Privately Clipping Heavy-Tailed Data

Matthew Joseph
Google Research
mtjoseph@google.com

Alex Kulesza
Google Research
kulesza@google.com

Abstract

Summing vectors is a basic task in differentially private data analysis. Standard algorithms clip vectors to some bound, sum them, and add noise scaled to the bound. The differential privacy literature has developed a deep library of such noise addition mechanisms but offers few tools for privately choosing the clipping bound. We introduce a method that privately chooses a clipping bound by optimizing *residual coherence*, the directional alignment of vectors exceeding the bound, against the variance cost of additional noise. We prove a relationship between residual coherence and bias for a general class of data regimes and show empirically that, across over a dozen datasets, our method outperforms existing baselines.

1 Introduction

A differentially private (DP) [13] computation controls any single data contributor's impact on the final result. For DP vector sums, this control is typically achieved by clipping vectors to a maximum norm C, summing the clipped vectors, and adding noise scaled to C. Larger clipping bounds reduce bias but increase noise, so C must be chosen carefully to minimize overall error.

The DP literature offers little guidance for this choice (see Section 1.1). Instead, it typically assumes that the algorithm operator has enough domain knowledge to choose a good clipping bound without examining the dataset. This is sometimes plausible. Most people, for example, can probably provide tight bounds on human age or height. When prior knowledge is lacking, a slightly more involved solution is to clip to a private estimate of an extreme (e.g., 0.99) quantile of the data norms. In general, choosing a clipping bound should be straightforward for any data *light-tailed* enough that typical points and outliers lie on similar scales.

This is not true for *heavy-tailed* data, where a significant minority of users may contribute vectors orders of magnitude larger than those of the "typical" core. A large body of empirical work has observed that heavy-tailed distributions arise frequently across a variety of real-world contexts [40, 15, 31], and are particularly common in internet activity [1]. Unfortunately, light-tailed heuristics fail in this setting. The 0.9, 0.95, and 0.99 norm quantiles likely lie on vastly different scales, and there is no principled reason to favor one in general: the 0.99 quantile may work well when a dataset's large vectors are well-aligned and poorly when they are not.

Perhaps not coincidentally, empirical work finds that selecting a clipping bound is a significant challenge to deploying DP [35, 11, 4]. This persists even though private sums (often in the form of private histograms) are a common primitive in public applications of DP [34, 6, 38, 18, 2].

Contributions. This work proposes a fully differentially private (and, we suggest, effectively parameter-free) method for choosing an ℓ_2 clipping bound C for computing a private d-dimensional sum. The method applies AboveThreshold [14], a canonical DP algorithm for privately answering a sequence of low-sensitivity binary queries with privacy cost scaling only with the number of "yes" responses. We use AboveThreshold to take a candidate sequence of clipping bounds $C_1 < C_2 < \cdots$ and check the current bound's *residual coherence* against a fixed threshold.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Definition 1. *Given a database of vectors $X = \{x_1, \ldots, x_n \in \mathbb{R}^d\}$ with norms $R_i = \|x_i\|_2$ and unit directions $U_i = x_i/R_i$, along with candidate clipping bound schedule $0 < C_1 < C_2 < \cdots$, the* residual coherence *at C_j is*

$$\mathsf{RC}(C_j; X) := \left\| \sum_{R_i > C_j} U_i \right\|_2 .$$

Importantly, residual coherence has bounded sensitivity of 2 under swap DP and 1 under add-remove DP. If the vectors are nonnegative, both sensitivities are 1. This makes it amenable to an algorithm like AboveThreshold.

$\mathsf{RC}(C; X)$ (which we will shorthand as $\mathsf{RC}(C)$) is large exactly when the database has many well-aligned vectors whose norms exceed C. A simple motivation is that heavy-tailed data often exhibits a two-phase trend: there is a large "core" of individuals with similar small vectors (e.g., casual movie viewers selecting a few popular movies) and a smaller "tail" of individuals diverging into more numerous niche interests. Increasing the clipping bound to shrink residual coherence can therefore approximate the point at which the signal of the clipped tail is outweighed by the noise it demands.

A full description of our algorithm, OptimizeRC, appears in Section 3.1, and we describe an efficient implementation in Section 3.2. Section 3.3 develops a simple utility result showing that residual coherence serves as a principled proxy for bias reduction in (a stylized class of) heavy-tailed data. Section 4 evaluates our method across over a dozen real-world datasets.

1.1 Related Work

Amin et al. [3] show that the optimal clipping bound for an ε-DP scalar sum is the $1 - \frac{1}{\varepsilon}$ quantile, but do not consider vector sums. Durfee [12] suggests using AboveThreshold to estimate a fixed quantile of the data norms as a clipping bound and calls the method UnboundedQuantile (which we shorthand UQ) as it does not require *a priori* bounds. While this avoids selecting C directly, it does so by shifting the hyperparameter burden to the choice of quantile. Furthermore, as we demonstrate empirically, the optimal quantile varies significantly across heavy-tailed datasets (Section 4) and no fixed quantile yields a good bias-variance tradeoff in general. Methods that simply attempt to privately estimate the maximum of a dataset [38, 26] face similar issues.

A private sum can also be computed, in a somewhat roundabout way, by privately estimating the mean and scaling its output by the database size n. Most mean estimation algorithms in the literature are ruled out by focusing on light-tailed (e.g., subgaussian) distributions or requiring their own careful parameter selections. One exception is the quantile-based algorithm of Huang et al. [23], which we discuss and evaluate in Section 4.

Several works have considered clipping gradients for DP-SGD. AdaCliP [32] and GeoClip [16] privately track gradient mean and variance statistics and use them to update subsequent clipping norm choices, but this consumes privacy budget over the course of many updates and introduces new hyperparameters that are tuned non-privately. Andrew et al. [5] propose an adaptive strategy that privately tracks a specified quantile (e.g., the median) of the update norm distribution during training, which shares the same general quantile drawbacks described above. Bu et al. [8] introduce a method that avoids choosing a clipping norm by normalizing each gradient to a unit vector, but this again shifts the choice to another hyperparameter (the learning rate), which is tuned non-privately.

2 Preliminaries

We use the "swap" notion of neighboring databases, primarily for consistency with the setup used by Durfee [12]. Our algorithm is easily modified for add-remove privacy.

Definition 2. *Two databases X, X' from domain $\mathcal{X}$ are* neighbors*, denoted $X \sim X'$, if one can be obtained from the other by replacing at most one data point.*

We use concentrated differential privacy (CDP) for similar reasons. The basic ideas of our method readily extend to other notions like pure and approximate DP, at some cost to composition neatness.

Definition 3 ([9])**.** *Randomized mechanism $M \colon \mathcal{X} \to \mathcal{O}$ is ρ-CDP if for all $X \sim X' \in \mathcal{X}$ and all $\alpha > 1$, $D_\alpha(M(X)||M(X')) \leq \rho\alpha$, where D_α denotes α-Rényi divergence.*

Our method is tuned assuming that the final private sum is computed by adding ρ-CDP (spherical) Gaussian noise. The derivation extends to any noise with a closed-form expression for the trace of its covariance matrix, but we use ρ-CDP Gaussian noise for its generality, widespread use, and clean privacy composition. We recall the basic statement for ρ-CDP Gaussian noise below.

Lemma 1 ([9]). *Given statistic $f\colon \mathcal{X} \to \mathbb{R}^d$ of ℓ_2 sensitivity C and database X, releasing $f(X) + Z$ where each $Z_j \sim_{iid} N(0, \frac{C^2}{2\rho})$ is ρ-CDP.*

Although AboveThreshold was introduced by Dwork et al. [14], we use the formulation (and improved analysis) of Durfee [12]. Pseudocode appears in Algorithm 1, where $\mathsf{Expo}(\lambda)$ has density $f_\lambda(x) \propto \frac{1}{\lambda}\exp(-x/\lambda)$ and support $\mathbb{R}_{\geq 0}$.

Algorithm 1 AboveThreshold

Require: Dataset X, queries $\{f_j : \mathcal{X} \to \mathbb{R}\}$ with sensitivity Δ, threshold T, privacy budget ρ

Set $\tilde{T} = T + \mathsf{Expo}\left(\Delta/\sqrt{2\rho}\right)$
for each query j **do**
    Set $\nu_j = \mathsf{Expo}\left(2\Delta/\sqrt{2\rho}\right)$
    if $f_j(X) + \nu_j \geq \tilde{T}$ **then return** $\top$ and halt **else return** $\perp$ and continue
end for

Lemma 2 (Corollary 3.1 and Theorem 1 [12]). AboveThreshold *is 4ρ-CDP.*

3 Optimizing Residual Coherence

3.1 Algorithm Description

The starting point for our algorithm is a straightforward decomposition of the final error of a DP sum. As mentioned earlier, for simplicity we assume that the private sum is computed by clipping each vector to ℓ_2 norm C, summing the clipped vectors, and then adding ρ_{SUM}-CDP Gaussian noise tuned to ℓ_2 sensitivity $2C$, though our method can handle any form of spherical noise with a closed expression for its variance.

We measure error using Mean Squared Error (MSE) of the final vector sum. Given database of vectors $\{x_1, \ldots, x_n \in \mathbb{R}^d\}$, define norms $R_i \coloneqq \|x_i\|_2$ and unit directions $U_i \coloneqq \frac{x_i}{\|x_i\|_2}$. For initial exposition, suppose additionally that the vectors are nonnegative (see Section 3.3 for analysis of the general case). A standard bias-variance decomposition yields

$$\begin{aligned}
\text{MSE}\,(C) &= \frac{1}{d}\mathbb{E}\left[\left\|\sum_{i=1}^n x_i - \left(\sum_{i=1}^n \mathsf{clip}\,(x_i, C) + Z\right)\right\|_2^2\right] \\
&= \frac{1}{d}\left\|\mathbb{E}\left[\sum_{i=1}^n x_i - \mathsf{clip}\,(x_i, C)\right]\right\|_2^2 + \frac{\mathsf{Tr}\,(\mathsf{Cov}\,(Z))}{d} \\
&= \frac{1}{d}\left\|\sum_{R_i > C}(R_i - C)U_i\right\|_2^2 + \frac{2C^2}{\rho_{\text{SUM}}}
\end{aligned}$$

since the sum clipped to C has swap sensitivity $2C$. To find the optimal bound C^*, we take the derivative of the MSE with respect to C[1]:

$$\begin{aligned}
\frac{d}{dC}\text{MSE}\,(C) &= \frac{2}{d}\sum_{R_i > C}(R_i - C)U_i \cdot \left(-\sum_{R_i > C} U_i\right) + \frac{4C}{\rho_{\text{SUM}}} \\
&= -\frac{2}{d}\sum_{R_i > C}\sum_{R_j > C}(R_i - C)\langle U_i, U_j\rangle + \frac{4C}{\rho_{\text{SUM}}}.
\end{aligned}$$

[1]The exact MSE is technically only piecewise differentiable due to the clipping operation, but evaluating its derivative where it exists provides the central intuition.

When the underlying vectors are nonnegative, so is each $\langle U_i, U_j \rangle$, and the overall MSE derivative goes from negative to positive as C grows. We therefore set the derivative expression to 0 and get

$$\sum_{R_i > C} \sum_{R_j > C} (R_i - C)\langle U_i, U_j \rangle = \frac{2Cd}{\rho_{\text{SUM}}}. \tag{1}$$

The LHS sum in Equation 1 is the bias derivative. Evaluating it requires knowing the private magnitude depths $R_i - C$. Since these have unbounded sensitivity without clipping, it is not obvious how to estimate them in a differentially private way. We instead approximate the bias derivative by using residual coherence (Definition 1) as a proxy:

$$\sum_{R_i > C} \sum_{R_j > C} (R_i - C)\langle U_i, U_j \rangle \approx C \cdot \mathsf{RC}\left(C\right)^2. \tag{2}$$

Unlike the bias derivative, $\mathsf{RC}\left(C\right)$ has bounded sensitivity; Section 3.3 describes other properties that justify its use. Our algorithm uses AboveThreshold to search for the threshold where $C \cdot \mathsf{RC}\left(C\right)^2$ balances the variance derivative $\frac{2Cd}{\rho_{\text{SUM}}}$, i.e., residual coherence equals $\sqrt{\frac{2d}{\rho_{\text{SUM}}}}$. Pseudocode appears in Algorithm 2. Note that we negate the queries and threshold because AboveThreshold stops when the query exceeds the threshold.

Algorithm 2 OptimizeRC

Require: Dataset X, ambient dimension d, bounds $\{C_j\}$, budgets $\rho_{\text{SUM}}, \rho_{\text{CLIP}}$

1: $T \leftarrow -\sqrt{\frac{2d}{\rho_{\text{SUM}}}}$

2: Return C_j where AboveThreshold $\left(X, -\{\mathsf{RC}\left(C_j\right)\}_{j=1}^{\infty}, \Delta = 1, T, \rho_{\text{CLIP}}/4\right)$ halts

To see that OptimizeRC is private, observe that $\mathsf{RC}\left(C\right)$ is 1-sensitive for nonnegative vectors and 2-sensitive for general vectors, then apply Lemma 2.

Lemma 3. *If the data vectors are nonnegative, then* OptimizeRC *is ρ_{CLIP}-CDP. For general vectors, it is $4\rho_{\text{CLIP}}$-CDP.*

3.2 Implementing OptimizeRC

We briefly describe an efficient implementation of OptimizeRC that uses pre-computation to speed up on-the-fly calculation of $\mathsf{RC}\left(C_j\right)$. We focus on the case where the candidate bounds C_j are powers of some base β, as used by UQ [12] and all of our experiments (see Section 4 for additional discussion).

Pseudocode appears in Algorithm 3. The algorithm makes a first pass through the data to compute each data point x_i's norm, unit vector U_i, and the index b_i of the largest clip bound C_{b_i} where it contributes to residual coherence. It then computes the last round $B = \max\{b_i\}$ with nonzero residual coherence and constructs a $(B+1) \times d$ matrix `dirs` where row j stores directions for points x_i with $b_i = j$. If $j > B$, then $\mathsf{RC}\left(C_j\right) = 0$. For the remaining clip bounds, we iterate backward to compute $\mathsf{RC}\left(C_B\right), \mathsf{RC}\left(C_{B-1}\right), \ldots$ using a running sum `Tail` of each entry of `dirs`.

The result is a complete set of the residual coherence values needed for AboveThreshold in time and space $O(nd + Bd)$, with additional possible savings using sparse matrices and parallelism. Empirical runtime results for our implementation, which uses Algorithm 3 and Scipy's sparse matrix libraries, appear in Section 4.3.3.

3.3 Residual Coherence and the Optimal Clipping Bound

This subsection proves that residual coherence at the optimal clipping bound is near our chosen threshold T. The result (Theorem 1) relies on Assumption 1, which precludes any clean linear relationship between how deep a vector lies in the tail and the direction that it points in. This can be viewed as a formalization of the rough idea that outliers tend to vary in different ways.

Assumption 1. *Given MSE-optimal clipping bound C^*, a tail vector's excess magnitude is uncorrelated with its direction:*

$$\mathbb{E}_{\mathcal{D}}\left[(R_i - C^*)U_i \mid R_i > C^*\right] = \mathbb{E}_{\mathcal{D}}\left[R_i - C^* \mid R_i > C^*\right] \cdot \mathbb{E}_{\mathcal{D}}\left[U_i \mid R_i > C^*\right]. \tag{3}$$

Algorithm 3 Efficient OptimizeRC Pre-Computation

Require: Dataset $X \in \mathbb{R}^{n \times d}$, geometric base β
for nonzero data points x_i **do**
  $R_i \leftarrow \|x_i\|_2$
  $U_i \leftarrow x_i / R_i$
  if $R_i \leq 1$ **then**
    $b_i \leftarrow -1$ ▷ -1 handles points which contribute to no $\mathsf{RC}\left(\beta^j\right)$
  else
    $b_i \leftarrow \lfloor \log_\beta(R_i) \rfloor$
  end if
end for
$B \leftarrow \max_{i \in [n]} b_i$
Initialize array `dirs` of size $(B+1) \times d$ to 0
for $i \in [n]$ **do**
  if $b_i \geq 0$ **then**
    $\texttt{dirs}[b_i] \leftarrow \texttt{dirs}[b_i] + U_i$
  end if
end for
$\texttt{Tail} \leftarrow \mathbf{0} \in \mathbb{R}^d$
Initialize array `RC` of size $B+1$
for $k = B, B-1, \ldots, 0$ **do**
  $\texttt{Tail} \leftarrow \texttt{Tail} + \texttt{dirs}[k]$
  $\texttt{RC}[k] \leftarrow \|\texttt{Tail}\|_2$
end for
return `RC`

Assumption 1 also rules out a pathological bad instance for OptimizeRC in which, for example, half of the data consists of small vectors in a shared direction, the other half of the data consists of much larger vectors in a shared opposite direction, and OptimizeRC stops prematurely because the unit vectors cancel at $C = \beta^0 = 1$. With Assumption 1, expected residual coherence is near $T = \sqrt{2d/\rho_{\text{SUM}}}$ at the optimal C^*.

Theorem 1. *Let $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ be a dataset drawn i.i.d. from a distribution $\mathcal{D}$ satisfying Assumption 1. Let C^* denote the MSE-optimal bound, let $p_{C^*} = \mathbb{P}_{\mathcal{D}}[R_i > C^*]$, and define γ by $\mathbb{E}_{\mathcal{D}}[R_i - C^* \mid R_i > C^*] = \gamma C^*$. Then $\mathbb{E}_{\mathcal{D}^n}\left[\mathsf{RC}(C^*)^2\right] = \frac{2d}{\gamma \rho_{\text{SUM}}}$, and*

$$\sqrt{\frac{n}{n-1}\left(\frac{2d}{\gamma\rho_{\text{SUM}}} - np_{C^*}\right)} \leq \mathbb{E}_{\mathcal{D}^n}\left[\mathsf{RC}(C^*)\right] \leq \sqrt{\frac{2d}{\gamma\rho_{\text{SUM}}}}.$$

Proof. Define $Y_i = \mathbb{I}(R_i > C^*)U_i$ and $Z_i = \mathbb{I}(R_i > C^*)(R_i - C^*)U_i$, so that $\mathsf{RC}(C^*) = \|\sum_{i=1}^n Y_i\|_2$ and the bias derivative at C^* is $\sum_{i=1}^n \sum_{j=1}^n \langle Z_i, Y_j \rangle$. By linearity of expectation over $[n]^2$,

$$\mathbb{E}_{\mathcal{D}^n}\left[\sum_{R_i > C^*}\sum_{R_j > C^*}(R_i - C^*)\langle U_i, U_j\rangle\right] = \sum_{i=1}^n\sum_{j=1}^n \mathbb{E}_{\mathcal{D}^n}\left[\langle Z_i, Y_j\rangle\right]. \tag{4}$$

For the diagonal terms ($i = j$), since $\langle U_i, U_i \rangle = 1$, the definition of γ yields

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}\left[\langle Z_i, Y_i\rangle\right] &= \mathbb{E}_{\mathcal{D}}\left[\mathbb{I}(R_i > C^*)(R_i - C^*)\right] \\
&= p_{C^*}\mathbb{E}_{\mathcal{D}}\left[R_i - C^* \mid R_i > C^*\right] \\
&= \gamma C^* p_{C^*} \\
&= \gamma C^* \mathbb{E}_{\mathcal{D}}\left[\|Y_i\|_2^2\right].
\end{aligned}$$

For the off-diagonal terms ($i \neq j$), we apply Assumption 1 and get

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}[Z_i] &= p_{C^*}\mathbb{E}_{\mathcal{D}}[(R_i - C^*)U_i \mid R_i > C^*] \\ &= p_{C^*}\mathbb{E}_{\mathcal{D}}[R_i - C^* \mid R_i > C^*] \cdot \mathbb{E}_{\mathcal{D}}[U_i \mid R_i > C^*] \\ &= \gamma C^* p_{C^*}\mathbb{E}_{\mathcal{D}}[U_i \mid R_i > C^*] \\ &= \gamma C^* \mathbb{E}_{\mathcal{D}}[Y_i].\end{aligned}$$

Then by independence of x_i and x_j for $i \neq j$,

$$\begin{aligned}\mathbb{E}_{\mathcal{D}^2}[\langle Z_i, Y_j\rangle] &= \langle \mathbb{E}_{\mathcal{D}}[Z_i], \mathbb{E}_{\mathcal{D}}[Y_j]\rangle \\ &= \gamma C^* \langle \mathbb{E}_{\mathcal{D}}[Y_i], \mathbb{E}_{\mathcal{D}}[Y_j]\rangle \\ &= \gamma C^* \mathbb{E}_{\mathcal{D}^2}[\langle Y_i, Y_j\rangle].\end{aligned}$$

We return to Equation 4 and get

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}_{\mathcal{D}^n}[\langle Z_i, Y_j\rangle] = \gamma C^* \sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}_{\mathcal{D}^n}[\langle Y_i, Y_j\rangle] = \gamma C^* \cdot \mathbb{E}_{\mathcal{D}^n}\left[\mathsf{RC}(C^*)^2\right]. \tag{5}$$

Setting the expected bias derivative to $\frac{2C^*d}{\rho_{\text{SUM}}}$ in Equation 5 gives $\mathbb{E}_{\mathcal{D}^n}\left[\mathsf{RC}(C^*)^2\right] = \frac{2d}{\gamma\rho_{\text{SUM}}}$, and by Jensen's inequality

$$\mathbb{E}_{\mathcal{D}^n}[\mathsf{RC}(C^*)] \leq \sqrt{\mathbb{E}_{\mathcal{D}^n}\left[\mathsf{RC}(C^*)^2\right]} = \sqrt{\frac{2d}{\gamma\rho_{\text{SUM}}}}.$$

For the lower bound, let $\mu = \mathbb{E}_{\mathcal{D}}[Y_i]$. Since $Y_1, \ldots, Y_n$ are i.i.d., we have $\mathbb{E}_{\mathcal{D}}\left[\|Y_i\|_2^2\right] = p_{C^*}$, $\mathbb{E}_{\mathcal{D}^2}[\langle Y_i, Y_j\rangle] = \langle \mathbb{E}_{\mathcal{D}}[Y_i], \mathbb{E}_{\mathcal{D}}[Y_j]\rangle = \|\mu\|_2^2$ for all $i \neq j$, and $\|\mathbb{E}_{\mathcal{D}^n}[\sum_{i=1}^n Y_i]\|_2^2 = n^2\|\mu\|_2^2$. Expanding $\mathbb{E}_{\mathcal{D}^n}\left[\mathsf{RC}(C^*)^2\right]$ therefore gives

$$\begin{aligned}\mathbb{E}_{\mathcal{D}^n}\left[\mathsf{RC}(C^*)^2\right] &= \sum_{i=1}^{n}\mathbb{E}_{\mathcal{D}}\left[\|Y_i\|_2^2\right] + \sum_{i\neq j}\mathbb{E}_{\mathcal{D}^2}[\langle Y_i, Y_j\rangle] \\ &= np_{C^*} + n(n-1)\|\mu\|_2^2 \\ &= np_{C^*} + \frac{n-1}{n}\left\|\mathbb{E}_{\mathcal{D}^n}\left[\sum_{i=1}^{n} Y_i\right]\right\|_2^2.\end{aligned}$$

Rearranging, substituting in $\mathbb{E}_{\mathcal{D}^n}\left[\mathsf{RC}(C^*)^2\right] = \frac{2d}{\gamma\rho_{\text{SUM}}}$, and applying Jensen's inequality again yields

$$\begin{aligned}\mathbb{E}_{\mathcal{D}^n}[\mathsf{RC}(C^*)] &= \mathbb{E}_{\mathcal{D}^n}\left[\left\|\sum_{i=1}^{n} Y_i\right\|_2\right] \\ &\geq \left\|\mathbb{E}_{\mathcal{D}^n}\left[\sum_{i=1}^{n} Y_i\right]\right\|_2 \\ &= \sqrt{\frac{n}{n-1}\left(\frac{2d}{\gamma\rho_{\text{SUM}}} - np_{C^*}\right)}.\end{aligned}$$

□

We note that Assumption 1 is sufficient but perhaps not necessary. The next section demonstrates that OptimizeRC performs well across a variety of datasets. No dataset was chosen to satisfy Assumption 1, and most only approximately satisfy it.

4 Experiments

This section evaluates the empirical performance of OptimizeRC on real-world datasets. For each dataset, we compare OptimizeRC against baselines across a range of ρ_{CLIP}, keeping $\rho_{\text{SUM}} = 0.1$ fixed.

In each experiment, we compute a ρ_{CLIP}-CDP clipping bound $\tilde{C}$, clip the vectors to $\tilde{C}$, sum them, apply ρ_{SUM}-CDP spherical Gaussian noise scaled to $2\tilde{C}$, and finally compute the MSE $\frac{1}{d}\sum_{j=1}^{d}(\tilde{y}_j - y_j)^2$ between the noisy vector sum $\tilde{y}$ and the true (unclipped) vector sum y. OptimizeRC uses the same candidate clip bounds $\{1.001^{j-1}\}_{j=1}^{\infty}$ throughout, and these bounds were taken from Durfee [12] before running any experiments.[2] Complete experiment code may be found on Github [19].

4.1 Baselines

We evaluate OptimizeRC against two baselines. The first is the UQ algorithm introduced by Durfee [12], which uses AboveThreshold to privately estimate a target quantile of the vectors' ℓ_2 norms. We use the same candidate clip bounds $\{1.001^{j-1}\}_{j=1}^{\infty}$ as OptimizeRC, again noting that the choice of 1.001 comes from Durfee [12]. To evaluate the role that the quantile choice plays in its performance, we run it for each target quantile in $\{0.9, 0.95, 0.99\}$, noting that Durfee [12] uses 0.99 as the single fixed choice in their experiments.[3]

The second baseline is a similar quantile-based algorithm due to Huang et al. [23], denoted HLY-UQ. It privately estimates the kth-highest ℓ_2 norm using $k = \sqrt{\frac{2d}{\rho_{\text{SUM}}}}$. Their justification is similar to the MSE decomposition presented in Section 3.1, with the simplifying (but loose and geometry-agnostic) application of $\|\sum_{R_i>C}(R_i - C)U_i\|_2 \leq \sum_{R_i>C}(R_i - C)$. Our implementation replaces their binary search-based quantile subroutine with Durfee's UQ algorithm, as this consistently and significantly reduced empirical error. We also omit their pre-quantile shifting step for the same reason.[4]

Finally, we also evaluated ApproxBounds [38], CoinPress [7], and CDPHDME [24]. However, they never beat the above baselines and are omitted from our experiments. A longer discussion of these omitted methods appears in Appendix A.

4.2 Datasets

Our experiments use ten real-world datasets (and, in the Appendix, an additional four light-tail datasets for completeness). For each dataset, we record the descriptive 4-tuple (dimensions $n \times d$, row, column, entry) below. Note that the first eight datasets only have nonnegative data, while the last two also have negative data. We set ρ_{CLIP} appropriately according to Lemma 3.

1. Steam [37]: (11350 x 3600, Steam user, game, # hours played)
2. Reddit [20]: (223388 x 153701, Reddit user, token, # times token used in comments)
3. LastFM [10]: (10237 x 49602, LastFM user, music artist, # plays), first 500,000 rows
4. HN [27]: (43679 x 18884, Hacker News user, story, comment or not), March 2026
5. MovieLens [22]: (6022 x 3043, MovieLens user, movie, whether or not user rated movie)
6. Pantry [28]: (247659 x 10814, Pantry user, item, whether or not user rated item)
7. Twitch [33]: (100000 x 739991, Twitch user, stream, whether or not user watched stream)
8. Foursquare [39]: (2293 x 61858, Foursquare Tokyo user, venue, # visits)
9. Jester [17]: (24983 x 100, Jester user, joke, rating in $[-10, 10]$)
10. Slashdot [25]: (44044 x 70284, Slashdot user, rated user, ± 1 tag)

We chose these datasets to represent a variety of contexts. Prior information consisted only of a dataset's brief qualitative description. Most datasets fit under the broad umbrella of "internet data". This is in part because we view such data as an important application for differential privacy, and in part because user-keyed internet data is most readily publicly available. Nonetheless, we suggest that the data contexts are heterogeneous enough to plausibly extend to other individual data settings.

As an aside, we note that a fully private histogram sometimes requires an initial private key selection stage to even generate the histogram domain. The domain in the Reddit dataset, for example, is a

[2] These bounds implicitly assume $C^* \geq 1$. If an end-user wishes to account for the possibility that $C^* \ll 1$, a simple solution is to additionally run OptimizeRC on a sequence of shrinking C_j until residual coherence *exceeds* the threshold.

[3] We also evaluated quantile 0.995, but it was never the best quantile choice, so we omit it.

[4] Their shifting step involves a division of the privacy budget across d. This may account for its relative success on their datasets ($d \ll 1000$) and failure on most of ours ($d \gg 1000$).

function of user data, as a user can introduce a token. We view private key selection as orthogonal to the private sum problem considered here and omit it from our pipeline, essentially treating the column headers of each dataset as public information. A closer investigation of the relationship between key selection and clipping bound selection may be an interesting direction for future work.

Finally, for completeness we also evaluated the methods on four datasets whose qualitative descriptions suggested light tails. As described in the introduction, most methods generally did well on these datasets, so we defer the plots to Appendix C. Here, we simply note that our conceptual tailoring of OptimizeRC to heavy-tailed data does not lead to poor performance on light-tailed data. Instead, OptimizeRC matches the best (low) error of the baselines.

4.3 Results

4.3.1 Error

MSE results by dataset, method, and ρ_{CLIP} value appear in Figure 1 and are summarized in Table 1. The primary metric is MSE ratio with respect to Optimal, which non-privately computes the MSE-optimal C^*. OptimizeRC achieves the strongest overall empirical performance. Its mean MSE ratio at $\rho_{\text{CLIP}} = 0.05$[5] is best on 6/10 datasets, second-best on the remaining 4/10, and suboptimal by at most a factor of 1.32 across datasets. In contrast, every other baseline is suboptimal by at least a factor of 2 (and often much more) on at least half the datasets.

Dataset	Steam	Reddit	LastFM	HN	MovieLens	Pantry	Twitch	Foursquare	Jester	Slashdot
OptimizeRC	1.02	1.03	**1.32**	1.04	1.12	**1.05**	**1.22**	**1.10**	**1.05**	**1.04**
HLY-UQ	1.23	1.08	5.45	3.97	1.43	2.94	8.40	4.31	1.27	6.93
UQ, 0.9	1.64	6.26	5.31	**1.01**	**1.05**	1.58	5.67	35.65	3.24	2.08
UQ, 0.95	**1.01**	2.76	11.15	1.41	1.32	1.06	7.74	67.95	1.22	3.96
UQ, 0.99	2.33	**1.01**	41.40	5.06	2.06	1.41	11.78	136.66	1.10	13.87
$\hat{\mathbf{F}}(\mathbf{C}^*)$	0.96	0.99	0.51	0.89	0.86	0.97	0.08	≈ 0.00	0.98	0.65

Table 1: First five rows: mean error ratio (relative to optimal clipping) across 100 trials at $\rho_{\text{CLIP}} = 0.05$ for each method (lower is better), with the lowest error ratio bolded for each dataset. Bottom row: empirical CDF at the optimal clipping bound C^* across datasets.

Figure 1 shows that some instances of UQ can achieve lower average error than OptimizeRC on some datasets. However, the empirical CDF at the optimal clipping bound C^* varies widely across datasets, from < 0.01 to 0.99 (see last row of Table 1). Consequently, even a non-private fixed quantile cannot perform well in general. In contrast, OptimizeRC incorporates the underlying data geometry to achieve low error across a range of datasets.

Additional plots in Appendix B (Figure 2) record the relationship between the residual coherence of successive clipping bounds, the threshold $\sqrt{2d/\rho_{\text{SUM}}}$, and the optimal clipping bound C^*. We find that residual coherence generally intersects the threshold relatively close to C^*, and even when it errs, it does not err as catastrophically as the quantile methods. Perhaps surprisingly, we also find that residual coherence is monotonically decreasing in C even on general data[6], as the chosen datasets lack the strongly opposing vectors that would be required to make residual coherence non-monotonic.

Plots of $\tilde{C}/C^*$ for different algorithms also appear in Appendix B (Figure 3). We find that HLY-UQ consistently overestimates the optimal C^*, and there is no (heavy-tailed dataset, ρ_{CLIP}) pair out of 100 on which HLY-UQ obtains lower error than OptimizeRC. This follows from HLY-UQ's simplifying bound $\|\sum_{R_i>C}(R_i - C)U_i\|_2 \leq \sum_{R_i>C}(R_i - C)$ (see Section 4.1), which consistently leads to overestimates relative to OptimizeRC.

[5]Relative performance does not vary much across ρ_{CLIP} (see Figure 1). $\rho_{\text{CLIP}} = 0.05$ is a representative point.

[6]It is always monotonically decreasing on nonnegative data.

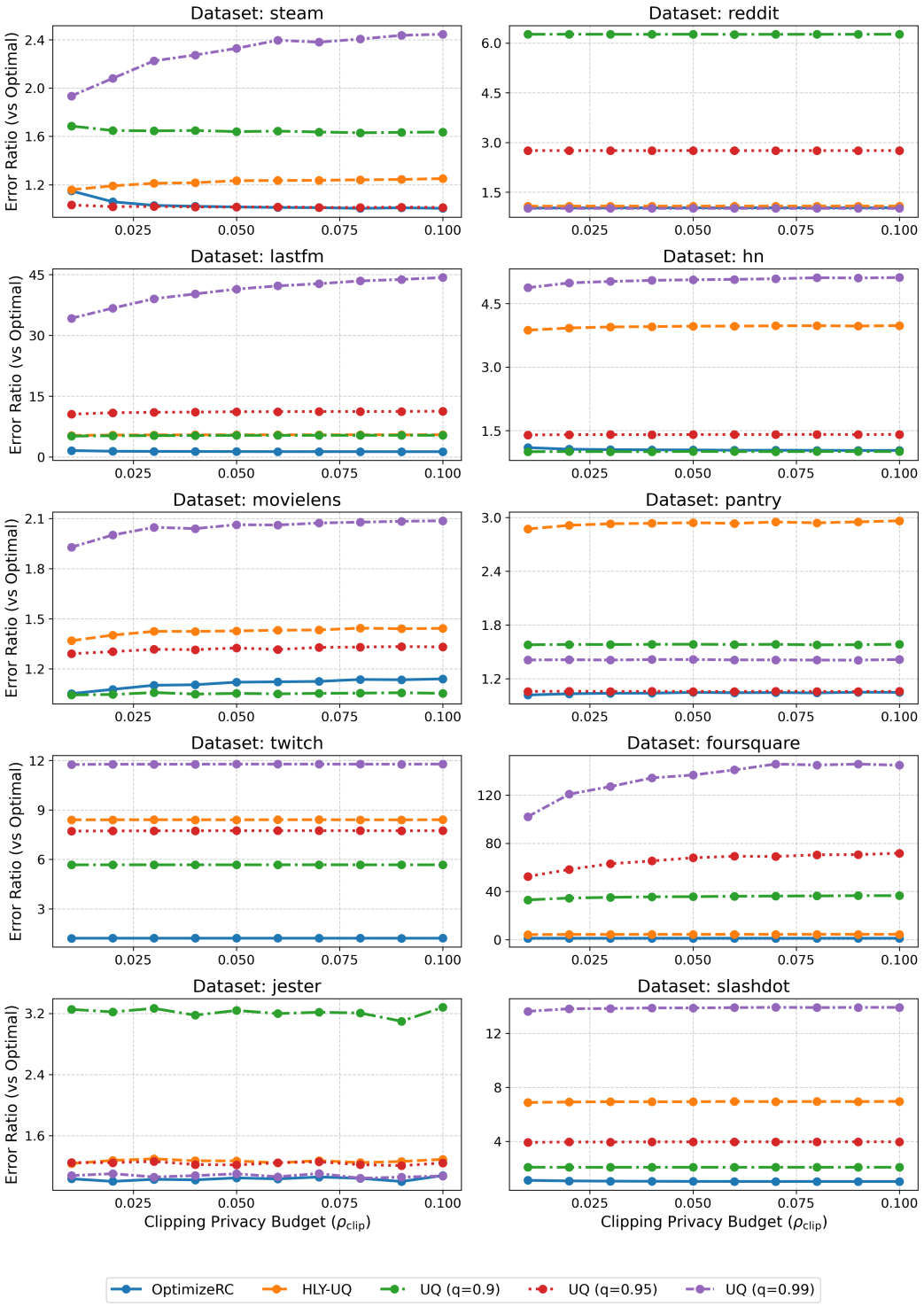

Figure 1: Plots of OptimizeRC's performance against UQ ($q = \{0.9, 0.95, 0.99\}$) and HLY-UQ. Each line is a different method, and each point on the line represents the mean of 100 trials per $\rho_{\text{CLIP}} \in \{0.01, 0.02, \ldots, 0.1\}$. We plot ratios of the mean MSE with respect to the (non-private) MSE-optimal Optimal, so a lower value is better.

4.3.2 Testing Assumption 1

Our utility result (Theorem 1) assumes that the underlying data distribution satisfies Assumption 1. The following experiment tests this assumption by evaluating the normalized ℓ_2 distance between the LHS and RHS of Equation 3 at C^* across datasets.

Dataset	Steam	Reddit	LastFM	HN	MovieLens	Pantry	Twitch	Foursquare	Jester	Slashdot
$\frac{\|\mathbf{L}-\mathbf{R}\|_2}{\|\mathbf{L}\|_2}$	0.10	0.04	0.27	0.27	0.13	0.29	0.27	0.18	0.14	0.38

Table 2: Normalized ℓ_2 distance between $\mathbf{L} = \hat{\mathbb{E}}[(R_i - C^*)U_i \mid R_i > C^*]$ and $\mathbf{R} = \hat{\mathbb{E}}[R_i - C^* \mid R_i > C^*]\hat{\mathbb{E}}[U_i \mid R_i > C^*]$ (LHS and RHS in Assumption 1). Values far from 0 correspond to worse assumption violations.

As shown in Table 2, Assumption 1 is only approximately satisfied on most of our datasets, suggesting that Theorem 1 is conservative: weaker properties may govern OptimizeRC's practical performance.

4.3.3 Time

Runtimes for evaluations on a typical personal computer appear in Table 3, where the pre-computation for a given ρ_{CLIP} is amortized over 10 trials of AboveThreshold. OptimizeRC is typically slower than UQ and HLY-UQ by approximately an order of magnitude, but it is still fast in absolute terms, consistently running in $\leq$ 1s even when the number and length of vectors are in the hundreds of thousands. This speedup is a consequence of the fast algorithm described in Section 3.2 and its implementation using SciPy's sparse matrices. We suggest that the mild increase in absolute runtime is an acceptable trade for OptimizeRC's strong error improvements.

Algorithm	Steam	Reddit	LastFM	HN	MovieLens	Pantry	Twitch	Foursquare	Jester	Slashdot
OptimizeRC	0.14	0.7	0.4	0.1	0.06	0.05	1	0.2	0.07	0.13
UQ, 0.9	0.01	0.03	0.02	0.007	0.006	0.02	0.02	0.008	0.01	0.007
UQ, 0.95	0.01	0.03	0.02	0.007	0.006	0.02	0.02	0.009	0.01	0.007
UQ, 0.99	0.02	0.03	0.02	0.009	0.006	0.02	0.02	0.01	0.01	0.009
HLY-UQ	0.01	0.03	0.02	0.009	0.006	0.02	0.02	0.006	0.01	0.008

Table 3: Average runtime (in seconds) of bound estimation algorithms.

5 Discussion

This paper proposes a method, OptimizeRC, for privately choosing a clipping bound for summing d-dimensional vectors. Incorporating the dataset's geometry makes it possible to privately select clipping bounds that yield lower MSE than those produced by the comparable baselines in the literature. We note that the small number of algorithmic choices required from the operator, such as the candidate clipping bound schedule $C_1, C_2, \ldots$, are fixed to the same untuned quantities across our experiments. We present this as evidence that it is empirically possible to privately choose good sum clipping bounds with no input from the algorithm operator.

We close with a few ways OptimizeRC might be improved. First, while Theorem 1 establishes sufficient conditions under which residual coherence intersects the desired threshold at the optimal C^*, it does not provide necessary conditions (as shown in Section 4.3.2) or prove that the output $\tilde{C}$ is near-optimal. Second, to the best of our knowledge, little is known about residual coherence. The material in Section 3.3 presents some evidence that it is a plausible low-sensitivity proxy for the bias derivative, but better identification of the data regimes in which it performs well – or better low-sensitivity alternatives – may be an interesting avenue for future work. Finally, we have not attempted to apply OptimizeRC to one of the main venues for private sums: gradient clipping in DP-SGD. Briefly, this is because improving DP-SGD does not cleanly reduce to minimizing the MSEs of gradient sums. As (non-DP) SGD demonstrates, clipping can improve convergence stability, and the appropriate degree of clipping is not well understood even absent privacy. Nonetheless, exploring OptimizeRC or similar clipping bound selection approaches in DP-SGD may be useful.

References

[1] Lada A Adamic and Bernardo A Huberman. Zipf's law and the Internet. *Glottometrics*, 2002.

[2] Temilola Adeleye, Skye Berghel, Damien Desfontaines, Michael Hay, Isaac Johnson, Cléo Lemoisson, Ashwin Machanavajjhala, Tom Magerlein, Gabriele Modena, David Pujol, et al. Publishing Wikipedia usage data with strong privacy guarantees. *arXiv preprint arXiv:2308.16298*, 2023.

[3] Kareem Amin, Alex Kulesza, Andres Munoz, and Sergei Vassilvitskii. Bounding user contributions: A bias-variance trade-off in differential privacy. In *International Conference on Machine Learning (ICML)*, 2019.

[4] Kareem Amin, Alex Kulesza, and Sergei Vassilvitskii. Practical considerations for differential privacy. *arXiv preprint arXiv:2408.07614*, 2024.

[5] Galen Andrew, Om Thakkar, Brendan McMahan, and Swaroop Ramaswamy. Differentially private learning with adaptive clipping. *Neural Information Processing Systems (NeurIPS)*, 2021.

[6] Shailesh Bavadekar, Andrew Dai, John Davis, Damien Desfontaines, Ilya Eckstein, Katie Everett, Alex Fabrikant, Gerardo Flores, Evgeniy Gabrilovich, Krishna Gadepalli, et al. Google COVID-19 search trends symptoms dataset: Anonymization process description (version 1.0). *arXiv preprint arXiv:2009.01265*, 2020.

[7] Sourav Biswas, Yihe Dong, Gautam Kamath, and Jonathan Ullman. Coinpress: Practical private mean and covariance estimation. *Neural Information Processing Systems (NeurIPS)*, 2020.

[8] Zhiqi Bu, Yu-Xiang Wang, Sheng Zha, and George Karypis. Automatic clipping: Differentially private deep learning made easier and stronger. *Neural Information Processing Systems*, 2023.

[9] Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference (TCC)*, 2016.

[10] O. Celma. Last.fm dataset. `https://zenodo.org/records/6090214`, 2010.

[11] Rachel Cummings, Damien Desfontaines, David Evans, Roxana Geambasu, Yangsibo Huang, Matthew Jagielski, Peter Kairouz, Gautam Kamath, Sewoong Oh, Olga Ohrimenko, et al. Advancing differential privacy: Where we are now and future directions for real-world deployment. *Harvard Data Science Review*, 2024.

[12] David Durfee. Unbounded differentially private quantile and maximum estimation. *Neural Information Processing Systems (NeurIPS)*, 2023.

[13] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference (TCC)*, 2006.

[14] Cynthia Dwork, Moni Naor, Omer Reingold, Guy N Rothblum, and Salil Vadhan. On the complexity of differentially private data release: efficient algorithms and hardness results. In *Symposium on the Theory of Computing (STOC)*, 2009.

[15] Xavier Gabaix. Zipf's law for cities: an explanation. *The Quarterly Journal of Economics*, 1999.

[16] Atefeh Gilani, Naima Tasnim, Lalitha Sankar, and Oliver Kosut. GeoClip: Geometry-Aware Clipping for Differentially Private SGD. In *Neural Information Processing Systems (NeurIPS)*, 2025.

[17] Ken Goldberg, Theresa Roeder, Dhruv Gupta, and Chris Perkins. Eigentaste: A constant time collaborative filtering algorithm. *Information Retrieval*, 2001. URL `https://goldberg.berkeley.edu/jester-data/`.

[18] Google. Privacy checks in ads data hub. `https://developers.google.com/ads-data-hub/guides/privacy-checks`, 2026.

[19] Google. dp_clip. https://github.com/google-research/google-research/tree/master/dp_clip, 2026.

[20] Sivakanth Gopi, Pankaj Gulhane, Janardhan Kulkarni, Judy Hanwen Shen, Milad Shokouhi, and Sergey Yekhanin. Differentially private set union. In *International Conference on Machine Learning (ICML)*, 2020.

[21] Claire C Gordon, Thomas Churchill, Charles E Clauser, Bruce Bradtmiller, John T McConville, Ilse Tebbetts, and Robert A Walker. 1988 Anthropometric survey of US army personnel: methods and summary statistics. *Technical Report Natick/TR-89/044*, 1988.

[22] F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *Transactions on Intelligent Systems*, 2015.

[23] Ziyue Huang, Yuting Liang, and Ke Yi. Instance-optimal mean estimation under differential privacy. *Neural Information Processing Systems (NeurIPS)*, 2021.

[24] Gautam Kamath, Vikrant Singhal, and Jonathan Ullman. Private mean estimation of heavy-tailed distributions. In *Conference on Learning Theory (COLT)*, 2020.

[25] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. Signed networks in social media. In *Conference on Human Factors in Computing Systems (CHI)*, 2010.

[26] Joseph P Near and Chiké Abuah. Programming differential privacy. *Foundations and Trends® in Programming Languages*, 2025.

[27] Duc-Tam Nguyen. Hacker news dataset. https://huggingface.co/datasets/open-index/hacker-news, 2026.

[28] Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

[29] US Bureau of Labor Statistics. American time use survey (atus). https://www.kaggle.com/datasets/bls/american-time-use-survey/data, 2016.

[30] OpenPowerlifting. Powerlifting database. https://www.kaggle.com/datasets/open-powerlifting/powerlifting-database?select=openpowerlifting.csv, 2023.

[31] Steven T Piantadosi. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, 2014.

[32] Venkatadheeraj Pichapati, Ananda Theertha Suresh, Felix X Yu, Sashank J Reddi, and Sanjiv Kumar. Adaclip: Adaptive clipping for private sgd. *arXiv preprint arXiv:1908.07643*, 2019.

[33] Jérémie Rappaz, Julian McAuley, and Karl Aberer. Recommendation on live-streaming platforms: Dynamic availability and repeat consumption. In *Conference on Recommender Systems (RecSys)*, 2021.

[34] Apple Machine Learning Research. Learning iconic scenes with differential privacy. https://machinelearning.apple.com/research/scenes-differential-privacy, 2023.

[35] Jayshree Sarathy, Sophia Song, Audrey Haque, Tania Schlatter, and Salil Vadhan. Don't look at the data! how differential privacy reconfigures the practices of data science. In *Conference on Human Factors in Computing Systems (CHI)*, 2023.

[36] Thomas Shaw. Nfl combine - 2010 to 2023 performance data. https://www.kaggle.com/datasets/thomassshaw/nfl-combine-performance-dataset, 2023.

[37] Tamber. Steam games dataset. https://www.kaggle.com/datasets/tamber/steam-video-games, 2016.

[38] Royce J Wilson, Celia Yuxin Zhang, William Lam, Damien Desfontaines, Daniel Simmons-Marengo, and Bryant Gipson. Differentially private SQL with bounded user contribution. In *Privacy Enhancing Technologies Symposium (PETS)*, 2020.

[39] Dingqi Yang, Daqing Zhang, Vincent W. Zheng, and Zhiyong Yu. Modeling User Activity Preference by Leveraging User Spatial Temporal Characteristics in LBSNs. *IEEE Transactions on Systems, Man, and Cybernetics*, 2015.

[40] George Kingsley Zipf. *Human behavior and the principle of least effort: An introduction to human ecology*. Addison-Wesley Press, 1949.

A Omitted Methods

This section discusses three omitted alternative methods for private clipping bound selection: ApproxBounds [38], CoinPress [7], and CDPHDME [24]. At a high level, the three methods pursue different algorithmic strategies, but they all include a step that deliberately chooses a clipping bound large enough to leave most points unclipped. This introduces the same problems as any method that privately estimates a large fixed quantile (see Section 4.3.1) and is in each case compounded by other algorithmic details. Explanations for each method follow.

ApproxBounds computes, e.g., private sums by constructing a private histogram over the data's ℓ_2 norms, computes a threshold count T unlikely to be crossed by any truly empty bin, taking the largest bin with count at least T as the clipping bound, and computing a standard clipped private sum. In addition to the aforementioned problems in taking a fixed quantile as the clipping bound, this histogram approach is additionally a less accurate private quantile primitive than UnboundedQuantile's AboveThreshold, so it is empirically dominated by UnboundedQuantile.

CoinPress computes a private mean by projecting the data onto some ball containing the true mean, estimating a simple private clipped mean using additive noise, and repeating this process with geometrically shrinking balls. The clipping bound chosen for each clipped mean is set using Gaussian tail bounds calibrated to probably contain all data points. This does not yield an obvious analogue when the data is heavy-tailed, and the general clipping strategy is effectively attempting to use the maximum norm as a clipping bound. Finally, we note that the experiments of Huang et al. [23] generally show HLY-UQ outperforming CoinPress outside of well-specified Gaussian data.

CDPHDME is designed for heavy-tailed private mean estimation. It constructs a private histogram over each dimension to obtain an initial coordinate-wise mean estimate, truncates data to a ball around this estimate, and finally computes a standard private clipped mean using the ball. However, the initial mean estimate requires an input interval $[-R, R]$ containing the true mean and a bound on the data distribution's k^{th} moment and further requires $n \gg d$ samples for the result to be accurate. It is not clear how to set either parameter in general, and only 4/10 of our primary datasets have $n \geq 2d$. Furthermore, the second estimation step chooses a truncated ball radius based on worst-case concentration bounds for distributions with bounded k^{th} moment, calibrated to contain nearly all data points. This again inherits the aforementioned problems of clipping to accommodate the largest data points.

Finally, we note that evaluations of the above methods on our datasets never beat the three methods discussed in the text, though the number of user-specified parameters in CoinPress and CDPHDME makes a standardized untuned comparison difficult.

B Additional Heavy-Tail Plots

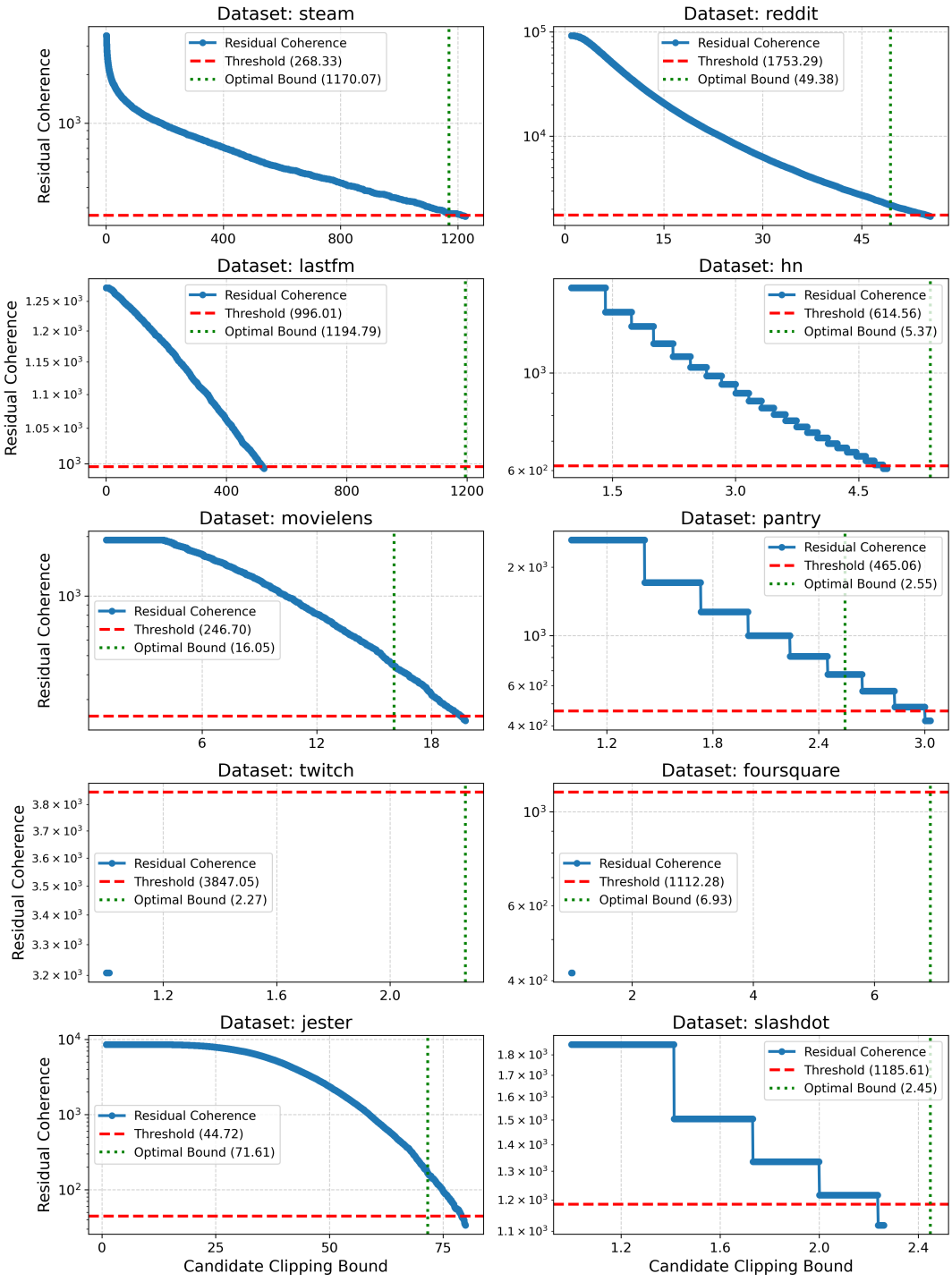

Figure 2: Plots of residual coherence vs. clipping bounds across all evaluated datasets. The horizontal line is $-T$ in OptimizeRC. Note that HN, Pantry, and Slashdot have a distinctive "staircase" shape because the small clipping bound range accentuates the underlying integer values. For the Twitch and Foursquare datasets, residual coherence is immediately below the threshold, and OptimizeRC outputs 1. This aggressively clipped bound nonetheless beats the UnboundedQuantile methods, which select bounds that are too large (see Figure 3).

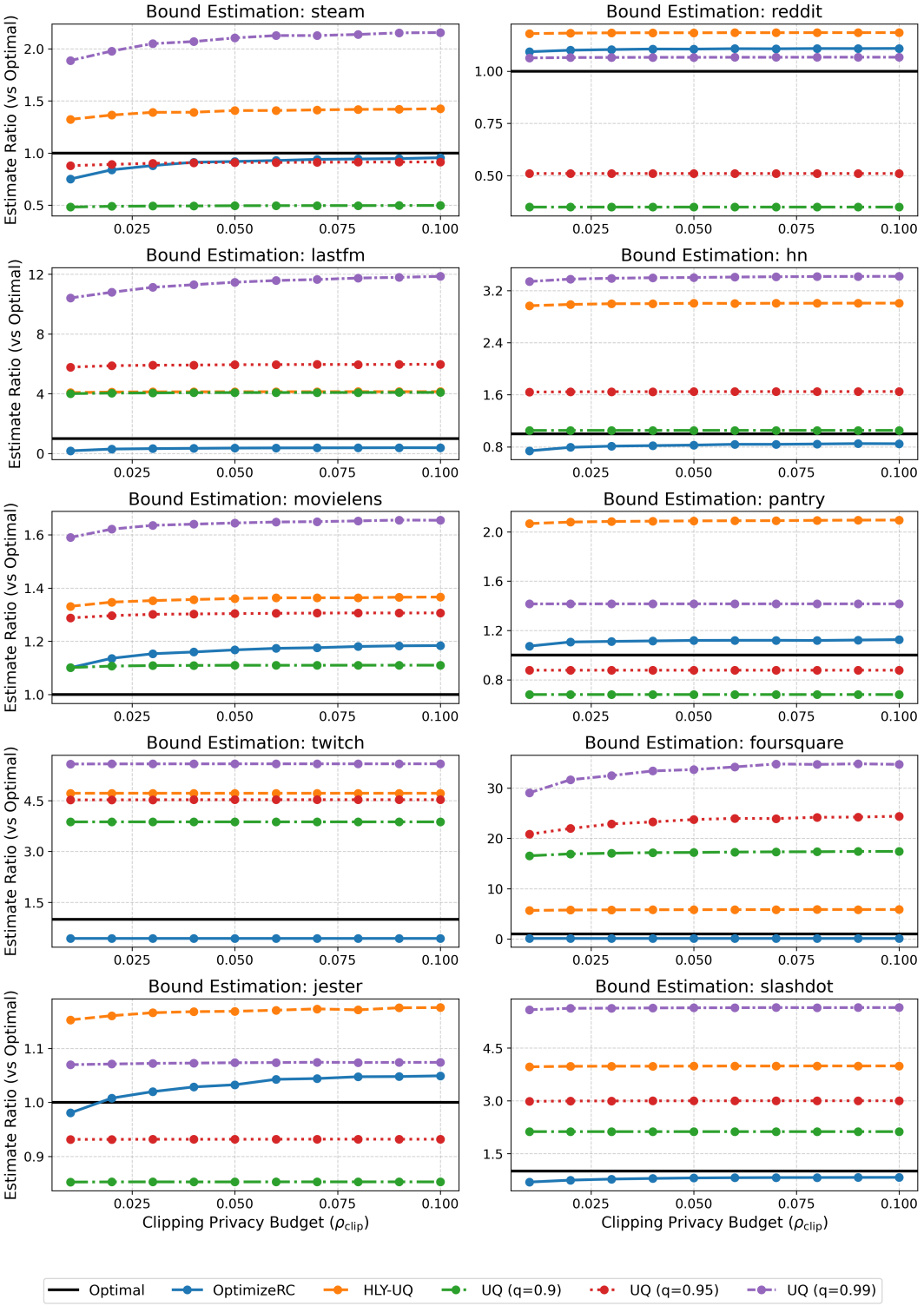

Figure 3: Plots illustrating the ratio between the estimated clipping bound and the optimal C^*. A value < 1 means the estimated bound is smaller, and a value > 1 means it is larger.

C Light-Tailed Experiments

The light-tailed experiments use the following four datasets. We again record descriptive 4-tuples (dimensions $n \times d$, row, column, entry) below.

1. ANSUR [21]: (1774 x 132, individual, measurement type, size in cm)
2. ATUS [29]: (5008 x 71, individual, ATUS activity, daily time spent in hours), restricted to first 100,000 rows
3. NFL [36]: (2555 x 6, NFL combine participant, drill, score)
4. Powerlifting [30]: (45268 x 5, lifter, metric, value), restricted to first 100,000 rows

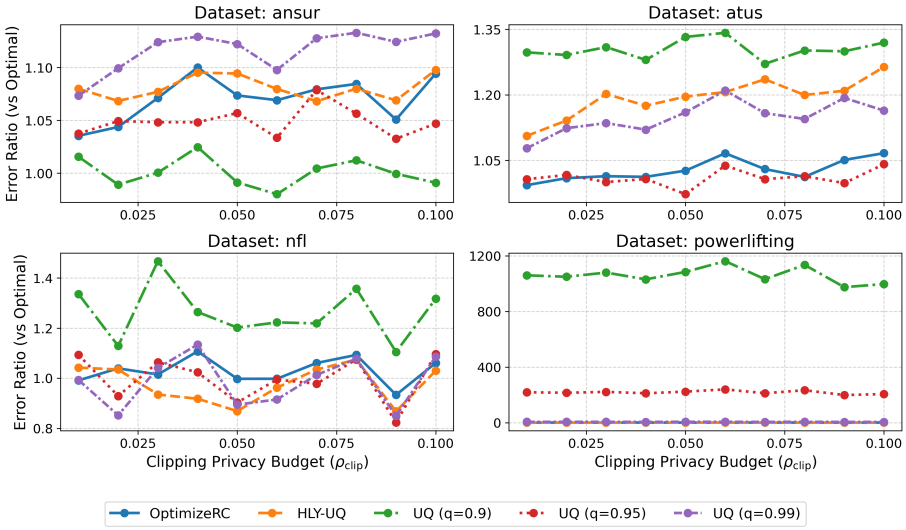

Figure 4: An analogue of Figure 1, but for light-tailed datasets. Note that methods occasionally beat Optimal on NFL because the low d increases the noise variance.

Note that OptimizeRC's utility is similar to that of the heavy-tail setting, though the UnboundedQuantile 0.99 and HLY-UQ baselines now often perform similarly well. Finally, we include clipping bound plots for completeness.

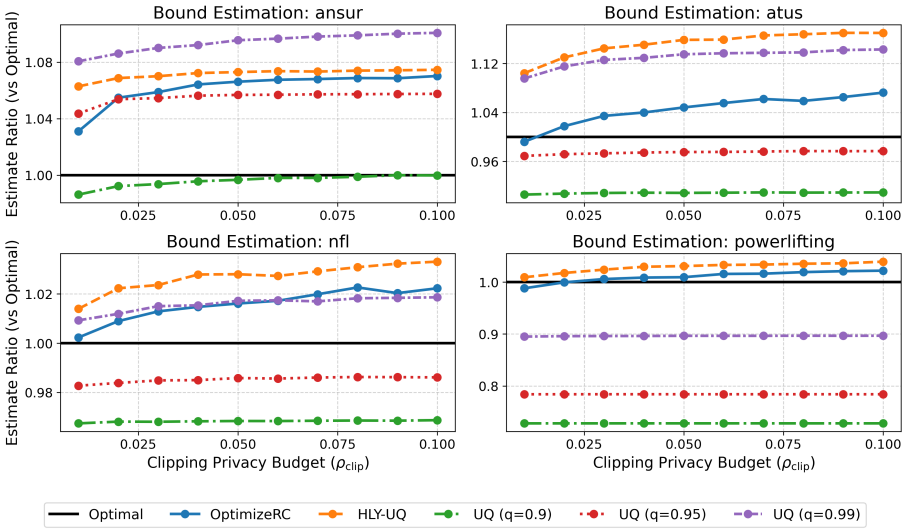

Figure 5: An analogue of Figure 3, but for light-tailed datasets.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   [Yes]

   Justification: The paper's main claims appear on the first page and are supported by the rest of the main body.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We pointed out limitations including: specificity to a class of heavy-tailed data, sufficiency rather than necessity of the assumption in the utility result, and slightly increased time and space requirements. More discussion and future work appears in the final Discussion section.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Assumptions are clearly stated before our main utility result.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We include a link to a public code repository with instructions for reproducing the (public data) experiments, in addition to implementation discussions within the paper itself.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We include a code supplement with documentation, and all experiments are on publicly available data.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes, all parameters and their choices (fixed across all datasets) are discussed in the Experiments section.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Error bars were typically small enough that they didn't add much over simply recording means.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All experiments were run on a typical personal computer, and this is mentioned.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: All datasets used are publicly available, and not in the list of prohibited datasets. The paper involves no new data or human experiments.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: The paper provides a method to help compute differentially private sums. We suggest that it does little beyond making differential privacy easier. Whatever possible negative societal impacts there are, they apply to every differential privacy paper.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: See previous response. This is just a differential privacy paper, and the scope for misuse appears quite low.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Every dataset used has an accompanying citation and link.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A].

    Justification: We do not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: We do not use crowdsourcing or human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A].

    Justification: We do not use human subjects

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: We only used LLMs for experiment code and minor math assistance and editing. The prose is human-written and AI-generated code has been human-checked.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.