CS 860: Algorithms for Private Data Analysis Fall 2020

Lecture 15 — Private ML and Stats: Mean Estimation

Prof. Gautam Kamath *Scribe: Gautam Kamath*

Today, we will study private mean estimation. We will distinguish between estimating the mean of a dataset, and estimating the mean of the underlying distribution. Note that there are likely to be many drawings to help exposit the material, so it might be worth consulting the handwritten notes or the lecture videos.

Binary Mean Estimation

First, let's start with one of the simplest problems in differential privacy. We want to estimate the mean of a dataset $X_1, \dots, X_n$ where $X_i \in \{0, 1\}$, i.e., the data is binary. As an example, each datapoint could represent whether the corresponding individual is male. We will use $\tilde{p} = \frac{1}{n}\sum X_i$ to represent the true empirical mean of the dataset. Since the sensitivity of the mean is 1, this is easily privatizable via the Laplace mechanism:

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i + \mathrm{Lap}\left(\frac{1}{\varepsilon n}\right).$$

Then a simple application of Chebyshev's inequality implies that

$$|\hat{p} - \tilde{p}| \leq O\left(\frac{1}{\varepsilon n}\right)$$

with reasonable probability.[1]

This computes the empirical mean of a fixed dataset. But often data comes from some underlying distribution or population, and we wish to estimate the parameter of this distribution. For our first example, suppose $X_1, \dots, X_n$ are sampled i.i.d. from Bernoulli(p), where $0 \leq p \leq 1$ is some unknown parameter which we are trying to estimate. We have that $\mathbf{E}[X_i] = p$, and thus $\mathbf{E}\left[\frac{1}{n}\sum X_i\right] = p$ as well. The variance X_i is $p(1-p)$, and thus the variance of $\frac{1}{n}\sum X_i$ is $p(1-p)/n \leq 1/4n$. Chebyshev's inequality allows us to conclude that

$$|p - \tilde{p}| \leq O\left(\frac{1}{\sqrt{n}}\right)$$

with reasonable probability. This describes the difference between the true parameter p and the empirical mean of the dataset $\tilde{p}$. The calculation above characterizes the error between the empirical mean of the dataset $\tilde{p}$ and the privatized mean $\hat{p}$. Combining the two, we get that the error between the true parameter p and the privatized estimate $\hat{p}$ is

$$|p - \hat{p}| \leq O\left(\frac{1}{\sqrt{n}} + \frac{1}{\varepsilon n}\right).$$

[1] For this lecture, we will be cavalier with concerns of failure probabilities, as they are mostly uninteresting. We will instead just say that events happen. While I will refer to Chebyshev's inequality as a method of proving many of the claims today (due to its simplicity), the Chernoff bound will often give sharper bound on the failure probabilities.

Stated differently, if

$$n \geq O\left(\frac{1}{\alpha^2} + \frac{1}{\alpha\varepsilon}\right),$$

then $|p - \hat{p}| \leq \alpha$.

The $1/\alpha^2$ term is the non-private estimation cost, and the $1/\alpha\varepsilon$ is the additional cost due to privacy. If ε is a constant (say, $\varepsilon = 1$), then we can see the main cost in the sample complexity is the non-private cost, and the cost of privacy is a lower order term. That said, this ignores constant factors which can be relevant in practice.

To summarize: we wish to bound

$$\left|\mu - \left(\frac{1}{n}\sum X_i + N\right)\right|,$$

where N is the noise added for privacy. Using triangle inequality, we upper bound it by

$$\left|\mu - \frac{1}{n}\sum X_i\right| + \left|\frac{1}{n}\sum X_i - \left(\frac{1}{n}\sum X_i + N\right)\right|,$$

where the former is the sampling error, and the latter is the noise error.

Unbounded Data

The result above crucially depended on the data being binary, or at least having bounded range. We will see in a brief example why this is the case. Suppose we wanted to estimate the mean of a dataset $X_1, \ldots, X_n$ where each $X_i \in \mathbb{R}$ is an arbitrary real number. This is more reasonable in other settings: suppose a datapoint represents an individual's height, and we wish to estimate the average height of the set of individuals. Using $\tilde{p} = \frac{1}{n}\sum X_i$ again for the empirical mean of the dataset, we can now see the sensitivity of the mean is unbounded. Thus, the Laplace mechanism would not be able to estimate the mean of the dataset to any finite accuracy. In fact, no differentially private algortihm would be able to – it is possible to formalize this via a packing argument. This is an issue – frequently, we want to estimate the mean of unbounded datasets. Is this fundamentally impossible?

The answer is no, and we will use some a priori information about the dataset to do better. For the height example above, we can say that no human height would be greater than (say) 300 cm. Thus, we can clip the dataset so that all heights are between 0 and 300 cm. This would allow us to bound the sensitivity of the empirical mean, and thus we can apply simple private methods such as the Laplace mechanism. Note that these bounds must be obtained *without* looking at the dataset at all – they must be data independent. We will try to exploit this idea in the distributional setting as well.

Private Parameter Estimation of a Distribution

We more precisely describe the setup for private parameter estimation. We have a dataset $X_1, \ldots X_n$, and we want an algorithm with the following guarantees.

- Privacy: The algorithm is differentially private, no matter what the dataset is.
- Accuracy: If the data is generated from a distribution which satisfies some assumptions, the algorithm should be accurate with high probability.

Note that the former requirement is worst-case, whereas the latter is stochastic. This can be justified for a number of different reasons. One is philosophical in nature: if we, as a data analyst, are incorrect with our assumptions about the data generation process, we do not want the users' privacy to be violated as a result. Another reason is technical, as Steinke and Ullman discuss [SU20a, SU20b], we lose a number of nice properties when we move to average-case definitions of privacy.

Our privacy guarantee is genuinely worst-case in nature. Even if the dataset would arise with exponentially low, or even zero, probability under the distributional assumptions.

Our accuracy guarantee will depend on certain distributional assumptions. For example, even in the non-private setting, if we want to estimate the mean of a distribution which generated a dataset X, we require certain conditions on the unknown distribution. Say, if X was generated i.i.d. from a Gaussian distribution, the empirical mean would be accurate, whereas if it instead came from a Cauchy distribution, it would not.[2]

Univariate Gaussian Estimation

We will first focus on the problem of estimating the mean of a univariate Gaussian distribution. This might be a reasonable abstraction of the mean height estimation problem described before. Specifically, we are given a dataset $X_1, \ldots, X_n$. We wish to design an ε-differentially private algorithm which operates on this dataset (we will later move from pure to approximate differential privacy). If the dataset is generated i.i.d. from $N(\mu, 1)$ where $|\mu| \leq R$ is some unknown parameter, we wish to estimate μ with high probability. While the condition $|\mu| \leq R$ might seem a bit odd, recall our discussion before on where we must exploit a priori information about the distribution. If we do not, then packing lower bounds will preclude any finite sample algorithm.

Note that this bound on the magnitude of the mean μ alone does not imply a bound on the magnitude of the datapoints. However, this bound in combination with the distributional assumption does. With high probability, if we draw n samples from $N(\mu, 1)$, they will all lie in the range $\left[\mu - O\left(\sqrt{\log n}\right), \mu + O\left(\sqrt{\log n}\right)\right]$.This is easy to derive using the following Gaussian tail bound:

$$\Pr_{X \sim N(\mu,1)}[|X - \mu| \geq t] \leq 2 \exp\left(-\frac{t^2}{2}\right).$$

Setting $t = \sqrt{20 \log n}$, this bounds the probability that a single sample falls outside this interval as $2/n^{10}$, and by a union bound, the probability that any point falls outside this interval is at most $2/n^9$. While we chose a generously large value of $t = O(\sqrt{\log n})$, in practice choosing t to be a small constant (say, 3) would suffice, as this interval would contain 99.7% of the datapoints. Combining this bound with the fact that $\mu \in [-R, R]$, we can say that all datapoints will be in the range $[-R - O(\sqrt{\log n}), R + O(\sqrt{\log n})]$ with high probability. All the approaches we describe today will crucially use this fact.

Given this, let us describe a very simple private estimator of the mean, which does not achieve the right sample complexity.

[2] Though technically, the Cauchy distribution doesn't even have a mean.

Theorem 1. *There exists an ε-differentially private algorithm which estimates the mean of $N(\mu, 1)$ (where $|\mu| \leq R$) to accuracy α, given*

$$n = \tilde{O}\left(\frac{1}{\alpha^2} + \frac{R}{\alpha\varepsilon}\right)$$

samples.

Proof. Consider the following two-step procedure.

1. Clip the dataset to the range $[-R - O(\sqrt{\log n}), R + O(\sqrt{\log n})]$: if any X_i lands outside this range, move it to the closest endpoint of this interval.
2. Compute $\frac{1}{n}\sum X_i + \text{Lap}\left(\frac{2R+O(\sqrt{\log n})}{n\varepsilon}\right)$.

It is clear that this statistic is ε-differentially private: due to the clipping step, the sensitivity of the empirical mean $\frac{2R+O(\sqrt{\log n})}{n}$, and privacy follows by the Laplace mechanism. Let us now reason about the accuracy. First, as argued before, if the data is actually from a Gaussian that satisfies our assumptions, the clipping step will not change any point in the dataset, as they will already satisfy these bounds. We use μ for the true mean of the distribution, $\tilde{\mu}$ for the true mean of the dataset, and $\hat{\mu}$ for the output of this algorithm. First, we quantify the non-private error:

$$|\tilde{\mu} - \mu| \leq O\left(\frac{1}{\sqrt{n}}\right).$$

This can be seen using a very similar analysis as the Bernoulli case earlier in this lecture. As for the additional error due to privacy, we have

$$|\tilde{\mu} - \hat{\mu}| = \left|\text{Lap}\left(\frac{2R + O(\sqrt{\log n})}{n\varepsilon}\right)\right| \leq O\left(\frac{R}{n\varepsilon}\right).$$

Putting the two together gives

$$|\mu - \hat{\mu}| \leq O\left(\frac{1}{\sqrt{n}} + \frac{R + O(\sqrt{\log n})}{n\varepsilon}.\right)$$

Upper bounding the right-hand side by α, we require that $n \geq \tilde{O}\left(\frac{1}{\alpha^2} + \frac{R}{\alpha\varepsilon}\right)$, as claimed. □

Another view of this proof, in language similar to before: we wish to bound

$$\left|\mu - \left(\frac{1}{n}\sum f(X_i) + N\right)\right|,$$

where f is the function that clips a datapoint to the interval $[-R - O(\sqrt{\log n}), R + O(\sqrt{\log n})]$, and N is the noise added for privacy. Using triangle inequality, we upper bound it by

$$\left|\mu - \frac{1}{n}\sum X_i\right| + \left|\frac{1}{n}\sum X_i - \frac{1}{n}\sum f(X_i)\right| + \left|\frac{1}{n}\sum f(X_i) - \left(\frac{1}{n}\sum f(X_i) + N\right)\right|.$$

The first term is the sampling error (which is $O(1/\sqrt{n})$) and the last term is the noise error (which is $\tilde{O}(R/n\varepsilon)$), but the middle term quantifies the bias introduced due to the clipping procedure. In this case, we set the clipping interval to be wide enough that no bias is introduced with high probability (this will be the case throughout the lecture, until the end when we talk about heavy-tailed distributions).

We can see that the additional cost due to privacy, $R/\alpha\varepsilon$, can be significant if R is large. For instance, if we have only a very poor guess about the range in where our data lies (for example, if it were a quantity less familiar than human heights), then this cost might overwhelm the non-private cost. Thus, our goal is to reduce this cost.

We will provide two different proofs of the following theorem.

Theorem 2. *There exists an ε-differentially private algorithm which estimates the mean of $N(\mu, 1)$ (where $|\mu| \leq R$) to accuracy α, given*

$$n = \tilde{O}\left(\frac{1}{\alpha^2} + \frac{1}{\alpha\varepsilon} + \frac{\log R}{\varepsilon}\right)$$

samples.

There is actually a third proof of this, which uses the exponential mechanism in place of the histogram-based approach in the next section, but we don't go into details of that today.

Observe that the first two terms in this sample complexity are the same as the simple problem of privately estimating the parameter of a Bernoulli distribution. The last term, $\frac{\log R}{\varepsilon}$, is the cost due to error in our a priori knowledge about the dataset. We have reduced the cost from linear to logarithmic in R, which is an exponential improvement – this proves to be significant improvement in practice, as we will see later. Note that all the terms in the statement are tight up to logarithmic factors. The first term is well known to be the non-private sample complexity, while the latter two can be proven via packing lower bounds.

Histogram-based Approach

This method is due to Karwa and Vadhan [KV18]. The procedure can be divided into two parts. First, we obtain a coarse estimate of the mean, a step which necessitates the $n = \tilde{O}\left(\frac{\log R}{\varepsilon}\right)$ term in the sample complexity. Then we exploit this coarse estimate to get the final fine estimate, which incurs the $n = \tilde{O}\left(\frac{1}{\alpha^2} + \frac{1}{\alpha\varepsilon}\right)$.

The first step splits the range $[-R - O(\sqrt{\log n}), R + O(\sqrt{\log n})]$ into $O(R/\sqrt{\log n}) \leq O(R)$ intervals of width $O(\sqrt{\log n})$. The intuition is $N(\mu, 1)$ should assign significant probability mass to at most one or two of these intervals, and all others should have negligible mass. In particular, one of them should have at least (roughly) $n/2$ datapoints in it, and all others should have roughly 0 points. We will use a private histogram to find which of the intervals has the most points. Recall that this approach adds $\mathrm{Lap}(1/\varepsilon)$ noise to the number of points which falls into each interval, and then outputs the maximum. This incurs a maximum error of $O\left(\frac{\log R}{\varepsilon}\right)$ to any count when there are R bins. In order to locate a bin with at least $n/2$ points, it suffices that $n/2 - \log R/\varepsilon \gg 0$, or $n \geq O\left(\frac{\log R}{\varepsilon}\right)$. Therefore, the interval with the maximum private count will either contain the mean, or be next to it – we can simply return the interval with the maximum, combined with the interval on either side.

We can summarize this step with the following lemma:

Lemma 3. *There exists an ε-differentially private algorithm which outputs an interval I which contains the mean of $N(\mu, 1)$ (where $|\mu| \leq R$), as long as $n \geq \tilde{O}\left(\frac{\log R}{\varepsilon}\right)$. The width of the interval I is at most $O(\sqrt{\log n})$.*

At this point, we can just run the algorithm which achieves the guarantees of Theorem 1. Note that we effectively reduced the problem to itself: while we started with a general value of R, Lemma 3 reduces this to $R = O(\sqrt{\log n})$. Let's unpack this argument. The coarse estimation step gives us an interval $I = [\ell, \ell + C\sqrt{\log n}]$: we clip the dataset to the interval $[\ell - C\sqrt{\log n}, \ell + 2C\sqrt{\log n}]$, compute the empirical mean of the resulting dataset, and add $\text{Lap}(3C\sqrt{\log n}/n\varepsilon)$ noise to the result. This gives us the desired sample complexity claimed by Theorem 2. Note that we must set the privacy budget to $\varepsilon/2$ for each step to achieve an overall ε-DP privacy guarantee.

This approach is nice, but it relies heavily upon the private histogram approach, which does not scale well to (say) the multivariate setting. The one we introduce next will be more flexible in these settings.

Shrinking Confidence Intervals Approach

This method is from Biswas, Dong, Kamath, and Ullman [BDKU20]. The approach is rather simple, relying on nothing except the basic Laplace mechanism. Rather than having a coarse and a fine estimate step, we instead make gradual refinements, iteratively reducing the width of our interval containing the mean.

Consider the naive method given above, which clips the data to some interval and noises the result. We will extend that approach, with the addition of a final step which returns a new *confidence interval*, rather than just outputting a point. We start with an interval $[-R, R]$ which contains the mean μ.

1. Clip the dataset to the range $[-R - O(\sqrt{\log n}), R + O(\sqrt{\log n})]$: if any X_i lands outside this range, move it to the closest endpoint of this interval.
2. Compute $Z = \frac{1}{n}\sum X_i + \text{Lap}\left(\frac{2R + O(\sqrt{\log n})}{n\varepsilon}\right)$.
3. Return an interval centered at Z, of width $O\left(\frac{1}{\sqrt{n}} + \frac{R + \sqrt{\log n}}{n\varepsilon}\right)$.

The idea is that if we wanted the result of step 2 to be highly accurate, we would need a significant amount of data. But just a little bit of data gives us a narrower interval than we started with.

Claim 4. *If $n \geq O\left(\frac{1}{\varepsilon}\right)$ and $R \geq C\sqrt{\log n}$ for some absolute constant C, then the interval returned will be a constant factor smaller than the initial interval. Also, if the data is truly Gaussian, then this interval will contain the true mean μ with high probability.*

Proof. The former claim is easy to see, just by inspecting the first and the last line. The original interval is of width $2R$, and the final interval is of width $O(1/\sqrt{n} + (R + \sqrt{\log n})/n\varepsilon)$. Choosing $n \geq O(1/\varepsilon)$ (with a sufficiently large hidden constant) would result in an interval of width at most $(R + \sqrt{\log n})/100$, which indeed makes a constant factor improvement if $R \geq C\sqrt{\log n}$.

The trickier part is to see why exactly we did this: if Z had no error introduced due to the sampling process and the noise, then Z would be a perfect estimate of the true mean μ. Let us instead try to account for this error:

$$\begin{aligned}
|Z - \mu| &= \left|\frac{1}{n}\sum X_i + \mathrm{Lap}\left(\frac{2R + O(\sqrt{\log n})}{n\varepsilon}\right) - \mu\right| \\
&\leq \left|\frac{1}{n}\sum X_i - \mu\right| + \left|\mathrm{Lap}\left(\frac{2R + O(\sqrt{\log n})}{n\varepsilon}\right)\right|
\end{aligned}$$

Exactly as we argued before, these two terms are bounded by $O\left(\frac{1}{\sqrt{n}}\right)$ and $O\left(\frac{R+\sqrt{\log n}}{n\varepsilon}\right)$, respectively, with high probability. Thus, if we draw an interval around Z of width equal to the sum of these terms, it will contain μ with high probability. □

This claim states that a single iteration results in a constant factor reduction in the width of the interval. Repeating this $t = \tilde{O}(\log R)$ times would result in an interval of constant width. Note that by basic composition, this would result in an overall privacy expenditure of $t\varepsilon$ – we must rescale the value of ε by a factor of $O(t)$ in order to maintain a privacy budget of ε. This results in the requirement that $n \geq O\left(\frac{1}{\varepsilon}\right)$ becoming $n \geq O\left(\frac{\log R}{\varepsilon}\right)$.

At this point, we could simply run Theorem 1, with a value of $R = O(\sqrt{\log n})$. This is exactly what we did before, in the histogram-based approach. This gives us the other two terms in the sample complexity, $\tilde{O}(1/\alpha^2 + 1/\alpha\varepsilon)$, completing the proof.

Beyond Univariate Gaussians

The approach we described is quite flexible. Let's try to describe it abstractly, and see what exactly we needed. The core idea is to constantly maintain a *confidence interval* which contains the unknown parameter of interest. This is initialized to be our a priori knowledge about the parameter – in the univariate mean estimation case, it would be the interval $[-R, R]$.

1. Clip the data. Do this based on the confidence interval containing the parameter, combined with the tail bounds of the distribution class.

2. Compute the empirical estimator for the quantity, and add noise proportional to the sensitivity (which should be bounded due to clipping).

3. Define a new confidence interval centered around this estimate, with a width based on the sampling error and the noise added.

4. Repeat.

Let's see how our univariate mean estimation fits into this framework. We first clip the data. This was done based on the confidence interval containing our parameter ($\mu \in [-R, R]$), and tail bounds on the class of interest (no sample from a Gaussian will land further than $O(\sqrt{\log n})$ away from its mean).

Next, we simply compute and noise the empirical mean of the dataset. The magnitude of the noise is based on the sensitivity, which is restricted based on the width of the confidence interval.

Finally, we have to quantify the error introduced by both the sampling and the noise for privacy. Specifically, how wide are the confidence intervals for both types of error? The former can be quantified by Gaussian tail bounds (using the fact that the average of Gaussians is also a Gaussian), and the latter by Laplace tail bounds.

We state a theorem for the multivariate case, and sketch how the same framework can be used to get near-optimal sample complexity in these cases as well.

Theorem 5. *There exists an (ε, δ)-differentially private algorithm which estimates the mean of $N(\mu, I)$ (where $\|\mu\|_2 \le R$) to ℓ_2-accuracy α, given*

$$n = \tilde{O}\left(\frac{d}{\alpha^2} + \frac{d\sqrt{\log(1/\delta)}}{\alpha\varepsilon} + \frac{\sqrt{d \log R \log(1/\delta)}}{\varepsilon}\right)$$

samples.

Note that it shifts from pure DP to approximate DP, which is common in the multivariate case. The ℓ_2-sensitivity is more natural in these settings, so we generally use the Gaussian mechanism.

The data is again clipped using the confidence set containing the mean, as well as Gaussian tail bounds. This time, rather than an interval, the mean μ is in an ℓ_2 ball of radius R. Furthermore, given n samples from a Gaussian in d dimensions, it is well-known that the largest point will have ℓ_2-norm $\Theta(\sqrt{d} + \sqrt{\log n})$. Thus, we can clip to the ball of radius $\tau = O(R + \sqrt{d} + \sqrt{\log n})$.

We then compute the empirical estimator for the mean. Due to the clipping, the ℓ_2-sensitivity is bounded by τ, and we can apply the Gaussian mechanism.

Finally, we quantify how wrong the privatized estimate could be, and draw a ball around it. Since both the data and the noise are Gaussian, this again follows by (multivariate) Gaussian tail bounds. Loosely speaking, it says the ball will be of radius $O\left(\sqrt{d} \cdot \frac{R+\sqrt{d}+\sqrt{\log n}}{n\varepsilon}\right)$ – as long as $n \ge \tilde{O}(\sqrt{d}/\varepsilon)$, we make a constant factor improvement (in the right parameter regime). Repeating this $\log R$ times (and rescaling the ε parameter appropriately) gives the requirement $n \ge \tilde{O}(\sqrt{d \log R}/\varepsilon)$. This will reduce the radius of the ball to $\tilde{O}(\sqrt{d})$, and at this point, a naive estimator (similar to Theorem 1) will give the result claimed in Theorem 5.

There is a similar result for Gaussian covariance estimation, but it's probably beyond the scope of these lecture notes.

Heavy-Tailed Data

So far, we've discussed only Gaussian data, which can be a very strong assumption for data in the real world – there are many cases where we might not have our data fall into such a nice parametric class. The same arguments also hold for *sub-Gaussian* distributions. Informally speaking, these are distributions whose tails decay faster than than of a Gaussian distribution. This implies that we have bounds on all the moments of the distribution: $\mathbf{E}[(P-\mu)^k]$ is bounded for all k.

However, this is frequently not the case in practice, where data may be heavy tailed. Consider a simple case where all we know about an unknown distribution is that $\mu = \mathbf{E}[P] \in [-1, 1]$ (intentionally chosen to be narrow to simplify the setting) and $\mathbf{E}[(P-\mu)^2] \leq 1$, and we wish to privately estimate μ. As we will see, the lack of bounded moments will require us to clip much more aggressively: our previous strategy of not clipping any point will introduce far too much noise.

We will find it helpful to recall the way we decomposed the error:

$$\left|\mu - \frac{1}{n}\sum X_i\right| + \left|\frac{1}{n}\sum X_i - \frac{1}{n}\sum f(X_i)\right| + \left|\frac{1}{n}\sum f(X_i) - \left(\frac{1}{n}\sum f(X_i) + N\right)\right|.$$

Recall that the three terms are the error due to sampling, bias from clipping, and noise addition for privacy. Before, our strategy was to choose the clipping function f such that the middle term will be 0. Let's see what happens if we try that again.

Consider clipping to the interval $[-1-\tau, 1+\tau]$, where τ is some parameter to be set. Chebyshev's inequality on P implies that $\Pr[|P-\mu| \geq \sqrt{n}] \leq 1/n$, and by a union bound, clipping at $\tau = \Theta(\sqrt{n})$ will not affect any points with reasonable probability, and the bias is likely to be 0. However, clipping at this threshold will introduce an unacceptable amount of noise: one must add $\text{Lap}\left(\frac{O(\sqrt{n})}{n\varepsilon}\right)$ noise. The errors from sampling, clipping, and privacy will be $O(1/\sqrt{n})$, 0, and $O(1/\sqrt{n}\varepsilon)$, respectively. Thus, to bound our error by α, we require that $n \geq O\left(\frac{1}{\alpha^2\varepsilon^2}\right)$.

Let's see if there's an improved way to clip, to have a better balance between these terms. Choosing a smaller value of τ will introduce some non-zero bias, but reduce the magnitude of noise addition. We will not prove it here, but it turns out that a clipping parameter of τ introduces bias of order $1/\tau$:

$$\left|\frac{1}{n}\sum X_i - \frac{1}{n}\sum f(X_i)\right| \leq O(1/\tau).$$

At the same time, a clipping parameter of τ requires us to add $\text{Lap}\left(\frac{O(\tau)}{n\varepsilon}\right)$ noise. This time, the sampling, clipping, and privacy errors are $O(1/\sqrt{n})$, $O(1/\tau)$, and $O(\tau/n\varepsilon)$. Bounding the sum of errors by α, the best setting of parameters is $\tau = \Theta(1/\alpha)$ and $n \geq O\left(\frac{1}{\alpha^2\varepsilon}\right)$. Note that this is more expensive than the $O(1/\alpha^2 + 1/\alpha\varepsilon)$ when we had sub-Gaussianity, and the result turns out to be tight. There are also further extensions to when we have bounds on higher order moments, and in the multivariate setting [KSU20].

References

[BDKU20] Sourav Biswas, Yihe Dong, Gautam Kamath, and Jonathan Ullman. Coinpress: Practical private mean and covariance estimation. *arXiv preprint arXiv:2006.06618*, 2020.

[KSU20] Gautam Kamath, Vikrant Singhal, and Jonathan Ullman. Private mean estimation of heavy-tailed distributions. In *Proceedings of the 33rd Annual Conference on Learning Theory*, COLT '20, 2020.

[KV18] Vishesh Karwa and Salil Vadhan. Finite sample differentially private confidence intervals. In *Proceedings of the 9th Conference on Innovations in Theoretical Computer Science*, ITCS '18, pages 44:1–44:9, Dagstuhl, Germany, 2018. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[SU20a] Thomas Steinke and Jonathan Ullman. The pitfalls of average-case differential privacy. https://differentialprivacy.org/average-case-dp/, July 2020.

[SU20b] Thomas Steinke and Jonathan Ullman. Why privacy needs composition. https://differentialprivacy.org/privacy-composition/, August 2020.