Preserving Statistical Validity in Adaptive Data Analysis*

Cynthia Dwork† Vitaly Feldman‡ Moritz Hardt§ Toniann Pitassi¶
Omer Reingold‖ Aaron Roth**

Abstract

A great deal of effort has been devoted to reducing the risk of spurious scientific discoveries, from the use of sophisticated validation techniques, to deep statistical methods for controlling the false discovery rate in multiple hypothesis testing. However, there is a fundamental disconnect between the theoretical results and the practice of data analysis: the theory of statistical inference assumes a fixed collection of hypotheses to be tested, or learning algorithms to be applied, selected non-adaptively before the data are gathered, whereas in practice data is shared and reused with hypotheses and new analyses being generated on the basis of data exploration and the outcomes of previous analyses.

In this work we initiate a principled study of how to guarantee the validity of statistical inference in adaptive data analysis. As an instance of this problem, we propose and investigate the question of estimating the expectations of m adaptively chosen functions on an unknown distribution given n random samples.

We show that, surprisingly, there is a way to estimate an *exponential* in n number of expectations accurately even if the functions are chosen adaptively. This gives an exponential improvement over standard empirical estimators that are limited to a linear number of estimates. Our result follows from a general technique that counter-intuitively involves actively perturbing and coordinating the estimates, using techniques developed for privacy preservation. We give additional applications of this technique to our question.

*Preliminary version of this work appears in the proceedings of the ACM Symposium on Theory of Computing (STOC), 2015

†Microsoft Research

‡IBM Almaden Research Center. Part of this work done while visiting the Simons Institute, UC Berkeley

§IBM Almaden Research Center

¶University of Toronto

‖Samsung Research America

**Department of Computer and Information Science, University of Pennsylvania

1 Introduction

Throughout the scientific community there is a growing recognition that claims of statistical significance in published research are frequently invalid [Ioa05b, Ioa05a, PSA11, BE12]. The past few decades have seen a great deal of effort to understand and propose mitigations for this problem. These efforts range from the use of sophisticated validation techniques and deep statistical methods for controlling the false discovery rate in multiple hypothesis testing to proposals for preregistration (that is, defining the entire data-collection and data-analysis protocol ahead of time). The statistical inference theory surrounding this body of work assumes a fixed procedure to be performed, selected before the data are gathered. In contrast, the practice of data analysis in scientific research is by its nature an adaptive process, in which new hypotheses are generated and new analyses are performed on the basis of data exploration and observed outcomes on the same data. This disconnect is only exacerbated in an era of increased amounts of open access data, in which multiple, mutually dependent, studies are based on the same datasets.

It is now well understood that adapting the analysis to data (*e.g.*, choosing what variables to follow, which comparisons to make, which tests to report, and which statistical methods to use) is an implicit multiple comparisons problem that is not captured in the reported significance levels of standard statistical procedures. This problem, in some contexts referred to as "p-hacking" or "researcher degrees of freedom", is one of the primary explanations of why research findings are frequently false [Ioa05b, SNS11, GL14].

The "textbook" advice for avoiding problems of this type is to collect fresh samples from the same data distribution whenever one ends up with a procedure that depends on the existing data. Getting fresh data is usually costly and often impractical so this requires partitioning the available dataset randomly into two or more disjoint sets of data (such as a training and testing set) prior to the analysis. Following this approach conservatively with m adaptively chosen procedures would significantly (on average by a factor of m) reduce the amount of data available for each procedure. This would be prohibitive in many applications, and as a result, in practice even data allocated for the sole purpose of testing is frequently reused (for example to tune parameters). Such abuse of the holdout set is well known to result in significant overfitting to the holdout or cross-validation set [Reu03, RF08].

Clear evidence that such reuse leads to overfitting can be seen in the data analysis competitions organized by Kaggle Inc. In these competitions, the participants are given training data and can submit (multiple) predictive models in the course of competition. Each submitted model is evaluated on a (fixed) test set that is available only to the organizers. The score of each solution is provided back to each participant, *who can then submit a new model*. In addition the scores are published on a public leaderboard. At the conclusion of the competition the best entries of each participant are evaluated on an additional, hitherto unused, test set. The scores from these final evaluations are published. The comparison of the scores on the adaptively reused test set and one-time use test set frequently reveals significant overfitting to the reused test set (e.g. [Win, Kaga]), a well-recognized issue frequently discussed on Kaggle's blog and user forums [Kagb, Kagc].

Despite the basic role that adaptivity plays in data analysis we are not aware of previous general efforts to address its effects on the statistical validity of the results (see Section 1.4 for an overview of existing approaches to the problem). We show that, surprisingly, the challenges of adaptivity can be addressed using insights from *differential privacy*, a definition of privacy tailored to privacy-preserving data analysis. Roughly speaking, differential privacy ensures that the probability of observing any outcome from an analysis is "essentially unchanged" by modifying any single

dataset element (the probability distribution is over the randomness introduced by the algorithm). Differentially private algorithms permit a data analyst to learn about the dataset as a whole (and, by extension, the distribution from which the data were drawn), while simultaneously protecting the privacy of the individual data elements. Strong composition properties show this holds even when the analysis proceeds in a sequence of adaptively chosen, individually differentially private, steps.

1.1 Problem Definition

We consider the standard setting in statistics and statistical learning theory: an analyst is given samples drawn randomly and independently from some unknown distribution $\mathcal{P}$ over a discrete universe $\mathcal{X}$ of possible data points. While our approach can be applied to any output of data analysis, we focus on real-valued functions defined on $\mathcal{X}$. Specifically, for a function $\psi\colon \mathcal{X} \to [0,1]$ produced by the analyst we consider the task of estimating the expectation $\mathcal{P}[\psi] = \mathbb{E}_{x\sim\mathcal{P}}[\psi(x)]$ up to some additive error τ usually referred to as *tolerance*. We require the estimate to be within this error margin with high probability.

We choose this setup for three reasons. First, a variety of quantities of interest in data analysis can be expressed in this form for some function ψ. For example, true means and moments of individual attributes, correlations between attributes and the generalization error of a predictive model or classifier. Next, a request for such an estimate is referred to as a *statistical query* in the context of the well-studied statistical query model [Kea98, FGR+13], and it is known that using statistical queries in place of direct access to data it is possible to implement most standard analyses used on i.i.d. data (see [Kea98, BDMN05, CKL+06] for examples). Finally, the problem of providing accurate answers to a large number of queries for the average value of a function on the dataset has been the subject of intense investigation in the differential privacy literature.[1]

We address the following basic question: how many adaptively chosen statistical queries can be correctly answered using n samples drawn i.i.d. from $\mathcal{P}$? The conservative approach of using fresh samples for each adaptively chosen query would lead to a sample complexity that scales linearly with the number of queries m. We observe that such a bad dependence is inherent in the standard approach of estimating expectations by the exact empirical average on the samples. This is directly implied by the techniques from [DN03] who show how to make linearly many non-adaptive counting queries to a dataset, and reconstruct nearly all of it. Once the data set is nearly reconstructed it is easy to make a query for which the empirical average on the dataset is far from the true expectation. Note that this requires only a single round of adaptivity! A simpler and more natural example of the same phenomenon is known as "Freedman's paradox" [Fre83] and we give an additional simple example in the Appendix. This situation is in stark contrast to the non-adaptive case in which $n = O\left(\frac{\log m}{\tau^2}\right)$ samples suffice to answer m queries with tolerance τ using empirical averages. Below we refer to using empirical averages to evaluate the expectations of query functions as the *naïve* method.

1.2 Our Results

Our main result is a broad *transfer theorem* showing that any adaptive analysis that is carried out in a differentially private manner must lead to a conclusion that generalizes to the underlying

[1] The average value of a function ψ on a set of random samples is a natural estimator of $\mathcal{P}[\psi]$. In the differential privacy literature such queries are referred to as *(fractional) counting queries*.

distribution. This theorem allows us to draw on a rich body of results in differential privacy and to obtain corresponding results for our problem of guaranteeing validity in adaptive data analysis. Before we state this general theorem, we describe a number of important corollaries for the question we formulated above.

Our primary application is that, remarkably, it is possible to answer nearly *exponentially many* adaptively chosen statistical queries (in the size of the data set n). Equivalently, this reduces the *sample complexity* of answering m queries from *linear* in the number of queries to *polylogarithmic*, nearly matching the dependence that is necessary for non-adaptively chosen queries.

Theorem 1 (Informal)**.** *There exists an algorithm that given a dataset of size at least* $n \geq \min(n_0, n_1)$*, can answer any* m *adaptively chosen statistical queries so that with high probability, each answer is correct up to tolerance* τ*, where:*

$$n_0 = O\left(\frac{(\log m)^{3/2}\sqrt{\log|\mathcal{X}|}}{\tau^{7/2}}\right) \quad \textit{and} \quad n_1 = O\left(\frac{\log m \cdot \log|\mathcal{X}|}{\tau^4}\right).$$

The two bounds above are incomparable. Note that the first bound is larger than the sample complexity needed to answer non-adaptively chosen queries by only a factor of $O\left(\sqrt{\log m \log|\mathcal{X}|}/\tau^{3/2}\right)$, whereas the second one is larger by a factor of $O\left(\log(|\mathcal{X}|)/\tau^2\right)$. Here $\log|\mathcal{X}|$ should be viewed as roughly the *dimension* of the domain. For example, if the underlying domain is $\mathcal{X} = \{0,1\}^d$, the set of all possible vectors of d-boolean attributes, then $\log|\mathcal{X}| = d$.

The above mechanism is not computationally efficient (it has running time linear in the size of the data universe $|\mathcal{X}|$, which is *exponential* in the dimension of the data). A natural question raised by our result is whether there is an efficient algorithm for the task. This question was addressed in [HU14, SU14] who show that under standard cryptographic assumptions any algorithm that can answer more than $\approx n^2$ adaptively chosen statistical queries must have running time exponential in $\log|\mathcal{X}|$.

We show that it is possible to match this quadratic lower bound using a simple and practical algorithm that perturbs the answer to each query with independent noise.

Theorem 2 (Informal)**.** *There exists a computationally efficient algorithm for answering* m *adaptively chosen statistical queries, such that with high probability, the answers are correct up to tolerance* τ*, given a data set of size at least* $n \geq n_0$ *for:*

$$n_0 = O\left(\frac{\sqrt{m}(\log m)^{3/2}}{\tau^{5/2}}\right).$$

Finally, we show a computationally efficient method which can answer *exponentially many* queries so long as they were generated using $o(n)$ rounds of adaptivity, even if we do not know where the rounds of adaptivity lie. Another practical advantage of this algorithm is that it only pays the price for a round if adaptivity actually causes overfitting. In other words, the algorithm does not pay for the adaptivity itself but only for the actual harm to statistical validity that adaptivity causes. This means that in many situations it would be possible to use this algorithm successfully with a much smaller "effective" r (provided that a good bound on it is known).

Theorem 3 (Informal)**.** *There exists a computationally efficient algorithm for answering m adaptively chosen statistical queries, generated in r rounds of adaptivity, such that with high probability, the answers are correct up to some tolerance τ, given a data set of size at least $n \geq n_0$ for:*

$$n_0 = O\left(\frac{r \log m}{\tau^2}\right) .$$

Formal statements of these results appear in Section 5.

1.3 Overview of Techniques

We consider a setting in which an *arbitrary* adaptive data analyst chooses queries to ask (as a function of past answers), and receives answers from an algorithm referred to as an *oracle* whose input is a dataset $\boldsymbol{S}$ of size n randomly drawn from $\mathcal{P}^n$. To begin with, the oracles we use will only guarantee accuracy with respect to the empirical average on their input dataset $\boldsymbol{S}$, but they will simultaneously guarantee differential privacy. We exploit a crucial property about differential privacy, known as its post-processing guarantee: *Any* algorithm that can be described as the (possibly randomized) post-processing of the output of a differentially private algorithm is itself differentially private. Hence, although we do not know how an arbitrary analyst is adaptively generating her queries, we do know that if the only access she has to $\boldsymbol{S}$ is through a differentially private algorithm, then her method of producing query functions must be differentially private with respect to $\boldsymbol{S}$. We can therefore, without loss of generality, think of the oracle and the analyst as a *single* algorithm $\mathcal{A}$ that is given a random data set $\boldsymbol{S}$ and returns a differentially private output query $\boldsymbol{\phi} = \mathcal{A}(\boldsymbol{S})$. Note that $\boldsymbol{\phi}$ is random both due to the internal randomness of $\mathcal{A}$ and the randomness of the data $\boldsymbol{S}$. This picture is the starting point of our analysis, and allows us to study the generalization properties of queries which are *generated* by differentially private algorithms, rather than estimates returned by them.

Our results then follow from a strong connection we make between *differential privacy* and *generalization*, which will likely have applications beyond those that we explore in this paper. At a high level, we prove that if $\mathcal{A}$ is a differentially private algorithm then the empirical average of a function that it outputs on a random dataset will be close to the true expectation of the function with high probability[2] (over the choice of the dataset and the randomness of $\mathcal{A}$). More formally, for a dataset $S = (x_1, \ldots, x_n)$ and a function $\psi : \mathcal{X} \to [0, 1]$, let $\mathcal{E}_S[\psi] = \frac{1}{n}\sum_{i=1}^{n} \psi(x_i)$ denote the empirical average of ψ. We denote a random dataset chosen from $\mathcal{P}^n$ by $\boldsymbol{S}$. For any fixed function ψ, the empirical average $\mathcal{E}_{\boldsymbol{S}}[\psi]$ is strongly concentrated around its expectation $\mathcal{P}[\psi]$. However, this statement is no longer true if ψ is allowed to depend on $\boldsymbol{S}$ (which is what happens if we choose functions adaptively, using previous estimates on $\boldsymbol{S}$). However for a hypothesis output by a differentially private $\mathcal{A}$ on $\boldsymbol{S}$ (denoted by $\boldsymbol{\phi} = \mathcal{A}(\boldsymbol{S})$), we show that $\mathcal{E}_{\boldsymbol{S}}[\boldsymbol{\phi}]$ is close to $\mathcal{P}[\boldsymbol{\phi}]$ with high probability.

High probability bounds are necessary to ensure that valid answers can be given to an exponentially large number of queries. To prove these bounds, we show that differential privacy roughly preserves the moments of $\mathcal{E}_{\boldsymbol{S}}[\boldsymbol{\phi}]$ even when conditioned on $\boldsymbol{\phi} = \psi$ for any fixed ψ. Now using strong concentration of the k-th moment of $\mathcal{E}_{\boldsymbol{S}}[\psi]$ around $\mathcal{P}[\psi]^k$, we can obtain that $\mathcal{E}_{\boldsymbol{S}}[\boldsymbol{\phi}]$ is concentrated around $\mathcal{P}[\boldsymbol{\phi}]$. Such an argument works only for $(\varepsilon, 0)$-differential privacy due to conditioning on the

[2] A weaker connection that gives closeness *in expectation* over the dataset and algorithm's randomness was known to some experts and is considered folklore. We give a more detailed comparison in Sec. 1.4 and Sec. 2.1.

event $\boldsymbol{\phi} = \psi$ which might have arbitrarily low probability. We use a more delicate conditioning to obtain the extension to (ε, δ)-differential privacy. We note that (ε, δ)-differential privacy is necessary to obtain the stronger bounds that we use for Theorems 1 and 2.

We give an alternative, simpler proof for $(\varepsilon, 0)$-differential privacy that, in addition, extends this connection beyond expectations of functions. We consider a differentially private algorithm $\mathcal{A}$ that maps a database $\boldsymbol{S} \sim \mathcal{P}^n$ to elements from some arbitrary range Z. Our proof shows that if we have a collection of events $R(y)$ defined over databases, one for each element $y \in Z$, and each event is individually unlikely in the sense that for all y, the probability that $\boldsymbol{S} \in R(y)$ is small, then the probability remains small that $\boldsymbol{S} \in R(\boldsymbol{Y})$, where $\boldsymbol{Y} = \mathcal{A}(\boldsymbol{S})$. Note that this statement involves a re-ordering of quantifiers. The hypothesis of the theorem says that the probability of event $R(y)$ is small for each y, where the randomness is taken over the choice of database $\boldsymbol{S} \sim \mathcal{P}^n$, which is independent of y. The conclusion says that the probability of $R(\boldsymbol{Y})$ remains small, even though $\boldsymbol{Y}$ is chosen as a function of $\boldsymbol{S}$, and so is no longer independent. The upshot of this result is that *adaptive* analyses, if performed via a differentially private algorithm, can be thought of (almost) as if they were non-adaptive, with the data being drawn *after* all of the decisions in the analysis are fixed.

To prove this result we note that it would suffice to establish that for every $y \in Z$, $\mathbb{P}[\boldsymbol{S} \in R(y) \mid \boldsymbol{Y} = y]$ is not much larger than $\mathbb{P}[\boldsymbol{S} \in R(y)]$. By Bayes' rule, for every dataset S,

$$\frac{\mathbb{P}[\boldsymbol{S} = S \mid \boldsymbol{Y} = y]}{\mathbb{P}[\boldsymbol{S} = S]} = \frac{\mathbb{P}[\boldsymbol{Y} = y \mid \boldsymbol{S} = S]}{\mathbb{P}[\boldsymbol{Y} = y]}.$$

Therefore, to bound the ratio of $\mathbb{P}[\boldsymbol{S} \in R(y) \mid \boldsymbol{Y} = y]$ to $\mathbb{P}[\boldsymbol{S} \in R(y)]$ it is sufficient to bound the ratio of $\mathbb{P}[\boldsymbol{Y} = y \mid \boldsymbol{S} = S]$ to $\mathbb{P}[\boldsymbol{Y} = y]$ for *most* $S \in R(y)$. Differential privacy implies that $\mathbb{P}[\boldsymbol{Y} = y \mid \boldsymbol{S} = S]$ does not change fast as a function of S. From here, using McDiarmid's concentration inequality, we obtain that $\mathbb{P}[\boldsymbol{Y} = y \mid \boldsymbol{S} = S]$ is strongly concentrated around its mean, which is exactly $\mathbb{P}[\boldsymbol{Y} = y]$.

1.4 Related Work

Numerous techniques have been developed by statisticians to address common special cases of adaptive data analysis. Most of them address a single round of adaptivity such as variable selection followed by regression on selected variables or model selection followed by testing and are optimized for specific inference procedures (the literature is too vast to adequately cover here, see Ch. 7 in [HTF09] for a textbook introduction and [TT15] for a survey of some recent work). In contrast, our framework addresses multiple stages of adaptive decisions, possible lack of a predetermined analysis protocol and is not restricted to any specific procedures.

The traditional perspective on why adaptivity in data analysis invalidates the significance levels of statistical procedures given for the non-adaptive case is that one ends up disregarding all the other possible procedures or tests that would have been performed had the data been different (see *e.g.* [SNS11]). It is well-known that when performing multiple tests on the same data one cannot use significance levels of individual tests and instead it is necessary to control measures such as the false discovery rate [BH95]. This view makes it necessary to explicitly account for all the possible ways to perform the analysis in order to provide validity guarantees for the adaptive analysis. While this approach might be possible in simpler studies, it is technically challenging and often impractical in more complicated analyses [GL14].

There are procedures for controlling false discovery in a sequential setting in which tests arrive one-by-one [FS08, ANR11, AR14]. However the analysis of such tests crucially depends on tests maintaining their statistical properties despite conditioning on previous outcomes. It is therefore unsuitable for the problem we consider here, in which we place no restrictions on the analyst.

The classical approach in theoretical machine learning to ensure that empirical estimates generalize to the underlying distribution is based on the various notions of complexity of the set of functions output by the algorithm, most notably the VC dimension (see [KV94] or [SSBD14] for a textbook introduction). If one has a sample of data large enough to guarantee generalization for all functions in some class of bounded complexity, then it does not matter whether the data analyst chooses functions in this class adaptively or non-adaptively. Our goal, in contrast, is to prove generalization bounds *without* making any assumptions about the class from which the analyst can choose query functions. In this case the adaptive setting is very different from the non-adaptive setting.

An important line of work [BE02, MNPR06, PRMN04, SSSSS10] establishes connections between the *stability* of a learning algorithm and its ability to generalize. Stability is a measure of how much the output of a learning algorithm is perturbed by changes to its input. It is known that certain stability notions are necessary and sufficient for generalization. Unfortunately, the stability notions considered in these prior works are not robust to post-processing, and so the stability of a query answering procedure would not guarantee the stability of the query *generating* procedure used by an arbitrary adaptive analyst. They also do not compose in the sense that running multiple stable algorithms sequentially and adaptively may result in a procedure that is not stable. Differential privacy is stronger than these previously studied notions of stability, and in particular enjoys strong post-processing and composition guarantees. This provides a calculus for building up complex algorithms that satisfy stability guarantees sufficient to give generalization. Past work has considered the generalization properties of one-shot learning procedures. Our work can in part be interpreted as showing that differential privacy implies generalization in the adaptive setting, and beyond the framework of learning.

Differential privacy emerged from a line of work [DN03, DN04, BDMN05], culminating in the definition given by [DMNS06]. It defines a stability property of an algorithm developed in the context of data privacy. There is a very large body of work designing differentially private algorithms for various data analysis tasks, some of which we leverage in our applications. Most crucially, it is known how to accurately answer *exponentially many* adaptively chosen queries on a fixed dataset while preserving differential privacy [RR10, HR10], which is what yields the main application in our paper, when combined with our main theorem. See [Dwo11] for a short survey and [DR14] for a textbook introduction to differential privacy.

For differentially private algorithms that output a hypothesis it has been known as folklore that differential privacy implies stability of the hypothesis to replacing (or removing) an element of the input dataset. Such stability is long known to imply generalization *in expectation* (*e.g.* [SSSSS10]). See Section 2.1 for more details. Our technique can be seen as a substantial strengthening of this observation: from expectation to high probability bounds (which is crucial for answering many queries), from pure to approximate differential privacy (which is crucial for our improved efficient algorithms), and beyond the expected error of a hypothesis.

Further Developments: Our work has attracted substantial interest to the problem of statistical validity in adaptive data analysis and its relationship to differential privacy. Hardt and Ullman [HU14] and Steinke and Ullman [SU14] have proven complementary *computational* lower bounds for

the problem formulated in this work. They show that, under standard cryptographic assumptions, the exponential running time of the algorithm instantiating our main result is unavoidable. Specifically, that the square-root dependence on the number of queries in the sample complexity of our efficient algorithm is nearly optimal, among all computationally efficient mechanisms for answering arbitrary statistical queries.

In [DFH+15a] we discuss approaches to the problem of adaptive data analysis more generally. We demonstrate how differential privacy and description-length-based analyses can be used in this context. In particular, we show that the bounds on n_1 obtained in Theorem 1 can also be obtained by analyzing the transcript of the median mechanism for query answering [RR10] (even without adding noise). Further, we define a notion of approximate max-information between the dataset and the output of the analysis that ensures generalization with high probability, composes adaptively and unifies (pure) differential privacy and description-length-based analyses. We also demonstrate an application of these techniques to the problem of reusing the holdout (or testing) dataset. An overview of this work and [DFH+15a] intended for a broad scientific audience appears in [DFH+15b].

Blum and Hardt [BH15] give an algorithm for reusing the holdout dataset specialized to the problem of maintaining an accurate leaderboard for a machine learning competition (such as those organized by Kaggle Inc. and discussed earlier). Their generalization analysis is based on the description length of the algorithm's transcript.

Our results for approximate ($\delta > 0$) differential privacy apply only to statistical queries (see Thm. 10). Bassily, Nissim, Smith, Steinke, Stemmer and Ullman [BNS+15] give a novel, elegant analysis of the $\delta > 0$ case that gives an exponential improvement in the dependence on δ and generalizes it to arbitrary low-sensitivity queries. This leads to stronger bounds on sample complexity that remove an $O(\sqrt{\log(m)/\tau})$ factor from the bounds on n_0 we give in Theorems 1 and 2. It also implies a similar improvement and generalization to low-sensitivity queries in the reusable holdout application [DFH+15a].

Another implication of our work is that composition and post-processing properties (which are crucial in the adaptive setting) can be ensured by measuring the effect of data analysis on the probability space of the analysis outcomes. Several additional techniques of this type have been recently analyzed. Bassily *et al.* [BNS+15] show that generalization in expectation (as discussed in Cor. 7) can also be obtained from two additional notions of stability: KL-stability and TV-stability that bound the KL-divergence and total variation distance between output distributions on adjacent datasets, respectively. Russo and Zou [RZ15] show that generalization in expectation can be derived by bounding the mutual information between the dataset and the output of analysis. They give applications of their approach to analysis of adaptive feature selection procedures. We note that these techniques do not imply high-probability generalization bounds that we obtain here and in [DFH+15a].

2 Preliminaries

Let $\mathcal{P}$ be a distribution over a discrete universe $\mathcal{X}$ of possible data points. For a function $\psi \colon \mathcal{X} \to [0,1]$ let $\mathcal{P}[\psi] = \mathbb{E}_{x\sim\mathcal{P}}[\psi(x)]$. Given a dataset $S = (x_1, \ldots, x_n)$, a natural estimator of $\mathcal{P}[\psi]$ is the empirical average $\frac{1}{n}\sum_{i=1}^{n}\psi(x_i)$. We let $\mathcal{E}_S$ denote the empirical distribution that assigns weight $1/n$ to each of the data points in S and thus $\mathcal{E}_S[\psi]$ is equal to the empirical average of ψ on S.

Definition 4. *A* statistical query *is defined by a function $\psi : \mathcal{X} \to [0,1]$ and tolerance τ. For distribution $\mathcal{P}$ over $\mathcal{X}$ a valid response to such a query is any value v such that $|v - \mathcal{P}(\psi)| \leq \tau$.*

The standard Hoeffding bound implies that for a fixed query function (chosen independently of the data) the probability over the choice of the dataset that $\mathcal{E}_S[\psi]$ has error greater than τ is at most $2 \cdot \exp(-2\tau^2 n)$. This implies that an exponential in n number of statistical queries can be evaluated within τ as long as the hypotheses do not depend on the data.

We now formally define differential privacy. We say that datasets S, S' are *adjacent* if they differ in a single element.

Definition 5. [DMNS06, DKM+06] *A randomized algorithm $\mathcal{A}$ with domain $\mathcal{X}^n$ is (ε, δ)-differentially private if for all $\mathcal{O} \subseteq \mathrm{Range}(\mathcal{A})$ and for all pairs of adjacent datasets $S, S' \in \mathcal{X}^n$:*

$$\mathbb{P}[\mathcal{A}(S) \in \mathcal{O}] \leq \exp(\varepsilon)\, \mathbb{P}[\mathcal{A}(S') \in \mathcal{O}] + \delta,$$

where the probability space is over the coin flips of the algorithm $\mathcal{A}$. The case when $\delta = 0$ is sometimes referred to as pure *differential privacy, and in this case we may say simply that $\mathcal{A}$ is ε-differentially private.*

Appendix B contains additional background that we will need later on.

2.1 Review of the known connection between privacy and generalization

We now briefly summarize the basic connection between differential privacy and generalization that is considered folklore. This connection follows from an observation that differential privacy implies stability to replacing a single sample in a dataset together with known connection between stability and *on-average* generalization. We first state the form of stability that is immediately implied by the definition of differential privacy. For simplicity, we state it only for $[0,1]$-valued functions. The extension to any other bounded range is straightforward.

Lemma 6. *Let $\mathcal{A}$ be an (ϵ, δ)-differentially private algorithm ranging over functions from $\mathcal{X}$ to $[0,1]$. For any pair of adjacent datasets S and S' and $x \in \mathcal{X}$:*

$$\mathbb{E}\left[\mathcal{A}(S)(x)\right] \leq e^{\varepsilon} \cdot \mathbb{E}\left[\mathcal{A}(S')(x)\right] + \delta,$$

and, in particular,

$$\left|\mathbb{E}\left[\mathcal{A}(S)(x)\right] - \mathbb{E}\left[\mathcal{A}(S')(x)\right]\right| \leq e^{\varepsilon} - 1 + \delta. \tag{1}$$

Algorithms satisfying equation (1) are referred to as *strongly-uniform-replace-one stable* with rate $(e^{\varepsilon} - 1 + \delta)$ by Shalev-Schwartz *et al.* [SSSSS10]. It is easy to show and is well-known that replace-one stability implies generalization in expectation, referred to as *on-average* generalization [SSSSS10, Lemma 11]. In our case this connection immediately gives the following corollary.

Corollary 7. *Let $\mathcal{A}$ be an (ϵ, δ)-differentially private algorithm ranging over functions from $\mathcal{X}$ to $[0,1]$, let $\mathcal{P}$ be a distribution over $\mathcal{X}$ and let $\boldsymbol{S}$ be an independent random variable distributed according to $\mathcal{P}^n$. Then*

$$\left|\mathbb{E}[\mathcal{E}_{\boldsymbol{S}}[\mathcal{A}(\boldsymbol{S})]] - \mathbb{E}[\mathcal{P}[\mathcal{A}(\boldsymbol{S})]]\right| \leq e^{\varepsilon} - 1 + \delta.$$

This corollary was observed in the context of functions expressing the loss of the hypothesis output by a (private) learning algorithm, that is, $\phi(x) = L(h(x), x)$, where x is a sample (possibly including a label), h is a hypothesis function and L is a non-negative loss function. When applied to such a function, Corollary 7 implies that the expected true loss of a hypothesis output by an

(ε, δ)-differentially private algorithm is at most $e^{\varepsilon} - 1 + \delta$ larger than the expected empirical loss of the output hypothesis, where the expectation is taken over the random dataset and the randomness of the algorithm. A special case of this corollary is stated in a recent work of Bassily *et al.* [BST14]. More recently, Wang *et al.* [WLF15] have similarly used the stability of differentially private learning algorithms to show a general equivalence of differentially private learning and differentially private empirical loss minimization.

A standard way to obtain a high-probability bound from a bound on expectation in Corollary 7 is to use Markov's inequality. Using this approach, a bound that holds with probability $1 - \beta$ will require a polynomial dependence of the sample size on $1/\beta$. While this might lead to a useful bound when the expected empirical loss is small it is less useful in the common scenario when the empirical loss is relatively large. In contrast, our results in Sections 3 and 4 directly imply generalization bounds with logarithmic dependence of the sample size on $1/\beta$. For example, in Theorem 9 we show that for any $\varepsilon, \beta > 0$ and $n \geq O(\ln(1/\beta)/\varepsilon^2)$, the output of an ε-differentially private algorithm $\mathcal{A}$ satisfies $\mathbb{P}\left[|\mathcal{P}[\mathcal{A}(\boldsymbol{S})] - \mathcal{E}_{\boldsymbol{S}}[\mathcal{A}(\boldsymbol{S})]| > 2\varepsilon\right] \leq \beta$.

3 Differential Privacy and Preservation of Moments

We now prove that if a function $\boldsymbol{\phi}$ is output by an (ε, δ)-differentially private algorithm $\mathcal{A}$ on input of a random dataset $\boldsymbol{S}$ drawn from $\mathcal{P}^n$, then the average of $\boldsymbol{\phi}$ on $\boldsymbol{S}$, that is, $\mathcal{E}_{\boldsymbol{S}}[\boldsymbol{\phi}]$, is concentrated around its true expectation $\mathcal{P}[\boldsymbol{\phi}]$.

The statement we wish to prove is nontrivial due to the apparent dependency between the function $\boldsymbol{\phi}$ and the dataset $\boldsymbol{S}$ that arises because $\boldsymbol{\phi} = \mathcal{A}(\boldsymbol{S})$. If instead $\boldsymbol{\phi}$ was evaluated on a fresh dataset $\boldsymbol{T}$ drawn independently of $\boldsymbol{\phi}$, then indeed we would have $\mathbb{E}\,\mathcal{E}_{\boldsymbol{T}}[\boldsymbol{\phi}] = \mathcal{P}[\boldsymbol{\phi}]$. At a high level, our goal is therefore to resolve the dependency between $\boldsymbol{\phi}$ and $\boldsymbol{S}$ by relating the random variable $\mathcal{E}_{\boldsymbol{S}}[\boldsymbol{\phi}]$ to the random variable $\mathcal{E}_{\boldsymbol{T}}[\boldsymbol{\phi}]$. To argue that these random variables are close with high probability we relate the moments of $\mathcal{E}_{\boldsymbol{S}}[\boldsymbol{\phi}]$ to the moments of $\mathcal{E}_{\boldsymbol{T}}[\boldsymbol{\phi}]$. The moments of $\mathcal{E}_{\boldsymbol{T}}[\boldsymbol{\phi}]$ are relatively easy to bound using standard techniques.

Our proof is easier to execute when $\delta = 0$ and we start with this case for the sake of exposition.

3.1 Simpler case where $\delta = 0$

Our main technical tool relates the moments of the random variables that we are interested in.

Lemma 8. *Assume that $\mathcal{A}$ is an $(\epsilon, 0)$-differentially private algorithm ranging over functions from $\mathcal{X}$ to $[0,1]$. Let $\boldsymbol{S}, \boldsymbol{T}$ be independent random variables distributed according to $\mathcal{P}^n$. For any function $\psi : \mathcal{X} \to [0,1]$ in the support of $\mathcal{A}(\boldsymbol{S})$,*

$$\mathbb{E}\left[\mathcal{E}_{\boldsymbol{S}}[\boldsymbol{\phi}]^k \;\middle|\; \boldsymbol{\phi} = \psi\right] \leq e^{k\varepsilon} \cdot \mathbb{E}\left[\mathcal{E}_{\boldsymbol{T}}[\psi]^k\right]. \tag{2}$$

Proof. We use I to denote a k-tuple of indices $(i_1, \ldots, i_k) \in [n]^k$ and use $\boldsymbol{I}$ to denote a k-tuple chosen randomly and uniformly from $[n]^k$. For a data set $T = (y_1, \ldots, y_n)$ we denote by $\Pi_T^I(\psi) = \prod_{j \in [k]} \psi(y_{i_j})$. We first observe that for any ψ,

$$\mathcal{E}_T[\psi]^k = \mathbb{E}[\Pi_T^{\boldsymbol{I}}(\psi)]. \tag{3}$$

For two datasets $S, T \in \mathcal{X}^n$, let $S_{I \leftarrow T}$ denote the data set in which for every $j \in [k]$, element i_j in S is replaced with the corresponding element from T. We fix I. Note that the random variable

$\boldsymbol{S}_{I\leftarrow\boldsymbol{T}}$ is distributed according to $\mathcal{P}^n$ and therefore

$$\begin{aligned}
\mathbb{E}\left[\Pi^I_{\boldsymbol{S}}(\boldsymbol{\phi}) \mid \boldsymbol{\phi}=\psi\right] &= \mathbb{E}\left[\Pi^I_{\boldsymbol{S}_{I\leftarrow\boldsymbol{T}}}(\mathcal{A}(\boldsymbol{S}_{I\leftarrow\boldsymbol{T}})) \mid \mathcal{A}(\boldsymbol{S}_{I\leftarrow\boldsymbol{T}})=\psi\right] \\
&= \mathbb{E}\left[\Pi^I_{\boldsymbol{T}}(\mathcal{A}(\boldsymbol{S}_{I\leftarrow\boldsymbol{T}})) \mid \mathcal{A}(\boldsymbol{S}_{I\leftarrow\boldsymbol{T}})=\psi\right] \\
&= \int_0^1 \frac{\mathbb{P}\left[\Pi^I_{\boldsymbol{T}}(\mathcal{A}(\boldsymbol{S}_{I\leftarrow\boldsymbol{T}}))\geq t \text{ and } \mathcal{A}(\boldsymbol{S}_{I\leftarrow\boldsymbol{T}})=\psi\right]}{\mathbb{P}\left[\mathcal{A}(\boldsymbol{S}_{I\leftarrow\boldsymbol{T}})=\psi\right]}\mathrm{d}t \\
&= \int_0^1 \frac{\mathbb{P}\left[\Pi^I_{\boldsymbol{T}}(\mathcal{A}(\boldsymbol{S}_{I\leftarrow\boldsymbol{T}}))\geq t \text{ and } \mathcal{A}(\boldsymbol{S}_{I\leftarrow\boldsymbol{T}})=\psi\right]}{\mathbb{P}\left[\boldsymbol{\phi}=\psi\right]}\mathrm{d}t \qquad (4)
\end{aligned}$$

Now for any fixed t, S and T consider the event $\Pi^I_T(\mathcal{A}(S))\geq t$ and $\mathcal{A}(S)=\psi$ (defined on the range of $\mathcal{A}$). Data sets S and $S_{I\leftarrow T}$ differ in at most k elements. Therefore, by the ε-differential privacy of $\mathcal{A}$ and Lemma 24, the distribution $\mathcal{A}(S)$ and the distribution $\mathcal{A}(S_{I\leftarrow T})$ satisfy:

$$\mathbb{P}\left[\Pi^I_T(\mathcal{A}(S_{I\leftarrow T}))\geq t \text{ and } \mathcal{A}(S_{I\leftarrow T})=\psi\right] \leq e^{k\varepsilon}\cdot\mathbb{P}\left[\Pi^I_T(\mathcal{A}(S))\geq t \text{ and } \mathcal{A}(S)=\psi\right].$$

Taking the probability over $\boldsymbol{S}$ and $\boldsymbol{T}$ we get:

$$\mathbb{P}\left[\Pi^I_{\boldsymbol{T}}(\mathcal{A}(\boldsymbol{S}_{I\leftarrow\boldsymbol{T}}))\geq t \text{ and } \mathcal{A}(\boldsymbol{S}_{I\leftarrow\boldsymbol{T}})=\psi\right] \leq e^{k\varepsilon}\cdot\mathbb{P}\left[\Pi^I_T(\boldsymbol{\phi})\geq t \text{ and } \boldsymbol{\phi}=\psi\right].$$

Substituting this into eq. (4) we get

$$\begin{aligned}
\mathbb{E}\left[\Pi^I_{\boldsymbol{S}}(\boldsymbol{\phi}) \mid \boldsymbol{\phi}=\psi\right] &\leq e^{k\varepsilon}\int_0^1 \frac{\mathbb{P}\left[\Pi^I_{\boldsymbol{T}}(\boldsymbol{\phi})\geq t \text{ and } \boldsymbol{\phi}=\psi\right]}{\mathbb{P}\left[\boldsymbol{\phi}=\psi\right]}dt \\
&= e^{k\varepsilon}\,\mathbb{E}\left[\Pi^I_{\boldsymbol{T}}(\boldsymbol{\phi}) \mid \boldsymbol{\phi}=\psi\right] \\
&= e^{k\varepsilon}\,\mathbb{E}\left[\Pi^I_{\boldsymbol{T}}(\psi) \mid \boldsymbol{\phi}=\psi\right] \\
&= e^{k\varepsilon}\,\mathbb{E}\left[\Pi^I_{\boldsymbol{T}}(\psi)\right]
\end{aligned}$$

Taking the expectation over $\boldsymbol{I}$ and using eq. (3) we obtain that

$$\mathbb{E}\left[\mathcal{E}_{\boldsymbol{S}}[\boldsymbol{\phi}]^k \;\middle|\; \boldsymbol{\phi}=\psi\right] \leq e^{k\varepsilon}\,\mathbb{E}\left[\mathcal{E}_{\boldsymbol{T}}[\psi]^k\right],$$

completing the proof of the lemma. □

We now turn our moment inequality into a theorem showing that $\mathcal{E}_{\boldsymbol{S}}[\boldsymbol{\phi}]$ is concentrated around the true expectation $\mathcal{P}[\boldsymbol{\phi}]$.

Theorem 9. *Let $\mathcal{A}$ be an ε-differentially private algorithm that given a dataset S outputs a function from $\mathcal{X}$ to $[0,1]$. For any distribution $\mathcal{P}$ over $\mathcal{X}$ and random variable $\boldsymbol{S}$ distributed according to $\mathcal{P}^n$ we let $\boldsymbol{\phi}=\mathcal{A}(\boldsymbol{S})$. Then for any $\beta>0,\tau>0$ and $n\geq 12\ln(4/\beta)/\tau^2$, setting $\varepsilon\leq\tau/2$ ensures $\mathbb{P}\left[|\mathcal{P}[\boldsymbol{\phi}]-\mathcal{E}_{\boldsymbol{S}}[\boldsymbol{\phi}]|>\tau\right]\leq\beta$, where the probability is over the randomness of $\mathcal{A}$ and $\boldsymbol{S}$.*

Proof. Consider an execution of $\mathcal{A}$ with $\varepsilon=\tau/2$ on a data set $\boldsymbol{S}$ of size $n\geq 12\ln(4/\beta)/\tau^2$. By Lemma 29 we obtain that RHS of our bound in Lemma 8 is at most $e^{\varepsilon k}\mathcal{M}_k[B(n,\mathcal{P}[\psi])]$. We use Lemma 31 with $\varepsilon=\tau/2$ and $k=4\ln(4/\beta)/\tau$ (noting that the assumption $n\geq 12\ln(4/\beta)/\tau^2$ ensures the necessary bound on n) to obtain that

$$\mathbb{P}\left[\mathcal{E}_{\boldsymbol{S}}[\boldsymbol{\phi}]\geq\mathcal{P}[\psi]+\tau \mid \boldsymbol{\phi}=\psi\right]\leq\beta/2.$$

This holds for every ψ in the range of $\mathcal{A}$ and therefore,

$$\mathbb{P}\left[\mathcal{E}_{\boldsymbol{S}}[\boldsymbol{\phi}] \geq \mathcal{P}[\boldsymbol{\phi}] + \tau\right] \leq \beta/2.$$

We can apply the same argument to the function $1 - \boldsymbol{\phi}$ to obtain that

$$\mathbb{P}\left[\mathcal{E}_{\boldsymbol{S}}[\boldsymbol{\phi}] \leq \mathcal{P}[\boldsymbol{\phi}] - \tau\right] \leq \beta/2.$$

A union bound over the above inequalities implies the claim. □

3.2 Extension to $\delta > 0$

We now extend our proof to the case when $\mathcal{A}$ satisfies (ε, δ)-differential privacy for sufficiently small but nonzero $\delta > 0$. The main difficulty in extending the previous proof is that the condition $\{\boldsymbol{\phi} = \psi\}$ appearing in Lemma 8 may have arbitrarily small probability. A simple extension of the previous proof would lead to an error of $\delta / \mathbb{P}[\boldsymbol{\phi} = \psi]$. We avoid this issue by using a more carefully chosen condition. Specifically, instead of restricting $\boldsymbol{\phi}$ to be equal to a particular function ψ, we only constrain $\mathcal{P}[\boldsymbol{\phi}]$ to be in a certain interval of length τ. This conditioning still gives us enough information about $\boldsymbol{\phi}$ in order to control $\mathcal{E}_{\boldsymbol{T}}[\boldsymbol{\phi}]$, while allowing us to ignore events of exceedingly small probability.

Theorem 10. *Let $\mathcal{A}$ be an (ε, δ)-differentially private algorithm that given a dataset S outputs a function from $\mathcal{X}$ to $[0,1]$. For any distribution $\mathcal{P}$ over $\mathcal{X}$ and random variable $\boldsymbol{S}$ distributed according to $\mathcal{P}^n$ we let $\boldsymbol{\phi} = \mathcal{A}(\boldsymbol{S})$. Then for any $\beta > 0, \tau > 0$ and $n \geq 48\ln(4/\beta)/\tau^2$, setting $\varepsilon \leq \tau/4$ and $\delta = \exp(-4 \cdot \ln(8/\beta)/\tau)$ ensures $\mathbb{P}\left[|\mathcal{P}[\boldsymbol{\phi}] - \mathcal{E}_{\boldsymbol{S}}[\boldsymbol{\phi}]| > \tau\right] \leq \beta$, where the probability is over the randomness of $\mathcal{A}$ and $\boldsymbol{S}$.*

Proof. We use the notation from the proof of Theorem 9 and consider an execution of $\mathcal{A}$ with ε and δ satisfying the conditions of the theorem.

Let $L = \lceil 1/\tau \rceil$. For a value $\ell \in [L]$ we use B_ℓ to denote the interval set $[(\ell - 1)\tau, \ell\tau]$.

We say that $\ell \in [L]$ is *heavy* if $\mathbb{P}\left[\mathcal{P}[\boldsymbol{\phi}] \in B_\ell\right] \geq \beta/(4L)$ and we say that ℓ is *light* otherwise. The key claim that we prove is an upper bound on the k-th moment of $\mathcal{E}_{\boldsymbol{S}}[\boldsymbol{\phi}]$ for heavy ℓ's:

$$\mathbb{E}\left[\mathcal{E}_{\boldsymbol{S}}[\boldsymbol{\phi}]^k \;\middle|\; \mathcal{P}[\boldsymbol{\phi}] \in B_\ell\right] \leq e^{k\varepsilon} \cdot \mathcal{M}_k[B(n, \tau\ell)] + \delta e^{(k-1)\varepsilon} \cdot 4L/\beta. \tag{5}$$

We use the same decomposition of the k-th moment as before:

$$\mathbb{E}\left[\mathcal{E}_{\boldsymbol{S}}[\boldsymbol{\phi}]^k \;\middle|\; \mathcal{P}[\boldsymbol{\phi}] \in B_\ell\right] = \mathbb{E}\left[\Pi^{\boldsymbol{I}}_{\boldsymbol{S}}(\boldsymbol{\phi}) \;\middle|\; \mathcal{P}[\boldsymbol{\phi}] \in B_\ell\right].$$

Now for a fixed $I \in [n]^k$, exactly as in eq. (4), we obtain

$$\mathbb{E}\left[\Pi^I_{\boldsymbol{S}}(\boldsymbol{\phi}) \mid \mathcal{P}[\boldsymbol{\phi}] \in B_\ell\right] = \int_0^1 \frac{\mathbb{P}\left[\Pi^I_{\boldsymbol{T}}(\mathcal{A}(\boldsymbol{S}_{I \leftarrow \boldsymbol{T}})) \geq t \text{ and } \mathcal{P}[\mathcal{A}(\boldsymbol{S}_{I \leftarrow \boldsymbol{T}})] \in B_\ell\right]}{\mathbb{P}\left[\mathcal{P}[\boldsymbol{\phi}] \in B_\ell\right]} dt \tag{6}$$

Now for fixed values of t, S and T we consider the event $\Pi^I_T(\mathcal{A}(S)) \geq t$ and $\mathcal{P}[\mathcal{A}(S)] \in B_\ell$ defined on the range of $\mathcal{A}$. Datasets S and $S_{I \leftarrow T}$ differ in at most k elements. Therefore, by the (ε, δ)-differential

privacy of $\mathcal{A}$ and Lemma 24, the distribution over the output of $\mathcal{A}$ on input S and the distribution over the output of $\mathcal{A}$ on input $S_{I\leftarrow T}$ satisfy:

$$\begin{aligned}&\mathbb{P}\left[\Pi_T^I(\mathcal{A}(S_{I\leftarrow T}))\geq t \text{ and } \mathcal{P}[\mathcal{A}(S_{I\leftarrow T})]\in B_\ell\right]\\ \leq e^{k\varepsilon}\cdot\mathbb{P}&\left[\Pi_T^I(\mathcal{A}(S))\geq t \text{ and } \mathcal{P}[\mathcal{A}(S)]\in B_\ell\right]+e^{(k-1)\varepsilon}\delta.\end{aligned}$$

Taking the probability over $\boldsymbol{S}$ and $\boldsymbol{T}$ and substituting this into eq. (6) we get

$$\begin{aligned}\mathbb{E}\left[\Pi_{\boldsymbol{S}}^I(\boldsymbol{\phi})\mid\mathcal{P}[\boldsymbol{\phi}]\in B_\ell\right]&\leq e^{k\varepsilon}\int_0^1\frac{\mathbb{P}\left[\Pi_{\boldsymbol{T}}^I(\boldsymbol{\phi})\geq t \text{ and } \mathcal{P}[\boldsymbol{\phi}]\in B_\ell\right]}{\mathbb{P}\left[\mathcal{P}[\boldsymbol{\phi}]\in B_\ell\right]}dt+\frac{e^{(k-1)\varepsilon}\delta}{\mathbb{P}\left[\mathcal{P}[\boldsymbol{\phi}]\in B_\ell\right]}\\ &=e^{k\varepsilon}\,\mathbb{E}\left[\Pi_{\boldsymbol{T}}^I(\boldsymbol{\phi})\mid\mathcal{P}[\boldsymbol{\phi}]\in B_\ell\right]+\frac{e^{(k-1)\varepsilon}\delta}{\mathbb{P}\left[\mathcal{P}[\boldsymbol{\phi}]\in B_\ell\right]}\end{aligned}$$

Taking the expectation over $\boldsymbol{I}$ and using eq. (3) we obtain:

$$\mathbb{E}\left[\mathcal{E}_{\boldsymbol{S}}[\boldsymbol{\phi}]^k\;\middle|\;\mathcal{P}[\boldsymbol{\phi}]\in B_\ell\right]\leq e^{k\varepsilon}\,\mathbb{E}\left[\mathcal{E}_{\boldsymbol{T}}[\boldsymbol{\phi}]^k\;\middle|\;\mathcal{P}[\boldsymbol{\phi}]\in B_\ell\right]+\frac{e^{(k-1)\varepsilon}\delta}{\mathbb{P}\left[\mathcal{P}[\boldsymbol{\phi}]\in B_\ell\right]}. \tag{7}$$

Conditioned on $\mathcal{P}[\boldsymbol{\phi}]\in B_\ell$, $\mathcal{P}[\boldsymbol{\phi}]\leq\tau\ell$ and therefore by Lemma 29,

$$\mathbb{E}\left[\mathcal{E}_{\boldsymbol{T}}[\boldsymbol{\phi}]^k\;\middle|\;\mathcal{P}[\boldsymbol{\phi}]\in B_\ell\right]\leq\mathcal{M}_k[B(n,\tau\ell)].$$

In addition, by our assumption, ℓ is heavy, that is $\mathbb{P}\left[\mathcal{P}[\boldsymbol{\phi}]\in B_\ell\right]\geq\beta/(4L)$. Substituting these values into eq. (7) we obtain the claim in eq. (5).

As before, we use Lemma 31 with $\varepsilon=\tau/2$ and $k=4(\tau\ell)\ln(4/\beta)/\tau=4\ell\ln(4/\beta)$ (noting that condition $n\geq 12\ln(4/\beta)/\tau^2$ ensures the necessary bound on n) to obtain that

$$\mathbb{P}\left[\mathcal{E}_{\boldsymbol{S}}[\boldsymbol{\phi}]\geq\tau\ell+\tau\mid\mathcal{P}[\boldsymbol{\phi}]\in B_\ell\right]\leq\beta/2+\frac{\delta e^{(k-1)\varepsilon}\cdot 4L}{\beta(\tau\ell+\tau)^k}, \tag{8}$$

Using condition $\delta=\exp(-2\cdot\ln(4/\beta)/\tau)$ and inequality $\ln(x)\leq x/e$ (for $x>0$) we obtain

$$\begin{aligned}\frac{\delta e^{(k-1)\varepsilon}\cdot 4L}{\beta((\ell+1)\tau)^k}&\leq\frac{\delta\cdot e^{2\ln(4/\beta)}\cdot 4/\tau}{\beta e^{4\ln((\ell+1)\tau)\cdot\ell\ln(4/\beta)}}\\ &\leq\frac{\delta\cdot e^{4\ln(4/\beta)}}{\tau\cdot e^{4\ln((\ell+1)\tau)\cdot\ell\ln(4/\beta)}}\cdot\frac{\beta}{4}\\ &\leq\delta\cdot\exp\left(4\ln(1/((\ell+1)\tau))\cdot\ell\ln(4/\beta)+4\ln(4/\beta)+\ln(1/\tau)\right)\cdot\frac{\beta}{4}\\ &\leq\delta\cdot\exp\left(\frac{4}{e}\cdot\frac{1}{(\ell+1)\tau}\cdot\ell\ln(4/\beta)+4\ln(4/\beta)+\ln(1/\tau)\right)\cdot\frac{\beta}{4}\\ &\leq\delta\cdot\exp\left(\frac{4}{e}\cdot\ln(4/\beta)/\tau+4\ln(4/\beta)+\ln(1/\tau)\right)\cdot\frac{\beta}{4}\\ &\leq\delta\cdot\exp\left(2\cdot\ln(4/\beta)/\tau\right)\cdot\frac{\beta}{4}\leq\beta/4.\end{aligned}$$

Substituting this into eq. (8) we get

$$\mathbb{P}\left[\mathcal{E}_{\boldsymbol{S}}[\boldsymbol{\phi}]\geq\tau\ell+\tau\mid\mathcal{P}[\boldsymbol{\phi}]\in B_\ell\right]\leq 3\beta/4.$$

Note that, conditioned on $\mathcal{P}[\boldsymbol{\phi}] \in B_\ell$, $\mathcal{P}[\boldsymbol{\phi}] \geq \tau(\ell - 1)$, and therefore

$$\mathbb{P}\left[\mathcal{E}_{\boldsymbol{S}}[\boldsymbol{\phi}] \geq \mathcal{P}[\boldsymbol{\phi}] + 2\tau \mid \mathcal{P}[\boldsymbol{\phi}] \in B_\ell\right] \leq 3\beta/4.$$

This holds for every heavy $\ell \in [L]$ and therefore,

$$\begin{aligned}\mathbb{P}\left[\mathcal{E}_{\boldsymbol{S}}[\boldsymbol{\phi}] \geq \mathcal{P}[\boldsymbol{\phi}] + 2\tau\right] &\leq 3\beta/4 + \sum_{\ell \in [L] \text{ is light}} \mathbb{P}\left[\mathcal{P}[\boldsymbol{\phi}] \in B_\ell\right] \\ &\leq 3\beta/4 + L\beta/(4L) = \beta.\end{aligned}$$

Apply the same argument to $1 - \boldsymbol{\phi}$ and use a union bound. We obtain the claim after rescaling τ and β by a factor 2.

□

4 Beyond statistical queries

The previous section dealt with statistical queries. A different way of looking at our results is to define for each function ψ a set $R(\psi)$ containing all datasets S such that ψ is far from the correct value $\mathcal{P}[\psi]$ on S. Formally, $R(\psi) = \{S\colon\ |\mathcal{E}_S[\psi] - \mathcal{P}[\psi]| > \tau\}$. Our results showed that if $\boldsymbol{\phi} = \mathcal{A}(\boldsymbol{S})$ is the output of a differentially private algorithm $\mathcal{A}$ on a random dataset $\boldsymbol{S}$, then $\mathbb{P}[\boldsymbol{S} \in R(\boldsymbol{\phi})]$ is small.

Here we prove a broad generalization that allows the differentially private algorithm to have an arbitrary output space Z. The same conclusion holds for any collection of sets $R(y)$ where $y \in Z$ provided that $\mathbb{P}[\boldsymbol{S} \in R(y)]$ is small for all $y \in Z$.

Theorem 11. *Let $\mathcal{A}$ be an $(\varepsilon, 0)$-differentially private algorithm with range Z. For a distribution $\mathcal{P}$ over $\mathcal{X}$, let $\boldsymbol{S}$ be a random variable drawn from $\mathcal{P}^n$. Let $\boldsymbol{Y} = \mathcal{A}(\boldsymbol{S})$ be the random variable output by $\mathcal{A}$ on input $\boldsymbol{S}$. For each element $y \in Z$ let $R(y) \subseteq \mathcal{X}^n$ be some subset of datasets and assume that $\max_y \mathbb{P}[\boldsymbol{S} \in R(y)] \leq \beta$. Then, for $\varepsilon \leq \sqrt{\frac{\ln(1/\beta)}{2n}}$ we have $\mathbb{P}[\boldsymbol{S} \in R(\boldsymbol{Y})] \leq 3\sqrt{\beta}$.*

Proof. Fix $y \in Z$. We first observe that by Jensen's inequality,

$$\mathop{\mathbb{E}}_{S \sim \mathcal{P}^n}[\ln(\mathbb{P}[\boldsymbol{Y} = y \mid \boldsymbol{S} = S])] \leq \ln\left(\mathop{\mathbb{E}}_{S \sim \mathcal{P}^n}[\mathbb{P}[\boldsymbol{Y} = y \mid \boldsymbol{S} = S]]\right) = \ln(\mathbb{P}[\boldsymbol{Y} = y]).$$

Further, by definition of differential privacy, for two databases S, S' that differ in a single element,

$$\mathbb{P}[\boldsymbol{Y} = y \mid \boldsymbol{S} = S] \leq e^{\varepsilon} \cdot \mathbb{P}[\boldsymbol{Y} = y \mid \boldsymbol{S} = S'].$$

Now consider the function $g(S) = \ln\left(\frac{\mathbb{P}[\boldsymbol{Y}=y \mid \boldsymbol{S}=S]}{\mathbb{P}[\boldsymbol{Y}=y]}\right)$. By the properties above we have that $\mathbb{E}[g(\boldsymbol{S})] \leq \ln(\mathbb{P}[\boldsymbol{Y} = y]) - \ln(\mathbb{P}[\boldsymbol{Y} = y]) = 0$ and $|g(S) - g(S')| \leq \varepsilon$. This, by McDiarmid's inequality (Lemma 28), implies that for any $t > 0$,

$$\mathbb{P}[g(\boldsymbol{S}) \geq \varepsilon t] \leq e^{-2t^2/n}. \tag{9}$$

For an integer $i \geq 1$ let

$$B_i \doteq \left\{ S \;\middle|\; \varepsilon\sqrt{n\ln(2^i/\beta)/2} \leq g(S) \leq \varepsilon\sqrt{n\ln(2^{i+1}/\beta)/2} \right\}$$

and let $B_0 \doteq \{S \mid g(S) \leq \varepsilon\sqrt{n\ln(2/\beta)/2}\}$.

By inequality (9) we have that for $i \geq 1$, $\mathbb{P}[g(\boldsymbol{S}) \geq \varepsilon\sqrt{n\ln(2^i/\beta)/2}] \leq \beta/2^i$. Therefore, for all $i \geq 0$,

$$\mathbb{P}[\boldsymbol{S} \in B_i \cap R(y)] \leq \beta/2^i,$$

where the case of $i = 0$ follows from the assumptions of the lemma.

By Bayes' rule, for every $S \in B_i$,

$$\frac{\mathbb{P}[\boldsymbol{S} = S \mid \boldsymbol{Y} = y]}{\mathbb{P}[\boldsymbol{S} = S]} = \frac{\mathbb{P}[\boldsymbol{Y} = y \mid \boldsymbol{S} = S]}{\mathbb{P}[\boldsymbol{Y} = y]} = \exp(g(S)) \leq \exp\left(\varepsilon\sqrt{n\ln(2^{i+1}/\beta)/2}\right).$$

Therefore,

$$\begin{aligned}
\mathbb{P}[\boldsymbol{S} \in B_i \cap R(y) \mid \boldsymbol{Y} = y] &= \sum_{S \in B_i \cap R(y)} \mathbb{P}[\boldsymbol{S} = S \mid \boldsymbol{Y} = y] \\
&\leq \exp\left(\varepsilon\sqrt{n\ln(2^{i+1}/\beta)/2}\right) \cdot \sum_{S \in B_i \cap R(y)} \mathbb{P}[\boldsymbol{S} = S] \\
&= \exp\left(\varepsilon\sqrt{n\ln(2^{i+1}/\beta)/2}\right) \cdot \mathbb{P}[\boldsymbol{S} \in B_i \cap R(y)] \\
&\leq \exp\left(\varepsilon\sqrt{n\ln(2^{i+1}/\beta)/2} - \ln(2^i/\beta)\right). \qquad (10)
\end{aligned}$$

The condition $\varepsilon \leq \sqrt{\frac{\ln(1/\beta)}{2n}}$ implies that

$$\begin{aligned}
\varepsilon\sqrt{\frac{n\ln(2^{i+1}/\beta)}{2}} - \ln(2^i/\beta) &\leq \sqrt{\frac{\ln(1/\beta)\ln(2^{i+1}/\beta)}{4}} - \ln(2^i/\beta) \\
&\leq \frac{\ln(2^{(i+1)/2}/\beta)}{2} - \ln(2^i/\beta) = -\ln\left(\frac{2^{(3i-1)/4}}{\sqrt{\beta}}\right)
\end{aligned}$$

Substituting this into inequality (10), we get

$$\mathbb{P}[\boldsymbol{S} \in B_i \cap R(y) \mid \boldsymbol{Y} = y] \leq \frac{\sqrt{\beta}}{2^{(3i-1)/4}}.$$

Clearly, $\cup_{i\geq 0} B_i = \mathcal{X}^{[n]}$. Therefore

$$\mathbb{P}[\boldsymbol{S} \in R(y) \mid \boldsymbol{Y} = y] = \sum_{i\geq 0} \mathbb{P}[\boldsymbol{S} \in B_i \cap R(y) \mid \boldsymbol{Y} = y] \leq \sum_{i\geq 0} \frac{\sqrt{\beta}}{2^{(3i-1)/4}} = \sqrt{\beta} \cdot \frac{2^{1/4}}{1 - 2^{-3/4}} \leq 3\sqrt{\beta}.$$

Finally, let $\mathcal{Y}$ denote the distribution of $\boldsymbol{Y}$. Then,

$$\mathbb{P}[\boldsymbol{S} \in R(\boldsymbol{Y})] = \mathop{\mathbb{E}}_{y\sim\mathcal{Y}}[\mathbb{P}[\boldsymbol{S} \in R(y) \mid \boldsymbol{Y} = y]] \leq 3\sqrt{\beta}.$$

□

Our theorem gives a result for statistical queries that achieves the same bound as our earlier result in Theorem 9 up to constant factors in the parameters.

Corollary 12. *Let $\mathcal{A}$ be an ε-differentially private algorithm that outputs a function from $\mathcal{X}$ to $[0,1]$. For a distribution $\mathcal{P}$ over $\mathcal{X}$, let $\boldsymbol{S}$ be a random variable distributed according to $\mathcal{P}^n$ and let $\boldsymbol{\phi} = \mathcal{A}(\boldsymbol{S})$. Then for any $\tau > 0$, setting $\varepsilon \leq \sqrt{\tau^2 - \ln(2)/2n}$ ensures $\mathbb{P}\left[|\mathcal{P}[\boldsymbol{\phi}] - \mathcal{E}_{\boldsymbol{S}}[\boldsymbol{\phi}]| > \tau\right] \leq 3\sqrt{2}e^{-\tau^2 n}$.*

Proof. By the Chernoff bound, for any fixed query function $\psi : \mathcal{X} \to [0,1]$,

$$\mathbb{P}[|\mathcal{P}[\psi] - \mathcal{E}_{\boldsymbol{S}}[\psi]| \geq \tau] \leq 2e^{-2\tau^2 n}.$$

Now, by Theorem 11 for $R(\psi) = \{S \in \mathcal{X}^n \mid |\mathcal{P}[\psi] - \mathcal{E}_{\boldsymbol{S}}[\psi]| > \tau\}$, $\beta = 2e^{-2\tau^2 n}$ and any $\varepsilon \leq \sqrt{\tau^2 - \ln(2)/2n}$,

$$\mathbb{P}\left[|\mathcal{P}[\boldsymbol{\phi}] - \mathcal{E}_{\boldsymbol{S}}[\boldsymbol{\phi}]| > \tau\right] \leq 3\sqrt{2}e^{-\tau^2 n}.$$

□

5 Applications

To obtain algorithms for answering adaptive statistical queries we first note that if for a query function ψ and a dataset S, $|\mathcal{P}[\psi] - \mathcal{E}_S[\psi]| \leq \tau/2$ then we can use an algorithm that outputs a value v that is $\tau/2$-close to $\mathcal{E}_S[\psi]$ to obtain a value that is τ-close to $\mathcal{P}[\psi]$. Differentially private algorithms that for a given dataset S and an adaptively chosen sequence of queries $\phi_1, \ldots, \phi_m$ produce a value close to $\mathcal{E}_S[\phi_i]$ for each query $\phi_i \colon \mathcal{X} \to [0,1]$ have been the subject of intense investigation in the differential privacy literature (see [DR14] for an overview). Such queries are usually referred to as (fractional) *counting queries* or *linear queries* in this context. This allows us to obtain statistical query answering algorithms by using various known differentially private algorithms for answering counting queries.

The results in Sections 3 and 4 imply that $|\mathcal{P}[\psi] - \mathcal{E}_{\boldsymbol{S}}[\psi]| \leq \tau$ holds with high probability whenever ψ is generated by a differentially private algorithm $\mathcal{M}$. This might appear to be inconsistent with our application since there the queries are generated by an arbitrary (possibly adversarial) adaptive analyst and we can only guarantee that the query answering algorithm is differentially private. The connection comes from the following basic fact about differentially private algorithms:

Fact 13 (Postprocessing Preserves Privacy (see e.g. [DR14]))**.** *Let $\mathcal{M} : \mathcal{X}^n \to \mathcal{O}$ be an (ϵ, δ) differentially private algorithm with range $\mathcal{O}$, and let $\mathcal{F} : \mathcal{O} \to \mathcal{O}'$ be an arbitrary randomized algorithm. Then $\mathcal{F} \circ \mathcal{M} : \mathcal{X}^n \to \mathcal{O}'$ is (ϵ, δ)-differentially private.*

Hence, an *arbitrary* adaptive analyst $\mathcal{A}$ is guaranteed to generate queries in a manner that is differentially private in $\boldsymbol{S}$ so long as the only access that she has to $\boldsymbol{S}$ is through a differentially private query answering algorithm $\mathcal{M}$. We also note that the bounds we state here give the probability of correctness for each individual answer to a query, meaning that the error probability β is for each query ϕ_i and not for all queries at the same time. The bounds we state in Section 1.2 hold with high probability for all m queries and to obtain them from the bounds in this section, we apply the union bound by setting $\beta = \beta'/m$ for some small β'.

We now highlight a few applications of differentially private algorithms for answering counting queries to our problem.

5.1 Laplacian Noise Addition

The Laplacian Mechanism on input of a dataset S answers m adaptively chosen queries $\phi_1, \ldots, \phi_m$ by responding with $\phi_i(S) + \mathrm{Lap}(0, \sigma)$ when given query ϕ_i. Here, $\mathrm{Lap}(0, \sigma)$ denotes a Laplacian random variable of mean 0 and scale σ. For suitably chosen σ the algorithm has the following guarantee.

Theorem 14 (Laplace)**.** *Let $\tau, \beta, \epsilon > 0$ and define*

$$n_L(\tau, \beta, \epsilon, m) = \frac{m \log(1/\beta)}{\epsilon\tau}\,.$$

$$n_L^\delta(\tau, \beta, \epsilon, \delta, m) = \frac{\sqrt{m \log(1/\delta)} \log(1/\beta)}{\epsilon\tau}\,.$$

There is computationally efficient algorithm called Laplace *which on input of a data set S of size n accepts any sequence of m adaptively chosen functions $\phi_1, \ldots, \phi_m \in \mathcal{X}^{[0,1]}$ and returns estimates $a_1, \ldots, a_m$ such that for every $i \in [m]$ we have $\mathbb{P}\left[|\mathcal{E}_S[\phi_i] - a_i| > \tau\right] \leq \beta$. To achieve this guarantee under $(\epsilon, 0)$-differential privacy, it requires $n \geq C n_L(\tau, \beta, \epsilon, m)$, and to achieve this guarantee under (ϵ, δ)-differential privacy, it requires $n \geq C n_L^\delta(\tau, \beta, \epsilon, \delta, m)$ for sufficiently large constant C.*

Applying our main generalization bound for $(\epsilon, 0)$-differential privacy directly gives the following corollary.

Corollary 15. *Let $\tau, \beta > 0$ and define*

$$n_L(\tau, \beta, m) = \frac{m \log(1/\beta)}{\tau^2}\,.$$

There is a computationally efficient algorithm which on input of a data set S of size n sampled from $\mathcal{P}^n$ accepts any sequence of m adaptively chosen functions $\phi_1, \ldots, \phi_m \in \mathcal{X}^{[0,1]}$ and returns estimates $a_1, \ldots, a_m$ such that for every $i \in [m]$ we have $\mathbb{P}\left[|\mathcal{P}[\phi_i] - a_i| > \tau\right] \leq \beta$ provided that $n \geq C n_L(\tau, \beta, m)$ for sufficiently large constant C.

Proof. We apply Theorem 9 with $\epsilon = \tau/2$ and plug this choice of ϵ into the definition of n_L in Theorem 14. We note that the stated lower bound on n implies the lower bound required by Theorem 9. □

The corollary that follows the (ϵ, δ) bound gives a quadratic improvement in m compared with Corollary 15 at the expense of a slightly worse dependence on τ and $1/\beta$.

Corollary 16. *Let $\tau, \beta > 0$ and define*

$$n_L^\delta(\tau, \beta, m) = \frac{\sqrt{m} \log^{1.5}(1/\beta)}{\tau^{2.5}}\,.$$

There is a computationally efficient algorithm which on input of a data set S of size n sampled from $\mathcal{P}^n$ accepts any sequence of m adaptively chosen functions $\phi_1, \ldots, \phi_m \in \mathcal{X}^{[0,1]}$ and returns estimates $a_1, \ldots, a_m$ such that for every $i \in [m]$ we have $\mathbb{P}\left[|\mathcal{P}[\phi_i] - a_i| > \tau\right] \leq \beta$ provided that $n \geq C n_L^\delta(\tau, \beta, m)$ for sufficiently large constant C.

Proof. We apply Theorem 10 with $\epsilon = \tau/2$ and $\delta = \exp(-4\ln(8/\beta)/\tau)$. Plugging these parameters into the definition of n_L^δ in Theorem 14 gives the stated lower bound on n. We note that the stated lower bound on n implies the lower bound required by Theorem 10. □

5.2 Multiplicative Weights Technique

The private multiplicative weights algorithm [HR10] achieves an exponential improvement in m compared with the Laplacian mechanism. The main drawback is a running time that scales linearly with the domain size in the worst case and is therefore not computationally efficient in general.

Theorem 17 (Private Multiplicative Weights)**.** *Let $\tau, \beta, \epsilon > 0$ and define*

$$n_{MW}(\tau, \beta, \epsilon) = \frac{\log(|\mathcal{X}|)\log(1/\beta)}{\epsilon\tau^3}.$$

$$n^{\delta}_{MW}(\tau, \beta, \epsilon, \delta) = \frac{\sqrt{\log(|\mathcal{X}|)\log(1/\delta)}\log(1/\beta)}{\epsilon\tau^2}.$$

There is algorithm called PMW *which on input of a data set S of size n accepts any sequence of m adaptively chosen functions $\phi_1, \ldots, \phi_m \in \mathcal{X}^{[0,1]}$ and with probability at least $1 - (n \log |\mathcal{X}|)\beta$ returns estimates $a_1, \ldots, a_m$ such that for every $i \in [m]$ we have $\mathbb{P}\left[|\mathcal{E}_S[\phi_i] - a_i| > \tau\right] \leq \beta$. To achieve this guarantee under $(\epsilon, 0)$ differential privacy, it requires that $n \geq Cn_{MW}(\tau, \beta, \epsilon)$ and to achieve it under (ϵ, δ)-differential privacy it requires $n \geq Cn^{\delta}_{MW}(\tau, \beta, \epsilon, \delta)$ for sufficiently large constant C.*

Corollary 18. *Let $\tau, \beta > 0$ and define*

$$n_{MW}(\tau, \beta) = \frac{\log(|\mathcal{X}|)\log(1/\beta)}{\tau^4}.$$

There is an algorithm which on input of a data set S of size n sampled from $\mathcal{P}^n$ accepts any sequence of m adaptively chosen functions $\phi_1, \ldots, \phi_m \in \mathcal{X}^{[0,1]}$ and with probability at least $1 - (n \log |\mathcal{X}|)\beta$ returns estimates $a_1, \ldots, a_m$ such that for every $i \in [m]$ we have $\mathbb{P}\left[|\mathcal{P}[\phi_i] - a_i| > \tau\right] \leq \beta$ provided that $n \geq Cn_{MW}(\tau, \beta)$ for sufficiently large constant C.

Proof. We apply Theorem 9 with $\epsilon = \tau/2$ and plug this choice of ϵ into the definition of n_{MW} in Theorem 17. We note that the stated lower bound on n implies the lower bound required by Theorem 9. □

Under (ϵ, δ) differential privacy we get the following corollary that improves the dependence on τ and $\log |\mathcal{X}|$ in Corollary 18 at the expense of a slightly worse dependence on β.

Corollary 19. *Let $\tau, \beta > 0$ and define*

$$n^{\delta}_{MW}(\tau, \beta) = \frac{\sqrt{\log(|\mathcal{X}|)}\log(1/\beta)^{3/2}}{\tau^{3.5}}.$$

There is an algorithm which on input of a data set S of size n sampled from $\mathcal{P}^n$ accepts any sequence of m adaptively chosen functions $\phi_1, \ldots, \phi_m \in \mathcal{X}^{[0,1]}$ and with probability at least $1 - (n \log |\mathcal{X}|)\beta$ returns estimates $a_1, \ldots, a_m$ such that for every $i \in [m]$ we have $\mathbb{P}\left[|\mathcal{P}[\phi_i] - a_i| > \tau\right] \leq \beta$ provided that $n \geq Cn^{\delta}_{MW}(\tau, \beta)$ for sufficiently large constant C.

Proof. We apply Theorem 10 with $\epsilon = \tau/2$ and $\delta = \exp(-4\ln(8/\beta)/\tau)$. Plugging these parameters into the definition of n^{δ}_{MW} in Theorem 17 gives the stated lower bound on n. We note that the stated lower bound on n implies the lower bound required by Theorem 10. □

5.3 Sparse Vector Technique

In this section we give a computationally efficient technique for answering exponentially many queries $\phi_1, \ldots, \phi_m$ in the size of the data set n so long as they are chosen using only $o(n)$ rounds of adaptivity. We say that a sequence of queries $\phi_1, \ldots, \phi_m \in \mathcal{X}^{[0,1]}$, answered with numeric values $a_1, \ldots, a_m$ is generated with r rounds of adaptivity if there are r indices $i_1, \ldots, i_r$ such that the procedure that generates the queries as a function of the answers can be described by $r+1$ (possibly randomized) algorithms $f_0, f_1, \ldots, f_r$ satisfying:

$$(\phi_1, \ldots, \phi_{i_1-1}) = f_0(\emptyset)$$

$$(\phi_{i_1}, \ldots, \phi_{i_2-1}) = f_1((\phi_1, a_1), \ldots, (\phi_{i_1-1}, a_{i_1-1}))$$

$$(\phi_{i_2}, \ldots, \phi_{i_3-1}) = f_2((\phi_1, a_1), \ldots, (\phi_{i_2-1}, a_{i_2-1}))$$

$$\vdots$$

$$(\phi_{i_r}, \ldots, \phi_m) = f_r((\phi_1, a_1), \ldots, (\phi_{i_r-1}, a_{i_r-1}))$$

We build our algorithm out of a differentially private algorithm called SPARSE that takes as input an adaptively chosen sequence of queries together with *guesses of the answers to those queries.* Rather than always returning numeric valued answers, it compares the error of our guess to a *threshold* T and returns a numeric valued answer to the query only if (a noisy version of) the error of our guess was above the given threshold. SPARSE is computationally efficient, and has the remarkable property that its accuracy has polynomial dependence only on the number of queries for which the error of our guesses are close to being above the threshold.

Theorem 20 (Sparse Vector $(\epsilon, 0)$)**.** *Let $\tau, \beta, \epsilon > 0$ and define*

$$n_{SV}(\tau, \beta, \epsilon) = \frac{9r\left(\ln(4/\beta)\right)}{\tau\epsilon}.$$

$$n^{\delta}_{SV}(\tau, \beta, \epsilon, \delta) = \frac{\left(\sqrt{512}+1\right)\sqrt{r\ln(2/\delta)}\left(\ln(4/\beta)\right)}{\tau\epsilon}.$$

There is an algorithm called SPARSE *parameterized by a real valued threshold T, which on input of a data set S of size n accepts any sequence of m adaptively chosen queries together with guesses at their values $g_i \in \mathbb{R}$: $(\phi_1, g_1), \ldots, (\phi_m, g_m)$ and returns answers $a_1, \ldots, a_m \in \{\perp\} \cup \mathbb{R}$. It has the property that for all $i \in [m]$, with probability $1-\beta$: if $a_i = \perp$ then $|\mathcal{E}_S[\phi_i] - g_i| \leq T + \tau$ and if $a_i \in \mathbb{R}, |\mathcal{E}_S[\phi_i] - a_i| \leq \tau$. To achieve this guarantee under $(\epsilon, 0)$-differential privacy it requires $n \geq n_{SV}(\tau, \beta, \epsilon)$ and to achieve this guarantee under (ϵ, δ)-differential privacy, it requires $n \geq n^{\delta}_{SV}(\tau, \beta, \epsilon, \delta)$. In either case, the algorithm also requires that $|\{i : |\mathcal{E}_S[\phi_i] - g_i| \geq T - \tau\}| \leq r$. (If this last condition does not hold, the algorithm may halt early and stop accepting queries)*

We observe that the *naïve* method of answering queries using their empirical average allows us to answer each query up to accuracy τ with probability $1-\beta$ given a data set of size $n_0 \geq \ln(2/\beta)/\tau^2$ so long as the queries are non-adaptively chosen. Thus, with high probability, problems only arise between rounds of adaptivity. If we knew when these rounds of adaptivity occurred, we could refresh our sample between each round, and obtain total sample complexity linear in the number of rounds of adaptivity. The method we present (using $(\epsilon, 0)$-differential privacy) lets us get a comparable

bound *without* knowing where the rounds of adaptivity appear. Using (ϵ, δ) privacy would allow us to obtain constant factor improvements if the number of queries was large enough, but does not get an asymptotically better dependence on the number of rounds r (it would allow us to reuse the round testing set *quadratically* many times, but we would still potentially need to refresh the training set after each round of adaptivity, in the worst case).

The idea is the following: we obtain r different estimation samples $S_1, \ldots, S_r$ each of size sufficient to answer non-adaptively chosen queries to error $\tau/8$ with probability $1 - \beta/3$, and a separate round detection sample S_h of size $n_{SV}(\tau/8, \beta/3, \epsilon)$ for $\epsilon = \tau/16$, which we access only through a copy of SPARSE we initialize with threshold $T = \tau/4$. As queries ϕ_i start arriving, we compute their answers $a_i^t = \mathcal{E}_{S_1}[\phi_i]$ using the naïve method on estimation sample S_1 which we use as our *guess* of the correct value on S_h when we feed ϕ_i to SPARSE. If the answer SPARSE returns is $a_i^h = \perp$, then we know that with probability $1 - \beta/3$, a_i^t is accurate up to tolerance $T + \tau/8 = 3\tau/8$ with respect to S_h, and hence statistically valid up to tolerance $\tau/2$ by Theorem 9 with probability at least $1 - 2\beta/3$. Otherwise, we discard our estimation set S_1 and continue with estimation set S_2. We know that with probability $1 - \beta/3$, a_i^h is accurate with respect to S_h up to tolerance $\tau/8$, and hence statistically valid up to tolerance $\tau/4$ by Theorem 9 with probability at least $1 - 2\beta/3$. We continue in this way, discarding and incrementing our estimation set whenever our guess g_i is incorrect. This succeeds in answering every query so long as our guesses are not incorrect more than r times in total. Finally, we know that except with probability at most $m\beta/3$, by the accuracy guarantee of our estimation set for *non-adaptively* chosen queries, the only queries i for which our guesses g_i will deviate from $\mathcal{E}_{S_h}[\phi_i]$ by more than $T - \tau/8 = \tau/8$ are those queries that lie *between rounds of adaptivity*. There are at most r of these by assumption, so the algorithm runs to completion with probability at least $1 - m\beta/3$. The algorithm is given in figure 1.

Algorithm EffectiveRounds

Input: A database $\boldsymbol{S}$ of size $|\boldsymbol{S}| \geq \frac{1156r\ln(\frac{12}{\beta})}{\tau^2}$.

Initialization: Randomly split $\boldsymbol{S}$ into $r+1$ sets: r sets $S_1, \ldots, S_r$ with size $|S_i| \geq \frac{4\ln(\frac{12}{\beta})}{\tau^2}$, and one set S_h with size $|S_h| = \frac{1152 \cdot r \cdot \ln(\frac{12}{\beta})}{\tau^2}$. Instantiate SPARSE with input S_h and parameters $T = \tau/4$, $\tau' = \tau/8$, $\beta' = \beta/3$, and $\epsilon = \tau/16$. Let $c \leftarrow 1$.

Query stage For each query ϕ_i do:

1. Compute $a_i^t = \mathcal{E}_{S_c}[\phi_i]$. Let $g_i = a_i^t$ and feed (ϕ_i, g_i) to sparse and receive answer a_i^h.
2. If $a_i^h = \perp$ then return answer $a_i = a_i^t$.
3. Else return answer $a_i = a_i^h$. Set $c \leftarrow c + 1$.
4. If $c > r$ HALT.

Figure 1: The EffectiveRounds algorithm

This algorithm yields the following theorem:

Theorem 21. *Let $\tau, \beta > 0$ and define*

$$n_{SV}(\tau, \beta) = \frac{r\ln(\frac{1}{\beta})}{\tau^2}.$$

There is an algorithm which on input of a data set S of size n sampled from $\mathcal{P}^n$ accepts any sequence of m adaptively chosen queries $\phi_1, \ldots, \phi_m$ generated with at most r rounds of adaptivity. With probability at least $1 - m\beta$ the algorithm runs to completion and returns estimates $a_1, \ldots, a_m$ for each query. These estimates have the property that for all $i \in [m]$ we have $\mathbb{P}\left[|\mathcal{P}[\phi_i] - a_i| > \tau\right] \leq \beta$ provided that $n \geq C n_{SV}(\tau, \beta)$ for sufficiently large constant C.

Remark 22. *Note that the accuracy guarantee of* SPARSE *depends only on the number of incorrect guesses that are actually made. Hence, EffectiveRounds does not halt until the actual number of instances of over-fitting to the estimation samples S_i is larger than r. This could be equal to the number of rounds of adaptivity in the worst case (for example, if the analyst is running the Dinur-Nissim reconstruction attack within each round [DN03]), but in practice might achieve a much better bound (if the analyst is not fully adversarial).*

Acknowledgements We would like to thank Sanjeev Arora, Nina Balcan, Avrim Blum, Dean Foster, Michael Kearns, Jon Kleinberg, Sasha Rakhlin, and Jon Ullman for enlightening discussions and helpful comments. We also thank the Simons Institute for Theoretical Computer Science at Berkeley where part of this research was done.

References

[ANR11] Ehud Aharoni, Hani Neuvirth, and Saharon Rosset. The quality preserving database: A computational framework for encouraging collaboration, enhancing power and controlling false discovery. *IEEE/ACM Trans. Comput. Biology Bioinform.*, 8(5):1431–1437, 2011.

[AR14] Ehud Aharoni and Saharon Rosset. Generalized a-investing: definitions, optimality results and application to public databases. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(4):771–794, 2014.

[BDMN05] Avrim Blum, Cynthia Dwork, Frank McSherry, and Kobbi Nissim. Practical privacy: the SuLQ framework. In *PODS*, pages 128–138, 2005.

[BE02] Olivier Bousquet and André Elisseeff. Stability and generalization. *JMLR*, 2:499–526, 2002.

[BE12] C. Glenn Begley and Lee Ellis. Drug development: Raise standards for preclinical cancer research. *Nature*, 483:531–533, 2012.

[BH95] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate – a practical and powerful approach to multiple testing. *Journal of the Royal Statistics Society: Series B (Statistical Methodology)*, 57:289–300, 1995.

[BH15] Avrim Blum and Moritz Hardt. The ladder: A reliable leaderboard for machine learning competitions. *CoRR*, abs/1502.04585, 2015.

[BNS+15] Raef Bassily, Kobbi Nissim, Adam D. Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. *CoRR*, abs/1511.02513, 2015.

[BST14] Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization, revisited. *CoRR*, abs/1405.7085, 2014.

[CKL+06] C. Chu, S. Kim, Y. Lin, Y. Yu, G. Bradski, A. Ng, and K. Olukotun. Map-reduce for machine learning on multicore. In *Proceedings of NIPS*, pages 281–288, 2006.

[DFH+15a] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. *CoRR*, abs/1506, 2015.

[DFH+15b] 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, 2015.

[DKM+06] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *EUROCRYPT*, pages 486–503, 2006.

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*, pages 265–284. Springer, 2006.

[DN03] Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In *PODS*, pages 202–210. ACM, 2003.

[DN04] Cynthia Dwork and Kobbi Nissim. Privacy-preserving datamining on vertically partitioned databases. In *CRYPTO*, pages 528–544, 2004.

[DR14] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(34):211–407, 2014.

[Dwo11] Cynthia Dwork. A firm foundation for private data analysis. *CACM*, 54(1):86–95, 2011.

[FGR+13] Vitaly Feldman, Elena Grigorescu, Lev Reyzin, Santosh Vempala, and Ying Xiao. Statistical algorithms and a lower bound for planted clique. In *STOC*, pages 655–664. ACM, 2013.

[Fre83] David A. Freedman. A note on screening regression equations. *The American Statistician*, 37(2):152–155, 1983.

[FS08] D. Foster and R. Stine. Alpha-investing: A procedure for sequential control of expected false discoveries. *J. Royal Statistical Soc.: Series B (Statistical Methodology)*, 70(2):429–444, 2008.

[GL14] Andrew Gelman and Eric Loken. The statistical crisis in science. *The American Statistician*, 102(6):460, 2014.

[HR10] Moritz Hardt and Guy N. Rothblum. A multiplicative weights mechanism for privacy-preserving data analysis. In *51st IEEE FOCS 2010*, pages 61–70, 2010.

[HTF09] Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer series in statistics. Springer, 2009.

[HU14] Moritz Hardt and Jonathan Ullman. Preventing false discovery in interactive data analysis is hard. In *FOCS*, pages 454–463, 2014.

[Ioa05a] John A. Ioannidis. Contradicted and initially stronger effects in highly cited clinical research. *The Journal of American Medical Association*, 294(2):218–228, 2005.

[Ioa05b] John P. A. Ioannidis. Why Most Published Research Findings Are False. *PLoS Medicine*, 2(8):124, August 2005.

[Kaga] Five lessons from Kaggle's event recommendation engine challenge. http://www.rouli.net/2013/02/five-lessons-from-kaggles-event.html. Accessed: 2014-10-07.

[Kagb] Kaggle blog: No free hunch. http://blog.kaggle.com/. Accessed: 2014-10-07.

[Kagc] Kaggle user forums. https://www.kaggle.com/forums. Accessed: 2014-10-07.

[Kea98] Michael Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM (JACM)*, 45(6):983–1006, 1998.

[KV94] Michael J Kearns and Umesh Virkumar Vazirani. *An introduction to computational learning theory*. MIT press, 1994.

[MNPR06] Sayan Mukherjee, Partha Niyogi, Tomaso Poggio, and Ryan Rifkin. Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics*, 25(1-3):161–193, 2006.

[PRMN04] Tomaso Poggio, Ryan Rifkin, Sayan Mukherjee, and Partha Niyogi. General conditions for predictivity in learning theory. *Nature*, 428(6981):419–422, 2004.

[PSA11] Florian Prinz, Thomas Schlange, and Khusru Asadullah. Believe it or not: how much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery*, 10(9):712–712, 2011.

[Reu03] Juha Reunanen. Overfitting in making comparisons between variable selection methods. *Journal of Machine Learning Research*, 3:1371–1382, 2003.

[RF08] R. Bharat Rao and Glenn Fung. On the dangers of cross-validation. an experimental evaluation. In *International Conference on Data Mining*, pages 588–596. SIAM, 2008.

[RR10] Aaron Roth and Tim Roughgarden. Interactive privacy via the median mechanism. In 42*nd ACM STOC*, pages 765–774. ACM, 2010.

[RZ15] Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. *CoRR*, abs/1511.05219, 2015.

[SNS11] Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11):1359–1366, 2011.

[SSBD14] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[SSSSS10] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 11:2635–2670, 2010.

[SU14] Thomas Steinke and Jonathan Ullman. Interactive fingerprinting codes and the hardness of preventing false discovery. *arXiv preprint arXiv:1410.1228*, 2014.

[TT15] Jonathan Taylor and Robert J. Tibshirani. Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, 112(25):7629–7634, 2015.

[Win] David Wind. Learning from the best. http://blog.kaggle.com/2014/08/01/learning-from-the-best/. Accessed: 2014-10-07.

[WLF15] Yu-Xiang Wang, Jing Lei, and Stephen E. Fienberg. Learning with differential privacy: Stability, learnability and the sufficiency and necessity of ERM principle. *CoRR*, abs/1502.06309, 2015.

A Adaptivity in fitting a linear model

In this section, we give a very simple example to illustrate how a data analyst could end up overfitting to a dataset by asking only a small number of (adaptively) chosen queries to the dataset, if they are answered using the naïve method.

The data analyst has n samples $D = \{x_1, \ldots, x_n\}$ over d real-valued attributes sampled from an unknown distribution $\mathcal{D}$. The analyst's goal is to find a linear model ℓ that maximizes the average correlation with the unknown distribution. Formally, the goal is to find a unit vector that maximizes the function

$$f(u) = \mathop{\mathbb{E}}_{x \sim \mathcal{D}}[\langle u, x \rangle] .$$

Not knowing the distribution the analyst decides to solve the corresponding optimization problem on her finite sample:

$$\tilde{f}_D(u) = \frac{1}{n} \sum_{x \in D} \langle u, x \rangle .$$

The analyst attempts to solve the problem using the following simple but *adaptive strategy:*

1. For $i = 1, \ldots, d$, determine $s_i = \text{sign}\Big(\sum_{x \in D} x_i \Big)$.
2. Let $\tilde{u} = \frac{1}{\sqrt{d}}(s_1, \ldots, s_d)$.

Intuitively, this natural approach first determines for each attribute whether it is positively or negatively correlated. It then aggregates this information across all d attributes into a single linear model.

The next lemma shows that this adaptive strategy has a terrible generalization performance (if d is large). Specifically, we show that even if there is no linear structure whatsoever in the underlying distribution (namely it is normally distributed), the analyst's strategy falsely discovers a linear model with large objective value.

Lemma 23. *Suppose $\mathcal{D} = N(0,1)^d$. Then, every unit vector $u \in \mathbb{R}^d$ satisfies $f(u) = 0$. However, $\mathbb{E}_D[\tilde{f}_D(\tilde{u})] = \sqrt{2/\pi} \cdot \sqrt{d/n}$.*

Proof. The first claim follows because $\langle u, x\rangle$ for $x \sim N(0,1)^d$ is distributed like a Gaussian random variable $N(0,1)$. Let us now analyze the objective value of $\tilde{u}$.

$$\tilde{f}_D(\tilde{u}) = \frac{1}{n}\sum_{x\in D} \frac{s_i}{\sqrt{d}} \sum_{i=1}^{d} x_i = \frac{1}{\sqrt{d}} \sum_{i=1}^{d} \left| \frac{1}{n} \sum_{x\in D} x_i \right|$$

Hence,

$$\mathop{\mathbb{E}}_{D}[\tilde{f}_D(\tilde{u})] = \sum_{i=1}^{d} \frac{1}{\sqrt{d}} \mathop{\mathbb{E}}_{D}\left[\left| \frac{1}{n} \sum_{x\in D} x_i \right| \right].$$

Now, $(1/n)\sum_{x\in D} x_i$ is distributed like a gaussian random variable $g \sim N(0, 1/n)$, since each x_i is a standard gaussian. It follows that

$$\mathop{\mathbb{E}}_{D} \tilde{f}_D(\tilde{u}) = \sqrt{\frac{2d}{\pi n}}.$$

□

Note that all the operations performed by the analyst are based on empirical averages of real-valued functions. To determine the bias, the function is just $\phi_i(x) = x_i$ and to determine the final correlation it is $\psi(x) = \langle u, x\rangle$. These functions are not bounded to the range $[0,1]$ as required by the formal definition of our model. However, it is easy to see that this is a minor issue. Note that both x_i and $\langle u, x\rangle$ are distributed according to $N(0,1)$ whenever $x \sim N(0,1)^d$. This implies that for every query function ϕ we used, $\mathbb{P}[|\phi(x)| \geq B] \leq 1/\mathsf{poly}(n,d)$ for some $B = O(\log(dn))$. We can therefore truncate and rescale each query as $\phi'(x) = P_B(\phi(x))/(2B) + 1/2$, where P_B is the truncation of the values outside $[-B, B]$. This ensures that the range of $\phi'(x)$ is $[0,1]$. It is easy to verify that using these $[0,1]$-valued queries does not affect the analysis in any significant way (aside from scaling by a logarithmic factor) and we obtain overfitting in the same way as before (for large enough d).

B Background on Differential Privacy

When applying (ϵ, δ)-differential privacy, we are typically interested in values of δ that are very small compared to n. In particular, values of δ on the order of $1/n$ yield no meaningful definition of privacy as they permit the publication of the complete records of a small number of data set participants—a violation of any reasonable notion of privacy.

Theorem 24. *Any (ε, δ)-differentially private mechanism $\mathcal{A}$ satisfies for all pairs of data sets S, S' differing in at most k elements, and all $\mathcal{O} \subseteq \mathrm{Range}(\mathcal{A})$:*

$$\mathbb{P}[\mathcal{A}(S) \in \mathcal{O}] \leq \exp(k\varepsilon)\, \mathbb{P}[\mathcal{A}(S') \in \mathcal{O}] + e^{\epsilon(k-1)}\delta,$$

where the probability space is over the coin flips of the mechanism $\mathcal{A}$.

Differential privacy also degrades gracefully under composition. It is easy to see that the independent use of an $(\varepsilon_1, 0)$-differentially private algorithm and an $(\varepsilon_2, 0)$-differentially private algorithm, when taken together, is $(\varepsilon_1 + \varepsilon_2, 0)$-differentially private. More generally, we have

Theorem 25. *Let $\mathcal{A}_i : \mathcal{X}^n \to \mathcal{R}_i$ be an $(\varepsilon_i, \delta_i)$-differentially private algorithm for $i \in [k]$. Then if $\mathcal{A}_{[k]} : \mathcal{X}^n \to \prod_{i=1}^k \mathcal{R}_i$ is defined to be $\mathcal{A}_{[k]}(S) = (\mathcal{A}_1(S), \ldots, \mathcal{A}_k(S))$, then $\mathcal{A}_{[k]}$ is $(\sum_{i=1}^k \varepsilon_i, \sum_{i=1}^k \delta_i)$-differentially private.*

A more sophisticated argument yields significant improvement when $\varepsilon < 1$:

Theorem 26. *For all $\varepsilon, \delta, \delta' \geq 0$, the composition of k arbitrary (ε, δ)-differentially private mechanisms is $(\varepsilon', k\delta + \delta')$-differentially private, where*

$$\varepsilon' = \sqrt{2k \ln(1/\delta')}\varepsilon + k\varepsilon(e^\varepsilon - 1),$$

even when the mechanisms are chosen adaptively.

Theorems 25 and 26 are very general. For example, they apply to queries posed to overlapping, but not identical, data sets. Nonetheless, data utility will eventually be consumed: the Fundamental Law of Information Recovery states that overly accurate answers to too many questions will destroy privacy in a spectacular way (see [DN03] *et sequelae*). The goal of algorithmic research on differential privacy is to stretch a given privacy "budget" of, say, ε_0, to provide as much utility as possible, for example, to provide useful answers to a great many counting queries. The bounds afforded by the composition theorems are the first, not the last, word on utility.

C Concentration and moment bounds

C.1 Concentration inequalities

We will use the following statement of the multiplicative Chernoff bound:

Lemma 27 (Chernoff's bound)**.** *Let $Y_1, Y_2, \ldots, Y_n$ be i.i.d. Bernoulli random variables with expectation $p > 0$. Then for every $\gamma > 0$,*

$$\mathbb{P}\left[\sum_{i\in[n]} Y_i \geq (1+\gamma)np\right] \leq \exp\left(-np((1+\gamma)\ln(1+\gamma) - \gamma)\right).$$

Lemma 28 (McDiarmid's inequality)**.** *Let $X_1, X_2, \ldots, X_n$ be independent random variables taking values in the set $\mathcal{X}$. Further let $f : \mathcal{X}^n \to \mathbb{R}$ be a function that satisfies, for all $i \in [n]$ and $x_1, x_2, \ldots, x_n, x_i' \in \mathcal{X}$,*

$$f(x_1, \ldots, x_i, \ldots, x_n) - f(x_1, \ldots, x_i', \ldots, x_n) \leq c.$$

Then for all $\alpha > 0$, and $\mu = \mathbb{E}[f(X_1, \ldots, X_n)]$,

$$\mathbb{P}[f(X_1, \ldots, X_n) - \mu \geq \alpha] \leq \exp\left(\frac{-2\alpha^2}{n \cdot c^2}\right).$$

C.2 Moment Bounds

Lemma 29. *Let $Y_1, Y_2, \ldots, Y_n$ be i.i.d. Bernoulli random variables with expectation p. We denote by $\mathcal{M}_k[B(n,p)] \doteq \mathbb{E}\left[\left(\frac{1}{n}\sum_{i\in[n]} Y_i\right)^k\right]$. Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with values in $[0,1]$ and expectation p. Then for every $k > 0$,*

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{i\in[n]} X_i\right)^k\right] \leq \mathcal{M}_k[B(n,p)].$$

Proof. We use I to denote a k-tuple of indices $(i_1, \ldots, i_k) \in [n]^k$ (not necessarily distinct). For I like that we denote by $\{\ell_1, \ldots, \ell_{k'}\}$ the set of distinct indices in I and let $k_1, \ldots, k_{k'}$ denote their multiplicities. Note that $\sum_{j\in[k']} k_j = k$. We first observe that

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{i\in[n]} X_i\right)^k\right] = \mathop{\mathbb{E}}_{I\sim[n]^k}\left[\mathbb{E}\left[\prod_{j\in[k]} X_{i_j}\right]\right] = \mathop{\mathbb{E}}_{I\sim[n]^k}\left[\mathbb{E}\left[\prod_{j\in[k']} X_{\ell_j}^{k_j}\right]\right] = \mathop{\mathbb{E}}_{I\sim[n]^k}\left[\prod_{j\in[k']} \mathbb{E}\left[X_{\ell_j}^{k_j}\right]\right], \tag{11}$$

where the last equality follows from independence of X_i's. For every j, the range of X_{ℓ_j} is $[0,1]$ and thus

$$\mathbb{E}\left[X_{\ell_j}^{k_j}\right] \leq \mathbb{E}\left[X_{\ell_j}\right] = p.$$

Moreover the value p is achieved when X_{ℓ_j} is Bernoulli with expectation p. That is

$$\mathbb{E}\left[X_{\ell_j}^{k_j}\right] \leq \mathbb{E}\left[Y_{\ell_j}^{k_j}\right],$$

and by using this in equality (11) we obtain that

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{i\in[n]} X_i\right)^k\right] \leq \mathbb{E}\left[\left(\frac{1}{n}\sum_{i\in[n]} Y_i\right)^k\right] = \mathcal{M}_k[B(n,p)].$$

□

Lemma 30. *For all integers $n \geq k \geq 1$ and $p \in [0,1]$,*

$$\mathcal{M}_k[B(n,p)] \leq p^k + (k\ln n + 1)\cdot\left(\frac{k}{n}\right)^k.$$

Proof. Let U denote $\frac{1}{n}\sum_{i\in[n]} X_i$, where X_i's are i.i.d. Bernoulli random variables with expectation $p > 0$ (the claim is obviously true if $p = 0$). Then

$$\mathbb{E}[U^k] \leq p^k + \int_{p^k}^1 \mathbb{P}[U^k \geq t]dt. \tag{12}$$

We substitute $t = (1+\gamma)^k p^k$ and observe that Lemma 27 gives:

$$\mathbb{P}[U^k \geq t] = \mathbb{P}[U^k \geq ((1+\gamma)p)^k] = \mathbb{P}[U \geq (1+\gamma)p] \leq \exp\left(-np((1+\gamma)\ln(1+\gamma) - \gamma)\right).$$

Using this substitution in eq.(12) together with $\frac{dt}{d\gamma} = k(1+\gamma)^{k-1} \cdot p^k$ we obtain

$$\begin{aligned}
\mathbb{E}[U^k] &\leq p^k + \int_0^{1/p-1} \exp\left(-np((1+\gamma)\ln(1+\gamma) - \gamma)\right) \cdot k(1+\gamma)^{k-1} d\gamma \\
&= p^k + p^k k \int_0^{1/p-1} \frac{1}{1+\gamma} \cdot \exp\left(k\ln(1+\gamma) - np((1+\gamma)\ln(1+\gamma) - \gamma)\right) d\gamma \\
&\leq p^k + p^k k \max_{\gamma \in [0,1/p-1]} \left\{\exp\left(k\ln(1+\gamma) - np((1+\gamma)\ln(1+\gamma) - \gamma)\right)\right\} \cdot \int_0^{1/p-1} \frac{1}{1+\gamma} d\gamma \\
&= p^k + p^k k \ln(1/p) \cdot \max_{\gamma \in [0,1/p-1]} \left\{\exp\left(k\ln(1+\gamma) - np((1+\gamma)\ln(1+\gamma) - \gamma)\right)\right\}. \qquad (13)
\end{aligned}$$

We now find the maximum of $g(\gamma) \doteq k\ln(1+\gamma) - np((1+\gamma)\ln(1+\gamma) - \gamma)$. Differentiating the expression we get $\frac{k}{1+\gamma} - np\ln(1+\gamma)$ and therefore the function attains its maximum at the (single) point γ_0 which satisfies: $(1+\gamma_0)\ln(1+\gamma_0) = \frac{k}{np}$. This implies that $\ln(1+\gamma_0) \leq \ln\left(\frac{k}{np}\right)$. Now we observe that $(1+\gamma)\ln(1+\gamma) - \gamma$ is always non-negative and therefore $g(\gamma_0) \leq k\ln\left(\frac{k}{np}\right)$. Substituting this into eq.(13) we conclude that

$$\mathbb{E}[U^k] \leq p^k + p^k k\ln(1/p) \cdot \exp\left(k\ln\left(\frac{k}{np}\right)\right) = p^k + k\ln(1/p) \cdot \left(\frac{k}{n}\right)^k.$$

Finally, we observe that if $p \geq 1/n$ then clearly $\ln(1/p) \leq \ln n$ and the claim holds. For any $p < 1/n$ we use monotonicity of $\mathcal{M}_k[B(n,p)]$ in p and upper bound the probability by the bound for $p = 1/n$ that equals

$$\left(\frac{1}{n}\right)^k + (k\ln n) \cdot \left(\frac{k}{n}\right)^k \leq (k\ln n + 1) \cdot \left(\frac{k}{n}\right)^k.$$

□

Lemma 31. *Let $n > k > 0, \varepsilon > 0, p > 0, \delta \geq 0$ and let V be a non-negative random variable that satisfies $\mathbb{E}[V^k] \leq e^{\varepsilon k}\mathcal{M}_k[B(n,p)] + \delta$. Then for any $\tau \in [0, 1/3]$, $\beta \in (0, 2/3]$ if*

- $\varepsilon \leq \tau/2$,
- $k \geq \max\{4p\ln(2/\beta)/\tau,\ 2\log\log n\}$,
- $n \geq 3k/\tau$ *then*

$$\mathbb{P}[V \geq p + \tau] \leq \beta + \delta/(p+\tau)^k.$$

Proof. Observe that by Markov's inequality:

$$\mathbb{P}[V \geq p+\tau] = \mathbb{P}[V^k \geq (p+\tau)^k] \leq \frac{\mathbb{E}[V^k]}{(p+\tau)^k} \leq \frac{e^{\varepsilon k}\mathcal{M}_k[B(n,p)]}{p^k(1+\tau/p)^k} + \frac{\delta}{(p+\tau)^k}.$$

Using Lemma 30 we obtain that

$$\mathbb{P}[V \geq p+\tau] \leq \frac{p^k + (k\ln n+1)\cdot\left(\frac{k}{n}\right)^k}{e^{-\varepsilon k}p^k(1+\tau/p)^k} + \frac{\delta}{(p+\tau)^k} = \frac{1+(k\ln n+1)\cdot\left(\frac{k}{pn}\right)^k}{(e^{-\varepsilon}(1+\tau/p))^k} + \frac{\delta}{(p+\tau)^k}. \quad (14)$$

Using the condition $\varepsilon \leq \tau/2$ and $\tau \leq 1/3$ we first observe that

$$e^{-\varepsilon}(1+\tau/p) \geq (1-\varepsilon)(1+\tau/p) = 1+\tau/p-\varepsilon-\varepsilon\tau/p \geq 1+\tau/(3p).$$

Hence, with the condition that $k \geq 4p\ln(2/\beta)/\tau$ we get

$$(e^{-\varepsilon}(1+\tau/p))^k \geq (1+\tau/(3p))^k \geq e^{k\tau/(4p)} \geq \frac{2}{\beta}. \quad (15)$$

Using the condition $n \geq 3k/\tau$.

$$e^{-\varepsilon}\tau/p \geq 3e^{-\varepsilon}k/(np) > 2k/(np).$$

Together with the condition $k \geq \max\{4\ln(2/\beta)/\tau,\ 2\log\log n\}$, we have

$$\log(2/\beta)+\log(k\ln n+1) \leq \log(2/\beta)+\log(k+1)+\log\log n \leq k$$

since $k/2 \geq \log\log n$ holds by assumption and for $k \geq 12\ln(2/\beta)$, $k/6 \geq \log(2/\beta)$ and $k/3 \geq \log(k+1)$ (whenever $\beta < 2/3$). Therefore we get

$$(e^{-\varepsilon}(1+\tau/p))^k \geq (e^{-\varepsilon}\tau/p)^k \geq 2^k\cdot\left(\frac{k}{pn}\right)^k \geq \frac{2}{\beta}\cdot(k\ln n+1)\cdot\left(\frac{k}{pn}\right)^k. \quad (16)$$

Combining eq.(15) and (16) we obtain that

$$\frac{1+(k\ln n+1)\cdot\left(\frac{k}{pn}\right)^k}{(e^{-\varepsilon}(1+\tau/p))^k} \leq \beta/2+\beta/2 = \beta.$$

Substituting this into eq.(14) we obtain the claim. □