Mean Estimation in the Add-Remove Model of Differential Privacy

Alex Kulesza [*1] **Ananda Theertha Suresh** [*1] **Yuyan Wang** [*1]

Abstract

Differential privacy is often studied under two different models of neighboring datasets: the *add-remove* model and the *swap* model. While the swap model is frequently used in the academic literature to simplify analysis, many practical applications rely on the more conservative add-remove model, where obtaining tight results can be difficult. Here, we study the problem of one-dimensional mean estimation under the add-remove model. We propose a new algorithm and show that it is min-max optimal, achieving the best possible constant in the leading term of the mean squared error for all ε, and that this constant is the same as the optimal algorithm under the swap model. These results show that the add-remove and swap models give nearly identical errors for mean estimation, even though the add-remove model cannot treat the size of the dataset as public information. We also demonstrate empirically that our proposed algorithm yields at least a factor of two improvement in mean squared error over algorithms frequently used in practice. One of our main technical contributions is a new *hourglass* mechanism, which might be of independent interest in other scenarios.

1. Introduction

Mean estimation is one of the simplest and most widely used techniques in statistics, and is often deployed as a subroutine for more complex analyses. However, the mean of a dataset can reveal private information, and so a variety of differentially private methods have been proposed to estimate it. These include private mean estimation with robustness (Liu et al., 2021), mean estimation under statistical models (Kamath et al., 2020), and techniques with instance-specific guarantees (Huang et al., 2021; Dick et al., 2023).

[*]In alphabetical order [1]Google Research, NYC. Correspondence to: Yuyan Wang <wangyy@google.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

However, despite the ubiquitous nature of differentially private mean estimation, fundamental questions remain about the optimal error that can be achieved under commonly used variants of differential privacy. We begin with the general definition of a differentially private mechanism.

Definition 1.1 (Differential privacy (Dwork et al., 2014))**.** A randomized, real-valued algorithm A satisfies ε-differential privacy if for any two neighboring datasets D, D' and for any output $\mathcal{S} \subseteq \mathcal{R}$, it holds that

$$\Pr[A(D) \in \mathcal{S}] \leq e^{\varepsilon} \cdot \Pr[A(D') \in \mathcal{S}].$$

It remains to specify what makes two datasets *neighbors*. Two definitions are commonly used. The first, called the *swap* model (Dwork et al., 2006; Vadhan, 2017), defines two datasets D and D' as neighboring if and only if

$$|D \setminus D'| = 1 \text{ and } |D' \setminus D| = 1.$$

The second, called the *add-remove* model (Dwork et al., 2014, Definition 2.4), defines D and D' as neighboring if and only if

$$|D \setminus D'| + |D' \setminus D| = 1.$$

Intuitively, under the swap model, D' is obtained from D by changing a value, while under the add-remove model, it is obtained by adding or removing a value.

While both neighborhood models of differential privacy are studied in the literature, the add-remove model is more frequently used for statistical queries in practice (McSherry, 2009; Wilson et al., 2019; Rogers et al., 2020; Amin et al., 2022), likely because it is more conservative: the add-remove model protects the *size* of the input dataset, while the swap model does not. Furthermore, a ε-DP algorithm under the add-remove model is also a 2ε-DP algorithm under the swap model, so in this sense the relationship is strict.

Here, we revisit the problem of scalar mean estimation in the add-remove model of differential privacy, proposing a new algorithm that is min-max optimal, including the constant on the leading term of the error, and showing that the add-remove and swap models give nearly identical errors despite the former's additional protections.

1.1. Setting

Let $\mathcal{D}_n(\ell, u)$ denote the set of all datasets consisting of n real values in the range $[\ell, u]$, let $\mathcal{D}_{\geq n}(\ell, u)$ denote the set of

all datasets consisting of *at least* n real values in the range $[\ell, u]$, and let $\mathcal{D}^*(\ell, u)$ denote the set of datasets of all sizes with values in $[\ell, u]$.

Given a dataset $D = \{x_1, x_2, \dots, \}$, the mean is given by

$$\mu(D) \triangleq \frac{1}{|D|} \sum_{x \in D} x.$$

We measure the utility of a mean estimator $\hat{\mu} : \mathcal{D}^*(\ell, u) \to [\ell, u]$ for a dataset D in terms of mean squared error (MSE),

$$L(\hat{\mu}, D) \triangleq \mathbb{E}[(\hat{\mu}(D) - \mu(D))^2],$$

where the expectation is over the randomization of the estimator. Typically, for private estimators, $L(\hat{\mu}, D)$ decreases with the size of D as $O(1/|D|^2)$. Hence, we measure the normalized mean squared error as

$$L_{\text{norm}}(\hat{\mu}, D) \triangleq |D|^2 L(\hat{\mu}, D).$$

Let $\mathcal{A}_\varepsilon^{\text{sw}}$ denote the set of all ε-differentially private algorithms under the swap model, and let $\mathcal{A}_\varepsilon^{\text{ar}}$ denote the set of all ε-differentially private algorithms under the add-remove model. The min-max normalized mean squared error for sufficiently large datasets (of size at least n_0) in the swap model is defined as

$$R_{\text{sw}}(\varepsilon, n_0, \ell, u) \triangleq \inf_{\hat{\mu} \in \mathcal{A}_\varepsilon^{\text{sw}}} \sup_{D \in \mathcal{D}_{\geq n_0}(\ell, u)} L_{\text{norm}}(\hat{\mu}, D),$$

and similarly in the add-remove model is

$$R_{\text{ar}}(\varepsilon, n_0, \ell, u) \triangleq \inf_{\hat{\mu} \in \mathcal{A}_\varepsilon^{\text{ar}}} \sup_{D \in \mathcal{D}_{\geq n_0}(\ell, u)} L_{\text{norm}}(\hat{\mu}, D).$$

1.2. Optimality in the swap model

Geng and Viswanath (2014) showed that

$$R_{\text{sw}}(\varepsilon, n_0, \ell, u) = (u - \ell)^2 \cdot \sigma^2(\varepsilon) \cdot (1 \pm o(1)), \quad (1)$$

where the $o(1)$ term (here and throughout the paper) tends to zero as $n_0 \to \infty$ for fixed ε, ℓ and u, and

$$\sigma^2(\varepsilon) = \frac{2^{-2/3} e^{-2\varepsilon/3} (1 + e^{-\varepsilon})^{2/3} + e^{-\varepsilon}}{(1 - e^{-\varepsilon})^2}. \quad (2)$$

As $\varepsilon \to 0$, $\sigma^2(\varepsilon) \to 2/\varepsilon^2$, and hence for small values of ε

$$R_{\text{sw}}(\varepsilon, n_0, \ell, u) \overset{\varepsilon \to 0}{\approx} \frac{2(u - \ell)^2}{\varepsilon^2} (1 \pm o(1)). \quad (3)$$

The Laplace mechanism matches this mean squared error up to the $o(1)$ term, and in this sense is optimal as $\varepsilon \to 0$.

However, for larger ε, the Laplace mechanism is not optimal. Geng and Viswanath (2014) defined a class of differentially private mechanisms called *staircase* mechanisms whose density is parameterized by $\gamma \in [0, 1]$ and showed that, for any monotonic loss, there exists a γ such that the staircase mechanism is the optimal differentially private mechanism for that loss. We provide a definition of the staircase mechanism in Definition A.1 for completeness.

1.3. Optimality in the add-remove model

The story is a bit more complicated under the add-remove model. Since the mean is the ratio of the sum ($s \triangleq \sum_{x \in D} x$) to the count ($n \triangleq |D|$), one simple algorithm for private mean estimation is to use a fraction of the privacy budget (say, $\varepsilon/2$) to estimate the sum as $\hat{s}$, use the remaining privacy budget ($\varepsilon/2$) to estimate the count as $\hat{n}$, and finally estimate the mean as $\hat{s}/\hat{n}$. Since the true mean always lies in the range $[\ell, u]$, we can additionally clip the result to $[\ell, u]$ to improve accuracy. This standard algorithm is shown in Algorithm 1, where $\text{Clip}(x, [a, b]) = \max(a, \min(x, b))$.

Algorithm 1 Independent noise addition.

Input: Multiset $D \subset [l, u]$, $\varepsilon > 0$.

1 Let $w = \max(|\ell|, |u|)$.
2 Let $s = \sum_{x \in D} x$.
3 Let $n = |D|$.
4 Let $\hat{s} = s + Z_s$, where $Z_s \sim \text{Lap}(\frac{2w}{\varepsilon})$.
5 Let $\hat{n} = n + Z_n$, where $Z_n \sim \text{Lap}(\frac{2}{\varepsilon})$.
6 Output $\hat{\mu} = \text{Clip}(\frac{\hat{s}}{\hat{n}}, [\ell, u])$.

Algorithm 2 Shifted noise addition.

Input: Multiset $D \subset [l, u]$, $\varepsilon > 0$.

1 Let $w = u - l$ and $m = \frac{l+u}{2}$.
2 Let $D' = D - m$.
3 Let $s = \sum_{x \in D'} x$.
4 Let $n = |D'|$.
5 Let $\hat{s} = s + Z_s$, where $Z_s \sim \text{Lap}(\frac{w}{\varepsilon})$.
6 Let $\hat{n} = n + Z_n$, where $Z_n \sim \text{Lap}(\frac{2}{\varepsilon})$.
7 Output $\hat{\mu} = \text{Clip}(\frac{\hat{s}}{\hat{n}}, [-\frac{w}{2}, \frac{w}{2}]) + m$.

The noise added in Algorithm 1 is proportional to $\max(|\ell|, |u|)$, which can be badly suboptimal, for instance if $l = 10^6$ and $u = 10^6 + 1$. Algorithm 2 modifies Algorithm 1 by shifting the inputs by $(\ell + u)/2$ before computing the sum, reducing the sensitivity to $(u - \ell)/2$. This improves the error, sometimes dramatically. It can be shown that, for any dataset $D \in \mathcal{D}^*(\ell, u)$,

$$L_{\text{norm}}(\text{Algorithm 2}, \ell, u) \leq \frac{4(u - \ell)^2}{\varepsilon^2} (1 + o(1))$$

and hence

$$R_{\text{ar}}(\varepsilon, n_0, \ell, u) \leq \frac{4(u - \ell)^2}{\varepsilon^2} (1 + o(1)). \quad (4)$$

For simplicity, we drop n_0 in the notation of R_{sw} and R_{ar} for the rest of the paper.

This result is still a factor of two larger than the swap model lower bound in (3), and the loss is even further from R_{sw}

in the low-privacy regime where ε is large. Recently, Kamath et al. (2023) proposed a generic mechanism to convert a differentially private algorithm in the swap model to a differentially private algorithm in the add-remove model and instantiated it for the unbiased mean estimation problem. However, their focus was not the constant in the mean squared error, and the stated result (Kamath et al., 2023, Theorem D.6) has a constant of 9. We note that swap and add-remove models are also referred to as bounded and unbounded models of differential privacy, respectively (Tramèr et al., 2015; Takagi et al., 2023).

Thus, a natural question remains: is mean estimation in the add-remove model inherently harder than in the swap model? We show in this paper that the answer is no.

2. Our contributions

We propose a new mean estimation algorithm in Algorithm 3, introducing two key improvements over Algorithm 2:

Transformed noise addition. Instead of estimating the sum and count directly as in Algorithm 2, the new algorithm estimates a linear transformation of the sum and count to reduce ℓ_1 sensitivity. This is sufficient to achieve optimal error using the vector Laplace mechanism in the high-privacy regime where ε is small.

The hourglass mechanism. To achieve optimal error in the low-privacy regime where ε is large, we propose a new two-dimensional noise distribution called the *hourglass* mechanism. It has the desirable property that the marginal distribution over either dimension is the optimal univariate staircase mechanism (Geng and Viswanath, 2014).

We show that the hourglass mechanism can be sampled efficiently, and prove a bound on the mean squared error of Algorithm 3 when the noise is drawn from the hourglass mechanism. Combined with an information-theoretic lower bound on $R_{\text{ar}}(\varepsilon, \ell, u)$, this shows that Algorithm 3 is optimal for all ε, ℓ, and u.

These bounds also match the result of Geng and Viswanath (2014) for $R_{\text{sw}}(\varepsilon, \ell, u)$, establishing that the swap model and the add-remove model give the same mean squared error (up to $o(1)$ terms).

We note in passing that, when the bounds ℓ and u are unknown, Algorithm 3 can be combined with standard clipping algorithms to perform on unbounded domains (Amin et al., 2019).

2.1. Overview of technical results

We first analyze Algorithm 2 as a baseline and provide a dataset specific upper bound in Lemma 3.2, proving that for any dataset D, its mean squared error is upper bounded by

$$\left(\frac{2(u-l)^2}{|D|^2\varepsilon^2} + \frac{8(\mu - \frac{u+\ell}{2})^2}{|D|^2\varepsilon^2}\right)(1+o(1)).$$

We then analyze Algorithm 3 with Laplace noise in Theorem 3.3, showing that its error is at most

$$\left(\frac{(u-l)^2}{|D|^2\varepsilon^2} + \frac{4(\mu - \frac{u+\ell}{2})^2}{|D|^2\varepsilon^2}\right)(1+o(1)), \tag{5}$$

and hence

$$R_{\text{ar}}(\varepsilon, \ell, u) \leq \frac{2(u-\ell)^2}{\varepsilon^2}(1+o(1)).$$

This is the same as the min-max MSE of the swap model for small values of ε.

We next analyze Algorithm 3 with noise drawn from the two-dimensional staircase mechanism proposed by Geng et al. (2015) in Lemma 4.4, and show that its error is at most

$$\left(\frac{(u-l)^2\widetilde{\sigma}^2(\varepsilon)}{|D|^2}\right)(1+o(1)),$$

where $\widetilde{\sigma}^2(\varepsilon)$ is the variance of the two-dimensional staircase mechanism optimized for mean squared error with privacy guarantee ε. While the above result is better than (5) for large values of ε, it still does not match the error of the swap model in general. This is due to the fact that for large values of ε, the error of the swap model scales as

$$\sigma^2(\varepsilon) = \Theta\left((u-\ell)^2 e^{-2\varepsilon/3}\right),$$

while the error of the two dimensional staircase mechanism applied in Algorithm 3 scales as

$$\widetilde{\sigma}^2(\varepsilon) = \Theta\left((u-\ell)^2 e^{-\varepsilon/2}\right).$$

(See Lemma A.3 in the Appendix for details.)

We finally analyze Algorithm 3 with noise drawn from the hourglass mechanism in Theorem 4.5, showing that its error is upper bounded by

$$\left(\frac{(u-l)^2\sigma^2(\varepsilon)}{|D|^2}\right)(1+o(1)),$$

where $\sigma^2(\varepsilon)$ is given in (2), matching the swap model lower bound in (1).

In addition, we prove an information-theoretic lower bound for the add-remove model in Theorem 5.2:

$$R_{\text{ar}}(\varepsilon, \ell, u) \geq (u-\ell)^2 \cdot \sigma^2(\varepsilon) \cdot (1-o(1)),$$

Algorithm 3 Transformed noise addition.

Input: Multiset $D \subset [l, u]$, $\varepsilon > 0$.

Let $w = u - l$.
Let $D' = D - l$.
Let $s_1 = \sum_{x \in D'} x/w$.
Let $s_2 = \sum_{x \in D'} (1 - x/w)$.
Let $Z = (Z_1, Z_2) \sim$ two-dim. noise mechanism(ε).
Let $\hat{s}_1 = s_1 + Z_1$.
Let $\hat{s}_2 = s_2 + Z_2$.
Output $\hat{\mu} = w \cdot \text{Clip}(\frac{\hat{s}_1}{\hat{s}_1 + \hat{s}_2}, [0, 1]) + l$.

establishing that

$$R_{\text{ar}}(\varepsilon, \ell, u) = (u - \ell)^2 \cdot \sigma^2(\varepsilon) \cdot (1 \pm o(1)),$$

and therefore that Algorithm 3 is optimal, as well as that the add-remove and swap models have equivalent mean squared error for mean estimation.

The rest of the paper is organized as follows. In Section 3 we discuss the high-privacy regime, showing how a linear transformation on the sum and count leads to optimal error using the Laplace mechanism. In Section 4, we generalize our results result to the low-privacy regime, introducing the hourglass mechanism in Section 4.1 and applying it to mean estimation in Section 4.2. In Section 5 we prove an information-theoretic lower bound showing that our results with the hourglass mechanism are optimal in the add-remove model and match the optimal error in the swap model as well. Finally, in Section 6, we empirically demonstrate the performance of our algorithm.

3. High privacy regime

We first build intuition by viewing Algorithms 1 and 2 geometrically, drawing on the framework of Hardt and Talwar (2010); Awan and Slavković (2021). To simplify, we will assume $\ell = 0$ and $u = 1$. Let

$$q(D) = \left(\sum_{x \in D} x, \sum_{x \in D} 1 \right)$$

be the two-dimensional vector containing the sum and count for dataset D. Define the *sensitivity space* $S(q)$ to be the set of possible values for $q(D) - q(D')$ when D and D' are neighboring datasets. Under the add-remove model, the sensitivity space for q is $\cup_{x \in [0,1]} \{(x, 1), (-x, -1)\}$, depicted by the two bold line segments in the middle plot of Figure 1. Throughout the paper, we use (x, y) to denote a two-dimensional vector, and $[a, b]$ to denote a closed interval. We use $[(x_1, y_1), (x_2, y_2)]$ to denote a segment in the two-dimensional space with end points (x_1, y_1) and (x_2, y_2), with both ends included.

The standard vector Laplace mechanism can be used to obtain a differentially private estimate of $q(D)$ based on its sensitivity. In particular, the mechanism adds noise scaled to the maximum ℓ_1 norm of the sensitivity space—that is, the smallest constant a such that the ℓ_1 ball $\{x : \|x\|_1 = a\}$ contains $S(q)$. Here, the minimum value of a is 2, as shown by the red diamond in the middle plot of Figure 1. And, indeed, adding Laplace noise scaled to $a = 2$ is precisely what Algorithm 1 does.

However, noise added in this way actually supports a much larger sensitivity space than just $S(q)$, as can be seen in the figure. This means that Algorithm 1 is effectively "wasting" noise to protect against changes to q that cannot occur, unnecessarily increasing error. (The problem is even worse when the range $[\ell, u]$ is far from zero, as noted earlier.)

Compared to this naive approach, Algorithm 2 is significantly better, since the shift in Line 2 maps values in the dataset from $[0, 1]$ to $[-\frac{1}{2}, \frac{1}{2}]$. The bold line segments in the right plot of Figure 1 depict the new sensitivity space, which is contained by the ℓ_1 ball with smaller scale $a = 1$ (shown in orange).

In fact, one can show that Algorithm 2 is equivalent to the following procedure:

1. Apply a linear transformation given by the matrix $\begin{bmatrix} 1 & -\frac{1}{2} \\ 0 & \frac{1}{2} \end{bmatrix}$ to the sum and count vector $q(D)$.
2. Add vector Laplace noise to the transformed vector according to its (reduced) sensitivity ($a = 1$).
3. Reverse the transformation by applying the inverse matrix $\begin{bmatrix} 1 & 1 \\ 0 & 2 \end{bmatrix}$, then divide to estimate the mean and truncate as before.

For a detailed proof of the above claim, see Appendix B.

In the original (untransformed) q-space, depicted in the middle plot of Figure 1, the region protected by the resulting noise distribution is a parallelogram, shown in orange. Compared with the red diamond of Algorithm 1, this region more tightly encloses the sensitivity space, reducing noise and increasing accuracy. However, it still does not perfectly enclose $S(q)$.

3.1. Optimizing the Laplace mechanism via linear transformation

The approach in Algorithm 3 is to transform q so that the ℓ_1 unit ball encloses it as tightly as possible. To do this, it maps the two segments of the sensitivity space $[(0, 1), (1, 1)]$ and $[(-1, -1), (0, -1)]$ onto two sides of the ℓ_1 ball using the transformation shown in the left plot of Figure 1. The enclosing ℓ_1 ball is depicted in blue.

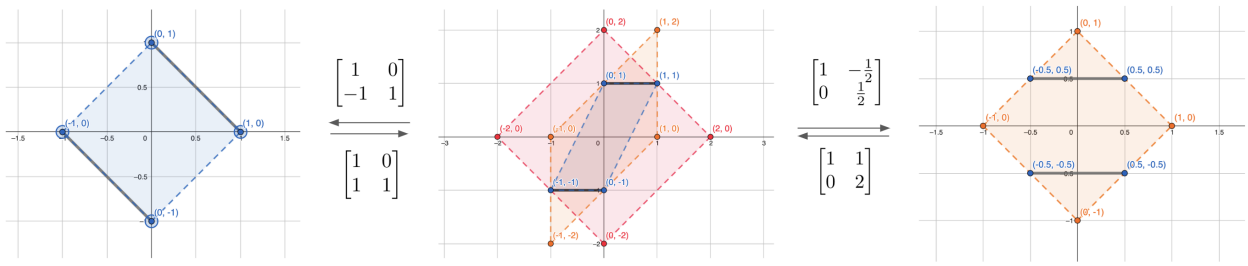


Figure 1. The Laplace mechanism applied to different linear transformations of $S(q)$. The middle plot shows the original sensitivity space, where red denotes the noise ball used by Algorithm 1; orange the ball used by Algorithm 2, and blue the ball used by Algorithm 3, which is the smallest convex shape possible. The left plot shows the transformed space used by Algorithm 3, and the right plot shows the transformed space used by Algorithm 2.

More concretely, Algorithm 3 computes the the scaled sum (denoted by s_1) and the difference of count and scaled sum (denoted by s_2). It then privatizes both s_1 and s_2 and uses them to compute mean. Next, we prove that Algorithm 3 instantiated with Laplace noise is not just intuitively better but in fact optimal for minimizing the mean squared error of the mean when ε is sufficiently small.

3.2. Analysis of Laplace noise algorithms

We first state a technical result which we use in proving upper bounds.

Lemma 3.1. *Let $b \geq 0$. Let a be such that $|a|/b \leq M$. Let $C = \frac{Z_a}{b} - \frac{aZ_b}{b^2}$ and $F = \left|\frac{2MZ_b^2}{b^2}\right| + \left|\frac{2Z_aZ_b}{b^2}\right|$. Then,*

$$\mathbb{E}\left[\left(Clip\left(\frac{a+Z_a}{b+Z_b}\right) - \frac{a}{b}\right)^2\right] \leq \mathbb{E}\left[C^2\right] + \mathbb{E}\left[F^2\right] + 2\sqrt{\mathbb{E}\left[C^2\right]\mathbb{E}\left[F^2\right]} + 4M^2 Pr(Z_b < -b/2).$$

We provide the proof in Appendix C. We next state an upper bound on the mean squared error of Algorithm 2.

Lemma 3.2. *The output of Algorithm 2 is ε-differentially private. For any dataset $D \in \mathcal{D}^*(\ell, u)$, the mean squared error of Algorithm 2 is upper bounded by*

$$\left(\frac{2(u-l)^2}{|D|^2\varepsilon^2} + \frac{8(\mu-m)^2}{|D|^2\varepsilon^2}\right)(1+o(1)) \leq \frac{4(u-l)^2}{|D|^2\varepsilon^2}(1+o(1)),$$

where $m = \frac{\ell+u}{2}$.

We provide the proof in Appendix D. We finally prove the upper bound on the mean squared error of Algorithm 3 when the noise distribution is Laplace.

Theorem 3.3. *Let Z_1 and Z_2 be sampled from independent Laplace distributions with parameter ε. Then the output of Algorithm 3 is ε-differentially private. Furthermore, for any dataset $D \in \mathcal{D}^*(\ell, u)$, the mean squared error of Algorithm 3 is upper bounded by*

$$\left(\frac{(u-l)^2}{|D|^2\varepsilon^2} + \frac{4(\mu-m)^2}{|D|^2\varepsilon^2}\right)(1+o(1)) \leq \frac{2(u-l)^2}{|D|^2\varepsilon^2}(1+o(1)),$$

where $m = \frac{\ell+u}{2}$.

Proof. Let $s = (s_1, s_2)$ be a two-dimensional vector. Let s' and s'' be vectors corresponding to two neighboring datasets. Then the ℓ_1 sensitivity is bounded by $\|s' - s''\|_1 \leq 1$.

In lines 5 and 6 of the algorithm, we add $\text{Lap}(1/\varepsilon)$ noise to each coordinate. Hence $(\hat{s}_1, \hat{s}_2)$ is an ε differentially private vector and, by the post-processing lemma, the output $\hat{\mu}$ is ε differentially private.

The proof of the error bound relies heavily on Lemma 3.1. Let $n = |D|$. Since clipping only reduces the error,

$$\mathbb{E}[(\hat{\mu} - \mu)^2] \leq (u-\ell)^2 \mathbb{E}\left[\left(\frac{\hat{s}_1}{\hat{s}_1 + \hat{s}_2} - \frac{s_1}{n}\right)^2\right].$$

Let $a = s_1$, $b = n$, $Z_a = Z_1 \sim \text{Lap}(1/\varepsilon)$, $Z_b = Z_1 + Z_2$, and $M = 1$. To usefully apply Lemma 3.1, we need bounds for $\mathbb{E}[C^2]$, $\mathbb{E}[F^2]$, and $\Pr(Z_b < -b/2)$. We bound each of these below.

$$\begin{aligned}\mathbb{E}[C^2] &= \mathbb{E}\left[\left(\frac{Z_1}{n} - \frac{(\frac{s_1}{n})(Z_1+Z_2)}{n}\right)^2\right] \\ &= \frac{1}{n^2}\mathbb{E}\left[\left(\left(\left(1-\frac{s_1}{n}\right)Z_1 - \frac{s_1}{n}Z_2\right)^2\right]\right. \\ &= \frac{1}{n^2\varepsilon^2} + \frac{4(\frac{s_1}{n} - \frac{1}{2}))^2}{n^2\varepsilon^2}. \end{aligned} \tag{6}$$

Since $(x+y)^2 \le 2x^2 + 2y^2$,

$$\mathbb{E}[F^2] \le \frac{8}{n^4}\mathbb{E}[(Z_1+Z_2)^4] + \frac{8}{n^4}\mathbb{E}[(Z_1+Z_2)^2 Z_1^2] = O\left(\frac{1}{n^4\varepsilon^4}\right) = o\left(\frac{1}{n^2\varepsilon^2}\right),$$

where the last equality follows from the moments of the Laplace distribution (Kotz et al., 2012). Finally,

$$\begin{aligned}\Pr(Z_b < -b/2) &\le \Pr(Z_1 + Z_2 < -n/2) \\ &\le \Pr(Z_1 < -n/4) + \Pr(Z_2 < -n/4) \\ &= O\left(e^{-O(n\varepsilon)}\right) = o\left(\frac{1}{n^2\varepsilon^2}\right),\end{aligned}$$

where the last equality follows from the tail bounds of the Laplace distribution. Combining the three bounds above with Lemma 3.1 and observing the fact that $\mu = \ell + \frac{s_1 w}{n}$ yields the theorem. □

Theorem 3.3 implies the following corollary, which shows that, in the high-privacy regime where ε is small, the error of the add-remove model matches the swap model (Equation 3).

Corollary 3.4.

$$R_{ar}(\varepsilon, \ell, u) \le \frac{2(u-l)^2}{\varepsilon^2}\left(1 + o(1)\right).$$

4. Generalization to all values of ε

In this section we design an optimal ε-DP mechanism for private mean estimation in the add-remove model for *any* ε, dropping the high-privacy assumption. We prove the optimality of the new mechanism with respect to $R_{\text{ar}}(\varepsilon, \ell, u)$, and show that the optimal min-max error for the add-remove model is equivalent to that for the swap model, for any ε, up to a $(1 + o(1))$ constant factor.

We first motivate the new mechanism, which we call the *hourglass* mechanism. Observe that, in Algorithm 3, $s_1 + s_2$ always sums to an integer, and hence the sensitivity space of (s_1, s_2) is just the two bold segments on the left graph in Figure 1. Previously, we used the Laplace mechanism to protect the convex hull of these segments, but this is (still) a strict superset of the actual sensitivity space, which is non-convex. Figure 2a shows how many points in the ℓ_1 unit ball actually cannot be reached by taking a single step to a neighboring database. The hourglass mechanism is designed to protect this sensitivity space more precisely, allowing for less noise. We formalize this notion and show that when the hourglass mechanism is used in Algorithm 3 to noise (s_1, s_2), the result is an optimal private mean estimator for all values of ε.

In particular, it is known that for privately computing one-dimensional sums (such as s_1 or s_2) the optimal mechanism is the staircase mechanism (Geng and Viswanath, 2014). The hourglass mechanism is constructed so that its marginal distributions both exactly match the univariate hourglass mechanism, with no change in ε.

4.1. Hourglass mechanism

The hourglass mechanism adds two-dimensional noise drawn from a distribution parameterized by $\gamma \in (0, 1]$. Its density $f_\gamma(x, y)$ is supported on $\{(x, y) : x + y = k\Delta, k \in \mathbb{Z}\}$. For $x \ge 0$, we divide each diagonal line $x + y = k\Delta$ into regions according to the value of x. For integers $i \ge 1$:

$$\begin{aligned} A_k &: x \in [0, (k+\gamma)\Delta) \\ B_k(i) &: x \in [(k+\gamma+i-1)\Delta, (k+\gamma+i)\Delta) \end{aligned}$$

Note that, for $k < 0$, A_k is always empty, and $B_k(i)$ does not contain any points with $x \ge 0$ unless $i \ge -k$. See Figure 2b for an illustration.

The density of the hourglass noise distribution is given by

$$f_\gamma(x, y) \propto \begin{cases} e^{-k\cdot\varepsilon}, & (x, y) \in A_k \\ e^{-(2i+k)\cdot\varepsilon}, & (x, y) \in B_k(i) \end{cases}. \tag{7}$$

For $x < 0$, $f_\gamma(x, y) = f_\gamma(-x, -y)$. See Figure 2c for an illustration of the density when $\Delta = 1$ and $\gamma = 0.4$.

Note that the hourglass mechanism is different from the natural extension of the univariate staircase mechanism to two dimensions as proposed by Geng et al. (2015). The latter is known to be optimal for an ℓ_1-ball sensitivity space, but does not have the properties shown in Lemma 4.2, which are key to the optimality of the hourglass mechanism, and does not perform as well in practice (see Section 6).

Theorem 4.1. *Let $q : \mathcal{D} \to R^2$ be a query such that for any two neighboring datasets D and D',*

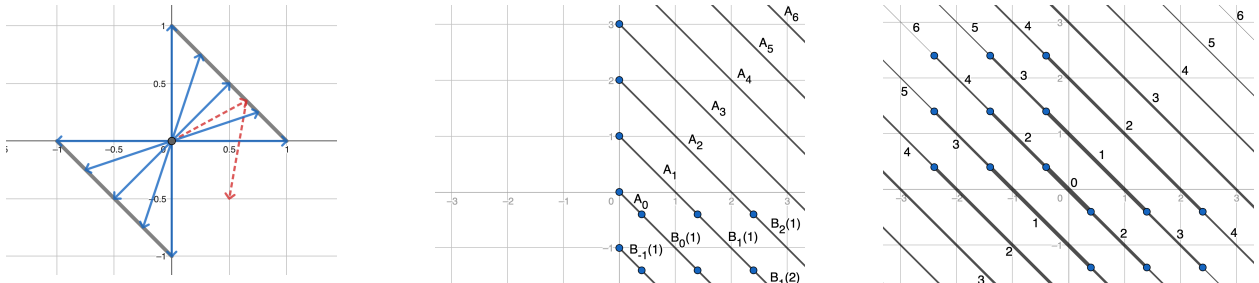
$$q(D) - q(D') \in \{(x_0, \Delta - x_0) : x_0 \in [0, \Delta]\}.$$

Then the hourglass mechanism given by

$$q(D) + (Z_1, Z_2),$$

where Z_1, Z_2 are sampled according to the density f_γ, is ε-DP.

Proof. We provide an intuition based on Figure 2c here and a detailed proof in Appendix E.2. Recall that a segment marked by integer j in Figure 2c has density proportional to $e^{-j\varepsilon}$, and on neighboring databases the noise distribution is effectively shifted by $(x_0, 1 - x_0)$ for some $x_0 \in [0, 1]$. Thus, to ensure privacy, it must be the case that the integers marking any pair of points (x, y) and $(x + x_0, y + 1 - x_0)$, which are on adjacent lines in Figure 2c and not more than one unit apart along either axis, differ by at most one. By construction, this is true across the support of f_γ. □



(a) The "hourglass" sensitivity space. Blue solid arrows show how, when moving to a neighboring database, the origin can shift only to points along the line segments $(0, 1)$ to $(1, 0)$ and $(0, -1)$ to $(-1, 0)$. Red dashed arrows show how points like $(0.5, -0.5)$ are only reachable via a chain of two neighboring steps, even though though they lie within the ℓ_1 unit ball.

(b) Illustration of the line segments A_k and $B_k(i)$ used to define the hourglass distribution for $\Delta = 1$ and $\gamma = 0.4$. Note that the segments are only active for $x \geq 0$.

(c) The hourglass density for $\Delta = 1$ and $\gamma = 0.4$. Bolder segments have higher density; in particular, points on a line segment marked j have density proportional to $e^{-j\varepsilon}$.

Figure 2. The hourglass mechanism.

We now show that the marginal distributions of the hourglass mechanism are univariate staircase mechanisms, which is the key ingredient to proving optimal utility guarantees.

Lemma 4.2 (Marginal distribution)**.** *The marginal densities $f_\gamma(x)$ and $f_\gamma(y)$ match the ε-DP univariate staircase mechanism with parameter γ. That is, $f_\gamma(x) \propto e^{-\varepsilon\lceil|x|-\gamma\rceil}$, and $f_\gamma(y) \propto e^{-\varepsilon\lceil|y|-\gamma\rceil}$.*

Proof. Assume $x \geq 0$, and let $j \geq 0$ be the unique integer for which $x \in [j-1+\gamma, j+\gamma)$. Then the marginal density for x is

$$f_\gamma(x) \propto \sum_{k=-\infty}^{\infty} f_\gamma(x, k-x) \tag{8}$$

$$= \sum_{k=-\infty}^{j-1} e^{-(2j-k)\varepsilon} + \sum_{k=j}^{\infty} e^{-k\varepsilon} \tag{9}$$

The first term follows because $j - k \geq 1$ in this range, and so $x \in B_k(j-k)$. The second term follows because $x \in A_k$ in this range. Summing the geometric series, the above is equal to

$$e^{-(j+1)\varepsilon}/(1-e^{-\varepsilon}) + e^{-j\varepsilon}/(1-e^{-\varepsilon}) \propto e^{-j\varepsilon}\ ,$$

which is precisely the staircase density. The densities for $x < 0$ and y are handled symmetrically. □

Next we derive the conditional distribution of y given x.

Lemma 4.3 (Conditional distribution)**.** *The conditional probability under f_γ of y given a fixed x is a geometric distribution with ratio $e^{-\varepsilon}$. In particular, for any y such that $y + x$ is an integer,*

$$f_\gamma(Y = y|x) = \frac{1-e^{-\varepsilon}}{1+e^{-\varepsilon}} e^{-\varepsilon|y-y_0(x)|},$$

where

$$y_0(x) = \begin{cases} -x + \lfloor x + (1-\gamma)\rfloor, & x \geq 0 \\ -x - \lfloor -x + (1-\gamma)\rfloor, & x < 0 \end{cases} \tag{10}$$

Using the above lemmas, we get a simple sampling algorithm for the hourglass distribution: first sample x from a univariate staircase mechanism with γ and ε, and then sample y from the geometric distribution in Lemma 4.3.

4.2. Implications for mean estimation

Before proceeding to analyze the accuracy of the hourglass mechanism for mean estimation, we first consider the two-dimensional staircase mechanism.

Lemma 4.4. *Let Z_1 and Z_2 be sampled by the two-dimensional staircase mechanism with parameter ε and with parameter γ optimized for mean squared error. Then the output of Algorithm 3 is ε-DP. Furthermore, for any dataset $D \in \mathcal{D}^*(\ell, u)$, the MSE of Algorithm 3 is upper bounded by*

$$\left(\frac{(u-l)^2\widetilde{\sigma}^2(\varepsilon)}{|D|^2}\right)(1+o(1))\,,$$

where $\widetilde{\sigma}^2(\varepsilon)$ is the variance of the two-dimensional staircase mechanism optimized for MSE.

The proof is provided in Appendix F. We now state our main upper bound for all values of ε. The proof is similar to that of Lemma 4.4 together with the fact that the marginal distribution along each dimension is the same as the univariate staircase mechanism (Lemma 4.2). We provide the full proof in Appendix G.

Theorem 4.5. *Let Z_1 and Z_2 be sampled by the hourglass mechanism with parameter ε and γ as given by Geng and Viswanath (2014, Equation 50). Then the output of Algorithm 3 is ε-differentially private. Furthermore, for any dataset $D \in \mathcal{D}^*(\ell, u)$, the mean squared error of Algorithm 3 is upper bounded by*

$$\left(\frac{(u-l)^2\sigma^2(\varepsilon)}{|D|^2}\right)(1+o(1)),$$

where $\sigma^2(\varepsilon)$ is defined in (2).

5. Lower bound

We next state a well-known result of Geng and Viswanath (2014) that we will use to prove our lower bound.

Lemma 5.1 (Geng and Viswanath (2014, Section VI.C))**.** *Let $k > 0$ and let $\mathcal{D}^k$ denote the collection of all datasets of size at most k where each value lies in the range $[0, 1]$. For $D \in \mathcal{D}^k$, let $S(D)$ denote the sum of the elements in D. Let $\hat{S}$ denote an estimator of the same sum. Then, for any ε,*

$$\inf_{\hat{S}\in\mathcal{A}_\varepsilon^{ar}} \sup_{D\in\mathcal{D}^k} \mathbb{E}\left[(\hat{S}-S(D))^2\right] \geq \sigma^2(\varepsilon)(1-o(1)),$$

where $\sigma^2(\varepsilon)$ is given by (2). Here the $o(1)$ term goes to zero as k tends to infinity.

Lemma 5.1 provides a lower bound for estimating the sum. We construct a class of datasets $\mathcal{D}'$ with varying size such that for every dataset in $\mathcal{D}'$ there exists a corresponding dataset in $\mathcal{D}^k$ (as defined in Lemma 5.1) with the same sum. Hence, any differentially private mean estimator for the datasets in $\mathcal{D}'$ can be modified to obtain a differentially private sum estimator for the datasets in $\mathcal{D}^k$. We use this observation to show the following lower bound on the minmax MSE of mean estimation under the add-remove model of differential privacy.

Theorem 5.2.

$$R_{ar}(\varepsilon, \ell, u) \geq (u-\ell)^2 \cdot \sigma^2(\varepsilon)(1-o(1)).$$

Combining Theorem 4.5 together with Theorem 5.2 and (2) yields the following result.

Corollary 5.3. *For any $\varepsilon > 0$, $R_{ar}(\varepsilon, \ell, u)$ is equal to*

$$\frac{2^{-2/3}e^{-2\varepsilon/3}(1+e^{-\varepsilon})^{2/3}+e^{-\varepsilon}}{(1-e^{-\varepsilon})^2}(1\pm o(1)).$$

Furthermore, the hourglass mechanism with suitable value of γ achieves this normalized MSE.

6. Experiments

In Figure 3 we plot the empirical performance of the algorithms discussed in Sections 3 and 4.1 on synthetic datasets and explore how the performance changes with parameters such as the privacy budget ε and the true mean μ. The underlying datasets are generated i.i.d. with varying μ in the range $[\ell = 0, u = 1]$. All datasets have $10,000$ points, and mean squared error is computed over $100,000$ runs of each algorithm. The errors are normalized by $\frac{|D|^2\varepsilon^2}{2}$ to keep them in a similar range across a wide range of parameters.

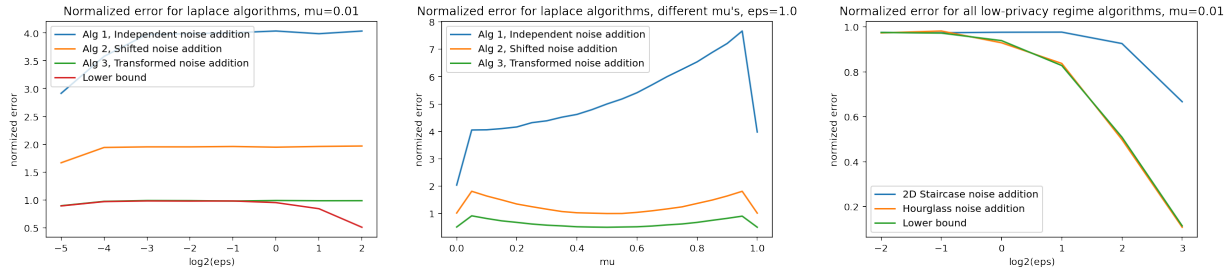
Figures 3a and 3b compare the performance of Algorithms 1, 2 and 3 using the Laplace mechanism, varying ε and μ, respectively. These plots focus on the high-privacy regime, where Algorithm 3 was shown to be optimal. Indeed, Algorithm 3 outperforms the others, reducing error over Algorithms 2 by roughly a factor of two, which matches the ratio of upper bounds in Lemma 3.2 and Theorem 3.3. Similarly, in Figure 3a, Algorithm 3 closely matches the lower bound $\sigma^2(\varepsilon)$ from (2) when $\varepsilon < 1$.

In Figure 3b, we explore how the error changes with the true mean of the data. As analyzed in (6), the error is largest when μ approaches 0 or 1. However, the error also drops when μ becomes very close to 0 or 1, since $\frac{\hat{s}_1}{\hat{n}}$ falls out of the range $[\ell, u]$ with probability approaching 50%. In these cases the clipping operation significantly reduces the mean squared error.

Finally, we compare the two-dimensional staircase mechanism to Algorithm 3 with noise from the hourglass mechanism in Figure 3c, focusing on the low-privacy regime where ε is large. The lower bound in Figure 3c is again $\sigma^2(\varepsilon)$, but notice now that as ε grows, the hourglass mechanism continues to match the lower-bound indefinitely, and increasingly outperforms the two-dimensional staircase mechanism. In Figure 4 (Appendix I), we show that both mechanisms also have a similar M-shaped error curve over μ. In Appendix I, we provide additional experiments to demonstrate that the proposed algorithm outperforms existing algorithms on small datasets.

Impact statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.



(a) All Laplace mechanisms, high privacy regime, varying ε.

(b) All Laplace mechanisms, high privacy regime, varying μ.

(c) Hourglass and 2D staircase mechanisms, low privacy regime, varying ε.

Figure 3. Error comparison of different algorithms with varying ε or μ.

Acknowledgements

The authors thank Travis Dick, Gautam Kamath, and Ziteng Sun for helpful comments and suggestions.

References

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

K. Amin, J. Gillenwater, M. Joseph, A. Kulesza, and S. Vassilvitskii. Plume: differential privacy at scale. *arXiv preprint arXiv:2201.11603*, 2022.

J. Awan and A. Slavković. Structure and sensitivity in differential privacy: Comparing k-norm mechanisms. *Journal of the American Statistical Association*, 116(534):935–954, 2021.

T. Dick, A. Kulesza, Z. Sun, and A. T. Suresh. Subset-based instance optimality in private estimation. In *International Conference on Machine Learning*, pages 7992–8014. PMLR, 2023.

C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pages 265–284. Springer, 2006.

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

Q. Geng and P. Viswanath. The optimal mechanism in differential privacy. In *2014 IEEE international symposium on information theory*, pages 2371–2375. IEEE, 2014.

Q. Geng, P. Kairouz, S. Oh, and P. Viswanath. The staircase mechanism in differential privacy. *IEEE Journal of Selected Topics in Signal Processing*, 9(7):1176–1184, 2015.

M. Hardt and K. Talwar. On the geometry of differential privacy. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 705–714, 2010.

Z. Huang, Y. Liang, and K. Yi. Instance-optimal mean estimation under differential privacy. *Advances in Neural Information Processing Systems*, 34:25993–26004, 2021.

G. Kamath, V. Singhal, and J. Ullman. Private mean estimation of heavy-tailed distributions. In J. Abernethy and S. Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2204–2235. PMLR, 09–12 Jul 2020.

G. Kamath, A. Mouzakis, M. Regehr, V. Singhal, T. Steinke, and J. Ullman. A bias-variance-privacy trilemma for statistical estimation. *arXiv preprint arXiv:2301.13334*, 2023.

S. Kotz, T. Kozubowski, and K. Podgorski. *The Laplace distribution and generalizations: a revisit with applications to communications, economics, engineering, and finance*. Springer Science & Business Media, 2012.

X. Liu, W. Kong, S. Kakade, and S. Oh. Robust and differentially private mean estimation. *Advances in neural information processing systems*, 34:3887–3901, 2021.

F. D. McSherry. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, pages 19–30, 2009.

R. Rogers, S. Subramaniam, S. Peng, D. Durfee, S. Lee, S. K. Kancha, S. Sahay, and P. Ahammad. Linkedin's audience engagements API: A privacy preserving data analytics system at scale. *arXiv preprint arXiv:2002.05839*, 2020.

S. Takagi, F. Kato, Y. Cao, and M. Yoshikawa. From bounded to unbounded: Privacy amplification via shuffling with dummies. In *2023 IEEE 36th Computer Se-*

curity Foundations Symposium (CSF), pages 457–472. IEEE, 2023.

F. Tramèr, Z. Huang, J.-P. Hubaux, and E. Ayday. Differential privacy with bounded priors: reconciling utility and privacy in genome-wide association studies. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1286–1297, 2015.

S. Vadhan. The complexity of differential privacy. *Tutorials on the Foundations of Cryptography: Dedicated to Oded Goldreich*, pages 347–450, 2017.

R. J. Wilson, C. Y. Zhang, W. Lam, D. Desfontaines, D. Simmons-Marengo, and B. Gipson. Differentially private SQL with bounded user contribution. *arXiv preprint arXiv:1909.01917*, 2019.

A. Properties of staircase mechanisms

We first define both the one-dimensional and two-dimensional staircase mechanisms.

Definition A.1 (Univariate staircase mechanism)**.** Let Δ be the sensitivity of the underlying query. The univariate staircase mechanism (Geng and Viswanath, 2014) is parameterized by $\gamma \in [0, 1]$ and is given by

$$f_\gamma(x) = \begin{cases} a(\gamma), & x \in [0, \gamma\Delta) \\ a(\gamma)e^{-\varepsilon}, & x \in [\gamma\Delta, \Delta) \\ e^{-k\varepsilon} f_\gamma(x - k\Delta), & x \in [k\Delta, (k+1)\Delta) \\ f_\gamma(-x), & x < 0. \end{cases} \tag{11}$$

Here $a(\gamma)$ is the normalization factor and given by

$$a(\gamma) = \frac{1 - e^{-\varepsilon}}{2\Delta\left(\gamma + e^{-\varepsilon}(1-\gamma)\right)}.$$

Definition A.2 (Two-dimensional staircase mechanism)**.** Let Δ be the sensitivity of the underlying query. The two-dimensional staircase mechanism (Geng et al., 2015) is parameterized by $\gamma \in [0, 1]$ and is given by

$$f_\gamma(x, y) = \begin{cases} a(\gamma)e^{-\varepsilon k}, & |x| + |y| \in [k\Delta, (k+\gamma)\Delta) \\ a(\gamma)e^{-\varepsilon(k+1)}, & |x| + |y| \in [(k+\gamma)\Delta + k + 1)\Delta), \end{cases} \tag{12}$$

Here $a(\gamma)$ is the normalization factor and given by

$$a(\gamma) = \frac{(1 - e^{-\varepsilon})^2}{2\left(e^{-\varepsilon}(e^{-\varepsilon} + (1 - e^{-\varepsilon})\gamma) + (1 - e^{-\varepsilon})(e^{-\varepsilon} + (1 - e^{-\varepsilon})\gamma^2)\right)}.$$

We prove the following result for the MSE of the two-dimensional staircase mechanism.

Lemma A.3. *Let $\varepsilon \geq \varepsilon_0$ for a sufficiently large value of ε_0. Then, for the two-dimensional stair-case mechanism,*

$$\min_\gamma \mathbb{E}_{f_\gamma}[x^2] = \min_\gamma \mathbb{E}_{f_\gamma}[y^2] = \Delta^2 \cdot \Theta\left(e^{-\varepsilon/2}\right),$$

Proof. Without loss of generality, we assume the sensitivity to be one. By symmetry,

$$\min_\gamma \mathbb{E}_{f_\gamma}[x^2] = \min_\gamma \mathbb{E}_{f_\gamma}[y^2],$$

and it suffices to consider $\mathbb{E}_{f_\gamma}[x^2]$. For $k \geq 1$, it can be shown that

$$\int_{|x|+|y| \in [k, k+1)} x^2 dxdy = \Theta\left(k^3\right).$$

Similarly,

$$\int_{|x|+|y| \in [\gamma, 1)} x^2 dxdy = \Theta\left(1\right),$$

and

$$\int_{|x|+|y| \in [0, \gamma)} x^2 dxdy = \Theta\left(\gamma^4\right).$$

Furthermore,

$$a(\gamma) = \Theta\left(\frac{1}{e^{-\varepsilon} + \gamma^2}\right).$$

Combining the above four equations yields,

$$\begin{aligned}\mathbb{E}_{f_\gamma}[x^2] &= a(\gamma)\Theta\left(\gamma^4 + e^{-\varepsilon} + \sum_{k\geq 1}(e^{-k\varepsilon} + e^{(k+1)\varepsilon})k^3\right)\\ &= a(\gamma)\Theta\left(\gamma^4 + e^{-\varepsilon}\right)\\ &= \Theta\left(\frac{\gamma^4 + e^{-\varepsilon}}{e^{-\varepsilon} + \gamma^2}\right).\end{aligned}$$

Minimizing over γ yields the desired result. □

We next prove a result on the moments of both one-dimensional and two-dimensional staircase mechanisms.

Lemma A.4. *Let $16 \geq m \geq 0$. For both the one-dimensional and two-dimensional staircase mechanisms that guarantees ε-differential privacy for sensitivity one queries,*

$$\mathbb{E}[x^m] = \frac{1}{\varepsilon^{m-2}} \cdot O\left(\mathbb{E}[x^2]\right)$$

Proof. We provide the proof for the univariate staircase mechanism. The proof for the two-dimensional stair case mechanism is similar and omitted. Let $a(\gamma)$ denote the normalization term in the definition of univariate staircase mechanism Geng et al. (2015, equation 26). Since $\gamma \leq 1$,

$$\begin{aligned}\mathbb{E}_{f_\gamma}[x^m] &= a(\gamma)\Theta\left(\gamma^m + \sum_{k\geq 1} e^{-k\varepsilon}k^m\right)\\ &\overset{(a)}{=} a(\gamma)O\left(\gamma^2 + \sum_{k\geq 1} e^{-k\varepsilon}k^m\right)\\ &\overset{(b)}{=} a(\gamma)O\left(\gamma^2 + \frac{1}{(1-e^{-\varepsilon})^{m-2}}\sum_{k\geq 1} e^{-k\varepsilon}k^2\right)\\ &= \frac{1}{(1-e^{-\varepsilon})^{m-2}}a(\gamma)O\left(\gamma^2 + \sum_{k\geq 1} e^{-k\varepsilon}k^2\right)\\ &= \frac{1}{(1-e^{-\varepsilon})^{m-2}} \cdot O\left(\mathbb{E}_{f_\gamma}[x^2]\right)\\ &\overset{(c)}{=} \frac{1}{\varepsilon^{m-2}} \cdot O\left(\mathbb{E}_{f_\gamma}[x^2]\right),\end{aligned}$$

where (a) uses the fact that $\gamma \leq 1$, (c) follows from the fact that $1 - e^{-\varepsilon} \leq \varepsilon$. To prove (b) notice that

$$(1-e^{-\varepsilon})\sum_{k\geq 1} e^{-k\cdot\varepsilon}k^m = \sum_{k\geq 1} e^{-k\cdot\varepsilon}(k^m - (k-1)^m) = O\left(\sum_{k\geq 1} e^{-k\cdot\varepsilon}k^{m-1}\right).$$

□

We next state a concentration property of both one-dimensional and two-dimensional staircase mechanisms.

Lemma A.5. *Let $n \geq 10$. For both the one-dimensional and two-dimensional staircase mechanisms that guarantees ε-differential privacy for sensitivity one,*

$$Pr(x \leq -n) \leq e^{-n\varepsilon/2}.$$

Proof. We provide the proof for the univariate staircase mechanism. The proof for the two-dimensional stair case mechanism is similar and omitted. Let $a(\gamma)$ denote the normalization term in the definition of univariate staircase mechanism Geng et al. (2015, equation 26). Note that for the univariate staircase mechanism,

$$a(\gamma) \leq \frac{1 - e^{-\varepsilon}}{2e^{-\varepsilon}}.$$

We now bound the desired quantity as follows.

$$\begin{aligned}
\Pr(x \leq -n) &\leq \sum_{k \geq n} a(\gamma) e^{-k\varepsilon} \\
&= a(\gamma) \frac{e^{-n\varepsilon}}{1 - e^{-\varepsilon}} \\
&\leq (1 - e^{-\varepsilon}) e^{\varepsilon} \frac{e^{-n\varepsilon}}{1 - e^{-\varepsilon}} \\
&= e^{-(n-1)\varepsilon},
\end{aligned}$$

where the last inequality follows from the bound on $a(\gamma)$. □

B. Viewing Algorithm 2 via the lens of linear transformation

In Section 3, we have claimed that Algorithm 2 is equivalent to the following procedure:

1. Apply a linear transformation given by the matrix $\begin{bmatrix} 1 & -\frac{1}{2} \\ 0 & \frac{1}{2} \end{bmatrix}$ to the sum and count vector $q(D)$.
2. Add vector Laplace noise to the transformed vector according to its (reduced) sensitivity ($a = 1$).
3. Reverse the transformation by applying the inverse matrix $\begin{bmatrix} 1 & 1 \\ 0 & 2 \end{bmatrix}$, then divide to estimate the mean and truncate as before.

Here we prove this claim by showing that both the procedures sample a two-dimensional vector (x, y) that they use to noise the sum and count is sampled from the same joint distribution. Since the private mean is computed by taking the division of noisy sum over noisy count, the resulting private mean has the same distribution.

Proof. Without loss of generality, we only have to prove it for $[\ell, u] = [0, 1]$. The proof easily extends to general range bound $[\ell, u]$.

Given D, let $s_0 = \sum_{x \in D} x$ denote the true sum of the values in D, and let n denote the true size. Algorithm 2 first computes $s = s_0 - \frac{n}{2}$ and then adds noise $Z_s \sim \text{Lap}(\frac{1}{\varepsilon})$ to obtain $\hat{s} = s_0 - \frac{n}{2} + Z_s$. It also computes $\hat{n} = n + Z_n$, where $Z_n \sim \text{Lap}(\frac{2}{\varepsilon})$. Finally, before clipping, it computes the noisy mean as:

$$\begin{aligned}
\frac{\hat{s}}{\hat{n}} + \frac{1}{2} &= \frac{s_0 - \frac{n}{2} + Z_s}{n + Z_n} + \frac{1}{2} \\
&= \frac{s_0 - \frac{n}{2} + Z_s + \frac{n}{2} + \frac{1}{2} Z_n}{n + Z_n} \\
&= \frac{s_0 + Z_s + \frac{1}{2} Z_n}{n + Z_n}
\end{aligned}$$

On the other hand, for the algorithm using matrix transformation:

1. Step 1, computes the vector $(s_0 - \frac{n}{2}, \frac{n}{2})$.

2. Step 2 adds random vector noise (Z_1, Z_2, where $Z_1 \sim \text{Lap}(\frac{1}{\varepsilon})$, $Z_2 \sim Lap(\frac{1}{\varepsilon})$).
3. Step 3 applies the inverse transformation to obtain $((s_0 - \frac{n}{2} + Z_1) + (\frac{n}{2} + Z_2), n + 2Z_2) = (s_0 + Z_1 + Z_2, n + 2Z_2)$.

The private mean is computed by taking the ratio of noisy sum and noisy count and applying clipping.

Since Z_1 has the same distribution as Z_s, Z_2 has the same distribution as $\frac{1}{2}Z_n$, this yields the same result as Algorithm 2. □

C. Proof of Lemma 3.1

We first focus on the upper bound.

$$\begin{aligned}\left(\text{Clip}\left(\frac{a+Z_a}{b+Z_b}\right) - \frac{a}{b}\right)^2 &= \left(\text{Clip}\left(\frac{a+Z_a}{b+Z_b}\right) - \frac{a}{b}\right)^2 1_{Z_b \geq -b/2} + \left(\text{Clip}\left(\frac{a+Z_a}{b+Z_b}\right) - \frac{a}{b}\right)^2 1_{Z_b < -b/2} \\ &\leq \left(\text{Clip}\left(\frac{a+Z_a}{b+Z_b}\right) - \frac{a}{b}\right)^2 1_{Z_b \geq -b/2} + 4M^2 1_{Z_b < -b/2} \\ &\leq \left(\text{Clip}\left(\frac{a+Z_a}{b+Z_b}\right) - \frac{a}{b}\right)^2 1_{Z_b \geq -b/2} + 4M^2 1_{Z_b < -b/2},\end{aligned}$$

where the first inequality uses the fact that both $\text{Clip}\left(\frac{a+Z_a}{b+Z_b}\right)$ and $\frac{a}{b}$ lie in $[0,1]$ and the last inequality uses the fact that clipping is a projection operator. Taking expectation on both sides yield,

$$\mathbb{E}\left[\left(\text{Clip}\left(\frac{a+Z_a}{b+Z_b}\right) - \frac{a}{b}\right)^2\right] \leq \mathbb{E}\left[\left(\frac{a+Z_a}{b+Z_b} - \frac{a}{b}\right)^2 1_{Z_b \geq -b/2}\right] + 4M^2 \text{Pr}(Z_b < -b/2).$$

We now use algebraic manipulation to simplify $\frac{a+Z_a}{b+Z_b} - \frac{a}{b}$.

$$\begin{aligned}\frac{a+Z_a}{b+Z_b} - \frac{a}{b} &= \frac{Z_a}{b} + \frac{a+Z_a}{b+Z_b} - \frac{a+Z_a}{b} \\ &= \frac{Z_a}{b} - \frac{(a+Z_a)(Z_b)}{(b+Z_b)(b)} \\ &= \frac{Z_a}{b} - \frac{aZ_b}{(b+Z_b)(b)} - \frac{Z_aZ_b}{(b+Z_b)(b)} \\ &= \frac{Z_a}{b} - \frac{aZ_b}{b^2} + \frac{aZ_b}{b^2} - \frac{aZ_b}{(b+Z_b)(b)} - \frac{Z_aZ_b}{(b+Z_b)(b)} \\ &= \frac{Z_a}{b} - \frac{aZ_b}{b^2} + \frac{aZ_b^2}{b^2(b+Z_b)} - \frac{Z_aZ_b}{(b+Z_b)(b)}.\end{aligned}$$

Let $C = \frac{Z_a}{b} - \frac{aZ_b}{b^2}$ and $D = \frac{aZ_b^2}{b^2(b+Z_b)} - \frac{Z_aZ_b}{(b+Z_b)(b)}$.

$$\begin{aligned}\mathbb{E}\left[\left(\frac{a+Z_a}{b+Z_b} - \frac{a}{b}\right)^2 1_{Z_b \geq -b/2}\right] &= \mathbb{E}\left[(C+D)^2 1_{Z_b \geq -b/2}\right] \\ &= \mathbb{E}\left[C^2 1_{Z_b \geq -b/2}\right] + \mathbb{E}\left[D^2 1_{Z_b \geq -b/2}\right] + \mathbb{E}\left[2CD 1_{Z_b \geq -b/2}\right] \\ &\leq \mathbb{E}\left[C^2 1_{Z_b \geq -b/2}\right] + \mathbb{E}\left[D^2 1_{Z_b \geq -b/2}\right] + 2\sqrt{\mathbb{E}\left[C^2\right]\mathbb{E}\left[D^2 1_{Z_b \geq -b/2}\right]},\end{aligned}$$

where the last inequality uses Cauchy-Schwarz inequality. We next upper bound D. If $Z_b \geq -b/2$, and $|a| \leq bM$, then

$$\begin{aligned}|D| = \left|\frac{aZ_b^2}{b^2(b+Z_b)} - \frac{Z_aZ_b}{(b+Z_b)(b)}\right| &\leq \left|\frac{aZ_b^2}{b^2(b+Z_b)}\right| + \left|\frac{Z_aZ_b}{(b+Z_b)(b)}\right| \\ &\leq \left|\frac{MZ_b^2}{b(b+Z_b)}\right| + \left|\frac{Z_aZ_b}{(b+Z_b)(b)}\right| \\ &\leq \left|\frac{2MZ_b^2}{b^2}\right| + \left|\frac{2Z_aZ_b}{b^2}\right|.\end{aligned}$$

Let $F = \left|\frac{2MZ_b^2}{b^2}\right| + \left|\frac{2Z_aZ_b}{b^2}\right|$. Combing the above bound with previous equations yields the upper bound:

$$\begin{aligned}\mathbb{E}\left[\left(\frac{a+Z_a}{b+Z_b} - \frac{a}{b}\right)^2 1_{Z_b \geq -b/2}\right] &\leq \mathbb{E}\left[C^2 1_{Z_b \geq -b/2}\right] + \mathbb{E}\left[D^2 1_{Z_b \geq -b/2}\right] + 2\sqrt{\mathbb{E}\left[C^2\right]\mathbb{E}\left[D^2 1_{Z_b \geq -b/2}\right]} \\ &\leq \mathbb{E}\left[C^2 1_{Z_b \geq -b/2}\right] + \mathbb{E}\left[F^2 1_{Z_b \geq -b/2}\right] + 2\sqrt{\mathbb{E}\left[C^2\right]\mathbb{E}\left[F^2 1_{Z_b \geq -b/2}\right]} \\ &\leq \mathbb{E}\left[C^2\right] + \mathbb{E}\left[F^2\right] + 2\sqrt{\mathbb{E}\left[C^2\right]\mathbb{E}\left[F^2\right]}.\end{aligned}$$

D. Proof of Lemma 3.2

The differential privacy guarantee follows by the properties of Laplace mechanism and composition theorem and in the rest of the proof, we focus on the MSE guarantees. The analysis of MSE heavily relies in Lemma 3.1. Let $n = |D|$, $a = s$, $b = n$, $Z_a = Z_s \sim \text{Lap}(w/\varepsilon)$, $Z_b = Z_n = \text{Lap}(2/\varepsilon)$, and $M = (u - \ell)/2$. With these definitions, to apply Lemma 3.1, we need to bound $\mathbb{E}[C^2]$, $\mathbb{E}[F^2]$, and $\Pr(Z_b < -b/2)$. We bound each of the terms below.

$$\begin{aligned}\mathbb{E}[C^2] &= \mathbb{E}\left[\left(\frac{Z_s}{n} - \frac{(\mu - m)Z_n}{n}\right)^2\right] \\ &= \frac{2(u-l)^2}{n^2\varepsilon^2} + \frac{8(\mu - m)^2}{n^2\varepsilon^2}.\end{aligned}$$

Since $(x+y)^2 \leq 2x^2 + 2y^2$,

$$\begin{aligned}\mathbb{E}[F^2] &\leq 8\frac{(u-\ell)^2}{n^4}\mathbb{E}\left[Z_n^4\right] + 8\frac{1}{n^4}\mathbb{E}\left[Z_a^2 Z_b^2\right] \\ &= o\left(\frac{(u-l)^2}{n^2\varepsilon^2}\right).\end{aligned}$$

Finally, by the tail bounds of the Laplace mechanism,

$$\begin{aligned}M^2\Pr(Z_b < -b/2) &\leq (u-\ell)^2\Pr(Z_n < -n/2) \\ &= o\left(\frac{(u-l)^2}{n^2\varepsilon^2}\right).\end{aligned}$$

Combining the above three equations together with Lemma 3.1 yields the lemma.

E. Analysis of the hourglass mechanism

E.1. Computing normalization constant

Let $c(\gamma)$ be the normalizing factor that ensures the sum of the probabilities is one.

From the paper Geng and Viswanath (2014), we know the normalizing factor for the univariate staircase mechanism to be

$$a(\gamma) \triangleq \frac{1 - e^{-\varepsilon}}{2(\gamma +^{-\varepsilon} (1-\gamma))}.$$

From Lemma 4.2, the marginal of the hourglass distribution is the univariate staircase distribution. hence,

$$\sum_y f_\gamma(x, y) = a(\gamma),$$

where $a(\gamma)$ is the normalization factor in the univariate staircase distribution. Observe that if $x = 0$, then for all integer y, $(x, y) \in A_y$ and hence

$$\sum_y f_\gamma(x, y) = \sum_y c(\gamma)e^{-\varepsilon y} = c(\gamma)\frac{1 - e^{-\varepsilon}}{1 + e^{-\varepsilon}}.$$

Hence,

$$c(\gamma) = \frac{1 - e^{-\varepsilon}}{1 + e^{-\varepsilon}}a(\gamma) = \frac{(1 - e^{-\varepsilon})^2}{2\Delta(1 + e^{-\varepsilon})(\gamma + e^{-\varepsilon}(1-\gamma))}.$$

E.2. Proof of Theorem 4.1

Let D and D' be two neighboring datasets such that $q(D) - Q(D_0) = -(x_0, 1 - x_0)$. Let let $\tilde{q}(D)$ denote the output of the hourglass mechanism. To provide differential privacy guarantee, it suffices to prove upper and lower bounds for

$$\frac{\Pr(\tilde{q}(D) = (x, y))}{\Pr(\tilde{q}(D') = (x, y))}.$$

Let $\bar{Z}$ be a sample from the hourglass mechanism, then

$$\begin{aligned}
\frac{\Pr(\tilde{q}(D) = (x, y))}{\Pr(\tilde{q}(D') = (x, y))} &= \frac{\Pr(q(D) + \bar{Z} = (x, y))}{\Pr(q(D') + \bar{Z} = (x, y))} \\
&= \frac{\Pr(q(D') - (x_0, 1 - x_0) + \bar{Z} = (x, y))}{\Pr(q(D') + \bar{Z} = (x, y))} \\
&= \frac{\Pr(\bar{Z} = (x, y) - q(D') + (x_0, 1 - x_0))}{\Pr(\bar{Z} = (x, y) - q(D'))} \\
&= \frac{\Pr(\bar{Z} = (x', y') + (x_0, 1 - x_0))}{\Pr(\bar{Z} = (x', y')} \\
&= \frac{f_\gamma(x' + x_0, y' + 1 - x_0))}{f_\gamma(x', y')},
\end{aligned}$$

where $(x', y') = (x, y) - q(D')$. Hence, to prove the privacy guarantee, it suffices to prove upper and lower bounds on the ratio,

$$R(x, y, x_0) \triangleq \frac{f_\gamma(x + x_0, y + 1 - x_0))}{f_\gamma(x, y)},$$

for all x, y and $x_0 \in [0, 1]$. Without loss of generality, we assume that $\Delta = 1$. Let $\mathcal{S} = \{A_k, B_k(i) : \forall k, i\}$. Note that $\mathcal{S}$ partitions the domain of f_γ into disjoint partitions. We observe that if $(x, y) \in A_{-1}$, then $f_\gamma(x, y) = c(\gamma)e^{-\varepsilon}$ and for all x_0, $f_\gamma(x + x_0, y + 1 - x_0) \in \{c(\gamma), c(\gamma)e^{-2\varepsilon}\}$, hence if $(x, y) \in A_{-1}$ then $R(x, y, x_0) \in \{e^{-\varepsilon}, e^{\varepsilon}\}$ for all x_0. If $(x, y) \notin A_{-1}$, then $(x + x_0, y + 1 - x_0)$ belongs to at most two sets in $\mathcal{S}$ and hence $f_\gamma(x + x_0, y + 1 - x_0)$ is monotonic in x_0 and proving the result for $x_0 \in \{0, 1\}$ suffices.

We now focus on the case when $x_0 \in \{0, 1\}$. Of these two cases, by symmetry it suffices to consider $x_0 = 0$. Furthermore, we show the result when $x + y = k$ for some $x \geq 0$. The proof for the other side is similar and omitted. We prove the result by dividing the problem into subcases depending the value of (x, y). Subcase (A): If $(x, y) \in A_k$ for some k, then $(x, y + 1) \in A_{k+1}$ and hence $R(x, y, 0) = e^{-\varepsilon}$. Subcase $(B1)$: If $(x, y) \in B_k(1)$, then $(x, y) \in A_{k+1}$ and hence $R(x, y, 0) = e^{\varepsilon}$. Subcase $(B2)$: If $(x, y) \in B_k(i)$ for $i \geq 2$, then $(x, y) \in B_{k+1}(i - 1)$ and hence $R(x, y, 0) = e^{\varepsilon}$. Hence, we have shown that for all x, y and $x_0 \in \{0, 1\}$,

$$R(x, y, x_0) \in \{e^{-\varepsilon}, e^{\varepsilon}\}.$$

F. Proof of Lemma 4.4

The privacy guarantee is similar to that of Theorem 3.3 and is omitted. As before, the proof of utility heavily relies in Lemma 3.1. Let $n = |D|$. Observe that

$$\mathbb{E}[(\hat{\mu} - \mu)^2] \leq (u - \ell)^2 \mathbb{E}\left[\left(\frac{\hat{s}_1}{\hat{s}_1 + \hat{s}_2} - \frac{s_1}{n}\right)^2\right].$$

Let $a = s_1$, $b = n$, $Z_a = Z_1$, $Z_b = Z_1 + Z_2$, where Z_1, Z_2 are from the two-dimensional staircase mechanism. Let $M = 1$. With these definitions, to apply Lemma 3.1, we need to bound $\mathbb{E}[C^2]$, $\mathbb{E}[F^2]$, and $\Pr(Z_b < -b/2)$. We bound each of the

terms below. Let $\alpha = \frac{s_1}{n}$.

$$\begin{aligned}
\mathbb{E}[C^2] &= \mathbb{E}\left[\left(\frac{Z_1}{n} - \frac{(\frac{s_1}{n})(Z_1+Z_2)}{n}\right)^2\right] \\
&= \frac{1}{n^2}\mathbb{E}\left[(1-\alpha)^2 Z_1^2 + \alpha^2 Z_2^2 - 2(1-\alpha)\alpha Z_1, Z_2\right] \\
&\leq \frac{1}{n^2}\mathbb{E}\left[(1-\alpha)^2 Z_1^2 + \alpha^2 Z_2^2\right] + 2(1-\alpha)\alpha\sqrt{\mathbb{E}\left[Z_1^2, Z_2^2\right]} \\
&\leq \frac{1}{n^2}(1-\alpha)^2\widetilde{\sigma}^2(\varepsilon) + \alpha^2\widetilde{\sigma}^2(\varepsilon) + 2(1-\alpha)\alpha\widetilde{\sigma}^2(\varepsilon) \\
&= \frac{\widetilde{\sigma}^2(\varepsilon)}{n^2},
\end{aligned}$$

where the first inequality follows by the Cauchy-Schwarz inequality. Since $(x+y)^2 \leq 2x^2 + 2y^2$,

$$\begin{aligned}
\mathbb{E}[F^2] &\leq \frac{8}{n^4}\mathbb{E}[(Z_1+Z_2)^4] + \frac{8}{n^4}\mathbb{E}[(Z_1+Z_2)^2 Z_1^2] \\
&= o\left(\frac{\widetilde{\sigma}^2(\varepsilon)}{n^2}\right),
\end{aligned}$$

where the last equality follows from Lemma A.4. Finally,

$$\begin{aligned}
\Pr(Z_b < -b/2) &\leq \Pr(Z_1 + Z_2 < -n/2) \\
&\leq \Pr(Z_1 < -n/4) + \Pr(Z_2 < -n/4) \\
&= o\left(\frac{\widetilde{\sigma}^2(\varepsilon)}{n^2}\right),
\end{aligned}$$

where the last equality follows from Lemma A.5. Combining the above three equations together with Lemma 3.1 and observing the fact that $\mu = \ell + \frac{s_1 w}{n}$ yields the result.

G. Proof of Theorem 4.5

The privacy guarantee is similar to that of Theorem 3.3 and is omitted. As before, the proof of utility heavily relies in Lemma 3.1. Let $n = |D|$. Observe that

$$\mathbb{E}[(\hat{\mu} - \mu)^2] \leq (u-\ell)^2\mathbb{E}\left[\left(\frac{\hat{s}_1}{\hat{s}_1 + \hat{s}_2} - \frac{s_1}{n}\right)^2\right].$$

Let $a = s_1$, $b = n$, $Z_a = Z_1$, $Z_b = Z_1 + Z_2$, where Z_1, Z_2 are from the hourglass mechanism. Let $M = 1$. With these definitions, to apply Lemma 3.1, we need to bound $\mathbb{E}[C^2]$, $\mathbb{E}[F^2]$, and $\Pr(Z_b < -b/2)$. We bound each of the terms below. Let $\alpha = \frac{s_1}{n}$.

$$\begin{aligned}
\mathbb{E}[C^2] &= \mathbb{E}\left[\left(\frac{Z_1}{n} - \frac{(\frac{s_1}{n})(Z_1+Z_2)}{n}\right)^2\right] \\
&= \frac{1}{n^2}\mathbb{E}\left[(1-\alpha)^2 Z_1^2 + \alpha^2 Z_2^2 - 2(1-\alpha)\alpha Z_1, Z_2\right] \\
&\leq \frac{1}{n^2}\mathbb{E}\left[(1-\alpha)^2 Z_1^2 + \alpha^2 Z_2^2\right] + 2(1-\alpha)\alpha\sqrt{\mathbb{E}\left[Z_1^2, Z_2^2\right]} \\
&\leq \frac{1}{n^2}(1-\alpha)^2\widetilde{\sigma}^2(\varepsilon) + \alpha^2\widetilde{\sigma}^2(\varepsilon) + 2(1-\alpha)\alpha\widetilde{\sigma}^2(\varepsilon) \\
&= \frac{\sigma^2(\varepsilon)}{n^2},
\end{aligned}$$

where the first inequality follows by the Cauchy-Schwarz inequality. Since $(x+y)^2 \le 2x^2 + 2y^2$,

$$\begin{aligned}\mathbb{E}[F^2] &\le \frac{8}{n^4}\mathbb{E}[(Z_1+Z_2)^4] + \frac{8}{n^4}\mathbb{E}[(Z_1+Z_2)^2 Z_1^2] \\ &= o\left(\frac{\sigma^2(\varepsilon)}{n^2}\right),\end{aligned}$$

where the last equality follows from Lemma A.4 and the fact that the marginal distribution of hourglass mechanism is same as the staircase mechanism.. Finally,

$$\begin{aligned}\Pr(Z_b < -b/2) &\le \Pr(Z_1 + Z_2 < -n/2) \\ &\le \Pr(Z_1 < -n/4) + \Pr(Z_2 < -n/4) \\ &= o\left(\frac{\sigma^2(\varepsilon)}{n^2}\right),\end{aligned}$$

where the last equality follows from Lemma A.5 and the fact that the marginal distribution of hourglass mechanism is same as the staircase mechanism. Combining the above three equations together with Lemma 3.1 and observing the fact that $\mu = \ell + \frac{s_1 w}{n}$ yields the theorem.

H. Proof of Theorem 5.2

We first state the following lemma, which removes the dependence on ℓ and u.

Lemma H.1.

$$R_{ar}(\varepsilon, \ell, u) = (u-\ell)^2 R_{ar}(\varepsilon, 0, 1).$$

Proof. Given a dataset from $D \in \mathcal{D}^*(\ell, u)$, one can create a dataset in $f(D) \in \mathcal{D}^*(0,1)$ by applying the transformation $f(x) = \frac{x-\ell}{u-\ell}$ to each of the points. Let $\hat{\mu}$ be a mean estimation algorithm for datasets in $\mathcal{D}^*(0,1)$, then given a dataset from $D \in \mathcal{D}^*(\ell, u)$, one can scale all points by applying f and compute the output as $\hat{\mu}'(D) = f^{-1}(\hat{\mu}(f(D))$. If $\hat{\mu}$ is an ε-differentially private algorithm, then $\hat{\mu}'$ is also an ε-differentially private algorithm. Furthermore, the utilities are related by

$$L(\hat{\mu}', D) \le (u-\ell)^2 L(\hat{\mu}, f(D)).$$

Taking supremum over datasets D and infimum over all differentially private algorithms yields

$$R_{\text{ar}}(\varepsilon, \ell, u) \le (u-\ell)^2 R_{\text{ar}}(\varepsilon, 0, 1).$$

The proof for the other direction is similar and omitted. □

Proof of Theorem 5.2. By Lemma H.1, it suffices to consider the scenario when $\ell = 0$ and $u = 1$. Let $\mathcal{D}^k$ and $S(D)$ be the same as those be defined as in Lemma 5.1. Let $\widetilde{\mathcal{D}}^k$ be the set of all datasets obtaining by combining each dataset in $\mathcal{D}^k$ with n values of zeros. Observe for any dataset $\widetilde{D}$ in $\widetilde{\mathcal{D}}^k$,

$$\mu(\widetilde{D}) = \frac{S(\widetilde{D})}{|\widetilde{D}|} \le \frac{k}{n}.$$

Suppose we have an ε-differentially private estimator $\hat{\mu}$ on $\widetilde{\mathcal{D}}^k$. We convert it to an estimator of sum for datasets in $\mathcal{D}^k$ as

$$\hat{S}(D) = n\hat{\mu}(D \cup \{0\}^n),$$

where $\{0\}^n$ is a dataset with n zeros. For this estimator, based on the Lemma 5.1, there exists a $D \in \mathcal{D}^k$ such that

$$\mathbb{E}\left[(\hat{S}(D) - S(D))^2\right] \ge \sigma^2(\varepsilon)(1 - o(1))$$

and hence

$$\mathbb{E}\left[(n\hat{\mu}(D \cup \{0\}^n) - S(D))^2\right] \ge \sigma^2(\varepsilon)(1 - o(1))$$

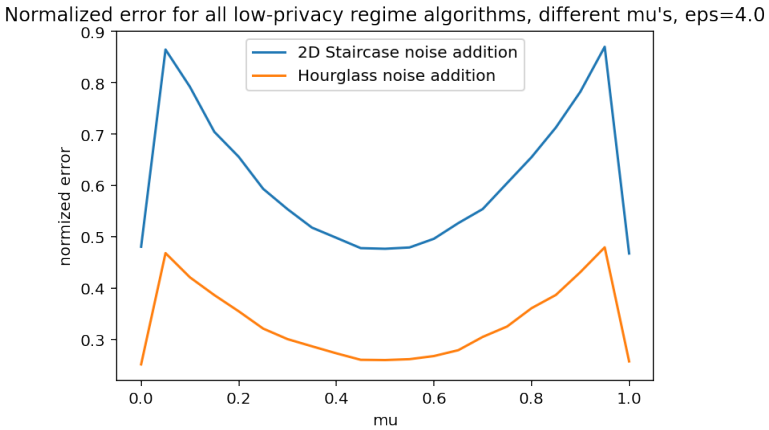


Figure 4. Hourglass and two dimensional staircase mechanisms, low privacy regime, varying μ.

We now upper bound the left hand side of the above expression. For brevity, let $\hat{\mu} = \hat{\mu}(D \cup \{0\}^n)$ and $\mu = \mu(D \cup \{0\}^n)$, and $S = S(D) = S(D \cup \{0\}^n)$. Observe that

$$\begin{aligned}
\mathbb{E}\left[(n\hat{\mu} - S)^2\right] &= \mathbb{E}\left[(n\hat{\mu} - n\mu + n\mu - S)^2\right] \\
&= \mathbb{E}\left[(n\hat{\mu} - n\mu)^2\right] + \mathbb{E}\left[(n\mu - S)^2\right] - 2\mathbb{E}\left[(n\hat{\mu} - n\mu)(n\mu - S)\right] \\
&\leq \mathbb{E}\left[(n\hat{\mu} - n\mu)^2\right] + \mathbb{E}\left[(n\mu - S)^2\right] + 2\sqrt{\mathbb{E}\left[(n\hat{\mu} - n\mu)^2\right]\mathbb{E}\left[(n\mu - S)^2\right]} \\
&\leq \mathbb{E}\left[(n\hat{\mu} - n\mu)^2\right] + \frac{k^4}{n^2} + 2\sqrt{\mathbb{E}\left[(n\hat{\mu} - n\mu)^2\right]}\frac{k^2}{n} \\
&\leq \mathbb{E}\left[(n\hat{\mu} - n\mu)^2\right] + \frac{k^4}{n^2} + \frac{k^3}{n} \\
&\leq |D|^2\mathbb{E}\left[(\hat{\mu} - \mu)^2\right] + \frac{k^4}{n^2} + \frac{k^3}{n},
\end{aligned}$$

where the first inequality follows by Cauchy-Schwarz inequality, the second inequality follows by observing that $|n\mu - S| \leq k\mu \leq k^2/n$, and the third inequality follows by observing that both $\hat{\mu}$ and μ lie in $[0, k/n]$. Setting $k = o(n^{1/3})$ yields the following theorem. □

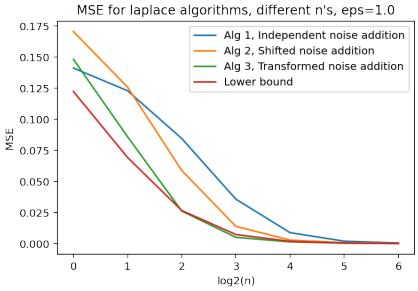
I. Additional experiments

In this section, we explore how the MSEs of different algorithms change when the the size of the dataset $|D|$ changes. The errors are normalized by $\frac{\varepsilon^2}{2}$ to keep them in a similar range across different privacy regimes. We choose $|D| \in \{2^i\}_{i=0,\ldots,6}$ and observe the trend of changes in MSEs for all listed algorithms.

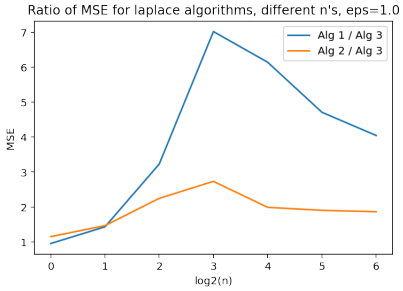
Figure 5a shows how the MSEs (normalized by $\frac{\varepsilon^2}{2}$) of different Laplace mechanisms vary as a function of $|D|$ for $\varepsilon = 1.0$ and $\mu = 0.01$. The lower bound (red curve), as in Section 6, is developed from using the univariate staircase mechanism with optimal γ on private mean in the swap model. One can see that only Algorithm 3 converges to the lower bound as $|D|$ grows bigger, being about two times better than Algorithm 2 when $|D| \geq 16$.

On the other hand, Figure 5c and 5d compare the MSEs (normalized by $\frac{\varepsilon^2}{2}$) of the two-dimensional staircase and the hourglass mechanism in the low privacy regime when $\varepsilon = 4.0$ and $\mu = 0.01$. The ratio of the former over latter is found to be always bigger than one, showing hourglass mechanism to be better.

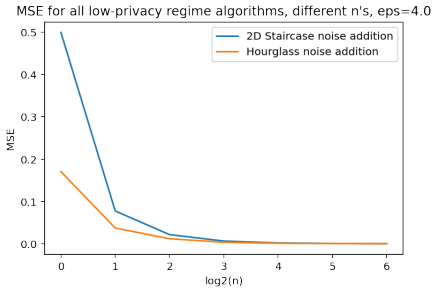
Table 1 lists the MSEs of all the algorithms listed in the paper for very large values of ε. Although some of these budget values are unlikely to be used in practice, we provide these results for completeness. One can see that hourglass performs exponentially better compared to all the other algorithms as ε grows larger.



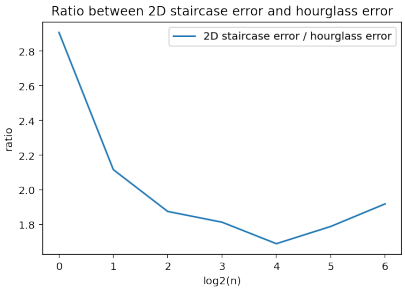
(a) High privacy regime, all Laplace mechanisms, varying $n = |D|$.



(b) High privacy regime, ratio of MSE, varying $n = |D|$.



(c) Low privacy regime, 2D staircase and hourglass mechanisms, varying $n = |D|$.



(d) Low privacy regime, ratio of MSE, varying $n = |D|$.

Figure 5. Experiments for different dataset sizes n.

Table 1. MSEs of different algorithms for extremely high privacy budgets

ε	Alg 1	Alg 2	Alg 3	2D staircase	hourglass	lower bound
4	3.99	1.95	0.98	0.93	0.52	0.51
8	3.99	1.95	0.99	0.66	0.11	0.10
16	4.00	1.97	0.97	1.70	1.14e-2	1.10e-2
32	4.01	1.94	0.98	1.20e-3	5.86e-8	5.73e-8