Building Tools for Controlling Overfitting in Adaptive Data Analysis

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

In this paper, we take the first steps for providing practical tools that help in bounding overfitting in adaptive data analysis. We provide new upper bounds on the error of some well-known mechanisms for answering adaptively-selected linear queries. Our bounds are obtained via a careful combination of key ideas from several different lines of work; they improve significantly on any of the existing techniques in isolation. We also initiate an empirical study of how an analyst's query-selection strategy affects the performance of different mechanisms. Along the way, we demonstrate empirically that our upper bounds on error are tight in a range of settings.

1 Introduction

Modern scientific data analysis is messy and what analysts do in practice often violates the standard modeling assumptions of statistics and machine learning. In many settings, it is not clear *a priori* what are the most salient analyses and hypotheses to test on the data; the analyst performs a significant amount of data exploration to make sense of the data and to identify the interesting hypotheses. This data exploration is a fundamentally *adaptive* process, meaning that the questions asked at lated stages depend on the results of earlier ones. The exploration process creates a coupling between the dataset and the choice of analyses, violating basic statistical and machine learning assumptions. Standard tools for controlling overfitting, such as false discovery rate (FDR) control and crossvalidation, assume that all the hypotheses to be tested, and the procedure for testing them, are chosen independently of the (validation) dataset.

Adaptivity arises even more starkly when multiple research groups make use of a single data set (as is common in fields where data are expensive to collect), or when a hidden data set is used to evaluate many sequentially selected submissions to a machine learning contest Blum and Hardt [2015], Hardt [2017]. The difficulty of analyzing an adaptively selected workflow (a "garden of forking paths") has been blamed as a central cause ongoing statistical issues the sciences Gelman and Loken [2014].

Adaptivity's prevalence drives the need for methods that account for the dependencies among stages of an analysis; this is particulary difficult when the analyst's decision process is not known. The statistics community began to address issues arising from adaptivity long ago (e.g., [Buehler and Feddersen, 1963, Freedman, 1983]). This area, often called *selective* (or *post-selection*) *inference* has seen a surge of recent interest—see Bi et al. [2017] for an overview. It produces inference algorithms that are specific to a particular workflow, since it involves explicit conditioning on the past outcomes. Unfortunately, the conditional

Submitted to 32nd Conference on Neural Information Processing Systems (NIPS 2018). Do not distribute.

distribution required in this approach may well be computationally intractable, especially when may previous analyses must be considered.

A different line of work in the computer science community develops techniques for bounding the error of adaptively chosen statistical estimators by limiting the types of algorithms ("mechanisms") that can be used to produce estimates [Dwork et al., 2015c,a,b, Hardt and Ullman, 2014, Blum and Hardt, 2015, Russo and Zou, 2016, Cummings et al., 2016, Rogers et al., 2016, Hardt, 2017, Steinke and Ullman, 2015, Bassily et al., 2016, Feldman and Steinke, 2017, Fish et al., 2017]. These works fall into two broad, intertwined categories: those that bound the additional error introduced by selection using *distributional stability* or *privacy* properties of the mechanisms, and those that bound the error using upper bounds on different measures of *information* leaked about the data by earlier analyses. This line of work has the advantage of being directly prescriptive, providing tools to choose how to answer queries at early stages in order to maximize the accuracy of queries at later stages.

This paper takes the first steps towards practical tools for adaptive data analysis based on these ideas. The paper has two thrusts. First, we develop new upper bounds on the error of specific algorithms for adaptive data analysis. These bounds allow us to provide concrete confidence intervals along with query answers, and improve the sample size required for given error goals by orders of magnitude. Second, we initiate an empirical study of adaptive data analysis. We design and implement several adaptive query-selection strategies—workloads, in effect—in order to evaluate the tightness of our upper bounds on error, and in order to understand the role that the workload structure plays in the error of different query-answering mechanisms.

Our broader goal is to develop tools and techniques to manage the tradeoff between allowing unfettered access to a data for exploratory analysis, on one hand, and providing a long-lived resource for statistically valid answers.

We focus our attention on the setting of a data analyst who asks a sequence of *linear* queries (also called *statistical queries* [Kearns, 1998]). We assume there is a data set $\boldsymbol{X}$ of size n drawn i.i.d. from an underlying distribution $\mathcal{D}$ (the population being studied) over a base set $\mathcal{X}$. The analyst specifies a function $\phi : \mathcal{X} \to [0, 1]$, and wishes to learn the expectation of ϕ in the population, denoted $\phi(\mathcal{D}) \stackrel{\text{def}}{=} \mathbb{E}_{X\sim\mathcal{D}}[\phi(X)]$. The difficulty in answering such queries lies in overfitting: when queries are asked adaptively, the population average $\phi(\mathcal{D})$ may very different from empirical average $\phi(\boldsymbol{X}) = \frac{1}{n}\sum_i \phi(X_i)$. Our goal is to find mechanisms that, given $\boldsymbol{X}$ and an adaptively chosen sequence of queries $\phi_1, ...\phi_k$, return answers $a_1, ..., a_k$ such that $a_j \approx \phi_j(\mathcal{D})$.

Linear queries capture a wide range of analyses, from the very basic—frequencies, histograms and other basic counting tasks—to sophisticated algorithms—such as first- and second-order optimization. Many quantities used for validation (misclassification rates, goodness-of-fit tests) can be written in terms of linear queries. Linear queries are thus interesting in their own right, and important stepping stones towards more general types of queries. They are sufficiently structured, however, to provide a clean model for analysis. The line of computer science work cited above establishes asymptotic upper and lower bounds on the sample size required to answer a given number k of queries at a given level of accuracy, showing that $n \approx \sqrt{k}$ samples are necessary and sufficient. This paper initiates an empirically-grounded study of these questions, aiming for both worst-case results andx workload-aware methods.

1.1 Contributions

Confidence intervals via new upper bounds We provide new, tighter upper bounds on the error of certain types of algorithms for answering adaptive linear queries. We obtain these bounds by abstracting out key ideas from the literature, and showing how they can be plugged together and optimized. We end up with a collection of mutually incomparable bounds which can be numerically optimized and selected from for any given setting of parameters. For all the settings we study empirically, we are able to obtain the best bounds by combining tools from the stability literature [Bassily et al., 2016, Bun and Steinke, 2016] (KL-stability and one of two "monitor" arguments) and information bounds [Russo and Zou, 2016].

Reporting values together with confidence intervals is an important component of any analysis tool. It is also one that, in the adaptive setting, cannot obviously be done empirically—e.g. via resampling, crossvalidation, or explicit conditioning—in anything resembling polynomial time, since one must be able to perform resampling or conditioning consistently with previous answers. Our methods are simple and efficient, and illustrated with working code.

Our combined bounds improve significantly on any of the existing techniques in isolation. In many cases, the new bounds improve on the required sample complexity by orders of magnitude. As an extreme example, suppose one wants to be able to answer a number k of queries equal to the sample size n while providing answers within 30% of the population value. The theory states this is possible, via a mechanism that reports noisy empirical means, for sufficiently large n. However, previous bounds required n almost 10 million, whereas our techniques give such bounds with n less than $100,000$.

The new bounds also vastly improve on the bounds one could get via naïve approaches such as splitting the sample into k separate batches of size n/k (to answer k queries). (In particular, when $n = k$, sample splitting provides no nontrivial accuracy, regardless of n.)

Workload / analyst strategies We implement specific query-selection strategies to test the tightness of our upper bounds, and to help understand the role of the workload's structure on different mechanisms' accuracy. The two strategies we consider are drawn from the impossibility results in the theory literature, operationalizing those results as actual workflows.

The first strategy is a simple, two-round strategy in which the analyst asks for marginal distributions of the features in a dataset and then asks a single "hard query" based on the most significant features; this strategy comes from the earliest statistical work on adaptivity Freedman [1983], Pötscher [1991], Leeb and Pötscher [2005] and was also considered in Dwork et al. [2015b], Hardt [2017]. The second strategy is a "tracing" attack, adapted from the fingerprinting lower bounds of Bun et al. [2014], Hardt and Ullman [2014], Steinke and Ullman [2015] and simplified. These two strategies are simple and implemented in essentially linear time.

For the simplest mechanism, Gaussian noise addition, the first strategy achieves root mean squared error that matches, within small constants, the error bounds and confidence widths provided by our methods, showing that they are close to tight. In other settings, the bounds are loose, highlighting the possibility of providing for tighter, workload- and mechanism-aware intervals.

The simplicity of these mechanisms highlights the difficulty—and importance, going forward—of modeling "benign" analyst behavior and workloads.

Initiating a broader empirical study In addition to measuring the tightness of our error estimates, we initiate a broader empirical investigation of adaptive data analysis. We consider two query-selection strategies and four nontrivial query-answering mechanisms (in addition to a number of naïve mechanisms as benchmarks).

Our results highlight that we do not yet have a one-size-fits-all mechanisms that adapts to different workloads automatically: the order of the error of the mechanisms we consider is inverted for the two query-selection mechanisms—those that do better on one do worse on the other. The simple Gaussian mechanism does poorly with the single " hard query" strategy; in contrast, it provides the longest-lived nontrivial answers for queries selected by the tracing strategy. The Thresholdout mechanism (due to Dwork et al. [2015a]) does does nearly optimally against the "one hard query", but fares poorly for tracing strategy's queries.

The two strategies we consider vary greatly in the extent to which they use adaptivity: the first makes only one query that depends on previous oututs, while the second incorporates information continuously. We conjecture that it is possible to come up with a single strategy that incorporates the "worst" (i.e. hardest to answer accurately) of both strategies, and pushes all three known mechanisms to the limits of their effectiveness.

Our results also highlight the challenges of parameter-tuning for the mechanisms we consider. All the mechanisms involve a noise magnitude parameter (typically denoted σ, though it is not always exactly the standard deviation), and some also involve thresholds and train/test split sizes. These parameters all play a role in the mechanism's error. To tune the parameters, one may choose to minimize either the width of the confidence intervals reported by the mechanism, or the error achived on a particular workload. Our results demonstrate that these two criteria lead to significantly different parameter settings, even when the achieved error bounds are within small factors of the reported widths. The message of the results is mixed: on one hand, the errors (both upper and lower) suggest that there is some leeway in parameter choice for the right query-selection strategy; on the other hand, they point out that optimizing for the wrong workload can lead to significant problems for current mechanisms.

1.2 Preliminaries

Throughout, we assume that the data $\boldsymbol{X} = (X_1, \cdots, X_n) \sim \mathcal{D}^n$ comes from a product distribution over $\mathcal{X}$. We will denote the true average of statistical query ϕ as $\phi(\mathcal{D}) = \mathbb{E}_{x\sim\mathcal{D}}[\phi(x)]$ and the empirical average as $\phi(\boldsymbol{X}) = \frac{1}{n}\sum_{i=1}^{n}\phi(X_i)$. Also, if X_i is a k-element vector, we denote the jth element of X_i by $X_i(j)$, for $j \in [k]$. We defer additional preliminaries related to confidence intervals, stability measures, and existing work, to Appendix A.

2 Confidence Intervals for Specific Mechanisms

In this section, we show how we can obtain valid confidence intervals when mechanisms like Gaussian noise addition, or Thresholdout Dwork et al. [2015a], are used to answer a sequence of adaptively chosen statistical queries. We will first see how one can obtain tighter confidence bounds than previous works for the Gaussian mechanism by carefully combining existing techniques. Along the way, we also provide bounds for the mean squared error (MSE) for the Gaussian mechanism. Using similar ideas but a more involved analysis, we also provide confidence bounds for the Thresholdout technique.

Confidence Bounds for the Gaussian Mechanism We now show how we can use results from Bun and Steinke [2016] for bounding the mutual information of an algorithm when Gaussian noise is added to each query answer. Similar to techniques in Russo and Zou [2016], we will bound the bias between the empirical average of a statistical query and its true average when the query is chosen adaptively. We then use Chebyshev's inequality and the monitor argument from Bassily et al. [2016] (Algorithm 1) to obtain a high probability accuracy bound. Our accuracy guarantee can be stated as follows.

Theorem 2.1. *Given confidence level $1-\beta$ and using the Gaussian mechanism for each algorithm $\mathcal{M}_i$ for $i \in [k]$, then $(\mathcal{M}_1, \cdots, \mathcal{M}_k)$ is $(\tau^\star, \beta)$-accurate. We define $\tau^\star$ to be the solution to the following program: minimize τ such that $\tau \geq \sqrt{\frac{2}{n\beta} \cdot \min_{\lambda\in[0,1)} \left(\frac{2\rho' kn - \ln(1-\lambda)}{\lambda}\right)}$, and $\tau \geq \frac{2}{n}\sqrt{\frac{\ln(4k/\beta)}{\rho'}}$, for $\rho' kn \geq 0$.*

See Appendix C.1 for a proof of Theorem 2.1.

Comparison with Prior Work One can also get a high-probability bound on the sample accuracy of $\mathcal{M}(\boldsymbol{X})$ using Theorem 3 in Xu and Raginsky [2017], resulting in $\tau \geq \sqrt{\frac{8}{n}\left(\frac{2\rho' kn}{\beta} + \log\left(\frac{4}{\beta}\right)\right)}$ instead of the first inequality in Theorem 2.1; the proof is similar to the proof of Theorem 2.1. If the mutual information bound $B = \rho' kn \geq 1$, then it is easy to see that the current Theorem 2.1 results in a tighter bound than t via Xu and Raginsky [2017] for any $\beta \in (0, 1)$. For very small B, there exist small β for which the result obtained via Xu and Raginsky [2017] is better.

In Figure 1, we show the widths of the valid confidence intervals for k adaptively selected statistical queries where each answer has Gaussian noise added to it. We fix $n/k \in \{10, 100\}$, and obtain the 0.95 confidence interval for various values of adaptive queries k. The label "DFH-PRR" plots the bounds derived from Theorem B.1, "BNSSSU" plots the bounds from Theorem B.4, and "CDP+Monitor+RZ" plots the bound from Theorem 2.1. "CDP+Monitor+XR"

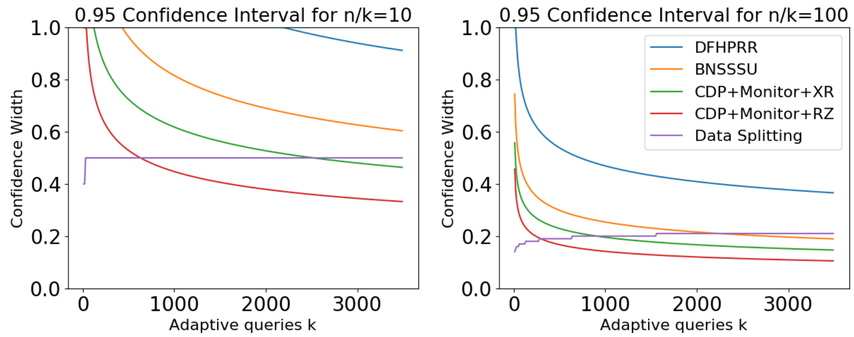

Figure 1: Widths of confidence intervals for k adaptively chosen statistical queries via data-splitting or Gaussian noise addition on the same dataset.

plots the bound derived via the results of Xu and Raginsky [2017] combined with the proof technique of Theorem 2.1. The traditional approach of splitting the data and running each analysis on each chunk is exhibited in the plot called “Data Splitting", where we are applying a Chernoff bound on each n/k chunk of data and applying a union bound over all k chunks.

From Figure 1, we can see that the best confidence bounds are obtained via “RZ+Monitor" (Theorem 2.1), and they provide an improvement over datasplitting when $k \approx 500$ and $n \approx 5000$. We notice this improvement for smaller k as n increases. Note that we get meaningful bounds (confidence width ≤ 0.5) for $k > 500$ via “RZ+Monitor" even when $n/k = 10$, which is not the case with the other techniques. A strong reason for Figure 1 providing the tightest bounds is that it is obtained via a careful combination of existing strategies, thus performing better than its components.

Bounds for Mean Squared Error We present here a bound on the mean squared error (MSE) of answering adaptively chosen statistical queries $\mathcal{Q}_{SQ}$ by adding Gaussian noise to the empirical answers. We consider an analyst $\mathcal{A}$ that selects a query $\phi_1 : \mathcal{X} \to [0,1]$ for which $\mathcal{M}_1$ will report answer $a_1 = \phi_1(\boldsymbol{X}) + N(0, \frac{1}{2n^2\rho})$, where $\boldsymbol{X} \sim \mathcal{D}^n$ is the sampled dataset. Then at future rounds, the analyst selects query ϕ_i based on the queries already asked and the received answers.

We then want to bound the MSE of the worst statistical query, where the expectation is over the entire sequence of algorithms $\mathcal{M} = (\mathcal{M}_1, \cdots, \mathcal{M}_k)$ and the adversary $\mathcal{A}$.

$$\begin{aligned}\mathop{\mathbb{E}}_{\substack{\boldsymbol{X}\sim\mathcal{D}^n,\\ \mathcal{A},\mathcal{M}}}\left[\max_{i\in[k]}\left(\phi_i(\mathcal{D}) - a_i\right)^2\right] &\leq 2\cdot \mathop{\mathbb{E}}_{\substack{\boldsymbol{X}\sim\mathcal{D}^n,\\ \mathcal{A},\mathcal{M}}}\left[\max_{i\in[k]}\left\{\left(\phi_i(\mathcal{D}) - \phi_i(\boldsymbol{X})\right)^2 + \left(\phi_i(\boldsymbol{X}) - a_i\right)^2\right\}\right] \\ &= 2\cdot\mathbb{E}\left[\max_{i\in[k]}\left(\phi_i(\mathcal{D}) - \phi_i(\boldsymbol{X})\right)^2\right] + 2\cdot \mathop{\mathbb{E}}_{Z_i\sim N\left(0,\frac{1}{2n^2\rho}\right)}\left[\max_{i\in[k]} Z_i^2\right] \qquad (1)\end{aligned}$$

To bound $\mathbb{E}\left[\max_{i\in[k]}\left(\phi_i(\mathcal{D}) - \phi_i(\boldsymbol{X})\right)^2\right]$, we obtain the following using the monitor from Bassily et al. [2016] along with results from Russo and Zou [2016], Bun and Steinke [2016].

Theorem 2.2. *Using the Gaussian mechanism for each algorithm $\mathcal{M}_i$ with reported answers $a_1, \cdots, a_k$, we have for $\rho > 0$,*

$$\mathop{\mathbb{E}}_{\boldsymbol{X}\sim\mathcal{D}^n,\mathcal{M},\mathcal{A}}\left[\max_{i\in[k]}\left(\phi_i(\mathcal{D}) - a_i\right)^2\right] \leq \frac{1}{2n}\cdot\min_{\lambda\in[0,1)}\left(\frac{2\rho kn - \ln(1-\lambda)}{\lambda}\right) + 2\cdot\mathop{\mathbb{E}}_{Z_i\sim N\left(0,\frac{1}{2n^2\rho}\right)}\left[\max_{i\in[k]} Z_i^2\right]$$

See Appendix C.2 for a proof of Theorem 2.2.

In Figure 2, we show the highest value of k that guarantees an RMSE at most 0.1 when each of the k statistical queries is adaptively chosen, and when only one query is chosen adaptively. The label “CDP+MonitorRZ+" shows the results obtained via Theorem 2.2

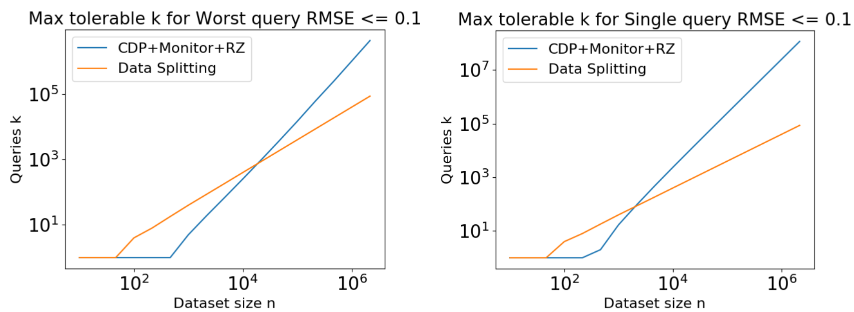

Figure 2: Maximum tolerable adaptive queries k, w.r.t. n, for RMSE ≤ 0.1 over all the queries (left), and for any single query (right) via data-splitting or Gaussian noise addition.

(left), and the corresponding result for any single query (right). "Data Splitting" shows the results obtained via data-splitting. From both the plots, we can see that at large but realistic dataset sizes, Gaussian noise addition provides better guarantees than data-splitting. As evident from the plot on the right of Figure 2, Gaussian noise addition also provides such guarantees when k exceeds n, which is not possible with data-splitting.

3 The Interplay between Mechanisms and Analyst Strategies

In this section, we empirically investigate how the performance of mechanisms is affected under different query workloads and different levels of adaptivity. We consider some well-known mechanisms, and also some variants from them (we provide detailed descriptions for them in Appendix D.1). We first observe the performance of the mechanisms under a two-round analyst strategy, where the analyst is asks multiple non-adaptive queries in the first round, and then asks an adaptive query in the second round. With such a strategy in hand, we also show that this strategy almost closes the gap between the performance achieved via the Gaussian mechanism and its upper bound (Theorem 2.2). Next, we introduce a strategy with multiple rounds of adaptivity, and observe that the performance of the mechanisms is very different than on the two-round strategy.

3.1 A two-round analyst strategy

Now, we consider a strategy for obtaining the maximum RMSE for an adaptively chosen statistical query when each sample in the dataset is drawn u.a.r. from $\{-1, 1\}^{k+1}$, and the adaptive query is asked after the knowledge of the empirical correlations of each of the first k features with the $(k+1)^{th}$ feature. We provide a pseudocode of the strategy in Algorithm 4.

Theorem 3.1. *The output by the two-round analyst strategy above results in the maximum possible RMSE for an adaptively chosen statistical query when each sample in the dataset is drawn uniformly at random from $\{-1, 1\}^{k+1}$, and $\mathcal{M}$ is the Naive Empirical Estimator, i.e., $\mathcal{M}$ provides the empirical correlation of each of the first k features with the $(k+1)^{th}$ feature.*

See Appendix C for a proof of Theorem 3.1.

Next, we will see how the two-round analyst strategy compares with the upper bound for the Gaussian mechanism (Theorem 2.2) for one adaptive query. First, we explore the effect of the noise scale σ in the Gaussian mechanism (shown on the left in Figure 3). We fix the number of samples $n = 5000$ and the number of non-adaptive queries $k = 500$, and then we see how setting σ affects the RMSE of the last adaptive query. We can observe from this experiment that tuning for σ is non-trivial; the value of σ for which the upper bound obtained via Theorem 2.2 guarantees the least possible RMSE is different than the value of σ for which the two-round analyst strategy achieves the least possible RMSE. On the right in Figure 3, we set $n = 5000$ and see how close is the two-round strategy to the upper bound. We vary k from 1 to 50000, and for each value, we set σ for the Gaussian mechanism to answer the two-round strategy as suggested by the minimization in Theorem 2.2. For all

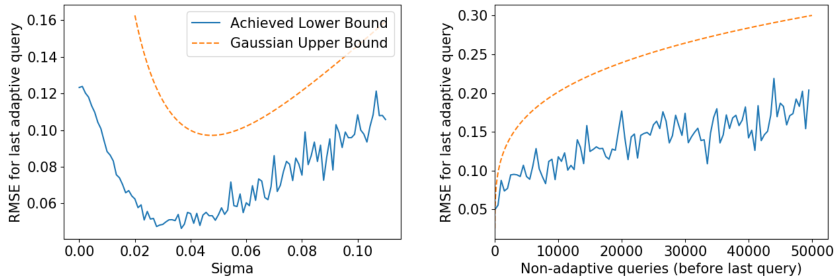

Figure 3: Dependence of RMSE for Gaussian mechanism on noise scale σ for $n = 5000, k = 500$ (left), and on k for $n = 5000$ and σ set as suggested by Theorem 2.2 (right).

values of k, we observe that the RMSE achieved by the two-round strategy is within a factor of 2.5 of the upper bound. This provides evidence that the upper bound for the Gaussian mechanism provided by Theorem 2.2 is essentially tight as the two-round strategy effectively provides a lower bound which is close to the upper bound over a wide range of k. The RMSE is averaged over 100 independent runs in the plots for the two-round strategy in Figure 3.

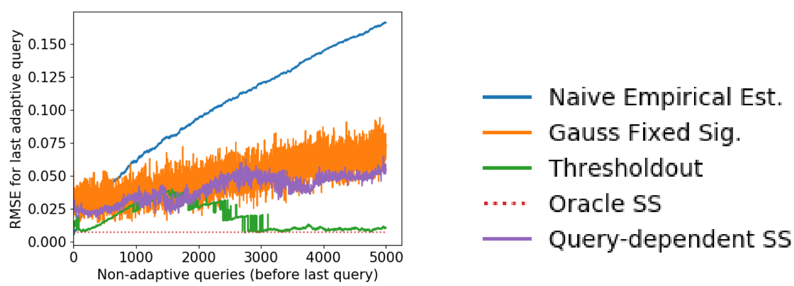

Figure 4: The performance of different mechanisms for the two-round strategy for $n = 5000$.

Figure 4 shows the performance of some of the mechanisms for the two-round strategy for $n = 5000$ samples. The plot labelled “Gaussian Fixed Sig.” is for the Gaussian mechanism having input a fixed $\sigma = 0.03$, Query-dependent Sample Splitting (SS) has input parameters $B = n/3, H = n/3, T = 0.1, \sigma = 0.001$, and Thresholdout has inputs $h = n/2, T = 0.05, \sigma = 0.0001$. Oracle SS represents a variant of Sample Splitting having an oracle that it can use to determine when is it best to switch to the next batch. For this strategy, the optimal point to switch for the mechanism is the last query as it is the only adaptive query. Thus, we present the result for “Oracle SS” with input batch size $B = n/2$. We tune the input parameters for all the mechanisms, and the results presented are averaged over 30 independent runs.

We can observe from Figure 4 that although all the mechanisms improve upon the Naive Empirical Estimator, there is a wide difference in performance while comparing among the mechanisms. The gaussian mechanism, even with a tuned value of σ, provides the largest RMSE. However, Thresholdout does perform extremely well for this strategy; it provides RMSE almost as low as that by the Oracle SS mechanism once k is greater than ≈ 3000. It is important to note that Oracle SS achieves the lowest possible RMSE due to the knowledge of when best to answer from a different batch.

3.2 A multi-round analyst strategy

Since the analyst strategy described in section 3.1 had effectively only one round of adaptivity, a mechanism like Oracle Sample Splitting could succesfully answer queries with an extremely low RMSE. In other words, a simple mechanism like sample splitting can be extremely effective if it has knowledge of the analyst strategy. However, there are two potential downsides to this. First, it can be unrealistic to expect a mechanism to be designed with the knowledge of all the possible analyst strategies it might be run against. Second, mechanisms that limit the analyst to only one round of adaptivity would constrain the reuse of data, and thus, could be very expensive. As we will see in this section, it can be non-trivial to design well-performing mechanisms when the analyst can have multiple rounds of adaptivity, and more importantly, a mechanism that is successful against a two-round strategy can perform very poorly against a multi-round strategy (and vice-versa).

We provide an analyst strategy in which the analyst tries to "trace" the hidden dataset of a mechanism, and consequently, force the mechanism to give query answers that do not generalize well to the underlying population. The hidden dataset D_0 of size n given to the mechanism $\mathcal{M}$ is sampled i.i.d and u.a.r. from a universe $\{0, 1, \ldots, N-1\}$. The analyst adaptively poses statistical queries q_j, for $j \in [k]$ to $\mathcal{M}$, and $\mathcal{M}$'s goal is to provide answers a_j that are close to the population mean $\mathbb{E}(q_j) = \frac{1}{N}\sum_{i=0}^{N-1} q_j(i)$. However, for each query q_j, the mechanism is only allowed to access $\{q_j(i)\}_{i \in D_j}$, i.e., evaluations of q_j on the currently "untraced" sample points $D_j \subseteq D_0$. We provide a pseudocode of the strategy in Algorithm 5.

Figure 5 shows the performance of various mechanisms for this strategy for $n = 5000$ samples, $N = 100n, c = 100$, and $k = 4 \times 10^5$. To reduce the variance in performance, we set the bias $p_j = 0.5$ in every round. The Gaussian mechanism shown has input $\sigma = 0.035$. Query-dependent Sample Splitting has inputs $B = n/10, H = n/10, T = 0.25, \sigma = 0.03$, and Sample Splitting has inputs $B = n/8, M = kB/n$. Both switch batches when their respective switching conditions are triggered, or all samples in the current batch have been traced. Both Thresholdout and Noisy Thresholdout have inputs $h = n/4, T = 0.05, \sigma = 0.03$. All the mechanisms start to provide random answers if all the samples in their input dataset have been traced. We tune the input parameters for all the mechanisms.

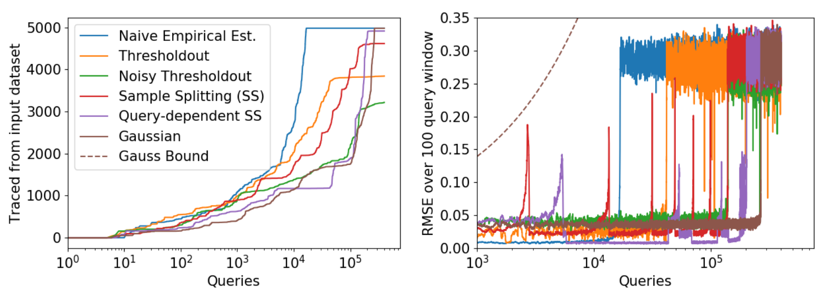

Figure 5: The performance of different mechanisms for the multi-round tracing strategy.

We can observe from Figure 5 that for this strategy, the Gaussian mechanism provides the best performance among all the mechanisms. Even after 10^5 queries, we see on the left that less than 2000 samples have been traced in the run with the Gaussian mechanism, and as a result, it provides a low and stable average RMSE until almost 3×10^5 queries. We also see that Noisy Thresholdout is able to provide answers with low average RMSE for a magnitude larger than Thresholdout, indicating that it can be highly beneficial to add noise to the answers via the training set in Thresholdout.[1] Query-dependent Sample Splitting provides the second-best performance for both the analyst strategies that we consider; we consider providing analytical bounds for it would be informative, and leave it for future work. We also plot the upper bound (Theorem 2.2) for the Gaussian mechanism with $\sigma = 0.035$, same as that used in the Gaussian mechanism for the strategy. Although we can see from the fluctuating behavior of both the sample splitting variants that this strategy does utilize multiple rounds of adaptivity, we can observe from the upper bound that the Gaussian mechanism is very far from it for all values of k.

[1]Note: One thing we are doing, but had not completed at submission time, is to provide a comparison of the achieved error for Thresholdout with our upper bound (Theorem B.9), and a comparison of our upper bound with those that follow from previous work.

References

Raef Bassily, Kobbi Nissim, Adam Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1046–1059. ACM, 2016.

Nan Bi, Jelena Markovic, Lucy Xia, and Jonathan Taylor. Inferactive data analysis. *arXiv:1707.06692 [math.ST]*, 2017.

Avrim Blum and Moritz Hardt. The ladder: A reliable leaderboard for machine learning competitions. *arXiv preprint arXiv:1502.04585*, 2015.

Robert J Buehler and Alan P Feddersen. Note on a conditional property of student's t. *The Annals of Mathematical Statistics*, 34(3):1098–1100, 1963.

Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In Martin Hirt and Adam Smith, editors, *Theory of Cryptography*, pages 635–658, Berlin, Heidelberg, 2016. Springer Berlin Heidelberg. ISBN 978-3-662-53641-4.

Mark Bun, Jonathan Ullman, and Salil P. Vadhan. Fingerprinting codes and the price of approximate differential privacy. In *STOC*, pages 1–10. ACM, May 31 – June 3 2014.

Rachel Cummings, Katrina Ligett, Kobbi Nissim, Aaron Roth, and Zhiwei Steven Wu. Adaptive learning with robust generalization guarantees. In *29th Annual Conference on Learning Theory*, pages 772–814, 2016.

Cynthia Dwork and Guy N. Rothblum. Concentrated differential privacy. *CoRR*, abs/1603.01887, 2016.

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Advances in Cryptology - EUROCRYPT 2006, 25th Annual International Conference on the Theory and Applications of Cryptographic Techniques, St. Petersburg, Russia, May 28 - June 1, 2006, Proceedings*, pages 486–503, 2006a. doi: 10.1007/11761679_29.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, pages 265–284. Springer, 2006b.

Cynthia Dwork, Guy N. Rothblum, and Salil P. Vadhan. Boosting and differential privacy. In *51th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2010, October 23-26, 2010, Las Vegas, Nevada, USA*, pages 51–60, 2010. doi: 10.1109/FOCS.2010.12.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toni Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. In *Advances in Neural Information Processing Systems*, pages 2350–2358, 2015a.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349 (6248):636–638, 2015b.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Leon Roth. Preserving statistical validity in adaptive data analysis. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, STOC '15, pages 117–126, New York, NY, USA, 2015c. ACM. ISBN 978-1-4503-3536-2. doi: 10.1145/2746539.2746580.

Vitaly Feldman and Thomas Steinke. Generalization for adaptively-chosen estimators via stable median. In *Proceedings of the 30th Conference on Learning Theory, COLT 2017, Amsterdam, The Netherlands, 7-10 July 2017*, pages 728–757, 2017. URL http://proceedings.mlr.press/v65/feldman17a.html.

Benjamin Fish, Lev Reyzin, and Benjamin I. P. Rubinstein. Sublinear-time adaptive data analysis. *CoRR*, abs/1709.09778, 2017. URL http://arxiv.org/abs/1709.09778.

David A Freedman. A note on screening regression equations. *The American Statistician*, 37 (2):152–155, 1983.

Andrew Gelman and Eric Loken. The statistical crisis in science. *American Scientist*, 102(6): 460, 2014.

Robert M. Gray. *Entropy and Information Theory*. Springer-Verlag, Berlin, Heidelberg, 1990. ISBN 0-387-97371-0.

Moritz Hardt. Climbing a shaky ladder: Better adaptive risk estimation. *CoRR*, abs/1706.02733, 2017. URL http://arxiv.org/abs/1706.02733.

Moritz Hardt and Jonathan Ullman. Preventing false discovery in interactive data analysis is hard. In *Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on*, pages 454–463. IEEE, 2014.

Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. *IEEE Trans. Information Theory*, 63(6):4037–4049, 2017.

S.P. Kasiviswanathan and A. Smith. On the 'Semantics' of Differential Privacy: A Bayesian Formulation. *Journal of Privacy and Confidentiality*, Vol. 6: Iss. 1, Article 1, 2014.

Michael Kearns. Efficient noise-tolerant learning from statistical queries. *J. ACM*, 45(6): 983–1006, November 1998. ISSN 0004-5411. doi: 10.1145/293347.293351.

Hannes Leeb and Benedikt M Pötscher. Model selection and inference: Facts and fiction. *Econometric Theory*, 21(01):21–59, 2005.

Benedikt M Pötscher. Effects of model selection on inference. *Econometric Theory*, pages 163–185, 1991.

Ryan Rogers, Aaron Roth, Adam Smith, and Om Thakkar. Max-information, differential privacy, and post-selection hypothesis testing. In *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*, 2016.

D. Russo and J. Zou. How much does your data exploration overfit? Controlling bias via information usage. *ArXiv e-prints*, November 2015.

Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS*, 2016.

Thomas Steinke and Jonathan Ullman. Interactive fingerprinting codes and the hardness of preventing false discovery. In *Proceedings of The 28th Conference on Learning Theory*, pages 1588–1628, 2015.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In *NIPS 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 2521–2530, 2017.

A Omitted Definitions

Here, we present the definitions that were omitted from the main body due to space constraints.

A.1 Confidence Interval Preliminaries

In our implementation, we are comparing the true average $\phi(\mathcal{D})$ to the answer a, which will be the true answer on the sample with additional noise to ensure each query is stably answered. We then use the following string of inequalities to find the width τ of the confidence interval.

$$\begin{aligned}\Pr\left[|\phi(\mathcal{D}) - a| \geq \tau\right] &\leq \Pr\left[|\phi(\mathcal{D}) - \phi(\boldsymbol{X})| + |\phi(\boldsymbol{X}) - a| \geq \tau\right] \\ &\leq \underbrace{\Pr\left[|\phi(\mathcal{D}) - \phi(\boldsymbol{X})| \geq \tau/2\right]}_{\text{Population Accuracy}} + \underbrace{\Pr\left[|\phi(\boldsymbol{X}) - a| \geq \tau/2\right]}_{\text{Sample Accuracy}}\end{aligned} \tag{2}$$

We will then use this connection to get a bound in terms of the accuracy on the sample and the error in the empirical average to the true mean. Many of the results in this line of work use a *transfer theorem* which states that if a query is selected via a private method, then the query evaluated on the sample is close to the true population answer, thus providing a bound on *population accuracy*. However, we also need to control the *sample accuracy* which is affected by the amount of noise that is added to ensure stability. We then seek a balance between the two terms, where too much noise will give terrible sample accuracy but great accuracy on the population – due to the noise making the choice of query essentially independent of the data – and too little noise makes for great sample accuracy but bad accuracy to the population. We will consider Gaussian noise, and use the composition theorems to determine the scale of noise to add to achieve a target accuracy after k adaptively selected statistical queries.

We then define accuracy with respect to the population. Note that the analyst $\mathcal{A}$ selects a statistical query ϕ_i as a function of what queries $\mathcal{A}$ has already asked and what answers she has seen.

Definition A.1 (Accuracy)**.** A sequence of algorithms $\mathcal{M} = (\mathcal{M}_1, \cdots, \mathcal{M}_k)$, where each $\mathcal{M}_i$ may be adaptively chosen, is (τ, β) accurate (with respect to the population) if for all analysts $\mathcal{A}$ we have $\Pr\left[\max_{i\in[k]} |\phi_i(\mathcal{D}) - \mathcal{M}_i(\boldsymbol{X})| \leq \tau\right] \geq 1 - \beta$, where the probability is over the dataset $\boldsymbol{X} \sim \mathcal{D}^n$ as well as any randomness from the algorithms $\mathcal{M}_1 \cdots, \mathcal{M}_k$ and the adversary $\mathcal{A}$.

Given the size of our dataset n, number of adaptively chosen statistical queries k, and confidence level $1 - \beta$, we want to find what *confidence width* τ ensures $\mathcal{M} = (\mathcal{M}_1, \cdots, \mathcal{M}_k)$ is (τ, β)-accurate with respect to the population when each algorithm $\mathcal{M}_i$ adds either Laplace or Gaussian noise to the answers computed on the sample with some yet to be determined variance. To bound the sample accuracy, we can use the following theorem that gives the accuracy guarantees of the Gaussian mechanism.

Theorem A.2. *If* $\{Z_i : i \in [k]\} \overset{i.i.d.}{\sim} N(0, \sigma^2)$ *then for* $\beta \in (0, 1]$ *we have:*

$$\Pr\left[|Z_i| \geq \sigma\sqrt{2\ln(2/\beta)}\right] \leq \beta \implies \Pr\left[\exists i \in [k] \ s.t. \ |Z_i| \geq \sigma\sqrt{2\ln(2k/\beta)}\right] \leq \beta \tag{3}$$

A.2 Stability Measures

It turns out that privacy preserving algorithms give strong stability guarantees which allows for the rich theory of differential privacy to extend to adaptive data analysis [Dwork et al., 2015c,a, Bassily et al., 2016, Rogers et al., 2016]. In order to define these privacy notions, we define two datasets $\boldsymbol{x} = (x_1, \cdots, x_n), \boldsymbol{x}' = (x'_1, \cdots, x'_n) \in \mathcal{X}^n$ to be *neighboring* if they differ in at most one entry, i.e. there is some $i \in [n]$ where $x_i \neq x'_i$, but $x_j = x'_j$ for all $j \neq i$. We first define *differential privacy*.

Definition A.3 (Differential Privacy [Dwork et al., 2006b,a])**.** A randomized algorithm (or mechanism) $\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}$ is (ϵ, δ)-differentially private (DP) if for all neighboring datasets $\boldsymbol{x}$ and $\boldsymbol{x}'$ and each outcome $S \subseteq \mathcal{Y}$, we have $\Pr\left[\mathcal{M}(\boldsymbol{x}) \in S\right] \leq e^{\epsilon}\Pr\left[\mathcal{M}(\boldsymbol{x}') \in S\right] + \delta$. If $\delta = 0$, we simply say $\mathcal{M}$ is ϵ-DP or pure DP. Otherwise for $\delta > 0$, we say *approximate* DP.

We then give a more recent notion of privacy, called concentrated differential privacy (CDP), which can be thought of as being "in between" pure and approximate DP. In order to define CDP, we define the privacy loss random variable which quantifies how much the output distributions of an algorithm on two neighboring datasets can differ.

Definition A.4 (Privacy Loss)**.** Let $\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}$ be a randomized algorithm. For neighboring datasets $\boldsymbol{x}, \boldsymbol{x'} \in \mathcal{X}^n$, let $Z(y) = \ln\left(\frac{\Pr[\mathcal{M}(\boldsymbol{x})=y]}{\Pr[\mathcal{M}(\boldsymbol{x'})=y]}\right)$. We then define the privacy loss variable $\mathrm{PrivLoss}\left(\mathcal{M}(\boldsymbol{x})||\mathcal{M}(\boldsymbol{x'})\right)$ to have the same distribution as $Z(\mathcal{M}(\boldsymbol{x}))$.

Note that if we can bound the privacy loss random variable with certainty over all neighboring datasets, then the algorithm is pure DP. Otherwise, if we can bound the privacy loss with high probability then it is approximate DP (see Kasiviswanathan and Smith [2014] for a more detailed discussion on this connection).

We can now define *zero concentrated differential privacy* (zCDP), given by Bun and Steinke [2016] (Note that Dwork and Rothblum [2016] initially gave a definition of CDP which Bun and Steinke [2016] then modified).

Definition A.5 (zCDP)**.** An algorithm $\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}$ is ρ-zero concentrated differentially private (zCDP), if for all neighboring datasets $\boldsymbol{x}, \boldsymbol{x'} \in \mathcal{X}^n$ and all $\lambda > 0$ we have

$$\mathbb{E}\left[\exp\left(\lambda\left(\mathrm{PrivLoss}\left(\mathcal{M}(\boldsymbol{x})||\mathcal{M}(\boldsymbol{x'})\right) - \rho\right)\right)\right] \le e^{\lambda^2 \rho}.$$

We then give the Laplace and Gaussian mechanism for statistical queries.

Theorem A.6. *Let $\phi : \mathcal{X} \to [0,1]$ be a statistical query and $\boldsymbol{X} \in \mathcal{X}^n$. The Laplace mechanism $\mathcal{M}_{\mathrm{Lap}} : \mathcal{X}^n \to \mathbb{R}$ is the following $\mathcal{M}_{\mathrm{Lap}}(\boldsymbol{X}) = \frac{1}{n}\sum_{i=1}^n \phi(X_i) + \mathrm{Lap}\left(\frac{1}{\epsilon n}\right)$, which is ϵ-DP. Further, the Gaussian mechanism $\mathcal{M}_{\mathrm{Gauss}} : \mathcal{X}^n \to \mathbb{R}$ is the following $\mathcal{M}_{\mathrm{Gauss}}(X) = \frac{1}{n}\sum_{i=1}^n \phi(X_i) + N\left(0, \frac{1}{2\rho n^2}\right)$, which is ρ-zCDP.*

We now give the advanced composition theorem for k-fold adaptive composition.

Theorem A.7 (Dwork et al. [2010],Kairouz et al. [2017])**.** *The class of ϵ'-DP algorithms is (ϵ, δ)-DP under k-fold adaptive composition where $\delta > 0$ and*

$$\epsilon = \left(\frac{e^{\epsilon'} - 1}{e^{\epsilon'} + 1}\right)\epsilon' k + \epsilon'\sqrt{2k\ln(1/\delta)} \tag{4}$$

We will also use the following results from zCDP.

Theorem A.8 (Bun and Steinke [2016])**.** *The class of ρ-zCDP algorithms is $k\rho$-zCDP under k-fold adaptive composition. Further if $\mathcal{M}$ is ϵ-DP then $\mathcal{M}$ is $\epsilon^2/2$-zCDP and if $\mathcal{M}$ is ρ-zCDP then $\mathcal{M}$ is $(\rho + 2\sqrt{\rho\ln(\sqrt{\pi\rho}/\delta)}, \delta)$-DP for any $\delta > 0$.*

Another notion of stability that we will use is mutual information (in nats) between two random variables: the input $\boldsymbol{X}$ and output $\mathcal{M}(\boldsymbol{X})$.

Definition A.9 (Mutual Information)**.** Consider two random variables X and Y and let $Z(x,y) = \ln\left(\frac{\Pr[(X,Y)=(x,y)]}{\Pr[X=x]\Pr[Y=y]}\right)$. We then denote the mutual information as $I(X;Y) = \mathbb{E}\left[Z(X,Y)\right]$, where the expectation is taken over the joint distribution of (X,Y).

A.3 Monitor Argument

For the population accuracy term in (2), we will use the *monitor argument* from Bassily et al. [2016]. Roughly, this analysis allows us to obtain a bound on the population accuracy over k rounds of interaction between adversary $\mathcal{A}$ and algorithm $\mathcal{M}$ by only considering the difference $|\phi(\boldsymbol{X}) - \phi(\mathcal{D})|$ for the two stage interaction where ϕ is chosen by $\mathcal{A}$ based on outcome $\mathcal{M}(\boldsymbol{X})$. We present the monitor $\mathcal{W}_{\mathcal{D}}[\mathcal{M}, \mathcal{A}]$ in Algorithm 1.

Since our stability definitions are closed under post-processing, we can substitute the monitor $\mathcal{W}_{\mathcal{D}}[\mathcal{M}, \mathcal{A}]$ as our post-processing function f in the above theorem. We then get the following result.

Corollary A.10. *Let $\mathcal{M} = (\mathcal{M}_1, \cdots, \mathcal{M}_k)$, where each $\mathcal{M}_i$ may be adaptively chosen, satisfy any stability condition that is closed under post-processing. For each $i \in [k]$, let ϕ_i be*

Algorithm 1 Monitor $\mathcal{W}_{\mathcal{D}}[\mathcal{M}, \mathcal{A}](\boldsymbol{X})$

Require: $\boldsymbol{x} \in \mathcal{X}^n$

We simulate $\mathcal{M}(\boldsymbol{x})$ and $\mathcal{A}$ interacting. We write $q_1, \cdots, q_k \in \mathcal{Q}_{SQ}$ as the queries chosen by $\mathcal{A}$ and write $a_1, \cdots, a_k \in \mathbb{R}$ as the corresponding answers of $\mathcal{M}$.

Let

$$j^* = \underset{j \in [k]}{\operatorname{argmax}} \left| q_j(\mathcal{D}) - a_j \right|.$$

Ensure: q_{j^*}

the statistical query chosen by adversary $\mathcal{A}$ based on answers $a_j = \mathcal{M}_j(\boldsymbol{X}), \forall j \in [i-1]$, and let ϕ be any function of $(a_1, \cdots, a_k)$. Then, we have

$$\Pr_{\boldsymbol{X} \sim \mathcal{D}^n, (\mathcal{M}_1, \cdots, \mathcal{M}_k)} \left[\max_{i \in [k]} |\phi_i(\mathcal{D}) - a_i| \geq \tau \right] \leq \Pr_{\boldsymbol{X} \sim \mathcal{D}^n, \phi \leftarrow \mathcal{M}(\boldsymbol{X})} \left[|\phi(\mathcal{D}) - \phi(\boldsymbol{X})| \geq \tau/2 \right] + \Pr_{\boldsymbol{X} \sim \mathcal{D}^n, (\mathcal{M}_1, \cdots, \mathcal{M}_k)} \left[\max_{i \in [k]} |\phi_i(\boldsymbol{X}) - a_i| \geq \tau/2 \right]$$

Proof. From the monitor in Algorithm 1 and the fact that $\mathcal{M}$ is closed under post-processing, we have

$$\begin{aligned}
\Pr_{\boldsymbol{X} \sim \mathcal{D}^n, (\mathcal{M}_1, \cdots, \mathcal{M}_k)} \left[\max_{i \in [k]} |\phi_i(\mathcal{D}) - a_i| \geq \tau \right] &= \Pr_{\boldsymbol{X} \sim \mathcal{D}^n, \phi_{j^*} \leftarrow \mathcal{W}_{\mathcal{D}}[\mathcal{M}, \mathcal{A}](\boldsymbol{X})} \left[|\phi_{j^*}(\mathcal{D}) - a_{j^*}| \geq \tau \right] \\
&\leq \Pr_{\boldsymbol{X} \sim \mathcal{D}^n, \phi_{j^*} \leftarrow \mathcal{W}_{\mathcal{D}}[\mathcal{M}, \mathcal{A}](\boldsymbol{X})} \left[|\phi_{j^*}(\mathcal{D}) - \phi_{j^*}(\boldsymbol{X})|| \geq \tau/2 \right] \\
&\quad + \Pr_{\boldsymbol{X} \sim \mathcal{D}^n, \phi_{j^*} \leftarrow \mathcal{W}_{\mathcal{D}}[\mathcal{M}, \mathcal{A}](\boldsymbol{X})} \left[|\phi_{j^*}(\boldsymbol{X}) - a_{j^*}|| \geq \tau/2 \right] \\
&\leq \Pr_{\boldsymbol{X} \sim \mathcal{D}^n, \phi \leftarrow \mathcal{M}(\boldsymbol{X})} \left[|\phi(\mathcal{D}) - \phi(\boldsymbol{X})| \geq \tau/2 \right] \\
&\quad + \Pr_{\boldsymbol{X} \sim \mathcal{D}^n, (\mathcal{M}_1, \cdots, \mathcal{M}_k)} \left[\max_{i \in [k]} |\phi_i(\boldsymbol{X}) - a_i| \geq \tau/2 \right]
\end{aligned}$$

□

We can then use the above corollary to obtain an accuracy guarantee by union bounding over the sample accuracy for all k rounds of interaction and then bounding the population error for a single adaptively chosen statistical query.

B Omitted Confidence Interval Bounds

Here we present the bounds derived via prior work, and the novel bounds for Thresholdout Dwork et al. [2015a].

B.1 Confidence Bounds from Dwork et al. Dwork et al. [2015a]

We start by deriving confidence bounds using results from Dwork et al. [2015a], which uses the following transfer theorem (see Theorem 10 in Dwork et al. [2015a]).

Theorem B.1. *If $\mathcal{M}$ is (ϵ, δ)-DP where $\phi \leftarrow \mathcal{M}(\boldsymbol{X})$ and $\tau \geq \sqrt{\frac{48}{n} \ln(4/\beta)}$, $\epsilon \leq \tau/4$ and $\delta = \exp\left(\frac{-4\ln(8/\beta)}{\tau}\right)$, then $\Pr\left[|\phi(\mathcal{D}) - \phi(\boldsymbol{X})| \geq \tau\right] \leq \beta$.*

We pair this together with the accuracy from either the Gaussian mechanism or the Laplace mechanism along with Corollary A.10 to get the following result

Theorem B.2. *Given confidence level $1 - \beta$ and using the Laplace or Gaussian mechanism for each algorithm $\mathcal{M}_i$, then $(\mathcal{M}_1, \cdots, \mathcal{M}_k)$ is $(\tau^{(1)}, \beta)$-accurate.*

- **Laplace Mechanism**: *We define $\tau^{(1)}$ to be the solution to the following program*

$$\begin{aligned} \min \quad & \tau \\ s.t. \quad & \tau \geq 2\sqrt{\frac{48}{n}\ln(8/\beta)} \\ & \tau \geq \frac{2\ln(2k/\beta)}{n\epsilon'} \\ & \tau \geq 8\left(\epsilon' k \cdot \left(\frac{e^{\epsilon'}-1}{e^{\epsilon'}+1}\right) + 4\epsilon' \cdot \sqrt{\frac{k \ln \frac{16}{\beta}}{\tau}}\right) \\ \textit{for} \quad & \epsilon' > 0 \end{aligned}$$

- **Gaussian Mechanism**: *We define $\tau^{(1)}$ to be the solution to the following program*

$$\begin{aligned} \min \quad & \tau \\ s.t. \quad & \tau \geq 2\sqrt{\frac{48}{n}\ln(8/\beta)} \\ & \tau \geq \frac{1}{n}\sqrt{\frac{1}{\rho'}\ln(4k/\beta)} \\ & \tau \geq 8\rho' k + \sqrt{256\rho' k\left(\ln\left(\sqrt{\pi\rho' k}\right) + \frac{\ln\frac{16}{\beta}}{\tau}\right)} \\ \textit{for} \quad & \rho' > 0 \end{aligned}$$

To bound the sample accuracy, we will use the following lemma that gives the accuracy guarantees of Laplace mechanism.

Lemma B.3. *If $\{Y_i : i \in [k]\} \overset{i.i.d.}{\sim} \mathrm{Lap}(b)$, then for $\beta \in (0,1]$ we have:*

$$\Pr\left[|Y_i| \geq \ln(1/\beta)b\right] \leq \beta \implies \Pr\left[\exists i \in [k] \;\; s.t. \;\; |Y_i| \geq \ln(k/\beta)b\right] \leq \beta \tag{5}$$

Proof of Theorem B.2. We will focus on the Laplace mechanism part first, so that we add $\mathrm{Lap}\left(\frac{1}{n\epsilon'}\right)$ noise to each answer. After k adaptively selected queries, the entire sequence of noisy answers is (ϵ, δ)-DP where

$$\epsilon = k\epsilon' \cdot \frac{e^{\epsilon'}-1}{e^{\epsilon'}+1} + \epsilon' \cdot \sqrt{2k\ln(1/\delta)}.$$

We then want to bound the two terms in (2). To bound the sample accuracy, we then use (5) so that

$$\tau \geq \frac{2}{n\epsilon'}\ln(2k/\beta)$$

For the population accuracy, we need to apply Theorem B.1, which requires us to have the following, where we take a union bound over all selected statistical queries:

$$\delta = \exp\left(\frac{-8\ln(16/\beta)}{\tau}\right) \qquad \text{and} \qquad \tau \geq \max\left\{2\sqrt{\frac{48}{n}\ln(8/\beta)}, 8\epsilon\right\}.$$

We then write ϵ in terms of δ to get:

$$\epsilon = k\epsilon' \cdot \frac{e^{\epsilon'}-1}{e^{\epsilon'}+1} + 4\epsilon' \cdot \sqrt{k\frac{\ln(16/\beta)}{\tau}}.$$

We are then left to pick $\epsilon' > 0$ to obtain the smallest value of τ.

When can then follow a similar argument when we add Gaussian noise with variance $\frac{1}{2n^2\rho'}$. The only modification we make is using Theorem A.8 to get a composed DP algorithm with parameters in terms of ρ', and the accuracy guarantee in (3). $\square$

B.2 Confidence Bounds from Bassily et al. Bassily et al. [2016]

We now go through the argument of Bassily et al. [2016] to improve the constants as much as we can via their analysis to get a decent confidence bound on k adaptively chosen statistical queries. This requires presenting their *monitoring*, which is similar to the monitor presented in Algorithm 1 but takes as input several independent datasets. We first present the result.

Theorem B.4. *Given confidence level $1-\beta$ and using the Laplace or Gaussian mechanism for each algorithm $\mathcal{M}_i$, then $(\mathcal{M}_1, \cdots, \mathcal{M}_k)$ is $(\tau^{(2)}, \beta)$-accurate.*

- **Laplace Mechanism**: *We define $\tau^{(2)}$ to be the following quantity:*

$$\frac{1}{1-(1-\beta)^{\left\lfloor \frac{1}{\beta} \right\rfloor}} \cdot \inf_{\substack{\epsilon'>0,\\ \delta\in(0,1)}} \left\{ e^{\psi} - 1 + 2\left\lfloor \frac{1}{\beta} \right\rfloor \delta + \frac{\ln\left(\frac{k}{2\delta}\right)}{\epsilon' n} \right\},$$

$$\text{where } \psi = \left(\frac{e^{\epsilon'}-1}{e^{\epsilon'}+1}\right) \cdot \epsilon' k + \epsilon' \sqrt{2k \ln\left(\frac{1}{\delta}\right)}$$

- **Gaussian Mechanism**: *We define $\tau^{(2)}$ to be the following quantity:*

$$\frac{1}{1-(1-\beta)^{\left\lfloor \frac{1}{\beta} \right\rfloor}} \cdot \inf_{\substack{\rho>0,\\ \delta\in(0,1)}} \left\{ e^{\xi} - 1 + 2\left\lfloor \frac{1}{\beta} \right\rfloor \delta + \sqrt{\frac{\ln\left(\frac{k}{\delta}\right)}{n^2 \rho}} \right\},$$

$$\text{where } \xi = k\rho + 2\sqrt{k\rho \ln\left(\frac{\sqrt{\pi\rho}}{\delta}\right)}$$

In order to prove this result, we begin with a technical lemma which considers an algorithm $\mathcal{W}$ that takes as input a collection of s samples and outputs both an index in $[s]$ and a statistical query, where we denote $\mathcal{Q}_{SQ}$ as the set of all statistical queries $q : \mathcal{X} \to [0,1]$ and their negation.

Lemma B.5 ([Bassily et al., 2016]). *Let $\mathcal{W} : (\mathcal{X}^n)^s \to \mathcal{Q}_{SQ} \times [s]$ be (ϵ, δ)-DP. If $\vec{\boldsymbol{X}} = (\boldsymbol{X}^{(1)}, \cdots, \boldsymbol{X}^{(s)}) \sim (\mathcal{D}^n)^s$ then*

$$\left| \mathop{\mathbb{E}}_{\vec{\boldsymbol{X}}, (q,t) = \mathcal{W}(\vec{\boldsymbol{X}})} \left[q(\mathcal{D}) - q(\boldsymbol{X}^{(t)}) \right] \right| \le e^{\epsilon} - 1 + s\delta$$

We then define what we will call the *extended monitor* in Algorithm 2.

Algorithm 2 Extended Monitor $\mathcal{W}_{\mathcal{D}}[\mathcal{M}, \mathcal{A}](\vec{\boldsymbol{X}})$

Require: $\vec{\boldsymbol{x}} = (\boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}^{(s)}) \in (\mathcal{X}^n)^s$
  for $t \in [s]$ **do**
    We simulate $\mathcal{M}(\boldsymbol{X}^{(t)})$ and $\mathcal{A}$ interacting. We write $q_{t,1}, \cdots, q_{t,k} \in \mathcal{Q}_{SQ}$ as the queries chosen by $\mathcal{A}$ and write $a_{t,1}, \cdots, a_{t,k} \in \mathbb{R}$ as the corresponding answers of $\mathcal{M}$.
  Let

$$(j^*, t^*) = \mathop{\mathrm{argmax}}_{j\in[k], t\in[s]} \left| q_{t,j}(\mathcal{D}) - a_{t,j} \right|.$$

  if $a_{t^*,j^*} - q_{t^*,j^*}(\mathcal{D}) \ge 0$ **then**
    $q^* \leftarrow q_{t^*,j^*}$
  else
    $q^* \leftarrow -q_{t^*,j^*}$
Ensure: (q^*, t^*)

We then present a series of lemmas that leads to an accuracy bound from Bassily et al. [2016].

Lemma B.6 ([Bassily et al., 2016]). *For each $\epsilon, \delta \geq 0$, if $\mathcal{M}$ is (ϵ, δ)-DP for k adaptively chosen queries from $\mathcal{Q}_{SQ}$, then for every data distribution $\mathcal{D}$ and analyst $\mathcal{A}$, the monitor $\mathcal{W}_{\mathcal{D}}[\mathcal{M}, \mathcal{A}]$ is (ϵ, δ)-DP.*

Lemma B.7 ([Bassily et al., 2016]). *If $\mathcal{M}$ fails to be (τ, β)-accurate, then $q^*(\mathcal{D}) - a^* \geq 0$, where a^* is the answer to q^* during the simulation ($\mathcal{A}$ can determine a^* from output (q^*, t^*)) and*

$$\Pr_{\substack{\vec{\boldsymbol{X}} \sim (\mathcal{D}^n)^s, \\ (q^*, t^*) = \mathcal{W}_{\mathcal{D}}[\mathcal{M}, \mathcal{A}](\vec{\boldsymbol{X}})}} \left[\max_{j \in [k]} |q_{t,j}(\mathcal{D}) - a_{t,j}| > \tau \right] > 1 - (1 - \beta)^s.$$

The following result is not stated exactly the same as in Bassily et al. [2016], but it follows the same analysis. We just do not simplify the expressions in the inequalities.

Lemma B.8. *If $\mathcal{M}$ is (τ', β') accurate on the sample but not (τ, β)-accurate for the population, then*

$$\left| \mathop{\mathbb{E}}_{\vec{\boldsymbol{X}} \sim (\mathcal{D}^n)^s, (q,t) = \mathcal{W}[\mathcal{M}, \mathcal{A}](\vec{\boldsymbol{X}})} \left[q(\mathcal{D}) - q(\boldsymbol{X}^{(t)}) \right] \right| \geq \tau \left(1 - (1 - \beta)^s\right) - \left(\tau' + 2s\beta'\right).$$

We now put everything together to get our result.

Proof of Theorem B.4. We ultimately want a contradiction between the result given in Lemma B.5 and Lemma B.8. Thus, we want to find the parameter values that minimizes τ but satisfies the following inequality

$$\tau \left(1 - (1 - \beta)^s\right) - \left(\tau' + 2s\beta'\right) > e^{\epsilon} - 1 + s\delta. \tag{6}$$

We first analyze the case when we add noise $\text{Lap}\left(\frac{1}{n\epsilon'}\right)$ to each query answer on the sample to preserve ϵ'-DP of each query and then use advanced composition Theorem A.7 to get a bound on ϵ.

$$\epsilon = \left(\frac{e^{\epsilon'} - 1}{e^{\epsilon'} + 1}\right) \epsilon' k + \epsilon' \sqrt{2k \ln(1/\delta)} = \psi.$$

Further, we obtain (τ', β')-accuracy on the sample, where for $\beta' > 0$ we have $\tau' = \frac{\ln(k/\beta')}{\epsilon' n}$. We then plug these values into (6) to get the following bound on τ

$$\tau \geq \left(\frac{1}{1 - (1 - \beta)^s}\right) \left(\frac{\ln\left(\frac{k}{\beta'}\right)}{\epsilon' n} + 2s\beta' + e^{\psi} - 1 + s\delta\right)$$

We then choose some of the parameters to be the same as in Bassily et al. [2016], like $s = \lfloor 1/\beta \rfloor$ and $\beta' = 2\delta$. We then want to find the best parameters ϵ', δ that makes the right hand side as small as possible. Thus, the best confidence width τ that we can get with this approach is the following

$$\frac{1}{1 - (1 - \beta)^{\lfloor \frac{1}{\beta} \rfloor}} \cdot \inf_{\substack{\epsilon' > 0, \\ \delta \in (0,1)}} \left\{ e^{\psi} - 1 + 2 \left\lfloor \frac{1}{\beta} \right\rfloor \delta + \frac{\ln\left(\frac{k}{2\delta}\right)}{\epsilon' n} \right\}$$

Using the same analysis but with Gaussian noise added to each statistical query answer with variance $\frac{1}{2\rho' n^2}$ (so that $\mathcal{M}$ is $\rho' k$-zCDP), we get the following confidence width τ,

$$\frac{1}{1 - (1 - \beta)^{\lfloor \frac{1}{\beta} \rfloor}} \cdot \inf_{\substack{\rho > 0, \\ \delta \in (0,1)}} \left\{ e^{\xi} - 1 + 2 \left\lfloor \frac{1}{\beta} \right\rfloor \delta + \sqrt{\frac{\ln\left(\frac{k}{\delta}\right)}{n^2 \rho}} \right\}$$

$\square$

B.3 Confidence bounds for Thresholdout Dwork et al. [2015a]

Using techniques similar to those in Section 2 but a more involved analysis, we also provide confidence bounds for the well-known technique Thresholdout from Dwork et al. [2015a]. This is significant because the bounds only have a dependence on the number of queries that overfit to the training set, and not on the total number of queries asked. We present the bounds for Thresholdout in

Here, we present the bounds for Thresholdout Dwork et al. [2015a].

Theorem B.9. *Given confidence level $1-\beta$ and using the Thresholdout mechanism $\mathcal{M}$ with budget B, noise scale σ, and threshold T, for answering queries q_i, $i \in [k]$, such that $\mathcal{M}$ uses the holdout set of size h to answer at most $B-1$ queries, then the answer for each query q_i is (τ, β)-accurate, where*

$$\tau = \max\left\{\sqrt{\frac{2\xi}{h\beta}}, 2\sigma\ln\left(\frac{\beta}{2}\right), \sqrt{\frac{1}{\beta}}\cdot\left(\sqrt{T^2+56\sigma^2}+\sqrt{\frac{\xi}{4h}}\right)\right\}, \textit{for } \xi = \min_{\lambda\in[0,1)}\left(\frac{\frac{2B}{\sigma^2 h}-\ln(1-\lambda)}{\lambda}\right).$$

Proof of Theorem B.9. Let us denote the holdout set in $\mathcal{M}$ by X_h and the remaining set as X_t. We know that for at most B queries q_i, the output of $\mathcal{M}$ was $a_i = \phi_i(X_h) + Lap(\sigma)$, whereas it was $a_i = \phi_i(X_t)$ for at least $k-B$ queries. Let us start by considering the queries answered via X_h. We define a set S_h which contains the indices of the queries answered via X_h.

For every $i \in S_h$, there are two costs induced due to privacy: the Sparse Vector component, and the noise addition to $q_i(X_h)$. By the proof of Lemma 23 in Dwork et al. [2015a], each individually provides a guarantee of $\left(\frac{1}{\sigma h}, 0\right)$-DP. Using Theorem A.8, this translates to each providing a $\left(\frac{1}{2\sigma^2h^2}\right)$-zCDP guarantee. Since there are at most B such instances of each, by Theorem A.8 we get that $\mathcal{M}$ is $\left(\frac{B}{\sigma^2h^2}\right)$-zCDP. Thus, by Lemma C.5 we have

$$I(\mathcal{M}(X_h); X_h) \le \frac{B}{\sigma^2 h}$$

Proceeding similar to the proof of Theorem 2.2, we use the sub-Gaussian parameter for statistical queries in Lemma C.4 to obtain the following bound from Theorem C.1:

$$\begin{aligned}\mathop{\mathbb{E}}_{\substack{\boldsymbol{X}\sim\mathcal{D}^n,\\ q^*\sim\mathcal{W}_{\mathcal{D}}[\mathcal{M},\mathcal{A}](\boldsymbol{X})}}\left[(q^*(\boldsymbol{X}_h)-q^*(\mathcal{D}))^2\right] &= \mathop{\mathbb{E}}_{\boldsymbol{X}\sim\mathcal{D}^n,\mathcal{M},\mathcal{A}}\left[\max_{i\in S_h}\left\{(q_i(\boldsymbol{X}_h)-q_i(\mathcal{D}))^2\right\}\right]\\ &\le \frac{1}{4h}\cdot\min_{\lambda\in[0,1)}\left(\frac{\frac{2B}{\sigma^2h}-\ln(1-\lambda)}{\lambda}\right) \end{aligned} \tag{7}$$

Using Chebyshev's inequality, we can get a high probability bound on the population accuracy in (2) for the answers on the holdout set as:

$$\Pr_{\substack{\boldsymbol{X}\sim\mathcal{D}^n,\\ \phi\leftarrow\mathcal{M}(\boldsymbol{X})}}\left[|\phi(\boldsymbol{X}_h)-\phi(\mathcal{D})|\ge\frac{\tau}{2}\right] \le \frac{1}{h\tau^2}\cdot\min_{\lambda\in[0,1)}\left(\frac{\frac{2B}{\sigma^2h}-\ln(1-\lambda)}{\lambda}\right) \tag{8}$$

Now, since $\mathcal{M}$ releases $a_i = q_i(X_h) + Lap(\sigma)$ for every $i \in S_h$, we bound the sample error via the guarantees of the Laplace distribution (Lemma B.3):

$$\Pr_{\substack{\boldsymbol{X}\sim\mathcal{D}^n,\\ \phi\leftarrow\mathcal{M}(\boldsymbol{X})}}\left[|a_i-\phi(\boldsymbol{X}_h)|\ge\frac{\tau}{2}\right] \le \exp\left(\frac{-\tau}{2\sigma}\right) \tag{9}$$

Thus, for the accuracy bound for each $i \in S_h$, using (8), (9), and Corollary A.10, we want to find the smallest value of τ such that the following conditions are satisfied:

$$\tau \ge \sqrt{\frac{2}{h\beta}\cdot\min_{\lambda\in[0,1)}\left(\frac{\frac{2B}{\sigma^2h}-\ln(1-\lambda)}{\lambda}\right)} \qquad \text{and} \qquad \tau \ge 2\sigma\ln(\beta/2) \tag{10}$$

Next, we consider the queries answered via X_t. Define a set $S_t = [k] \setminus S_h$ containing the indices of the queries answered via X_t.

For every $i \in S_t$, we have $|q_i(X_t) - q_i(X_h)| \leq T + Lap(2\sigma) + Lap(4\sigma)$, and the output of $\mathcal{M}$ is $q_i(X_t)$. Thus, we have:

$$
\begin{aligned}
\underset{\substack{\boldsymbol{X}\sim\mathcal{D}^n,\\ q^*\sim\mathcal{W}_{\mathcal{D}}[\mathcal{M},\mathcal{A}](\boldsymbol{X})}}{\mathbb{E}}\left[(q^*(\boldsymbol{X}_t) - q^*(\boldsymbol{X}_h))^2\right] &\leq \mathbb{E}\left[(T + Lap(2\sigma) + Lap(4\sigma))^2\right] \\
&\leq T^2 + \mathbb{E}\left[(Lap(2\sigma) + Lap(4\sigma))^2\right] \\
&\leq T^2 + 8\sigma^2 + 32\sigma^2 + \sqrt{\mathbb{E}\left[(Lap(2\sigma))^2\right]\cdot\mathbb{E}\left[(Lap(4\sigma))^2\right]} \\
&= T^2 + 56\sigma^2 \qquad (11)
\end{aligned}
$$

where the last inequality follows from the Cauchy-Schwarz inequality.

As a result, we have:

$$
\begin{aligned}
\underset{\substack{\boldsymbol{X}\sim\mathcal{D}^n,\\ q^*\sim\mathcal{W}_{\mathcal{D}}[\mathcal{M},\mathcal{A}](\boldsymbol{X})}}{\mathbb{E}}\left[(q^*(\boldsymbol{X}_t) - q^*(\mathcal{D}))^2\right] &= \underset{\substack{\boldsymbol{X}\sim\mathcal{D}^n,\\ q^*\sim\mathcal{W}_{\mathcal{D}}[\mathcal{M},\mathcal{A}](\boldsymbol{X})}}{\mathbb{E}}\left[(q^*(\boldsymbol{X}_t) - q^*(\boldsymbol{X}_h) + q^*(\boldsymbol{X}_h) - q^*(\mathcal{D}))^2\right] \\
&\leq \underset{\substack{\boldsymbol{X}\sim\mathcal{D}^n,\\ q^*\sim\mathcal{W}_{\mathcal{D}}[\mathcal{M},\mathcal{A}](\boldsymbol{X})}}{\mathbb{E}}\left[(q^*(\boldsymbol{X}_t) - q^*(\boldsymbol{X}_h))^2\right] \\
&\quad + \underset{\substack{\boldsymbol{X}\sim\mathcal{D}^n,\\ q^*\sim\mathcal{W}_{\mathcal{D}}[\mathcal{M},\mathcal{A}](\boldsymbol{X})}}{\mathbb{E}}\left[(q^*(\boldsymbol{X}_h) - q^*(\mathcal{D}))^2\right] \\
&\quad + 2\cdot\sqrt{\underset{\substack{\boldsymbol{X}\sim\mathcal{D}^n,\\ q^*\sim\mathcal{W}_{\mathcal{D}}[\mathcal{M},\mathcal{A}](\boldsymbol{X})}}{\mathbb{E}}\left[(q^*(\boldsymbol{X}_t) - q^*(\boldsymbol{X}_h))^2\right]} \\
&\qquad \cdot\sqrt{\underset{\substack{\boldsymbol{X}\sim\mathcal{D}^n,\\ q^*\sim\mathcal{W}_{\mathcal{D}}[\mathcal{M},\mathcal{A}](\boldsymbol{X})}}{\mathbb{E}}\left[(q^*(\boldsymbol{X}_h) - q^*(\mathcal{D}))^2\right]} \\
&\leq T^2 + 56\sigma^2 + \frac{1}{4h}\cdot\min_{\lambda\in[0,1)}\left(\frac{\frac{2B}{\sigma^2 h} - \ln(1-\lambda)}{\lambda}\right) \\
&\quad + 2\sqrt{(T^2 + 56\sigma^2)\cdot\left(\frac{1}{4h}\cdot\min_{\lambda\in[0,1)}\left(\frac{\frac{2B}{\sigma^2 h} - \ln(1-\lambda)}{\lambda}\right)\right)} \\
&= \left(\sqrt{T^2 + 56\sigma^2} + \sqrt{\frac{1}{4h}\cdot\min_{\lambda\in[0,1)}\left(\frac{\frac{2B}{\sigma^2 h} - \ln(1-\lambda)}{\lambda}\right)}\right)^2
\end{aligned}
$$

where the first inequality follows by linearity of expectation and the Cauchy-Schwarz inequality, and the second inequality follows from Equation (7) and Equation (11).

Using Chebyshev's inequality, we can get a high probability bound on the population accuracy in (2) for the answers on X_t as:

$$
\underset{\substack{\boldsymbol{X}\sim\mathcal{D}^n,\\ \phi\leftarrow\mathcal{M}(\boldsymbol{X})}}{\Pr}\left[|\phi(\boldsymbol{X}_t) - \phi(\mathcal{D})| \geq \tau\right] \leq \frac{1}{\tau^2}\cdot\left(\sqrt{T^2 + 56\sigma^2} + \sqrt{\frac{1}{4h}\cdot\min_{\lambda\in[0,1)}\left(\frac{\frac{2B}{\sigma^2 h} - \ln(1-\lambda)}{\lambda}\right)}\right)^2 \qquad (12)
$$

We get the statement of the theorem by the conditions in Equation (10), and Equation (12). □

In Figure 6, we give the widths of the valid confidence intervals for each query for Thresholdout when B queries evaluated on the holdout set of size $h = 100B$. We present plots for three different values of the threshold T, and we tune over the noise scale σ for every pair (T, B) to obtain the best width guaranteed by Theorem B.9.

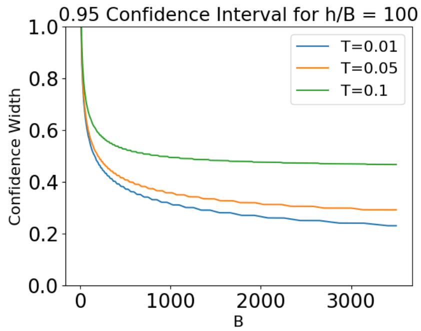

Figure 6: Widths of confidence intervals for each query for Thresholdout

C Omitted Proofs

In this section, we provide detailed proofs that have been omitted from the main body of the paper.

Proof of Theorem 3.1. Consider a dataset $X \in \mathcal{X}^n$, where $\mathcal{X}$ is the uniform distribution over $\{-1, 1\}^{k+1}$. Now, $\forall i \in [k]$, we have that:

$$\mathop{\mathbb{E}}_{X}\left[\boldsymbol{x}(i) \cdot \boldsymbol{x}(k)\right] = \Pr\left(\boldsymbol{x}(i) = \boldsymbol{x}(k)\right) - \Pr\left(\boldsymbol{x}(i) \neq \boldsymbol{x}(k)\right) = a_i$$

$$\therefore \Pr_X\left(\boldsymbol{x}(i) = \boldsymbol{x}(k)\right) = \frac{1+a_i}{2} \qquad \text{and} \qquad \Pr_X\left(\boldsymbol{x}(i) \neq \boldsymbol{x}(k)\right) = \frac{1-a_i}{2}$$

Now,

$$\begin{aligned}
\ln\left(\frac{\Pr_X\left(\boldsymbol{x}(k) = 1 \,|\, \wedge_{i \in [k]}\, \boldsymbol{x}(i) = x_i\right)}{\Pr_X\left(\boldsymbol{x}(k) = -1 \,|\, \wedge_{i \in [k]}\, \boldsymbol{x}(i) = x_i\right)}\right) &= \ln\left(\frac{\Pr_X\left(\boldsymbol{x}(k) = 1 \wedge (\wedge_{i \in [k]} \boldsymbol{x}(i) = x_i)\right)}{\Pr_X\left(\boldsymbol{x}(k) = -1 \wedge (\wedge_{i \in [k]} \boldsymbol{x}(i) = x_i)\right)}\right) \\
&= \ln\left(\prod_{i \in [k]} \frac{\Pr_X\left(\boldsymbol{x}(k) = 1 \wedge \boldsymbol{x}(i) = x_i\right)}{\Pr_X\left(\boldsymbol{x}(k) = -1 \wedge \boldsymbol{x}(i) = x_i\right)}\right) \\
&= \ln\left(\prod_{i \in [k]} \left(\frac{\Pr_X\left(\boldsymbol{x}(k) = \boldsymbol{x}(i)\right)}{\Pr_X\left(\boldsymbol{x}(k) \neq \boldsymbol{x}(i)\right)}\right)^{x_i}\right) \\
&= \ln\left(\prod_{i \in [k]} \left(\frac{1+a_i}{1-a_i}\right)^{x_i}\right) = \sum_{i \in [k]} \left(x_i \cdot \ln\frac{1+a_i}{1-a_i}\right)
\end{aligned}$$

Thus, $q_k(\boldsymbol{x}) = sign\left(\ln\left(\frac{\Pr_X\left(\boldsymbol{x}(k)=1 | \wedge_{i\in[k]} \boldsymbol{x}(i)=x_i\right)}{\Pr_X\left(\boldsymbol{x}(k)=-1 | \wedge_{i\in[k]} \boldsymbol{x}(i)=x_i\right)}\right)\right)$.

As a result, the query q_k in Algorithm 4 is a naive Bayes classifier of $\boldsymbol{x}(k)$, and given that $\mathcal{X}$ is the uniform distribution over $\{-1, 1\}^{k+1}$, this is the best possible classifier for $\boldsymbol{x}(k)$. This results answer a_k achieving the maximum possible deviation from the answer on the population, which is 0 as $\mathcal{X}$ is uniformly distributed over $\{-1, 1\}^{k+1}$. Thus, a_k results in the maximum possible RMSE. □

C.1 Proof of Theorem 2.1

Rather than use the stated result in Russo and Zou [2016], we use a modified "corrected" version and provide a proof for it here. The result stated here and the one in Russo and Zou [2016] are incomparable.

Theorem C.1. *Let $\mathcal{Q}_\sigma$ be the class of queries $q : \mathcal{X}^n \to \mathbb{R}$ such that $q(\boldsymbol{X}) - q(\mathcal{D}^n)$ is σ-subgaussian where $\boldsymbol{X} \sim \mathcal{D}^n$. If $\mathcal{M} : \mathcal{X}^n \to \mathcal{Q}_\sigma$ is a randomized mapping from datasets to queries such that $I(\mathcal{M}(\boldsymbol{X}); \boldsymbol{X}) \leq B$ then*

$$\mathop{\mathbb{E}}_{\substack{\boldsymbol{X}\sim\mathcal{D}^n,\\ q\leftarrow\mathcal{M}(\boldsymbol{X})}} \left[(q(\boldsymbol{X}) - q(\mathcal{D}^n)))^2\right] \leq \sigma^2 \cdot \min_{\lambda\in[0,1)} \left(\frac{2B - \ln(1-\lambda)}{\lambda}\right).$$

In order to prove the theorem, we need the following results.

Lemma C.2 (Russo and Zou [2015], Gray [1990])**.** *Given two probability measures P and Q defined on a common measurable space and assuming that P is absolutely continuous with respect to Q, then*

$$\mathrm{D}_{KL}[P||Q] = \sup_X \left\{\mathop{\mathbb{E}}_P[X] - \log \mathop{\mathbb{E}}_Q[\exp(X)]\right\}$$

Lemma C.3 (Russo and Zou [2015])**.** *If X is a zero-mean subgaussian random variable with parameters σ then*

$$\mathbb{E}\left[\exp\left(\frac{\lambda X^2}{2\sigma^2}\right)\right] \leq \frac{1}{\sqrt{1-\lambda}} \qquad \forall\lambda \in [0,1)$$

Proof of Theorem C.1. Proceeding similar to the proof of Proposition 3.1 in Russo and Zou [2015], we write $\boldsymbol{\phi}(\boldsymbol{X}) = (\phi(\boldsymbol{X}) : \phi \in \mathcal{Q}_\sigma)$,

$$\begin{aligned}
I(\mathcal{M}(\boldsymbol{X}); \boldsymbol{X}) &\geq I(\mathcal{M}(\boldsymbol{X}); \boldsymbol{\phi}(\boldsymbol{X})) \\
&= \sum_{\boldsymbol{a},\ q\in\mathcal{Q}_\sigma} \ln\left(\frac{\Pr[(\boldsymbol{\phi}(\boldsymbol{X}), \mathcal{M}(\boldsymbol{X})) = (\boldsymbol{a}, q)]}{\Pr[\boldsymbol{\phi}(\boldsymbol{X}) = \boldsymbol{a}]\Pr[\mathcal{M}(\boldsymbol{X}) = q]}\right) \cdot \Pr[(\boldsymbol{\phi}(\boldsymbol{X}), \mathcal{M}(\boldsymbol{X})) = (\boldsymbol{a}, q)] \\
&= \sum_{\boldsymbol{a},\ q\in\mathcal{Q}_\sigma} \ln\left(\frac{\Pr[\boldsymbol{\phi}(\boldsymbol{X}) = \boldsymbol{a}|\mathcal{M}(\boldsymbol{X}) = q]}{\Pr[\boldsymbol{\phi}(\boldsymbol{X}) = \boldsymbol{a}]}\right) \cdot \Pr[\mathcal{M}(\boldsymbol{X}) = q]\Pr[\boldsymbol{\phi}(\boldsymbol{X}) = \boldsymbol{a}|\mathcal{M}(\boldsymbol{X}) = q] \\
&\geq \sum_{a,\ q\in\mathcal{Q}_\sigma} \ln\left(\frac{\Pr[q(\boldsymbol{X}) = a|\mathcal{M}(\boldsymbol{X}) = q]}{\Pr[q(\boldsymbol{X}) = a]}\right) \cdot \Pr[\mathcal{M}(\boldsymbol{X}) = q]\Pr[q(\boldsymbol{X}) = a|\mathcal{M}(\boldsymbol{X}) = q] \\
&= \sum_{q\in\mathcal{Q}_\sigma} \Pr[\mathcal{M}(\boldsymbol{X}) = q] \cdot \mathrm{D}_{KL}[(q(\boldsymbol{X})|\mathcal{M}(\boldsymbol{X}) = q)||q(\boldsymbol{X})] \qquad (13)
\end{aligned}$$

where the first inequality follows from post processing of mutual information, i.e. the data processing inequality. Consider the function $f_q(x) = \frac{\lambda}{2\sigma^2}(x - q(\mathcal{D}^n))^2$ for $\lambda \in [0, 1)$. We have

$$\begin{aligned}
\mathrm{D}_{KL}[(q(\boldsymbol{X})|\mathcal{M}(\boldsymbol{X}) = q)||q(\boldsymbol{X})] &\geq \mathop{\mathbb{E}}_{\boldsymbol{X}\sim\mathcal{D}^n,\mathcal{M}}[f_q(q(\boldsymbol{X}))|\mathcal{M}(\boldsymbol{X}) = q] - \ln \mathop{\mathbb{E}}_{\substack{\boldsymbol{X}\sim\mathcal{D}^n,\\ q\sim\mathcal{M}(\boldsymbol{X})}}[\exp(f_q(q(\boldsymbol{X})))] \\
&\geq \frac{\lambda}{2\sigma^2} \mathop{\mathbb{E}}_{\boldsymbol{X}\sim\mathcal{D}^n,\mathcal{M}}\left[(q(\boldsymbol{X}) - q(\mathcal{D}^n))^2|\mathcal{M}(\boldsymbol{X}) = q\right] - \ln\left(\frac{1}{\sqrt{1-\lambda}}\right)
\end{aligned}$$

where the first and second inequalities follows from Lemmas C.2 and C.3, respectively.

Therefore, from eq. (13), we have

$$I(\mathcal{M}(\boldsymbol{X}); \boldsymbol{X}) \geq \frac{\lambda}{2\sigma^2} \mathop{\mathbb{E}}_{\boldsymbol{X}\sim\mathcal{D}^n,\ q\sim\mathcal{M}(\boldsymbol{X})}\left[(q(\boldsymbol{X}) - q(\mathcal{D}^n))^2\right] - \ln\left(\frac{1}{\sqrt{1-\lambda}}\right)$$

Rearranging terms, we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{X}\sim\mathcal{D}^n,\ q\sim\mathcal{M}(\boldsymbol{X})}\left[(q(\boldsymbol{X}) - q(\mathcal{D}))^2\right] &\leq \frac{2\sigma^2}{\lambda}\left(I(\mathcal{M}(\boldsymbol{X}); \boldsymbol{X}) + \ln\left(\frac{1}{\sqrt{1-\lambda}}\right)\right) \\
&= \sigma^2 \cdot \frac{2I(\mathcal{M}(\boldsymbol{X}); \boldsymbol{X}) - \ln(1-\lambda)}{\lambda}
\end{aligned}$$

□

In order to apply this result, we need to know the subgaussian parameter for statistical queries and the mutual information for private algorithms.

Lemma C.4. *For statistical queries ϕ and $\boldsymbol{X} \sim \mathcal{D}^n$, we have $\phi(\boldsymbol{X}) - \phi(\mathcal{D}^n)$ is $\frac{1}{2\sqrt{n}}$-subgaussian.*

We also use the following bound on the mutual information for zCDP mechanisms:

Lemma C.5 (Bun and Steinke [2016])**.** *If $\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}$ is ρ-zCDP and $\boldsymbol{X} \sim \mathcal{D}^n$, then $I\left(\mathcal{M}(\boldsymbol{X}); \boldsymbol{X}\right) \leq \rho n$.*

Proof of Theorem 2.1. We follow the same analysis for proving Theorem B.4 where we add Gaussian noise with variance $\frac{1}{2\rho' n^2}$ to each query answer so that the algorithm $\mathcal{M}$ is $\rho' k$-zCDP, which (using Lemma C.5) makes the mutual information bound $B = \rho' kn$. We then use the sub-Gaussian parameter for statistical queries in Lemma C.4 to obtain the following bound from Theorem C.1.

$$\mathop{\mathbb{E}}_{\substack{\boldsymbol{X}\sim\mathcal{D}^n, \\ \phi\leftarrow\mathcal{M}(\boldsymbol{X})}} \left[\left(\phi(\boldsymbol{X}) - \phi(\mathcal{D})\right)^2 \right] \leq \frac{1}{4n} \cdot \min_{\lambda \in [0,1)} \left(\frac{2\rho' kn - \ln\left(1-\lambda\right)}{\lambda} \right).$$

We can then bound the population accuracy in (2) using Chebyshev's inequality to obtain the following high probability bound,

$$\Pr_{\substack{\boldsymbol{X}\sim\mathcal{D}^n, \\ \phi\leftarrow\mathcal{M}(\boldsymbol{X})}} \left[|\phi(\boldsymbol{X}) - \phi(\mathcal{D})| \geq \tau \right] \leq \frac{1}{4n\tau^2} \cdot \min_{\lambda \in [0,1)} \left(\frac{2\rho' kn - \ln\left(1-\lambda\right)}{\lambda} \right)$$

We then use the result of Corollary A.10 to obtain our accuracy bound. Thus we want to find $\rho' kn > 0$ that minimizes τ such that the following conditions are satisfied:

$$\tau \geq \sqrt{\frac{2}{n\beta} \cdot \min_{\lambda \in [0,1)} \left(\frac{2\rho' kn - \ln\left(1-\lambda\right)}{\lambda} \right)} \qquad \text{and} \qquad \tau \geq \frac{2}{n}\sqrt{\frac{\ln(4k/\beta)}{\rho'}}.$$

□

C.2 Proof of Theorem 2.2

We use the same monitor from Algorithm 1 in which there is a single dataset as input to the monitor and it outputs the query whose answer had largest error with the true query answer. We first need to show that the monitor has bounded mutual information as long as $\mathcal{M}$ does, which follows from mutual information being preserved under post-processing.

Lemma C.6. *If $I\left(\boldsymbol{X}; \mathcal{M}(\boldsymbol{X})\right) \leq B$ where $\boldsymbol{X} \sim \mathcal{D}^n$, then $I\left(\boldsymbol{X}; \mathcal{W}_{\mathcal{D}}[\mathcal{M}, \mathcal{A}](\boldsymbol{X})\right) \leq B$.*

We are now ready to prove our result.

Proof of Theorem 2.2. Recall that we add Gaussian noise with variance $\frac{1}{2\rho n^2}$ to each query answer so that the algorithm $\mathcal{M}$ is ρ-zCDP, which (using Lemma C.5 and the post-processing property of zCDP) makes the mutual information bound $B = \rho kn$. We then use the sub-Gaussian parameter for statistical queries in Lemma C.4 to obtain the following bound from Theorem C.1.

$$\begin{aligned} \mathop{\mathbb{E}}_{\substack{\boldsymbol{X}\sim\mathcal{D}^n, \\ q^*\sim\mathcal{W}_{\mathcal{D}}[\mathcal{M},\mathcal{A}](\boldsymbol{X})}} \left[\left(q^*(\boldsymbol{X}) - q^*(\mathcal{D})\right)^2 \right] &= \mathop{\mathbb{E}}_{\boldsymbol{X}\sim\mathcal{D}^n, \mathcal{M}, \mathcal{A}} \left[\max_{i \in [k]} \left\{ \left(q_i(\boldsymbol{X}) - q_i(\mathcal{D})\right)^2 \right\} \right] \\ &\leq \frac{1}{4n} \cdot \min_{\lambda \in [0,1)} \left(\frac{2\rho kn - \ln\left(1-\lambda\right)}{\lambda} \right) \end{aligned}$$

We then combine this result with (1) to get the statement of the theorem.

□

D Omitted Descriptions and Pseudocodes

D.1 Mechanisms considered

Here, we provide detailed descriptions of the mechanisms we consider for the analyst strategies in Section 3. We consider the following mechanisms to answer queries on a dataset containing n samples:

1. Naive Empirical Estimator: For each query, this mechanism provides the exact answer of the query when evaluated on the input dataset.
2. Gaussian mechanism: Input- noise scale σ. Adds Gaussian noise with standard deviation σ to each query answer provided by the Naive Empirical Estimator.
3. Thresholdout Dwork et al. [2015a]: Input- holdout size h, threshold T, noise scale σ.
4. Noisy Thresholdout: Input- holdout size h, threshold T, noise scale σ. Variant which also adds Laplace noise with scale σ to answers obtained via the training set split.
5. Sample splitting: Input- batch size B, maximum number of queries per batch m. Randomly partitions the input dataset into n/B batches of size at most B each, and provides the query answer on a batch for m queries before moving to the next batch.
6. Query-dependent Sample Splitting: Input- batch size B, holdout size H, threshold T, noise scale σ. Randomly partitions the input dataset into $(n-H)/B$ batches of size at most B each, and a holdout set D_H of size H. For a query, it starts by comparing the query answers on the first batch and D_H and, it releases the answer on the batch if the difference is below a "noisy" version of T, else it repeats the test for the next batch, and so on. For reference, we provide a pseudocode in Algorithm 3.

Algorithm 3 Query-dependent Sample Splitting

Require: batch size B, holdout size H, threshold T, noise scale σ
  Randomly partition dataset D into $(n-H)/B$ batches $D_i, i \in [1, \lceil (n-H)/B \rceil]$, of size at most B each, and a holdout set D_H of size H
  Initialize $\widehat{T} \leftarrow T + N(0, \sigma^2)$
  for each query q **do**
    Initialize $success = False$, and $i = 1$
    while $success$ is $False$ and $i \leq \lceil (n-H)/B \rceil$ **do**
      $\widehat{q}(D_i) = q(D_i) + N(0, \sigma^2)$
      if $\widehat{q}(D_i) - q(D_H) \leq \widehat{T}$ **then**
        Output $\widehat{q}(D_i)$
        $success = True$
      else
        $i \leftarrow i + 1$, $\widehat{T} \leftarrow T + N(0, \sigma^2)$
    if $success$ is $False$ **then**
      Output $q(D_H) + N(0, \sigma^2)$

D.2 Omitted Pseudocodes

Algorithm 4 A two-round analyst strategy for random data

Require: Mechanism $\mathcal{M}$ with a hidden dataset $X \in \{-1, 1\}^{n \times (k+1)}$
  for $j \in [k]$ **do**
    Define $q_j(x) = x(i) \cdot x(k)$. Give q_j to $\mathcal{M}$
    Receive $a_j \in [-1, 1]$ from $\mathcal{M}$
  Give q_k to $\mathcal{M}$ s.t. $q_k(\boldsymbol{x}) = sign\left(\sum_{i \in [k]} \left(\boldsymbol{x}(i) \cdot \ln \frac{1+a_i}{1-a_i}\right)\right)$, where $sign(y) = \begin{cases} 1 & \text{if } y \geq 0 \\ -1 & \text{otherwise} \end{cases}$
  Output $a_k \in [-1, 1]$ received from $\mathcal{M}$

Algorithm 5 A multi-round tracing analyst strategy

Require: Mechanism $\mathcal{M}$ with a hidden dataset $D_0 \in [N]^n$ sampled u.a.r., control set size c

Initialize $score_{-1,i} = 0$ for $i \in [N]$, and $I_{-1} = \emptyset$

Define a control dataset $C = \{0, 1, \ldots, c-1\}$. Initialize $score_{-1,C(i)} = 0$ for $i \in [c]$

for each round $j \in [k]$ **do**

  Select a *bias* $p_j \in [0, 1]$ u.a.r.

  Select a query $q_j : [N] \to \{0, 1\}$ where each entry is 1 independently with probability p_j

  Define query $q_{j,C} : [c] \to \{0, 1\}$ for the control set C analogous to q_j

  Let $q_j\big|_{D_j}$ be the restriction of q^j to the elements in D_j

  Give $q_j\big|_{D_j}$ to $\mathcal{M}$ (so $\mathcal{M}$ cannot access the values of q_j outside D_j)

  Receive $a_j \in [0, 1]$ from $\mathcal{M}$

  Update $score_{j,i} \leftarrow \begin{cases} score_{j-1,C(i)} + (a_j - p_j)(q_j(i) - p_j) & \text{if } i \in [N \setminus I_{j-1}] \\ score_{j-1,C(i)} & \text{if } i \in I_{j-1} \end{cases}$

  For each $i \in [c]$, update $score_{j,C(i)} \leftarrow score_{j-1,C(i)} + (a_j - p_j)(q_{j,C}(i) - p_j)$

  Let $I_j = \left\{ i \in [N] \text{ s.t. } score_{j,i} > \max\limits_{\ell \in [c]}(score_{j,C(\ell)}) \right\}$. $\mathcal{M}$ sets $D_{j+1} := D_j \setminus I_j$