + Z_j\right) 1_{\bar{h}_j+Z_j>t} - \bar{h}_j\right| \middle| Z_j < 0\right] \\
&= \sum_{j:\bar{h}_j>t} \mathbb{E}\left[\left|\left(\bar{h}_j + Z_j\right) 1_{\bar{h}_j+Z_j>t} - \bar{h}_j\right| \middle| Z_j < 0\right] + \sum_{j:\bar{h}_j\le t} \mathbb{E}\left[\left|\left(\bar{h}_j + Z_j\right) 1_{\bar{h}_j+Z_j>t} - \bar{h}_j\right| \middle| Z_j < 0\right] \\
&= \sum_{j:\bar{h}_j>t} \mathbb{E}\left[\left|\left(\bar{h}_j + Z_j\right) 1_{\bar{h}_j+Z_j>t} - \bar{h}_j\right| \middle| Z_j < 0\right] + \sum_{j:\bar{h}_j\le t} \bar{h}_j \\
&= \sum_{j:\bar{h}_j>t} \left(\bar{h}_j \Pr[Z_j \le t - \bar{h}_j | Z_j < 0] + \mathbb{E}[|Z_j| \mid t - \bar{h}_j < Z_j < 0] \Pr[Z_j > t - \bar{h}_j | Z_j < 0]\right) + \sum_{j:\bar{h}_j\le t} \bar{h}_j \qquad (10) \\
&= \sum_{j:\bar{h}_j>t} \left(\bar{h}_j e^{-\frac{\varepsilon}{C}(\bar{h}_j - t)} + \frac{C}{\varepsilon} - \left(\bar{h}_j - t + \frac{C}{\varepsilon}\right) e^{-\frac{\varepsilon}{C}(\bar{h}_j - t)}\right) + \sum_{j:\bar{h}_j\le t} \bar{h}_j \\
&= \sum_{j:\bar{h}_j>t} \left(\frac{C}{\varepsilon} + \left(t - \frac{C}{\varepsilon}\right) e^{-\frac{\varepsilon}{C}(\bar{h}_j - t)}\right) + \sum_{j:\bar{h}_j\le t} \bar{h}_j \qquad (11)
\end{aligned}
$$

From (10) we used Lemma B.2.

$$
\begin{aligned}
(*) &= \sum_{j:\overline{N}_j>0} \mathbb{E}\left[\left|\left(\bar{h}_j + Z_j\right) 1_{\bar{h}_j+Z_j>t} - \overline{N}_j\right| \middle| Z_j \ge 0\right] \\
&= \sum_{j:\bar{h}_j\ge t} \mathbb{E}\left[\left|\left(\bar{h}_j + Z_j\right) 1_{\bar{h}_j+Z_j>t} - \overline{N}_j\right| \middle| Z_j \ge 0\right] + \sum_{j:\bar{h}_j< t} \mathbb{E}\left[\left|\left(\bar{h}_j + Z_j\right) 1_{\bar{h}_j+Z_j>t} - \overline{N}_j\right| \middle| Z_j \ge 0\right] \\
&= \sum_{j:\bar{h}_j\ge t} \mathbb{E}\left[\left|\bar{h}_j + Z_j - \overline{N}_j\right| \middle| Z_j \ge 0\right] + \sum_{j:\bar{h}_j< t} \mathbb{E}\left[\left|\left(\bar{h}_j + Z_j\right) 1_{\bar{h}_j+Z_j>t} - \overline{N}_j\right| \middle| Z_j \ge 0\right] \\
&= \sum_{j:\bar{h}_j\ge t} \mathbb{E}\left[\left|\bar{h}_j + Z_j - \overline{N}_j\right| \middle| Z_j \ge 0\right] \\
&\quad + \sum_{j:\bar{h}_j< t} \mathbb{E}\left[\left|\left(\bar{h}_j + Z_j\right) - \overline{N}_j\right| \middle| Z_j \ge t - \bar{h}_j\right] \Pr[Z_j \ge t - \bar{h}_j \mid Z_j \ge 0] + \Pr(Z_j < t - \bar{h}_j \mid Z_j \ge 0)\overline{N}_j \\
&= \sum_{j;\bar{h}_j>t} \left(\overline{N}_j - \bar{h}_j + \frac{C}{\varepsilon}(2e^{-\frac{\varepsilon}{C}(\overline{N}_j - \bar{h}_j)} - 1)\right) \qquad (12) \\
&\quad + \sum_{j:\bar{h}_j< t} \left(\overline{N}_j(1 - e^{-\frac{\varepsilon}{C}(t-\bar{h}_j)}) + e^{-\frac{\varepsilon}{C}(t-\bar{h}_j)}\mathbb{E}[|h_j + Z_j - N_j| | h_j + Z_j > t]\right) \qquad (13) \\
&= \sum_{j:\bar{h}_j>t} \Theta(\overline{N}_j - \bar{h}_j + \frac{C}{\varepsilon}) + \sum_{j:\bar{h}_j<t} \left(\overline{N}_j(1 - e^{-\frac{\varepsilon}{C}(t-\bar{h}_j)}) + e^{-\frac{\varepsilon}{C}(t-\bar{h}_j)}\Theta(|N_j - t| + \frac{C}{\varepsilon})\right) \qquad (14)
\end{aligned}
$$

(12) is due to when $\bar{h}_j > t$,

$$
\begin{aligned}
&\mathbb{E}\left[\left|\bar{h}_j + Z_j - \overline{N}_j\right| \middle| Z_j \ge 0\right] \\
&= \Pr[Z_j > \overline{N}_j - \bar{h}_j \mid Z_j \ge 0]\mathbb{E}\left[\bar{h}_j + Z_j - \overline{N}_j \middle| Z_j > \overline{N}_j - \bar{h}_j\right] \\
&\quad + \Pr[Z_j \le \overline{N}_j - \bar{h}_j \mid Z_j \ge 0]\mathbb{E}\left[\overline{N}_j - \bar{h}_j - Z_j \middle| 0 \le Z_j \le \overline{N}_j - \bar{h}_j\right] \\
&= \frac{C}{\varepsilon}e^{-\frac{\varepsilon}{C}(\overline{N}_j - \bar{h}_j)} + (\overline{N}_j - \bar{h}_j)(1 - e^{-\frac{\varepsilon}{C}(\overline{N}_j - \bar{h}_j)}) - \left(\frac{C}{\varepsilon} - e^{-\frac{\varepsilon}{C}(\overline{N}_j - \bar{h}_j)}\left(\overline{N}_j - \bar{h}_j + \frac{C}{\varepsilon}\right)\right) \\
&= \frac{C}{\varepsilon}\left(2e^{-\frac{\varepsilon}{C}(\overline{N}_j - \bar{h}_j)} - 1\right) + \overline{N}_j - \bar{h}_j
\end{aligned}
$$

The conditional expectation in (13) is evaluated as,

$$\begin{aligned}
&\mathbb{E}[|h_j + Z_j - N_j||h_j + Z_j > t] \\
&= \mathbb{E}[|h_j + Z_j - N_j + t - h_j||Z_j > 0] \\
&= \mathbb{E}[|t + Z_j - N_j||Z_j > 0] = \begin{cases} (\overline{N}_j - t + \frac{C}{\varepsilon}(2e^{-\frac{\varepsilon}{C}(\overline{N}_j - t)} - 1), & t < \overline{N}_j \\ t - N_j + \frac{C}{\varepsilon}, & t \geq \overline{N}_j \end{cases} \\
&= \Theta(|N_j - t| + \frac{C}{\varepsilon}),
\end{aligned}$$

The final equality is due to Lemma B.3. Combining all the parts leads to (14).

Lemma B.3. *Let $a > 0$ be constant. The function $g(x) = x + \frac{1}{a}(2e^{-ax} - 1) = \Theta(x + \frac{1}{a})$ for $x \geq 0$.*

Proof. It is obvious that $g(x) \leq x + \frac{1}{a}$ since $2e^{-ax} - 1 \leq 1$. It remains to prove $g(x) \geq c(x + 1/a)$ for some constant c. Consider the function $h(x) = g(x)/(x + 1/a)$. Then,

$$h'(x) = 2\frac{\frac{1}{a} - e^{-ax}(x + 2/a)}{(x + 1/a)^2}$$

Since the numerator $\phi(x) = \frac{1}{a} - e^{-ax}(x + 2/a)$ is monotonically increasing for $x \geq 0$, and $\phi(0) = -1/a$, $\phi(\infty) = 1/a$, there must be a unique $\xi > 0$ such that $\phi(\xi) = 0$. Thus $h(x)$ is decreasing in $(0, \xi)$ and increasing in (ξ, ∞). The minimum of h is reached when $x = \xi$, with a minimum value of

$$h(\xi) = 1 - \frac{2}{a\xi + 2} > 0$$

Rearranging the terms in equation $\phi(x) = 0$, we easily note that $a\xi$ is the unique solution to the equation $e^{-x}(x+2) - 1 = 0$. Note that

$$e^{-1}(1 + 2) > 1, \quad e^{-2}(2 + 2) < 1$$

Therefore, we must have $a\xi > 1$. Note that $1 - 2/(a\xi + 2)$ increases with ξ, thus,

$$h(\xi) \geq 1 - \frac{2}{1 + 2} = \frac{1}{3}.$$

This implies $g(x) \geq \frac{1}{3}(x + 1/a)$.

□

Now we can combine (11), (14) to prove the theorem.

$$
\begin{aligned}
&\mathbb{E}[\|\widehat{N}-\overline{N}\|_1] \\
&= \sum_{j:\overline{N}_j>0} \mathbb{E}\left[\left|\left(\bar{h}_j+Z_j\right)1_{\bar{h}_j+Z_j>t}-\sum_i N_{ij}\right|\right] \\
&= \frac{1}{2}\sum_{j:\bar{h}_j>t}\left(\frac{C}{\varepsilon}+\left(t-\frac{C}{\varepsilon}\right)e^{-\frac{\varepsilon}{C}(\bar{h}_j-t)}\right)+\frac{1}{2}\sum_{j:\bar{h}_j\le t}\bar{h}_j+\frac{1}{2}\sum_{i:||N_i||_1>C}(||N_i||_1-C) \\
&\quad+\frac{1}{2}\sum_{j:\bar{h}_j>t}\left(\frac{C}{\varepsilon}\left(2e^{-\frac{\varepsilon}{C}(\overline{N}_j-\bar{h}_j)}-1\right)+\overline{N}_j-\bar{h}_j\right)+\frac{1}{2}\sum_{j:\bar{h}_j\le t}\left(\overline{N}_j(1-e^{-\frac{\varepsilon}{C}(t-\bar{h}_j)})+e^{-\frac{\varepsilon}{C}(t-\bar{h}_j)}\Theta(|N_j-t|+\frac{C}{\varepsilon})\right) \\
&= \frac{1}{2}\sum_{j:\bar{h}_j>t}\left(\frac{C}{\varepsilon}+\left(t-\frac{C}{\varepsilon}\right)e^{-\frac{\varepsilon}{C}(\bar{h}_j-t)}+\Theta(\overline{N}_j-\bar{h}_j+\frac{C}{\varepsilon})\right)+\frac{1}{2}\sum_{i:||N_i||_1>C}(||N_i||_1-C) \\
&\quad+\frac{1}{2}\sum_{j:\bar{h}_j\le t}\left(\bar{h}_j+\overline{N}_j(1-e^{-\frac{\varepsilon}{C}(t-\bar{h}_j)})+e^{-\frac{\varepsilon}{C}(t-\bar{h}_j)}\Theta\left(|\overline{N}_j-t|+\frac{C}{\varepsilon}\right)\right)
\end{aligned}
$$

From Lemma B.3, the Θ expressions are upper bounded by a factor of 1 and lower bounded by a factor of 1/3. Therefore,

$$
\begin{aligned}
&\frac{1}{2}\sum_{i=1}^{n}\max\{\|N_i\|_1-C,0\}+\frac{1}{2}\sum_{j:\bar{h}_j>t}\left(\frac{C}{\varepsilon}+e^{-\frac{\varepsilon}{C}(\bar{h}_j-t)}\left(t-\frac{C}{\varepsilon}\right)\right) \\
&\quad+\frac{1}{6}\sum_{j:\bar{h}_j\le t}\left(\bar{h}_j+\overline{N}_j+e^{-\frac{\varepsilon}{C}(t-\bar{h}_j)}\left(|\overline{N}_j-t|-\overline{N}_j+\frac{C}{\varepsilon}\right)\right) \\
&\le \mathbb{E}[\|\widehat{N}-\overline{N}\|_1] \le \sum_{i=1}^{n}\max\{\|N_i\|_1-C,0\}+\sum_{j:\bar{h}_j>t}\left(\frac{C}{\varepsilon}+e^{-\frac{\varepsilon}{C}(\bar{h}_j-t)}\left(t-\frac{C}{\varepsilon}\right)\right) \\
&\quad+\frac{1}{2}\sum_{j:\bar{h}_j\le t}\left(\bar{h}_j+\overline{N}_j+e^{-\frac{\varepsilon}{C}(t-\bar{h}_j)}\left(|\overline{N}_j-t|-\overline{N}_j+\frac{C}{\varepsilon}\right)\right)
\end{aligned}
$$

□

Corollary B.4. *Assume that $\delta \le 1/8$, $C \ge 1$ and $C \le e^{\varepsilon/3\delta}$. Then*

$$
\begin{aligned}
&\frac{1}{2}\sum_{i=1}^{n}\max\{\|N_i\|_1-C,0\}+\frac{1}{12}\sum_{j:\bar{h}_j<t}\overline{N}_j+\frac{1}{2}\sum_{j:\bar{h}_j>t}\frac{C}{\epsilon} \\
&\le \mathbb{E}[\|\widehat{N}-\overline{N}\|_1 \mid \bar{h}_j] \le \sum_{i=1}^{n}\max\{\|N_i\|_1-C,0\}+\sum_{j:\bar{h}_j<t}\overline{N}_j+\sum_{j:\bar{h}_j>t}t
\end{aligned}
$$

Proof. For terms with $\bar{h}_j > t$,

$$
\frac{C}{\epsilon} \le \frac{C}{\varepsilon}+e^{-\frac{\varepsilon}{C}(\bar{h}_j-t)}\left(t-\frac{C}{\varepsilon}\right) \le t = \frac{C}{\epsilon}\log\frac{C}{2\delta}.
$$

For the terms with $\bar{h}_j \le t$,

$$
\frac{\overline{N}_j}{2} \le W := \bar{h}_j+\overline{N}_j+e^{-\frac{\varepsilon}{C}(t-\bar{h}_j)}\left(|\overline{N}_j-t|-\overline{N}_j+\frac{C}{\varepsilon}\right) \le 2\overline{N}_j.
$$

To prove this, we consider $\overline{N}_j \ge t$ and $\overline{N}_j < t$ separately.

1. $\overline{N}_j \geq t$ Since $t \geq C/\varepsilon$ and $0 \leq \bar{h}_j \leq \overline{N}_j$,

$$W = \bar{h}_j + \overline{N}_j + e^{-\frac{\varepsilon}{C}(t-\bar{h}_j)}\left(\frac{C}{\varepsilon} - t\right) \leq 2\overline{N}_j.$$

W is concave with respect to $\bar{h}_j$, thus the minimum of W must occur at either $\bar{h}_j = 0$ or $\bar{h}_j = t$. When $\bar{h}_j = 0$,

$$W = \overline{N}_j + \frac{2\delta}{e^{\varepsilon}}\left(t - \frac{C}{\varepsilon}\right) \geq \frac{\overline{N}_j}{2}. \tag{15}$$

as long as $\overline{N}_j \geq \frac{4\delta}{Ce^{\varepsilon}}(t - \frac{C}{\varepsilon})$. Since $\delta \leq 1/4$ and $C \geq 1$, we must have $\overline{N}_j \geq t \geq \frac{4\delta}{Ce^{\varepsilon}}(t - \frac{C}{\varepsilon})$, so the condition is satisfied. When $\bar{h}_j = t$,

$$\overline{N}_j \leq W = \overline{N}_j + \frac{C}{\varepsilon} \leq 2\overline{N}_j.$$

The final inequality is due to $C/\varepsilon \leq t \leq \overline{N}_j$.

2. $\overline{N}_j < t$

$$W = \bar{h}_j + \overline{N}_j + e^{-\frac{\varepsilon}{C}(t-\bar{h}_j)}\left(t - 2\overline{N}_j + \frac{C}{\varepsilon}\right)$$

If $\overline{N}_j > (t + C/\varepsilon)/2$, then

$$\frac{\overline{N}_j}{2} \leq \bar{h}_j + \overline{N}_j + e^{-\frac{\varepsilon}{C}(t-\bar{h}_j)}\left(\frac{C}{\varepsilon} - t\right) \leq W \leq \bar{h}_j + \overline{N}_j \leq 2\overline{N}_j$$

The first inequality is proved similarly as (15).

If $\overline{N}_j \leq (t + C/\varepsilon)/2$, then

$$W \geq \bar{h}_j + \overline{N}_j \geq \overline{N}_j.$$

It remains to prove that $W \leq (1 + \log \frac{C}{2\delta})\overline{N}_j$. We consider the following function

$$f(x) = x + \beta e^{ax}(\gamma - x), \quad a = \frac{\varepsilon}{C}, \gamma = \frac{1}{2}\left(t + \frac{C}{\varepsilon}\right), \beta = \frac{2\delta}{Ce^{\varepsilon}}.$$

Note that since $\bar{h}_j \leq \overline{N}_j$, we have $W \leq 2f(\overline{N}_j)$. We just need to upper bound $g(x) = f(x)/x$ when $x \in [1, \gamma]$. Taking the derivative,

$$g'(x) = -\beta e^{ax}\frac{ax^2 - a\gamma x + \gamma}{x^2}.$$

When $a\gamma \leq 4$, $g'(x) \leq 0$, thus

$$g(x) \leq g(1) = 1 + \frac{\delta}{e^{\varepsilon(1-1/C)}}\left(\frac{1}{\varepsilon}\left(1 + \log\frac{C}{2\delta}\right) + 1 - \frac{1}{C}\right),$$

which is at most 2 given the assumption that $C \leq e^{\varepsilon/3\delta}$.

When $a\gamma > 4$, then consider the root of $g'(x) = 0$,

$$x_1 = \frac{\gamma}{2}\left(1 - \sqrt{1 - \frac{4}{\alpha\gamma}}\right), x_2 = \frac{\gamma}{2}\left(1 + \sqrt{1 - \frac{4}{\alpha\gamma}}\right).$$

x_1 is a local minimum, and x_2 is a local maximum of $g(x)$. Thus the maximum of $g(x)$ on $[1, \gamma]$ must be either $x = 1$ or $x = x_2$. Note that $x_1 < \gamma/2$ and $x_2 < \gamma$. We evaluate and upper bound $g(x_2)$,

$$\begin{aligned}
g(x_2) &= 1 + \beta e^{ax_2}(ax_1 - 1) \\
&\leq 1 + \beta e^{a\gamma}\frac{a\gamma}{2} \\
&= 1 + \frac{2\delta}{Ce^{\varepsilon}}\frac{1}{4}\left(1 + \varepsilon + \log\frac{C}{2\delta}\right)\sqrt{e\frac{e^{\varepsilon}C}{2\delta}} \\
&= 1 + \sqrt{\frac{2\delta}{Ce^{\varepsilon}}}\frac{1}{4}\left(1 + \varepsilon + \log\frac{C}{2\delta}\right) \leq 2
\end{aligned}$$

Combining with the upper bound for $g(1)$ completes the proof. □

We simplify the above proof further to prove Theorem 5.2.

Corollary B.5. *Let C_m be an upper bound on C.*

$$\frac{1}{2}\sum_{i=1}^{n}\max\{\|N_i\|_1 - C, 0\} + \frac{1}{12\log\frac{C_m}{2\delta}}\sum_j \min(h_j, t)$$
$$\leq \mathbb{E}[\|\widehat{N} - \overline{N}\|_1 \mid \bar{h}_j] \leq$$
$$2\sum_{i=1}^{n}\max\{\|N_i\|_1 - C, 0\} + \sum_j \min(\bar{h}_j, t).$$

Proof. First observe that

$$\begin{aligned}\sum_{j:\bar{h}_j<t}\overline{N}_j + \sum_{j:\bar{h}_j>t} t &= \sum_{j:\bar{h}_j<t}\bar{h}_j + \sum_{j:\bar{h}_j>t} t + \sum_{j:\bar{h}_j<t}(\overline{N}_j - \bar{h}_j)\\ &\leq \sum_{j:\bar{h}_j<t}\bar{h}_j + \sum_{j:\bar{h}_j>t} t + \sum_{j:\overline{N}_j>0}(\overline{N}_j - \bar{h}_j)\\ &= \sum_{j:\bar{h}_j<t}\bar{h}_j + \sum_{j:\bar{h}_j>t} t + \sum_i \max(\|N_i\|_1 - C, 0)\\ &= \sum_j \min(\bar{h}_j, t) + \sum_i \max(\|N_i\|_1 - C, 0)\end{aligned}$$

Similarly, for the lower bound, observe that

$$\begin{aligned}\frac{1}{12}\sum_{j:\bar{h}_j<t}\overline{N}_j + \frac{1}{2}\sum_{j:\bar{h}_j>t}\frac{C}{\epsilon} &\geq \frac{1}{12}\sum_{j:\bar{h}_j<t}\overline{N}_j + \sum_{j:\bar{h}_j>t}\frac{t}{2\log\frac{C_m}{2\delta}}\\ &\geq \frac{1}{12\log\frac{C_m}{2\delta}}\sum_j \min(\bar{h}_j, t).\end{aligned}$$

□

C. Bias reduction

We first note that in many datasets, counts of most symbols appear very few times. For example, in the Sentiment140 dataset, which contains counts for a total of roughly $6\cdot 10^5$ words distributed across $6\cdot 10^5$ users, the average counts of all words among the users are no more than two. Therefore we analyze the debiasing step when the λ_i's are small and prove the following result for our desbiasing algorithm given in Algorithm 5.

Theorem C.1. *Suppose $N_i \sim Poi(\lambda_i)$. Let $\bar{\lambda} = \frac{1}{n}\sum_i \lambda_i$, $\Sigma = \frac{1}{n}\sum_{i=1}^n(\lambda_i - \bar{\lambda})^2$ and $\hat{\lambda} = \min\{1, \max\{0, \widehat{N}/n\}\}$, where $\widehat{N}$ is the output of Algorithm 5. If $\bar{\lambda} \leq 1$, then*

$$\mathbb{E}[(\bar{\lambda} - \hat{\lambda})^2] \leq \gamma_C^2\left(\frac{C^2}{n^2\varepsilon^2} + \frac{\bar{\lambda}}{n} + \min\left\{1, \frac{1}{8\pi(C-1)}\right\}\Sigma^2\right). \tag{16}$$

where $\gamma_C = \Pr[Poi(1) < C]^{-1} \leq \max\left\{e, \frac{1}{1-e^{-(C-1)^2/2C}}\right\}$. If we further assume that $\lambda_i \leq 1$, then

$$\mathbb{E}[(\bar{\lambda} - \hat{\lambda})^2] \leq \gamma_C^2\left(\frac{C^2}{n^2\varepsilon^2} + \frac{\bar{\lambda}}{n} + \frac{1}{4((C-1)!)^2}\cdot\Sigma^2\right). \tag{17}$$

The error consists of three terms. The first term is the error due to added noise, which is proportional to the clipping threshold C. The second term is essentially the variance of the random variable $\frac{1}{n}\sum_i N_i$, which is an inherent error due

to the randomness in the counts N_i's. The third term is a bias term which depends on the closeness of user distributions, characterized by Σ, the variance of λ_i's. If the users' distributions are similar, then we can expect the estimation error to be small. Note that the bias term has a $1/C$ rate. The detailed proof is provided in Appendix C.1 We provide a bound with a better dependence on C in Appendix C.2.

Next we analyze the the optimal choice of threshold C after the debiasing step. Observe that to minimize the error upper bound in (16), roughly we want to choose

$$C \propto (n\Sigma)^{2/3} + 1.$$

Therefore, when user's distribution are similar, we can use a smaller clipping threshold. This implies that as long as

$$\Sigma = O(n^{-1/4}),$$

we can find a C that ensures a squared error of $O(1/n)$, which matches the error for the i.i.d. case.

From (17), when $\lambda_i \leq 1$ for any user i, we can choose

$$C \propto 1 + \log(1 + n\Sigma).$$

This choice of C always guarantees $O(1/n)$ error since when $\lambda_i \leq 1$ for all i, $\Sigma \leq 1$.

In practice, we can also privately choose C as the top $\lceil 1/\varepsilon \rceil$ count as suggested by Lemma 4.2.

So far we have characterized the effect of debiasing under heterogeneous data. We next show that under mild assumptions debiasing helps even if data is non i.i.d.. The formal result is stated in Theorem C.2. The proof is in Appendix C.3.

Theorem C.2. *Let $\bar{h} = \frac{1}{n}\sum_{i=1}^{n} h(\lambda_i)$. Let $\hat{\lambda}_L = \widehat{N}_L/n$ be the average count obtained by Algorithm 1 with Laplace noise. Assume that $\bar{h} \geq h_{\min} := h(\bar{\lambda}) - \frac{\bar{\lambda} - h(\bar{\lambda})}{\gamma_C - 1}$. Write $\bar{h} = h_{\min} + \alpha\frac{\bar{\lambda} - h(\bar{\lambda})}{\gamma_C - 1}$ where $\alpha \in (0, 1]$. Then the gap between Algorithm 1 and Algorithm 5 is*

$$\begin{aligned} &\mathbb{E}[(\bar{\lambda} - \hat{\lambda}_L)^2] - \mathbb{E}[(\bar{\lambda} - \hat{\lambda})^2] \\ &\geq \frac{\alpha(2\gamma_C - (\gamma_C + 1)\alpha)}{\gamma_C - 1}(\bar{\lambda} - h(\bar{\lambda}))^2 - O_C\left(\frac{1}{n}\right). \end{aligned} \tag{18}$$

This implies that for any fixed C, under the assumptions stated in the theorem, with n sufficiently large, there is always a constant gap between the two algorithms and debiasing helps even if the data is not i.i.d.. This result justifies the choice of C as the optimal quantile suggested by Lemma 4.2.

We argue that the assumption of $\bar{h} \geq h_{\min}$ is not too restrictive. It essentially requires that either λ_i's are sufficiently similar, or C is sufficiently larger than $\bar{\lambda}$. Indeed, if $\lambda_i = \bar{\lambda}$ for all user i, then $h(\bar{\lambda}) = \bar{h}$; if C is sufficiently large, then h is almost linear near $\bar{\lambda}$ and hence $\bar{h}$ is close to $h(\bar{\lambda})$.

As a specific example, set $C = 2, \bar{\lambda} = 1$. If all $\lambda_i \in [0, 2]$, due to concavity of h, we have $\bar{h} \geq h(2)/2 \geq 0.729$. With some arithmetic, $h_{\min} \leq 0.61$, and the first term in (18) is at least 0.0217. Note that this is the difference between squared errors. The gap between absolute errors could well be of order 0.1, which is significant considering that $\bar{\lambda} = 1$. This example shows that Algorithm 5 can achieve significant improvement even when the variance of λ_is is constant.

C.1. Proof of Theorem C.1

Proof. Let $Y = \frac{1}{n}Y_i$. Then, when $N_i \sim \text{Poi}(\lambda_i)$,

$$\begin{aligned}
\mathbb{E}[Y] &= \frac{1}{n}\sum_{i=1}^{n} h(\lambda_i) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=0}^{C-1} j\cdot \Pr[N_i=j] + \sum_{j=C}^{\infty} C\cdot \Pr[N_i=j]\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(C - \sum_{j=0}^{C-1}(C-j)\Pr[N_i=j]\right) \\
&= C - \frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{C-1}(C-j)\, e^{-\lambda_i}\frac{\lambda_i^j}{j!}
\end{aligned}$$

Let $Z = \text{Lap}(C/n\varepsilon)$, then we have $\hat{\lambda} = h^{-1}(Y+Z)$. We first bound the error for estimating $h(\lambda)$,

$$\begin{aligned}
\mathbb{E}\left[(h(\bar{\lambda}) - h(\hat{\lambda}))^2\right] &= \mathbb{E}\left[(h(\bar{\lambda}) - \mathbb{E}[Y] + \mathbb{E}[Y] - h(\hat{\lambda}))^2\right] \\
&= \mathbb{E}[(h(\hat{\lambda}) - \mathbb{E}[Y])^2] + (\mathbb{E}[Y] - h(\bar{\lambda}))^2 \\
&= \mathbb{E}[Z^2] + \mathbb{E}[(Y-\mathbb{E}Y)^2] + (\mathbb{E}[Y] - h(\bar{\lambda}))^2 \\
&\le \frac{C^2}{n^2\varepsilon^2} + \mathbb{E}[(Y-\mathbb{E}Y)^2] + (\mathbb{E}[Y] - h(\bar{\lambda}))^2. \qquad (19)
\end{aligned}$$

We first bound $\mathbb{E}[(Y-\mathbb{E}[Y])^2]$.

$$\begin{aligned}
\mathbb{E}[(Y-\mathbb{E}[Y])^2] &= \frac{1}{n^2}\mathbb{E}\left[\left(\sum_{i=1}^{n}(Y_i - \mathbb{E}[Y_i])\right)^2\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}[Y_i] \qquad (20) \\
&\le \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}[N_i] = \frac{1}{n^2}\sum_{i=1}^{n}\lambda_i.
\end{aligned}$$

The final inequality is due to $\text{Var}[Y_i] \le \text{Var}[N_i]$. To see this we use the symmetrization trick. Let N_i' be an independent copy of N_i and $Y_i' = \text{clip}(N_i', C)$. Then by definition $|Y_i - Y_i'| \le |N_i - N_i'|$. Hence,

$$\text{Var}[Y_i] = \mathbb{E}[(Y_i - Y_i')^2/2] \le \mathbb{E}[(N_i - N_i')^2/2] = \text{Var}[N_i].$$

We then bound $|\mathbb{E}[Y] - h(\bar{\lambda})|$. We compute Taylor expansion of $\mathbb{E}[Y]$ at $\bar{\lambda}$ with the Lagrangian remainder,

$$\begin{aligned}
\mathbb{E}[Y] &= \frac{1}{n}\sum_{i=1}^{n}\left(h(\bar{\lambda}) + h'(\bar{\lambda})(\lambda_i - \bar{\lambda}) + \frac{h''(\xi_i)}{2}(\bar{\lambda}-\lambda_i)^2\right) \\
&= h(\bar{\lambda}) + \frac{1}{n}\sum_{i=1}^{n}\frac{h''(\xi_i)}{2}(\bar{\lambda}-\lambda_i)^2,
\end{aligned}$$

where $\xi_i \in (\min\{\bar{\lambda},\lambda_i\}, \max\{\bar{\lambda},\lambda_i\})$. We move on to compute $h''(\lambda)$,

$$\begin{aligned}
h''(\lambda) &= e^{-\lambda}\left(-\sum_{j=0}^{C-1}\frac{C-j}{j!}\lambda^j + 2\sum_{j=0}^{C-2}\frac{C-(j+1)}{j!}\lambda^j - \sum_{j=0}^{C-3}\frac{C-(j+2)}{j!}\lambda^j\right) \\
&= -e^{-\lambda}\frac{\lambda^{C-1}}{(C-1)!}
\end{aligned}$$

We can verify that for $C \geq 1$, $|h''(\lambda)|$ increases when $\lambda \leq C-1$ and decreases when $\lambda \geq C-1$. Therefore by Sterling's approximation,

$$|h''(\xi_i))| \leq \frac{(C-1)^{C-1}}{e^{C-1}(C-1)!} \leq \frac{1}{\sqrt{2\pi(C-1)}}.$$

Thus,

$$|\mathbb{E}[Y] - h(\bar{\lambda})| \leq \frac{1}{2\sqrt{2\pi(C-1)}} \cdot \frac{1}{n}\sum_{i=1}^{n}(\lambda_i - \bar{\lambda})^2.$$

Combining all the parts we have

$$\mathbb{E}[|h(\hat{\lambda}) - h(\bar{\lambda})|] \leq \frac{C}{n\varepsilon} + \frac{1}{n}\sqrt{\sum_{i=1}^{n}\lambda_i} + \frac{1}{2\sqrt{2\pi(C-1)}} \cdot \frac{1}{n}\sum_{i=1}^{n}(\lambda_i - \bar{\lambda})^2.$$

To bound the estimation error $\mathbb{E}[|\bar{\lambda} - \hat{\lambda}|]$, we first compute $h'(\lambda)$,

$$\begin{aligned} h'(\lambda) &= -e^{-\lambda}\left(\sum_{j=0}^{C-2}\frac{C-(j+1)}{j!}\lambda^j - \sum_{j=0}^{C-1}\frac{C-j}{j!}\lambda^j\right) \\ &= e^{-\lambda}\sum_{j=0}^{C-1}\frac{\lambda^j}{j!} \end{aligned}$$

Let $X \sim \text{Poi}(1)$. We note that for $\lambda \in [0,1]$,

$$|h'(\lambda)| \geq |h'(1)| = 1 - \Pr[X \geq C] = \frac{1}{\gamma_C} \geq 1 - e^{-\frac{(C-1)^2}{2C}}.$$

Hence, we can proceed to bound the error for estimating $\bar{\lambda}$,

$$\begin{aligned} &\mathbb{E}[(\bar{\lambda} - \hat{\lambda})^2] \\ &\leq \sup_{\lambda\in[0,1]}\frac{1}{|h'(\lambda)|^2}\mathbb{E}[(h(\bar{\lambda}) - h(\hat{\lambda}))^2] \qquad (21) \\ &\leq \gamma_C^2\left(\frac{C^2}{n^2\varepsilon^2} + \frac{1}{n^2}\sum_{i=1}^{n}\lambda_i + \frac{1}{8\pi(C-1)} \cdot \left(\frac{1}{n}\sum_{i=1}^{n}(\lambda_i - \bar{\lambda})^2\right)^2\right). \qquad (22) \end{aligned}$$

If we assume that $\lambda_i \leq 1$ for all i we can bound $|h''(\xi_i)|$ as

$$|h''(\xi_i))| \leq \max_{\lambda\in[0,1]}|h''(\lambda)| \leq \frac{1}{(C-1)!}.$$

Hence the last term can be bounded as

$$|\mathbb{E}[Y] - h(\lambda)| \leq \frac{1}{2(C-1)!} \cdot \frac{1}{n}\sum_{i=1}^{n}(\lambda_i - \lambda)^2.$$

□

C.2. Improved bound of Theorem C.1

Theorem C.3. *Suppose λ_i are i.i.d. drawn from a distribution F with mean $\bar{\lambda}$ and variance Σ. Then for $C \geq 3$, Algorithm 5 yields an error of*

$$\mathbb{E}[(\bar{\lambda} - \hat{\lambda})^2] \leq \gamma_C^2 \left(\frac{C^2}{n^2\varepsilon^2} + \frac{\bar{\lambda}}{n} + \frac{1}{8\pi(C-1)} \left(0.83^{C-1}\Sigma + \Pr_{\lambda \sim F}\left(\lambda \geq \frac{C-1}{2} \right) \right)^2 \right)$$

The theorem shows that if the distribution of user's average counts are concentrated and has small tail probability, then we can obtain small estimation error.

Suppose that $C \geq 3$. Let $\eta = 1/2$. We bound $|\mathbb{E}[Y] - h(\bar{\lambda})|$ by analyzing λ_i close to $C-1$ and far from $C-1$ separately.

$$\begin{aligned}
&|\mathbb{E}[Y] - h(\bar{\lambda})| \\
&= \frac{1}{n}\sum_{i=1}^{n} \frac{|h''(\xi_i)|}{2}(\bar{\lambda} - \lambda_i)^2 \\
&= \frac{1}{n} \sum_{i:\, \frac{\lambda_i}{C-1} \in (\eta, 1/\eta)} \frac{|h''(\xi_i)|}{2}(\bar{\lambda} - \lambda_i)^2 + \frac{1}{n} \sum_{i:\, \frac{\lambda_i}{C-1} \notin (\eta, 1/\eta)} \frac{|h''(\xi_i)|}{2}(\bar{\lambda} - \lambda_i)^2
\end{aligned}$$

If $\lambda_i \leq \eta(C-1)$ and $C \geq 3$, then we also have $\xi_i \leq \eta(C-1)$. In this case,

$$\begin{aligned}
|h''(\xi_i)| &\leq |h''(\eta(C-1))| = \frac{(\eta(C-1))^{C-1}}{e^{\eta(C-1)}(C-1)!} \\
&\leq \frac{1}{\sqrt{2\pi(C-1)}} \left(\frac{\eta e}{e^{\eta}} \right)^{C-1} \leq \frac{1}{\sqrt{2\pi(C-1)}} \cdot 0.83^{C-1}
\end{aligned}$$

The last inequality follows by ex/e^x is increasing for $x \in [0,1]$. Using a similar argument, we can verify that the above inequality also holds when $\lambda_i \geq (C-1)/\eta$. Therefore,

$$|\mathbb{E}[Y] - h(\bar{\lambda})| \leq \frac{1}{2\sqrt{2\pi(C-1)}} \cdot \frac{1}{n} \left(\sum_{i:\, \frac{\lambda_i}{C-1} \notin (1/2,2)} 0.83^{C-1}(\lambda_i - \bar{\lambda})^2 + \sum_{i:\, \frac{\lambda_i}{C-1} \in (1/2,2)} (\lambda_i - \bar{\lambda})^2 \right).$$

For convenience let $\Sigma = \frac{1}{n}\sum_{i=1}^{n}(\lambda_i - \bar{\lambda})^2$. We can further bound the bias by

$$|\mathbb{E}[Y] - h(\bar{\lambda})| \leq 0.83^{C-1}\Sigma + \frac{\sqrt{2}(C-1)^{3/2}}{\sqrt{\pi}} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\lambda_i > (C-1)/2}. \tag{23}$$

We can see that the bias term depends on C, Σ, and $\sum_{i=1}^{n} \mathbf{1}_{\lambda_i > 2(C-1)}$ (which depends on the distribution of $\{\lambda_i\}_{i=1}^{n}$). To ensure an $O(1/n)$ rate, the two terms should be at most $O(1/\sqrt{n})$,

$$0.83^{C-1}\Sigma \leq \frac{1}{\sqrt{n}} \quad \frac{4(C-1)^{3/2}}{2\sqrt{2\pi}} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\lambda_i > (C-1)/2} \leq \frac{1}{\sqrt{n}}.$$

Note that in the worst case

$$\frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\lambda_i > (C-1)/2} \leq O\left(\frac{\Sigma}{(C-1)^2} \right),$$

which recovers (16). The bound could have a better dependence on C if λ_is are more concentrated. For example, if $\lambda_i \leq 1$, then the above quantity is 0 as long as $C > 3$, and we can choose $C = 3 + O(\log(1 + n\Sigma))$ to achieve $O(1/n)$ mean squared error.

More generally, if λ_is are from a distribution with exponential tail, i.e. $\Pr[\lambda_i \geq x] = O(\exp(-\Omega(x)))$, then

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\lambda_i > 2(C-1)} \simeq \Pr[\lambda_i \geq (C-1)/2] = O(\exp(-\Omega(C-1))).$$

Choosing $C = 3 + O(\log n)$ gives $O(1/n)$ mean squared error.

C.3. Proof of Theorem C.2

Before we proceed to the proof, we first characterize the error of Algorithm 1 with Laplace noise.

Lemma C.4. *Let $\hat{\lambda}_L = \widehat{N}_L/n$ where $\widehat{N}_L$ is the output of Algorithm 1 with Laplace noise. Then*

$$\mathbb{E}\left[(\bar{\lambda} - \hat{\lambda}_L)^2\right] = \left(\bar{\lambda} - \frac{1}{n}\sum_{i=1}^{n} h(\lambda_i)\right)^2 + \frac{1}{n^2}\sum_{i=1}^{n} \mathit{Var}(Y_i) + \frac{C^2}{n^2\varepsilon^2}.$$

Proof.

$$\begin{aligned}
\mathbb{E}\left[(\bar{\lambda} - \hat{\lambda}_L)^2\right] &= \mathbb{E}\left[(\bar{\lambda} - \mathbb{E}[\hat{\lambda}_L] + \mathbb{E}[\hat{\lambda}_L] - \hat{\lambda}_L)^2\right] \\
&= \mathbb{E}\left[(\bar{\lambda} - \mathbb{E}[\hat{\lambda}_L])^2\right] + \mathbb{E}\left[(\mathbb{E}[\hat{\lambda}_L] - \hat{\lambda}_L)^2\right] \\
&= \left(\bar{\lambda} - \frac{1}{n}\sum_{i=1}^{n} h(\lambda_i)\right)^2 + \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(Y_i) + \frac{C^2}{n^2\varepsilon^2}.
\end{aligned}$$

□

Now we have all the ingredients to complete Theorem C.2

Proof. Combining (19), (20), and (21) we have

$$\mathbb{E}[(\bar{\lambda} - \hat{\lambda})^2] \leq \gamma_C^2 \left(\frac{C^2}{n^2\varepsilon^2} + \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}[Y_i] + \left(h(\bar{\lambda}) - \frac{1}{n}\sum_{i=1}^{n} h(\lambda_i)\right)^2\right).$$

Let $\bar{h} = -\frac{1}{n}\sum_{i=1}^{n} h(\lambda_i)$.Therefore,

$$\begin{aligned}
&\mathbb{E}[(\bar{\lambda} - \hat{\lambda}_L)^2] - \mathbb{E}[(\bar{\lambda} - \hat{\lambda})^2] \\
&= -(\gamma_C^2 - 1)\left(\frac{C^2}{n^2\varepsilon^2} + \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}[Y_i]\right) + \left(\bar{\lambda} - \bar{h}\right)^2 - \gamma_C^2\left(h(\bar{\lambda}) - \bar{h}\right)^2
\end{aligned}$$

We bound the terms separately. Note that $\mathrm{Var}[Y_i] \leq \mathrm{Var}[N_i] = \lambda_i$. Hence,

$$(\gamma_C^2 - 1)\left(\frac{C^2}{n^2\varepsilon^2} + \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}[Y_i]\right) \leq (\gamma_C^2 - 1)\left(\frac{C^2}{n^2\varepsilon^2} + \frac{1}{n}\right).$$

Recall the assumption that $\bar{h} \geq h_{\min} := h(\bar{\lambda}) - \frac{\bar{\lambda} - h(\bar{\lambda})}{\gamma_C - 1}$, and write $\bar{h} = h_{\min} + \alpha\frac{\bar{\lambda} - h(\bar{\lambda})}{\gamma_C - 1}$. The remaining part simplifies to

$$\begin{aligned}
(\bar{\lambda} - \bar{h})^2 - \gamma_C^2(h(\bar{\lambda}) - \bar{h})^2 &= \frac{(\gamma_C - \alpha)^2 - \gamma_C^2(1-\alpha)^2}{(\gamma_C - 1)^2}(\bar{\lambda} - h(\bar{\lambda}))^2 \\
&= \frac{\alpha(2\gamma_C - (\gamma_C + 1)\alpha)}{\gamma_C - 1}(\bar{\lambda} - h(\bar{\lambda}))^2
\end{aligned}$$

Combining the two parts completes the proof. □

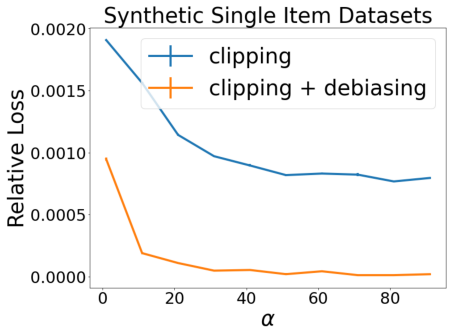

Figure 4. Total counts estimation for single item on synthetic Poisson datasets. Larger α means that user data distributions are more similar.

Table 2. Single word in Sentiment140. The "debiasing" columns are results of clipping + debiasing.

word	clipping avg loss±std	debiasing avg loss±std
the	0.0289 $\pm 2.40 \cdot 10^{-5}$	0.0257 $\pm 2.48 \cdot 10^{-5}$
today	0.0381 $\pm 4.34 \cdot 10^{-5}$	0.0155 $\pm 3.34 \cdot 10^{-5}$
you	0.0745 $\pm 2.83 \cdot 10^{-5}$	0.0671 $\pm 7.26 \cdot 10^{-5}$

C.4. Experiments

In this section, we run experiments for Algorithm 5 with both synthetic datasets and words in Sentiment140. For the synthetic part, we generate $n = 10^6$ users with $\lambda_1, \ldots, \lambda_n$ from a Dirichlet distribution with parameter α. Larger α means that the λ_is are closer. In this experiment, we set C to the privately estimated top $1/\varepsilon$ count among users, as discussed in Amin et al. (2019). Figure 4 shows that the debiased output of Algorithm 5 greatly reduces the error compared to the original output of Algorithm 1, especially for large α (meaning the dataset is more i.i.d. like).

We also run experiments on three population words in Sentiment140: "the", "today" and "you". Table 2 shows that Algorithm 5 performs better than Algorithm 1, but the gain is not as much as synthetic datasets that are close to i.i.d. distributions.

C.5. Extension to $d > 1$

We now discuss two possible extensions to the general d.

1. A natural extension to the entire histogram is to apply Algorithm 5 to each symbol in the histogram separately. To ensure (ε, δ) differential privacy, we assign each symbol a privacy budget of $O(\varepsilon/\sqrt{d \log(1/\delta)})$ by strong composition (Kairouz et al., 2017). The main disadvantage is that when d is large, clipping each coordinate separately may perform poorly compared to clipping the ℓ_1 or ℓ_2 norm of the entire histogram.

2. We can generalize Algorithm 5 to $d > 1$ by replacing 1-d clipping with the high dimensional clipping functions as defined in Algorithm 1. Then choose a suitable function g that essentially inverts the expectation of the clipped vector Y_i. However finding such inverse may be difficult in high dimensions as it likely involves non-convex optimization.

D. Additional experiments

D.1. Algorithm 2 with fixed $C_m = 150$

Fixing C_m performs better than $C_m = DP\text{-}M\text{-}quantile$, because in the latter case, we need to split half of the privacy budget to estimate $DP\text{-}M\text{-}quantile$. We note that the output perturbation method has a large variance for the SNAP dataset and investigating the reason behind is an interesting future direction.

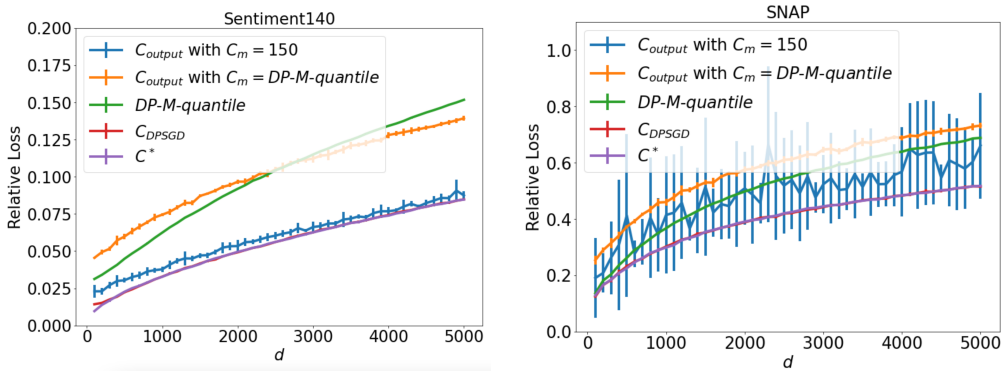

Figure 5. Histogram estimation over bounded domains. Left: Sentiment140 dataset. Right: SNAP dataset.

D.2. Setting $s = d$

In this section, we demonstrate the performance of the bounded domain algorithm in Section 4 when $s = d$, i.e., bound on sparsity is not known. We can see that setting $s = d$ still yields a performance close to the true 2-approximation threshold, and much better than the quantile suggested by (Amin et al., 2019).

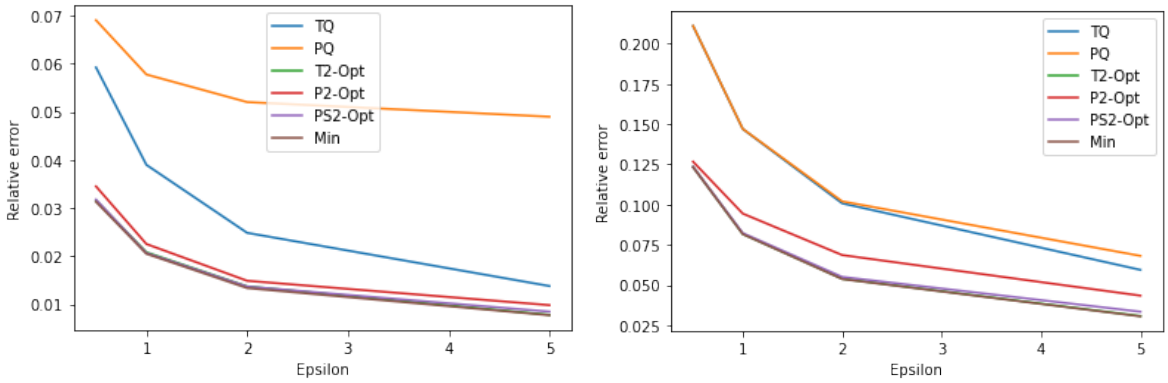

Figure 6. Bounded domain histogram estimation with $d = 1000$ (left) and $d = 10000$ (right).
TQ is the true (non-private) quantile suggested by (Amin et al., 2019). **PQ** is the private estimate of the quantile. **P2-Opt** is the result of DP-SGD with $s = d$. **PS2-Opt** is DP-SGD with $s = 0.1d$. **T2-Opt** is the true 2-approximation threshold. **Min** is the best threshold (computed by a linear search).