Preserving Statistical Validity in Adaptive Data Analysis

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

October 30, 2014

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.

While being able to explore the data and adapt the analysis to it is often advantageous, it is also known to easily lead to false discovery and overfitting [SNS11, GL13]. In this work we initiate the study of techniques to preserve the validity of statistical inference in adaptive data analysis. As an instance of this problem we propose the question of estimating the expectations on the underlying data distribution of m adaptively chosen functions given a dataset sampled i.i.d. from the distribution.

We observe that the number of samples necessary might grow linearly in m when standard empirical estimates are used to evaluate the expectations. However we show that remarkably, there is a way to estimate the expectations that allows an analyst – who is free to choose the functions on the basis of the estimates she has already obtained – to obtain an estimate of the expectation for each of the *exponentially many* functions with high probability. This technique counter-intuitively involves actively perturbing and coordinating the answers given to the queries of the analyst, using techniques developed for privacy preservation.

Finally, we show a computationally efficient method that given n samples, can answer exponentially many such queries so long as the number of rounds of adaptivity is $o(n^2)$. This matches computational lower bounds recently proved for our question [HU14, SU14].

*Microsoft Research
†IBM Almaden Research Center
‡IBM Almaden Research Center
§University of Toronto
¶omer.reingold@gmail.com
‖Department of Computer and Information Science, University of Pennsylvania, aaroth@cis.upenn.edu

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 recognized 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, GL13].

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 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).

[Vitaly's Note: This example could be omitted from this version but seems potentially useful in making the issue a bit more concrete] 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 nature 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 1.3 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 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 as a sequence of adaptively chosen, individually differentially private, steps.

1.1 Our Results

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 tolerance τ that is correct with high probability (or, equivalently, a confidence interval with high confidence level).

We make this choice 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], 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, BDMN05b, CKL+06] for examples). Finally, the problem of providing accurate answers to a large number of queries for the average value of a hypothesis 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 sample complexity that scales linearly with the number of queries m. We observe that such bad dependence is inherent in the standard approach of estimating expectations by (exact) empirical average on the samples. This is directly implied by the techniques from [DN03b] who show how to make linearly many non-adaptive counting queries to a dataset, and reconstruct nearly all of it. Once the dataset 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 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.

Our main result is that, remarkably, it is possible to evaluate *exponentially many* adaptively

[1]The average value of a hypothesis 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*.

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 *logarithmic*, nearly matching the dependence that is necessary for non-adaptively chosen queries.

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

$$n_0 = O\left(\frac{\log m \sqrt{\log |\mathcal{X}|}}{\tau^{7/2}}\right).$$

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

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 $|\mathcal{X}|$.

At the same time, our techniques can be used to give an efficient algorithm that can answer an exponential number of statistical queries as long as the total number of rounds of adaptivity is $n^{2-\omega(1)}$. Importantly, this can be achieved even without the analyst explicitly specifying when each of the rounds starts. This gives an almost quadratic improvement over the naïve method which, in addition, would require knowing the boundaries of rounds.

[Vitaly's Note: I've changed the description to the stronger, round-based result but the theorem is old and needs to be updated.]

Theorem 2 (Informal, some parameters left out)**.** *There exists a computationally efficient algorithm for evaluating m adaptively chosen hypotheses, such that with high probability, the answers to every hypothesis are valid up to tolerance τ, given a data set of size at least $n \geq n_0$ for:*

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

1.2 Overview of Techniques

Our results follow from a basic connection we make between *differential privacy* and *generalization*, which might 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 (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 expected value $\mathcal{E}_{\boldsymbol{S}}[\psi]$ is exactly equal to 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 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 argument works only for $(\varepsilon, 0)$-differential privacy due to conditioning on the event $\boldsymbol{\phi} = \psi$ which might have arbitrarily low probability. We use a more delicate conditioning to obtain the extension to the more general (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 differentially private algorithms $\mathcal{A}$ that map a database $\boldsymbol{S} \sim \mathcal{P}^n$ to elements from some arbitrary range Z. We prove says 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 the adaptivity is 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.

[Vitaly's Note: Need to add something about the proof here.]

Note the seeming disconnect between these theorems and our applications: the theorems hold for a function generated by a differentially private algorithm. On the other hand, we want to estimate expectation of queries generated by an analyst whom we do not assume to be restricted in any way. The connection comes from the post-processing guarantee of differential privacy: *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 in $\boldsymbol{S}$. Therefore it is sufficient to ensure that the algorithm that answers the queries is differentially private and apply our theorems.

1.3 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 examples and references). In contrast, our framework addresses multiple stages of adaptive decisions, possible lack of 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 *family-wise* measures of error 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 [GL13].

False discovery controlling procedures have been developed for 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 e.g. [KV94] 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 than 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 which there are several variants) 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 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 composition guarantees. This in particular 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 classification.

Differential privacy emerged from a line of work [DN03b, DN04, BDMN05a], 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 hypotheses on

a fixed dataset while preserving differential privacy [RR10, HR10], which is what yields the main application in our paper, when combined with out main theorem. See [Dwo11] for a short survey and [DR14] for a textbook introduction to differential privacy. It has been known as folklore that a hypothesis output by a differentially private learning algorithm generalizes *in expectation* [2]. Our technique can be seen as a substantial strengthening and generalization of these observations: from expectation to high probability bounds (that is crucial for answering many queries) and beyond the expected error of a hypothesis.

Finally, inspired by this work [HU14] and [SU14] have proven complementary *computational* lower bounds for the problem formulated in this paper. In short, the work of [HU14, SU14] shows that the exponential running time of the algorithm instantiating our main result is unavoidable, and that the sample complexity of our efficient algorithm is nearly optimal, among all computationally efficient mechanisms for evaluating hypotheses.

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 3. *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 4. [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] We are aware of these folklore results via conversations with Kunal Talwar and Frank McSherry, who originally discussed these ideas. Frank McSherry also has a blog post on related ideas: `http://windowsontheory.org/2014/02/04/differential-privacy-for-measure-concentration/`

3 Differential Privacy and Preservation of Moments

We now prove that if a hypothesis $\boldsymbol{\phi}$ 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 5. *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 $\mathcal{A}(\boldsymbol{S})$,*

$$\mathbb{E}\left[\mathcal{E}_{\boldsymbol{S}}[\boldsymbol{\phi}]^k \;\middle|\; \boldsymbol{\phi} = \psi\right] \leq e^{k\varepsilon} \cdot \mathbb{E}\left[\mathcal{E}_{\boldsymbol{T}}[\psi]^k\right]. \tag{1}$$

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,

$$\mathcal{E}_T[\psi]^k = \mathbb{E}[\Pi_T^{\boldsymbol{I}}(\psi)]. \tag{2}$$

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

$$\begin{aligned}
\mathbb{E}\left[\Pi_{\boldsymbol{S}}^I(\boldsymbol{\phi}) \mid \boldsymbol{\phi} = \psi\right] &= \mathbb{E}\left[\Pi_{\boldsymbol{S}_{I \leftarrow \boldsymbol{T}}}^I(\mathcal{A}(\boldsymbol{S}_{I \leftarrow \boldsymbol{T}})) \mid \mathcal{A}(\boldsymbol{S}_{I \leftarrow \boldsymbol{T}}) = \psi\right] \\
&= \mathbb{E}\left[\Pi_{\boldsymbol{T}}^I(\mathcal{A}(\boldsymbol{S}_{I \leftarrow \boldsymbol{T}})) \mid \mathcal{A}(\boldsymbol{S}_{I \leftarrow \boldsymbol{T}}) = \psi\right] \\
&= \int_0^1 \frac{\mathbb{P}\left[\Pi_{\boldsymbol{T}}^I(\mathcal{A}(\boldsymbol{S}_{I \leftarrow \boldsymbol{T}})) \geq t \text{ and } \mathcal{A}(\boldsymbol{S}_{I \leftarrow \boldsymbol{T}}) = \psi\right]}{\mathbb{P}\left[\mathcal{A}(\boldsymbol{S}_{I \leftarrow \boldsymbol{T}}) = \psi\right]} \mathrm{d}t \\
&= \int_0^1 \frac{\mathbb{P}\left[\Pi_{\boldsymbol{T}}^I(\mathcal{A}(\boldsymbol{S}_{I \leftarrow \boldsymbol{T}})) \geq t \text{ and } \mathcal{A}(\boldsymbol{S}_{I \leftarrow \boldsymbol{T}}) = \psi\right]}{\mathbb{P}\left[\boldsymbol{\phi} = \psi\right]} \mathrm{d}t
\end{aligned} \tag{3}$$

Now for any fixed t, S and T consider the event $\Pi_T^I(\mathcal{A}(S)) \geq t$ 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 19, the distribution $\mathcal{A}(S)$ and the distribution $\mathcal{A}(S_{I\leftarrow T})$:

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

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

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

Substituting this into eq. (3) we get

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

Taking the expectation over $\boldsymbol{I}$ and using eq. (2) 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 6. *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 Lemmas 24 and 25 we obtain that RHS of our bound in Lemma 1 is at most $e^{\varepsilon k}\mathcal{M}_k[B(n, \mathcal{P}[\psi])]$. We use Lemma 26 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 $\{\boldsymbol{\phi} = \phi\}$ appearing in Lemma 5 may have arbitrarily small probability. A simple extension of the previous proof would lead to an error of $\delta / \mathbb{P}[\boldsymbol{\phi} = \phi]$. We avoid this issue by using a more carefully chosen condition. Specifically, instead of restricting $\boldsymbol{\phi}$ to be equal to a particular function ϕ, 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 7. *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 6 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:

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

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. (3), we obtain

$$\mathbb{E}\left[\Pi^I_{\boldsymbol{S}}(\boldsymbol{\phi}) \mid \mathcal{P}[\boldsymbol{\phi}] \in B_\ell\right] = \int_0^1 \frac{\mathbb{P}\left[\Pi^I_{\boldsymbol{T}}(\mathcal{A}(\boldsymbol{S}_{I \leftarrow \boldsymbol{T}})) \geq t \text{ and } \mathcal{P}[\mathcal{A}(\boldsymbol{S}_{I \leftarrow \boldsymbol{T}})] \in B_\ell\right]}{\mathbb{P}\left[\mathcal{P}[\boldsymbol{\phi}] \in B_\ell\right]} dt \tag{5}$$

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 19, 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:

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

Taking the probability over $\boldsymbol{S}$ and $\boldsymbol{T}$ and substituting this into eq. (5) we get

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

Taking the expectation over $\boldsymbol{I}$ and using eq. (2) we obtain:

$$\mathbb{E}\left[\mathcal{E}_{\boldsymbol{S}}[\boldsymbol{\phi}]^k \;\middle|\; \mathcal{P}[\boldsymbol{\phi}] \in B_\ell\right] \leq e^{k\varepsilon}\,\mathbb{E}\left[\mathcal{E}_{\boldsymbol{T}}[\boldsymbol{\phi}]^k \;\middle|\; \mathcal{P}[\boldsymbol{\phi}] \in B_\ell\right] + \frac{e^{(k-1)\varepsilon}\delta}{\mathbb{P}\left[\mathcal{P}[\boldsymbol{\phi}] \in B_\ell\right]}. \tag{6}$$

Conditioned on $\mathcal{P}[\boldsymbol{\phi}] \in B_\ell$, $\mathcal{P}[\boldsymbol{\phi}] \leq \tau\ell$ and therefore by Lemma 24,

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

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. (6) we obtain the claim in eq. (4).

As before, we use Lemma 26 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

$$\mathbb{P}\left[\mathcal{E}_{\boldsymbol{S}}[\boldsymbol{\phi}] \geq \tau\ell + \tau \mid \mathcal{P}[\boldsymbol{\phi}] \in B_\ell\right] \leq \beta/2 + \frac{\delta e^{(k-1)\varepsilon}\cdot 4L}{\beta(\tau\ell+\tau)^k}, \tag{7}$$

Using condition $\delta = \exp(-2\cdot\ln(4/\beta)/\tau)$ and inequality $\ln(x) \leq x/e$ 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. (7) 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 the empirical average

Our previous results dealt with the empirical average of a hypotheses $\phi\colon \mathcal{X} \to [0,1]$. A different way of looking at our result is to define for each hypothesis ϕ a set $R(\phi)$ containing all datasets S such that ϕ is far from the correct value $\mathcal{P}[\phi]$ on S. Formally, $R(\phi) = \{S\colon |\mathcal{E}_S[\phi] - \mathcal{P}[\phi]| > \tau\}$. Our result 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 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 8. *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 23), implies that for any $t > 0$,

$$\mathbb{P}[g(\boldsymbol{S}) \geq \varepsilon t] \leq e^{-2t^2/n}. \tag{8}$$

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

By inequality (8) we have that for $i \ge 1$, $\mathbb{P}[g(\boldsymbol{S}) \ge \varepsilon\sqrt{n\ln(2^i/\beta)/2}] \le \beta/2^i$. Therefore, for all $i \ge 0$,

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

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

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

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

Therefore,

$$\begin{aligned}
\mathbb{P}[\boldsymbol{S} \in B_i \cap R(y) \mid \boldsymbol{Y} = y] &= \sum_{S \in B_i \cap R(y)} \mathbb{P}[\boldsymbol{S} = S \mid \boldsymbol{Y} = y] \\
&\le \exp\left(\varepsilon\sqrt{n\ln(2^{i+1}/\beta)/2}\right) \cdot \sum_{S \in B_i \cap R(y)} \mathbb{P}[\boldsymbol{S} = S] \\
&= \exp\left(\varepsilon\sqrt{n\ln(2^{i+1}/\beta)/2}\right) \cdot \mathbb{P}[\boldsymbol{S} \in B_i \cap R(y)] \\
&\le \exp\left(\varepsilon\sqrt{n\ln(2^{i+1}/\beta)/2} - \ln(2^i/\beta)\right). \qquad (9)
\end{aligned}$$

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

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

Substituting this into inequality (9), we get

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

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

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

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

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

□

Our theorem gives a result for statistical queries that achieves the same bound as our earlier result in Theorem 6 up to constant factors in the parameters.

Corollary 9. *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 8 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

We now spell out the corollaries of our connection to differential privacy (Theorems 6 and 7) formally.

5.1 Laplacian Noise Addition

Theorem 10 (Laplace)**.** *Let $\tau, \beta, \epsilon > 0$ and define*

$$n_0(\tau, \beta, \epsilon, k) = \frac{k \log(1/\beta)}{\epsilon\tau}.$$

There is an $(\epsilon, 0)$-differentially private algorithm called Laplace *which on input of a data set S of size n accepts any sequence of k adaptively chosen hypotheses $\phi_1, \ldots, \phi_k$ and returns estimates $a_1, \ldots, a_k$ such that for all $i \in [k]$ we have $\mathbb{P}\left[|\mathcal{E}_S[\phi_i] - a_i| > \tau\right] \leq \beta$ provided that $n \geq Cn_0(\tau, \beta, \epsilon, k)$ for sufficiently large constant C.*

Corollary 11. *Let $\tau, \beta > 0$ and define*

$$n_0(\tau, \beta, k) = \frac{k \log(1/\beta)}{\tau^2}.$$

There is an algorithm which on input of a data set S of size n sampled from $\mathcal{P}^n$ accepts any sequence of k adaptively chosen hypotheses $\phi_1, \ldots, \phi_k$ and returns estimates $a_1, \ldots, a_k$ such that for all $i \in [k]$ we have $\mathbb{P}\left[|\mathcal{P}[\phi_i] - a_i| > \tau\right] \leq \beta$ provided that $n \geq Cn_0(\tau, \beta, k)$ for sufficiently large constant C.

Proof. We apply Theorem 6 with $\epsilon = \tau/2$ and plug this choice of ϵ into the definition of n_0 in Theorem 14. We note that the stated lower bound on n implies the lower bound required by Theorem 6. □

Theorem 12 (Laplace)**.** *Let $\tau, \beta, \epsilon, \delta > 0$ and define*

$$n_0(\tau, \beta, \epsilon, \delta, k) = \frac{\sqrt{k \log(1/\delta)} \log(1/\beta)}{\epsilon \tau}.$$

There is an (ϵ, δ)-differentially private algorithm called Laplace *which on input of a data set S of size n accepts any sequence of k adaptively chosen hypotheses $\phi_1, \ldots, \phi_k$ and returns estimates $a_1, \ldots, a_k$ such that for all $i \in [k]$ we have $\mathbb{P}\left[|\mathcal{E}_S[\phi_i] - a_i| > \tau\right] \leq \beta$ provided that $n \geq C n_0(\tau, \beta, \epsilon, \delta, k)$ for sufficiently large constant C.*

Corollary 13. *Let $\tau, \beta > 0$ and define*

$$n_0(\tau, \beta, k) = \frac{\sqrt{k} \log^{1.5}(1/\beta)}{\tau^{2.5}}.$$

There is an algorithm which on input of a data set S of size n sampled from $\mathcal{P}^n$ accepts any sequence of k adaptively chosen hypotheses $\phi_1, \ldots, \phi_k$ and returns estimates $a_1, \ldots, a_k$ such that for all $i \in [k]$ we have $\mathbb{P}\left[|\mathcal{P}[\phi_i] - a_i| > \tau\right] \leq \beta$ provided that $n \geq C n_0(\tau, \beta, k)$ for sufficiently large constant C.

Proof. We apply Theorem 7 with $\epsilon = \tau/2$ and $\delta = \exp(-4 \ln(8/\beta)/\tau)$. Plugging thse parameters into the the definition of n_0 in Theorem 16 gives the stated lower bound on n. We note that the stated lower bound on n implies the lower bound required by Theorem 7. □

5.2 Multiplicative Weights Technique

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

$$n_0(\tau, \beta, \epsilon) = \frac{\log(|\mathcal{X}|) \log(n \log(|\mathcal{X}|)/\beta)}{\epsilon \tau^3}.$$

There is an $(\epsilon, 0)$-differentially private algorithm called PMW *which on input of a data set S of size n accepts any sequence of k adaptively chosen hypotheses $\phi_1, \ldots, \phi_k$ and returns estimates $a_1, \ldots, a_k$ such that for all $i \in [k]$ we have $\mathbb{P}\left[|\mathcal{E}_S[\phi_i] - a_i| > \tau\right] \leq \beta$ provided that $n \geq C n_0(\tau, \beta, \epsilon)$ for sufficiently large constant C.*

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

$$n_0(\tau, \beta) = \frac{\log(|\mathcal{X}|) \log(n \log(|\mathcal{X}|)/\beta)}{\tau^4}.$$

There is an algorithm which on input of a data set S of size n sampled from $\mathcal{P}^n$ accepts any sequence of k adaptively chosen hypotheses $\phi_1, \ldots, \phi_k$ and returns estimates $a_1, \ldots, a_k$ such that for all $i \in [k]$ we have $\mathbb{P}\left[|\mathcal{P}[\phi_i] - a_i| > \tau\right] \leq \beta$ provided that $n \geq C n_0(\tau, \beta)$ for sufficiently large constant C.

Proof. We apply Theorem 6 with $\epsilon = \tau/2$ and plug this choice of ϵ into the definition of n_0 in Theorem 14. We note that the stated lower bound on n implies the lower bound required by Theorem 6. □

PMW also satisfies (ϵ, δ)-differential privacy with the following quantitative bound.

Theorem 16 (Private Multiplicative Weights)**.** *Let $\tau, \beta, \epsilon, \delta > 0$ and define*

$$n_0(\tau, \beta, \epsilon, \delta) = \frac{\sqrt{\log(|\mathcal{X}|)\log(1/\delta)}\log(n/\beta)}{\epsilon\tau^2}.$$

There is an (ϵ, δ)-differentially private algorithm called PMW *which on input of a data set S of size n accepts any sequence of k adaptively chosen hypotheses $\phi_1, \ldots, \phi_k$ and returns estimates $a_1, \ldots, a_k$ such that for all $i \in [k]$ we have $\mathbb{P}\left[|\mathcal{E}_S[\phi_i] - a_i| > \tau\right] \leq \beta$, provided that $n \geq Cn_0(\tau, \beta, \epsilon, \delta)$ for sufficiently large constant C.*

The previous guarantee gives the following corollary that improves the dependence on τ and $\log|\mathcal{X}|$ in Corollary 15 at the expense of a slightly worse dependence on β.

Corollary 17. *Let $\tau, \beta > 0$ and define*

$$n_0(\tau, \beta) = \frac{\sqrt{\log(|\mathcal{X}|)\log(1/\beta)}\log(n/\beta)}{\tau^{3.5}}.$$

There is an algorithm which on input of a data set S of size n sampled from $\mathcal{P}^n$ accepts any sequence of k adaptively chosen hypotheses $\phi_1, \ldots, \phi_k$ and returns estimates $a_1, \ldots, a_k$ such that for all $i \in [k]$ we have $\mathbb{P}\left[|\mathcal{P}[\phi_i] - a_i| > \tau\right] \leq \beta$ provided that $n \geq Cn_0(\tau, \beta)$ for sufficiently large constant C.

Proof. We apply Theorem 7 with $\epsilon = \tau/2$ and $\delta = \exp(-4\ln(8/\beta)/\tau)$. Plugging thse parameters into the the definition of n_0 in Theorem 16 gives the stated lower bound on n. We note that the stated lower bound on n implies the lower bound required by Theorem 7. □

Acknowledgements We would like to particularly thank Jon Ullman for many enlightening discussions about this work. We would also like to thank Sanjeev Arora, Avrim Blum, Dean Foster, Michael Kearns, Jon Kleinberg, and seminar audiences at Cornell, Johns Hopkins, MIT, Microsoft Research, and Yahoo Research for helpful comments.

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.

[BDMN05a] A. Blum, C. Dwork, F. McSherry, and K. Nissim. Practical privacy: the sulq framework. In *PODS*, 2005.

[BDMN05b] Avrim Blum, Cynthia Dwork, Frank McSherry, and Kobbi Nissim. Practical privacy: the SuLQ framework. In Chen Li, editor, *PODS*, pages 128–138. ACM, 2005.

[BE02] Olivier Bousquet and André Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 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] Y. Benjamini and Y. 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.

[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.

[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.

[DN03a] I. Dinur and K. Nissim. Revealing information while preserving privacy. In *Proceedings of the Twenty-Second ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, pages 202–210, 2003.

[DN03b] 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 Matthew K. Franklin, editor, *CRYPTO*, volume 3152 of *Lecture Notes in Computer Science*, pages 528–544. Springer, 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. *Communications of the ACM*, 54(1):86–95, 2011.

[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.

[GL13] Andrew Gelman and Eric Loken. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time, 2013.

[HR10] Moritz Hardt and Guy N Rothblum. A multiplicative weights mechanism for privacy-preserving data analysis. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 61–70. IEEE, 2010.

[HTF09] T. Hastie, R. Tibshirani, and J.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. *arXiv preprint arXiv:1408.1655*, 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.

[RR10] Aaron Roth and Tim Roughgarden. Interactive privacy via the median mechanism. In *Proceedings of the 42nd ACM symposium on Theory of computing*, pages 765–774. ACM, 2010.

[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.

[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.

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

A Adaptivity in fitting a linear model

The data analyst has n samples $D = \{x_1, \dots, x_n\}$ 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 $i = 1, \dots, d$, determine $s_i = \operatorname{sign}\Big(\sum_{x \in D} x_i \Big)$.
2. Let $\tilde{u} = \frac{1}{\sqrt{d}}(s_1, \dots, s_d)$.

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 objecive value.

Lemma 18. *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,

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

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

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

□

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 19. *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 20. *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 21. *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 20 and 21 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 [DN03a] *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 22 (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 23 (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 24. *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

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{i\in[n]} X_i\right)^k\right] = \mathop{\mathbb{E}}_{I\sim[n]^k}\left[\mathbb{E}\left[\prod_{j\in[k]} X_{i_j}\right]\right] = \mathop{\mathbb{E}}_{I\sim[n]^k}\left[\mathbb{E}\left[\prod_{j\in[k']} X_{\ell_j}^{k_j}\right]\right] = \mathop{\mathbb{E}}_{I\sim[n]^k}\left[\prod_{j\in[k']} \mathbb{E}\left[X_{\ell_j}^{k_j}\right]\right], \tag{10}$$

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 (10) 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 25. *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

$$\mathbb{E}[U^k] \leq p^k + \int_{p^k}^{1} \mathbb{P}[U^k \geq t]dt. \tag{11}$$

We substitute $t = (1+\gamma)^k p^k$ and observe that Lemma 22 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.(11) together with $\frac{dt}{d\gamma} = k(1+\gamma)^{k-1}\cdot p^k$ we obtain

$$\begin{aligned}
\mathbb{E}[U^k] &\leq p^k + \int_0^{1/p-1} \exp\left(-np((1+\gamma)\ln(1+\gamma)-\gamma)\right)\cdot k(1+\gamma)^{k-1}d\gamma \\
&= p^k + p^k k\int_0^{1/p-1} \frac{1}{1+\gamma}\cdot\exp\left(k\ln(1+\gamma) - np((1+\gamma)\ln(1+\gamma)-\gamma)\right)d\gamma \\
&\leq p^k + p^k k \max_{\gamma\in[0,1/p-1]}\left\{\exp\left(k\ln(1+\gamma) - np((1+\gamma)\ln(1+\gamma)-\gamma)\right)\right\}\cdot\int_0^{1/p-1}\frac{1}{1+\gamma}d\gamma \\
&= p^k + p^k k\ln(1/p)\cdot\max_{\gamma\in[0,1/p-1]}\left\{\exp\left(k\ln(1+\gamma) - np((1+\gamma)\ln(1+\gamma)-\gamma)\right)\right\}. \qquad (12)
\end{aligned}$$

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.(12) 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 26. *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 25 we obtain that

$$\mathbb{P}[V \geq p+\tau] \leq \frac{p^k + (k\ln n + 1)\cdot\left(\frac{k}{n}\right)^k}{e^{-\varepsilon k}p^k(1+\tau/p)^k} + \frac{\delta}{(p+\tau)^k} = \frac{1 + (k\ln n + 1)\cdot\left(\frac{k}{pn}\right)^k}{(e^{-\varepsilon}(1+\tau/p))^k} + \frac{\delta}{(p+\tau)^k}. \quad (13)$$

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

$$(e^{-\varepsilon}(1+\tau/p))^k \geq (1+\tau/(3p))^k \geq e^{k\tau/(4p)} \geq \frac{2}{\beta}. \quad (14)$$

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

$$(e^{-\varepsilon}(1+\tau/p))^k \geq (e^{-\varepsilon}\tau/p)^k \geq 2^k \cdot \left(\frac{k}{pn}\right)^k \geq \frac{2}{\beta} \cdot (k \ln n + 1) \cdot \left(\frac{k}{pn}\right)^k. \tag{15}$$

Combining eq.(14) and (15) 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.(13) we obtain the claim. □