Subset-Based Instance Optimality in Private Estimation

Travis Dick [1] **Alex Kulesza** [1] **Ziteng Sun** [1] **Ananda Theertha Suresh** [1]

Abstract

We propose a new definition of instance optimality for differentially private estimation algorithms. Our definition requires an optimal algorithm to compete, simultaneously for every dataset D, with the best private benchmark algorithm that (a) knows D in advance and (b) is evaluated by its worst-case performance on large subsets of D. That is, the benchmark algorithm need not perform well when potentially extreme points are *added* to D; it only has to handle the removal of a small number of real data points that already exist. This makes our benchmark significantly stronger than those proposed in prior work. We nevertheless show, for real-valued datasets, how to construct private algorithms that achieve our notion of instance optimality when estimating a broad class of dataset properties, including means, quantiles, and ℓ_p-norm minimizers. For means in particular, we provide a detailed analysis and show that our algorithm simultaneously matches or exceeds the asymptotic performance of existing algorithms under a range of distributional assumptions.

1. Introduction

Differentially private algorithms intentionally inject noise to obscure the contributions of individual data points (Dwork et al., 2006; 2014). This noise, of course, reduces the accuracy of the result, so it is natural to ask whether we can derive a private algorithm that minimzes the cost to accuracy, "optimally" estimating some property $\theta(D)$ of a dataset D given the constraint of differential privacy.

But what do we mean by optimal? Optimal worst-case loss is often achievable by algorithms that add noise calibrated to the global sensitivity of θ (Dwork et al., 2006; Aldà & Simon, 2017; Balle & Wang, 2018; Fernandes et al., 2021). However, realistic datasets may have local sensitivities that are much smaller, making this a weak notion of optimality that does not reward reducing loss on typical "easy" examples (Nissim et al., 2007; Bun & Steinke, 2019). At the other extreme, we might hope for instance optimality, where a single algorithm competes, simultaneously for all datasets D, with the best possible benchmark algorithm chosen with knowledge of D. It is easy to see that this is too strong: a constant algorithm that always returns exactly $\theta(D)$, regardless of its input, is perfectly private and gives zero loss on D. Of course, we cannot hope to achieve zero loss for all datasets at once.

Thus, there has been recent interest in finding variations of instance optimality that are strong and yet still achievable. Asi & Duchi (2020b) gave a variant defined using local minimax risk, which softens the definition above by allowing an optimal algorithm's performance to degrade on D whenever there is a second dataset D' such that no private algorithm can simultaneously achieve low loss on both D and D'. Huang et al. (2021) gave a different variant of instance optimality for mean estimation in which the worst-case loss across nearby datasets sharing support with D defines the risk for D. Related ideas were also explored by Błasiok et al. (2019); Brunel & Avella-Medina (2020); McMillan et al. (2022), and others.

In this work we propose a new definition of instance optimality. We refer to our definition as *subset optimality* because it calibrates the risk of a dataset D using only subsets of D.

To motivate this approach, consider the basic semantics of differential privacy. A DP algorithm is required to give similar output distributions on D and D' whenever D' is a *neighbor* of D—that is, whenever D' can be obtained from D by adding or removing a single point. Technically, this definition is symmetric, but in practice addition and removal can have very different implications. For typical real datasets where data points tend to be similar to one another, removing a point may change the target property $\theta(D)$ only modestly, and a correspondingly modest amount of noise might ensure sufficiently similar output distributions. On the other hand, an *added* point could potentially be an extreme outlier that dramatically changes $\theta(D)$, requiring a large amount of noise to conceal its presence.

Concretely, imagine we have a database reporting the annual

[1]Google Research, New York. Correspondence to: Ziteng Sun <zitengsun@google.com>.

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

incomes for n households in a particular neighborhood, the largest of which happens to be \$100,000. We wish to privately compute the mean household income. Removing one of the existing households from the database can change the mean by at most roughly $\$100{,}000/n$. Adding a *new* household, on the other hand, could have a dramatically larger effect—a new household's income might theoretically be, say, \$100,000,000, requiring 1000x more noise. Thus, an algorithm that is required to perform well only on subsets of the real dataset intuitively has an advantage over one that must perform well everywhere.

The surprising implication of our results is that this is not always true. We demonstrate how to construct subset-optimal differentially private algorithms that simultaneously compete, for all datasets D, with the best private algorithm that (a) knows D in advance and (b) is evaluated by its worst-case performance over only (large) subsets of D. Subset optimality is achievable for a class of monotone properties that include means, quantiles, and other common estimators, despite being stronger than prior definitions in the literature.

We begin by describing our setting in more detail, defining subset optimality, and comparing our definition to those found in related work. We then show how to achieve subset optimality for mean estimation, giving an algorithm that simultaneously matches or exceeds the asymptotic performance of existing algorithms under a range of distributional assumptions (see Table 1). Finally, we generalize the result for means and show how to construct optimal algorithms for a broad class of monotone properties.

2. Problem formulation

A dataset D is a collection of points from the domain $[-R, R]$. Let $|D|$ be the cardinality of the set D. For two datasets D and D' we define the distance between them to be the number of points that need to be added to and/or removed from D to obtain D'. More formally,

$$d(D, D') \triangleq |D \setminus D'| + |D' \setminus D|.$$

For example, $d(\{1,2,3\},\{2,3,4\}) = 2$. We refer to D and D' as *neighboring* datasets if $d(D, D') = 1$ (Dwork et al., 2014, Chapter 2).

Note that this notion of neighboring datasets is sometimes called the *add-remove* model, in contrast to the *swap* model where all datasets are the same size and datasets are neighbors if they differ in a single point. We use the add-remove model here since we study algorithms that accept input datasets of various sizes. Because the add-remove model leads to stronger privacy than the swap model, our algorithms also provide guarantees in the swap model, with at most a factor of two increase in the privacy parameter ε.

Definition 2.1 (Differential privacy)**.** A randomized algorithm A with range $\mathcal{R}$ satisfies ε-differential privacy if for any two neighboring datasets D, D' and for any output $\mathcal{S} \subseteq \mathcal{R}$, it holds that

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

Let $\mathcal{A}_\varepsilon$ be the set of all ε-DP algorithms. Our goal is to estimate some property of D, denoted by $\theta(D)$, and we use a loss function $\ell : R \times R \to R_{\geq 0}$ to measure the performance of an algorithm A on a dataset D as

$$\ell(A(D), \theta(D)).$$

Given ℓ, we would like to identify an algorithm A that performs well not just on average or for certain datasets, but simultaneously for every dataset D. For each D, we will aim to compete with a benchmark algorithm that can be *selected* using knowledge of D but is *evaluated* according to its performance on large subsets of D. In particular, we adopt the following definition of subset-based risk.

$$R(D, \varepsilon) \triangleq \inf_{A \in \mathcal{A}_\varepsilon} \sup_{\substack{D' \subseteq D \\ |D'| \geq |D| - 1/\varepsilon}} \mathbb{E}\left[\ell(A(D'), \theta(D'))\right] \quad (1)$$

The benchmark algorithm for D is the one with the lowest worst-case loss on datasets obtained by removing up to $1/\varepsilon$ elements from D. Intuitively, if it is feasible to perform well on all of these subsets at once, then the benchmark risk is small and an optimal algorithm will be expected to perform correspondingly well, even if *adding* elements to D would dramatically increase the benchmark algorithm's loss. When no private algorithm performs well on all large subsets of D, then an optimal algorithm will also be permitted to have larger loss.

Definition 2.2. We say an algorithm A is *subset-optimal* with respect to $\mathcal{A}_\varepsilon$ if there exist constants α, β, and c such that, for all D, we have

$$\mathbb{E}\left[\ell(A(D), \theta(D))\right] \leq \alpha \cdot R(D, \varepsilon) + \beta,$$

and A is $c \cdot \varepsilon$-DP.

Ideally, we want the constants α and c to be close to 1 and β to be close to 0.

Remark 2.3. The constraint $|D'| \geq |D| - 1/\varepsilon$ in Equation (1) could be generalized to $|D| - \tau$ for any $\tau \geq 0$. Intuitively, smaller τ would lead to a stronger notion of optimality, since the benchmark algorithm would need to perform well on fewer datasets. Our choice of $\tau = 1/\varepsilon$ gives the strongest possible optimality definition that remains achievable: for any $\tau \ll 1/\varepsilon$, it is impossible to

[1] We focus on the case when $\varepsilon \leq 1$ in this paper. To simply our notation, we often treat $1/\varepsilon$ as a positive integer. If this is not the case, we can set $\varepsilon' = 1/\lceil 1/\varepsilon \rceil \in (\varepsilon/2, \varepsilon]$. This will affect our results by at most constant factors.

compete with the benchmark algorithm. To see this, consider the task of real mean estimation. Let D_1 contain $(n - 1/\varepsilon)$ copies of 0 and $1/\varepsilon$ copies of 1, and let D_2 contain $(n - 1/\varepsilon)$ copies of 0 and $1/\varepsilon$ copies of -1. Since the add/remove distance between D_1 and D_2 and is $2/\varepsilon$, standard packing arguments (e.g., Lemma 4.4 or Lemma 8.4 in Dwork et al. (2014)) show that an (ε, δ)-DP algorithm cannot give significantly different answers on D_1 and D_2, and therefore must incur an error of at least $\Theta(1/n\varepsilon)$ on one of them. On the other hand, a benchmark algorithm that knows the dataset can always output $1/n\varepsilon$ for D_1, and on subsets of D_1 with size at least $|D_1| - \tau$, the error will be at most τ/n (and similarly for D_2). Thus $\tau \ll 1/\varepsilon$ does not yield an achievable definition of instance-optimality.

Notation. For a dataset $D = \{x_1 \leq x_2 \leq \ldots \leq x_n\}$, let $L_m(D)$ denote the multiset $\{x_1, x_2, \ldots, x_m\}$ (the m lower elements of D) and $U_m(D)$ denote the multiset $\{x_{n-m+1}, \ldots, x_n\}$ (the m upper elements of D).

2.1. Our Contributions.

Subset-based instance optimality. We propose a new notion of instance-optimality (Definition 2.2) for private estimation. The notion has the advantage of only considering the effect of removing values from the dataset, which leads to tighter (or as tight) rates compared to other instance-optimal formulations that need to handle extreme data points. See Section 2.2 for a detailed discussion. Moreover, we propose Algorithm 4 based on private threshold estimation and the inverse sensitivity mechanism of (Asi & Duchi, 2020a). For real-valued datasets, the algorithm is *subset-optimal* for a wide range of monotone properties with arbitrary $\beta > 0$ and α and c at most logarithmic in problem-specific parameters (see Section 4 and Theorem 4.3 for the precise definition and statement).

Improvement on mean estimation. For the task of private mean estimation (Section 3), we propose an efficient algorithm (Algorithm 3) that is *subset-optimal*. In the statistical setting (Section 5) we show how this algorithm obtains distribution-specific rate guarantees that depend on all centralized absolute moments (Corollary 5.3). To the best of our knowledge, the rate improves upon the previously best-known distribution-specific results for distributions whose kth-moment is much smaller than its best sub-Gaussian proxy for some $k \geq 2$. For distribution families with concentration assumptions, the distribution-specific rate recovers (up to logarithmic factors) the min-max optimal rate for each corresponding family. Moreover, our proposed algorithm achieves this rate without explicit assumptions on which family the distribution comes from. See Table 1 for a detailed comparison.

2.2. Related Work

Instance-optimality in private estimation. Several variations of instance optimality for differential privacy have been studied recently.

Asi & Duchi (2020b) initiated the study of instance optimality using the following notion of *local minimax risk*:

$$R_1(D, \varepsilon) = \sup_{D'} \inf_{A \in A_\varepsilon} \sup_{\tilde{D} \in \{D, D'\}} \mathbb{E}[\ell(A(\tilde{D}), \theta(\tilde{D}))].$$

They showed that the inverse sensitivity mechanism gives nearly optimal results with respect to this notion of risk. However, for mean estimation in one dimension, with all values bounded in $[-R, R]$, it can be shown that

$$R_1(D, \varepsilon) \gtrapprox \frac{R}{n\varepsilon}$$

for every dataset D. In contrast, subset optimality provides much tighter guarantees for mean estimation, roughly replacing the full range R with the range actually spanned by D, as described in the sections below. For general losses, Asi & Duchi (2020b) showed that

$$R_1(D, \varepsilon) \approx \max_{D' : d(D, D') \leq 1/\varepsilon} \ell(\theta(D), \theta(D')), \qquad (2)$$

while we show that

$$R(D, \varepsilon) \approx \max_{D' : d(D, D') \leq 1/\varepsilon, D' \subseteq D} \ell(\theta(D), \theta(D')),$$

which is strictly tighter due to the subset constraint. (As illustrated above, the difference can be dramatic for realistic D.) McMillan et al. (2022) studied a quantity similar to $R_1(D, \varepsilon)$ in the setting where D is drawn from a distribution. We focus on the comparison to Asi & Duchi (2020a) since it is most relevant to our setting.

Huang et al. (2021) considered a slightly different notion of instance optimality given by $R_2(D, \varepsilon) =$

$$\inf_{A \in \mathcal{A}_\varepsilon} \sup_{\substack{\text{supp}(D') \subseteq \text{supp}(D) \\ d(D, D') = 1}} \inf_{\eta} \left\{ \Pr\left(\ell(A(D') \theta(D')) > \eta\right) < \frac{2}{3} \right\},$$

where $\text{supp}(D)$ denotes the set of unique elements in the dataset D. They proposed an algorithm for d-dimensional mean estimation and showed that it is $O(\sqrt{d/\rho})$-optimal for ρ-zCDP (concentrated differential privacy). Their definition and results differ from $R(D, \varepsilon)$ and $R_1(D, \varepsilon)$ in two basic ways. First, their definition is a high probability definition, whereas the others are in expectation. Second, even for one-dimensional mean estimation, their proposed algorithm is only $O(1/\sqrt{\rho})$ competitive with the lower bound, and they further show that no algorithm can achieve a better competitive guarantee. Our definition (modulo expectation)

Table 1. Results for statistical mean estimation. $R(A,p)$ is the expected absolute error of A given n *i.i.d.* samples from p (Equation (3)). $M_k(p)$ denotes the kth absolute central moment of p (Definition 5.2). $\sigma_{\mathcal{G}}(p)$ denotes the best sub-Gaussian proxy of p. † is due to (Huang et al., 2021). ‡ is due to Kamath et al. (2020) and * is due to Karwa & Vadhan (2017).

Assumption	Metric	Prior work	This work
N.A.	$R(A,p)$	$\tilde{O}\left(\frac{\sqrt{M_2(p)}}{\sqrt{n}}+\frac{\sigma_{\mathcal{G}}(p)}{n\varepsilon}\right)$†	$\tilde{O}\left(\frac{\sqrt{M_2(p)}}{\sqrt{n}}+\min_k \frac{M_k(p)^{1/k}}{(n\varepsilon)^{1-1/k}}\right)$
Bounded Moment	$\max_{p:M_k(p)\le m_k} R(A,p)$	$O\left(\frac{m_k^{1/k}}{\sqrt{n}}+\frac{m_k^{1/k}}{(n\varepsilon)^{1-1/k}}\right)$‡	$\tilde{O}\left(\frac{m_k^{1/k}}{\sqrt{n}}+\frac{m_k^{1/k}}{(n\varepsilon)^{1-1/k}}\right)$
Sub-Gaussian	$\max_{p:\sigma_{\mathcal{G}}(p)\le\sigma} R(A,p)$	$\tilde{O}\left(\frac{\sigma}{\sqrt{n}}+\frac{\sigma}{n\varepsilon}\right)^*$	$\tilde{O}\left(\frac{\sigma}{\sqrt{n}}+\frac{\sigma}{n\varepsilon}\right)$

coincides with this definition for $\varepsilon = 1$; however, we are able to construct constant optimal algorithms for general ε. We also show that our definition of optimality is achievable for target properties beyond just means.

Remark 2.4. The definition of Huang et al. (2021) can be extended by changing $d(D,D') = 1$ to $d(D,D') \le 1/\varepsilon$, bringing it closer to our definition. However, this leads to an instance-dependent risk of

$$\tilde{R}_2(D,\varepsilon) \approx \sup_{\substack{\mathrm{supp}(D')\subseteq\mathrm{supp}(D)\\ d(D,D')\le 1/\varepsilon}} \ell(\theta(D),\theta(D')),$$

which can be much larger than bound in Lemma 3.2.

In Appendix F, we use ℓ_p minimization as a concrete example to compare these instance-dependent risks and show that our new definition gives significant quantitative improvements for specific datasets.

It is worth remarking here that both Asi & Duchi (2020a) and Huang et al. (2021) provide achievability results when the dataset is supported over a high-dimensional space while our result mainly focuses on one-dimensional datasets. Whether our subset-based instance optimality can be achieved in the high-dimensional setting is an interesting future direction to explore.

Private statistical mean estimation. Private mean estimation has been widely studied in the statistical setting, where the dataset is assumed to be generated *i.i.d.* from an underlying distribution. Classic methods such as the Laplace or Gaussian mechanism (Dwork et al., 2014) incur a privacy cost that scales with the worst-case sensitivity. However, recently, by assuming certain concentration properties of the underlying distribution (e.g., sub-Gaussianity (Karwa & Vadhan, 2017; Cai et al., 2019; Smith, 2011; Kamath et al., 2019; Bun & Steinke, 2019; Biswas et al., 2020), bounded moments (Feldman & Steinke, 2018; Kamath et al., 2020; Hopkins et al., 2022) and high probability concentration (Levy et al., 2021; Huang et al., 2021)), it has been shown that the privacy cost can be improved to scale with the concentration radius of the underlying distribution. These algorithms are in some cases known to be (nearly) optimal for distributions satisfying these specific concentration properties.

It is worth remarking here that most of these works consider high-dimensional distributions while our work only focuses on real-valued datasets. Moreover, for moment-bounded distributions, (Kamath et al., 2020) achieves constant-optimal estimation risk while our obtained rate is only optimal up to logarithmic factors. The results are outlined in Table 1.

Our mean estimation algorithm is similar to that of (Huang et al., 2021). However, we choose a tighter threshold in the clipping bound estimation step, which is crucial in the analysis to achieve the instance-optimal bound.

3. Subset-Optimal Private Means

We start by presenting an efficient subset-optimal algorithm for estimating means; in Section 4 we will generalize this approach to a larger class of properties.

Let $\mu(D)$ denote the mean of a dataset D. We will use $\ell(x,y) = |x-y|$ as our loss function. Our main result is stated in the theorem below.

Theorem 3.1. *Let $D \subset [-R,R]$ be a multiset of points and let $\hat{\mu}$ be the output of Algorithm 3 with parameters R and $\varepsilon > 0$. Publishing $\hat{\mu}$ is 3ε-differentially private and for any $\gamma > 0$, we have*

$$\mathbb{E}\left[|\hat{\mu}-\mu(D)|\right] = O\left(R\left(D,\varepsilon\right)\log\frac{R\varepsilon}{\gamma}+\frac{\gamma}{\varepsilon}\right).$$

We begin by establishing an up-to-constant characterization of the subset-based risk for mean estimation under this loss.

Lemma 3.2. *For any $\varepsilon \in (0,1]$ and multiset $D \subset \mathbb{R}$ with $|D| > \frac{1}{\varepsilon}$,*

$$R(D,\varepsilon) = c_{D,\varepsilon}\cdot\left(\mu(D\setminus L_{\frac{1}{\varepsilon}}) - \mu(D\setminus U_{\frac{1}{\varepsilon}})\right),$$

where $c_{D,\varepsilon} \in [1/(2e^2), 1]$.

The upper bound is straightforward; the lower bound proof is based on standard packing arguments (Dwork et al., 2014) and appears in Appendix B.1.

Now we turn to designing a subset-optimal algorithm that is competitive with this lower bound for every dataset D and prove Theorem 3.1.

First, we describe a straightforward algorithm for private mean estimation of bounded datasets. Pseudocode is given in Algorithm 1, and privacy and utility analyses are given in Lemma 3.3. We use $D + m$ to denote $\{x + m \mid x \in D\}$, and $\text{clip}(D, [l, u])$ to denote $\{\min(\max(x, l), u) \mid x \in D\}$.

Algorithm 1 Bounded mean estimation

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

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

We provide the proof in Appendix B.2.

Lemma 3.3. *Let $[l, u]$ be any interval and $\varepsilon > 0$ be any privacy parameter. Publishing $\hat{\mu}$ output by Algorithm 1 is ε-differentially private. Furthermore,*

$$\mathbb{E}\left[|\hat{\mu} - \mu(D)|\right] \leq \frac{3(u-l)}{|D|\varepsilon}.$$

In order to construct a subset-optimal mean algorithm from Algorithm 1, the high level idea is to first find an interval $[\hat{l}, \hat{u}]$ that contains all but a small number of outliers from the dataset D, clip D to $[\hat{l}, \hat{u}]$, and finally apply Algorithm 1. The error of this algorithm will have two main components: error incurred by clipping the data to $[\hat{l}, \hat{h}]$, and error due to the noise added by Algorithm 1. Our analysis shows that both error components are not significantly larger than the lower bound given by Lemma 3.2.

We start by describing the subroutine that we use to choose $\hat{l}$ and $\hat{u}$; following Lemma 3.2, our goal will be to find $\hat{l}$ and $\hat{u}$ that delineate approximately $1/\varepsilon$ elements of D each.

3.1. Private Thresholds

We present an ε-differentially private algorithm that, given a multiset D of real numbers and a target rank r, outputs a threshold $\tau \in \mathbb{R}$ that is approximately a rank-r threshold for D.

Roughly, τ being a rank-r threshold for D means that r points in D are less than or equal to τ. However, if D has repeated points then there can be ranks for which no such threshold exists. (In the extreme, consider the dataset $D = \{x, \ldots, x\}$, containing n copies of the same point; exactly 0 or n points are less than or equal to any threshold τ.)

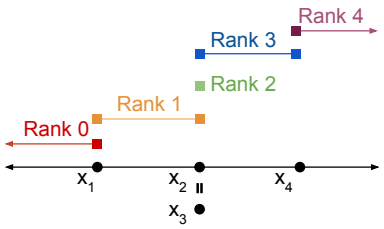

Figure 1. Example of ranks as defined in Definition 3.4. There are 4 points with $x_2 = x_3$. For each rank $r \in \{0, \ldots, 4\}$ we show the interval of points that are rank-r thresholds. For $r = 0, 1, \ldots, 4$, the intervals of rank-r thresholds are given by $I_0 = [-\infty, x_1]$, $I_1 = [x_1, x_2]$, $I_2 = \{x_2\}$, $I_3 = [x_3, x_4]$, and $I_4 = [x_4, \infty]$, respectively.

We will therefore define a rank-r threshold in a slightly more general way so that there is at least one rank-r threshold for every $r \in \{0, \ldots, |D|\}$. This definition is consistent with the standard definition of quantiles for distributions with point-masses.

Definition 3.4. Let D be any multiset of real numbers. We say that $\tau \in \mathbb{R}$ is a rank-r threshold for D if $\sum_{x \in D} \mathbb{I}\{x < \tau\} \leq r$ and $\sum_{x \in D} \mathbb{I}\{x \leq \tau\} \geq r$. That is, there are at most r points strictly smaller than x, and at least r points that are greater than or equal to x. For any threshold τ, let $\text{ranks}(\tau, D)$ denote the set of ranks r such that τ is a rank-r threshold for D. See Figure 1 for an example of this rank definition in a dataset with repeated points.

Definition 3.5. The *rank error* of a threshold τ for dataset D and target rank r is

$$\text{err}_{\text{rank}}(\tau, r, D) = \min_{r' \in \text{ranks}(\tau, D)} |r - r'|.$$

The rank error measures how close τ is to being a rank-r threshold for D and is equal to 0 if and only if τ is a rank-r threshold.

When privately estimating rank-r thresholds for a dataset D, we will incur a bicriteria error: our output will be close (on the real line) to a threshold with low rank error.

Definition 3.6. Let D be any multiset of real numbers. We say that τ is an (α, β)-approximate rank-r threshold for D if there exists $\tau' \in \mathbb{R}$ so that $|\tau - \tau'| \leq \alpha$ and $\text{err}_{\text{rank}}(\tau', r, D) \leq \beta$.

Our algorithm for finding approximate rank-r thresholds is an instance of the exponential mechanism. The loss (or

Algorithm 2 Threshold estimation

Input: Dataset $D \subset [a,b]$, target rank r, data range $[a,b]$, distance $\alpha > 0$, privacy parameter $\varepsilon > 0$.

1 Define $\ell(\tau) = \min_{\tau-\alpha \le \tau' \le \tau+\alpha} \mathrm{err}_{\mathrm{rank}}(\tau', r, D)$.
2 Let $f(\tau) = \exp(-\frac{\varepsilon}{2}\ell(\tau)) / \int_a^b \exp(-\frac{\varepsilon}{2}\ell(\tau))\, d\tau$.
3 Output sample $\hat{\tau}$ in $[a,b]$ drawn from density f.

negative utility) function that we minimize is parameterized by the distance error $\alpha > 0$, and the loss of a threshold τ is defined to be the minimum rank-error of any threshold τ' within distance α of τ.

Intuitively, by allowing the loss function to "search" in a window around τ, we guarantee that there is always an interval of width α with loss zero (namely, any interval centered around a true rank-r threshold). This is sufficient to argue that the exponential mechanism outputs a low-loss threshold with high probability and, by definition of the loss, this implies that there is a nearby threshold with low rank-error. Pseudocode is given in Algorithm 2. Since the density f is piecewise constant with at most $2|D|$ discontinuities at locations $x \pm \alpha$ for $x \in D$, it is possible to sample from f in time $O(|D| \log |D|)$ (see the proof of Theorem 3.7 for details). Theorem 3.7, which is proved in Appendix A, shows that this algorithm outputs an (α, β)-approximate threshold.

Theorem 3.7. *Fix data range $[a,b]$, $\varepsilon > 0$, and distance error $0 < \alpha \le \frac{b-a}{2}$. For any dataset $D \subset [a,b]$ and rank $1 \le r \le |D|$, let $\hat{\tau}$ be the output of Algorithm 2 run on D with parameters $[a,b]$, α, and ε. Publishing $\hat{\tau}$ satisfies ε-DP and, for any $\zeta > 0$, with probability at least $1-\zeta$, $\hat{\tau}$ is an (α, β)-approximate rank-r threshold for D with $\beta = \frac{2}{\varepsilon} \log \frac{b-a}{\alpha\zeta}$. Moreover, Algorithm 2 can be implemented with $O(|D| \log |D|)$ running time.*

Next, we argue that when the dataset is supported on a grid $\mathcal{Z} = \{z_1, \ldots, z_m\}$, where $z_{i+1} = z_i + \gamma$, running Algorithm 2 with a sufficiently small distance parameter α and rounding to the nearest grid point results in a $(0, \beta)$-approximate rank-r threshold with high probability. The proof of Corollary 3.8 is in Appendix A.

Corollary 3.8. *Let $\mathcal{Z} = \{z_1 \le \cdots \le z_m\}$ be such that $z_{i+1} = z_i + \gamma$ for all $i = 1, \ldots, m-1$ and let D be any multiset supported on $\mathcal{Z}$. Let $\hat{\tau}$ be the output of Algorithm 2 run on D with parameters $[a,b] = [z_1, z_m]$, $\alpha = \gamma/3$, and $\varepsilon > 0$, and let $\tilde{\tau}$ the closest point in $\mathcal{Z}$ to $\hat{\tau}$. Then for any $\zeta > 0$, with probability at least $1-\zeta$ we have that $\tilde{\tau}$ is a $(0, \beta)$-approximate rank-r quantile for D with $\beta = \frac{2}{\varepsilon} \log \frac{3m}{\zeta}$.*

3.2. Mean Estimation

We now present the pseudo-code for our subset-optimal mean estimation algorithm in Algorithm 3 and give the proof sketch of Theorem 3.1.

Proof sketch of Algorithm 3. We provide the proof sketch here and provide a detailed proof in Appendix B.3. Our goal is to show that the expected error of Algorithm 3 is not much larger than

$$R(D, \varepsilon) \ge \frac{\mu(D \setminus L_{\frac{1}{\varepsilon}}) - \mu(D \setminus U_{\frac{1}{\varepsilon}})}{2e^2}.$$

With probability at least $1-\zeta$, by Theorem 3.7 we are guaranteed that $\hat{l}$ and $\hat{u}$ are (α, β)-approximate rank-t_l and rank-t_u thresholds, respectively for the values of α and β defined in Algorithm 3. In particular, this implies that there exist l' and t_l' such that $|\hat{l} - l'| \le \alpha$, $|t_l' - t_l| \le \beta$, and l' is a rank-t_l' threshold for D. Similarly, there exist u' and t_u' such that $|\hat{u} - u'| \le \alpha$, $|t_u' - t_u| \le \beta$, and u' is a rank-t_u' threshold for D. Let G denote this high probability event. We first argue that conditioned on G, the expected loss of $\hat{\mu}$ is small (where the expectation is taken only over the randomness of Algorithm 1). To convert this bound into a bound that holds in expectation, we bound the error when G does not hold by $2R$.

Let $\hat{\mu}$ be the output of Algorithm 3. We decompose the error of $\hat{\mu}$ into three terms:

$$\begin{aligned}|\hat{\mu} - \mu(D)| \le\ & |\hat{\mu} - \mu(\mathrm{clip}(D, [\hat{l}, \hat{u}]))| \\ & + |\mu(\mathrm{clip}(D, [\hat{l}, \hat{u}]) - \mu(\mathrm{clip}(D, [l', u'])| \\ & + |\mu(\mathrm{clip}(D, [l', u'])) - \mu(D)|.\end{aligned}$$

Roughly speaking, the first term captures the variance incurred by using Algorithm 1 to estimate the mean of the clipped data, the second term measures our bias due to α in our (α, β)-approximate thresholds, and the third term measures the bias due to β. Our goal is to prove that all of these terms are not much larger than $R(D, \varepsilon)$.

Bounding first term. At a high level, we argue that all points in $L_{\frac{1}{\varepsilon}}$ are to the left of $\hat{l} + \alpha$, and all points in $R_{\frac{1}{\varepsilon}}$ are to the right of $\hat{u} - \alpha$. It follows that the distance from any point in $L_{\frac{1}{\varepsilon}}$ to any point in $U_{\frac{1}{\varepsilon}}$ is at least $\hat{u} - \hat{l} - 2\alpha$. In particular, this guarantees that the difference between the means of $D \setminus L_{\frac{1}{\varepsilon}}$ and $D \setminus U_{\frac{1}{\varepsilon}}$ must be at least $\frac{\hat{u} - \hat{l} - 2\alpha}{\varepsilon(|D| - \frac{1}{\varepsilon})}$, since we move $1/\varepsilon$ points a distance at least $\hat{u} - \hat{l} - 2\alpha$. This expression is close to the loss incurred by Algorithm 1 when run on the clipped dataset.

[2]Note that when τ is a rank-r threshold for a dataset D, $-\tau$ is a rank $|D| - r$ threshold for $-D$. Therefore, we can use Algorithm 2 to find an approximate rank-$(|D| - r)$ threshold for D by negating an approximate rank-r threshold for $-D$.

Algorithm 3 Subset optimal mean estimation

Input: Range R, dataset $D \subset [-R, R]$, privacy parameter $\varepsilon' > 0$ and $\alpha > 0$.

Define $\alpha = \frac{\gamma}{|D|}$, $\zeta = \frac{\alpha}{R|D|\varepsilon}$, $\beta = \frac{2}{\varepsilon} \log \frac{2R}{\alpha\zeta}$, $t_l = \frac{1}{\varepsilon} + \beta$, and $t_u = |D| - \frac{1}{\varepsilon} - \beta$.[2]
Let $\hat{l}$ be the output of Algorithm 2 run on D with parameters $[a, b] = [-R, R]$, $r = t_l$, α, and ε.
Let $\hat{u}$ be the output of Algorithm 2 run on D with parameters $[a, b] = [-R, R]$, $r = t_u$, α, and ε.
Let $\hat{\mu}$ be the output of Algorithm 1 run on D with interval $[a, b] = [\hat{l}, \hat{u}]$ and privacy parameter ε.
Output $\hat{\mu}$.

Bounding the second term. The key idea behind bounding the second term is that, whenever $\hat{l}$ is close to l' and $\hat{u}$ is close to u', then clipping a point x to $[\hat{l}, \hat{u}]$ is approximately the same as clipping it to $[l', u']$.

Bounding the third term. Our bound for the third term is the most involved. At a high level, we show that the bias introduced by clipping D to the interval $[l', u']$ is at most the worst "one-sided" clipping bias incurred clipping the points to the left of l' or to the right of u'. To see this, observe that when we clip from both sides, the left and right biases cancel out. Next, we argue that clipping points to the left of l' (or to the right of u') introduces less bias than *removing* those points. This step bridges the gap between Algorithm 1 which clips points and the lower bound on $R(D, \varepsilon)$, which removes points. We argue that the number of points removed to the left of l' or to the right of u' is not much larger than $\frac{1}{\varepsilon}$ and use Lemma B.1 to show that the resulting bias is not much larger than if we had removed exactly $\frac{1}{\varepsilon}$ points instead. Finally, to finish the bound, combine our two "one-sided" bias bounds to show that the overall bias is never much larger than $R(D, \varepsilon)$. □

3.3. Intuition

The lower bound in Lemma 3.2 is obtained by showing (roughly) that no private algorithm can reliably determine whether $O(1/\varepsilon)$ outliers have been removed from D. In proving the upper bound in Theorem 3.1, then, the challenge is to show that those outliers can be identified and removed *privately* without introducing asymptotically larger error even when D is not known in advance.

This is possible in Algorithm 3 due to a careful choice of the rank targets t_l and t_u. In particular, if we are overly aggressive in trying to privately remove outliers, we run the risk of adding too much bias, since we are clipping away important information about the mean. On the other hand, if we are too tentative, we may end up with wide clipping thresholds that require adding too much variance (in the form of noise) when calling Algorithm 1. The key to our construction, therefore, is choosing rank targets such that the risk of excess bias and the risk of excess variance both exactly balance with the lower bound; that is, they match the error incurred by removing outliers in the first place.

There is no reason to think *a priori* that this should be possible. Indeed, for certain properties (such as the mode), such a result does not seem to exist—errors due to over- or underestimating outliers can change the property arbitrarily. However, we show in the next section that the result for mean estimation can be extended to a relatively large class of common properties.

4. Instance-optimal algorithm for monotone properties

We now show that subset-optimal estimation algorithms can be constructed for any "monotone" property. We start by defining our notion of monotonicity.

Definition 4.1 (First-order stochastic dominance (Lehmann, 1955; Mann & Whitney, 1947))**.** Let D and D' both be multisets of real numbers. D' is said to *dominate* D (denoted $D' \succ D$) if, $\forall v \in \mathbb{R}$,

$$\frac{\sum_{x' \in D'} \mathbb{1}\{x' \leq v\}}{|D'|} \leq \frac{\sum_{x \in D} \mathbb{1}\{x \leq v\}}{|D|}.$$

In other words, first-order stochastic dominance requires the cumulative density function (CDF) of D to be larger than the CDF of D' for all points on the real line.

Definition 4.2 (Monotone property)**.** A property is called *monotone* if, for all D', D with $D' \succ D$, we have

$$\theta(D') \geq \theta(D)$$

or, for all D', D with $D' \succ D$, we have

$$\theta(D') \leq \theta(D).$$

Intuitively, the definition requires that if we move points from a dataset in one direction, we will always increase (or always decrease) the property. The family of monotone properties includes natural functions such as the mean, median, and quantiles. It also includes minimizers of ℓ_p distances, i.e.,

$$\theta_p(D) = \arg\min_y \sum_{x \in D} |x - y|^p,$$

and other common estimators.

Algorithm 4 Subset-optimal monotone property estimation

Input: Range R, dataset $D \subset [-R, R]$, privacy parameter $\varepsilon > 0$, and discretization parameter β.
Algorithms: Private threshold algorithm **PrvThreshold** (Algorithm 2). Inverse sensitivity algorithm **InvSen** (Asi & Duchi, 2020a) (Algorithm 5).

1: Quantize each value in the dataset D to the nearest multiple of β and denote the quantized dataset by D_{quant}.
2: Set error probability $\eta = \frac{L\beta}{B}$, rank $r = \frac{32\log(6R/\eta\beta)}{\varepsilon}$.
3: $l = \textbf{PrvThreshold}(D_{\text{quant}}, r/4, [-R, R], \beta/3, \varepsilon/4)$.
4: $u = \textbf{PrvThreshold}(D_{\text{quant}}, |D| - r/4, [-R, R], \beta/3, \varepsilon/4)$.
5: Let $D_{\text{quant}} = \{x_1 \le x_2 \le, \ldots, \le x_n\}$. For $i \le 3r/2$, let $y_i = x_i - \beta$, for $i \ge n - 3r/2$ $y_i = x_i + \beta$ and otherwise $y_i = x_i$. Let $D'_{\text{quant}} = \{y_1 \le y_2 \le, \ldots, \le y_n\}$.
6: Prune the dataset to obtain
$$D_{[l,u]} = D'_{\text{quant}} \cap [l, u].$$
7: Return the output of **InvSen** on $D_{[l,u]}$ with range $[l, u]$, granularity β, and privacy parameter $\varepsilon/2$.

We also make the following assumptions on the property θ and loss function ℓ:

- For any dataset D supported on $[-R, R]$, $\theta(D) \in [-R, R]$[3].
- ℓ is a metric; that is, it is commutative, it satisfies the triangle inequality, and $\ell(\theta, \theta) = 0$.
- ℓ is finite and bounded for all datasets under consideration. Let $B = \sup_{D,D'} \ell(\theta(D), \theta(D'))$.
- Whenever $\theta \ge \theta_1 \ge \theta_2$,
$$\ell(\theta, \theta_1) \le \ell(\theta, \theta_2).$$
- ℓ is L-Lipschitz, as defined below[4].
  - Let $x_i(D)$ denote the i^{th} largest element in D. For all D, D' such that $|D| = |D'|$, we have
  $$\ell(\theta(D), \theta(D')) \le L \max_{i \le |D|} |x_i(D) - x_i(D')|.$$
  - For all θ and $\theta_1 \ne \theta_2$, we have
  $$\ell(\theta, \theta_1) \le \ell(\theta, \theta_2) + L|\theta_1 - \theta_2|.$$

  Observe that both mean and median are 1-Lipschitz when $\ell(\theta, \theta') = |\theta - \theta'|$.

Our main result is stated below.

Theorem 4.3. *For any $\varepsilon \in (0, c_\varepsilon^{-1})$, there exists a $c_\varepsilon \cdot \varepsilon$-DP Algorithm (Algorithm 4) with $c_\varepsilon = 128\log(6RB/L\beta^2)$, whose output $A(D)$ satisfies*
$$\mathbb{E}[\ell(A(D), \theta(D))] \le 2e^2 R(D, \varepsilon) + 7L\beta.$$

We first show a simple lower bound on $R(D, \varepsilon)$ for general properties. This bound generalizes the mean estimation lower bound in Lemma 3.2; the proof is given in Appendix D.1.

[3]In general, we only need to assume the property is bounded and our result only depends on the bound logarithmically. We use the same R here to simplify notations.

[4]The constants in the two conditions below need not be the same. We use L here for both to keep notations simple.

Lemma 4.4. *For $\varepsilon \in [0, 1]$, let $S = \{(D_1, D_2) : \min(|D_1|, |D_2|) \ge |D| - 1/\varepsilon, d(D_1, D_2) \le 2/\varepsilon\}$. If $S \ne \emptyset$, then $R(D, \varepsilon)$ is at least*
$$\frac{1}{2e^2} \cdot \max_{(D_1, D_2) \in S} \ell(\theta(D_1), \theta(D_2)).$$

We show that the above lower bound can be achieved (up to logarithmic factors). The algorithm is given in Algorithm 4. It is similar in spirit to Algorithm 3, but we need to make a few modifications to ensure the algorithm works for general monotone properties. We briefly describe the steps in the algorithm below.

Discretization: As before, we will use the private threshold algorithm to remove outliers. The approximation guarantee in Theorem 3.7 has an additive rank error β and an additive threshold error α; however, for general properties, it is technically challenging to bound the effects of nonzero α. To work around this, we first discretize the interval into steps of size β. This allows us to use Corollary 3.8 to get a guarantee with $\alpha = 0$.

Private thresholds: We then find the private thresholds l and u as in Algorithm 3. As noted above, these estimates come with $(0, \beta)$ approximation guarantees due to discretization.

Pruning outliers: In Algorithm 3, we clip outliers outside the thresholds. However, the effect of clipping is difficult to analyze generally. Instead, in Algorithm 4 we simply prune outliers. Technically, it is possible for all values to lie on the thresholds, in which case we might not prune any elements. Hence, for ease of analysis, we deliberately move a small fraction of points outside the thresholds.

Inverse sensitivity mechanism. Finally, while in Algorithm 3 we directly computed the private mean of the clipped dataset, here we use the inverse sensitivity mechanism (Asi & Duchi, 2020a) to estimate the desired property.

5. Implications on private statistical mean estimation.

In this section, we apply our mean estimation algorithm to the statistical mean estimation (SME) task where $D = X^n$, which are *i.i.d.* samples from a distribution p with mean μ. And the performance of the algorithm is measured by the expected distance from the mean,

$$R_{\mathrm{SME}}(A, p) := \mathbb{E}\left[|A(D) - \mu|\right]. \tag{3}$$

We apply Algorithm 3 on D and obtain obtain distribution-specific bounds on $R_{\mathrm{SME}}(A, p)$. For distribution families with various concentration assumptions, we show that our instance-based bound is almost as tight (up to logarithmic factors) as algorithms designed for specific distribution families. We first state a generic result for statistical mean estimation.

Theorem 5.1. *Let $D = X^n$ be i.i.d. samples from a distribution p with mean μ and A be Algorithm 3. We have*

$$\begin{aligned} R_{\mathrm{SME}}(A, p) \le{}& \mathbb{E}\left[|\mu(D) - \mu|\right] \\ &+ C \cdot \mathbb{E}\left[\left|\mu(D \setminus L_{\frac{1}{\varepsilon}}) - \mu(D \setminus U_{\frac{1}{\epsilon}})\right|\right], \end{aligned}$$

where C hides logarithmic factors in the problem parameters.

Proof.

$$\begin{aligned} R_{\mathrm{SME}}(A, p) &= \mathbb{E}\left[|A(D) - \mu|\right] \\ &\le \mathbb{E}\left[|\mu(D) - \mu|\right] + \mathbb{E}\left[|A(D) - \mu(D)|\right]. \end{aligned}$$

Applying Theorem 3.1 to the second term directly leads to the claim. □

The bound in Theorem 5.1 can be hard to compute for a specific distribution. For distributions with bounded moments, we can obtain explicit upper bounds on the quantities above.

Definition 5.2. Let p be a distribution supported on $\mathbb{R}$ with mean μ. Its kth absolute central moment is denoted as

$$M_k(p) := \mathbb{E}_{X \sim p}\left[|X - \mu(p)|^k\right]$$

if it is finite; otherwise $M_k(p) = \infty$.

In Appendix E.1, we prove the following result for statistical mean estimation on distributions with bounded moments.

Corollary 5.3. *For any distribution p over $[-R, R]$, Algorithm 3 satisfies*

$$R_{\mathrm{SME}}(A, p) = \tilde{O}\left(\frac{M_2(p)^{1/2}}{\sqrt{n}} + \min_{k \ge 2} \frac{M_k(p)^{1/k}}{(n\varepsilon)^{1-1/k}}\right).$$

Note that Algorithm 3 obtains the above instance-specific rate without any knowledge on the underlying distribution p. Moreover, for specific distribution families such as sub-gaussian distributions and distributions with bounded kth moments ($k \ge 2$), Corollary 5.3 implies almost tight min-max rates.

Subgaussian distributions. A distribution p is subgaussian with proxy σ if $\forall t \ge 0$,

$$\mathbb{P}(|X - \mu| \ge t) \le 2 \exp\left(-\frac{t^2}{\sigma^2}\right).$$

We denote all such distributions as $\mathcal{G}_\sigma$. For such distributions, we have

$$\max_{p \in \mathcal{G}_\sigma} R_{\mathrm{SME}}(A, p) = \tilde{O}\left(\frac{\sigma}{\sqrt{n}} + \frac{\sigma}{n\varepsilon}\right).$$

This matches the optimal rate for sub-Gaussian distributions (*e.g.,*, in Karwa & Vadhan (2017); Kamath et al. (2019)).

Distributions with bounded moments. Let $\mathcal{M}_{k,m}$ be the family of distributions with $M_k(p) \le m$, we have

$$\max_{p \in \mathcal{M}_{k,m}} R_{\mathrm{SME}}(A, p) = \tilde{O}\left(\frac{M_k^{1/k}}{\sqrt{n}} + \frac{M_k^{1/k}}{(n\varepsilon)^{1-1/k}}\right).$$

This matches the optimal rate for distributions with bounded kth moment (*e.g.,*, in Kamath et al. (2020)). We list the detailed comparisons in Table 1.

Extending to higher dimensions. For (ε, δ)-DP mean estimation in the high-dimensional case, algorithms in Levy et al. (2021); Huang et al. (2021) rely on pre-processing techniques (*e.g.,* random rotation) and apply a one-dimensional estimation algorithm to each dimension. Our algorithm can also be combined with this procedure to obtain similar bounds since our algorithm provably provides an instance-optimal solution to each one-dimensional problem. We leave exploring better instance-specific bounds in high dimension as a direction for future work.

6. Conclusion

We proposed a new definition of instance optimality for differentially private estimation and showed that our notion of instance optimality is stronger than those proposed in prior work. We furthermore constructed private algorithms that achieve our notion of instance optimality when estimating a broad class of monotone properties. We also showed that our algorithm matches the asymptotic performance of prior work under a range of distributional assumptions on dataset generation.

References

Aldà, F. and Simon, H. U. On the optimality of the exponential mechanism. In *Cyber Security Cryptography and Machine Learning: First International Conference, CSCML 2017, Beer-Sheva, Israel, June 29-30, 2017, Proceedings 1*, pp. 68–85. Springer, 2017.

Asi, H. and Duchi, J. C. Instance-optimality in differential privacy via approximate inverse sensitivity mechanisms. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 14106–14117. Curran Associates, Inc., 2020a. URL https://proceedings.neurips.cc/paper/2020/file/a267f936e54d7c10a2bb70dbe6ad7a89-Paper.pdf.

Asi, H. and Duchi, J. C. Near instance-optimality in differential privacy. *arXiv preprint arXiv:2005.10630*, 2020b.

Balle, B. and Wang, Y.-X. Improving the gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *International Conference on Machine Learning*, pp. 394–403. PMLR, 2018.

Biswas, S., Dong, Y., Kamath, G., and Ullman, J. Coinpress: Practical private mean and covariance estimation. *Advances in Neural Information Processing Systems*, 33: 14475–14485, 2020.

Błasiok, J., Bun, M., Nikolov, A., and Steinke, T. Towards instance-optimal private query release. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 2480–2497. SIAM, 2019.

Brunel, V.-E. and Avella-Medina, M. Propose, test, release: Differentially private estimation with high probability. *arXiv preprint arXiv:2002.08774*, 2020.

Bun, M. and Steinke, T. Average-case averages: Private algorithms for smooth sensitivity and mean estimation. In *Advances in Neural Information Processing Systems*, pp. 181–191, 2019.

Cai, T. T., Wang, Y., and Zhang, L. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *arXiv preprint arXiv:1902.04495*, 2019.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. 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*, pp. 265–284. Springer, 2006.

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

Feldman, V. and Steinke, T. Calibrating noise to variance in adaptive data analysis. In *Conference On Learning Theory*, pp. 535–544. PMLR, 2018.

Fernandes, N., McIver, A., and Morgan, C. The laplace mechanism has optimal utility for differential privacy over continuous queries. In *2021 36th Annual ACM/IEEE Symposium on Logic in Computer Science (LICS)*, pp. 1–12. IEEE, 2021.

Hopkins, S. B., Kamath, G., and Majid, M. Efficient mean estimation with pure differential privacy via a sum-of-squares exponential mechanism. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2022, pp. 1406–1417, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392648. doi: 10.1145/3519935.3519947. URL https://doi.org/10.1145/3519935.3519947.

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

Kamath, G., Li, J., Singhal, V., and Ullman, J. Privately learning high-dimensional distributions. In *Conference on Learning Theory*, pp. 1853–1902. PMLR, 2019.

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

Karwa, V. and Vadhan, S. Finite sample differentially private confidence intervals, 2017.

Lehmann, E. L. Ordered families of distributions. *The Annals of Mathematical Statistics*, pp. 399–419, 1955.

Levy, D. A. N., Sun, Z., Amin, K., Kale, S., Kulesza, A., Mohri, M., and Suresh, A. T. Learning with user-level privacy. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=G1jmxFOtY_.

Mann, H. B. and Whitney, D. R. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pp. 50–60, 1947.

McMillan, A., Smith, A., and Ullman, J. Instance-optimal differentially private estimation. *arXiv preprint arXiv:2210.15819*, 2022.

Nissim, K., Raskhodnikova, S., and Smith, A. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pp. 75–84, 2007.

Smith, A. Privacy-preserving statistical estimation with optimal convergence rates. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pp. 813–822, 2011.

A. Proofs for Section 3.1

We will make use of the following characterization of rank-r thresholds.

Lemma A.1. *Let D be any multiset of real numbers. Then $\tau \in \mathbb{R}$ is a rank-r threshold for D if and only if every point $x \in L_r$ satisfies $x \le \tau$ and every point $x \in U_{|D|-r}$ satisfies $x \ge \tau$.*

Proof. First we show that if τ is a rank-r threshold for D, then every point $x \in L_r$ satisfies $x \le \tau$ and every point $x \in U_{|D|-r}$ satisfies $x \ge \tau$. By definition, we have that $\sum_{x \in D} \mathbb{I}\{x \le \tau\} \ge r$, which means that there are at least r points in D that are less than or equal to τ. Since L_r contains the r smallest points in D, every point in L_r must also be less than or equal to τ. Similarly, by definition we have that $r \ge \sum_{x \in D} \mathbb{I}\{x < \tau\} = |D| - \sum_{x \in D} \mathbb{I}\{x \ge \tau\}$, which means that there are at least $|D| - r$ points in D that are greater than or equal to τ. Since $U_{|D|-r}$ contains the largest $|D| - r$ points in D, they must all be greater than or equal to τ. This proves the first implication.

Now suppose that τ is a threshold with the property that every $x \in L_r$ satisfies $x \le \tau$ and every $x \in U_{|D|-r}$ satisfies $x \ge \tau$. We will show that this implies that τ is a rank-r threshold. Let $x \in D$ be any point with $x < \tau$. We know that all such points must belong to L_r (since $x \in U_{|D|-r}$ would violate our assumption). It follows that $\sum_{x \in D} \mathbb{I}\{x < \tau\} \le |L_r| = r$. Now let $x \in D$ be any point with $x > \tau$. We know that all such points must belong to $U_{|D|-r}$ (since $x \in L_r$ would violate our assumption). It follows that $\sum_{x \in D} \mathbb{I}\{x \le \tau\} = |D| - \sum_{x \in D} \mathbb{I}\{x > \tau\} \ge |D| - |U_{|D|-r}| = r$. Together, these arguments show that $\sum_{x \in D} \mathbb{I}\{x < \tau\} \le r$ and $\sum_{x \in D} \mathbb{I}\{x \le \tau\} \ge r$, which proves the second implication.

It follows that τ is a rank-r threshold if and only if every point $x \in L_r$ satisfies $x \le \tau$ and every point $x \in U_{|D|-r}$ satisfies $x \ge \tau$. □

Lemma A.2. *Let D be a multiset of real numbers and $\tau \in \mathbb{R}$ be any threshold. Then*

$$\operatorname{ranks}(\tau, D) = \left\{ \sum_{x \in D} \mathbb{I}\{x < \tau\}, \ldots, \sum_{x \in D} \mathbb{I}\{x \le \tau\} \right\}.$$

Proof. By definition, τ is a rank-r threshold if $\sum_{x \in D} \mathbb{I}\{x < \tau\} \le r$ and $\sum_{x \in D} \mathbb{I}\{x \le \tau\} \ge r$. Let $r_{\min} = \sum_{x \in D} \mathbb{I}\{x < \tau\}$ and $r_{\max} = \sum_{x \in D}\{x \le \tau\}$. We will show that τ is a rank-r threshold for D if and only if $r_{\min} \le r \le r_{\max}$.

First, since $r_{\min} = \sum_{x \in D} \mathbb{I}\{x < \tau\}$, we have that $\sum_{x \in D} \mathbb{I}\{x < \tau\} \le r$ if and only if $r_{\min} \le$. Similarly, since $r_{\max} = \sum_{x \in D} \mathbb{I}\{x \le \tau\}$, we have that $\sum_{x \in D} \mathbb{I}\{x \le \tau\} \ge r$ if and only if $r \le r_{\max}$. Since τ is a rank-r threshold if and only if both of the above inequalities hold, it follows that τ is a rank-r threshold iff $r_{\min} \le r \le r_{\max}$, as required. □

Theorem 3.7. *Fix data range $[a, b]$, $\varepsilon > 0$, and distance error $0 < \alpha \le \frac{b-a}{2}$. For any dataset $D \subset [a, b]$ and rank $1 \le r \le |D|$, let $\hat{\tau}$ be the output of Algorithm 2 run on D with parameters $[a, b]$, α, and ε. Publishing $\hat{\tau}$ satisfies ε-DP and, for any $\zeta > 0$, with probability at least $1 - \zeta$, $\hat{\tau}$ is an (α, β)-approximate rank-r threshold for D with $\beta = \frac{2}{\varepsilon} \log \frac{b-a}{\alpha\zeta}$. Moreover, Algorithm 2 can be implemented with $O(|D| \log |D|)$ running time.*

Proof. Algorithm 2 is an instance of the exponential mechanism, so to prove that it satisfies ε-differential privacy, it is sufficient to argue that the sensitivity of the loss $\ell(\tau)$ is bounded by one. Let D and D' be any neighboring datasets. Since adding or removing a point from D changes the value of each sum in the expression for $\operatorname{ranks}(\tau', D)$ given by Lemma A.2 by at most one, we are guaranteed that whenever $r' \in \operatorname{ranks}(\tau', D)$, then at least one of $r' - 1$, r', or $r' + 1$ belongs to $\operatorname{ranks}(\tau', D')$. From this, it follows that $|\operatorname{err}_{\text{rank}}(\tau', r, D) - \operatorname{err}_{\text{rank}}(\tau', r, D')| \le 1$. Taking the minimum of both sides of this inequality with respect to $\tau' \in [\tau - \alpha, \tau + \alpha]$ shows that the sensitivity of ℓ is bounded by one, as required.

Next we argue that there exists an interval $I^* \subset [a, b]$ of width at least α so that for every $\tau \in I^*$ we have $\ell(\tau) = 0$. Let $\tau^* \in [a, b]$ be any rank-r threshold for the dataset D. Next, define $I^* = [\tau^* - \alpha, \tau^* + \alpha] \cap [a, b]$. The width of I^* is at least α (since at least half of it is contained in $[a, b]$). Moreover, for every $\tau \in I^*$ we have that $\ell(\tau) = 0$, since τ is within distance α of an exact rank-r threshold, as required.

Finally, we follow the standard analysis of the exponential mechanism to prove that this is sufficient to find an approximate rank-r threshold. For any $c \ge 0$, define $S_c = \{\tau \in [a, b] \mid \ell(\tau) \ge c\}$ to be the set of thresholds whose loss is at least c. We

have that

$$\int_{\tau \in S_c} \exp\left(-\frac{\varepsilon}{2}\ell(\tau)\right) d\tau \le \int_{\tau \in S_c} \exp\left(-\frac{\varepsilon c}{2}\right) d\tau \le (b-a) \cdot \exp\left(-\frac{\varepsilon c}{2}\right).$$

On the other hand, we have

$$\int_a^b \exp\left(-\frac{\varepsilon}{2}\ell(\tau)\right) d\tau \ge \int_{I^*} \exp(0)\, d\tau \ge \alpha.$$

Together, it follows that

$$\Pr(\hat{\tau} \in S_c) = \int_{\tau \in S_c} f(\tau)\, d\tau \le \frac{b-a}{\alpha} \exp\left(-\frac{\varepsilon c}{2}\right).$$

Choosing $c = \frac{2}{\varepsilon} \log \frac{b-a}{\alpha\zeta}$ results in $\Pr(\hat{\tau} \in S_c) \le \zeta$.

It follows that with probability at least $1-\zeta$, we have that $\ell(\hat{\tau}) < \frac{2}{\varepsilon} \log \frac{b-a}{\alpha\zeta}$. From the definition of the loss, it follows that $\hat{\tau}$ is an (α, β)-approximate rank-r threshold for S.

Finally, we prove the running time guarantee. First, the loss function $\ell(\tau)$ is piecewise constant with at most $2|D|$ discontinuities. This is because, as we slide a threshold τ from left to right, the minimum rank error within the interval $[\tau - \alpha, \tau + \alpha]$ only changes when an endpoint of the interval crosses a datapoint in D, which can hapen at most $2|D|$ times. It follows that the output distribution of Algorithm 2 is also piecewise constant with discontinuities (potentially) occurring at $x \pm \alpha$ for each $x \in D$. Let the constant intervals be $I_1, \ldots, I_M$ and let $p_1, \ldots, p_M$ be the value of the exponential mechanism density on the intervals, respectively. Now let $\hat{\tau}$ be a sample from the output distribution and $\hat{I} \in \{I_1, \ldots, I_M\}$ be the interval that contains $\hat{\tau}$. The key idea behind the sampling strategy is to sample $\hat{I}$ first, and then sample $\hat{\tau}$ conditioned on the choice of $\hat{I}$. Since the density is constant on each interval $I_1, \ldots, I_M$, the second step is equivalent to outputting a uniformly random sample from $\hat{I}$. This works as long as the probability we choose $\hat{I} = I_i$ is equal to the marginal distribution of $\hat{I}$, which is given by $\mathbb{P}(\hat{I} = I_i) \propto \mathrm{width}(I_i) \cdot p_i$.

The running time of the above approach is dominated by the cost of computing the piecewise constant representation of the output density. This can be accomplished by constructing the set of $2|D|$ candidate discontinuity locations $x \pm \alpha$ for $x \in D$, sorting them, and then making a linear pass from left to right computing the constant intervals and the value of $\ell(\tau)$ on each interval. The overall running time of this is $O(|D| \log |D|)$. □

Corollary 3.8. *Let $\mathcal{Z} = \{z_1 \le \cdots \le z_m\}$ be such that $z_{i+1} = z_i + \gamma$ for all $i = 1, \ldots, m-1$ and let D be any multiset supported on $\mathcal{Z}$. Let $\hat{\tau}$ be the output of Algorithm 2 run on D with parameters $[a, b] = [z_1, z_m]$, $\alpha = \gamma/3$, and $\varepsilon > 0$, and let $\tilde{\tau}$ the closest point in $\mathcal{Z}$ to $\hat{\tau}$. Then for any $\zeta > 0$, with probability at least $1 - \zeta$ we have that $\tilde{\tau}$ is a $(0, \beta)$-approximate rank-r quantile for D with $\beta = \frac{2}{\varepsilon} \log \frac{3m}{\zeta}$.*

Proof. For any threshold τ, Lemma A.2 guarantees that

$$\mathrm{ranks}(\tau, D) = \left\{ \sum_{x \in D} \mathbb{I}\{x < \tau\}, \ldots, \sum_{x \in D} \mathbb{I}\{x \le \tau\} \right\}.$$

When the dataset D is supported on a grid $\mathcal{Z}$, moving τ to its nearest grid point never removes ranks from $\mathrm{ranks}(\tau, D)$, since the left sum is either the same or decreases, and the right sum is either the same or increases. This implies that moving τ to its closest grid point never increases its rank error.

By Theorem 3.7 we are guaranteed that with probability at least $1-\zeta$, there exists τ' with $|\hat{\tau} - \tau'| \le \alpha$ and $\mathrm{err}_{\mathrm{rank}}(\tau', r, D) \le \frac{2}{\varepsilon} \log \frac{b-a}{\alpha\zeta} \le \beta$ (where we used the fact that $(b-a)/\alpha \le 3m$). Assume this high probability event holds for the remainder of the proof.

Let $\tilde{\tau}$ and $\tilde{\tau}'$ be the closest grid points to $\hat{\tau}$ and τ', respectively. If $\tilde{\tau} = \tilde{\tau}'$ then we have that $\mathrm{err}_{\mathrm{rank}}(\tilde{\tau}, r, D) = \mathrm{err}_{\mathrm{rank}}(\tilde{\tau}', r, D) \le \mathrm{err}_{\mathrm{rank}}(\tau', r, D) \le \beta$ and we have shown that $\tilde{\tau}$ is a $(0, \beta)$-approximate rank-r threshold for D.

Now suppose that $\tilde{\tau} \ne \tilde{\tau}'$. First we argue that $D \cap [\hat{\tau}, \tau']$ must be the empty set. Suppose for contradiction that there is $x \in D \cap [\hat{\tau}, \tau']$. Then, x is a grid point, and we have that $\max\{|\hat{\tau} - x|, |\tau' - x|\} \le |\hat{\tau} - \tau'| \le \alpha = \gamma/3$. In particular, this

implies that both $\hat{\tau}$ and τ' are within distance $\gamma/3$ of the same grid point, which means we must have $\tilde{\tau} = \tilde{\tau}'$, which is a contradiction. Since there are no datapoints in $[\hat{\tau}, \tau']$, by Lemma A.2 we have that $\text{ranks}(\hat{\tau}, D) = \text{ranks}(\tau', D)$. It follows that $\text{err}_{\text{rank}}(\tilde{\tau}, r, D) \leq \text{err}_{\text{rank}}(\hat{\tau}, r, D) = \text{err}_{\text{rank}}(\tau', r, D) \leq \beta$.

In either case, we showed that the rank-error of $\tilde{\tau}$ is bounded by β. □

B. Proofs of mean estimation

B.1. Proof of Lemma 3.2

Upper bound. $c_{D,\varepsilon} \leq 1$**.** This can be seen since for all $D' \subset D'$, $|D'| \geq |D| - 1/\varepsilon$, we have

$$\mu(D \setminus L_{\frac{1}{\varepsilon}}) \leq \mu(D') \leq \mu(D \setminus U_{\frac{1}{\varepsilon}}).$$

Hence a fixed algorithm that outputs $\mu(D)$ will always have

$$|\mu(D) - \mu(D')| \leq \mu(D \setminus L_{\frac{1}{\varepsilon}}) - \mu(D \setminus U_{\frac{1}{\varepsilon}}).$$

Lower bound. $c_{D,\varepsilon} \geq 1/(2e^2)$**.** The proof follows from substituting $D_1 = D \setminus L_{\frac{1}{\varepsilon}}$ and $D_2 = D \setminus U_{\frac{1}{\varepsilon}}$ in Lemma 4.4.

B.2. Proof of Lemma 3.3

First we prove the privacy guarantee. Let D_1 and D_2 be any pair of datasets and let D_1' and D_2' be the shifted and clipped versions of them for the interval $[l, u]$. The size of the symmetric difference between D_1' and D_2' cannot be larger than between D_1 and D_2, so whenever D_1 and D_2 are neighbors, so are D_1' and D_2'. It follows that we can ignore the shifting and clipping step in the privacy analysis.

Since the add/remove sensitivity of n is 1, step 3 estimates n using the Laplace mechanism with a budget of $\varepsilon/2$. The add/remove sensitivity of the sum s of the shifted data is $w/2$, so step 4 estimates s using the Laplace mechanism with a budget of $\varepsilon/2$. The overall privacy guarantee of the algorithm then follows from basic composition and post-processing, since w and m are public quantities (i.e., they depend on the algorithm parameters, not on the actual dataset).

Now we turn to the utility analysis. Recall that $n = |D'| = |D|$. Since $D' = D - m$,

$$\hat{\mu} - \mu(D) = \text{clip}\left(\frac{\hat{s}}{\hat{n}}, \left[-\frac{w}{2}, \frac{w}{2}\right]\right) - \mu(D').$$

Since all all elements of D' lie in $[-\frac{w}{2}, \frac{w}{2}]$,

$$\left|\text{clip}\left(\frac{\hat{s}}{\hat{n}}, \left[-\frac{w}{2}, \frac{w}{2}\right]\right) - \mu(D')\right| \leq \left|\frac{\hat{s}}{\hat{n}} - \mu(D')\right| = \left|\frac{\hat{s}}{\hat{n}} - \frac{s}{n}\right|.$$

Hence, we can bound the desired expectation as

$$\begin{aligned}
\mathbb{E}\left[|\hat{\mu} - \mu(D)|\right] &= \Pr(Z_n < -n/2)\mathbb{E}\left[|\hat{\mu} - \mu(D)||Z_n < -n/2\right] + \Pr(Z_n \geq -n/2)\mathbb{E}\left[|\hat{\mu} - \mu(D)||Z_n \geq -n/2\right] \\
&= \Pr(Z_n < -n/2)\mathbb{E}\left[|\hat{\mu} - \mu(D)||Z_n < -n/2\right] + \mathbb{E}\left[\left|\frac{\hat{s}}{\hat{n}} - \frac{s}{n}\right| |Z_n \geq -n/2\right] \\
&\leq \frac{1}{2}e^{-n\varepsilon/4}|l - u| + \mathbb{E}\left[\left|\frac{\hat{s}}{\hat{n}} - \frac{s}{n}\right| |Z_n \geq -n/2\right], \qquad (4)
\end{aligned}$$

where the last inequality follows by observing that both $\hat{\mu}$ and $\mu(D)$ lie in $[l, u]$. We now simplify the second term in (4).

$$\begin{aligned}
\mathbb{E}\left[\left|\frac{\hat{s}}{\hat{n}} - \frac{s}{n}\right| | Z_n \geq -n/2\right] &= \mathbb{E}\left[\left|\frac{s + Z_s}{n + Z_n} - \frac{s}{n}\right| | Z_n \geq -n/2\right] \\
&= \mathbb{E}\left[\left|\frac{nZ_s}{(n + Z_n)n}\right| | Z_n \geq -n/2\right] \\
&= \mathbb{E}\left[\left|\frac{Z_s}{(n + Z_n)}\right| | Z_n \geq -n/2\right] \\
&\overset{(a)}{=} \mathbb{E}[|Z_s|] \cdot \mathbb{E}\left[\left|\frac{1}{(n + Z_n)}\right| | Z_n \geq -n/2\right] \\
&\leq \mathbb{E}[|Z_s|] \cdot \frac{2}{n} \\
&= \frac{(u - l)}{\varepsilon} \cdot \frac{2}{n}, \qquad (5)
\end{aligned}$$

where (a) uses the fact that Z_s and Z_n are independent of each other. Combining (4) and (5) yields,

$$\mathbb{E}\left[|\hat{\mu} - \mu(D)|\right] \leq \frac{1}{2} e^{-n\varepsilon/4} |l - u| + \frac{2(u - l)}{n\varepsilon} \leq \frac{(u - l)}{n\varepsilon}\left(\frac{2}{e} + 2\right) \leq \frac{3(u - l)}{n\varepsilon},$$

where the penultimate inequality uses the fact that $e^{-x}x \leq e^{-1}$ for all $x \geq 0$.

B.3. Proof of Theorem 3.1

Our goal is to show that the expected error of Algorithm 3 is not much larger than

$$R(D, \varepsilon) \geq \frac{\mu(D \setminus L_{\frac{1}{\varepsilon}}) - \mu(D \setminus U_{\frac{1}{\varepsilon}})}{2e^2}.$$

With probability at least $1 - \zeta$, by Theorem 3.7 we are guaranteed that $\hat{l}$ and $\hat{u}$ are (α, β)-approximate rank-t_l and rank-t_u thresholds, respectively for the values of α and β defined in Algorithm 3. In particular, this implies that there exist l' and t'_l such that $|\hat{l} - l'| \leq \alpha$, $|t'_l - t_l| \leq \beta$, and l' is a rank-t'_l threshold for D. Similarly, there exist u' and t'_u such that $|\hat{u} - u'| \leq \alpha$, $|t'_u - t_u| \leq \beta$, and u' is a rank-t'_u threshold for D. Let G denote this high probability event. We first argue that conditioned on G, the expected loss of $\hat{\mu}$ is small (where the expectation is taken only over the randomness of Algorithm 1). To convert this bound into a bound that holds in expectation, we bound the error when G does not hold by $2R$.

Let $\hat{\mu}$ be the output of Algorithm 3. We decompose the error of $\hat{\mu}$ into three terms:

$$\begin{aligned}
|\hat{\mu} - \mu(D)| \leq\ & |\hat{\mu} - \mu(\text{clip}(D, [\hat{l}, \hat{u}]))| \\
&+ |\mu(\text{clip}(D, [\hat{l}, \hat{u}]) - \mu(\text{clip}(D, [l', u'])| \\
&+ |\mu(\text{clip}(D, [l', u'])) - \mu(D)|
\end{aligned}$$

Roughly speaking, the first term captures the variance incurred by using Algorithm 1 to estimate the mean of the clipped data, the second term measures our bias due to α in our (α, β)-approximate thresholds, and the third term measures the bias due to β. Our goal is to prove that all of these terms are not much larger than $R(D, \varepsilon)$.

Bounding first term. At a high level, we argue that all points in $L_{\frac{1}{\varepsilon}}$ are to the left of $\hat{l} + \alpha$, and all points in $R_{\frac{1}{\varepsilon}}$ are to the right of $\hat{u} - \alpha$. It follows that the distance from any point in $L_{\frac{1}{\varepsilon}}$ to any point in $U_{\frac{1}{\varepsilon}}$ is at least $\hat{u} - \hat{l} - 2\alpha$. In particular, this guarantees that the difference between the means of $D \setminus L_{\frac{1}{\varepsilon}}$ and $D \setminus U_{\frac{1}{\varepsilon}}$ must be at least $\frac{\hat{u} - \hat{l} - 2\alpha}{\varepsilon(|D| - \frac{1}{\varepsilon})}$, since we move $1/\varepsilon$ points a distance at least $\hat{u} - \hat{l} - 2\alpha$. This expression is close to the loss incurred by Algorithm 1 when run on the clipped dataset.

Formally, let $S = (D \setminus L_{\frac{1}{\varepsilon}}) \cap (D \setminus U_{\frac{1}{\varepsilon}})$ be the set of common points in the two means from the lower bound. Then we

have that

$$\begin{aligned}\mu(D \setminus L_{\frac{1}{\varepsilon}}) - \mu(D \setminus U_{\frac{1}{\varepsilon}}) &= \frac{1}{|D| - \frac{1}{\varepsilon}} \left(\sum_{x \in S} x + \sum_{x \in U_{\frac{1}{\varepsilon}}} x - \sum_{x \in S} x - \sum_{x \in L_{\frac{1}{\varepsilon}}} x \right) \\ &= \frac{1}{|D| - \frac{1}{\varepsilon}} \left(\sum_{x \in U_{\frac{1}{\varepsilon}}} x - \sum_{x \in L_{\frac{1}{\varepsilon}}} x \right)\end{aligned}$$

Next, since l' is a rank-t'_l threshold with $t'_l \geq \frac{1}{\varepsilon}$, we know that every $x \in L_{t'_l} \supset L_{\frac{1}{\varepsilon}}$ satisfies $x \leq l' \leq \hat{l} + \alpha$. Similarly, since u' is a rank-t'_u threshold with $t'_u \leq |D| + \beta = |D| - \frac{1}{\varepsilon}$, we are guaranteed that every $x \in U_{|D|-t'_u} \supset U_{|D|-|D|+\frac{1}{\varepsilon}} = U_{\frac{1}{\varepsilon}}$ satisfies $x \geq u' \geq \hat{u} - \alpha$. Substituting these bounds into the above expression gives

$$\begin{aligned}2e^2 R(D, \varepsilon) &\geq \mu(D \setminus L_{\frac{1}{\varepsilon}}) - \mu(D \setminus R_{\frac{1}{\varepsilon}}) \\ &\geq \frac{\hat{u} - \hat{l} - 2\alpha}{\varepsilon |D| - 1} \\ &\geq \frac{\hat{u} - \hat{l}}{\varepsilon |D|} - \frac{2\alpha}{\varepsilon |D|}.\end{aligned}$$

By Lemma 3.3, we have that the expectation of the first term in the error decomposition conditioned on the choice of $\hat{l}$ and $\hat{u}$ is bounded by $\frac{3(\hat{u}-\hat{l})}{|D|\varepsilon}$. It follows that

$$\mathbb{E}\left[|\hat{\mu} - \mu(\mathrm{clip}(D, [\hat{l}, \hat{u}]))| \,\middle|\, G \right] \leq 6e^2 R(D, \varepsilon) + \frac{6\alpha}{\varepsilon |D|}.$$

Bounding the second term. The key idea behind bounding the second term is that, whenever $\hat{l}$ is close to l' and $\hat{u}$ is close to u', then clipping a point x to $[\hat{l}, \hat{u}]$ is approximately the same as clipping it to $[l', u']$. Formally, we have

$$\begin{aligned}|\mu(\mathrm{clip}(D, [\hat{l}, \hat{u}])) - \mu(\mathrm{clip}(D, [l', u']))| &\leq \frac{1}{|D|} \sum_{x \in D} |\,\mathrm{clip}(x, [\hat{l}, \hat{u}]) - \mathrm{clip}(x, [l', u'])| \\ &= \frac{1}{|D|} \sum_{x \in D} |\min(\hat{u}, \max(x, \hat{l})) - \min(u', \max(x, l'))| \\ &\leq \frac{1}{|D|} \sum_{x \in D} 2\alpha \\ &= 2\alpha\end{aligned}$$

Bounding the third term. Our bound for the third term is the most involved. At a high level, we show that the bias introduced by clipping D to the interval $[l', u']$ is at most the worst "one-sided" clipping bias incurred clipping the points to the left of l' or to the right of u'. To see this, observe that when we clip from both sides, the left and right biases cancel out. Next, we argue that clipping points to the left of l' (or to the right of u') introduces less bias than *removing* those points. This step bridges the gap between Algorithm 1 which clips points and the lower bound on $R(D, \varepsilon)$, which removes points. We argue that the number of points removed to the left of l' or to the right of u' is not much larger than $\frac{1}{\varepsilon}$ and use Lemma B.1 to show that the resulting bias is not much larger than if we had removed exactly $\frac{1}{\varepsilon}$ points instead. Finally, to finish the bound, combine our two "one-sided" bias bounds to show that the overall bias is never much larger than $R(D, \varepsilon)$.

We begin by showing that the bias is bounded by the worst "one-sided" bias. We have that

$$\mu(D) = \frac{1}{|D|} \left(\sum_{\substack{x \in D \\ x < l'}} x + \sum_{\substack{x \in D \\ l' \leq x \leq u'}} x + \sum_{\substack{x \in D \\ x > u'}} x \right)$$

and

$$\mu(\text{clip}(D,[l',r'])) = \frac{1}{|D|}\left(\sum_{\substack{x\in D\\ x<l'}} l' + \sum_{\substack{x\in D\\ l'\le x\le u'}} x + \sum_{\substack{x\in D\\ x>u'}} u'\right).$$

Therefore, we have that

$$\begin{aligned}|\mu(\text{clip}(D,[l',u'])) - \mu(D)| &= \frac{1}{|D|}\left|\sum_{\substack{x\in D\\ x<l'}} l' - x + \sum_{\substack{x\in D\\ x>u'}} u' - x\right| \\ &\le \frac{1}{|D|}\max\left\{\sum_{\substack{x\in D\\ x<l'}} l' - x, \sum_{\substack{x\in D\\ x>u'}} x - u'\right\},\end{aligned}$$

where the inequality follows because the two sums have opposite signs. This expression is the maximum bias we introduce if we only cliped either points to the left of l' or to the right of u'.

Next, we relate the bias of clipping the points to the left of l' to the bias of removing those points instead. Let $q = \sum_{x\in D}\mathbb{I}\{x < l'\}$ be the number of points that are clipped to l'. Next, since adding copies of $\mu(D\setminus L_q)$ to $D\setminus L_q$ does not change its mean, we have that

$$\begin{aligned}\mu(D\setminus L_q) - \mu(D) &= \frac{1}{|D|}\left(\sum_{x\in D\setminus L_q}(x-x) + \sum_{x\in L_q}\mu(D\setminus L_q) - x\right) \\ &= \frac{1}{|D|}\sum_{x\in L_q}\mu(D\setminus L_q) - x \\ &\ge \frac{1}{|D|}\sum_{x\in L_q} l' - x,\end{aligned}$$

where the final inequality follows from the fact that $\mu(D\setminus L_q)\ge l'$, since every element of $D\setminus L_q$ is at least l'. We have shown that the bias from clipping the points to l' is at most the bias from deleting them.

Next, we use the fact that l' is a rank-t_l' threshold for D to argue that the number of points clipped, q, cannot be too large. Since l' is a rank-t_l' threshold for D, we know that every $x\in U_{|D|-t_l'}$ satisfies $x\ge l'$. Therefore,

$$\sum_{x\in D}\mathbb{I}\{x\ge\tau_l'\} \ge |U_{|D|-t_l'} = |D| - t_l'.$$

Since $q = |D| - \sum_{x\in D}\mathbb{I}\{x\ge\tau'\}$, we have that $q\le t_l'\le\frac{1}{\varepsilon}+2\beta$. Putting these bounds together and using Lemma B.1 to handle the fact that q may be larger than $\frac{1}{\varepsilon}$, we have

$$\begin{aligned}\frac{1}{|D|}\sum_{x\in L_q} l' - x &\le \mu(D\setminus L_q) - \mu(D) \\ &\le \frac{2q}{1/\varepsilon}\big(\mu(D\setminus L_{\frac{1}{\varepsilon}}) - \mu(D)\big) \\ &\le (2+4\beta\varepsilon)\big(\mu(D\setminus L_{\frac{1}{\varepsilon}}) - \mu(D)\big)\end{aligned}$$

Next we turn to the bias of clipping points to the right of u'. Let $p = \sum_{x\in D}\mathbb{I}\{x>u'\}$ be the number of points in D that are strictly greater than u'. A similar argument to the above shows that $p\le\frac{1}{\varepsilon}+2\beta$ and that

$$\frac{1}{|D|}\sum_{x\in U_q} x - u' \le \mu(D) - \mu(D\setminus U_p) \le (2+4\beta\varepsilon)\big(\mu(D) - \mu(D\setminus R_{\frac{1}{\varepsilon}})\big).$$

Putting it all together, the third term of the error decomposition is upper bounded by

$$\begin{aligned}
&|\mu(\mathrm{clip}(D,[l',u'])) - \mu(D)| \\
&\le (2+4\beta\varepsilon)\max\{\mu(D\setminus L_{\frac{1}{\varepsilon}}) - \mu(D), \mu(D) - \mu(D\setminus R_{\frac{1}{\varepsilon}})\\
&\le (2+4\beta\varepsilon)\big(\mu(D\setminus L_{\frac{1}{\varepsilon}}) - \mu(D\setminus U_{\frac{1}{\varepsilon}})\big)\\
&\le 2e(2+4\beta\varepsilon)R(D,\varepsilon),
\end{aligned}$$

where the second inequality follows from the fact that the maximum of two numbers is not larger than the sum.

Finally, conditioned on the good event G, we have shown that the expected loss of $\hat{\mu}$ is bounded by

$$\big(2e(2+4\beta\varepsilon^2)+6e^2\big)R(D,\varepsilon) + \alpha\left(\frac{6}{\varepsilon|D|}+2\right).$$

Using the fact that $\beta = \frac{2}{\varepsilon}\log\frac{2R}{\alpha\zeta}$ and bounding the error when the good event G fails to hold by $2R$, we have that

$$\begin{aligned}
&\mathbb{E}\left[|\hat{\mu}-\mu(D)|\right]\\
&\le \left(56+44\log\frac{2R}{\alpha\zeta}\right)R(D,\varepsilon) + \alpha\left(\frac{6}{\varepsilon|D|}+1\right) + 2R\zeta,
\end{aligned}$$

as required.

Lemma B.1. *Let D be any multiset of real numbers of size n, and let $n_1 \le n_2 \le n/2$. Then we have*

$$\mu(D\setminus L_{n_2}) - \mu(D) \le \frac{2n_2}{n_1}\big(\mu(D\setminus L_{n_1}) - \mu(D)\big)$$

and

$$\mu(D) - \mu(D\setminus R_{n_2}) \le \frac{2n_2}{n_1}\big(\mu(D) - \mu(D\setminus R_{n_1})\big)$$

Proof. We only prove the first inequality and the second will follow similarly. For any $n' \le n/2$, we have

$$\begin{aligned}
\big|\mu_D - \mu_{D\setminus L_{n'}}\big| &= \frac{1}{n}\sum_{i\in[n]}(\mu_{D\setminus L_{n'}} - X_i)\mathbf{1}\{X_i \in L_{n'}\}\\
&= \sum_{i\in[n]}(\mu_{D\setminus L_{n'}} - X_i)\mathbf{1}\{X_i < a\} + (n-n')\mu_{D\setminus L_{n'}} - (n-n')\mu_{D\setminus L_{n'}}\\
&= n'\mu_{D\setminus L_{n'}} + (n-n')\mu_{D\setminus L_{n'}} - \sum_{i\in[n]}\mathbf{1}\{X_i < a\}X_i - \sum_{i\in[n]}\mathbf{1}\{X_i \ge a\}X_i\\
&= n\big(\mu_{D\setminus L_{n'}} - \mu_D\big).
\end{aligned}$$

Setting n' to be n_1, n_2 respectively, it would be enough to prove that

$$\mu_{D\setminus L_{n_2}} - \mu_{L_{n_2}} \le 2\Big(\mu_{D\setminus L_{n_1}} - \mu_{L_{n_1}}\Big).$$

This follows since

$$\begin{aligned}
2\Big(\mu_{D\setminus L_{n_1}} - \mu_{L_{n_1}}\Big) - \Big(\mu_{D\setminus L_{n_2}} - \mu_{L_{n_2}}\Big) &\ge 2\mu_{D\setminus L_{n_1}} - \mu_{D\setminus L_{n_2}} - \mu_{L_{n_1}}\\
&= 2\mu_{D\setminus L_{n_1}} - \frac{(n-n_1)\mu_{D\setminus L_{n_1}} - \sum_{i=n_1+1}^{n_2}X_i}{n-n_2} - \mu_{L_{n_1}}\\
&= \left(2-\frac{n-n_1}{n-n_2}\right)\mu_{D\setminus L_{n_1}} + \frac{n_2-n_1}{n-n_2}\mu_{L_{n_2}\setminus L_{n_1}} - \mu_{L_{n_1}}\\
&\ge 0.
\end{aligned}$$

□

C. Inverse sensitivity mechanism

In this section, we state the inverse sensitivity mechanism and its guarantee in the add/remove model of DP. Most of the proof follows from (Asi & Duchi, 2020a) in the replacement model of DP. We include its add/remove variant here for completeness.

We first provide a few definitions. Let D be a dataset supported over $\mathcal{Z}$ and $\forall k \in \mathbb{N}_+$, let $\omega_\ell(D,k)$ be defined as

$$\omega_\ell(D,k) := \max\left\{\ell(\theta(D),\theta(D')) \mid D' \in \mathcal{Z}^n, d(D,D') \le k\right\}$$

i.e., the maximum change in the parameter by k add/remove operations on D. Let

$$\mathrm{len}_\theta(D,t) := \min\left\{d(D',D) \mid \theta(D') = t\right\},$$

which is the minimum number of add/remove operations needed to change D to some D' with $\theta(D') = t$. Then the add/remove version of inverse sensitivity mechanism is stated below.

Algorithm 5 Inverse Sensitivity Mechanism (Asi & Duchi, 2020a)

Input: Range R, dataset $D \subset [-R,R]$, privacy parameter $\varepsilon > 0$ and granularity $\beta > 0$.

1: Let $\mathcal{T}$ be the set of points in $[-R,R]$ that are also mulitples of β.
2: Output $t \in \mathcal{T}$ with distribution

$$\Pr\left(A(D) = t\right) = \frac{\exp\left(-\mathrm{len}_\theta(D,t)\right)}{\sum_{t' \in \mathcal{T}} \exp\left(-\mathrm{len}_\theta(D,t')\right)},$$

The following guarantee holds for the inverse sensitivity mechanism.

Theorem C.1 ((Asi & Duchi, 2020a)). *For any $\beta \in (0,B)$, the inverse sensitivity mechanism A with privacy parameter $\varepsilon' = 2\varepsilon \log \frac{2BR}{\beta}$ satisfies that*

$$\mathbb{E}\left[\ell(A(D),\theta(D))\right] \le \omega_\ell(D,1/\varepsilon) + L\beta.$$

D. Proofs for the general algorithm

D.1. Proof of the Lemma 4.4

For any algorithm A,

$$\begin{aligned}\max_{D' \subseteq D : |D'| \ge |D| - 1/\varepsilon} \mathbb{E}\left[\ell(A(D'),\theta(D'))\right] &\ge \max_{(D_1,D_2)\in S} \max\left(\mathbb{E}\left[\ell(A(D_1),\theta(D_1))\right], \mathbb{E}\left[\ell(A(D_2),\theta(D_2))\right]\right) \\ &\ge \max_{(D_1,D_2)\in S} 0.5 \cdot \left(\mathbb{E}\left[\ell(A(D_1),\theta(D_1))\right] + \mathbb{E}\left[\ell(A(D_2),\theta(D_2))\right]\right).\end{aligned}$$

Since $d(D_1,D_2) \le 2/\epsilon$, if $A \in A_\epsilon$,

$$\begin{aligned}\mathbb{E}\left[\ell(A(D_1),\theta(D_1))\right] + \mathbb{E}\left[\ell(A(D_2),\theta(D_2))\right] &\ge e^{-(2/\varepsilon)\varepsilon}\mathbb{E}\left[\ell(A(D_2),\theta(D_1))\right] + \mathbb{E}\left[\ell(A(D_2),\theta(D_2))\right] \\ &\ge e^{-2}\left(\mathbb{E}\left[\ell(A(D_2),\theta(D_1))\right] + \mathbb{E}\left[\ell(A(D_2),\theta(D_2))\right]\right) \\ &\ge e^{-2}\ell(\theta(D_1),\theta(D_2)) \\ &\ge e^{-2}\ell(\theta(D_1),\theta(D_2)).\end{aligned}$$

Hence, for any algorithm $A \in A_\epsilon$,

$$\max_{D' \subseteq D : |D'| \ge |D| - 1/\varepsilon} \mathbb{E}\left[\ell(A(D'),\theta(D'))\right] \ge \max_{\max_{(D_1,D_2)\in S}} \frac{1}{2e^2}\ell(\theta(D_1),\theta(D_2)).$$

D.2. Proof of Theorem 4.3

We first prove a general result on stochastic dominance which will be helpful later in our results. For a dataset D and thresholds l,u, let $D_{[l,u]} = D \cap [l,u]$.

Lemma D.1. *Let l and u satisfy,*

$$2r \geq |D \cap (-\infty, l)| \geq \frac{3}{2}r,$$
$$2r \geq |D \cap (u, \infty)| \geq \frac{3}{2}r.$$

For all D' such that all of their elements are in $[l, u]$ and $d(D_{[l,u]}, D') \leq r/2$,

$$D \setminus L_{4r}(D) \succ D' \succ D \setminus U_{4r}(D).$$

Proof. Let L_{4r} denote $L_{4r}(D)$ and U_{4r} denote $U_{4r}(D)$ for convenience. We present the proof for $D \setminus L_{4r} \succ D'$ and the other relation will follow similarly. We divide the proof into three cases depending on the value of v in Definition 4.1.

Case $v < \min_{x \in D \setminus L_{4r}} x$**:** Since there are no points in $D \setminus L_{4r}$ in this range,

$$\frac{\sum_{x \in D \setminus L_{4r}} \mathbb{1}\{x \leq v\}}{|D \setminus L_{4r}|} = 0 \leq \frac{\sum_{x \in D'} \mathbb{1}\{x \leq v\}}{|D'|}.$$

Case $v \geq u$**:** Since all points of D' lie below u,

$$\frac{\sum_{x \in D \setminus L_{4r}} \mathbb{1}\{x \leq v\}}{|D \setminus L_{4r}|} \leq 1 = \frac{\sum_{x \in D'} \mathbb{1}\{x \leq v\}}{|D'|}$$

Case $v \in [\min_{x \in D \setminus L_{4r}} x, u)$**:** Since $d(D_{[l,h]}, D') \leq r/2$,

$$\begin{aligned}\frac{\sum_{x \in D'} \mathbb{1}\{x \leq v\}}{|D'|} &\geq \frac{\sum_{x \in D_{[l,u]}} \mathbb{1}\{x \leq v\} - r/2}{|D_{[l,u]}| + r/2} \\ &\geq \frac{\sum_{x \in D_{[l,u]}} \mathbb{1}\{x \leq v\} - r/2}{|D \setminus L_{4r}| + 3r/2},\end{aligned}$$

where the last inequality follows by observing that the assumptions in the lemma imply,

$$|D_{[l,u]}| + r/2 \leq |D| - 5r/2 \leq |D \setminus L_{4r}| + 3r/2.$$

For $v \in [\min_{x \in D \setminus L_{4r}} x, u)$,

$$\begin{aligned}\sum_{x \in D_{[l,u]}} \mathbb{1}\{x \leq v\} - r/2 &\geq \sum_{x \in D} \mathbb{1}\{x \leq v\} - 2r - r/2 \\ &= \sum_{x \in D} \mathbb{1}\{x \leq v\} - 5r/2 \\ &= \sum_{x \in D \setminus L_{4r}} \mathbb{1}\{x \leq v\} + 4r - 5r/2 \\ &= \sum_{x \in D \setminus L_{4r}} \mathbb{1}\{x \leq v\} + 3r/2.\end{aligned}$$

Hence,

$$\begin{aligned}\frac{\sum_{x \in D'} \mathbb{1}\{x \leq v\}}{|D'|} &\geq \frac{\sum_{x \in D \setminus L_{4r}} \mathbb{1}\{x \leq v\} + 3r/2}{|D \setminus L_{4r}| + 3r/2} \\ &\geq \frac{\sum_{x \in D \setminus L_{4r}} \mathbb{1}\{x \leq v\}}{|D \setminus L_{4r}|}.\end{aligned}$$

□

Proof of Theorem 4.3. In this proof, we show that Algorithm 4 is ε-DP and achieves

$$\mathbb{E}[\ell(A(D), \theta(D))] \le 2e^2 R(D, \varepsilon') + 7L\beta,$$

where $\varepsilon' = \frac{\varepsilon}{128\log(6RB/L\beta^2)}$. Theorem 4.3 can be obtained by applying Algorithm 4 with privacy parameter $128\log(6RB/L\beta^2)\varepsilon$.

By composition theorem, the overall privacy budget is ε. In the rest of the proof, we focus on the utility guarantee. We first quantify the effect of quantization. Observe that the output of the algorithm does not change for inputs D and the corresponding quantized dataset D_{quant}. Hence together with the Lipschitz property of θ,

$$\begin{aligned}\mathbb{E}[\ell(A(D), \theta(D)] &= \mathbb{E}[\ell(A(D_{\text{quant}}), \theta(D)] \\ &\le \mathbb{E}[\ell(A(D_{\text{quant}}), \theta(D_{\text{quant}})] + \ell(\theta(D_{\text{quant}}), \theta(D)), \\ &\le \mathbb{E}[\ell(A(D_{\text{quant}}), \theta(D_{\text{quant}})] + L\beta.\end{aligned}$$

By Corollary 3.8, with probability at least $1-\eta$, there exists a r' such that $|r' - 7r/4| \le \frac{8}{\varepsilon}\log\frac{6R}{\eta\beta}$ and $r' \in R(l, D_{\text{quant}})$. In other words,

$$R(l, D_{\text{quant}}) \cap [3r/2, 2r] \neq \emptyset.$$

Hence,

$$|D_{\text{quant}} \cap (-\infty, l)| \le 2r,$$

and hence,

$$3r/2 \le |D'_{\text{quant}} \cap (-\infty, l)| \le 2r.$$

Similarly,

$$3r/2 \le |D'_{\text{quant}} \cap (u, \infty)| \le 2r.$$

Let E be the event where both the above equations hold. By triangle inequality,

$$\mathbb{E}[\ell(A(D'_{\text{quant}}), \theta(D'_{\text{quant}})] \le \mathbb{E}[\ell(A(D'_{\text{quant}}), \theta(D_{[l,h]})] + \mathbb{E}[\ell(\theta(D'_{\text{quant}}), \theta(D_{[l,h]})], .$$

Let $L'_{4r} = L_{4r}(D'_{\text{quant}})$ and $U'_{4r} = U_{4r}(D'_{\text{quant}})$. By Lemma D.1, $D'_{\text{quant}} \setminus L_{4r} \succ D_{[l,u]} \succ D'_{\text{quant}} \setminus U_{4r}$. Hence, conditioned on the event E, by the monotonicity property

$$\mathbb{E}[\ell(\theta(D'_{\text{quant}}), \theta(D_{[l,h]})] \le \max\left(\ell(\theta(D'_{\text{quant}}), \theta(D'_{\text{quant}} \setminus L'_{4r}), \ell(\theta(D'_{\text{quant}}), \theta(D'_{\text{quant}} \setminus U'_{4r})\right).$$

Furthermore by triangle inequality,

$$\begin{aligned}\ell(\theta(D_{\text{quant}}), \theta(D'_{\text{quant}} \setminus L'_{4r}) &\le \ell(\theta(D), \theta(D \setminus L_{4r}) + \ell(\theta(D), D'_{\text{quant}})) + \ell(\theta(D \setminus L_{4r}, \theta(D'_{\text{quant}} \setminus L'_{4r}) \\ &\le \ell(\theta(D), \theta(D \setminus L_{4r}) + 4L\beta.\end{aligned}$$

Similarly,

$$\ell(\theta(D'_{\text{quant}}), \theta(D'_{\text{quant}} \setminus U'_{4r}) \le \ell(\theta(D), \theta(D \setminus U_{4r}) + 4L\beta.$$

Hence,

$$\begin{aligned}\ell(\theta(D'_{\text{quant}}), \theta(D_{[l,h]}) &\le \max\left(\ell(\theta(D), \theta(D \setminus L'_{4r}), \ell(\theta(D), \theta(D \setminus U'_{4r})\right) + 4L\beta \\ &\le \max_{D_1, D_2 \subseteq D : d(D_1, D_2) \le 4r} \ell(\theta(D_1), \theta(D_2)) + 4L\beta.\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{E}[\ell(\theta(D_{\text{quant}}), \theta(D_{[l,h]})] &\le \max_{D_1, D_2 \subseteq D : d(D_1, D_2) \le 4r} \ell(\theta(D_1), \theta(D_2)) + 4L\beta + \Pr(E)B \\ &= \max_{D_1, D_2 \subseteq D : d(D_1, D_2) \le 4r} \ell(\theta(D_1), \theta(D_2)) + 6L\beta.\end{aligned}$$

Since inverse sensitivity mechanism is applied on $D_{[l,u]}$,

$$\mathbb{E}[\ell(A(D'_{\text{quant}}), \theta(D_{[l,h]})] = \mathbb{E}[\ell(\text{InvSen}(D_{[l,h]}), \theta(D_{[l,h]})].$$

By the guarantee of the inverse sensitivity mechanism,

$$\mathbb{E}[\ell(\text{InvSen}(D_{[l,h]}), \theta(D_{[l,h]})] \leq \max_{D_1, D_2 \subseteq D: d(D_1, D_2) \leq r'} \ell(\theta(D_1), \theta(D_2)) + L\beta,$$

where $r' = \left(\frac{4\log((8BR)/L\beta^2)}{\varepsilon}\right)$. Combining the above equations yield

$$\begin{aligned}\mathbb{E}[\ell(A(D), \theta(D))] &\leq 2 \max_{D_1, D_2 \subseteq D: d(D_1, D_2) \leq \max(4r, r')} \ell(\theta(D_1), \theta(D_2)) + 7L\beta \\ &\leq 2eR(D, \varepsilon') + 7L\beta,\end{aligned}$$

where $\varepsilon' = \frac{\varepsilon}{128\log(6RB/L\beta^2)}$. □

E. Proofs for statistical mean estimation

E.1. Proof of Corollary 5.3

It is straightforward to see that

$$\mathbb{E}\left[|\mu(X^n) - \mu|\right] \leq \sqrt{\mathbb{E}\left[|\mu(X^n) - \mu|^2\right]} = \sqrt{\frac{M_2(p)}{n}}$$

Hence it remains to show that $\forall k \geq 2$,

$$\mathbb{E}\left[\left|\mu(X^n \setminus L_{\frac{1}{\epsilon}}) - \mu(X^n \setminus U_{\frac{1}{\epsilon}})\right|\right] O\left(\sqrt{\frac{M_2(p)}{n}} + \frac{M_k(p)^{1/k}}{(n\varepsilon)^{1-1/k}}\right). \tag{6}$$

Our proof will be based on the lemma below.

Lemma E.1. *For any sequence of samples X^n and $k \geq 2$, let $\widehat{M}_k(X^n) := \frac{1}{n}\sum_{i=1}^n |X_i - \mu(X^n)|^k$, we have*

$$\left|\mu(X^n \setminus L_{\frac{1}{\epsilon}}) - \mu(X^n \setminus U_{\frac{1}{\epsilon}})\right| = O\left(\frac{\widehat{M}_k(X^n)^{1/k}}{(n\varepsilon)^{1-1/k}}\right)$$

We first prove Equation (6) based on Lemma E.1 and then present the proof of Lemma E.1.

$$\mathbb{E}\left[\left|\mu(X^n \setminus L_{\frac{1}{\epsilon}}) - \mu(X^n \setminus U_{\frac{1}{\epsilon}})\right|\right] = O\left(\mathbb{E}\left[\frac{\widehat{M}_k(X^n)^{1/k}}{(n\varepsilon)^{1-1/k}}\right]\right)$$

Moreover,

$$\begin{aligned}\mathbb{E}\left[\widehat{M}_k(X^n)^{1/k}\right] &= \mathbb{E}\left[\left(\frac{\sum_{i=1}^n |X_i - \mu(X^n)|^k}{n}\right)^{1/k}\right] \\ &\leq 2\mathbb{E}\left[\left(\frac{\sum_{i=1}^n |X_i - \mu(p)|^k}{n}\right)^{1/k} + \left(\frac{\sum_{i=1}^n |\mu(X^n) - \mu(p)|^k}{n}\right)^{1/k}\right] \\ &= 2\mathbb{E}\left[\frac{\sum_{i=1}^n |X_i - \mu(p)|^k}{n}\right]^{1/k} + 2\mathbb{E}\left[|\mu(X^n) - \mu(p)|\right] \\ &\leq 2M_k(p)^{1/k} + \frac{2M_2(p)^{1/2}}{\sqrt{n}}.\end{aligned}$$

Proof of Lemma E.1: By definition,

$$\begin{aligned}\left|\mu(X^n \setminus L_{\frac{1}{\epsilon}}) - \mu(X^n \setminus U_{\frac{1}{\epsilon}})\right| &= \frac{1}{n - 1/\varepsilon}\left|\sum_{i\in L_{\frac{1}{\epsilon}}} X_i - \sum_{i\in U_{\frac{1}{\epsilon}}} X_i\right| \\ &\le \frac{2}{n}\sum_{i\in L_{\frac{1}{\epsilon}}\cup U_{\frac{1}{\epsilon}}} |X_i - \mu(X^n)| \\ &\le \frac{2}{n}\left(\sum_{i\in L_{\frac{1}{\epsilon}}\cup U_{\frac{1}{\epsilon}}} |X_i - \mu(X^n)|^k\right)^{1/k} \cdot \left(\sum_{i\in L_{\frac{1}{\epsilon}}\cup U_{\frac{1}{\epsilon}}} 1\right)^{(k-1)/k} \\ &\le \frac{2}{n}\left(\sum_{i\in[n]} |X_i - \mu(X^n)|^k\right)^{1/k} \cdot \left(\sum_{i\in L_{\frac{1}{\epsilon}}\cup U_{\frac{1}{\epsilon}}} 1\right)^{(k-1)/k} \\ &= \frac{4\widehat{M}_k(X^n)^{1/k}}{(n\varepsilon)^{1-1/k}}.\end{aligned}$$

□

F. Comparison between different instance-dependent risks for ℓ_p minimization.

Consider the task of estimating the ℓ_p minimizer of a dataset over $[0, R]$ with $R \gg 1$, *i.e.,* for $p > 1$,

$$\theta(D) = \min_{\mu\in\mathbb{R}} \sum_{x\in D} |x - \mu|^p.$$

Consider a dataset D consisting of $n - 1$ 0's and one 1. It can be shown that the minimizer is

$$\mu_p(D) = \frac{1}{1 + (n-1)^{1/(p-1)}}.$$

Let $\ell(x, x') = |x - x'|$. The definition in (Asi & Duchi, 2020a) (Equation (2)) will have

$$R_1(D, \varepsilon) \approx \max_{D': d(D,D')\le 1/\varepsilon} \ell(\theta(D), \theta(D')) \ge \frac{R}{1 + (n\varepsilon - 1)^{1/(p-1)}} - \frac{1}{1 + (n-1)^{1/(p-1)}},$$

where we take D'_1 to be the dataset with $n - 1/\varepsilon$ 0's and $1/\varepsilon$ R's.

The modified definition of (Huang et al., 2021) (Remark 2.4) will lead to a instance dependent risk of

$$\tilde{R}_2(D, \varepsilon) = \sup_{\substack{\mathrm{supp}(D')\subseteq\mathrm{supp}(D) \\ d(D,D')\le 1/\varepsilon}} \ell(\theta(D), \theta(D')) \ge \frac{1}{1 + (n\varepsilon - 1)^{1/(p-1)}} - \frac{1}{1 + (n-1)^{1/(p-1)}},$$

where we take D'_2 be the dataset with $n - 1/\varepsilon$ 0's and $1/\varepsilon$ 1's.

Our propose subset-risk will only allow $D' \subset D$. Hence $\mu_p(D') \in [0, \mu_p(D)]$, and

$$R(D, \varepsilon) \approx \sup_{D'\subset D, d(D,D')\le 1/\varepsilon} \ell(\theta(D), \theta(D')) \le \frac{1}{1 + (n-1)^{1/(p-1)}}.$$

In the regimen when $p < \log(n\varepsilon)$, $\varepsilon \ll 1$ and $R \gg 1$, we have

$$R_1(D, \varepsilon) \gg \tilde{R}_2(D, \varepsilon) \gg R(D, \varepsilon).$$