Differentially private estimation of U statistics

Kamalika Chaudhuri KAMALIKA@CS.UCSD.EDU
University of California, San Diego

Po-Ling Loh PLL28@CAM.AC.UK
University of Cambridge

Shourya Pandey SHOURYAP@UTEXAS.EDU
University of Texas at Austin

Purnamrita Sarkar PURNA.SARKAR@AUSTIN.UTEXAS.EDU
University of Texas at Austin

Editor: Under Review for COLT 2024

Abstract

In this paper, we consider the problem of private estimation of U statistics. U statistics are widely used estimators that naturally arise in a broad class of problems, from nonparametric signed rank tests to subgraph counts in random networks. They are simply averages of an appropriate kernel applied to all subsets of a given size (also known as the degree) of sample size n. However, despite the recent outpouring of interest in private mean estimation, private algorithms for more general U statistics have received little attention. We propose a framework where, for a broad class of U statistics, one can use existing tools in private mean estimation to obtain confidence intervals where the private error does not overwhelm the irreducible error resulting from the variance of the U statistics. However, in specific cases that arise when the U statistics degenerate or have vanishing moments, the private error may be of a larger order than the non-private error. To remedy this, we propose a new thresholding-based approach that uses Hajek projections to re-weight different subsets. As we show, this leads to more accurate inference in certain settings.

Keywords: U statistics, private estimation, mean estimation, degenerate kernel, sparse graphs, smoothed sensitivity

1. Introduction

U statistics are a well-established class of estimators that can be expressed as averages of functions of the form $h(X_1, \ldots, X_k)$, where h is a possibly vector-valued kernel and $\{X_i\}_{i=1}^n$ are i.i.d. draws from some underlying distribution. U statistics arise in many areas of statistics and machine learning, encompassing diverse estimators such as the sample mean and variance; the Mann-Whitney and Wilcoxon signed rank test statistics; Kendall's tau; the number of subgraphs in a random graph (Gilbert, 1961); the number of collisions in a stream of discrete data points; and applications to ranking and clustering (Clémençon et al., 2008; Clémençon, 2014).

Despite being a natural generalization of the sample mean, little work has been done on private estimation of U statistics. In comparison, private mean estimation has attracted a great deal of interest (Karwa and Vadhan, 2017; Kamath et al., 2019a; Cai et al., 2021; Kamath et al., 2019b; Biswas et al., 2020; Kamath et al., 2020). The few papers we are aware of are Ghazi et al. (2020) and Bell et al. (2020), but both of these papers focus on the setting of local differential privacy (Kasiviswanathan et al., 2011), whereas we are interested in privacy guarantees under the more basic central model. Moreover, much existing work focuses on discrete data, and relies on simple central differential privacy mechanisms (such as the Global Sensitivity mechanism (Dwork et al., 2006)) that are optimal in such settings.

In this paper, we ask the following questions:

1. Can we apply existing off-the-shelf tools for private mean estimation to privately estimate general U statistics?

2. What are specific cases where methods more aligned with the structure of a U statistic provide more accurate estimators?

Our proposed algorithms differ depending on certain properties of the underlying U statistics. A non-degenerate U statistics is one which, suitably scaled, converges to a limiting Gaussian distribution with variance $O(k/\sqrt{n})$; such a result is commonly used in hypothesis testing (Hoeffding, 1948; Arcones and Gine, 1993; Hoeffding, 1963). However, there are also cases of degenerate U statistics, where the limiting distribution is chi-squared and the variance of the statistic is $O(k^2/n^2)$. Another interesting type of U statistics arises in subgraph counts in random geometric graphs (Gilbert, 1961) when the probability of an edge being present tends to zero with n. In these cases, the kernel has mean and variance that also tends to 0, and one needs to be careful about the possibility of creating a private estimator that simply adds Laplace noise with some large scale.

We show that for general non-degenerate U statistics with sub-Gaussian kernels, an extension of the Coinpress algorithm (Biswas et al., 2020)—where we apply the algorithm on the kernel evaluated on different subsets of data—works suitably for private estimation. We provide accuracy bounds for two methods for aggregating data, the first being the vanilla version which averages over all subsets of size n, and the second being a faster, subsampled version which is less computationally expensive. Despite having a suboptimal dependence on the degree k, we show that these methods add Laplace noise that is of a smaller order than the main term in the confidence interval arising from the variance of the U statistic.

For degenerate kernels, although the natural variant of the Coinpress algorithm would immediately yield confidence intervals of length $O(1/n)$ rather than $O(1/\sqrt{n})$, the Laplace error would overpower the non-private part of the deviation. The same phenomenon happens for sparse graph applications. For these "atypical" situations, we present an interesting alternative algorithm, which uses the Hajek projections of U statistics to "decide" which data points are problematic, and then assigns corresponding weights to different subsets of the data. This method, inspired by work of Ullman and Sealfon (2019), provides confidence intervals where the Laplace error does not occlude the non-private part. However, this procedure is more computationally expensive and our current analysis requires the kernel to be bounded.

Our takeaway message is that, in many cases, one can indeed obtain a computationally efficient private U statistic using adaptations of existing private mean estimation algorithms. However, in special scenarios, one may need to design more sophisticated methods to adapt to the smaller variance. The remainder of our paper is organized as follows: Section 2 reviews background on U statistics and fundamental concepts in differential privacy. Section 3 presents the framework of our main private U statistic estimation algorithm and corresponding theory, which is then applied to various settings. Section 4 presents the alternative algorithm based on Hajek projections. We provide a short discussion of applications in Section 5. Section 6 concludes the paper.

2. Background and problem setup

We begin with some notation. Let k and n be positive integers with $k \leq n$. Let $\mathcal{D}$ be an unknown probability distribution over $\mathcal{X}$, and let $h : \mathcal{X}^k \to \mathbb{R}$ be a known function. Let $\mathcal{H}$ be the distribution of $h(X_1, X_2, \ldots, X_k)$, where $X_1, \ldots, X_k \sim \mathcal{D}$ are i.i.d. We use $[n]$ to denote $\{1, \ldots, n\}$. We will be interested in providing a differentially private confidence interval around the parameter $\theta = \mathbb{E}[h(X_1, \ldots, X_k)]$, which is the mean of $\mathcal{H}$ (Halmos, 1946). The classical minimum variance unbiased estimator is the U statistic (Hoeffding, 1948). Let $\mathcal{I}_{n,k}$ be the set of all k-element subsets of $[n]$. For any $S \in \mathcal{I}_{n,k}$, we also use X_S to denote $X_{i_1,\ldots,i_k}$, where $S = \{i_1, \ldots, i_k\}$. Define as $\mathcal{I}_{n,k}$ the set of all $\binom{n}{k}$ unordered subsets of size k from $[n]$. The U statistic U_n is then defined as

$$U_n = \frac{1}{\binom{n}{k}} \sum_{\{i_1,\ldots,i_k\} \in \mathcal{I}_{n,k}} h(X_{i_1}, \ldots, X_{i_k}). \tag{1}$$

The function h is known as the *kernel* and k is the *degree* of U_n. Note that U statistics can also be vector-valued. As a first step, we consider scalar U statistics in this paper. We also define variance of conditional means, which will be used to express the expansion of the variance of U_n:

$$\zeta_c := \text{Var}\left(\mathbb{E}[h(X_1, \ldots, X_k)|X_1 = x_1, \ldots, X_c = x_c]\right). \tag{2}$$

We have the following inequality from Lee (1990); Serfling (1980):

$$\frac{\zeta_i}{i} \leq \frac{\zeta_j}{j}, \text{ for } i < j. \tag{3}$$

We write $\text{proj}(X_i, \ell, r)$ to denote the scalar X_i projected to the interval (ℓ, r). For unbounded kernels, we will consider sub-Gaussian$(K\zeta_k)$ kernels, where we write $X \sim$ sub-Gaussian(σ^2) if $\mathbb{E}[\exp(\lambda X)] \leq \exp(\lambda^2\sigma^2/2)$ for all $\lambda \in \mathbb{R}$. For bounded kernels, we will use inequalities for sub-exponential variables. We write $X \sim$ sub-exponential(σ^2, b) when $\mathbb{E}[\exp(\lambda X)] \leq \exp(\sigma^2\lambda^2/2)$ for all $|\lambda| \leq 1/b$. Note that if $|h(X_1, \ldots, X_k)| \leq b$ almost surely, then $\mathcal{H}$ is sub-exponential(ζ_k, b). This distinction helps in dealing with sparse graphs, where $\zeta_k \to 0$.

Lemma A.1 in Appendix A.4 provides useful bounds on $\text{Var}(U_n)$, while Lemma A.2 provides Hoeffding and Bernstein type bounds for U statistics.

2.1. Classes of U statistics

Classical U statistics typically have small or fixed values of k. However, important estimators that appear in the context of subsampling (Politis et al., 2012) or Breiman's random forest algorithm (Song et al., 2019; Peng et al., 2019) have $k = o(n)$. These are sometimes referred to as *infinite-order* U statistics (Frees, 1989; Minsker, 2023)). While it is then natural to write the expectation θ and conditional variances $\{\zeta_c\}$ as indexed by n, we do not do this for simplicity.

n-dependent kernel: U statistics also frequently appear in the analysis of random geometric graphs (Gilbert, 1961). The difference between this setting and the examples above is that in the sparse setting, the conditional variances $\{\zeta_c\}$ also vanish with n. (See Section 5.1.)

Degenerate U statistics: A U statistic is *degenerate of order* $\ell < k$ if $\zeta_i = 0$ for all $i \in [\ell - 1]$ and $\zeta_\ell > 0$. Degenerate U statistics often arise in hypothesis testing, including Cramer-Von Mises and Pearson tests of goodness of fit (Gregory, 1977; Anderson and Darling, 1952), (Shorack and Wellner, 2009, Chapter 5). They also appear in tests for model misspecification in econometrics (Li and Fan, 2020; Linton and Gozalo, 2014). For more examples, see de Wet (1987); Weber (1981); Ho and Shieh (2006).

2.2. Differential privacy

In this work, we focus on the cryptographically-motivated notion of differential privacy (Dwork et al., 2006), which has emerged as the gold standard in private data analysis. The main idea is that the participation of a single person should not make a difference to the probability of any outcome:

Definition 1 *A (randomized) mechanism M that takes as input a dataset D and has outputs in a range space $\mathcal{S}$ is said to satisfy ϵ-differential privacy if for any pair of datasets D and D' that differ in the value of a single element, and for any subset $S \subseteq \mathcal{S}$ of the range space $\mathcal{S}$, we have* $\Pr(M(D) \in S) \leq e^\epsilon \Pr(M(D') \in S)$.

Sensitivity. The most common way to ensure that a mechanism satisfies differential privacy is through the Global Sensitivity Method (Dwork et al., 2006). Suppose we are trying to calculate a differentially private approximation to a function f; this method first computes its *global sensitivity*, which is the worst-case change in the function $f(D)$ when D and D' differ in a single value $GS(f) = \max_{D,D',|D\Delta D'|=1} |f(D) - f(D')|$. The mechanism outputs $M(D) = f(D) + \frac{GS(f)}{\epsilon} Z$, where Z is a Laplace random variable. A similar definition for a specific dataset D is the *local*

sensitivity: $LS(f, D) = \max_{D', |D \Delta D'|=1} |f(D) - f(D')|$. Unfortunately, adding noise proportional to the local sensitivity does not ensure differential privacy in general: variation in the noise magnitude itself could potentially leak private information.

Instead, Nissim et al. (2007) proposed an elegant way of computing the *smoothed sensitivity*, a smooth upper bound on the local sensitivity of a function f at a point D such that adding proportionate noise ensures differential privacy. A function $SS(D)$ is said to be an ϵ-smooth upper bound on the local sensitivity of f if (i) $SS(D) \geq LS(f, D)$ for all D, and (ii) $SS(D) \leq e^{\epsilon} SS(D')$ for all $|D \Delta D'| = 1$. Roughly speaking, the first condition ensures that enough noise is added, and the second condition ensures that the noise does not itself leak information about the data. The Smoothed Sensitivity Method outputs $M(D) = f(D) + \frac{SS(D)}{\epsilon} \cdot Z$, where Z is a Student's t-distribution with 3 degrees of freedom (Nissim et al., 2007; Bun and Steinke, 2019).

Private Mean Estimation. A fundamental task in private statistical inference is private mean estimation based on a set of i.i.d. observations. The most obvious way to do this is via the global sensitivity method; however, this means that the standard deviation of the noise scales with R, the range of the values. In the fairly realistic case where the range might be large while the typical value is small, this leads to highly noisy estimation. To remedy this effect, a body of work (Karwa and Vadhan, 2017; Kamath et al., 2019a; Cai et al., 2021; Kamath et al., 2019b) has looked into how to design better private mean estimators for (sub)-Gaussian vectors. Our work will build on one such method, known as CoinPress (Biswas et al., 2020). The main idea is to iteratively refine an estimate for the parameters, until one obtains a private range that contains most of the values; noise is then added proportional to this range. Observe that some dependence on range is inevitable, especially for estimation with pure differential privacy (Chaudhuri and Hsu, 2012).

3. Main results

Our goal is to provide a private estimator for an unknown, estimable parameter θ. We begin by discussing a (non-private) estimator that estimates θ by an average of independent quantities:

Definition 2 (Naive estimator) *Let $m = n/k$ and $\mathcal{I}_j := \{\{(j-1)k+1, \ldots, (j-1)k+k\}\}$ for all $j \in [m]$. Define $\mathcal{F}_{naive} = \{\mathcal{I}_1, \ldots, \mathcal{I}_m\}$, which we call the "naive" family. Estimate θ using $\hat{\theta}_{naive} := \frac{1}{n_k} \sum_{i=1}^{n_k} h(X_{(j-1)k+1}, \ldots, X_{(j-1)k+k})$.*

Remark 3 *Most private mean estimation algorithms (Karwa and Vadhan, 2017; Kamath et al., 2019b,a; Biswas et al., 2020) applied to our setting will essentially provide a confidence interval of width $O\left(\sqrt{\text{Var}\left(\hat{\theta}_{naive}\right)} + k\sqrt{\zeta_k}/n\epsilon\right) = O\left(\sqrt{k\zeta_k/n} + k\sqrt{\zeta_k}/n\epsilon\right)$, since $\text{Var}\left(\hat{\theta}_{naive}\right) = k\zeta_k/n$, where ζ_k is defined in Eq 2. Note that this is larger than the dominant term $k^2\zeta_1/n$ of $\text{Var}(U_n)$ (see Lemma A.1) and Eq 3. This stems from the fact that the naive estimator is a sub-optimal estimator of θ. The optimal estimator is the U statistic defined in Eq 1.*

In Algorithms 1 and 2, we present a general extension of the Coinpress algorithm (Biswas et al., 2020) for estimating θ, which will then be used to obtain a private estimate with the non-private term matching $\text{Var}(U_n)$. Originally, this algorithm was used for private mean and covariance estimation of i.i.d. (sub)-Gaussian data. We also provide an sketch of the idea behind the algorithm.

Consider the set $\{Y_j\}_{j \in [\binom{n}{k}]} = \{h(X_S) : S \in \mathcal{I}_{n,k}\}$, where $m = \binom{n}{k}$; let us call this *pseudo-data*. We want to apply the Coinpress algorithm on the pseudo-data by treating them as independent data. While this assumption is not true, the pseudo-data are only weakly related in the following sense: if a single data point X_i is changed, the only pseudo-data that potentially change are $h(X_S)$ for $S \in \mathcal{I}_{n,k}^{(i)}$; the fraction of such pseudo-data is $\binom{n-1}{k-1}/\binom{n}{k} = k/n$. This fact, along with sufficiently strong concentration of the Y_j's and their mean $\frac{1}{m}\sum_{j \in [m]} Y_j$ around θ, allows us to recover $\sqrt{\text{Var}(U_n)}$ in the non-private part of the length of the confidence interval.

Algorithm 1 U-StatMean $\left(n, k, h, \{X_i\}_{i\in[n]}, \mathcal{F} = \{\mathcal{I}_1, \ldots, \mathcal{I}_m\}, R, \epsilon, Q, Q^{\text{avg}}\right)$

$\alpha \leftarrow 0.99,\ t \leftarrow \log\left(R/Q_\alpha^{\text{avg}}\right),\ [l_0, r_0] \leftarrow [-R, R]$
for $j = 1, \ldots, m$ **do**
  $Y_{0,j} \leftarrow \frac{1}{|\mathcal{I}_j|}\sum_{S\in\mathcal{I}_j} h(X_S)$
end for
for $i = 1, 2, ..., t$ **do**
  $\{Y_{i,j}\}_{j\in[m]}, [l_i, r_i] \leftarrow$ U-StatOneStep $(n, k, \{Y_{i-1,j}\}, \mathcal{F}, [l_{i-1}, r_{i-1}], \epsilon/2t, \alpha/t, Q, Q^{\text{avg}})$
end for
$\{Y_{t+1,j}\}_{j\in[m]}, [l_{t+1}, r_{t+1}] \leftarrow$ U-StatOneStep $(n, k, \{Y_{t,j}\}, \mathcal{F}, [l_t, r_t], \epsilon/2, \alpha, Q, Q^{\text{avg}})$
return $(l_{t+1} + r_{t+1})/2$

Algorithm 2 U-StatOneStep $\left(n, k, \{Y_i\}_{i\in[m]}, \mathcal{F}, [l, r], \epsilon', \beta, Q, Q^{\text{avg}}\right)$

$Y_j \leftarrow \text{proj}_{l-Q_\beta, r+Q_\beta}(Y_j)$ for all $1 \le j \le m$.
$\Delta \leftarrow \text{dep}_{n,k}(\mathcal{F})(r - l + 2Q_\beta)$
$Z \leftarrow \frac{1}{m}\sum_{j=1}^m Y_j + W$, where $W \sim \text{Lap}\left(\frac{\Delta}{\epsilon'}\right)$
$[l, r] \leftarrow \left[Z - \left(Q_\beta^{\text{avg}} + \frac{\Delta}{\epsilon'}\log\frac{1}{\beta}\right), Z + \left(Q_\beta^{\text{avg}} + \frac{\Delta}{\epsilon'}\log\frac{1}{\beta}\right)\right]$
return $\{Y_j\}_{j\in[m]}, [l, r]$

More generally, this idea allows us to extend the algorithm to $\{Y_j\}_{j\in[m]}$ such that (i) each Y_j is a linear combination of the $h(X_S)$ with $\mathbb{E}[Y_j] = \theta$, (ii) the Y_j are weakly dependent, and (iii) all Y_j's and $\frac{1}{m}\sum_{j\in[m]} Y_j$ have sufficiently strong concentration around θ.

Setup. Let $m = m_{n,k}$ be a positive integer and let $\mathcal{F} = \mathcal{F}_{n,k} = \{\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_m\}$ be a family of non-empty subsets of $\mathcal{I}_{n,k}$, not necessarily distinct. For each $i \in [n]$, let

$$f_i := \frac{|\{j \in [m] : \exists\ S \in \mathcal{I}_j \text{ such that } i \in S\}|}{m} \tag{4}$$

be the fraction of indices j such that i is contained in some subset of $\mathcal{I}_j$. Let $\text{dep}_{n,k}(\mathcal{F}) := \max_{i\in[n]} f_i$. For each $j \in [m]$, let $Y_j := \frac{1}{|I_j|}\sum_{S\in\mathcal{I}_j} h(X_S)$. Clearly, $\mathbb{E}[Y_j] = \theta$. Moreover, $\text{dep}_{n,k}(\mathcal{F})$ is an upper bound on the fraction of Y_j that change if any single X_i is changed. To allow for small noise addition to ensure privacy, it will be desirable to choose $\mathcal{F}$ such that $\text{dep}_{n,k}(\mathcal{F})$ is small. For $\beta \in (0, 1]$, let $Q_\beta = Q_{n,k,h,\mathcal{D},\mathcal{F}}(\beta)$ and $Q_\beta^{\text{avg}} = Q_{n,k,h,\mathcal{D},\mathcal{F}}^{\text{avg}}(\beta)$ be defined such that

$$\mathbb{P}\left(\sup_{j\in[m]} |Y_j - \theta| > Q_\beta\right) < \beta, \text{ and } \mathbb{P}\left(\left|\frac{1}{m}\sum_{j=1}^m Y_j - \theta\right| > Q_\beta^{\text{avg}}\right) < \beta. \tag{5}$$

We will refer to Q_β and Q_β^{avg} as β-*confidence bounds* for $\sup_{j\in[m]} |Y_j - \theta|$ and $\left|\frac{1}{m}\sum_{j\in[m]} Y_j - \theta\right|$. Finally, let $\text{size}(\mathcal{F}) := \sum_{j\in[m]} |\mathcal{I}_j|$.

Proposition 4 *Let n, m, and k be positive integers with $k \le n$, and let $\alpha = 0.01$. Let $h : \mathcal{X}^k \to \mathbb{R}$ be a symmetric*[1] *function and let $\mathcal{D}$ be an unknown distribution over $\mathcal{X}$ with $\mathbb{E}\left[h(\mathcal{D}^k)\right] = \theta$. Moreover, let $R > 0$ be a known bound on $|\theta|$. Let $\mathcal{F} = \{\mathcal{I}_j\}_{j\in[m]}$ be a family of non-empty subsets of $\mathcal{I}_{n,k}$. For any $X_1, \ldots, X_n \in \mathcal{X}$, let $Y_j = \frac{1}{|\mathcal{I}_j|}\sum_{S\in\mathcal{I}_j} h(X_S)$ for all $j \in [m]$. Let Q and Q^{avg}, respectively, be known confidence bounds on $\sup_{j\in[m]} |Y_j - \theta|$ and $|\frac{1}{m}\sum_{j\in[m]} Y_j - \theta|$ as defined in Eq 5. Then, for all $\epsilon > 0$, Algorithm 1 with input $\left(n, k, h, \{X_i\}_{i\in[n]}, \mathcal{F}, R, \epsilon, Q, Q^{avg}\right)$ returns $\tilde{\theta}_n$ such that, with probability at least $\frac{3}{4}$,*

1. That is, the output does not change on permuting the inputs.

$$|\tilde{\theta}_n - \theta| \leq O\left(\underbrace{\frac{\sqrt{\text{Var}(\sum_{j\in[m]} Y_j)}}{\sqrt{m}}}_{\textit{non-private error}} + \underbrace{\frac{dep_{n,k}(\mathcal{F})\, Q_\alpha}{\epsilon}}_{\textit{private error}}\right), \tag{6}$$

as long as $dep_{n,k}(\mathcal{F}) \leq \frac{Q_\alpha \epsilon}{10 t Q_{\alpha/t} \log t/\alpha}$ *and* $Q^{avg}_{\alpha/t} < Q_\alpha$, *with* $t := \log\left(\frac{R}{Q_\alpha}\right)$. *Also, Algorithm 1 is* ϵ*-differentially private and runs in time* $O\left(m \log \frac{R}{n+Q_\alpha} + k \cdot size(\mathcal{F})\right)$.

Remark 5 *The proposition assumes that the concentration bounds* Q *and* Q^{avg} *are known, despite the mean* θ *being unknown. If these bounds are not known, we may first need to (privately) compute* Q_α *and* Q^{avg}_α, *and then use those privately computed bounds in the algorithm. We will see how to estimate these parameters for various families* $\mathcal{F}$ *of indices used in Algorithm 1.*

Boosting the error probability: Proposition 4 achieves a constant success probability of $\frac{3}{4}$. If we modify the algorithm parameters as is so that the error probability is at most α, we will incur a $1/\sqrt{\alpha}$ multiplicative factor in the non-private error. This stems from an application of Chebyshev's inequality to bound $|\frac{1}{m}\sum_{j=1}^m Y_{0,j} - \theta|$. Depending on the specific family $\mathcal{F}$, we may be able to provide a better concentration bound for $\frac{1}{m}\sum_{j=1}^m Y_{0,j}$ in Eq A.23. Instead, we complement the result of Proposition 4 with a median-of-means wrapper that allows for an improved dependence on α with only a $\log \frac{1}{\alpha}$ multiplicative blowup in the sample complexity to achieve the same error.

Lemma 6 *Let* $\alpha > 0$ *and let* $d := 12 \log \frac{1}{\alpha}$. *Perform* d *independent runs of Algorithm 1 to obtain* ϵ*-differentially private estimates* $\{\tilde{\theta}_{n,i}\}_{i\in[d]}$ *of* θ, *and let* $\tilde{\theta}^{med}_n$ *be the median of these values. Then,* $\tilde{\theta}^{med}_n$ *is* ϵ*-differentially private, and with probability at least* $1-\alpha$,

$$\left|\tilde{\theta}^{med}_n - \theta\right| = O\left(\sqrt{\text{Var}\left(\frac{1}{m}\sum_{j=1}^m Y_j\right)} + \frac{dep_{n,k}(\mathcal{F})\, Q_\alpha}{\epsilon}\right). \tag{7}$$

3.1. Main proposition applied to various families $\mathcal{F}$

In this section, we operate under the following assumption.

Setting. Let $\mathbf{X} = \{X_i\}_{i\in[n]}$ be i.i.d. draws from $\mathcal{D}$. Assume the distribution of $h(X_1, \ldots, X_k)$ is K-sub-Gaussian with unknown mean $\theta \in (-R, R)$ and unknown variance $\zeta_k \in (\sigma_{\min}, \sigma_{\max})$, for known parameters $K, R, \sigma_{\min}$, and $\sigma_{\max}$.

We now apply Proposition 4 (specifically, the form obtained in Lemma 6) to different $\mathcal{F}$ to obtain private estimates of θ, with statistical and computational tradeoffs depending on the family $\mathcal{F}$. As Remark 5 suggests, we will also need to privately estimate concentration bounds on the Y_j's and their average. Naturally, this requires a private estimate of the variance ζ_k. We provide the variance and mean estimation guarantees from Biswas et al. (2020) for variance estimation here, where we have translated the mean estimation guarantee to fit our setting.

Lemma 7 *There exists an algorithm* ***PrivateVariance*** $(\mathbf{X}, \sigma_{min}, \sigma_{max}, \epsilon, \alpha)$ *which is* ϵ*-differentially private and runs in time* $O(n \log \frac{\sigma_{max}}{\sigma_{min}})$, *such that with probability at least* $1-\alpha$, *the output* $\tilde{\zeta}_k$ *of the algorithm satisfies* $\zeta_k \leq \tilde{\zeta}_k \leq 2\zeta_k$ *as long as* $n = \tilde{\Omega}(\frac{K \log \frac{\sigma_{max}}{\sigma_{min}}}{\epsilon} \log \frac{1}{\alpha})$. *Moreover, this algorithm runs in time* $\tilde{O}(n \log \frac{\sigma_{max}}{\sigma_{min}})$.

Lemma 8 *Consider the naive estimator in Definition 2 corresponding to the following family of subsets of* $\mathcal{I}_{n,k}$: *let* $m = n/k$ *and* $\mathcal{I}_j := \{\{(j-1)k+1, \ldots, (j-1)k+k\}\}$ *for all* $j \in [m]$. *Let* $\mathcal{F}_{naive} = \{\mathcal{I}_1, \ldots, \mathcal{I}_m\}$. *There exist* ϵ*-differentially privately computable confidence bounds* Q

and Q^{avg} such that Algorithm 1 with input $\left(n, k, h, \{X_i\}_{i\in[n]}, \mathcal{F}_{naive}, R, \epsilon, \alpha, Q, Q^{avg}\right)$ returns an estimate $\hat{\theta}_{naive}$ of the mean θ such that with probability at least $1 - O(\alpha)$,

$$|\hat{\theta}_{naive} - \theta| \leq \frac{1}{\sqrt{\alpha}}\sqrt{\frac{k\zeta_k}{n}} + \tilde{O}\left(\frac{k\sqrt{K\zeta_k}}{n\epsilon}\log\frac{1}{\alpha}\right),$$

as long as $n = \tilde{\Omega}\left(\frac{k}{\epsilon}\left(\log R + \log\frac{\sigma_{max}}{\sigma_{min}}\right)\log\frac{1}{\alpha}\right)$. The estimate $\hat{\theta}_{naive}$ is 2ϵ-differentially private and the algorithm runs in time $n = \tilde{\Omega}\left(n + \frac{n}{k}\log\frac{R}{\zeta_k\sqrt{K}}\right)$.

Remark 9 *As discussed in Remark 3 above, this is a sub-optimal estimator, because the first term of the deviation in a non-private estimator is $O(\sqrt{\mathrm{Var}(U_n)})$. Indeed, using Lemma A.1 we can see that the variance of a non-degenerate U statistics is $k^2\zeta_1/n + O(k^2\zeta_k/n^2)$, which is smaller than the non-private part of the deviation provided in Corollary 8, due to (3).*

We now apply this algorithm to different estimators. We first introduce a more computationally intensive estimator, which simply uses each k-tuple from $\mathcal{I}_{n,k}$.

Definition 10 (All-tuples family) *Let $m = \binom{n}{k}$ and let $\mathcal{F}_{all} = \{\mathcal{I}_1, \ldots, \mathcal{I}_m\}$ be the m distinct singleton subsets of $\mathcal{I}_{n,k}$. Call this family the "all-tuples" family.*

Lemma 11 *Let $\mathcal{F}_{all}$ be the "all-tuples" family in Definition 10. Then, there exist ϵ-differentially private confidence bounds Q and Q^{avg} such that Algorithm 1 with $\mathcal{F} = \mathcal{F}_{all}$ returns an estimate $\tilde{\theta}_{all}$ of the mean θ such that, with probability at least $1 - O(\alpha)$,*

$$|\tilde{\theta}_{all} - \theta| \leq \tilde{O}\left(\left(\sqrt{\mathrm{Var}(U_n)} + \frac{k^{3/2}\sqrt{K\zeta_k}}{n\epsilon}\right)\log\frac{1}{\alpha}\right)^{2},$$

as long as $n = \tilde{\Omega}\left(\frac{kK}{\epsilon}\left(\log R + \log\frac{\sigma_{max}}{\sigma_{min}}\right)\log\frac{1}{\alpha}\right)$. The estimate $\tilde{\theta}_{all}$ is 2ϵ-differentially private and runs in time $\tilde{O}\left(\left(k + \log\frac{R}{Q_\alpha}\right)\binom{n}{k}\right)$.

Remark 12 *While Lemma 11 recovers the correct first term of the deviation, the private error term is a $\sqrt{k}$ factor worse. Moreover, for the private error to be a smaller order, one requires $k^2/n = o(1)$, so this only works for $k = o(\sqrt{n})$. Note, however, that existing concentration (Hoeffding, 1948; Arcones and Gine, 1993) or convergence in probability (Vaart, 1998; Minsker, 2023) results only require $k = o(n)$ (see Lemmas A.1 and A.2 in the Appendix).*

Given computational considerations, we now focus on subsampled U statistics. Previous work has shown how to use random subsampling to obtain computationally efficient, yet statistically accurate, approximations of U statistics (Janson, 1984; Politis et al., 2012; Chen and Kato, 2019), where the sum is replaced with m (with or without replacement) random samples from $\mathcal{I}_{n,k}$.

Definition 13 (Subsampled Estimator) *Draw m i.i.d. samples $S_1, \ldots, S_m$ from the uniform distribution over the elements of $\mathcal{I}_{n,k}$, and let $\mathcal{F}_{ss} := \{S_1, \ldots, S_m\}$. Define $\hat{\theta}_{ss} = \frac{1}{m}\sum_{j=1}^{m} h(X_{S_j})$.*

Note that unlike the other families of subsets of $\mathcal{I}_{n,k}$, the family $\mathcal{F}_{\mathrm{ss}}$ is randomized. Recall from our discussion before Proposition 4 that we want each of the $h(X_{S_j})$'s as well as $\hat{\theta}_{\mathrm{ss}}$, to have good concentration around θ, and we also want $\mathrm{dep}_{n,k}(\mathcal{F}_{\mathrm{ss}})$ to be small. The former concentrations hold in the same way it holds in the "all-tuples" case, and the latter holds with high probability.

2. The dependence on $\log 1/\alpha$ is $\tilde{O}\left((\log 1/\alpha)^{\ell+1}\right)$, where ℓ is the degeneracy order of the U statistic.

Lemma 14 *Let $M = M_{n,k,\alpha}$ be some parameter, and let $\mathcal{F}_{ss}$ be the subsampling family as in Definition 13, with $m := M$. Then there exist ϵ-differentially privately computable confidence bounds Q and Q^{avg} such that Algorithm 1 with input $\left(n, k, h, \{X_i\}_{i\in[n]}, \mathcal{F}_{ss}, R, \epsilon, \alpha, Q, Q^{avg}\right)$ returns an estimate $\tilde{\theta}_{ss}$ such that, with probability at least $1 - O(\alpha)$,*

$$|\tilde{\theta}_{ss} - \theta| \leq \tilde{O}\left(\sqrt{\text{Var}(U_n)} + \sqrt{\frac{\zeta_k}{M}} + \frac{k^{3/2}\sqrt{K\zeta_k}}{n\epsilon}\right),$$

as long as $M = \Omega(\frac{n}{k}\log\frac{n}{k\alpha})$, $M = poly(n, \log\frac{1}{\alpha})$, and $n = \tilde{\Omega}\left(\frac{k}{\epsilon}\left(\log R + \log\frac{\sigma_{max}}{\sigma_{min}}\right)\log\frac{1}{\alpha}\right)$. Moreover, the estimator $\tilde{\theta}_{ss}$ is 2ϵ-differentially private and runs in time $\tilde{O}\left(M\left(k + \log\frac{R}{Q_\alpha}\right)\right)$.

Remark 15 *The condition $M = poly(n, \log 1/\alpha)$ is needed to ensure that the exponent of $\log\frac{1}{\alpha}$ that the error blows up by due to the median-of-means argument is a constant. In particular, suppose $M = \Omega\left(\frac{n^2}{Kk^3\epsilon^2} + \frac{n}{k}\log\frac{n}{k\alpha}\right)$. Then, the second term in the bound of Lemma 14 can be absorbed into the third term and we recover the same bound as the all-tuples estimator but with an $O(n^3)$ computational overhead instead of an $O(n^k)$ overhead as in the "all-tuples" case.*

4. Private estimation for atypical U statistics

We have shown that for typical non-degenerate U statistics, standard private mean estimation ideas can be applied to estimate θ privately. However, in this section, we will focus on some "atypical" U statistics that are also relevant to applications. For this section, we will assume that $\|h\|_\infty \leq C < \infty$. We will first present a general algorithm for private U statistic estimation and provide its utility and privacy guarantees. We will then show that this type of analysis can be used to obtain finer deviations in the case of degenerate and non-degenerate U statistics.

4.1. Concentration of Hájek projections

In this section, we assume the U statistic is degenerate, i.e., $\zeta_1 = 0$. We also assume that the kernel h has absolutely bounded range: $\sup h - \inf h \leq C$ for some known $C > 0$. Let $\mathcal{I}_{n,k}^{(i)}$ denote the subset of $\mathcal{I}_{n,k}$ where every element contains i. Consider the projection $\hat{h}_1(X_i) := \frac{1}{\binom{n-1}{k-1}}\sum_{S\in\mathcal{I}_{n,k}^{(i)}} h(X_S)$. We show that the $\hat{h}_1(X_i)$'s are concentrated around the conditional mean:

Lemma 16 *Let $S_i \in \mathcal{I}_{n,k}$ be a set containing i. Define $\sigma_i^2 := \text{Var}\left(h(X_{S_i})|X_i = x_i\right)$. With probability at least $1 - \beta$, conditioned on $X_i = x_i$, we have for all $1 \leq i \leq n$,*

$$\left|\hat{h}_1(X_i) - \mathbb{E}\left[h(X_{S_i})|X_i = x_i\right]\right| \leq \sqrt{\frac{4kK}{n}\log\frac{2n}{\beta}}\sigma_i + \frac{4k}{3n}\log\frac{2n}{\beta}C. \tag{8}$$

4.2. Algorithm for atypical U statistics

We design an algorithm that builds upon ideas in (Ullman and Sealfon, 2019) and exploits the concentration of the Hájek projections $\hat{h}_1(X_i)$ in the case of degenerate U statistics. Let ξ be a parameter to be defined later; this parameter will be set so that with high probability, we have $\left|\hat{h}_1(X_i) - \theta\right| \leq \xi$, as in Lemma 16. For any n-tuple $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, let

$$U_n(\mathbf{X}) := \frac{1}{\binom{n}{k}}\sum_{S\in\mathcal{I}_{n,k}} h(X_S), \qquad \hat{h}_1^{\mathbf{X}}(X_i) := \frac{1}{\binom{n-1}{k-1}}\sum_{S\in\mathcal{I}_{n,k}^i} h(X_S) \;\; \forall i \in [n], \tag{9}$$

and let $\xi_{\mathbf{X}}$ be the smallest positive integer t such that at most t indices $i \in [n]$ satisfy the condition $\left|\hat{h}_1^{\mathbf{X}}(X_i) - U_n(\mathbf{X})\right| > \xi + \frac{(2r-1)Ct}{n-1}$; such an integer always exists because $t = n$ works. Define

$$\text{Good}(\mathbf{X}) := \left\{i : \left|\hat{h}_1^{\mathbf{X}}(X_i) - U_n(\mathbf{X})\right| \leq \xi + \frac{(2r-1)C\xi_{\mathbf{X}}}{n-1}\right\} \tag{10}$$

and $\text{Bad}(\mathbf{X}) := [n] \setminus \text{Good}(\mathbf{X})$. For each $S \in \mathcal{I}_{n,k}$, define

$$g^{\mathbf{X}}(X_S) := h(X_S)\mathbb{1}\left(S \subseteq \text{Good}(\mathbf{X})\right) + U_n(\mathbf{X})\mathbb{1}\left(S \not\subseteq \text{Good}(\mathbf{X})\right).$$

Finally, define $\tilde{U}_n(\mathbf{X})$ and $\hat{g}_1^{\mathbf{X}}(X_i)$ similar to Eq 9 but replacing the function h with a g.

Algorithm 3 PrivateMeanHajek$\left(n, k, h, \{X_i\}_{i\in[2n]}, C, \epsilon, \alpha = 0.999\right)$

$\tilde{\zeta}_k \leftarrow$**PrivateVariance**$\left(n, k, h, \{X_i\}_{i\in[n+1,2n]}, C, \epsilon, \alpha\right)$
$\xi \leftarrow 2\left(2\sqrt{\frac{4kK}{n}}\sqrt{\tilde{\zeta}_k} + \frac{4k}{3n}C\right)\log\frac{12n}{\alpha}$
$\mathbf{X} \leftarrow \{X_i\}_{i\in[n]}$; $U_n(\mathbf{X}) \leftarrow \sum_{S\in\mathcal{I}_{n,k}} h(X_S)/\binom{n}{k}$
for $i = 1, 2, \ldots, n$ **do**
  $\hat{h}_1^{\mathbf{X}}(X_i) \leftarrow \sum_{S\in\mathcal{I}_{n,k}^i} h(X_S)/\binom{n-1}{k-1}$
end for
Let $\xi_{\mathbf{X}}$ be the smallest positive integer such that at most $\xi_{\mathbf{X}}$ indices i satisfy $\left|\hat{h}_1^{\mathbf{X}}(X_i) - U_n(\mathbf{X})\right| > \xi + \frac{(2k-1)C\xi_{\mathbf{X}}}{n-1}$
$\text{Good}(\mathbf{X}) \leftarrow \left\{i : \left|\hat{h}_1^{\mathbf{X}}(X_i) - U_n(\mathbf{X})\right| \le \xi + \frac{(2k-1)C\xi_{\mathbf{X}}}{n-1}\right\}$; $\text{Bad}(\mathbf{X}) \leftarrow [n] \setminus \text{Good}(\mathbf{X})$
for $S \in \mathcal{I}_{n,k}$ **do**
  $g(X_S) \leftarrow h(X_S)\mathbb{1}\left(S \subseteq \text{Good}(\mathbf{X})\right) + U_n(\mathbf{X})\mathbb{1}\left(S \not\subseteq \text{Good}(\mathbf{X})\right)$
end for
$S(\mathbf{X}) \leftarrow \max_{0\le\ell\le n}\left(4k\left(\xi_{\mathbf{X}} + \ell\right)/n\right)\left(2\xi + \left(17k\left(\xi_{\mathbf{X}} + \ell\right)C/\left(n-1\right)\right)\right)e^{-\epsilon\ell}$
$\tilde{U}_n(\mathbf{X}) \leftarrow \sum_{S\in\mathcal{I}_{n,k}} g(X_S)/\binom{n}{k}$
return $\tilde{U}_n(\mathbf{X}) + S(\mathbf{X})/\epsilon \cdot Z$, where $Z \leftarrow t_3$

The idea of Algorithm 3 is as follows: If all $\hat{h}_1^{\mathbf{X}}(X_i)$ are within ξ of the mean $U_n(\mathbf{X})$, then $\text{Bad}(\mathbf{X}) = \varnothing$ and $\tilde{U}_n(\mathbf{X}) = U_n(\mathbf{X})$. Otherwise, for any $i \in \text{Bad}(\mathbf{X})$, the quantities $h(X_S)$ are replaced with the empirical mean $U_n(\mathbf{X})$ of the entire set, for any $i \in S$. As we will show, this averaging-out of the bad indices allows for a bound on the local sensitivity of $\tilde{U}_n$ in terms of $\xi_{\mathbf{X}} = |\text{Bad}(\mathbf{X})|$, which can be viewed as an indicator of how well-concentrated the data are. We will show that $\xi_{\mathbf{X}} = 1$ with high probability, which allows for a good utility guarantee.

Theorem 17 *Algorithm 3 is 2ϵ-differentially private. Moreover, if $\zeta_1 = 0$ and $|h|_\infty \le C$, then with probability at least 0.99, we have*

$$|\mathcal{A}(\mathbf{X}) - \theta| = O\left(\sqrt{var(U_n)} + \frac{k^{3/2}\sqrt{K}\left(1+1/\epsilon\right)\log n}{n^{3/2}\epsilon}\sqrt{\zeta_k} + \frac{k^2\left(1+1/\epsilon\right)^2\log n}{n^2\epsilon}C\right).$$

Proof The computation of $\tilde{\zeta}_k$ and ξ is ϵ-differentially private. By Lemma A.5, it suffices to consider a fixed $\tilde{\zeta}_k$ and show that the rest of the algorithm is ϵ-differentially private. Consider two adjacent datasets $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ and $\mathbf{X}' = (X_1', X_2', \ldots, X_n')$ differing only in the index i^*, that is, $X_i' = X_i$ for all $i \ne i^*$. Let $B := \text{Bad}(\mathbf{X}) \cup \text{Bad}(\mathbf{X}') \cup \{i^*\}$, and let $b := |B|$. Then,

$$\begin{aligned}\binom{n}{k}\left(\tilde{U}_n(\mathbf{X}) - \tilde{U}_n(\mathbf{X}')\right) &= \sum_{S\subseteq B^c}\left(g^{\mathbf{X}}(X_S) - g^{\mathbf{X}'}(X_S')\right) + \sum_{i\in B}\sum_{S\in\mathcal{I}_{n,k}^i}\left(g^{\mathbf{X}}(X_S) - g^{\mathbf{X}'}(X_S')\right) \\ &\quad - \sum_{a=2}^{k}\sum_{S\in\mathcal{I}_{n,k}:|S\cap B|=a}(a-1)\left(g^{\mathbf{X}}(X_S) - g^{\mathbf{X}'}(X_S')\right). \qquad (11)\end{aligned}$$

The first term in the above equation counts all subsets that are contained in $\text{Good}(\mathbf{X}) \cap \text{Good}(\mathbf{X}') \setminus \{i^*\}$. The second term now sums over all subsets with an element in B. However, this leads to

overcounting every subset with a elements in common with B exactly $a-1$ times. The third term corrects for this overcounting akin to an inclusion-exclusion type argument. We will now bound each of the three terms separately.

The first term in Eq 11 is 0; indeed, if $S \subseteq B^c$ then $g^{\mathbf{X}}(X_S) = h^{\mathbf{X}}(X_S) = h^{\mathbf{X}'}(X'_S) = g^{\mathbf{X}'}(X'_S)$. Next, note that

$$\left|U_n(\mathbf{X}) - U_n(\mathbf{X}')\right| = \left|\frac{1}{\binom{n}{k}} \sum_{S \in \mathcal{I}^{i^*}_{n,k}} \left(h(X_S) - h(X'_S)\right)\right| \leq \frac{kC}{n}. \tag{12}$$

Next, for any $i \neq i^*$,

$$\left|\hat{h}_1^{\mathbf{X}}(X_i) - \hat{h}_1^{\mathbf{X}'}(X'_i)\right| = \left|\frac{1}{\binom{n-1}{k-1}} \sum_{S \in \mathcal{I}^{i}_{n,k} \cap \mathcal{I}^{i^*}_{n,k}} \left(h(X_S) - h(X'_S)\right)\right| \leq \frac{(k-1)C}{n-1}. \tag{13}$$

Before we bound the other two terms, we make the following crucial claim (proof in Appendix).

Lemma 18 $|\xi_{\mathbf{X}} - \xi_{\mathbf{X}'}| \leq 1$.

For any index $i \in [n]$, if $i \notin \text{Good}(\mathbf{X})$, then $\hat{g}_1^{\mathbf{X}}(X_i) = U_n(\mathbf{X})$; and if $i \in \text{Good}(\mathbf{X})$, then

$$\begin{aligned}\hat{g}_1^{\mathbf{X}}(X_i) &= \frac{1}{\binom{n-1}{k-1}} \sum_{i \in S} h^{\mathbf{X}}(X_S) \mathbb{1}\left(S \subseteq \text{Good}(\mathbf{X})\right) + \frac{1}{\binom{n-1}{k-1}} \sum_{i \in S} U_n(\mathbf{X}) \mathbb{1}\left(S \not\subseteq \text{Good}(\mathbf{X})\right) \\ &= \hat{h}_1^{\mathbf{X}}(X_i) + \frac{1}{\binom{n-1}{k-1}} \sum_{i \in S} \left(U_n(\mathbf{X}) - h(X_S)\right) \mathbb{1}\left(S \not\subseteq \text{Good}(\mathbf{X})\right)\end{aligned}$$

which implies $|\hat{g}_1^{\mathbf{X}}(X_i) - U_n(\mathbf{X})| \leq \xi + \frac{(4k-3)C\xi_{\mathbf{X}}}{n-1}$. Here, we used

$$\frac{\binom{n-1}{k-1} - \binom{n-1-\xi_{\mathbf{X}}}{k-1}}{\binom{n-1}{k-1}} = 1 - \prod_{t=1}^{k-1}\left(1 - \frac{\xi_{\mathbf{X}}}{n-t}\right) \leq 1 - \left(1 - \frac{\xi_{\mathbf{X}}(k-1)}{n-k+1}\right) \leq \frac{2\xi_{\mathbf{X}}(k-1)}{n}.$$

Either way, $\left|\hat{g}_1^{\mathbf{X}}(X_i) - U_n(\mathbf{X})\right| \leq \xi + \frac{(4k-3)C\xi_{\mathbf{X}}}{n-1}$, so

$$\begin{aligned}\left|\hat{g}_1^{\mathbf{X}}(X_i) - \hat{g}_1^{\mathbf{X}'}(X'_i)\right| &\leq \left|U_n(\mathbf{X}) - U_n(\mathbf{X}')\right| + 2\xi + \frac{(4k-3)C\xi_{\mathbf{X}}}{n-1} + \frac{(4k-3)C\xi_{\mathbf{X}'}}{n-1} \\ &\leq 2\xi + \frac{8k\xi_{\mathbf{X}} + 5k}{n-1}C. \end{aligned} \tag{14}$$

Using Eq 14, the second term can be bounded as

$$\begin{aligned}\left|\sum_{i \in B} \sum_{S \in \mathcal{I}^i_{n,k}} \left(g^{\mathbf{X}}(X_S) - g^{\mathbf{X}'}(X'_S)\right)\right| &\leq \binom{n-1}{k-1} \sum_{i \in B} \left|\hat{g}_1^{\mathbf{X}}(X_i) - \hat{g}_1^{\mathbf{X}'}(X'_i)\right| \\ &\leq \binom{n-1}{k-1}\left(2\xi + \frac{8k\xi_{\mathbf{X}} + 5k}{n-1}C\right)(2\xi_{\mathbf{X}} + 2), \end{aligned} \tag{15}$$

where we used the fact that $b \leq \xi_{\mathbf{X}} + \xi_{\mathbf{X}'} + 1 \leq 2\xi_{\mathbf{X}} + 2$. Finally,

$$\begin{aligned}&\frac{1}{C}\left|\frac{1}{\binom{n}{k}} \sum_{a=2}^{k} \sum_{|S \cap B| = a} (a-1)\left(g^{\mathbf{X}}(X_S) - g^{\mathbf{X}'}(X'_S)\right)\right| \leq \frac{1}{\binom{n}{k}} \sum_{a=2}^{k} (a-1)\binom{b}{a}\binom{n-b}{k-a} \\ &= \frac{\binom{n-b}{k}}{\binom{n}{k}} - \left(1 - \frac{bk}{n}\right) \leq \left(1 - \frac{b}{n}\right)^k - \left(1 - \frac{bk}{n}\right) \leq \frac{b^2k^2}{n^2} \leq \frac{k^2(2\xi_{\mathbf{X}} + 2)^2}{n^2}. \end{aligned} \tag{16}$$

where we used the identities $\sum_{a=0}^{k}\binom{b}{a}\binom{n-b}{k-a}=\binom{n}{k}$ and $\sum_{a=0}^{k}a\binom{b}{a}\binom{n-b}{k-a}=\frac{bk}{n}\binom{n}{k}$. Combining Eqs 11, 15, 16, and the fact that the first term in Eq 11 is zero,

$$\left|U_n(\mathbf{X})-U_n(\mathbf{X}')\right|\le\frac{2k(\xi_{\mathbf{X}}+1)}{n}\left(2\xi+\frac{k\,(10\xi_{\mathbf{X}}+7)}{n-1}C\right)$$

which implies that $LS_{\tilde{U}_n}(\mathbf{X})\le\frac{4k\xi_{\mathbf{X}}}{n}\left(2\xi+\frac{17k\xi_{\mathbf{X}}}{n-1}C\right)$. Let $g(\xi,\xi_{\mathbf{X}},n):=\frac{4k\xi_{\mathbf{X}}}{n}\left(2\xi+\frac{17k\xi_{\mathbf{X}}}{n-1}C\right)$; note that it is strictly increasing in $\xi_{\mathbf{X}}$. Define

$$S(G)=\max_{\ell\in\mathbb{Z}_{\ge0}}e^{-\epsilon\ell}g(\xi,\xi_{\mathbf{X}}+\ell,m). \tag{17}$$

Lemma 19 *$S(G)$ is an ϵ-smooth upper bound on $LS_f(G)$. Moreover,*

$$S(G)=O\left(\frac{k\,(\xi_{\mathbf{X}}+1/\epsilon)\,(\xi+Ck(\xi_{\mathbf{X}}+1/\epsilon)/n)}{n}\right).$$

By Lemma 19, it is clear that the term $S(\mathbf{X})$ added in to $\tilde{U}_n(\mathbf{X})$ in Algorithm 3 is exactly the smoothed sensitivity defined in Eq 17. Therefore, the output $\tilde{U}_n(\mathbf{X})+S(\mathbf{X})/\epsilon\cdot Z$, where Z is sampled from a Student's t-distribution with three degrees of freedom, is ϵ-differentially private.

Utility. First, by Chebyshev's inequality, we have with probability at least $1-\alpha$ that $|U_n(\mathbf{X})-\theta|\le\frac{1}{\sqrt{\alpha}}\sqrt{\mathrm{var}(U_n)}$. Second, with probability at least $1-\alpha$ the estimate $\tilde{\zeta}_k$ of the variance of h satisfies $\zeta_k\le\tilde{\zeta}_k\le2\zeta_k$. First note that since $\mathrm{Var}(h(X_1,\ldots,X_k|X_i))=\zeta_1=0$, we have $\mathbb{E}[h(X_1,\ldots,X_k)|X_i]=\mathbb{E}[h(X_1,\ldots,X_k)]=\theta$.

Next, note that σ_i^2 are IID subexponential random variables with subexponential norm at most $K\zeta_k$, because by Jensen's inequality, $\mathbb{E}\left[\exp\left(t\sigma_i^2\right)\right]\le\mathbb{E}\left[\exp\left(t\left(h(X_S)-\theta\right)^2\right)\right]$. Conditioned on this event, and using Lem 16, with probability at least $1-\alpha$ all $\hat{h}_1^{\mathbf{X}}(X_i)$ satisfy

$$\left|\hat{h}_1^{\mathbf{X}}(X_i)-\theta\right|\le\left(2\sqrt{\frac{4kK}{n}}\sqrt{\tilde{\zeta}_k}+\frac{4k}{3n}C\right)\log\frac{12n}{\alpha}\le\frac{\xi}{2}.$$

Moreover, each of the projections $\hat{h}_1^{\mathbf{X}}(X_i)$ is within $\xi/2$ of the true mean θ with probability at least $1-\alpha$. If so, then each of the projections is also within ξ of the empirical mean of the projections. This means $\mathrm{Good}(\mathbf{X})=[n]$ and $\xi_{\mathbf{X}}=1$. Also, since all indices are good, $g(X_S)=h(X_S)$ for all S and $\tilde{U}_n(\mathbf{X})=U_n(\mathbf{X})$. Finally, with probability at least $1-\alpha$, the Z satisfies $Z\le\frac{1}{\sqrt{\alpha}}\mathrm{var}(t_3)=\frac{3}{\sqrt{\alpha}}$. Therefore,

$$\begin{aligned}|\mathcal{A}(\mathbf{X})-\theta|&\le\left|\tilde{U}_n(\mathbf{X})-U_n(\mathbf{X})\right|+|U_n(\mathbf{X})-\theta|+|S(\mathbf{X})/\epsilon\cdot Z|\\&=O\left(\sqrt{\mathrm{var}(U_n)}+\frac{k^{3/2}\sqrt{K}\,(1+1/\epsilon)\log n}{n^{3/2}\epsilon}\sqrt{\zeta_k}+\frac{k^2\,(1+1/\epsilon)^2\log n}{n^2\epsilon}C\right)\end{aligned} \tag{18}$$

conditioned on all the aforementioned events, which occur with probability 0.99. ■

5. Applications

We now discuss several applications illustrating the usefulness of our algorithmic framework.

5.1. Sparse graph statistics

Consider a geometric random graph as in (Gilbert, 1961). Here, each edge is of the form $g(X_i,X_j):=1(\|X_i-X_j\|_2\le r_n)$, where $X_i\in\mathbb{R}^d$. Assume for concreteness that X_i is uniformly distributed in the d-dimensional unit sphere $\mathcal{B}_d(1)$. If we apply Algorithm 1 with $Q_\alpha=1$, (since $Y_i\le1$ a.s.), there will be no clipping and the algorithm will simply return the mean of the Y_i's with suitable Laplace noise added. Take, for example, the all tuples estimator. The Laplace noise parameter

will be $O(k/n)$. For non-degenerate kernels, this may suffice since the main non-private element of the deviation $\hat{\theta} - \theta$ is $O(\sqrt{k^2\zeta_k/n})$. The catch is that for sparse graphs with $r_n \to 0$, we also have $\zeta_k \to 0$ with n.

Chapter 2 of (Gilbert, 1961) states that as long as the expected subgraph count for a subgraph with k vertices is $n^k(r_n)^{d(k-1)} \to \infty$, and $nr_n^d \to 0$, one has normal convergence of subgraph counts. Consider a concrete example with triangle counts with $k = 3$ and $d = 2$. Then $h(X_i, X_j, X_k) = g(X_i, X_j)g(X_j, X_k)g(X_k, X_\ell)$. Furthermore, $\mathbb{E}g(X_i, X_j) \propto r_n^2$, since this is the probability measure in a ball of radius r_n in $\mathbb{R}^2$, and $\mathbb{E}h(X_i, X_j, X_k) \propto r_n^4$, since this is the probability that i, j, and j, k are close.

Hence, the expected triangle count is $O(n^3 r_n^4)$, which tends to infinity if $r_n \gg n^{-3/4}$. The limiting variance of this U statistic will be $O(\zeta_3/n)$, where $\zeta_3 = O(\mathbb{E}g(X_i, X_j, X_k)) = O(r_n^4)$. Thus, the variance is $O(r_n^4/n)$, which may be smaller than the private error $O(1/n^2\epsilon^2)$ when the expected count is going to infinity.

This is why we will apply Algorithm 3 in this setting. Recall the definition of the Hajek projection from Eq 8. Since $\zeta_1 > 0$, Lemma 16 shows that $\hat{h}_1(X_i)$ concentrates around its conditional mean. Recall that $\sigma_i^2 := \mathrm{Var}(h(X_S)|X_i) \le \mathbb{E}[h(X_S)|X_i]$. Using Lemma 16, with probability at least $1 - \frac{\beta}{2n}$, conditioned on X_i, we have $|\hat{h}_1(X_i) - \theta| \le \sqrt{\mathbb{E}\left[h(X_S)|X_i\right]}(1 + \sqrt{\frac{4k}{n}}\sqrt{\log\frac{2n}{\beta}}) + \frac{4Ck}{3n}\log\frac{2n}{\beta} + \theta$.

Let $\mathcal{E}_1 := \{k \text{ points are within distance } r_n \text{ of } x\}$. We have $\mathbb{E}[h(X_S)|X_i] = P(\mathcal{E}_1|X_i = x)$, which can be bounded by $\sqrt{c_d r_n^{d(k-1)}}1(X_i \in \mathcal{B}_d(1+r_n))$. Hence, with probability at least $1-2\beta$, we have $|\hat{h}_1(X_i) - \theta| = \tilde{O}\left(\sqrt{c_d' r_n^{d(k-1)}} + \frac{4k}{3n}\log\frac{2n}{\beta}\right)$. Using this deviation bound with a similar argument as in Eq 18, we have $|\mathcal{A}(X) - \theta| = O\left(\sqrt{\mathrm{Var}(U_n)} + \frac{k}{n}\left(\sqrt{r_n^{d(k-1)}} + \frac{k}{n}\right)\right)$. Recall that $\mathrm{Var}(U_n) = O((r_n)^{d(k-1)}k/n)$ (using Eq 3). Hence, in situations like this, where all conditional expectations are small, we can obtain confidence intervals of length $O(1/n^2)$.

5.2. Statistial inference on random graphs

More generally, Biau and Bleakley (2006) consider a random graph as an i.i.d. sequence of random vectors $\{(X_i^n, X_j^n, Y_{ij}^n)\}_n$, taking values in $\mathcal{X} \times \mathcal{X} \times \{-1, 1\}$, where the pairs (X_i^n, X_j^n, Y_{ij}^n) and (X_k^n, X_l^n, Y_{kl}^n) are independent for $\{i, j\} \cap \{k, l\} = \emptyset$. For a reconstruction rule $g : \mathcal{X} \times \mathcal{X} \to \{-1, 1\}$, we can define the reconstruction risk $R_n(g) = \frac{1}{n(n-1)}\sum_{1\le i\ne j\le n} 1\{g(X_i^n, X_j^n) \ne Y_{ij}^n\}$. Note that this is a U statistic when g is a deterministic rule. In a typical statistical inference procedure, we might wish to minimize R_n over a candidate set of reconstruction rules, or test a hypothesis that $g = g_0$. The methods we have described in this paper would allow us to estimate $R_n(g)$ privately, to a certain level of accuracy.

5.3. Goodness-of-fit testing

The Cramer-Von Mises statistic for testing the hypothesis that the cumulative distribution function of a random variable is equal to a function F_0 is given by

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\int \left(1\{X_i \le x\} - F_0(x)\right)\left(1\{X_j \le x\} - F_0(x)\right) dF_0(x).$$

Under the null hypothesis $H_0 : X \sim F_0$, the distribution of the Cramer Von-Mises statistic is a degenerate U statistic (Vaart, 1998). Thus, our techniques presented in Section 4.2 provide a method for private goodness-of-fit testing based on the Cramer-Von Mises statistic. We note that private goodness-of-fit testing has so far mostly been studied in the setting of discrete data (Gaboardi et al., 2016; Acharya et al., 2018; Aliakbarpour et al., 2019). For continuous distributions, we are

only aware of work that analyzes the LDP framework (Dubois et al., 2019; Lam-Weil et al., 2022; Butucea et al., 2023), which is therefore not directly comparable to our proposed approach.

6. Discussion

In this paper, we have shown that for a broad class of standard U statistics, one can use existing private mean estimation tools to obtain error bounds where the error injected to preserve privacy does not occlude the irreducible error resulting from the variance of the nonprivate estimator. However, in atypical cases, where the U statistics has variance $O(1/n^2)$ as opposed to $O(1/n)$, the private error may overwhelm the true variance. We have proposed a new algorithm that uses Hajek projections to reweight different subsets of data appearing in a U statistic. This respects sensitivity requirements, while not differing too much from the original estimator. We have discussed a variety of applications in sparse random geometric graphs and goodness-of-fit testing.

References

J. Acharya, Z. Sun, and H. Zhang. Differentially private testing of identity and closeness of discrete distributions. *Advances in Neural Information Processing Systems*, 31, 2018.

M. Aliakbarpour, I. Diakonikolas, D. Kane, and R. Rubinfeld. Private testing of distributions via sample permutations. *Advances in Neural Information Processing Systems*, 32, 2019.

T. W. Anderson and D. Darling. Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *Annals of Mathematical Statistics*, 23:193–212, 1952. URL https://api.semanticscholar.org/CorpusID:120541257.

M. A. Arcones and E. Gine. Limit theorems for U-processes. *The Annals of Probability*, 21(3): 1494 – 1542, 1993. doi: 10.1214/aop/1176989128. URL https://doi.org/10.1214/aop/1176989128.

J. Bell, A. Bellet, A. Gascón, and T. Kulkarni. Private protocols for U-statistics in the local model and beyond. In *International Conference on Artificial Intelligence and Statistics*, pages 1573–1583. PMLR, 2020.

G. Biau and K. Bleakley. Statistical inference on graphs. *Statistics & Decisions*, 24(2):209–232, 2006.

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

M. Bun and T. Steinke. Average-case averages: Private algorithms for smooth sensitivity and mean estimation. *Advances in Neural Information Processing Systems*, 32, 2019.

C. Butucea, A. Rohde, and L. Steinberger. Interactive versus noninteractive locally differentially private estimation: Two elbows for the quadratic functional. *The Annals of Statistics*, 51(2): 464–486, 2023.

T. T. Cai, Y. Wang, and L. Zhang. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *The Annals of Statistics*, 49(5):2825–2850, 2021.

K. Chaudhuri and D. Hsu. Convergence rates for differentially private statistical estimation. In *Proceedings of the International Conference on Machine Learning*, volume 2012, page 1327. NIH Public Access, 2012.

X. Chen and K. Kato. Randomized incomplete U-statistics in high dimensions. *The Annals of Statistics*, 47(6):3127 – 3156, 2019. doi: 10.1214/18-AOS1773. URL https://doi.org/10.1214/18-AOS1773.

S. Clémençon. A statistical view of clustering performance through the theory of U-processes. *Journal of Multivariate Analysis*, 124:42–56, 2014.

S. Clémençon, G. Lugosi, and N. Vayatis. Ranking and empirical minimization of U-statistics. *The Annals of Statistics*, 36(2):844–874, 2008.

T. de Wet. Degenerate U- and V-statistics. *South African Statistical Journal*, 21:99–129, 1987. URL https://api.semanticscholar.org/CorpusID:125297203.

A. Dubois, T. B. Berrett, and C. Butucea. Goodness-of-fit testing for Hölder continuous densities under local differential privacy. In *Foundations of Modern Statistics*, pages 53–119. Springer, 2019.

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

E. W. Frees. Infinite order U-statistics. *Scandinavian Journal of Statistics*, 16(1):29–45, 1989. ISSN 03036898, 14679469. URL http://www.jstor.org/stable/4616120.

M. Gaboardi, H. Lim, R. Rogers, and S. Vadhan. Differentially private chi-squared hypothesis testing: Goodness of fit and independence testing. In *International conference on Machine Learning*, pages 2111–2120. PMLR, 2016.

B. Ghazi, P. Kamath, R. Kumar, P. Manurangsi, and A. Sealfon. On computing pairwise statistics with local differential privacy. *Advances in Neural Information Processing Systems*, 33:14475–14485, 2020.

E. N. Gilbert. Random plane networks. *Journal of The Society for Industrial and Applied Mathematics*, 9:533–543, 1961. URL https://api.semanticscholar.org/CorpusID:122310882.

G. Gregory. Large sample theory for U-statistics and tests of fit. *Annals of Statistics*, 5:110–123, 1977. URL https://api.semanticscholar.org/CorpusID:120342769.

P. R. Halmos. The theory of unbiased estimation. *The Annals of Mathematical Statistics*, 17(1):34 – 43, 1946. doi: 10.1214/aoms/1177731020. URL https://doi.org/10.1214/aoms/1177731020.

H.-C. Ho and G. S. Shieh. Two-stage U-statistics for hypothesis testing. *Scandinavian Journal of Statistics*, 33, 2006. URL https://api.semanticscholar.org/CorpusID:121782330.

W. Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293 – 325, 1948. doi: 10.1214/aoms/1177730196. URL https://doi.org/10.1214/aoms/1177730196.

W. Hoeffding. Probability inequalities for sum of bounded random variables. 1963. URL https://api.semanticscholar.org/CorpusID:121341745.

S. Janson. The asymptotic distributions of incomplete U-statistics. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 66(4):495–505, 1984.

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

G. Kamath, O. Sheffet, V. Singhal, and J. Ullman. Differentially private algorithms for learning mixtures of separated Gaussians. *Advances in Neural Information Processing Systems*, 32, 2019b.

G. Kamath, V. Singhal, and J. Ullman. Private mean estimation of heavy-tailed distributions. *ArXiv*, abs/2002.09464, 2020. URL https://api.semanticscholar.org/CorpusID:211252621.

V. Karwa and S. Vadhan. Finite sample differentially private confidence intervals. *arXiv preprint arXiv:1711.03908*, 2017.

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

J. Lam-Weil, B. Laurent, and J.-M. Loubes. Minimax optimal goodness-of-fit testing for densities and multinomials under a local differential privacy constraint. *Bernoulli*, 28(1):579–600, 2022.

A. J. Lee. *U-Statistics: Theory and Practice*. 1990. URL https://api.semanticscholar.org/CorpusID:125216198.

C. Li and X. Fan. On nonparametric conditional independence tests for continuous variables. *WIREs Computational Statistics*, 12(3):e1489, 2020. doi: https://doi.org/10.1002/wics.1489. URL https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wics.1489.

O. Linton and P. Gozalo. Testing conditional independence restrictions. *Econometric Reviews*, 33 (5-6):523–552, 2014. doi: 10.1080/07474938.2013.825135. URL https://doi.org/10.1080/07474938.2013.825135.

S. Minsker. U-statistics of growing order and sub-Gaussian mean estimators with sharp constants, 2023.

K. Nissim, S. Raskhodnikova, and A. Smith. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the Thirty-Ninth Annual ACM Symposium on Theory of Computing*, pages 75–84, 2007.

W. Peng, T. Coleman, and L. Mentch. Asymptotic distributions and rates of convergence for random forests via generalized U-statistics. *arXiv preprint arXiv:1905.10651*, 2019.

D. N. Politis, J. P. Romano, and M. Wolf. *Subsampling*. Springer Science & Business Media, 2012.

R. J. Serfling. *Approximation theorems of mathematical statistics*. Wiley Series in Probability and Mathematical Statistics : Probability and Mathematical Statistics. Wiley, New York, NY [u.a.], [nachdr.] edition, 1980. ISBN 0471024031. URL http://gso.gbv.de/DB=2.1/CMD?ACT=SRCHA&SRT=YOP&IKT=1016&TRM=ppn+024353353&sourceid=fbw_bibsonomy.

G. R. Shorack and J. A. Wellner. *Empirical Processes with Applications to Statistics*. 2009. URL https://api.semanticscholar.org/CorpusID:5891472.

Y. Song, X. Chen, and K. Kato. Approximating high-dimensional infinite-order U-statistics: Statistical and computational guarantees. 2019.

J. Ullman and A. Sealfon. Efficiently estimating erdos-renyi graphs with node differential privacy. *Advances in Neural Information Processing Systems*, 32, 2019.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

N. C. Weber. Incomplete degenerate u-statistics. *Scandinavian Journal of Statistics*, 8(2):120–123, 1981. ISSN 03036898, 14679469. URL http://www.jstor.org/stable/4615821.

Appendix A.1. Auxiliary results

Lemma A.1 *Suppose $k \leq n/2$.*

(i) If $\zeta_1 > 0$, we have

$$\mathrm{Var}(U_n) = \frac{k^2\zeta_1}{n} + O\left(\zeta_k \frac{k^2}{n^2}\right). \tag{A.19}$$

(ii) If $\zeta_1 = 0$ but $\zeta_2 > 0$, we have

$$\mathrm{Var}(U_n) = \frac{k^2(k-1)^2\zeta_2}{2n(n-1)} + O\left(\zeta_3 \frac{k^3}{n^3}\right). \tag{A.20}$$

Proof This result follows directly from a calculation appearing in the proof of Theorem 3.1 in Minsker (2023). Consider the kernels $h^{(1)}, h^{(2)}, \ldots, h^{(k)}$ of degrees $1, 2, \ldots, k$ respectively corresponding to the Hoeffding decomposition of the U statistic based on kernel h, and define

$$\delta_j^2 := \mathrm{var}(h^{(j)})$$

for all $j \in [k]$. Well-known properties of the Hoeffding decomposition gives

$$\zeta_k = \sum_{j=1}^{k} \binom{k}{j} \delta_j^2,$$

so $\delta_j^2 \leq \frac{\zeta_k}{\binom{k}{j}}$ for all j. Moreover,

$$\mathrm{Var}(U_n) = \sum_{j=1}^{k} \frac{\binom{k}{j}^2}{\binom{n}{j}} \delta_j^2.$$

For part (i), we write

$$\begin{aligned}
\mathrm{Var}(U_n) &= \frac{k^2\zeta_1}{n} + \sum_{j=2}^{k} \frac{\binom{k}{j}^2}{\binom{n}{j}} \delta_j^2 \leq \frac{k^2\zeta_1}{n} + \zeta_k \sum_{j=2}^{k} \frac{\binom{k}{j}}{\binom{n}{j}} \leq \frac{k^2\zeta_1}{n} + \zeta_k \sum_{j=2}^{k} \left(\frac{k}{n}\right)^j \\
&= \frac{k^2\zeta_1}{n} + \frac{k^2\zeta_k}{n^2}\left(1 - \frac{k}{n}\right)^{-1} \leq \frac{k^2\zeta_1}{n} + \frac{2k^2\zeta_k}{n^2},
\end{aligned}$$

whereas for part (ii), we write

$$\begin{aligned}
\mathrm{Var}(U_n) &= \frac{k^2\zeta_1}{n} + \frac{k^2(k-1)^2\zeta_2}{2n(n-1)} + \sum_{j=3}^{k} \frac{\binom{k}{j}^2}{\binom{n}{j}} \zeta_j \leq \frac{k^2(k-1)^2\zeta_2}{2n(n-1)} + \zeta_k \sum_{j=3}^{k} \frac{\binom{k}{j}}{\binom{n}{j}} \\
&\leq \frac{k^2(k-1)^2\zeta_2}{2n(n-1)} + \zeta_k \sum_{j=3}^{k} \left(\frac{k}{n}\right)^j = \frac{k^2(k-1)^2\zeta_2}{2n(n-1)} + \frac{2k^3\zeta_k}{n^3}.
\end{aligned}$$

■

Lemma A.2 *(Hoeffding, 1963)*

(i) If $\mathcal{H}$ is sub-Gaussian with variance proxy σ^2, then for all $t > 0$, we have

$$\mathbb{P}\left(|U_n - \theta| \geq t\right) \leq 2\exp\left(-\frac{\lfloor \frac{n}{k} \rfloor t^2}{2\sigma^2}\right). \tag{A.21}$$

(ii) If $\mathcal{H}$ is almost surely bounded in $(-C, C)$, then for all $t > 0$, we have

$$\mathbb{P}\left(|U_n - \theta| \geq t\right) \leq \exp\left(\frac{-\lfloor \frac{n}{k} \rfloor t^2}{2\zeta_k + 2Ct/3}\right). \tag{A.22}$$

Proof Without loss of generality, let $\theta = 0$. For any permutation σ of $[n]$, let

$$V_\sigma := \frac{1}{\lfloor n/k \rfloor} \sum_{i=1}^{\lfloor n/k \rfloor} h(X_{\sigma(k(i-1)+1)}, X_{\sigma(k(i-1)+2)}, \ldots, X_{\sigma(ki)})$$

By symmetry, $U_n = \frac{1}{n!}\sum_\sigma V_\sigma$. For any $s > 0$,

$$\begin{aligned}
\mathbb{P}\left(U_n \geq t\right) = \mathbb{P}\left(e^{sU_n} \geq e^{st}\right) &\leq e^{-st}\mathbb{E}\left[e^{sU_n}\right] = e^{-st}\mathbb{E}\left[\exp\left(\frac{s}{n!}\sum_\sigma V_n\right)\right] \\
&\leq e^{-st}\mathbb{E}\left[\frac{1}{n!}\sum_\sigma \exp(sV_n)\right] = e^{-st}\mathbb{E}\left[\exp(sV_{\text{id}})\right] \\
&= e^{-st}\mathbb{E}\left[\exp\left(\frac{s}{\lfloor n/k \rfloor} h(X_1, \ldots, X_k)\right)\right]^{\lfloor n/k \rfloor}
\end{aligned}$$

The first inequality used Markov's inequality, the second inequality used Jensen's inequality, the third equality is true by symmetry, and the last equality used independence of the $\lfloor n/k \rfloor$ summands. ■

Appendix A.2. Proofs of Proposition 4 and 8

Proof [Proof of Proposition 4] We will prove privacy and accuracy guarantees separately.

Privacy. Algorithm 1 makes $t+1$ calls to Algorithm 2; let Δ_i, W_i, and Z_i be the values taken by Δ, W, and Z in the ith call to Algorithm 2, for $1 \leq i \leq t+1$. Let $\beta := 0.01/t$. It can be shown inductively that the interval lengths $r_i - l_i$ and the values Δ_i do not depend on the dataset. For any $1 \leq i \leq t$, $Y_{i,j} = \text{proj}_{l_{i-1}-Q_\beta, r_{i-1}, Q_\beta}(Y_{i-1,j})$ for all $1 \leq j \leq m$. Suppose we change X_w to X'_w for some index w. For any $1 \leq i \leq t+1$, conditioned on the values of $Z_{i'}$ for $1 \leq i' < i$, at most an $\text{dep}_{n,k}(\mathcal{F})$ fraction of $\{Y_{i,j}\}_{j\in[m]}$ *depend* on w (this is true by the definition of $\text{dep}_{n,k}(\mathcal{F})$). Since $Y_{i,j} = \text{proj}_{l_{i-1}-Q_\beta, r_{i-1}+Q_\beta}(Y_{i-1,j})$ has range $r_{i-1} - l_{i-1} + 2Q_\beta$, the sensitivity of $\frac{1}{m}\sum_{j=1}^m Y_{i,j}$ is at most $\text{dep}_{n,k}(\mathcal{F})(r_{i-1} - l_{i-1} + 2Q_\beta) = \Delta_i$. Therefore, for all $1 \leq i \leq t$, the output Z_i (and therefore the interval $[l_i, r_i]$), conditioned on $Z_{i'}$ for $1 \leq i' < i$, is $\epsilon/2t$-differentially private. Similarly, the output $(l_{t+1} + r_{t+1})/2 = Z_{t+1}$, conditioned on $\{Z_i\}_{i\in[t]}$, is $\epsilon/2$-differentially private. By Basic Composition (see Lemma A.5), Algorithm 1 is ϵ-differentially private.

Utility. First, we show that if Algorithm 2 is invoked with $\theta \in [l, r]$, it returns an interval $[l', r']$ such that $\theta \in [l', r']$ with probability at least $1 - 3\beta$. Consider running a variant of Algorithm 1 with the projection step omitted in every call of Algorithm 2. Then, with probability at least $1 - \beta$, we have $\left|\frac{1}{m}\sum_{i=1}^m Y_i - \theta\right| \leq Q_\beta^{\text{avg}}$, and with probability at least $1 - \beta$, we have $|W| \leq \frac{\Delta}{\epsilon'}\log\frac{1}{\beta}$. Therefore, with probability at least $1 - 2\beta$, we have

$$|Z - \theta| \leq Q_\beta^{\text{avg}} + \frac{\Delta}{\epsilon}\log\frac{1}{\beta}.$$

Finally, bringing back the projection step does not project any of the m values Y_i with probability at least β. Therefore, $\theta \in [l', r']$ with probability at least $1 - 3\beta$.

Next, we claim that if $r - l > 10Q_\alpha$, then $r' - l' \le (r-l)/2$. Indeed, as long as

$$\mathrm{dep}_{n,k}(\mathcal{F}) \le \frac{Q_\alpha \epsilon}{10 t Q_{\alpha/t} \log t/\alpha} \le \min\left(\frac{\epsilon'}{10 \log 1/\beta}, \frac{Q_\alpha \epsilon'}{2 Q_\beta \log 1/\beta}\right),$$

and $Q_\beta^{\mathrm{avg}} < Q_\alpha$, we have

$$\begin{aligned} r' - l' &= \frac{2\mathrm{dep}_{n,k}(\mathcal{F}) \log 1/\beta}{\epsilon'}(r-l) + \left(2Q_\beta^{\mathrm{avg}} + \frac{4\mathrm{dep}_{n,k}(\mathcal{F}) Q_\beta \log 1/\beta}{\epsilon'}\right) \\ &\le \frac{r-l}{5} + (2Q_\alpha + Q_\alpha) \le \frac{r-l}{2}. \end{aligned}$$

Consider the for loop in Algorithm 1, which invokes Algorithm 2 t times. By a union bound, with probability at least $1 - 3t\beta = 0.97$, we have $\theta \in [l_t, r_t]$ and none of the projection operations in Algorithm operate on any element outside their projection intervals. Moreover, our parameter choices ensure that the length $r_t - l_t$ of the interval $[l_t, r_t]$ is at most $10Q_\alpha$. Condition on all these events. Finally, consider lines 8 and 9 of Algorithm 1. The algorithm returns the midpoint of the interval $[l_{t+1}, r_{t+1}]$, which is the Z_{t+1} in the final call of Algorithm 2. By Chebyshev's inequality,

$$\left|\frac{1}{m}\sum_{j=1}^{m} Y_{0,j} - \theta\right| \le 10\sqrt{\mathrm{Var}\left(\frac{1}{m}\sum_{i=1}^{m} Y_{0,j}\right)}, \tag{A.23}$$

with probability at least 0.99, and with probability at least 0.99, none of the Y_i's are truncated in the projection step in the final call of Algorithm 2. Finally, with probability at least 0.99, we have

$$W_{t+1} = O\left(\frac{\Delta_{t+1}}{\epsilon}\right) = O\left(\frac{\mathrm{dep}_{n,k}(\mathcal{F}) Q_\alpha}{\epsilon} \log \frac{1}{\alpha}\right).$$

The conclusion follows from a union bound over all events. ■

Proof [Proof of Corollary 8] First, suppose the variance ζ_k is known. For any index $i \in [n]$, there is exactly one index $j \in [m]$ such that i belongs to (the only set) $S \in \mathcal{I}_j$. Therefore, $\mathrm{dep}_{n,k}(\mathcal{F}_{\mathrm{naive}}) = \frac{k}{n}$. Next, by the assumption that $h(X_S)$ is K-subgaussian,

$$P(|h(X_S) - \theta| \ge y) \le 2\exp\left(-\frac{y^2}{2K\zeta_k}\right).$$

Hence, with probability $1 - \alpha/m$,

$$|Y_i - \theta| \le \sqrt{2K\zeta_k \log(2m/\alpha)}. \tag{A.24}$$

By a union bound, we get a valid bound $Q_\alpha = \sqrt{2K\zeta_k \log(2n/k\alpha)}$. Moreover, since the Y_i's are independent, $\frac{1}{m}\sum_{j\in[m]} Y_j$ is K-subgaussian with variance $\frac{\zeta_k}{m}$. Therefore,

$$P\left(\left|\frac{1}{m}\sum_i Y_i - \theta\right| \ge y\right) \le 2\exp\left(-\frac{my^2}{2K\zeta_k}\right)$$

This gives us a bound of $Q_\alpha^{\mathrm{avg}} = \sqrt{\frac{2Kk\zeta_k \log(2/\alpha)}{n}}$. It remains to verify the conditions of Proposition 4.

$$\frac{k}{n} \le \frac{Q_\alpha \epsilon}{10 t Q_{\alpha/t} \log(t/\alpha)} \iff \frac{k}{n} \le \frac{\epsilon}{10 t \log(t/\alpha)}\sqrt{\frac{\log(2n/k\alpha)}{\log(2nt/k\alpha)}},$$

and

$$Q^{\text{avg}}_{\alpha/t} < Q_\alpha \iff \sqrt{\frac{2Kk\zeta_k \log(2t/\alpha)}{n}} \leq \sqrt{2K\zeta_k \log(2n/k\alpha)}$$
$$\iff n \geq k\frac{\log(2t/\alpha)}{\log(2n/k\alpha)},$$

which are both true because $n = \tilde{\Omega}\left(\frac{k \log R}{\epsilon} \log \frac{1}{\alpha}\right)$. Therefore, with probability at least $1 - O(\alpha)$,

$$\left|\hat{\theta}_{\text{naive}} - \theta\right| \leq O\left(\frac{1}{\sqrt{\alpha}}\text{Var}(\hat{\theta}_{\text{naive}}) + \frac{k}{n\epsilon}\sqrt{2K\zeta_k \log(2n/k\alpha)} \log \frac{1}{\alpha}\right)$$

To fix the issue of ζ_k being unknown, note that we can instead use any privately computed upper bound on ζ_k. We will use half the data to estimate ζ_k and the other half to estimate θ, both differentially privately.

To estimate ζ_k, we split $[n]$ into n/k independent sets $S_1, \ldots, S_{n/k}$ of size k, and compute $h(X_{S_i})$ for each i. These give us $\frac{n}{k}$ IID data with variance $\zeta_k \in (\sigma_{\min}, \sigma_{\max})$. By Lemma 7 and the assumption on n, we can get an ϵ-differentially private estimate $\tilde{\zeta}_k$ of ζ_k such that with probability at least $1 - \alpha$, $\zeta_k \leq \tilde{\zeta}_k \leq 2\zeta_k$. The argument now goes through by using $\tilde{\zeta}_k$ instead of ζ_k for the bounds on Q_α and Q^{avg}_α.

■

Appendix A.3. Proof of Lemma 14

In order to prove this, we will need the following results.

Lemma A.3 *We have* $\text{Var}\left[\hat{\theta}_{ss}\right] = \left(1 - \frac{1}{m}\right)\text{Var}(U_n) + \frac{1}{m}\zeta_k$.

Proof Clearly, $\mathbb{E}[\hat{\theta}_{\text{ss}}] = \theta$.We compute both terms of the following decomposition of the variance of $\hat{\theta}_{\text{ss}}$ separately; recall that $\mathbf{X} = \{X_i\}_{i\in[n]}$:

$$\text{Var}\left(\hat{\theta}_{\text{ss}}\right) = \text{Var}\left(\mathbb{E}\left[\hat{\theta}_{\text{ss}}|\mathbf{X}\right]\right) + \mathbb{E}\left[\text{Var}\left(\hat{\theta}_{\text{ss}}|\mathbf{X}\right)\right].$$

Now,

$$\text{Var}\left(\mathbb{E}\left[\hat{\theta}_{\text{ss}}|\mathbf{X}\right]\right) = \text{Var}\left(\mathbb{E}\left[\frac{1}{m}\sum_{j\in[m]} h\left(X_{S_i}\right)\middle|\mathbf{X}\right]\right) = \text{Var}(U_n),$$

and

$$\mathbb{E}\left[\text{Var}\left(\hat{\theta}_{\text{ss}}\middle|\mathbf{X}\right)\right] = \mathbb{E}\left[\text{Var}\left(\frac{1}{m}\sum_{j=1}^{m} h(X_{S_j})\middle|\mathbf{X}\right)\right] = \frac{1}{m}\mathbb{E}\left[\text{Var}\left(h(X_S)|\,\mathbf{X}\right)\right]$$
$$= \frac{1}{m}\mathbb{E}\left[\frac{1}{\binom{n}{k}}\sum_{S\in\mathcal{I}_{n,k}} h(X_S)^2 - \left(\frac{1}{\binom{n}{k}}\sum_{S\in\mathcal{I}_{n,k}} h(X_S)\right)^2\right]$$
$$= \frac{1}{m}\left(\left(\zeta_k + \theta^2\right) - \left(\text{Var}(U_n) + \theta^2\right)\right) = \frac{\zeta_k - \text{Var}(U_n)}{m}.$$

Adding the two equalities yields the result. ■

Lemma A.4 *Let $\alpha > 0$, and let $M = \Omega(\frac{n}{k}\log\frac{n}{k\alpha})$ for a sufficiently large universal constant. Then $dep_{n,k}(\mathcal{F}_{ss}) \le \frac{4k}{n}$ with probability at least $1-\alpha$.*

Proof Let Z_i be the number of sampled subsets of which i is an element. Observe that Z_i is $\text{Binom}(M, k/n)$, with mean $\mu = Mk/n$. By a Chernoff bound, for any $\delta > 0$ and any $i \in [n]$,

$$\mathbb{P}\left(Z_i \ge (1+\delta)\mu\right) \le \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^\mu \tag{A.25}$$

By a union bound,

$$\mathbb{P}\left(\text{dep}_{n,k}\left(\mathcal{F}_{\text{ss}}\right) > \frac{4k}{n}\right) = \mathbb{P}\left(\max_i Z_i > 4\mu\right) \le M\left(\frac{e^3}{(1+3)^{1+3}}\right)^\mu \le M\exp\left(-\frac{Mk}{n}\right),$$

which is at most α by our choice of M. Let $\mathcal{G}_\alpha$ be the event that $\text{dep}_{n,k}(\mathcal{F}_{\text{ss}}) \le \frac{4k}{n}$. ■

Proof [Proof of Lemma 19] Clearly, $S(G) \ge g(\xi, \xi_{\mathbf{X}}, n) \ge LS_f(G)$, and for any two adjacent G and G',

$$\begin{aligned}
S(G') &= \max_{\ell\in\mathbb{Z}_{\ge 0}} e^{-\epsilon\ell} g(\xi, \xi_{\mathbf{X}'} + \ell, n) \le \max_{\ell\in\mathbb{Z}_{\ge 0}} e^{-\epsilon\ell} g(\xi, \xi_{\mathbf{X}} + \ell + 1, n) \\
&= \max_{\ell\in\mathbb{Z}_{>0}} e^{-\epsilon(\ell-1)} g(\xi, \xi_{\mathbf{X}} + \ell + 1, n) \le e^\epsilon \max_{\ell\in\mathbb{Z}_{\ge 0}} e^{-\epsilon\ell} g(\xi, \xi_{\mathbf{X}} + \ell + 1, n) = e^\epsilon S(G).
\end{aligned}$$

This shows that S is an ϵ-smooth upper bound on the local sensitivity of f. As for the upper bound, for any $\ell \ge 0$, we have

$$\begin{aligned}
e^{-\epsilon\ell} g(\xi, \xi_{\mathbf{X}} + \ell, n) &= e^{-\epsilon\ell}\frac{4k\left(\xi_{\mathbf{X}} + \ell\right)}{n}\left(2\xi + \frac{17k\left(\xi_{\mathbf{X}} + \ell\right)}{n-1}C\right) \\
&= \frac{4k\left(\xi_{\mathbf{X}} e^{-\epsilon\ell/2} + \ell e^{-\epsilon\ell/2}\right)}{n}\left(2\xi e^{-\epsilon\ell/2} + \frac{17k\left(\xi_{\mathbf{X}} e^{-\epsilon\ell/2} + \ell e^{-\epsilon\ell/2}\right)}{n-1}C\right) \\
&\le \frac{4k\left(\xi_{\mathbf{X}} + 1/\epsilon\right)}{n}\left(2\xi + \frac{17k\left(\xi_{\mathbf{X}} + 1/\epsilon\right)}{n-1}C\right).
\end{aligned}$$

■

Proof [Proof of Lemma 14] By the given sample complexity bound on n and using Lemma 7, we can obtain an ϵ-differentially private estimate $\tilde{\zeta}_k$ of the variance of ζ_k such that with probability at least $1-\alpha$, $\zeta_k \le \tilde{\zeta}_k \le 2\zeta_k$. Assume now that ζ_k is known; it will be apparent that using the privately estimated $\tilde{\zeta}_k$ instead of ζ_k in the proof does not affect the final error guarantee. Note also that conditioned on any family $\mathcal{F}_{\text{ss}}$ of subsets of $\mathcal{I}_{n,k}$, the run of Algorithm 1 is ϵ-differentially private. Since the randomness of $\mathcal{F}_{\text{ss}}$ is independent of the data, the algorithm (along with the private variance estimation) is still 2ϵ-differentially private.

Let $Q_\alpha = \sqrt{2K\zeta_k k\log\left(\frac{4n}{\alpha}\right)}$. Then, for any $S \in \mathcal{I}_{n,k}$ the probability that $|h(X_S) - \theta| \le Q_\alpha$ is at most $2\left(\frac{\alpha}{4n}\right)^k \le \frac{\alpha}{2n^k}$. By a union bound over all $\binom{n}{k}$ sets S, $|h(X_S) - \theta| \le Q_\alpha$ for all $S \in \mathcal{I}_{n,k}$ with probability at least $1 - \frac{\alpha}{2}$. Call this event $\mathcal{E}$; conditioned on this event, $\tilde{\theta}_{\text{ss}} = \hat{\theta}_{\text{ss}}$. As for Q_α^{avg}:

$$\mathbb{P}\left(|\hat{\theta}_{\text{ss}} - \theta| \ge t | \mathcal{G}_\alpha\right) \le \frac{P\left(|\hat{\theta}_{\text{ss}} - \theta| \ge t\right)}{P(\mathcal{G}_\alpha)}$$

Now, $\mathbb{E}[\hat{\theta}_{\text{ss}}|X_1, \dots X_n] = U_n$. Moreover, for any $t > 0$,

$$\begin{aligned}\mathbb{P}\left(|\hat{\theta}_{\text{ss}} - \theta| \geq t\right) &\leq \mathbb{P}\left(|\hat{\theta}_{\text{ss}} - U_n| \geq t/2\right) + P\left(|U_n - \theta| \geq t/2\right) \\ &\leq \mathbb{E}_{X_1,\dots,X_n}\mathbb{P}\left(|\hat{\theta}_{\text{ss}} - U_n| \geq t/2|X_1, \dots, X_n\right) + 2\exp\left(-\frac{nt^2}{8kK\zeta_k^2}\right) \quad \text{(A.26)}\end{aligned}$$

For the first term in Eq A.26, note that conditioned on the data $X_1, \dots, X_n$, the Y_j are independent draws from the uniform distribution over the $\binom{n}{k}$ values $\{h(X_S)\}_{S\in\mathcal{I}_{n,k}}$, have mean U_n, and the $|Y_j - \theta|$ are bounded by $\max_{S\in\mathcal{I}_{n,k}} |h(X_S) - \theta| \leq Q_\alpha$. Therefore, $Y_i - U_n$ is sub-Gaussian(Q_α^2). Therefore,

$$\begin{aligned}\mathbb{E}\mathbb{P}\left(|\hat{\theta}_{\text{ss}} - U_n| \geq t/2|X_1, \dots, X_n\right) &\leq 2\mathbb{E}\left[\exp\left(-\frac{Mt^2}{8Q_\alpha^2}\right)|\mathcal{E}\right] + P(\mathcal{E}^c) \\ &\leq 2\exp\left(-\frac{Mt^2}{16K\zeta_k k\log(4n/\alpha)}\right) + \frac{\alpha}{2} \quad \text{(A.27)}\end{aligned}$$

Combining Eqs A.26 and A.27, we get

$$\mathbb{P}\left(|\hat{\theta}_{\text{ss}} - \theta| \geq t\right) \leq 2\exp\left(-\frac{Mt^2}{16K\zeta_k k\log(2n/\alpha)}\right) + \alpha/2 + 2\exp\left(-\frac{nt^2}{8kK\zeta_k^2}\right) \leq \alpha,$$

for

$$Q_\alpha^{\text{avg}} = 4\sqrt{\frac{K\zeta_k k}{\min(M,n)}}\log\frac{8n}{\alpha}.$$

Thus, from Lemmas 6, A.3, and Eq A.26, with probability at least $1 - O(\alpha)$,

$$|\tilde{\theta}_{\text{ss}} - \theta| \leq \tilde{O}\left(\sqrt{\text{Var}(U_n)} + \sqrt{\frac{\zeta_k}{M}} + \sqrt{\frac{K\zeta_k k^3}{n^2\epsilon^2}}\right).$$

■

A.3.1. Proof of Lemma 16

Proof [Proof of Lemma 16] First, conditioned on X_i, the projection $\hat{h}_1(X_i)$ can be viewed as a U statistic on the other $n-1$ data. By Bernstein's inequality for U statistics (see Eq A.22), for all $t > 0$, we have

$$\mathbb{P}\left(\left|\hat{h}_1(X_i) - \mathbb{E}\left[h(X_S)|X_i\right]\right| \geq t\right) \leq 2\exp\left(\frac{-\left\lfloor\frac{n-1}{k-1}\right\rfloor t^2}{2\sigma_i^2 + 2Ct/3}\right). \quad \text{(A.28)}$$

Setting $t = \sigma_i\sqrt{\frac{4k}{n}}\sqrt{\log\frac{2n}{\beta}} + \frac{4Ck}{3n}\log\frac{2n}{\beta}$ in Eq A.28,

$$\mathbb{P}\left(\left|\hat{h}_1(X_i) - \theta\right| \geq t\right) \leq \exp\left(\frac{-nt^2/k}{2\sigma_i^2 + 2Ct/3}\right) \leq \exp\left(\min\left\{\frac{-nt^2}{4k\sigma_i^2}, \frac{-3nt}{4kC}\right\}\right) \leq \frac{\beta}{2n}. \quad \text{(A.29)}$$

Applying a union bound on the events in Eq A.29 over all $i \in [n]$ yields the result. ■

Proof of Lemma 6 **Proof** The output $\tilde{\theta}_n^{\text{med}}$ is ϵ-differentially private by Parallel Composition (see Lemma A.6). For each $i \in [d]$, let Z_i be a Bernoulli random variable which is 0 iff it satisfies Eq 6. By the guarantee on $\tilde{\theta}_{n,i}$, Z_i has mean $p \leq \frac{1}{4}$. Note that the median of these d estimates satisfies Eq 6 if more than $\frac{n}{2}$ of the estimates satisfy the equation, that is, if $\sum_{i\in[d]} Z_i < \frac{n}{2}$. Since increasing p can only decrease this probability, assume $p = \frac{1}{4}$. By a Chernoff bound,

$$\mathbb{P}\left(\sum_{i\in[d]} Z_i \geq \frac{n}{2}\right) = \mathbb{P}\left(\sum_{i\in[d]} Z_i \geq 2\mathbb{E}\sum_{i\in[d]} Z_i\right) \leq \exp\left(-\frac{1}{3}\cdot pd\right) = \alpha.$$

■

Proof of Lemma 11 **Proof** By the given sample complexity bound on n and using Lemma 7, we can obtain an ϵ-differentially private estimate $\tilde{\zeta}_k$ of the variance of ζ_k such that with probability at least $1-\alpha$,

$$\zeta_k \leq \tilde{\zeta}_k \leq 2\zeta_k$$

Assume now that ζ_k is known; it is easily seen that using the privately estimated $\tilde{\zeta}_k$ instead of ζ_k in the proof does not affect the rest of the argument and the error guarantee (up to constants).

For any $i \in [n]$, there are exactly $\binom{n-1}{k-1}$ sets $S \in \mathcal{I}_{n,k}$ such that $i \in S$. Following the notation from Eqs 4 and the definition of dep$_{,}$ (), $f_i = \binom{n-1}{k-1}/\binom{n}{k} = \frac{k}{n}$ for all $i \in [n]$, so $\text{dep}_{n,k}(\mathcal{F}_{\text{all}}) = \frac{k}{n}$. Moreover, for each $S \in \mathcal{I}_{n,k}$, $\mathbb{P}(|h(X_S) - \theta| \geq y) \leq 2\exp\left(\frac{-y^2}{2K\zeta_k}\right)$. Letting

$$Q_\delta := \sqrt{2K\zeta_k k \log\left(\frac{2n}{\delta}\right)} < \sqrt{2K\zeta_k \log\left(\frac{2}{\delta}\binom{n}{k}\right)},$$

we see that each Y_i is within Q_δ of θ with probability at least $\frac{\delta}{\binom{n}{k}}$. A union bound implies that this choice of Q_δ is valid. For the concentration of the average, $\frac{1}{m}\sum_{j\in[m]} Y_j$, which is simply the U statistic U_n, we will use the Hoeffding bound on U statistics (cf. Lemma A.2)

$$P\left(\left|\frac{1}{m}\sum_i Y_i - \theta\right| \geq y\right) \leq 2\exp\left(-\frac{\lfloor n/k\rfloor y^2}{2K\zeta_k}\right).$$

Thus, we can define

$$Q_\delta^{\text{avg}} := \sqrt{\frac{2K\zeta_k k \log\frac{2}{\delta}}{n}}.$$

To apply Proposition 4, we verify if the conditions in Proposition 4 hold.

$$\frac{k}{n} \leq \frac{Q_\alpha \epsilon}{10tQ_{\alpha/t}\log(t/\alpha)} = \frac{\epsilon}{10t\log(t/\alpha)}\sqrt{\frac{\log 2n/\alpha}{\log 2nt/\alpha}} \iff n \geq \frac{10kt\log(t/\alpha)}{\epsilon}\sqrt{\frac{\log 2nt/\alpha}{\log 2n/\alpha}},$$

and

$$Q_\beta^{\text{avg}} \leq Q_\alpha \iff \sqrt{\frac{2K\zeta_k k\log\frac{2t}{\alpha}}{n}} \leq \sqrt{2K\zeta_k k\log\left(\frac{n}{\alpha}\right)} \iff n \geq \frac{\log 2t/\alpha}{\log n/\alpha},$$

which are both true. Therefore, with probability at least $1 - O(\alpha)$, we have

$$|\tilde{\theta}_{\text{all}} - \theta| \leq O\left(\frac{1}{\sqrt{\alpha}}\sqrt{\text{Var}(U_n)} + \frac{k}{n\epsilon}\sqrt{2K\zeta_k k\log\left(\frac{2n}{\alpha}\right)}\right).$$

■

Proof [Proof of Lemma 18] By symmetry, it suffices to show $\xi_{\mathbf{X}'} \leq \xi_{\mathbf{X}} + 1$. So, if an index $i \neq i^*$ is in Good($\mathbf{X}$), then using Eqs 10, 12, and 13, we get

$$\begin{aligned}\left|\hat{h}_1^{\mathbf{X}'}(X_i') - U_n(\mathbf{X}')\right| &\leq \left|\hat{h}_1^{\mathbf{X}'}(X_i') - \hat{h}_1^{\mathbf{X}}(X_i)\right| + \left|\hat{h}_1^{\mathbf{X}}(X_i) - U_n(\mathbf{X})\right| + \left|U_n(\mathbf{X}) - U_n(\mathbf{X}')\right| \\ &\leq \xi + \frac{(2k-1)C(\xi_{\mathbf{X}}+1)}{n-1},\end{aligned}$$

which leaves at most $1+\xi_{\mathbf{X}}$ potential indices i for which $|\hat{h}_1^{\mathbf{X}'}(X_i') - U_n(\mathbf{X}')| > \xi + \frac{(2k-1)C(1+\xi_{\mathbf{X}})}{n-1}$: the bad indices in G and the index i^*. Therefore, $\xi_{\mathbf{X}'} \leq \xi_{\mathbf{X}} + 1$. ■

Appendix A.4. Composition Theorems for Differential Privacy

Lemma A.5 *(Basic Composition) Let $\mathcal{X}$ and $\mathcal{R}$ be non-empty sets. If $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_k : \mathcal{X}^n \to \mathcal{R}$ are each ϵ-differentially private algorithms, then the mechanism $\mathcal{A} : \mathcal{X}_n \to \mathcal{R}^k$ defined as*

$$\mathcal{A}(X_1, \ldots, X_n) = (\mathcal{A}_1(X_1, \ldots, X_n), \ldots, \mathcal{A}_k(X_1, \ldots, X_n))$$

is $k\epsilon$-differentially private.

Lemma A.6 *(Parallel Composition) Let $\mathcal{X}$ and $\mathcal{R}$ be non-empty sets. If $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_k : \mathcal{X}^n \to \mathcal{R}$ are each ϵ-differentially private algorithms, then the mechanism $\mathcal{A} : \mathcal{X}^{kn} \to \mathcal{R}^k$ defined as*

$$\mathcal{A}(X_1, \ldots, X_{kn}) = \left(\mathcal{A}_1(X_1, \ldots, X_n), \mathcal{A}_2(X_{n+1}, \ldots, X_{2n}), \ldots, \mathcal{A}_k\left(X_{(k-1)n+1}, \ldots, X_{kn}\right)\right)$$

is ϵ-differentially private.