Guaranteed Validity for Empirical Approaches to Adaptive Data Analysis

Ryan Rogers* **Aaron Roth†** **Adam Smith‡** **Nathan Srebro§** **Om Thakkar‡**

Blake Woodworth§

Abstract

We design a general framework for answering adaptive statistical queries that focuses on providing explicit confidence intervals along with point estimates. Prior work in this area has either focused on providing tight confidence intervals for specific analyses, or providing general worst-case bounds for point estimates. Unfortunately, as we observe, these worst-case bounds are loose in many settings — often not even beating simple baselines like sample splitting. Our main contribution is to design a framework for providing valid, instance-specific confidence intervals for point estimates that can be generated by heuristics. When paired with good heuristics, this method gives guarantees that are orders of magnitude better than the best worst-case bounds. We provide a Python library implementing our method.

1 Introduction

Many data analysis workflows are *adaptive*, that is, they re-use data over the course of a sequence of analyses, where the choice of analysis at any given stage depends on the results from previous stages. Such adaptive re-use of data is an important source of *overfitting* in machine learning and *false discovery* in the empirical sciences [Gelman and Loken, 2014]. Adaptive workflows arise, for example, when exploratory data analysis is mixed with confirmatory data analysis, when hold-out sets are re-used to search through large hyper-parameter spaces or to perform feature selection, and when datasets are repeatedly re-used within a research community.

A simple solution to this problem—that we can view as a naïve benchmark—is to simply not re-use data. More precisely, one could use *sample splitting*: partitioning the data set into k equal-sized pieces, and using a fresh piece of the data set for each of k adaptive interactions with the data. This allows us to treat each analysis as nonadaptive, and allows many quantities of interest to be accurately estimated with their empirical estimate, and paired with tight confidence intervals that come from classical statistics. This seemingly naive approach is wasteful in its use of data, however: the sample size needed to conduct a series of k adaptive analyses grows linearly with k.

A line of recent work [Dwork et al., 2015c,a,b, Russo and Zou, 2016, Bassily et al., 2016, Rogers et al., 2016, Feldman and Steinke, 2017a,b, Xu and Raginsky, 2017, Zrnic and Hardt, 2019, Mania et al., 2019] aims to improve on this baseline by using mechanisms which provide "noisy" answers to queries rather than exact empirical answers. The methods coming from this

*Previously at University of Pennsylvania. rrogers386@gmail.com

†University of Pennsylvania. aaroth@cis.upenn.edu

‡Boston University. {ads22, omthkkr}@bu.edu

§Toyota Technical Institute of Chicago. {nati, blake}@ttic.edu

Preprint. Under review.

line of work require that the sample size grow proportional to the *square root* of the number of adaptive analyses, dramatically beating the sample splitting baseline asymptotically. Unfortunately, the bounds proven in these papers—even when optimized—only beat the naïve baseline when both the data set size n, and the number of adaptive rounds k, are very large; see Figure 1 (left).

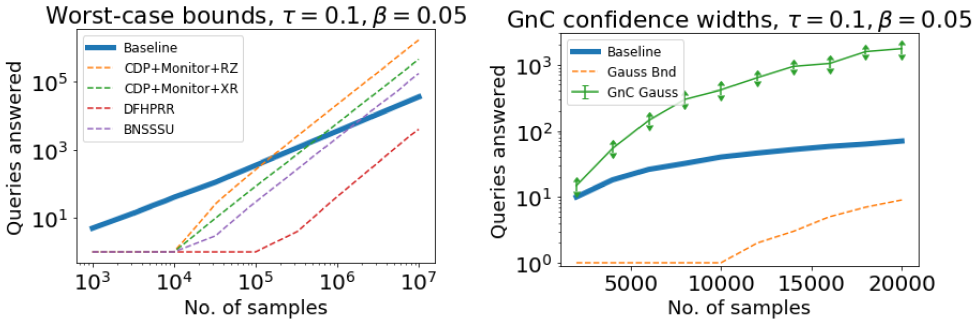

Figure 1: *Left:* Comparison of various worst-case bounds for the Gaussian mechanism with the sample splitting baseline. 'DFHPRR' and 'BNSSSU' refer to bounds given in prior work (Dwork et al. [2015d], Bassily et al. [2016]). The other lines plot improved worst-case bounds derived in this paper. (See Section 2 for the full model and parameter descriptions.) *Right:* Performance of Guess and Check with the Gaussian mechanism providing the guesses ("GnC Gauss") for a plausible query strategy (see Section 3.1), and comparing it with the best worst-case bounds for the Gaussian mechanism ("Gauss Bnd"), and the baseline.

The failure of these worst-case bounds to beat simple baselines in practice — despite their attractive asymptotics — has been a major obstacle to the practical adoption of techniques from this literature. There are two difficulties with directly improving this style of bounds. The first is that we are limited by what we can prove: mathematical analyses can often be loose by constants that are significant in practice. The more fundamental difficulty is that these bounds are guaranteed to hold even against a *worst-case* data analyst, who is adversarially attempting to find queries which over-fit the sample: one would naturally expect that when applied to a real workload of queries, such worst-case bounds would be extremely pessimistic. We address both difficulties in this paper.

Contributions In this paper, we move the emphasis from algorithms that provide point estimates to algorithms that explicitly manipulate and output *confidence intervals* based on the queries and answers so far, providing the analyst with both an estimated value and a measure of its actual accuracy. At a technical level, we have two types of contributions:

First, we give optimized worst-case bounds that carefully combine techniques from different pieces of prior work—plotted in Figure 1 (left). For certain mechanisms, our improved worst-case bounds are within small constant factors of optimal, in that we can come close to saturating their error bounds with a concrete, adversarial query strategy (Section 2). However, even these optimized bounds require extremely large sample sizes to improve over the naive sample splitting baseline, and their pessimism means they are often loose.

Our main result is the development of a simple framework called *Guess and Check*, that allows an analyst to pair *any method* for "guessing" point estimates and confidence interval widths for their adaptive queries, and then rigorously validate those guesses on an additional held-out data set. So long as the analyst mostly guesses correctly, this procedure can continue indefinitely. The main benefit of this framework is that it allows the analyst to guess confidence intervals whose guarantees *exceed what is guaranteed by the worst-case theory*, and still enjoy rigorous validity in the event that they pass the "check". This makes it possible to take advantage of the non-worst-case nature of natural query strategies, and avoid the need to "pay for" constants that seem difficult to remove from worst-case bounds. Our empirical evaluation demonstrates that our approach can improve on worst-case bounds by orders of magnitude, and that it improves on the naive baseline even for modest sample

sizes: see Figure 1 (right), and Section 3 for details. We also provide a Python library Rogers et al. [2019] containing an implementation of our Guess and Check framework.

Related Work Our "Guess and Check" (GnC) framework draws inspiration from the Thresholdout method of Dwork et al. [2015a], which uses a holdout set in a similar way. GnC has several key differences, which turn out to be crucial for practical performance: first, whereas the "guesses" in Thresholdout are simply the empirical query answers on a "training" portion of the dataset, we make use of other heuristic methods for generating guesses (including, in our experiments, Thresholdout itself) that empirically often seem to prevent overfitting to a substantially larger degree than their worst-case guarantees suggest. Second, we make confidence-intervals first-order objects: whereas the "guesses" supplied to Thresholdout are simply point estimates, the "guesses" supplied to GnC are point estimates along with confidence intervals. Finally, we use a more sophisticated analysis to track the number of bits leaked from the holdout, which lets us give tighter confidence intervals and avoids the need to a priori set an upper bound on the number of times the holdout is used. Gossmann et al. [2018] use a version of Thresholdout to get worst-case accuracy guarantees for values of the area under the receiver operating charateristic curve (AUC) for adaptively obtained queries. However, apart from being limited to binary classification tasks and the dataset being used only to obtain AUC values, their bounds require "unrealistically large" dataset sizes. Our results are complementary to theirs; by using appropriate concentration inequalities, GnC could also be used to provide confidence intervals for AUC values. Their technique could be used to provide the "guesses" to GnC.

Our improved worst-case bounds combine a number of techniques from the existing literature: namely the information theoretic arguments of Russo and Zou [2016], Xu and Raginsky [2017] together with the "monitor" argument of Bassily et al. [2016], and a more refined accounting for the properties of specific mechanisms using *concentrated differential privacy* (Dwork and Rothblum [2016], Bun and Steinke [2016b]).

Feldman and Steinke [2017a,b] give worst-case bounds that improve with the variance of the queries asked. We show in Section 3.1 how our techniques can also be used to give tighter bounds when the empirical query variance is small.

Mania et al. [2019] give an improved union bound for queries that have high overlap, that can be used to improve bounds for adaptively validating similar models, in combination with description length bounds. Zrnic and Hardt [2019] take a different approach to going beyond worst-case bounds in adaptive data analysis, by proving bounds that apply to data analysts that may only be adaptive in a constrained way. A difficulty with this approach in practice is that it is limited to analysts whose properties can be inspected and verified — but provides a potential explanation why worst-case bounds are not observed to be tight in real settings. Our approach is responsive to the degree to which the analyst actually overfits, and so will also provide relatively tight confidence intervals if the analyst satisfies the assumptions of Zrnic and Hardt [2019].

1.1 Preliminaries

As in previous work, we assume that there is a data set $X = (x_1, \cdots, x_n) \sim \mathcal{D}^n$ drawn i.i.d. from an unknown distribution $\mathcal{D}$ over a universe $\mathcal{X}$. This data set is the input to a mechanism $\mathcal{M}$ that also receives a sequence of queries $\phi_1, \phi_2, ...$ from an analyst $\mathcal{A}$ and outputs, for each one, an answer. Each ϕ_i is a *statistical query*, defined by a bounded function $\phi_i : \mathcal{X} \to [0, 1]$. We denote the expectation of a statistical query ϕ over the data distribution by $\phi(\mathcal{D}) = \mathbb{E}_{x\sim\mathcal{D}}[\phi(x)]$, and the empirical average on a dataset by $\phi(X) = \frac{1}{n}\sum_{i=1}^{n}\phi(x_i)$.

The mechanism's goal is to give estimates of $\phi_i(\mathcal{D})$ for query ϕ_i on the unknown $\mathcal{D}$. Previous work looked at analysts that produce a single point estimate a_i, and measured error based on the distances $|a_i - \phi_i(\mathcal{D})|$. As mentioned above, we propose a shift in focus: we ask mechanisms to produce a confidence interval specified by a point estimate a_i and width τ_i. The answer (a_i, τ_i) is *correct for* ϕ_i *on* $\mathcal{D}$ if $\phi_i(\mathcal{D}) \in (a_i - \tau_i, a_i + \tau_i)$. (Note that the data play no role in the definition of correctness—we measure only population accuracy.)

An interaction between randomized algorithms $\mathcal{M}$ and $\mathcal{A}$ on data set $X \in \mathcal{X}^n$ (denoted $\mathcal{M}(X) \rightleftharpoons \mathcal{A}$) consists of an unbounded number of query-answer rounds: at round i, $\mathcal{A}$ sends

ϕ_i, and $\mathcal{M}(X)$ replies with (a_i, τ_i). $\mathcal{M}$ receives X as input. $\mathcal{A}$ receives no direct input, but may select queries *adaptively*, based on the answers in previous rounds. The interaction ends when either the mechanism or the analyst stops. We say that the mechanism provides simultaneous coverage if, with high probability, *all* its answers are correct:

Definition 1.1 (Simultaneous Coverage)**.** Given $\beta \in (0,1)$, we say that $\mathcal{M}$ has *simultaneous coverage* $1-\beta$ if, for all $n \in \mathbb{N}$, all distributions $\mathcal{D}$ on $\mathcal{X}$ and all randomized algorithms $\mathcal{A}$,

$$\Pr_{\substack{X\sim\mathcal{D}^n \\ \{(\phi_i,a_i,\tau_i)\}_{i=1}^k \leftarrow (\mathcal{M}(X)\rightleftharpoons\mathcal{A})}} [\forall i \in [k] : \phi_i(\mathcal{D}) \in (a_i - \tau_i, a_i + \tau_i)] \geq 1-\beta\,.$$

We denote by k the (possibly random) number of rounds in a given interaction.

Definition 1.2 (Accuracy)**.** We say $\mathcal{M}$ is (τ,β)-*accurate*, if $\mathcal{M}$ has simultaneous coverage $1-\beta$ and its interval widths satisfy $\max_{i\in[k]} \tau_i \leq \tau$ with probability 1.

2 Confidence intervals from worst-case bounds

Our emphasis on explicit confidence intervals led us to derive worst-case bounds that are as tight as possible given the techniques in the literature. We discuss the Gaussian mechanism here, and defer the application to Thresholdout in Appendix B.4, and provide a pseudocode for Thresholdout in Algorithm 4.

The Gaussian mechanism is defined to be an algorithm that, given input dataset $X \sim \mathcal{D}^n$ and a query $\phi : \mathcal{X} \to [0,1]$, reports an answer $a = \phi(X) + N\left(0, \frac{1}{2n^2\rho}\right)$, where $\rho > 0$ is a parameter. It has existing analyses for simultaneous coverage (see Dwork et al. [2015d], Bassily et al. [2016]) — but these analyses involve large, sub-optimal constants. Here, we provide an improved worst-case analysis by carefully combining existing techniques. We use results from Bun and Steinke [2016a] to bound the mutual information of the Gaussian mechanism. We then apply an argument similar to that of Russo and Zou [2016] to bound the bias of the empirical average of a statistical query selected as a function of the perturbed outputs. Finally, we use Chebyshev's inequality, and the monitor argument from Bassily et al. [2016] to obtain high probability accuracy bound. Figure 1 (left) shows the improvement in the number of queries that can be answered with the Gaussian mechanism with (τ,β)-accuracy for $\tau = 0.1, \beta = 0.05$. Our guarantee is stated below, with its proof deferred to Appendix C.2.

Theorem 2.1. *Given input $X \sim \mathcal{D}^n$, confidence parameter β, and parameter ρ, the Gaussian mechanism is (τ,β)-accurate, where* $\tau = \max\left\{\sqrt{\frac{2}{n\beta}\cdot\min_{\lambda\in[0,1)}\left(\frac{2\rho kn - \ln(1-\lambda)}{\lambda}\right)}, \frac{2}{n}\sqrt{\frac{\ln(4k/\beta)}{\rho}}\right\}$.

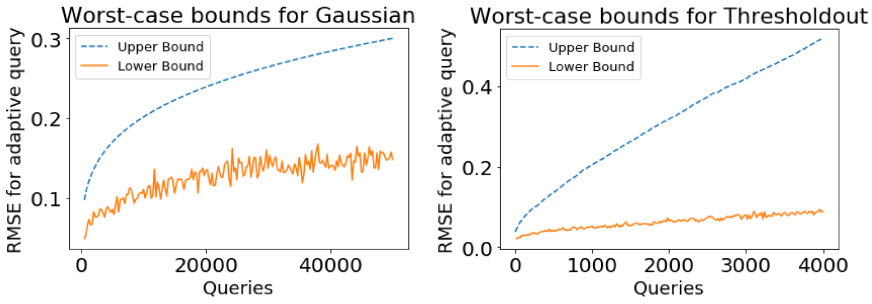

Figure 2: Worst-case upper (proven) and lower bounds (realized via the single-adaptive query strategy) on RMSE with $n = 5,000$ for Gaussian (left) and Thresholdout (right).

We now consider the extent to which our analyses are improvable for worst-case queries to the Gaussian and the Thresholdout mechanisms. To do this, we derive the worst query strategy in a particular restricted regime. We call it the "single-adaptive query strategy", and show that it maximizes the root mean squared error (RMSE) amongst all single query strategies under the assumption that each sample in the data set is drawn u.a.r. from $\{-1,1\}^{k+1}$, and

the strategy is given knowledge of the empirical correlations of each of the first k features with the $(k+1)$st feature (which can be obtained e.g. with k non-adaptive queries asked prior to the adaptive query). We provide a pseudocode for the strategy in Algorithm 5, and prove that our single adaptive query results in maximum error, in Appendix C.1. To make the bounds comparable, we translate our worst-case confidence upper bounds for both the mechanisms to RMSE bounds in Theorem C.8 and Theorem B.10. Figure 2 shows the difference between our best upper bound and the realized RMSE (averaged over 100 executions) for the two mechanisms using $n = 5,000$ and various values of k. (For the Gaussian, we set ρ separately for each k, to minimize the upper bound.) On the left, we see that the two bounds for the Gaussian mechanism are within a factor of 2.5, even for $k = 50,000$ queries. Our bounds are thus reasonably tight in one important setting. For Thresholdout (right side), however, we see a large gap between the bounds which grows with k, even for our best query strategy[5]. This result points to the promise for empirically-based confidence intervals for complex mechanisms that are harder to analyze.

3 The Guess and Check Framework

In light of the inadequacy of worst-case bounds, we here present our Guess and Check (GnC) framework which can go beyond the worst case. It takes as inputs *guesses* for both the point estimate of a query, and a confidence interval width. If GnC can validate a guess, it releases the guess. Otherwise, at the cost of widening the confidence intervals provided for future guesses, it provides the guessed confidence width along with a point estimate for the query using the holdout set such that the guessed width is valid for the estimate. An instance of GnC, $\mathcal{M}$, takes as input a data set X, desired confidence level $1-\beta$, and a mechanism $\mathcal{M}_g$ which operates on inputs of size $n_g < n$. $\mathcal{M}$ randomly splits X into two, giving one part X_g to $\mathcal{M}_g$, and reserving the rest as a holdout X_h. For each query ϕ_i, mechanism $\mathcal{M}_g$ uses X_g to make a "guess" $(a_{g,i}, \tau_i)$ to $\mathcal{M}$, for which $\mathcal{M}$ conducts a validity check. If the check succeeds, then $\mathcal{M}$ releases the guess as is, otherwise $\mathcal{M}$ uses the holdout X_h to provide a response containing a discretized answer that has τ_i as a valid confidence interval. This is closely related to Thresholdout. However, an important distinction is that the width of the target confidence interval, rather than just a point estimate, is provided as a guess. Moreover, the guesses themselves can be made by non-trivial algorithms. We provide pseudocode for GnC in Algorithm 1, and block schematic of how a query is answered by GnC in Figure 3.

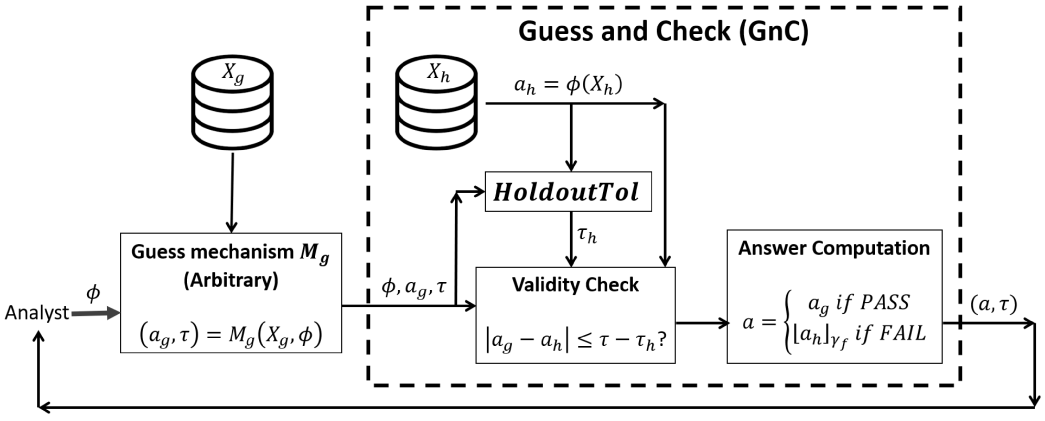

Figure 3: A schematic of how query ϕ is answered via our Guess and Check (GnC) framework. Dataset X_g is the guess set randomly partitioned by GnC. The dotted box represents computations that are previleged, and are not accessible to the analyst.

[5]We tweak the adaptive query in the single-adaptive query strategy to result in maximum error for Thresholdout. We also tried "tracing" attack strategies (adapted from the fingerprinting lower bounds of Bun et al. [2014], Hardt and Ullman [2014], Steinke and Ullman [2015]) that contained multiple adaptive queries, but gave similar results.

Algorithm 1 Guess and Check

Require: Data $X \in \mathcal{X}^n$, confidence parameter β, mechanism $\mathcal{M}_g$ with inputs of size $n_g < n$
  Define $c_j \leftarrow \frac{6}{\pi^2(j+1)^2}$ for $j \geq 0$ //only needs to satisfy $\sum_{j\geq 0} c_j \leq 1$
  Randomly partition X into a guess set X_g of size n_g, and a holdout X_h of size $n_h = n - n_g$
  Initialize $f \leftarrow 0$
  for $i = 1$ to ∞ **do**
    if $f > 0$ **then** $\nu_{i,f,\gamma_1^f} \leftarrow \binom{i-1}{f} \prod_{j\in[f]} \left(\frac{1}{\gamma_j}\right)$ **else** $\nu_{i,f,\gamma_1^f} \leftarrow 1$ //# possible transcripts
    $\beta_i \leftarrow (\beta \cdot c_{i-1} \cdot c_f)/\nu_{i,f,\gamma_1^f}$
    For query ϕ_i, receive guess $(a_{g,i}, \tau_i) \leftarrow \mathcal{M}_g(X_g, \phi_i)$
    Let holdout answer $a_{h,i} \leftarrow \phi_i(X_h)$
    Let $\tau_h \leftarrow HoldoutTol(\beta_i, a_{g,i}, a_{h,i})$ //$HoldoutTol$ returns a valid tolerance for $a_{h,i}$
    if $|a_{g,i} - a_{h,i}| \leq \tau_i - \tau_h$ **then**
      Output $(a_{g,i}, \tau_i)$
    else
      $f \leftarrow f + 1$
      Let $\gamma_f \leftarrow \max_{[0,\tau_i)} \gamma$ s.t. $2e^{-2(\tau_i-\gamma)^2 n_h} \leq \beta_i$ //max. discretization param. with validity
      if $\gamma_f > 0$ **then**
        Output $\left(\lfloor a_{h,i} \rfloor_{\gamma_f}, \tau_i\right)$, where $\lfloor y \rfloor_\gamma$ denotes y discretized to multiples of γ
      else
        Output $\perp$
        break //Terminate **for** loop

We provide coverage guarantees for GnC without any restrictions on the guess mechanism. To get the guarantee, we first show that for query ϕ_i, if function $HoldoutTol$ returns a $(1-\beta_i)$-confidence interval τ_h for holdout answer $a_{h,i}$, and GnC's output is the guess $(a_{g,i}, \tau_i)$, then τ_i is a $(1-\beta_i)$-confidence interval for $a_{g,i}$. We can get a simple definition for $HoldoutTol$ (formally stated in Appendix C.3.1), but we provide a slightly sophisticated variant below that uses the guess and holdout answers to get better tolerances, especially under *low-variance* queries. We defer the proof of Lemma 3.1 to Appendix C.3.2.

Lemma 3.1. *If the function HoldoutTol in GnC (Algorithm 1) is defined as*

$$HoldoutTol(\beta', a_g, a_h) = \begin{cases} \min \left\{ \tau' \in (0,\tau) : \begin{array}{c} \left(\frac{1+\mu(e^\ell-1)}{e^{\ell(\mu+\tau')}}\right)^n \leq \frac{\beta}{2}, \text{ where} \\ \ell \text{ solves } \frac{\mu e^\ell}{1+\mu(e^\ell+1)} = \mu + \tau' \end{array} \right\} & \text{if } a_g > a_h\,, \\ \min \left\{ \tau' \in (0,\tau) : \begin{array}{c} \left(\frac{1+\mu'(e^\ell-1)}{e^{\ell(\mu'+\tau')}}\right)^n \leq \frac{\beta}{2}, \text{ where} \\ \mu' = 1-\mu, \text{ and } \ell \text{ solves} \\ \frac{\mu' e^\ell}{1+\mu'(e^\ell+1)} = \mu' + \tau' \end{array} \right\} & \text{o.w.} \end{cases}$$

then for each query ϕ_i s.t. GnC's output is $(a_{g,i}, \tau_i)$, we have $\Pr\left(|a_{g,i} - \phi_i(\mathcal{D})| > \tau_i\right) \leq \beta_i$.

Next, if failure f occurs within GnC for query ϕ_i, by applying a Chernoff bound we get that γ_f is the maximum possible discretization parameter s.t. τ_i is a $(1-\beta_i)$-confidence interval for the discretized holdout answer $\lfloor a_{h,i} \rfloor_{\gamma_f}$. Finally, we get a simultaneous coverage guarantee for GnC by a union bound over the error probabilities of the validity over all possible transcripts between the mechanism and any analyst $\mathcal{A}$ with adaptive queries $\{\phi_1, \ldots, \phi_k\}$. The guarantee is stated below, with its proof deferred to Appendix C.3.1.

Theorem 3.2. *The Guess and Check mechanism (Algorithm 1), with inputs data set $X \sim \mathcal{D}^n$, confidence parameter β, and a mechanism $\mathcal{M}_g$ that, using inputs of size $n_g < n$, provides responses ("guesses") of the form $(a_{g,i}, \tau_i)$ for query ϕ_i, has simultaneous coverage $1 - \beta$.*

3.1 Experimental evaluation

Now, we provide details of our empirical evaluation of the Guess and Check framework. In our experiments, we use two mechanisms, namely the Gaussian mechanism and Thresholdout,

for providing guesses in GnC. For brevity, we refer to the overall mechanism as GnC Gauss when the Gaussian is used to provide guesses, and GnC Thresh when Thresholdout is used.

Strategy for performance evaluation: Some mechanisms evaluated in our experiments provide worst-case bounds, whereas the performance of others is instance-dependent and relies on the amount of adaptivity present in the querying strategy. To highlight the advantages of instance-dependent bounds, we design a query strategy called the quadratic-adaptive query strategy. It contains both adaptive and non-adaptive queries, where the adaptive queries become more sparsely distributed with time. "Hard" adaptive queries $\phi_i, i > 1$, are asked when i is a perfect square. They are computed using the answers to all the non-adaptive queries asked in prior rounds, using a strategy similar to that used in Figure 2. We provide pseudocode for the strategy in Algorithm 5.

Experimental Setup: We run the quadratic-adaptive strategy for up to $40,000$ queries. We tune the hyperparameters of each mechanism to optimize for this query strategy. We fix a confidence parameter β and set a target upper bound τ on the maximum allowable error we can tolerate, given our confidence bound. We evaluate each mechanism by the number of queries it can empirically answer with a confidence width of τ for our query strategy while providing a simultaneous coverage of $1-\beta$: i.e. the largest number of queries it can answer while providing (τ, β)-accuracy. We plot the average and standard deviation of the number of queries k answered before it exceeds its target error bound in 20 independent runs over the sampled data and the mechanism's randomness. When we plot the actual realized error for any mechanism, we denote it by dotted lines, whereas the provably valid error bounds resulting from the confidence intervals produced by GnC are denoted by solid lines. Note that the empirical error denoted by dotted lines is not actually possible to know without access to the distribution, and is plotted just to visualize the tightness of the provable confidence intervals. We compare to two simple baselines: sample splitting, and answer discretization: the better of these two is plotted as the thick solid line. For comparison, the best worst-case bounds for the Gaussian mechanism (Theorem 2.1) are shown as dashed lines. Note that we improve by roughly two orders of magnitude compared to the tightest bounds for the Gaussian. We improve over the baseline at data set sizes $n \geq 2,000$.

Boost in performance for low-variance queries: Since all the queries we construct take binary values on a sample $x \in \mathcal{X}$, the variance of query ϕ_i is given by $var(\phi_i) = \phi_i(\mathcal{D})(1-\phi_i(\mathcal{D}))$, as $\phi_i(\mathcal{D}) = \Pr(\phi_i(x) = 1)$. Now, $var(\phi_i)$ is maximized when $\phi_i(\mathcal{D}) = 0.5$. Hence, informally we call ϕ_i as low-variance if either $\phi_i(\mathcal{D}) \ll 0.5$, or $\phi_i(\mathcal{D}) \gg 0.5$. We want to be able to adaptively give tighter confidence intervals for low-variance queries (as e.g., the worst-case bounds of Feldman and Steinke [2017a,b] are able to). For instance, in Figure 4 (left), we show that in the presence of low-variance queries, using Lemma 3.1 for $HoldoutTol$ (plot labelled "GnC Check:MGF") results in a significantly better performance for GnC Gauss as compared to using Lemma C.9 (plot labelled "GnC Check:Chern"). We fix $\tau, \beta = 0.05$, and set $\phi_i(\mathcal{D}) = 0.9$ for $i \geq 1$. We can see that as the dataset size grows, using Lemma 3.1 provides an improvement of almost 2 orders of magnitude in terms of the number of queries k answered. This is due to Lemma 3.1 providing tighter holdout tolerances τ_h for low-variance queries (with guesses close to 0 or 1), compared to those obtained via Lemma C.9 (agnostic to the query variance). Thus, we use Lemma 3.1 for $HoldoutTol$ in all experiments with GnC below. Note that the worst-case bounds for the Gaussian don't promise a coverage of $1-\beta$ even for $k = 1$ in the considered parameter ranges. This is representative of a general phenomenon: switching to GnC-based bounds instead of worst-case bounds is often the difference between obtaining useful vs. vacuous guarantees.

Performance at high confidence levels: The bounds we prove for the Gaussian mechanism, which are the best known worst-case bounds for the considered sample size regime, have a substantially sub-optimal dependence on the coverage parameter $\beta : \sqrt{1/\beta}$. On the other hand, sample splitting (and the bounds from Dwork et al. [2015d], Bassily et al. [2016] which are asymptotically optimal but vacuous at small sample sizes) have a much better dependence on $\beta : \ln(1/2\beta)$. Since the coverage bounds of GnC are strategy-dependent, the dependence of τ on β is not clear a priori. In Figure 4 (right), we show the performance of GnC Gauss (labelled "GnC") when $\beta \in \{0.05, 0.005\}$. We see that reducing β by a factor of 10 has a negligible effect on GnC's performance. Note that this is the case even though the guesses are provided by the Gaussian, for which we do not have non-vacuous bounds

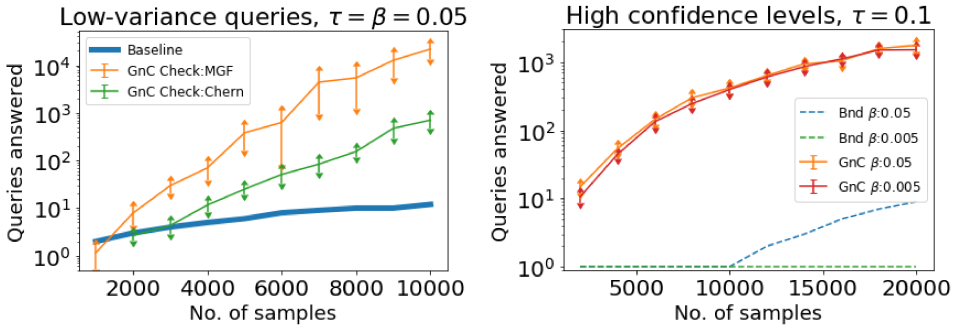

Figure 4: *Left:* Gain in performance for GnC Gauss by using Lemma 3.1 for *HoldoutTol* ("GnC Check:MGF"), as compared to using Lemma C.9 ("GnC Check:Chern"). *Right:* Performance of GnC Gauss ("GnC"), and the best Gaussian bounds ("Bnd"), for $\beta \in \{0.05, 0.005\}$.

with a mild dependence on β in the considered parameter range (see the worst-case bounds, plotted as "Bnd") — even though we might conjecture that such bounds exist. This gives an illustration of how GnC can correct deficiencies in our worst-case theory: conjectured improvements to the theory can be made rigorous with GnC's certified confidence intervals.

GnC with different guess mechanisms: GnC is designed to be modular, enabling it to take advantage of arbitrarily complex mechanisms to make guesses. Here, we compare the performance of two such mechanisms for making guesses, namely the Gaussian mechanism, and Thresholdout. In Figure 5 (left), we first plot the number of queries answered by the Gaussian ("Gauss Emp") and Thresholdout ("Thresh Emp") mechanisms, respectively, until the maximum empirical error of the query answers exceeds $\tau = 0.1$. It is evident that Thresholdout, which uses an internal holdout set to answer queries that likely overfit to its training set, provides better performance than the Gaussian mechanism. In fact, we see that for $n > 5000$, while Thresholdout is always able to answer $40,000$ queries (the maximum number of queries we tried in our experiments), the Gaussian mechanism isn't able to do so even for the largest data set size we consider. Note that the "empirical" plots are generally un-knowable in practice, since we do not have access to the underlying distributions. But they serve as upper bounds for the best performance a mechanism can provide.

Next, we fix $\beta = 0.05$, and plot the performance of GnC Gauss and GnC Thresh. We see that even though GnC Thresh has noticeably higher variance, it provides performance that is close to two orders of magnitude larger than GnC Gauss when $n \geq 8000$. Moreover, for $n \geq 8000$, it is interesting to see GnC Thresh gaurantees (τ, β)-accuracy for our strategy while consistently beating even the empirical performance of the Gaussian. We note that the best bounds for both the Gaussian and Thresholdout mechanisms alone (not used as part of GnC) do not provide any non-trivial guarantees in the considered parameter ranges.

Responsive widths that track the empirical error: The GnC framework is designed to certify guesses which represent both a point estimate and a desired confidence interval width for each query. Rather than having fixed confidence interval widths, this framework also provides the flexibility to incorporate guess mechanisms that provide increased interval widths as failures accumulate within GnC. This allows GnC to be able to re-use the holdout set in perpetuity, and answer an infinite number of queries (albeit with confidence widths that might grow to be vacuous). In Figure 5 (right), we fix $n = 30000, \beta = 0.05, \tau_1 = 0.06$, and plot the performance of GnC Gauss such that the guessed confidence width $\tau_{i+1} = \min(1.4\tau_i, 0.17)$ if the "check" for query ϕ_i results in a failure, otherwise $\tau_{i+1} = \tau_i$. For comparison, we also plot the actual maximum empirical error encountered by the answers provided by GnC ("GnC Gauss Emp"). It corresponds to the maximum empirical error of the answers of the Gaussian mechanism that is used as a guess mechanism within GnC, unless the check for a query results in a failure (which occurs 4 times in 40000 queries), in which case the error corresponds to the discretized answer on the holdout. We see that the statistically valid

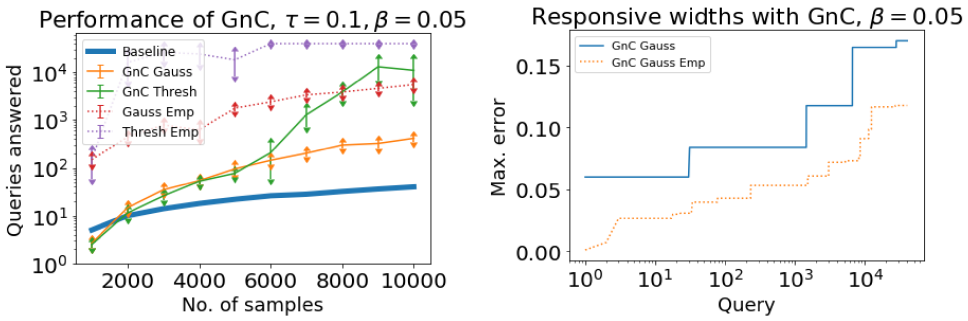

Figure 5: *Left:* Performance of GnC with Gaussian ("GnC Gauss"), and Thresholdout ("GnC Thresh") mechanisms, together with their empirical error. *Right:* The accuracy of GnC with Gaussian guesses provides "responsive" confidence interval widths that closely track the empirical error incurred by the guesses of the Gaussian mechanism ("GnC Gauss Emp").

accuracy guaranteed by GnC is "responsive" to the empirical error of the realized answers produced by the GnC, and is almost always within a factor of 2 of the actual error.

Acknowledgements

The authors would like to thank Omer Tamuz for his helpful comments regarding a conjecture that existed in a prior version of this work. A.R. acknowledges support in part by a grant from the Sloan Foundation, and NSF grants AF-1763314 and CNS-1253345. A.S. and O.T. were supported in part by a grant from the Sloan foundation and NSF grants IIS-1447700 and AF-1763314. B.W. is supported by the NSF GRFP (award No. 1754881). This work was done in part while R.R., A.S., and O.T. were visiting the Simons Institute for the Theory of Computing.

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.

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, 2016a. Springer Berlin Heidelberg. ISBN 978-3-662-53641-4.

Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. *CoRR*, abs/1605.02065, 2016b. URL http://arxiv.org/abs/1605.02065.

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.

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. doi: 10.1126/science.aaa9375. URL http://www.sciencemag.org/content/349/6248/636.abstract.

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*, pages 117–126. ACM, 2015c.

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, 2015d. 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, 2017a. URL http://proceedings.mlr.press/v65/feldman17a.html.

Vitaly Feldman and Thomas Steinke. Calibrating noise to variance in adaptive data analysis. *CoRR*, abs/1712.07196, 2017b. URL http://arxiv.org/abs/1712.07196.

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

Alexej Gossmann, Aria Pezeshk, and Berkman Sahiner. Test data reuse for evaluation of adaptive machine learning algorithms: over-fitting to a fixed'test'dataset and a potential solution. In *Medical Imaging 2018: Image Perception, Observer Performance, and Technology Assessment*, volume 10577, page 105770K. International Society for Optics and Photonics, 2018.

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

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.

Horia Mania, John Miller, Ludwig Schmidt, Moritz Hardt, and Benjamin Recht. Model similarity mitigates test set overuse. *arXiv preprint arXiv:1905.12580*, 2019.

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.

Ryan Rogers, Aaron Roth, Adam Smith, Nathan Srebro, Om Thakkar, and Blake Woodworth. Repository for empirical adaptive data analysis. https://github.com/omthkkr/empirical_adaptive_data_analysis, 2019.

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.

Tijana Zrnic and Moritz Hardt. Natural analysts in adaptive data analysis. *arXiv preprint arXiv:1901.11143*, 2019.

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.

$$\begin{aligned}\Pr\left[|\phi(\mathcal{D}) - a| \geq \tau\right] &\leq \Pr\left[|\phi(\mathcal{D}) - \phi(X)| + |\phi(X) - a| \geq \tau\right] \\ &\leq \underbrace{\Pr\left[|\phi(\mathcal{D}) - \phi(X)| \geq \tau/2\right]}_{\text{Population Accuracy}} + \underbrace{\Pr\left[|\phi(X) - a| \geq \tau/2\right]}_{\text{Sample Accuracy}} \end{aligned} \tag{1}$$

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.

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

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

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., 2015d,a, Bassily et al., 2016, Rogers et al., 2016]. In order to define these privacy notions, we define two datasets $X = (x_1, \cdots, x_n), X' = (x'_1, \cdots, x'_n) \in \mathcal{X}^n$ 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.2 (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 X and X' and each outcome $S \subseteq \mathcal{Y}$, we have $\Pr[\mathcal{M}(X) \in S] \leq e^{\epsilon} \Pr[\mathcal{M}(X') \in S] + \delta$. 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.3 (Privacy Loss)**.** Let $\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}$ be a randomized algorithm. For neighboring datasets $X, X' \in \mathcal{X}^n$, let $Z(y) = \ln\left(\frac{\Pr[\mathcal{M}(X)=y]}{\Pr[\mathcal{M}(X')=y]}\right)$. We then define the privacy loss variable $\text{PrivLoss}(\mathcal{M}(X)||\mathcal{M}(X'))$ to have the same distribution as $Z(\mathcal{M}(X))$.

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 [2016a] (Note that Dwork and Rothblum [2016] initially gave a definition of CDP which Bun and Steinke [2016a] then modified).

Definition A.4 (zCDP)**.** An algorithm $\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}$ is ρ-zero concentrated differentially private (zCDP), if for all neighboring datasets $X, X' \in \mathcal{X}^n$ and all $\lambda > 0$ we have

$$\mathbb{E}\left[\exp\left(\lambda\left(\text{PrivLoss}\left(\mathcal{M}(X)||\mathcal{M}(X')\right) - \rho\right)\right)\right] \leq e^{\lambda^2 \rho}.$$

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

Theorem A.5. *Let $\phi : \mathcal{X} \to [0,1]$ be a statistical query and $X \in \mathcal{X}^n$. The Laplace mechanism $\mathcal{M}_{\text{Lap}} : \mathcal{X}^n \to \mathbb{R}$ is the following $\mathcal{M}_{\text{Lap}}(X) = \frac{1}{n}\sum_{i=1}^n \phi(x_i) + \text{Lap}\left(\frac{1}{\epsilon n}\right)$, which is ϵ-DP. Further, the Gaussian mechanism $\mathcal{M}_{\text{Gauss}} : \mathcal{X}^n \to \mathbb{R}$ is the following $\mathcal{M}_{\text{Gauss}}(X) = \frac{1}{n}\sum_{i=1}^n \phi(x_i) + N\left(0, \frac{1}{2\rho n^2}\right)$, which is ρ-zCDP.*

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

Theorem A.6 (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*

$$\epsilon = \left(\frac{e^{\epsilon'} - 1}{e^{\epsilon'} + 1}\right)\epsilon' k + \epsilon'\sqrt{2k\ln(1/\delta)} \tag{3}$$

We will also use the following results from zCDP.

Theorem A.7 (Bun and Steinke [2016a])**.** *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 X and output $\mathcal{M}(X)$.

Definition A.8 (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 $I(X;Y) = \mathbb{E}[Z(X,Y)]$, where the expectation is taken over the joint distribution of (X,Y).

A.3 Monitor Argument

For the population accuracy term in (1), 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 $|\phi(X) - \phi(\mathcal{D})|$ for the two stage interaction where ϕ is chosen by $\mathcal{A}$ based on outcome $\mathcal{M}(X)$. We present the monitor $\mathcal{W}_{\mathcal{D}}[\mathcal{M}, \mathcal{A}]$ in Algorithm 2.

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

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

We simulate $\mathcal{M}(X)$ and $\mathcal{A}$ interacting. We write $\phi_1, \cdots, \phi_k \in \mathcal{Q}_{SQ}$ 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 $j^* = \operatorname{argmax}_{j \in [k]} |\phi_j(\mathcal{D}) - a_j|$.

Ensure: ϕ_{j^*}

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

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

Proof. From the monitor in Algorithm 2 and the fact that $\mathcal{M}$ is closed under post-processing, we have

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

□

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, provide a comparison of our bounds for the Gaussian mechanism (Theorem 2.1) with prior work.

B.1 Confidence Bounds from 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 $\phi \leftarrow \mathcal{M}(X)$ 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 $\Pr\left[|\phi(\mathcal{D}) - \phi(X)| \geq \tau\right] \leq \beta$.*

We pair this together with the accuracy from either the Gaussian mechanism or the Laplace mechanism along with Corollary A.9 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*

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

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

$$\begin{aligned} \min \quad & \tau \\ s.t. \quad & \tau \geq 2\sqrt{\frac{48}{n}\ln(8/\beta)} \\ & \tau \geq \frac{1}{n}\sqrt{\frac{1}{\rho}\ln(4k/\beta)} \\ & \tau \geq 8\rho k + \sqrt{256\rho k\left(\ln\sqrt{\pi\rho k} + \frac{\ln\frac{16}{\beta}}{\tau}\right)} \\ for \quad & \rho > 0 \end{aligned}$$

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

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

$$\Pr\left[|Y_i| \geq \ln(1/\beta)b\right] \leq \beta \implies \Pr\left[\exists i \in [k] \text{ s.t. } |Y_i| \geq \ln(k/\beta)b\right] \leq \beta. \tag{4}$$

Proof of Theorem B.2. We will focus on the Laplace mechanism part first, so that we add $\text{Lap}\left(\frac{1}{n\epsilon'}\right)$ 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 (1). To bound the sample accuracy, we then use (4) 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:

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

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.7 to get a composed DP algorithm with parameters in terms of ρ, and the accuracy guarantee in (2). □

B.2 Confidence Bounds from 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 2 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:*

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

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

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

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

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

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 $\phi : \mathcal{X} \to [0,1]$ 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 $\vec{X} = (X^{(1)}, \cdots, X^{(s)}) \sim (\mathcal{D}^n)^s$ then*

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

We then define what we will call the *extended monitor* in Algorithm 3.

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.*

Algorithm 3 Extended Monitor $\mathcal{W}_{\mathcal{D}}[\mathcal{M}, \mathcal{A}](\vec{X})$

Require: $\vec{X} = (X^{(1)}, \cdots, X^{(s)}) \in (\mathcal{X}^n)^s$
  for $t \in [s]$ **do**
    We simulate $\mathcal{M}(X^{(t)})$ and $\mathcal{A}$ interacting. We write $\phi_{t,1}, \cdots, \phi_{t,k} \in \mathcal{Q}_{SQ}$ 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 $(j^*, t^*) = \text{argmax}_{j \in [k], t \in [s]} \left| \phi_{t,j}(\mathcal{D}) - a_{t,j} \right|$.
  if $a_{t^*,j^*} - \phi_{t^*,j^*}(\mathcal{D}) \geq 0$ **then** $\phi^* \leftarrow \phi_{t^*,j^*}$
  else $\phi^* \leftarrow -\phi_{t^*,j^*}$
Ensure: (ϕ^*, t^*)

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

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

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*

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

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

$$\tau \left(1 - (1 - \beta)^s\right) - \left(\tau' + 2s\beta'\right) > e^{\epsilon} - 1 + s\delta. \tag{5}$$

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.6 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 (5) 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

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

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 ρk-zCDP), we get the following confidence width τ,

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

□

B.3 Comparison of Theorem 2.1 with Prior Work

One can also get a high-probability bound on the sample accuracy of $\mathcal{M}(X)$ using Theorem 3 in Xu and Raginsky [2017], resulting in

$$\tau \geq \max \left\{ \sqrt{\frac{8}{n}\left(\frac{2\rho k n}{\beta} + \log\left(\frac{4}{\beta}\right)\right)}, \frac{2}{n}\sqrt{\frac{\ln(4k/\beta)}{\rho}} \right\} \tag{6}$$

where i.i.d. Gaussian noise $N\left(0, \frac{1}{2\rho n^2}\right)$ has been added to each query. The proof is similar to the proof of Theorem 2.1. If the mutual information bound $B = \rho k n \geq 1$, then the first term in the expression of the confidence width in Theorem 2.1 is less than the first term in eq. (6). Furthermore, if $B \geq \frac{\ln k/\beta}{\rho n}$, then the first term dominates in the expression of the confidence width in Theorem 2.1, thus making Theorem 2.1 result in a tighter bound for any $\beta \in (0,1)$. For very small values of B, there exist sufficiently small β for which the result obtained via Xu and Raginsky [2017] is better.

B.4 Confidence Bounds for Thresholdout (Dwork et al. [2015a])

Theorem B.9. *If the Thresholdout mechanism $\mathcal{M}$ with noise scale σ, and threshold T is used for answering queries ϕ_i, $i \in [k]$, with reported answers $a_1, \cdots, a_k$ such that $\mathcal{M}$ uses the holdout set of size h to answer at most B queries, then given confidence parameter β, Thresholdout is (τ, β)-accurate, where*

$$\tau = \sqrt{\frac{1}{\beta} \cdot \left(T^2 + \psi + \frac{\xi}{4h} + \sqrt{\frac{\xi}{h} \cdot (T^2 + \psi)} \right)}$$

for $\psi = \mathbb{E}\left[(\max_{i\in[k]} W_i + \max_{j\in[B]} Y_j)^2 \right] + 2T \cdot \mathbb{E}\left[\max_{i\in[k]} W_i + \max_{j\in[B]} Y_j \right]$, *and* $\xi = \min_{\lambda\in[0,1)} \left(\frac{\frac{2B}{\sigma^2 h} - \ln(1-\lambda)}{\lambda} \right)$, *where* $W_i \sim Lap(4\sigma), i \in [k]$ *and* $Y_j \sim Lap(2\sigma), j \in [B]$.

Proof. Similar to the proof of Theorem 2.1, first we derive bounds on the mean squared error (MSE) for answers to statistical queries produced by Thresholdout. We want to bound the maximum MSE over all of the statistical queries, where the expectation is over the noise added by the mechanism and the randomness of the adversary.

Theorem B.10. *If the Thresholdout mechanism $\mathcal{M}$ with noise scale σ, and threshold T is used for answering queries ϕ_i, $i \in [k]$, with reported answers $a_1, \cdots, a_k$ such that $\mathcal{M}$ uses the holdout set of size h to answer at most B queries, then we have*

$$\mathop{\mathbb{E}}_{\substack{X\sim\mathcal{D}^n,\\ \phi_{j^*}\sim\mathcal{W}_{\mathcal{D}}[\mathcal{M},\mathcal{A}](X)}} \left[(a_{j^*} - \phi_{j^*}(\mathcal{D}))^2 \right] \leq T^2 + \psi + \frac{\xi}{4h} + \sqrt{\frac{\xi}{h} \cdot (T^2 + \psi)},$$

for $\psi = \mathbb{E}\left[(\max_{i\in[k]} W_i + \max_{j\in[B]} Y_j)^2 \right] + 2T \cdot \mathbb{E}\left[\max_{i\in[k]} W_i + \max_{j\in[B]} Y_j \right]$ *and* $\xi = \min_{\lambda\in[0,1)} \left(\frac{\frac{2B}{\sigma^2 h} - \ln(1-\lambda)}{\lambda} \right)$, *where* $W_i \sim Lap(4\sigma), i \in [k]$ *and* $Y_j \sim Lap(2\sigma), j \in [B]$.

Proof. Let us denote the holdout set in $\mathcal{M}$ by X_h and the remaining set as X_t. Let $\mathcal{O}$ denote the distribution $\mathcal{W}_{\mathcal{D}}[\mathcal{M}, \mathcal{A}](X)$, where $X \sim \mathcal{D}^n$. We have:

$$\begin{aligned}\mathop{\mathbb{E}}_{\phi_{j^*}\sim\mathcal{O}}\left[(a_{j^*} - \phi_{j^*}(\mathcal{D}))^2\right] &= \mathop{\mathbb{E}}_{\phi_{j^*}\sim\mathcal{O}}\left[(a_{j^*} - \phi_{j^*}(X_h) + \phi_{j^*}(X_h) - \phi_{j^*}(\mathcal{D}))^2\right] \\ &= \mathop{\mathbb{E}}_{\phi_{j^*}\sim\mathcal{O}}\left[(a_{j^*} - \phi_{j^*}(X_h))^2\right] + \mathop{\mathbb{E}}_{\phi_{j^*}\sim\mathcal{O}}\left[(\phi_{j^*}(X_h) - \phi_{j^*}(\mathcal{D}))^2\right] \\ &\quad + 2\sqrt{\mathop{\mathbb{E}}_{\phi_{j^*}\sim\mathcal{O}}\left[(a_{j^*} - \phi_{j^*}(X_h))^2\right]} \cdot \sqrt{\mathop{\mathbb{E}}_{\phi_{j^*}\sim\mathcal{O}}\left[(\phi_{j^*}(X_h) - \phi_{j^*}(\mathcal{D}))^2\right]}\end{aligned} \tag{7}$$

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

Now, define a set S_h which contains the indices of the queries answered via X_h. We know that for at most B queries $\phi_j \in S_h$, the output of $\mathcal{M}$ was $a_j = \phi_j(X_h) + Z_j$ where $Z_j \sim Lap(\sigma)$, whereas it was $a_i = \phi_i(X_t)$ for at least $k - B$ queries, $i \in [k \setminus S_h]$. Also, define $W_i \sim Lap(4\sigma), i \in [k]$ and $Y_j \sim Lap(2\sigma), j \in S_h$. Thus, for any $j^* \in [k]$, we have:

$$\begin{aligned}a_{j^*} - \phi_{j^*}(X_h) &\leq \max\left\{\max_{i\in[k\setminus S_h]} |\phi_i(X_h) - \phi_i(X_t)|, \max_{j\in S_h} Z_j\right\} \\ &\leq \max\left\{\max_{\substack{i\in[k\setminus S_h],\\ j(i)\in S_h}} T + Y_{j(i)} + W_i, \max_{j\in S_h} Z_j\right\} \\ &\leq \max\left\{\max_{\substack{i\in[k\setminus S_h],\\ j\in S_h}} T + Y_j + W_i, \max_{j\in S_h} Z_j\right\} \\ &\leq T + \max_{i\in[k]} W_i + \max_{j\in[B]} Y_j\end{aligned}$$

Thus,

$$\begin{aligned}\mathop{\mathbb{E}}_{\phi_{j^*}\sim\mathcal{O}}\left[(a_{j^*} - \phi_{j^*}(X_h))^2\right] &\leq \mathbb{E}\left[(T + \max_{i\in[k]} W_i + \max_{j\in[B]} Y_j)^2\right] \\ &= T^2 + \mathbb{E}\left[(\max_{i\in[k]} W_i + \max_{j\in[B]} Y_j)^2\right] + 2T \cdot \mathbb{E}\left[\max_{i\in[k]} W_i + \max_{j\in[B]} Y_j\right]\end{aligned} \tag{8}$$

We bound the 2nd term in (7) as follows. For every $i \in S_h$, there are two costs induced due to privacy: the Sparse Vector component, and the noise addition to $\phi_i(X_h)$. 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.7, this translates to each providing a $\left(\frac{1}{2\sigma^2 h^2}\right)$-zCDP guarantee. Since there are at most B such instances of each, by Theorem A.7 we get that $\mathcal{M}$ is $\left(\frac{B}{\sigma^2 h^2}\right)$-zCDP. Thus, by Lemma C.6 we have

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

Proceeding similar to the proof of Theorem C.8, we use the sub-Gaussian parameter for statistical queries in Lemma C.5 to obtain the following bound from Theorem C.2:

$$\mathop{\mathbb{E}}_{\phi_{j^*}\sim\mathcal{O}}\left[\left(\phi_{j^*}(X_h) - \phi_{j^*}(\mathcal{D})\right)^2\right] = \mathop{\mathbb{E}}_{\substack{X\sim\mathcal{D}^n,\\ \mathcal{M},\mathcal{A}}}\left[\max_{i\in S_h}\left\{\left(\phi_i(X_h) - \phi_i(\mathcal{D})\right)^2\right\}\right] \leq \frac{\xi}{4h} \tag{9}$$

Defining $\psi = \mathbb{E}\left[(\max_{i\in[k]} W_i + \max_{j\in[B]} Y_j)^2\right] + 2T \cdot \mathbb{E}\left[\max_{i\in[k]} W_i + \max_{j\in[B]} Y_j\right]$, and combining Equations (7), (8), and (9), we get:

$$\mathop{\mathbb{E}}_{\phi_{j^*}\sim\mathcal{O}}\left[(a_{j^*} - \phi_{j^*}(\mathcal{D}))^2\right] \leq T^2 + \psi + \frac{\xi}{4h} + \sqrt{\frac{\xi}{h} \cdot (T^2 + \psi)}$$

□

We can use the MSE bound from Theorem B.10, and Chebyshev's inequality, to get the statement of the theorem. □

C Omitted Results, and Proofs

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

C.1 RMSE analysis for the single-adaptive query strategy

Theorem C.1. *The output by the single-adaptive query 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.*

Proof. Consider a dataset $X \in \mathcal{X}^n$, where $\mathcal{X}$ is the uniform distribution over $\{-1,1\}^{k+1}$. We will denote the j^{th} element of $x_i \in X$ by $x_i(j)$, for $j \in [k+1]$. Now, $\forall j \in [k]$, we have that:

$$\mathop{\mathbb{E}}_{X}\left[\frac{1+x(j)\cdot x(k+1)}{2}\right] = \frac{1+\Pr\left(x(j)=x(k+1)\right)-\Pr\left(x(j)\neq x(k+1)\right)}{2} = a_j$$

$$\therefore \Pr_X\left(x(j)=x(k+1)\right)=a_j \qquad \text{and} \qquad \Pr_X\left(x(j)\neq x(k+1)\right)=1-a_j$$

Now,

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

Thus, $\phi_{k+1}(x) = \dfrac{sign\left(\ln\left(\frac{\Pr_X\left(x(k+1)=1|\wedge_{j\in[k]}x(j)=x_j\right)}{\Pr_X\left(x(k+1)=-1|\wedge_{j\in[k]}x(j)=x_j\right)}\right)\right)+1}{2}$.

As a result, the adaptive query ϕ_{k+1} in Algorithm 5 (setting input $S=\{k+1\}$) corresponds to a naive Bayes classifier of $x(k+1)$, and given that $\mathcal{X}$ is the uniform distribution over $\{-1,1\}^{k+1}$, this is the best possible classifier for $x(k+1)$. This results answer a_{k+1} achieving the maximum possible deviation from the answer on the population, which is 0.5 as $\mathcal{X}$ is uniformly distributed over $\{-1,1\}^{k+1}$. Thus, a_{k+1} results in the maximum possible RMSE. □

C.2 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.2. *Let $\mathcal{Q}_\sigma$ be the class of queries $\phi : \mathcal{X}^n \to \mathbb{R}$ such that $\phi(X)-\phi(\mathcal{D}^n)$ is σ-subgaussian where $X \sim \mathcal{D}^n$. If $\mathcal{M} : \mathcal{X}^n \to \mathcal{Q}_\sigma$ is a randomized mapping from datasets to*

queries such that $I(\mathcal{M}(X);X) \leq B$ *then*

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

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

Lemma C.3 (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.4 (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.2. Proceeding similar to the proof of Proposition 3.1 in Russo and Zou [2015], we write $\boldsymbol{\phi}(X) = (\phi(X) : \phi\in\mathcal{Q}_\sigma)$. We have:

$$\begin{aligned}
I(\mathcal{M}(X);X) &\geq I(\mathcal{M}(X);\boldsymbol{\phi}(X))\\
&= \sum_{\boldsymbol{a},\phi\in\mathcal{Q}_\sigma}\ln\left(\frac{\Pr[(\boldsymbol{\phi}(X),\mathcal{M}(X))=(\boldsymbol{a},\phi)]}{\Pr[\boldsymbol{\phi}(X)=\boldsymbol{a}]\Pr[\mathcal{M}(X)=\phi]}\right)\cdot\Pr[(\boldsymbol{\phi}(X),\mathcal{M}(X))=(\boldsymbol{a},\phi)]\\
&= \sum_{\boldsymbol{a},\phi\in\mathcal{Q}_\sigma}\ln\left(\frac{\Pr[\boldsymbol{\phi}(X)=\boldsymbol{a}|\mathcal{M}(X)=\phi]}{\Pr[\boldsymbol{\phi}(X)=\boldsymbol{a}]}\right)\cdot\Pr[\mathcal{M}(X)=\phi]\Pr[\boldsymbol{\phi}(X)=\boldsymbol{a}|\mathcal{M}(X)=\phi]\\
&\geq \sum_{a,\phi\in\mathcal{Q}_\sigma}\ln\left(\frac{\Pr[\phi(X)=a|\mathcal{M}(X)=\phi]}{\Pr[\phi(X)=a]}\right)\cdot\Pr[\mathcal{M}(X)=\phi]\Pr[\phi(X)=a|\mathcal{M}(X)=\phi]\\
&= \sum_{\phi\in\mathcal{Q}_\sigma}\Pr[\mathcal{M}(X)=\phi]\cdot\mathrm{D}_{KL}[(\phi(X)|\mathcal{M}(X)=\phi)||\phi(X)] \qquad (10)
\end{aligned}$$

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

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

where the first and second inequalities follows from Lemmas C.3 and C.4, respectively.

Therefore, from eq. (10), we have

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

Rearranging terms, we have

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

□

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

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

Lemma C.6 (Bun and Steinke [2016a])**.** *If $\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}$ is ρ-zCDP and $X \sim \mathcal{D}^n$, then $I(\mathcal{M}(X); X) \leq \rho n$.*

In order to prove Theorem C.8, we use the same monitor from Algorithm 2 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.7. *If $I(X; \mathcal{M}(X)) \leq B$ where $X \sim \mathcal{D}^n$, then $I(X; \mathcal{W}_{\mathcal{D}}[\mathcal{M}, \mathcal{A}](X)) \leq B$.*

Next, we derive bounds on the mean squared error (MSE) for answers to statistical queries produced by the Gaussian mechanism. We want to bound the maximum MSE over all of the statistical queries, where the expectation is over the noise added by the mechanism and the randomness of the adversary.

$$\begin{aligned}\mathop{\mathbb{E}}_{\substack{X\sim\mathcal{D}^n,\\ \mathcal{A},\mathcal{M}}}\left[\max_{i\in[k]}\left(\phi_i(\mathcal{D}) - a_i\right)^2\right] &\leq 2\mathop{\mathbb{E}}_{\substack{X\sim\mathcal{D}^n,\\ \mathcal{A},\mathcal{M}}}\left[\max_{i\in[k]}\left\{\left(\phi_i(\mathcal{D}) - \phi_i(X)\right)^2 + \left(\phi_i(X) - a_i\right)^2\right\}\right] \\ &= 2\cdot\mathbb{E}\left[\max_{i\in[k]}\left(\phi_i(\mathcal{D}) - \phi_i(X)\right)^2\right] + 2\cdot\mathop{\mathbb{E}}_{Z_i\sim N\left(0,\frac{1}{2n^2\rho}\right)}\left[\max_{i\in[k]} Z_i^2\right] \qquad (11)\end{aligned}$$

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

Theorem C.8. *For parameter $\rho \geq 0$, the answers provided by the Gaussian mechanism $a_1, \cdots, a_k$ against an adaptively selected sequence of queries satisfy:*

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

Proof. 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 ρ-zCDP, which (using Lemma C.6 and the post-processing property of zCDP) makes the mutual information bound $B = \rho kn$. We then use Lemma C.7 and the sub-Gaussian parameter for statistical queries in Lemma C.5 to obtain the following bound from Theorem C.2.

$$\begin{aligned}\mathop{\mathbb{E}}_{\substack{X\sim\mathcal{D}^n,\\ \phi^*\sim\mathcal{W}_{\mathcal{D}}[\mathcal{M},\mathcal{A}](X)}}\left[\left(\phi^*(X) - \phi^*(\mathcal{D})\right)\right)^2\right] &= \mathop{\mathbb{E}}_{X\sim\mathcal{D}^n,\mathcal{M},\mathcal{A}}\left[\max_{i\in[k]}\left\{\left(\phi_i(X) - \phi_i(\mathcal{D})\right)^2\right\}\right] \\ &\leq \frac{1}{4n}\cdot\min_{\lambda\in[0,1)}\left(\frac{2\rho kn - \ln(1-\lambda)}{\lambda}\right) \qquad (12)\end{aligned}$$

We then combine this result with (11) to get the statement of the theorem. □

Proof of Theorem 2.1. We can bound the population accuracy in (1) using Equation (12) and Chebyshev's inequality to obtain the following high probability bound,

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

We then use the result of Corollary A.9 to obtain our accuracy bound. □

C.3 Proofs from Section 3

C.3.1 Proof of Theorem 3.2

We start by proving the validity for query responses output by GnC that correspond to the responses provided by $\mathcal{M}_g$, i.e., each query ϕ_i s.t. the output of GnC is $(a_{g,i}, \tau_i)$.

Lemma C.9. *If the function* $HoldoutTol(\beta', a_g, a_h) = \sqrt{\frac{\ln 2/\beta'}{2n_h}}$ *in GnC (Algorithm 1), then for each query* ϕ_i *s.t. the output of GnC is* $(a_{g,i}, \tau_i)$*, we have* $\Pr\left(|a_{g,i} - \phi_i(\mathcal{D})| > \tau_i\right) \leq \beta_i$.

Proof. Consider a query ϕ_i for which the output of the GnC mechanism is $(a_{g,i}, \tau_i)$, and let $\tau_h = HoldoutTol(\beta_i, a_{g,i}, a_{h,i})$. Now, we have

$$\begin{aligned}\Pr\left(|a_{g,i} - \phi_i(\mathcal{D})| > \tau_i\right) &\leq \Pr\left(|a_{g,i} - a_{h,i}| + |a_{h,i} - \phi_i(\mathcal{D})| > \tau_i\right) \\ &= \Pr\left(|a_{h,i} - \phi_i(\mathcal{D})| > \tau_h\right) \\ &\leq \beta_i\end{aligned}$$

where the equality follows since $|a_{g,i} - a_{h,i}| \leq \tau_i - \tau_h$, and the last inequality follows from applying the Chernoff bound for statistical queries. □

Lemma C.9 is agnostic to the guesses and holdout answers while computing the holdout tolerance τ_h. However, GnC can provide a better tolerance τ_h in the presence of low-variance queries. We state the guarantee in Lemma 3.1, and provide a proof for it in Appendix C.3.2.

Next, we provide the accuracy for the query answers output by GnC that correspond to discretized empirical answers on the holdout. It is obtained by maximizing the discretization parameter such that applying the Chernoff bound on the discretized answer satisfies the required validity guarantee.

Lemma C.10. *If failure* f *occurs in GnC (Algorithm 1) for query* ϕ_i *and the output of GnC is* $\left(\lfloor a_{h,i} \rfloor_{\gamma_f}, \tau_i\right)$*, since we have* $\gamma_f = \max_{[0,\tau')} \gamma$ *s.t.* $2e^{-2(\tau'-\gamma)^2 n_h} \leq \beta'$*, we have* $\Pr\left(|a_{g,i} - \phi_i(\mathcal{D})| > \tau_i\right) \leq \beta_i$*. Here,* $\lfloor y \rfloor_\gamma$ *denotes* y *discretized to multiples of* γ

Proof of Theorem 3.2. Let an instance of the Guess and Check mechanism $\mathcal{M}$ encounter f failures while providing responses to k queries $\{\phi_1, \ldots, \phi_k\}$. We will consider the interaction between an analyst $\mathcal{A}$ and the Guess and Check mechanism $\mathcal{M}$ to form a tree T, where the nodes in T correspond to queries, and each branch of a node is a possible answer for the corresponding query. We first note a property about the structure of T:

Fact 1: For any query $\phi_{i'}$, if the check within $\mathcal{M}$ results in failure f', then there are $\frac{1}{\gamma_{f'}}$ possible responses for $\phi_{i'}$. On the other hand, if the check doesn't result in a failure, then there is only 1 possible response for $\phi_{i'}$, namely $(a_{g,i'}, \tau_{i'})$.

Next, notice that each node in T can be uniquely identified by the tuple $t = (i', f', \{j_1, \ldots, j_{f'}\}, \{\gamma_{j_1}, \ldots, \gamma_{j_{f'}}\})$, where i' is the depth of the node (also, the index of the next query to be asked), f' is the number of failures within $\mathcal{M}$ that have occurred in the path from the root to node t, and for $\ell \in [f']$, the value j_ℓ denotes the query index of the ℓth failure on this path, whereas γ_{j_ℓ} is the corresponding discretization parameter that was used to answer the query. We can now observe another property about the structure of T:

Fact 2: For any $i' \in [k], f' \in [i'-1]$, there are $\binom{i'-1}{f'} \prod_{\ell \in [f']} \left(\frac{1}{\gamma_{j_\ell}}\right)$ nodes in T of type $(i', f', ;, ;)$. This follows since there are $\binom{i'-1}{f'}$ possible ways that f' failures can occur in $i'-1$ queries, and from Fact 1 above, there are $\frac{1}{\gamma_{j_\ell}}$ possible responses for a failure occurring at query index $j_\ell, \ell \in [f']$.

Now, we have

$$
\begin{aligned}
\Pr\left(\exists i \in [k] : |\phi_i(\mathcal{D}) - a_i| > \tau_i\right) &\leq \sum_{\text{node } t \in T} \Pr\left(|\phi_t(\mathcal{D}) - a_t| > \tau_t\right) \\
&= \sum_{i' \in [k]} \sum_{f' \in [i'-1]} \sum_{\{j_1,\ldots,j_{f'}\}} \sum_{\{\gamma_{j_1},\ldots,\gamma_{j_{f'}}\}} \Pr\left(|\phi_{i'}(\mathcal{D}) - a_{i'}| > \tau_{i'} | t\right) \\
&= \sum_{i' \in [k]} \sum_{f' \in [i'-1]} \sum_{\{j_1,\ldots,j_{f'}\}} \sum_{\{\gamma_{j_1},\ldots,\gamma_{j_{f'}}\}} \frac{\beta \cdot c_{i'-1} \cdot c_{f'}}{\nu_{i',f',\gamma_{j_1}^{j_{f'}}}} \\
&= \beta \left(\sum_{i' \in [k]} \sum_{f' \in [i'-1]} \sum_{\{j_1,\ldots,j_{f'}\}} \sum_{\{\gamma_{j_1},\ldots,\gamma_{j_{f'}}\}} \frac{c_{i'-1} \cdot c_{f'}}{\binom{i'-1}{f'} \prod_{\ell \in [f']} \left(\frac{1}{\gamma_{j_\ell}}\right)} \right) \\
&= \beta \left(\sum_{i' \in [k]} c_{i'-1} \cdot \left(\sum_{f' \in [i'-1]} c_{f'} \right) \right) \leq \beta \sum_{i' \in [k]} c_{i'-1} \leq \beta
\end{aligned}
$$

where the second equality follows from Lemma 3.1, Lemma C.10, and substituting the values of β_i in Algorithm 1; the last equality follows from Fact 2 above; and the last two inequalities follow since $\sum_{j \geq 0} c_j \leq 1$. Thus, we have simultaneous coverage $1 - \beta$ for the Guess and Check mechanism $\mathcal{M}$. □

Lemma C.9 is agnostic to the guesses and holdout answers while computing the holdout tolerance τ_h. However, GnC can provide a better tolerance τ_h in the presence of low-variance queries. We provide a proof for it below.

C.3.2 Proof of Lemma 3.1

Lemma 3.1 uses the MGF of the binomial distribution to approximate the probabilities of deviation of the holdout's empirical answer from the true population mean (instead of, say, optimizing parameters in a large deviation bound). This is exact when the query only takes values in $\{0, 1\}$. To prove the lemma, we start by first proving the dominance of the binomial MGF.

Lemma C.11. *Let $X_1, X_2, ..., X_n$ be i.i.d. random variables in $[0, 1]$, distributed according to $\mathcal{D}$, and let $\mu = \mathbb{E}[X_i]$. Let $S = \sum_{i=1}^{n} X_i$, and $B \sim B(n, \mu)$ be a binomial random variable. Then, we have:*

$$
\Pr(S > t) \leq \min_{\lambda > 0} \frac{\mathbb{E}\left[e^{\lambda B}\right]}{e^{\lambda t}}.
$$

Proof. Consider some $\lambda > 0$. We have

$$
\begin{aligned}
\mathbb{E}\left[e^{\lambda S}\right] &= \left(\mathbb{E}\left[e^{\lambda X_1}\right]\right)^n \\
&\leq \left(\mathbb{E}\left[X_1 \cdot e^{\lambda} + (1 - X_1) \cdot e^{0}\right]\right)^n = \left(1 + \mathbb{E}[X_1]\left(e^{\lambda} - 1\right)\right)^n \\
&= \left(1 + \mu(e^{\lambda} - 1)\right)^n \\
&= \mathbb{E}\left[e^{\lambda B}\right] \qquad (13)
\end{aligned}
$$

where the first equality follows since $X_1, X_2, ..., X_n$ are i.i.d., and the last equality represents the MGF of the binomial distribution.

Now, we get:

$$
\begin{aligned}
\Pr(S > t) = \Pr\left(e^{\lambda S} > e^{\lambda t}\right) &\leq \frac{\mathbb{E}\left[e^{\lambda S}\right]}{e^{\lambda t}} \leq \frac{\mathbb{E}\left[e^{\lambda B}\right]}{e^{\lambda t}} \\
&\leq \min_{\lambda' > 0} \frac{\mathbb{E}\left[e^{\lambda' B}\right]}{e^{\lambda' t}}
\end{aligned}
$$

where the first inequality follows from the Chernoff bound, and the second inequality follows from Equation (13). □

Proof of Lemma 3.1. Consider a query ϕ_i for which the output of the GnC mechanism is $(a_{g,i}, \tau_i)$. Let $\tau_h = HoldoutTol(\beta_i, a_{g,i}, a_{h,i})$. For proving $\Pr\left(|a_{g,i} - \phi_i(\mathcal{D})| > \tau_i\right) \leq \beta_i$, it suffices to show that if $|a_{g,i} - \phi(\mathcal{D})| > \tau_i$, then

$$\sup_{\substack{\mathcal{D} \text{ s.t.} \\ \phi_i(\mathcal{D}) = a_{g,i} - \tau_i}} \Pr_{X_h \sim \mathcal{D}^{n_h}} \left(a_{h,i} \geq a_{g,i} - \tau_i + \tau_h\right) \leq \frac{\beta_i}{2} \tag{14}$$

$$\text{and} \quad \sup_{\substack{\mathcal{D} \text{ s.t.} \\ \phi_i(\mathcal{D}) = a_{g,i} + \tau_i}} \Pr_{X_h \sim \mathcal{D}^{n_h}} \left(a_{h,i} \leq a_{g,i} + \tau_i - \tau_h\right) \leq \frac{\beta_i}{2} \tag{15}$$

When $a_{g,i} > a_{h,i}$, we only require inequality 14 to hold. Let $B \sim B(n, \mu)$ be a binomial random variable. We have:

$$\begin{aligned} \sup_{\substack{\mathcal{D} \text{ s.t.} \\ \phi_i(\mathcal{D}) = \mu}} \Pr_{X_h \sim \mathcal{D}^{n_h}} \left(a_{h,i} \geq \mu + \tau'\right) &\leq \min_{\lambda > 0} \frac{\mathbb{E}\left[e^{\lambda B}\right]}{e^{\lambda n(\mu + \tau')}} = \min_{\lambda > 0} e^{\left\{\ln\left(\mathbb{E}\left[e^{\lambda B}\right]\right) - \lambda n(\mu + \tau')\right\}} \\ &= \frac{\mathbb{E}\left[e^{\ell B}\right]}{e^{\ell n(\mu + \tau')}} = \frac{\left(1 + \mu(e^{\ell} - 1)\right)^n}{e^{\ell n(\mu + \tau')}} \end{aligned} \tag{16}$$

where $\ell = \arg\min_{\lambda > 0} e^{\left\{\ln\left(\mathbb{E}\left[e^{\lambda B}\right]\right) - \lambda n(\mu + \tau')\right\}}$, i.e., $\frac{\mu e^{\ell}}{1 + \mu(e^{\ell} + 1)} = \mu + \tau'$. Here, the first inequality follows from Lemma C.11 by setting $t = (\mu + \tau')n$, and the last equality follows from the MGF of the binomial distribution. Thus, we get that inequality 14 holds.

Similarly, when $a_{g,i} \leq a_{h,i}$, we only require inequality 15 to hold. Let $\mu' = 1 - \mu$, and $B' \sim B(n, \mu')$. Therefore, we get

$$\sup_{\substack{\mathcal{D} \text{ s.t.} \\ \phi_i(\mathcal{D}) = \mu}} \Pr_{X_h \sim \mathcal{D}^{n_h}} \left(a_{h,i} \leq \mu - \tau'\right) = \sup_{\substack{\mathcal{D} \text{ s.t.} \\ \phi_i(\mathcal{D}) = \mu'}} \Pr_{X_h \sim \mathcal{D}^{n_h}} \left(a_{h,i} \geq \mu' + \tau'\right) \leq \frac{\left(1 + \mu'(e^{\ell'} - 1)\right)^n}{e^{\ell' n(\mu' + \tau')}}$$

where $\ell' = \arg\min_{\lambda > 0} e^{\left\{\ln\left(\mathbb{E}\left[e^{\lambda B'}\right]\right) - \lambda n(\mu' + \tau')\right\}}$, i.e., $\frac{\mu' e^{\ell'}}{1 + \mu'(e^{\ell'} + 1)} = \mu' + \tau'$. Here, the inequality follows from inequality 16. Thus, we get that inequality 15 holds.

□

D Omitted Pseudocodes

Algorithm 4 Thresholdout (Dwork et al. [2015a])

Require: train size t, threshold T, noise scale σ

Randomly partition dataset X into a training set X_t containing t samples, and a holdout set X_h containing $h = n - t$ samples

Initialize $\widehat{T} \leftarrow T + Lap(2\sigma)$

for each query ϕ **do**

  if $|\phi(X_h) - \phi(X_t)| > \widehat{T} + Lap(4\sigma)$ **then**

    $\widehat{T} \leftarrow T + Lap(2\sigma)$

    Output $\phi(X_h) + Lap(\sigma)$

  else

    Output $\phi(X_t)$

Algorithm 5 A custom adaptive analyst strategy for random data

Require: Mechanism $\mathcal{M}$ with a hidden dataset $X \in \{-1, 1\}^{n \times (k+1)}$, set $S \subseteq [k+1]$ denoting the indices of adaptive queries[6]

Define $j \leftarrow 1$, and $success \leftarrow True$

while $j \leq k$ and $success = True$ **do**

  if $j \in S$ **then**

    Define $\phi_j(x) = \frac{sign\left(\sum_{i \in [j-1] \setminus S} \left(x(i) \cdot \ln \frac{a_i}{a_i - 1}\right)\right) + 1}{2}$, where $sign(y) = \begin{cases} 1 & \text{if } y \geq 0 \\ -1 & \text{otherwise} \end{cases}$

  else

    Define $\phi_j(x) = \frac{1 + x(j) \cdot x(k+1)}{2}$

  Give ϕ_j to $\mathcal{M}$, and receive $a_j \in [0, 1] \cup \perp$ from $\mathcal{M}$

  if $a_j = \perp$ **then**

    $success = False$

  else

    $j \leftarrow j + 1$

[6] For the single-adaptive query strategy used in the plots in Figure 2, we set $S = \{k + 1\}$. For the quadratic-adaptive strategy used in the plots in Section 3, we set $S = \{i : 1 < i \leq k \text{ and } \exists \ell \in \mathbb{N} \text{ s.t. } \ell < i \text{ and } \ell^2 = i\}$.