LEVERAGING PRIVACY IN DATA ANALYSIS

Ryan Michael Rogers

A DISSERTATION

in

Applied Mathematics and Computational Science

Presented to the Faculties of the University of Pennsylvania in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

2017

Supervisor of Dissertation

Michael Kearns, Professor of Computer and Information Science and National Center Chair

Co-Supervisor of Dissertation

Aaron Roth, Associate Professor of Computer and Information Science

Graduate Group Chairperson

Charles L. Epstein, Thomas A. Scott Professor of Mathematics

Dissertation Committee

Salil Vadhan, Vicky Joseph Professor of Computer Science and Applied Mathematics,
Harvard University

Rakesh Vohra, George A. Weiss and Lydia Bravo Weiss University Professor

LEVERAGING PRIVACY IN DATA ANALYSIS

© COPYRIGHT

2017

Ryan Michael Rogers

This work is licensed under the
Creative Commons Attribution
NonCommercial-ShareAlike 3.0
License

To view a copy of this license, visit

`http://creativecommons.org/licenses/by-nc-sa/3.0/`

Dedicated to Daniela and Holden

ACKNOWLEDGEMENT

This dissertation would not be possible without the support of so many people. I am forever grateful for the incredible guidance and patience of both of my advisors, Michael Kearns and Aaron Roth. I am inspired by their genius and ability to explain complicated analyses in a comprehensible way. They made research enjoyable throughout my Ph.D. Before graduate school, I had never heard of differential privacy, but after seeing a talk from Aaron, I was hooked – Aaron is a great salesman for many research areas. I went to graduate school with the rough idea of working on algorithmic game theory. They helped me not just make contributions to algorithmic game theory but many other areas. I could not imagine graduate school without their encouragement and mentorship. Thank you!

I am thankful for the many professors in my undergraduate that persuaded me to go on to graduate school, including my research advisor at Stetson University, Thomas Vogel. Further, I would like to thank Richard Weber at Cambridge who encouraged me to get a Ph.D. and introduced me to game theory.

I am also very grateful to the other members in my dissertation committee. Thanks to Rakesh Vohra for teaching one of my favorite courses, *Submodularity and Discrete Convexity*. The techniques of this course turned out to be incredibly useful in future research, particularly in our work with Hsu et al. (2016). Also, thanks to Salil Vadhan, who helped me see the practicality of differential privacy when I interned at the Privacy Tools Project at Harvard. It was during this internship that I saw the widespread appeal of differential privacy to lawyers, social scientists, statisticians and other researchers. That internship was the first time that I considered the intersection of differential privacy and statistics, which is the focus of this dissertation.

There are many other collaborators that I would like to thank: Miro Dudík, Marco Gaboardi, Justin Hsu, Shahin Jabbari, Sampath Kannan, Daniel Kifer, Sebastién Lahaie, Hyun woo Lim, Jamie Morgenstern, Mallesh M. Pai, Adam Smith, Om Thakkar, Jonathan Ullman, Jennifer Wortman Vaughan, and Zhiwei Steven Wu. I would particularly like to thank Adam for his incredibly helpful advice. My visit to Penn State with Adam was one of the most productive weeks of my graduate school career. Further, I would like to thank my mentors and collaborators at Microsoft Research – Miro, Sebastién and Jennifer – for a wonderful internship.

Many postdocs and fellow graduate students are among the collaborators that I would like to thank. They helped me look at problems in different ways and allowed me to bounce ideas off them. Particularly, I would like to thank the “Happy Hour in One Hour” group, including: Kareem Amin, Rachel Cummings, Lili Dworkin, Justin Hsu, Hoda Heidari, Shahin Jabbari, Jamie Morgenstern, and Zhiwei Steven Wu.

It was only possible to keep my sanity in graduate school with the Wharton Crew group of rowers. They provided an escape from research and pushed me to the limit, physically and mentally. The many early mornings of rowing and coaching were demanding, but in hindsight totally worth it. Many of my highlights in graduate school were rowing with you

all, including the Head of the Schuylkill and Head of the Charles regattas as well as the Corvallis to Portland Race which lasted for a grueling 115 miles.

Last, but certainly not least, I would like to thank my family. Thanks to my Mom, Dad, and Brother for their love and support throughout my life, but particularly throughout my graduate school career. Special thanks to the loves of my life, my wife Daniela and our son Holden.[1]

---

[1] Holden takes full responsibility for all errors found in this dissertation, which was written mostly with me holding him with one arm.

ABSTRACT

LEVERAGING PRIVACY IN DATA ANALYSIS

Ryan Michael Rogers

Michael Kearns
Aaron Roth

Data analysis is inherently adaptive, where previous results may influence which tests are carried out on a single dataset as part of a series of exploratory analyses. Unfortunately, classical statistical tools break down once the choice of analysis may depend on the dataset, which leads to overfitting and spurious conclusions. In this dissertation we put constraints on what type of analyses can be used adaptively on the same dataset in order to ensure valid conclusions are made. Following a line of work initiated from Dwork et al. (2015c), we focus on extending the connection between differential privacy and adaptive data analysis.

Our first contribution follows work presented in Rogers et al. (2016a). We generalize and unify previous works in the area by showing that the generalization properties of (approximately) differentially private algorithms can be used to give valid p-value corrections in adaptive hypothesis testing while recovering results for statistical and low-sensitivity queries. One of the main benefits of differential privacy is that it composes, i.e. the combination of several differentially private algorithms is itself differentially private and the privacy parameters degrade sublinearly. However, we can only apply the composition theorems when the privacy parameters are all fixed up front. Our second contribution then presents a framework for obtaining composition theorems when the privacy parameters, along with the number of procedures that are to be used, need not be fixed up front and can be adjusted adaptively (Rogers et al., 2016b). These results are only useful if there are some differentially private procedures that an analyst would want to use. Hence, we present differentially private hypothesis tests (Gaboardi et al., 2016; Kifer and Rogers, 2016).

TABLE OF CONTENTS

ACKNOWLEDGEMENT iv

ABSTRACT vi

LIST OF TABLES ix

LIST OF ILLUSTRATIONS xi

LIST OF ALGORITHMS xii

I PROBLEM STATEMENT AND PREVIOUS WORK **1**

CHAPTER 1 :INTRODUCTION 3
1.1 Problem Formulation - Statistical Queries 5
1.2 Prior Results - Statistical Queries 7
1.3 Post Selection Hypothesis Testing 8
1.4 Handling More General Analyses - Max-information 10
1.5 Algorithms with Bounded Max-information 12
1.6 Contributions 14
1.7 Related Work 16

CHAPTER 2 :PRIVACY PRELIMINARIES 18
2.1 Differential Privacy 18
2.2 Concentrated Differential Privacy 20

CHAPTER 3 :COMPARISON TO DATA-SPLITTING 22
3.1 Preliminaries 22
3.2 Confidence Bounds from Dwork et al. (2015a) 23
3.3 Confidence Bounds from Bassily et al. (2016) 25
3.4 Confidence Bounds combining work from Russo and Zou (2016) and Bassily et al. (2016) 25
3.5 Confidence Bound Results 29

II DIFFERENTIAL PRIVACY IN ADAPTIVE DATA ANALYSIS: INFORMATION AND COMPOSITION **32**

CHAPTER 4 :MAX-INFORMATION, DIFFERENTIAL PRIVACY, AND POST-SELECTION HYPOTHESIS TESTING 34
4.1 Additional Preliminaries 35
4.2 Max-information for (ϵ, δ)-Differentially Private Algorithms 37
4.3 Comparison with Results from Bassily et al. (2016) 44

4.4 A Counterexample to Nontrivial Composition and a Lower Bound for Non-Product Distributions 45
4.5 Consequences of Lower Bound Result - Robust Generalization 48
4.6 Conversion between Mutual and Max-Information 50
4.7 Max-Information and Compression Schemes 54
4.8 Conclusion and Future Work 56

CHAPTER 5 :PRIVACY ODOMETERS AND FILTERS: PAY-AS-YOU-GO COMPOSITION 57
5.1 Results 57
5.2 Additional Preliminaries 58
5.3 Composition with Adaptively Chosen Parameters 60
5.4 Concentration Preliminaries 68
5.5 Advanced Composition for Privacy Filters 70
5.6 Advanced Composition for Privacy Odometers 71
5.7 zCDP Filters and Odometers 76
5.8 Conclusion and Future Work 79

III PRIVATE HYPOTHESIS TESTS **80**

CHAPTER 6 :PRIVATE CHI-SQUARE TESTS: GOODNESS OF FIT AND INDEPENDENCE TESTING 84
6.1 Goodness of Fit Testing 85
6.2 Independence Testing 93
6.3 Significance Results 99
6.4 Power Results 101
6.5 Conclusion 102

CHAPTER 7 :PRIVATE GENERAL CHI-SQUARE TESTS 105
7.1 General Chi-Square Tests 105
7.2 Private Goodness of Fit Tests 109
7.3 General Chi-Square Private Tests 122
7.4 General Chi-Square Tests with Arbitrary Noise Distributions 129
7.5 Conclusion 131

CHAPTER 8 :LOCAL PRIVATE HYPOTHESIS TESTS 132
8.1 Introduction 132
8.2 Local Private Chi-Square Tests 132
8.3 Ongoing Work 141

IV CONCLUSION **142**

APPENDIX 145

BIBLIOGRAPHY 155

LIST OF TABLES

TABLE 1 : Types of Errors in Hypothesis Testing 10

TABLE 2 : Contingency Table with Marginals. 93
TABLE 3 : GOF testing with $\alpha = 0.05$ and 0.00125-zCDP for $d = 100$. 100

LIST OF FIGURES

FIGURE 1 : Two models of adaptive data analysis. 5
FIGURE 2 : Interaction between analyst $\mathcal{A}$ and dataset $\boldsymbol{X}$ via algorithms $\mathcal{M}_1, \cdots, \mathcal{M}_k$. 6

FIGURE 3 : Widths of valid confidence intervals for k adaptively chosen statistical queries via data-splitting techniques or noise addition on the same dataset. 30
FIGURE 4 : Standard deviations of the Gaussian noise we added to each query to obtain the confidence widths. 31

FIGURE 5 : Empirical Type I Error in $10,000$ trials when using the classical `GOF` test without modification after incorporating noise due to privacy for $(\epsilon = 0.1)$-DP and $(\rho = \epsilon^2/8)$-zCDP. 87
FIGURE 6 : Empirical Type I Error of `AsymptGOF` with error bars corresponding to 1.96 times the standard error in 100,000 trials. 99
FIGURE 7 : (Log of) Critical values of `AsymptGOF` and `MCGOF` (with $m = 59$) with both Gaussian ($\rho = 0.00125$) and Laplace noise ($\epsilon = 0.1$) along with the classical critical value as the black line. 99
FIGURE 8 : Empirical Type I Error of `AsymptIndep` with error bars corresponding to 1.96 times the standard error in 1,000 trials. 101
FIGURE 9 : Empirical Type I Error of `MCIndep` using Gaussian noise with error bars corresponding to 1.96 times the standard error in 1,000 trials. 101
FIGURE 10 : Empirical Type I Error of `MCIndep` using Laplace noise with error bars corresponding to 1.96 times the standard error in $1,000$ trials. 102
FIGURE 11 : Comparison of empirical power of classical non-private test versus `AsymptGOF` (solid line) and `MCGOF` (dashed line) with Gaussian noise for alternate $\mathrm{H}_1 : \boldsymbol{p}^1 = \boldsymbol{p}^0 + \boldsymbol{\Delta}$ in $10,000$ trials. 103
FIGURE 12 : Comparison of empirical power of classical non-private test versus `MCGOF` with Laplace noise for alternate $\mathrm{H}_1 : \boldsymbol{p}^1 = \boldsymbol{p}^0 + \boldsymbol{\Delta}$ in $10,000$ trials. 103
FIGURE 13 : Comparison of empirical power of classical non-private test versus `AsymptIndep` (solid line) and `MCIndep` (dashed line) with Gaussian noise in $1,000$ trials. 104
FIGURE 14 : Comparison of empirical power of classical non-private test versus `MCIndep` with Laplace noise in $1,000$ trials. 104

FIGURE 15 : Empirical Type I Error for our goodness of fit tests in `NewStatAsymptGOF` with the nonprojected statistic $\mathrm{Q}_\rho^{(n)}$. 120
FIGURE 16 : Empirical Type I Error for our goodness of fit tests in `NewStatAsymptGOF` with the projected statistic $\mathcal{Q}_\rho^{(n)}$. 120

FIGURE 17 : Comparison of empirical power of classical non-private test versus `NewStatAsymptGOF` with both projected (solid line) and nonprojected statistics (dashed line). 121
FIGURE 18 : Comparison of empirical power between all zCDP hypothesis tests for goodness of fit and `NewStatAsymptGOF` with projected statistic. 122
FIGURE 19 : Empirical Type I Error for our new independence tests in `GenChiTest` with the nonprojected statistic. 126
FIGURE 20 : Empirical Type I Error for our new independence tests in `GenChiTest` with the projected statistic. 126
FIGURE 21 : Comparison of empirical power of classical non-private test versus `GenChiTest` in $1,000$ trials. The solid line is with the projected statistic and the dashed line is with the nonprojected statistic. . . 127
FIGURE 22 : Comparison of empirical power between all zCDP hypothesis tests for independence and `GenChiTest` with projected statistic. . . . 128
FIGURE 23 : A comparison of empirical power between `GenChiTest` with projected statistic and output perturbation from Yu et al. (2014) for independence testing for GWAS type datasets. 129

FIGURE 24 : Comparison of empirical power of classical non-private test versus local private tests `LocalGOF` (solid line) and `MC-GenChiTest` with projected statistic and Laplace noise (dashed line) for alternate $\mathrm{H}_1 : \boldsymbol{p}^1 = \boldsymbol{p}^0 + \boldsymbol{\Delta}$ in $1,000$ trials. We set $\boldsymbol{p}^0 = (1/4, 1/4, 1/4, 1/4)^\intercal$ and $\mathrm{H}_1 : \boldsymbol{p}^0 + \boldsymbol{\Delta}$ where $\boldsymbol{\Delta} = 0.01 \cdot (1, -1, -1, 1)^\intercal$. 136
FIGURE 25 : Comparison of empirical power of classical non-private test versus local private tests `LocalGOF` (solid line) and `MC-GenChiTest` with Laplace noise and projected statistic (dashed line) in $1,000$ trials. We set $\pi^{(1)} = \pi^{(2)} = 1/2$ and $\boldsymbol{\Delta} = 0.025 \cdot (1, -1)^\intercal(1, -1)$. 138

List of Algorithms

1 Monitor $\mathcal{W}_{\mathcal{D}}[\mathcal{M}, \mathcal{A}](\boldsymbol{X})$ 28
2 $\texttt{FixedParamComp}(\mathcal{A}, \mathcal{E}, b)$ 59
3 $\texttt{AdaptParamComp}(\mathcal{A}, k, b)$ 61
4 $\texttt{PrivacyFilterComp}(\mathcal{A}, k, b; \texttt{COMP}_{\epsilon_g,\delta_g})$ 63
5 $\texttt{SimulatedComp}(\mathcal{A}, k, b)$ 63
6 Stopping Time Adversary: $\mathcal{A}_{\epsilon,\delta}$ 73
7 Classical Goodness of Fit Test for Multinomial Data: `GOF` 85
8 MC Private Goodness of Fit: `MCGOF` 89
9 Private Chi-Squared Goodness of Fit Test: `AsymptGOF` 91
10 Pearson Chi-Squared Independence Test: `Indep` 94
11 Two Step MLE Calculation: `2MLE` 96
12 MC Independence Testing `MCIndep` 97
13 Private Independence Test for $r \times c$ tables: `AsymptIndep` 98
14 New Private Statistic Goodness of Fit Test: `NewStatAsymptGOF` 116
15 Private General Chi-Square Test: `GenChiTest` 124
16 Private Minimum Chi-Square Test using MC `MC-GenChiTest` 130
17 Exponential Mechanism $\mathcal{M}_{\texttt{EXP}}$ 134
18 Local DP GOF Test `LocalGOF` 135
19 Local DP General Chi-Square Test `LocalGeneral` 137
20 Local Randomizer $\mathcal{M}_{\texttt{FC}}$ 139
21 Extended Monitor $\mathcal{W}_{\mathcal{D}}[\mathcal{M}, \mathcal{A}](\vec{\boldsymbol{X}})$ 147
22 First Algorithm in Lower Bound Construction: $\mathcal{M}_1$ 150
23 Second Algorithm in Lower Bound Construction: $\mathcal{M}_2$ 150

Part I

PROBLEM STATEMENT AND PREVIOUS WORK

We first present the basic setup of adaptive data analysis including the motivation for why analysts that depend on statistical findings should be concerned with it and what some of the challenges are with adaptivity. We discuss some of the related work in this area, with the first paper in this line of research being Dwork et al. (2015c), which will be crucial to know when outlining our contributions. We then outline the rest of the dissertation, which largely follows work previously published in Rogers et al. (2016a,b); Gaboardi et al. (2016); Kifer and Rogers (2016), but does contain some results not previously published.

Some preliminaries in differential privacy are then presented which will be needed throughout the dissertation. We then conclude this first part of the thesis with empirical evaluations of valid confidence bounds that we can generate for adaptively chosen statistical queries using previous results and a new analysis, which have not been compared before. This then provides a link between the highly theoretical work in adaptive data analysis with some realistic settings and demonstrates an improvement over traditional data splitting techniques.

CHAPTER 1

INTRODUCTION

The goal of statistics and machine learning is to draw conclusions on a dataset that will *generalize* to the overall population, so that the same conclusion can be drawn from any new dataset that is collected from the same population. Tools from statistical theory have become ubiquitous in empirical science stretching across a myriad of disciplines.

However, classical statistical tools are only useful insofar as the original theory was intended. The scientific community has become increasingly aware that many of the "statistically significant" findings in published research are frequently invalid. In many replication studies, the published findings cannot be confirmed in a large proportion of them; much more than would be allowed by the theory, e.g. Ioannidis (2005); Gelman and Loken (2014). When a conclusion is made on a given dataset but cannot be replicated in other studies, then a *false discovery* has been committed. Similarly, in machine learning, the validity of models are based on how well the model generalizes to new instances, with the main concern being that the model *overfits* to the dataset and not the population.

Why is their an apparent disconnect between what theoretical statistics guarantees and the overwhelming number of *false discoveries* that are made from empirical studies? One of the crucial assumptions made in the classical theory is that the procedures that are to be conducted on the dataset are all known upfront, prior to actually seeing the data. In fact, one of the main suspects behind the prevalence of false discovery in replication studies is that the data analyst is *adaptively* selecting different analyses to run based on previous results on the same dataset and using the classical theory as if the tests were selected *independently* of the data (Gelman and Loken, 2014). This problem of adapting the analysis to the data is sometimes referred to as "p-hacking", "data snooping", and "researcher degrees of freedom" (Ioannidis, 2005; Simmons et al., 2011; Gelman and Loken, 2014). As soon as the analyst has looked at the data or some function of it and then selects a new analysis, the traditional theory is no longer valid. For example, it may be the case that the analyst wishes to select some variables for a model selection followed by some inference – note that the adaptive selection of the model invalidates the following inference using classical statistical theory. Over the past few decades there has been a significant amount of effort put into proposing fixes to this problem. Despite some techniques for preventing false discoveries, e.g. the Bonferroni Correction (Bonferroni, 1936; Dunn, 1961) and the Benjamini-Hochberg Procedure (Benjamini and Hochberg, 1995), the problem still persists.

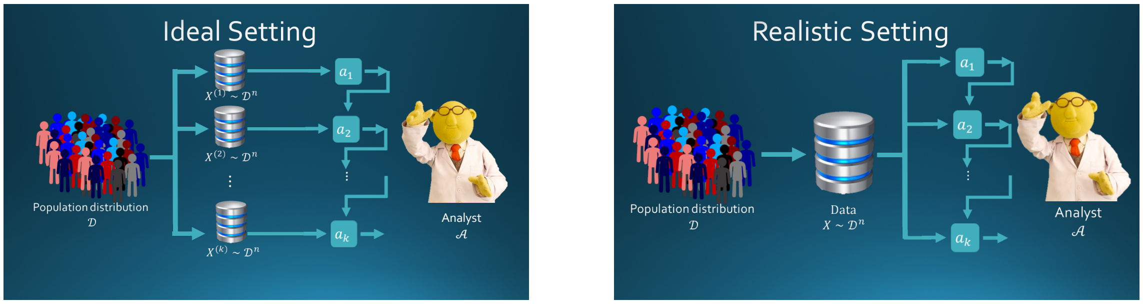
The practice of modern data analysis is inherently adaptive, where each analysis is conducted based on previous outcomes on the same data as part of an exploratory analysis. It may not be the case that a particular study would be thought of prior to running a test on the data, thus making preregistering what analyses you want to run useless. Typically, an analyst needs to use the data to find interesting analyses to perform and hypotheses to test. Further, as researchers increasingly allow open access of their data, multiple studies may be conducted on the same dataset where findings of different research groups may influence the studies performed by other research groups. Not taking into account the adaptivity in the separate research groups' analyses, this process can often lead to false conclusions drawn from a dataset, thus contributing to the crisis in reproducibility.

In order to use classical techniques in this adaptive setting, we would require the analyst to sample a fresh dataset with each new analysis to be conducted. Due to data collection being costly, this is certainly not an ideal solution. Instead, we would like to be able to consider several, adaptively chosen analyses on the same dataset and ensure valid conclusions are made that generalize to the population.

Recently, two lines of work have attempted to understand and mitigate the prevalence of false discoveries in adaptive data analysis. The first is to derive tight confidence intervals around parameter estimates from very specific types of analyses, such as LASSO model selection followed by regression (Fithian et al., 2014; Lee et al., 2013). The second line of work originated in the computer science community by Dwork et al. (2015c) and seeks to be very general by imposing conditions on the types of algorithms that carry out the analysis at each stage and makes no assumption on how the results are used by the analyst. Note that in the former line of work – called selective inference – the methods are focused on two stage problems: variable selection followed by significance testing and adjust for the inference in the second step. The main idea of the latter line of work aims to limit the amount of *information* (a notion which we will make precise later) that is released about the dataset with each analysis so that it is unlikely to commit a false discovery on a subsequent analysis. This dissertation is largely a continuation of the work initiated by Dwork et al. (2015c) and aims to further understand how we can correct for adaptivity in the classical theory. We then present some useful hypothesis tests which can be used in this adaptive setting while providing valid p-values.

Before we can discuss the specific contributions of this dissertation, we need to first discuss some of the previous work done in adaptive data analysis. We start by presenting a basic setup of the problem so that we can discuss the difficulty that arises when we need to consider analyses that are conducted adaptively on the same dataset as opposed to having the analyses known upfront. We will then discuss further advances in understanding adaptive data analysis and some of the results that have been shown.

Throughout this dissertation, we will write the data universe as $\mathcal{X}$, typically $\mathcal{X} = \{0,1\}^d$ where d is the dimensionality of the data, and some unknown data distribution $\mathcal{D}$ over $\mathcal{X}$ where a dataset $\boldsymbol{X} = (X_1, \cdots, X_n)$ of n subjects is typically sampled i.i.d. from $\mathcal{D}$, denoted as $\boldsymbol{X} \sim \mathcal{D}^n$. The analyst's goal is to infer something from the population rather than the dataset. An analyst will then select a sequence of analyses that she wants to conduct, receiving answers $a_1, \cdots, a_k$ that are computed using the dataset. Ideally, we would want the analyst to run each analysis separately on a fresh dataset $\boldsymbol{X}^{(i)}$ sampled from the same population distribution $\mathcal{D}^n$, thus ensuring that each analysis is independent of the data it is used on. Realistically, due to data collection being costly, we would like to reuse the same dataset for each analysis. See Figure 1 for a cartoon comparison between the ideal and realistic settings of adaptive data analysis.



(a) The ideal setting – each analysis is performed on a fresh dataset.

(b) A more realistic setting – a single dataset is reused for adaptively chosen analyses.

Figure 1: Two models of adaptive data analysis.

1.1. Problem Formulation - Statistical Queries

We start by formulating the problem of adaptivity for a simple setting that is standard in statistics and statistical learning theory, although later we will look at more general analyses that the analyst can conduct on the data. Here, we address the problem where the analyst wants to know $\mathbb{E}_{X\sim\mathcal{D}}\left[\phi(X)\right]$ where $\phi : \mathcal{X} \to [0,1]$. For notational convenience we will write this expectation as $\phi(\mathcal{D}) \overset{\text{defn}}{=} \mathbb{E}_{X\sim\mathcal{D}}\left[\phi(X)\right]$. The analyst then wants to obtain an estimate to $\phi(\mathcal{D})$ that is within tolerance τ with only access to sampled data $\boldsymbol{X} \sim \mathcal{D}^n$.

This formulation of the problem is nice because this estimate is a statistical query in the SQ model of Kearns (1993). We then model the interaction between an analyst $\mathcal{A}$ wanting to ask such queries $\phi_i : \mathcal{X} \to [0,1]$ and algorithms $\mathcal{M}_i$ for $i \in [k]$ which have direct access to the dataset $\boldsymbol{X} \sim \mathcal{D}^n$. Here we model the analyst $\mathcal{A}$ as first selecting ϕ_1 and receiving answer $a_1 = \mathcal{M}_1(\boldsymbol{X})$, then for $i = 2, \cdots, k$, we allow $\mathcal{A}$ to select ϕ_i as a function of $\phi_1, \cdots, \phi_{i-1}$ and answers $a_1, \cdots, a_{i-1}$ and she receives answer $a_i = \mathcal{M}_i(\boldsymbol{X})$. Note that the algorithms $\mathcal{M}_i$ may also depend on the previous queries and answers. One example for algorithm $\mathcal{M}_i(\boldsymbol{X})$ is to report the empirical average, $\phi_i(\boldsymbol{X}) \overset{\text{defn}}{=} \frac{1}{n}\sum_{j=1}^{n} \phi_i(X_i)$, which we know will be close to $\phi_i(\mathcal{D})$ when ϕ_i are chosen independently of the data. Thus the analyst makes decisions on what queries to ask based only on the outcomes of the algorithms $\mathcal{M}_1, \cdots, \mathcal{M}_k$. We then outline this interaction so that for each $i \in [k]$ (also see Figure 2):

- Analyst $\mathcal{A}$ selects query ϕ_i, which is based on previous queries $\phi_1, \cdots, \phi_{i-1}$ and corresponding answers $a_1, \cdots, a_{i-1}$
- $\mathcal{A}$ receives answer $a_i = \mathcal{M}_i(\boldsymbol{X})$, where $\mathcal{M}_i$ may also depend on $\phi_1, \cdots, \phi_{i-1}$ and $a_1, \cdots, a_{i-1}$.

We then define what we mean by accuracy in this setting.

Definition 1.1.1. *A sequence of algorithms* $\mathcal{M} = (\mathcal{M}_1, \cdots, \mathcal{M}_k)$ *is* (τ, β)-accurate with

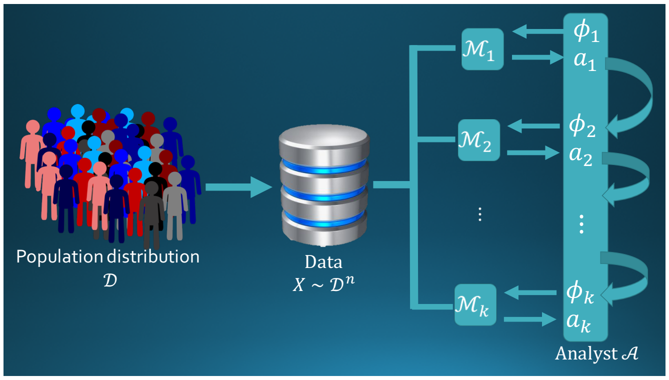


Figure 2: Interaction between analyst $\mathcal{A}$ and dataset $\boldsymbol{X}$ via algorithms $\mathcal{M}_1, \cdots, \mathcal{M}_k$.

respect to the population *if for all analysts* $\mathcal{A}$ *we have*

$$\Pr\left[\max_{i \in [k]} |\phi_i(\mathcal{D}) - \mathcal{M}_i(\boldsymbol{X})| \leq \tau\right] \geq 1 - \beta$$

where the probability is over the dataset $\boldsymbol{X} \sim \mathcal{D}^n$ *as well as any randomness from the algorithms* $\mathcal{M}_1 \cdots, \mathcal{M}_k$ *and the adversary* $\mathcal{A}$*.*

Before we dive deeper into the adaptive setting, we first consider the case where all the queries are asked up front, prior to any computations on the data. In this case we can apply a Chernoff bound and union bound to show that releasing the empirical average on the data is accurate,

$$\Pr_{\boldsymbol{X} \sim \mathcal{D}^n}\left[\max_{i \in [k]} |\phi_i(\mathcal{D}) - \phi_i(\boldsymbol{X})| \leq \sqrt{\frac{1}{2n}\log(2k/\beta)}\right] \geq 1 - \beta$$

A useful quantity for comparing the different methods in this section will be *sample complexity*, which gives a bound on the sample size n that is sufficient in answering k *nonadaptive* queries all with accuracy at most τ with constant probability, say $\beta = 0.05$. Thus, answering each (nonadaptively chosen) statistical query with the empirical average achieves sample complexity $n = \Theta\left(\log(k)/\tau^2\right)$.[1] Phrased another way, we can answer an exponential in n number of statistical queries and still achieve high accuracy on all of them. Further, it is a straightforward protocol that achieves this: simply answer with the empirical averages of each statistical query. However, this analysis crucially requires the statistical queries all be independent of the data.

---

[1] Throughout the dissertation, we will use $\log(\cdot)$ to denote the natural log unless we use another base, in which case we will make explicit the base b by writing $\log_b(\cdot)$

We now consider the setting where the statistical queries are adaptively chosen. One first approach might be to just answer each statistical query with the empirical average, as we did in the nonadaptive case. As pointed out in Hardt and Ullman (2014), we can use techniques from Dinur and Nissim (2003) to nearly reconstruct the entire database after seeing $O(\tau^2 n)$ many random empirical averages to statistical queries (nonadaptively chosen), so that the analyst can then find a statistical query q^* such that $|q^*(\boldsymbol{X}) - q^*(\mathcal{D})| > \tau$ with constant probability. This translates to empirical averages only having sample complexity $\Omega(k/\tau^2)$. Thus, only a linear number of statistical queries can be answered accurately using empirical estimates – an exponential blow up with a single round of adaptivity!

1.2. Prior Results - Statistical Queries

Given that empirical averages do not achieve good sample complexity with adaptively chosen statistical queries, we hope to find new ways in which to answer these queries accurately. There has then been a lot of work in developing algorithms that can answer much more than a linear number of adaptively selected statistical queries. The following result, which improves on an earlier result from Dwork et al. (2015c), shows that we can achieve a quadratic improvement on the number of adaptively selected queries, using techniques from Dwork et al. (2006b), which we will extensively go over in Chapter 3.

Theorem 1.2.1 [Bassily et al. (2016)]**.** *There is an algorithm that has the following sample complexity for k adaptively chosen statistical queries*

$$n \geq \widetilde{O}\left(\frac{\sqrt{k}}{\tau^2}\right).$$[2]

Further, the algorithm runs in time that is polynomial in n and $\log|\mathcal{X}|$ per query.

The following theorem, which also improves on an earlier result of Dwork et al. (2015c), shows that we can accurately answer an exponential in n number of adaptively selected statistical queries, but the algorithm which computes the answers is not run-time efficient. This result follows from the Private Multiplicative Weights algorithm from Hardt and Rothblum (2010).

Theorem 1.2.2 [Bassily et al. (2016)]**.** *There is an algorithm that has the following sample complexity for k adaptively chosen statistical queries*

$$n \geq \widetilde{O}\left(\frac{\sqrt{\log|\mathcal{X}|}\log(k)}{\tau^3}\right).$$

Further, the algorithm runs in time that is polynomial in n and $|\mathcal{X}|$ per query.

An immediate question that arises when comparing these results is why the gap between efficient and inefficient run-time algorithms (the second result requires $|\mathcal{X}|$ time per query) when improving on sample complexity? There was no such distinction in the nonadaptive setting. It turns out that this separation is actually inherent when answering adaptively

---

[2] We will use $\widetilde{O}(\cdot)$ throughout the dissertation to hide poly-logarithmic dependence on parameters that already appear, so that $\widetilde{O}(f(y)) = O(f(y)\text{polylog}(y))$ for some function f.

selected statistical queries accurately – adaptivity actually does come at a cost. The following result was first studied in Hardt and Ullman (2014) who gave the first computational barrier in answering adaptively selected queries and was then improved by Steinke and Ullman (2015).
Theorem 1.2.3 [Steinke and Ullman (2015)]. *Under a standard hardness assumption,*[3] *there is no computationally efficient algorithm that is accurate within constant tolerance with constant probability (over randomness of the sample and the algorithm) on* $k = O(n^2)$ *adaptively chosen statistical queries with* $\mathcal{X} = \{0, 1\}^d$.

It is worth pointing out here the dependence of this impossibility result on the dimensionality of the data. Note that if n were much larger than 2^d, then the empirical average of every possible statistical query could be answered accurately. So for these results to be interesting, we are considering $n \ll 2^d$.

The gap between the upper and lower bounds in sample complexity for adaptively chosen statistical queries is large. Hardt and Ullman (2014) and Steinke and Ullman (2015) showed that $n = \widetilde{O}\left(\min\{\sqrt{k}/\tau, \sqrt{\log(|\mathcal{X}|)}/\tau\}\right)$ samples are necessary for τ accuracy for k adaptively chosen statistical queries.

Although in this section we focused entirely on statistical queries, it is possible to obtain similar results for much richer classes of queries an analyst would like to ask about a dataset. Bassily et al. (2016) give results for the class of low sensitivity queries $q : \mathcal{X}^n \to \mathbb{R}$, defined as functions where for any two *neighboring* datasets $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}^n$, that is $\boldsymbol{x}$ and $\boldsymbol{x}'$ are the same in every entry except one element, we have

$$|q(\boldsymbol{x}) - q(\boldsymbol{x}')| \leq \Delta(q) = o(1) \qquad \text{as } n \to \infty.$$

We call $\Delta(q)$ the *sensitivity* of function q. For the particular results in Bassily et al. (2016) that we cite, we require $\Delta(q) = O(1/n)$, although their results do hold for more general sensitivities.

1.3. Post Selection Hypothesis Testing

The goal of this dissertation is to handle much more general types of analyses, rather than just statistical queries or low-sensitivity queries, in this adaptive setting. One specific type of analysis we might like to handle adaptively is hypothesis testing. In fact, the previous works (Dwork et al., 2015c; Bassily et al., 2016) are motivated by the problem of false discovery in empirical science despite the technical results being about estimating means of adaptively chosen statistical (or low-sensitivity) queries.

We will consider a simple model of one-sided hypothesis tests on real valued test statistics. A hypothesis test is defined by a *test statistic* $\phi^{(j)} : \mathcal{X}^n \to \mathbb{R}$ mapping datasets to a real value, where we use j to index different test statistics. Given an output $a = \phi^{(j)}(\boldsymbol{x})$, together with a distribution $\mathcal{D}$ over the data domain, the p-value associated with a and $\mathcal{D}$ is simply the probability of observing a value of the test statistic that is at least as extreme as a, assuming the data was drawn independently from $\mathcal{D}$: $p^{(j)}_{\mathcal{D}}(a) \stackrel{\text{defn}}{=} \Pr_{\boldsymbol{X}\sim\mathcal{D}^n}[\phi^{(j)}(\boldsymbol{X}) \geq a]$. Note that there may be multiple distributions $\mathcal{D}$ over the data that induce the same distribution over the test statistic. With each test statistic $\phi^{(j)}$, we associate a *null hypothesis* $\mathrm{H}_0^{(j)}$ as a collection of possible distributions over $\mathcal{X}$. The p-values are always computed with respect to a distribution $\mathcal{D} \in \mathrm{H}_0^{(j)}$, and hence from now on, we hide the dependence on $\mathcal{D}$ and simply write $p^{(j)}(a)$ to denote the p-value of a test statistic $\phi^{(j)}$ evaluated at a.

---

[3] The existence of one-way functions

The goal of a hypothesis test is to *reject the null hypothesis* if the data is not likely to have been generated from the proposed model, that is if the underlying distribution from which the data were drawn was not in $\mathrm{H}_0^{(j)}$. By definition, if $\boldsymbol{X}$ truly is drawn from $\mathcal{D}^n$ for some $\mathcal{D} \in \mathrm{H}_0^{(j)}$, then $p^{(j)}(\phi^{(j)}(\boldsymbol{X}))$ is uniformly distributed over $[0,1]$. A standard approach to hypothesis testing is to pick a *significance level* $\alpha \in [0,1]$ (often $\alpha = 0.05$), compute the value of the test statistic $a = \phi^{(j)}(\boldsymbol{X})$, and then *reject* the null hypothesis if $p^{(j)}(a) \leq \alpha$. Under this procedure, the probability of incorrectly rejecting the null hypothesis—i.e., of rejecting the null hypothesis when $\boldsymbol{X} \sim \mathcal{D}^n$ for some $\mathcal{D} \in \mathrm{H}_0^{(j)}$—is at most α. Note that an incorrect rejection of the null hypothesis is called a *false discovery.*

The discussion so far presupposes that $\phi^{(j)}$, the test statistic in question, was chosen independently of the dataset $\boldsymbol{X}$. Let $\mathcal{Y}$ denote a collection of test statistics, and suppose that we select a test statistic using a data-dependent selection procedure $\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}$. If $\phi^{(j)} = \mathcal{M}(\boldsymbol{X})$, then rejecting the null hypothesis when $p^{(j)}(\phi^{(j)}(\boldsymbol{X})) \leq \alpha$ may result in a false discovery with probability much larger than α. As we mentioned earlier, this kind of naïve approach to *post-selection* inference is suspected to be a primary culprit behind the prevalence of false discovery in empirical science (Gelman and Loken, 2014; Wasserstein and Lazar, 2016; Simmons et al., 2011). This is because even if the null hypothesis is true ($\boldsymbol{X} \sim \mathcal{D}^n$ for some $\mathcal{D} \in \mathrm{H}_0^{(j)}$), the distribution on $\boldsymbol{X}$ *conditioned on* $\phi^{(j)} = \mathcal{M}(\boldsymbol{X})$ *having been selected* need not be $\mathcal{D}^n$. Our goal in studying valid post-selection hypothesis testing is to then find a *valid* p-value correction function $\gamma : [0,1] \to [0,1]$, which we define as follows:
Definition 1.3.1 [Valid p-value Correction Function]**.** *A function* $\gamma : [0,1] \to [0,1]$ *is a* valid p-value correction function *for a selection procedure* $\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}$ *if for every significance level* $\alpha \in [0,1]$*, the procedure:*

1. *Select a test statistic* $\phi^{(j)} = \mathcal{M}(\boldsymbol{X})$ *using selection procedure* $\mathcal{M}$*.*
2. *Reject the null hypothesis* $\mathrm{H}_0^{(j)}$ *if* $p^{(j)}(\phi^{(j)}(\boldsymbol{X})) \leq \gamma(\alpha)$*.*

has probability at most α *of resulting in a false discovery.*

We will be interested in p-value corrections that are not too small – note that $\gamma(\alpha) = 0$ is a valid correction but not very interesting. We would like our tests to be able to correctly reject a wrong $\mathrm{H}_0^{(j)}$ with higher confidence as we increase the sample size. The ability for a hypothesis test to correct reject a null hypothesis is called the *power* of the test. We then model the various sources of error in hypothesis testing in Table 1 where $\mathrm{H}_1^{(j)}$ is some fixed alternate hypothesis, different from the null. Typically in hypothesis testing, we want to ensure the probability of a false discovery is at most some threshold α, and we would like to minimize the probability of *type II error*, i.e. failing to reject when the null hypothesis was false.

	$\mathrm{H}_0^{(j)}$ **True**	$\mathrm{H}_1^{(j)}$ **True**
Reject $\mathrm{H}_0^{(j)}$	False Discovery	Power
Fail to Reject $\mathrm{H}_0^{(j)}$	Significance	Type II Error

Table 1: Types of Errors in Hypothesis Testing

Necessarily, to give a nontrivial correction function γ, we will need to assume that the selection procedure $\mathcal{M}$ satisfies some useful property. We will discuss later the types of test selection procedures that will enable us to find valid correction functions. In Appendix A.1, we show that hypothesis testing is in general beyond the setting of statistical or low-sensitivity queries that we have already discussed above, which shows that we need new tools for handling adaptive hypothesis testing.

It is important to point out the role of the algorithm $\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}$ here. Before, we were considering an algorithm that was given access to a dataset and would release answers to adaptively chosen queries from the data analyst, then the analyst $\mathcal{A}$ would select any new analysis (or query) based on the answers she had already witnessed. However, now we are considering algorithms, or *test selection procedures* that release the analysis for the analyst to use. This is simply for mathematical convenience. The seeming discrepancy between these two models is resolved by the guarantees of the algorithms we will consider here – that they are closed under post-processing. Thus, we can combine the output of the algorithm $\mathcal{M}$ and the choice of analysis of $\mathcal{A}$ as a single procedure $\mathcal{M}$.

1.4. Handling More General Analyses - Max-information

There is one constraint on the selection procedure $\mathcal{M}$ that does allow us to give nontrivial p-value corrections—that $\mathcal{M}$ has bounded max-information. Max-information is a measure introduced by Dwork et al. (2015a), which we discuss next.

Given two (arbitrarily correlated) random variables X, Z, we let $X \otimes Z$ denote a random variable (in a different probability space) obtained by drawing independent copies of X and Z from their respective marginal distributions.
Definition 1.4.1 [Max-Information (Dwork et al., 2015a)]**.** *Let X and Z be jointly distributed random variables over the domain $(\mathcal{X}, \mathcal{Z})$. The max-information between X and Z, denoted by $I_\infty(X; Z)$, is the minimal value of m such that for every x in the support of X and z in the support of Z, we have* $\Pr[X = x | Z = z] \le 2^m \Pr[X = x]$*. Alternatively,*

$$I_\infty(X; Z) = \log_2 \sup_{(x,z)\in(\mathcal{X},\mathcal{Z})} \frac{\Pr[(X, Z) = (x, z)]}{\Pr[X \otimes Z = (x, z)]}.$$

The β-approximate max-information between X and Z is defined as

$$I_\infty^\beta(X; Z) = \log_2 \sup_{\substack{\mathcal{O}\subseteq(\mathcal{X}\times\mathcal{Z}),\\ \Pr[(X,Z)\in\mathcal{O}]>\beta}} \frac{\Pr[(X, Z) \in \mathcal{O}] - \beta}{\Pr[X \otimes Z \in \mathcal{O}]}.$$

We say that an algorithm $\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}$ *has* β*-approximate max-information of* m*, denoted as* $I_\infty^\beta(\mathcal{M}, n) \le m$*, if for every distribution* $\mathcal{S}$ *over elements of* $\mathcal{X}^n$*, we have* $I_\infty^\beta(\boldsymbol{X}; \mathcal{M}(\boldsymbol{X})) \le m$ *when* $\boldsymbol{X} \sim \mathcal{S}$*. We say that an algorithm* $\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}$ *has* β*-approximate max-information of* m *over* product distributions*, written* $I_{\infty,\Pi}^\beta(\mathcal{M}, n) \le m$*, if for every distribution* $\mathcal{D}$ *over* $\mathcal{X}$*, we have* $I_\infty^\beta(\boldsymbol{X}; \mathcal{M}(\boldsymbol{X})) \le m$ *when* $\boldsymbol{X} \sim \mathcal{D}^n$*.*

Max-information has several nice properties that are useful in adaptive data analysis. The first is that it composes, so that if an analyst uses an algorithm with bounded approximate max-information and then based on the output uses another algorithm with bounded approximate max-information, then the resulting analysis still has bounded approximate max-information.

Theorem 1.4.2 [Dwork et al. (2015a)]**.** *Let* $\mathcal{M}_1 : \mathcal{X}^n \to \mathcal{Y}$ *be an algorithm such that* $I_\infty^{\beta_1}(\mathcal{M}_1, n) \le m_1$ *and let* $\mathcal{M}_2 : \mathcal{X}^n \times \mathcal{Y} \to \mathcal{Z}$ *be an algorithm such that for each* $y \in \mathcal{Y}$*, we have* $I_\infty^{\beta_2}(\mathcal{M}_2(\cdot, y), n)$*. Then the composed algorithm* $\mathcal{M} : \mathcal{X}^n \to \mathcal{Z}$ *where* $\mathcal{M}(\boldsymbol{x}) = \mathcal{M}_2(\boldsymbol{x}, \mathcal{M}_1(\boldsymbol{x}))$ *has* $I_\infty^{\beta_1+\beta_2}(\mathcal{M}, n) \le m_1 + m_2$*. Further, if we also have* $I_{\infty,\Pi}^{\beta_1}(\mathcal{M}_1, n) \le m_1$*, then* $I_{\infty,\Pi}^{\beta_1+\beta_2}(\mathcal{M}, n) \le m_1 + m_2$

Note that this result can be iteratively applied to string together a sequence of adaptively chosen algorithms with approximate max-information and the resulting composed algorithm will also have bounded max-information. This composition theorem is crucial in controlling the probability of false discovery over a sequence of analyses when each analysis is individually known to have bounded max-information.

Another useful property is that max-information is preserved under post-processing. Thus, if our algorithm $\mathcal{M}$ answers adaptively chosen analyses (e.g. statistical queries) on dataset $\boldsymbol{X}$ and has bounded approximate max-information, then the analyst can take any function of the output $f(\mathcal{M}(\boldsymbol{X})) = \mathcal{M}'(\boldsymbol{X})$ to find a new analysis to run. The resulting algorithm $\mathcal{M}'$ then has max-information bound no larger than that of $\mathcal{M}$.

Theorem 1.4.3 [Dwork et al. (2015a)]**.** *If* $\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}$ *and* $\psi : \mathcal{Y} \to \mathcal{Y}$ *is any (possibly randomized) mapping, then* $\psi \circ \mathcal{M} : \mathcal{X}^n \to \mathcal{Y}'$ *satisfies the following for any random variable* $\boldsymbol{X}$ *over* $\mathcal{X}^n$ *and every* $\beta \ge 0$*,*

$$I_\infty^\beta(\boldsymbol{X}; \psi(\mathcal{M}(\boldsymbol{X}))) \le I_\infty^\beta(\boldsymbol{X}; \mathcal{M}(\boldsymbol{X})).$$

We now state some of the immediate consequences of max-information in adaptive data analysis. It follows from the definition that if an algorithm has bounded max-information, then we can control the probability of "bad events" that arise as a result of the dependence of $\mathcal{M}(\boldsymbol{X})$ on $\boldsymbol{X}$: for every event $\mathcal{O}$, we have $\Pr[(\boldsymbol{X}, \mathcal{M}(\boldsymbol{X})) \in \mathcal{O}] \le 2^m \Pr[\boldsymbol{X} \otimes \mathcal{M}(\boldsymbol{X}) \in \mathcal{O}] + \beta$.

For example, if $\mathcal{M}$ is a data-dependent selection procedure for selecting a test statistic, we can derive a valid p-value correction function γ as a function of a max-information bound on $\mathcal{M}$:

Theorem 1.4.4. *Let* $\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}$ *be a data-dependent algorithm for selecting a test statistic such that* $I_{\infty,\Pi}^\beta(\mathcal{M}, n) \le m$*. Then the following function* γ *is a valid p-value* *correction function for $\mathcal{M}$:*

$$\gamma(\alpha) = \max\left(\frac{\alpha-\beta}{2^m}, 0\right).$$

Proof. Fix a distribution $\mathcal{D}$ from which the dataset $\boldsymbol{X} \sim \mathcal{D}^n$. If $\frac{\alpha-\beta}{2^m} \leq 0$, then the theorem is trivial, so assume otherwise. Define $\mathcal{O} \subset \mathcal{X}^n \times \mathcal{Y}$ to be the event that $\mathcal{M}$ selects a test statistic for which the null hypothesis is true, but its p-value is at most $\gamma(\alpha)$:

$$\mathcal{O} = \{(\boldsymbol{x}, \phi^{(j)}) : \mathcal{D} \in \mathrm{H}_0^{(j)} \text{ and } p^{(j)}(\phi^{(j)}(\boldsymbol{x})) \leq \gamma(\alpha)\}$$

Note that the event $\mathcal{O}$ represents exactly those outcomes for which using γ as a p-value correction function results in a false discovery. Note also that, by definition of the null hypothesis, $\Pr[\boldsymbol{X} \otimes \mathcal{M}(\boldsymbol{X}) \in \mathcal{O}] \leq \gamma(\alpha) = \frac{\alpha-\beta}{2^m}$. Hence, by the guarantee that $I^{\beta}_{\infty,\Pi}(\mathcal{M}, n) \leq m$, we have that $\Pr[(\boldsymbol{X}, \mathcal{M}(\boldsymbol{X}) \in \mathcal{O})]$ is at most $2^m \cdot \left(\frac{\alpha-\beta}{2^m}\right) + \beta = \alpha$. $\square$

We can also use algorithms with small max-information to answer adaptively chosen low-sensitivity functions, using McDiarmid's inequality (given in Theorem A.1.1 in the appendix).

Theorem 1.4.5 [Dwork et al. (2015a)]**.** *Let $\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}$ be a data-dependent algorithm for selecting a function with sensitivity Δ and $I^{\beta}_{\infty,\Pi}(\mathcal{M}, n) \leq \log_2(e)\left(\tau^2/\Delta^2\right)$, then we have for $q = \mathcal{M}(\boldsymbol{X})$ where $\boldsymbol{X} \sim \mathcal{D}^n$*

$$\Pr_{\boldsymbol{X},\mathcal{M}}\left[|q(\mathcal{D}^n) - q(\boldsymbol{X})| \geq \tau\right] \leq \exp\left(\frac{-\tau^2}{n\Delta^2}\right) + \beta.$$

Max-information provides the correction factor in which we need to modify our analyses for the dependence on the data. Up to the correction factor, we can then use existing statistical theory as if the analysis were chosen independently of the data.

1.5. Algorithms with Bounded Max-information

Due to Theorems 1.4.4 and 1.4.5, we are then interested in finding test selection procedures $\mathcal{M}$ that have bounded approximate max-information. From Dwork et al. (2015a), there are two families of algorithms which are known to have bounded max-information, which we will discuss in turn. Note that these algorithms were known previously to give good generalization guarantees for adaptively chosen analyses, but the two are otherwise incomparable. Thus, max-information can be seen as a unifying measure for different types of analyses which have good generalization guarantees.

We first state the result that gives us a max-information bound in terms of the description length of the output of $\mathcal{M}$.

Theorem 1.5.1 [Dwork et al. (2015a)]**.** *Let $\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}$ be a randomized algorithm with finite output set $\mathcal{Y}$. Then for each $\beta > 0$, we have*

$$I^{\beta}_{\infty}(\mathcal{M}, n) \leq \log_2\left(|\mathcal{Y}|/\beta\right).$$

We can interpret this result as saying that the shorter the result of an analysis, the better it is for adaptive data analysis settings because it reveals less information about the dataset and leads to better generalization for subsequent analyses.

The second type of algorithms that were known to have bounded max-information are (pure) differentially private algorithms. At a high level, differential privacy is a stability guarantee on an algorithm in that it limits the sensitivity of the outcome to any individual data entry from an input dataset. Although differential privacy was introduced for private data analysis applications (Dwork et al., 2006b), where the dataset is assumed to contain sensitive information about its subjects, we can leverage the stability guarantees of differential privacy to answer new questions in various problems beyond privacy concerns. In fact, differential privacy has proven to be a powerful algorithmic property in game theoretical problems in economics, (Kearns et al., 2015; Rogers et al., 2015; Lykouris et al., 2016; Cummings et al., 2016b, 2015; Kannan et al., 2015). Similarly, differential privacy has been shown to be a useful tool in adaptive data analysis (Dwork et al., 2015c,a; Bassily et al., 2016), which is the connection we explore in this dissertation.

We will give the definition and useful properties of differential privacy in Chapter 2, but we state here the bound on max-information.

Theorem 1.5.2 [Pure Differential Privacy and Max-Information (Dwork et al., 2015a)]. *Let $\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}$ be an $(\epsilon, 0)$-differentially private algorithm. Then for every $\beta > 0$:*

$$I_\infty\left(\mathcal{M}, n\right) \leq \log_2(e) \cdot \epsilon n \quad \text{ and } \quad I_{\infty,\Pi}\left(\mathcal{M}, n\right) \leq \log_2(e) \cdot \left(\epsilon^2 n/2 + \epsilon\sqrt{n \ln(2/\beta)/2}\right)$$

Due to the composition property of max-information in Theorem 1.4.2, we can string pure differentially private algorithms together with bounded description length algorithms in arbitrary orders and still obtain generalization guarantees for the entire sequence.

The connection in Theorem 1.5.2 is powerful, because there are a vast collection of data analyses for which we have differentially private algorithms, including a growing literature – some which we will cover in this dissertation – on differentially private hypothesis tests (Johnson and Shmatikov, 2013; Uhler et al., 2013; Yu et al., 2014; Karwa and Slavković, 2016; Dwork et al., 2015d; Sheffet, 2015b; Wang et al., 2015; Gaboardi et al., 2016; Kifer and Rogers, 2016). However, there is an important gap: Theorem 1.5.2 holds only for *pure* $(\epsilon, 0)$-differential privacy, and not for the broader class, (approximate) (ϵ, δ)-differential privacy, where $\delta > 0$. Many statistical analyses can be performed much more accurately subject to approximate differential privacy, and it can be easier to analyze private hypothesis tests that satisfy approximate differential privacy, because the approximate privacy constraint is amenable to perturbations using Gaussian noise (rather than Laplace noise) (Gaboardi et al., 2016; Kifer and Rogers, 2016). Most importantly, for pure differential privacy, the privacy parameter ϵ degrades *linearly* with the number of analyses performed, whereas for approximate differential privacy, ϵ need only degrade with the *square root* of the number of analyses performed (Dwork et al., 2010). Hence, if the connection between max-information and differential privacy held also for approximate differential privacy, it would be possible to perform quadratically more adaptively chosen statistical tests without requiring a smaller

p-value correction factor. In fact, for the sample complexity results given in Theorems 1.2.1 and 1.2.2, in order to answer k adaptively chosen statistical queries accurately, we require that the overall composed algorithm $\mathcal{M}_k \circ \cdots \circ \mathcal{M}_1$ be approximately differentially private.

1.6. Contributions

We are now ready to discuss the specific contributions of this dissertation in understanding adaptive data analysis.

We will first demonstrate how we can use the previous results in adaptive data analysis for obtaining valid confidence intervals on adaptively chosen statistical queries using results from Dwork et al. (2015c), Bassily et al. (2016), and Russo and Zou (2016). Although, we know that we can asymptotically outperform data-splitting techniques, we show in Chapter 3 that we can improve even for reasonably sized datasets.

In Chapter 4 we extend the connection between differential privacy and max-information to approximate differential privacy, which follows from work published by Rogers et al. (2016a). We show the following (see Section 4.2 for a complete statement):

Theorem 4.2.1 (Informal). *Let $\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}$ be an (ϵ, δ)-differentially private algorithm. Then,*

$$I^{\beta}_{\infty,\Pi}(\mathcal{M}, n) = \widetilde{O}\left(n\epsilon^2 + n\sqrt{\tfrac{\delta}{\epsilon}}\right) \qquad \text{for } \beta = \widetilde{O}\left(n\sqrt{\tfrac{\delta}{\epsilon}}\right).$$

It is worth noting several things. First, this bound nearly matches the bound for max-information over product distributions from Theorem 1.5.2, except Theorem 4.2.1 extends the connection to the substantially more powerful class of (ϵ, δ)-differentially private algorithms. The bound is qualitatively tight in the sense that despite its generality, it can be used to nearly recover the tight bound on the generalization properties of differentially private mechanisms for answering low-sensitivity queries that was proven using a specialized analysis in Bassily et al. (2016), see Section 4.3 for a comparison.

We also only prove a bound on the max-information for product distributions on the input, and not for all distributions (that is, we bound $I^{\beta}_{\infty,\Pi}(\mathcal{M}, n)$ and not $I^{\beta}_{\infty}(\mathcal{M}, n)$). A bound for general distributions would be desirable, since such bounds compose, see Theorem 1.4.2. Unfortunately, a bound for general distributions based solely on (ϵ, δ)-differential privacy is impossible: a construction inspired by work from De (2012) implies the existence of (ϵ, δ)-differentially private algorithms for which the max-information between input and output on arbitrary distributions is much larger than the bound in Theorem 4.2.1.

One might nevertheless hope that bounds on the max-information under product distributions can be meaningfully composed. Our second main contribution is a negative result, showing that such bounds do not compose when algorithms are selected adaptively. Specifically, we analyze the adaptive composition of two algorithms, the first of which has a small finite range (and hence, by Dwork et al. (2015a), small bounded max-information), and the second of which is (ϵ, δ)-differentially private. We show that the composition of the two algorithms can be used to exactly recover the input dataset, and hence, the composition does not satisfy any nontrivial max-information bound. We then draw a connection between max-information and Shannon mutual information that allows us to improve on several prior results that dealt with mutual information in McGregor et al. (2011) and Russo and Zou (2016).

An important feature of differential privacy is that it is preserved under composition, that is combining many differentially private subroutines into a single algorithm preserves differential privacy and the privacy parameters degrade gracefully. However, there is a caveat to these composition theorems. In order to apply these results from differential privacy, we need the privacy parameters to all be fixed up front, prior to running any analysis on the data. In Chapter 5, we then give a framework for composition that allows for the types of composition theorems of differential privacy to work in this adaptive setting, which follows from work by Rogers et al. (2016b). We give a formal separation between the standard model of composition and our new setting, so that we cannot simply plug in the adaptively chosen privacy parameters into existing differential privacy composition theorems as if the *realized* parameters were fixed prior to running any analysis. Despite this result, we still give a bound on the privacy loss when the parameters can be adaptively selected which for datasets of size n is asymptotically no larger than $\sqrt{\log \log n}$ times the bound from the advanced composition theorem of Dwork et al. (2010), which assumes that the privacy parameters were selected beforehand.

After understanding the benefits of differential privacy in adaptive data analysis and how composition may be applied in this setting, we then present in Chapters 6 and 7 some primitives that an analyst may want to use at each round of interaction with the dataset. Specifically we give hypothesis tests that ensure statistical validity while satisfying differential privacy, focusing on categorical data and chi-square tests, such as tests for independence and goodness of fit. Chapter 6 follows from work by Gaboardi et al. (2016) and Chapter 7 is from Kifer and Rogers (2016).

1.6.1. New Results

For the reader that is interested in results that are not publised elsewhere, we outline the new contributions of this dissertation here:

- Chapter 3, which demonstrates the improvements we can obtain over data-splitting techniques for computing confidence intervals on adaptively chosen statistical queries, is new and based on ongoing work with Aaron Roth, Adam Smith, and Om Thakkar.
- We give a new implication of our lower bound result from Rogers et al. (2016a) in Section 4.5, which shows that *robustly generalizing* procedures, introduced by Cummings et al. (2016a) do not in general compose.
- We show in Section 4.7 that procedures with bounded max-information are not necessary to ensure generalization guarantees in the adaptive setting. Specifically, we show that compression schemes can have arbitrarily large max-information.
- We use a different concentration bound (Theorem 5.4.2) from the one used in Rogers

et al. (2016b) to obtain a privacy odometer with better constants than what appeared in Rogers et al. (2016b), which then follows from a more simplified analysis, presented in Theorem 5.6.5

- Section 5.7 extends privacy odometers and filters to include concentrated differentially private algorithms, which were defined by Bun and Steinke (2016).
- All of the experiments in Part III have been redone so that it is easier to directly compare empirical results of the different private hypothesis tests we propose.
- We give the variance of each of the chi-square statistics we consider in Theorem 7.2.12. This gives a more analytical reason for why some tests achieve better empirical power than others.
- We also give preliminary results on private hypothesis tests in the *local model* in Chapter 8, where each individual's data is reported in a private way, rather than in the traditional *curator* setting which assumes there is a trusted curator that collects everyone's raw data.

1.7. Related Work

This dissertation follows a line of work that was initiated by Dwork et al. (2015c) and Hardt and Ullman (2014), who formally modeled the problem in adaptive data analysis. Since these works, there has been several other contributions (Dwork et al., 2015a,b; Bassily et al., 2016; Russo and Zou, 2016; Cummings et al., 2016a; Wang et al., 2016), some in which we have already discussed many of their results. We then use this section to discuss some of the relevant work in this area that we have not already addressed.

One of the crucial observations of Dwork et al. (2015c) is that algorithms that are stable, i.e. *differentially private*, can be leveraged to obtain strong generalization guarantees in adaptive analysis. Stability measures the amount of change in the output of an algorithm if the input is perturbed. Another line of work (Bousquet and Elisseeff, 2002; Mukherjee et al., 2006; Poggio et al., 2004; Shalev-Shwartz et al., 2010) had established the connections between stability of a learning algorithm and its ability to generalize, although in nonadaptive settings. The problem with the stability notions that they consider is that they are *not* robust to post-processing or adaptive composition. That is, if individual algorithms are stable and known to generalize, then it is often not the case that stringing together a sequence of these algorithms will still ensure generalization. The main benefit of the type of stability that Dwork et al. (2015c) considers is that it is preserved under the operations of post-processing and adaptive composition.

Russo and Zou (2016) consider different types of exploratory analyses through the lens of information usage, similar to Dwork et al. (2015a). They study the bias that can result in adaptively chosen analyses, which are based on *subgaussian* statistics under the data distribution. They prove that the *bias* can be bounded by the dependence between the noise in the data and the choice of reported result using Shannon mutual information. To obtain the type of generalization bounds that we are concerned with – high probability guarantees – we can then apply Markov or Chebyshev's inequality. We compare some of our results with those of Russo and Zou (2016) in Section 4.6.

In a more restrictive setting, Wang et al. (2016) assumes that the analyst is selecting statistics which are jointly Gaussian. They show that adding Gaussian noise to the statistics is optimal in a *minimax* framework of adaptive data analysis. However, we are after generality of analyses, which comes at the cost that our results might be overly conservative.

In order to perform many adaptively chosen analyses and ensure good generalization over the entire sequence, we want each analysis to be robust to post-processing and adaptive composition. This is one of the main advantages with differential privacy, because these properties are known to hold already. Cummings et al. (2016a) then give three notions of generalizations that are closed under post-processing and amendable to adaptive composition, with each strictly stronger than the next: robust generalization, differential privacy, and perfect generalization. Although bounded description length and differentially private algorithms were known to be robustly generalizing, they demonstrate a third type – *compression schemes* – that also gives guarantees of robust generalization.

There have also been successful implementations of algorithms that guard against overfitting in adaptive data analysis. Specifically, Blum and Hardt (2015) gives a natural algorithm – *the Ladder* – to ensure that a leaderboard is accurate in machine learning competitions even when entries are allowed to evaluate their models several times on a holdout set, each time making modifications based on how they may rank on the leaderboard. They show that they can ensure a leaderboard accurately ranks the participants' models despite the adaptivity in real submission files and even give empirical results using real data from the Kaggle competition. Additionally, Dwork et al. (2015a,b) show how their methods can reduce overfitting to a holdout set, using the algorithm they call *Thresholdout*, when variables are selected for a model and then evaluated using the same holdout set.

CHAPTER 2

PRIVACY PRELIMINARIES

We use this section to present what has become the standard privacy benchmark, *differential privacy*, and then the more recent version called *concentrated differential privacy.* We then give some definitions and results that will be crucial for the rest of the dissertation.

We will define differential privacy in terms of indistinguishability, which measures the similarity between two random variables.

Definition 2.0.1 [Indistinguishability (Kasiviswanathan and Smith, 2014)]**.** *Two random variables X, Y taking values in a set $\mathcal{X}$ are (ϵ, δ)-indistinguishable, denoted $X \approx_{\epsilon,\delta} Y$, if for all $S \subseteq \mathcal{X}$,*

$$\Pr\left[X \in S\right] \leq e^{\epsilon} \cdot \Pr\left[Y \in S\right] + \delta \qquad \text{and} \qquad \Pr\left[Y \in S\right] \leq e^{\epsilon} \cdot \Pr\left[X \in S\right] + \delta.$$

2.1. Differential Privacy

Recall that when we defined *sensitivity* of a function, we say that two datasets $\boldsymbol{x} = (x_1, \cdots, x_n), \boldsymbol{x}' = (x'_1, \cdots, x'_n) \in \mathcal{X}^n$ are *neighboring* if they differ in at most one entry, i.e. there is some $i \in [n]$ where $x_i \neq x'_i$, but $x_j = x'_j$ for all $j \neq i$.

Definition 2.1.1 [Differential Privacy (Dwork et al., 2006b,a)]**.** *A randomized algorithm (or mechanism) $\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}$ is (ϵ, δ)-differentially private (DP) if for all neighboring datasets $\boldsymbol{x}$ and $\boldsymbol{x}'$ and each outcome $S \subseteq \mathcal{Y}$, we have $\mathcal{M}(\boldsymbol{x}) \approx_{\epsilon,\delta} \mathcal{M}(\boldsymbol{x}')$ or equivalently*

$$\Pr\left[\mathcal{M}(\boldsymbol{x}) \in S\right] \leq e^{\epsilon} \Pr\left[\mathcal{M}(\boldsymbol{x}') \in S\right] + \delta.$$

If $\delta = 0$, we simply say $\mathcal{M}$ is ϵ-DP or pure DP. Otherwise for $\delta > 0$, we say approximate *DP.*

Note that in the definition, the data is not assumed to be coming from a particular distribution. Rather, the probability in the definition statement is *only* over the randomness from the algorithm. In order to compute some statistic $f : \mathcal{X}^n \to \mathbb{R}^d$ on the data, a differentially private algorithm is to simply add symmetric noise to $f(\boldsymbol{x})$ with standard deviation that depends on the *global sensitivity* of f, which we define as

$$\Delta_p(f) = \max_{\text{neighboring } \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}^n} \{||f(\boldsymbol{x}) - f(\boldsymbol{x}')||_p\}. \tag{2.1}$$

We then give a commonly used differentially private algorithm, called the *Laplace Mechanism*, which releases an answer to a query on the dataset with appropriately scaled Laplace noise.

Theorem 2.1.2 [Laplace Mechanism (Dwork et al., 2006b)]**.** *Let $f : \mathcal{X}^n \to \mathbb{R}^d$. The algorithm $\mathcal{M}(\boldsymbol{x}) = f(\boldsymbol{x}) + \boldsymbol{L}$ where $\boldsymbol{L} \overset{i.i.d.}{\sim} \mathrm{Lap}(\Delta_1(f)/\epsilon)$, is ϵ-DP.*

A very useful fact about differentially private algorithms is that one cannot take the output of a differentially private mechanism and perform any modification to it that does not depend on the input itself and make the output any less private. This is precisely the same property that max-information enjoys in Theorem 1.4.3.

Theorem 2.1.3 [Post Processing (Dwork et al., 2006b)]**.** *Let $\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}$ be (ϵ, δ)-DP and $\psi : \mathcal{Y} \to \mathcal{Y}'$ be any function mapping to arbitrary domain $\mathcal{Y}'$. Then $\psi \circ \mathcal{M}$ is (ϵ, δ)-DP.*

One of the strongest properties of differential privacy is that it is preserved under *adaptive composition.* That is, combining many differentially private subroutines into a single algorithm preserves differential privacy and the privacy parameters degrade gracefully. We will discuss in Chapter 5 a caveat to these composition theorems and propose a new framework of composition, where the privacy parameters themselves may also be chosen adaptively.

Adaptive composition of algorithms models the way in which an analyst would interact with the same dataset multiple times, so that each algorithm may depend on the previous outcomes. We formalize *adaptive composition* in the following way:

- $\mathcal{M}_1 : \mathcal{X}^n \to \mathcal{Y}_1$.
- For each $i \in [k]$, $\mathcal{M}_i : \mathcal{X}^n \times \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_{i-1} \to \mathcal{Y}_i$.

We will then denote the entire composed algorithm as $\mathcal{M}_{1:k} : \mathcal{X}^n \to \mathcal{Y}_k$.

We first state a basic composition theorem which shows that the adaptive composition satisfies differential privacy where "the parameters just add up."

Theorem 2.1.4 [Basic Composition (Dwork et al., 2006b,a)]**.** *Let each $\mathcal{M}_i : \mathcal{X}^n \times \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_{i-1}$ be (ϵ_i, δ_i)-DP in its first argument for $i \in [k]$. Then $\mathcal{M}_{1:k} : \mathcal{X}^n \to \mathcal{Y}_k$ is (ϵ_g, δ_g)-differential privacy where*

$$\epsilon_g = \sum_{i=1}^k \epsilon_i, \qquad \textit{and} \qquad \delta_g = \sum_{i=1}^k \delta_i.$$

We now state the advanced composition bound originally given in Dwork et al. (2010) which gives a quadratic improvement to the basic composition bound. We state the bound as given by Kairouz et al. (2015), with improved constants and generalized it so that all the privacy parameters need not be the same.

Theorem 2.1.5 [Advanced Composition (Dwork et al., 2010; Kairouz et al., 2015)]**.** *Let each $\mathcal{M}_i : \mathcal{X}^n \times \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_{i-1}$ be (ϵ_i, δ_i)-DP in its first argument for $i \in [k]$. Then $\mathcal{M}_{1:k} : \mathcal{X}^n \to \mathcal{Y}_k$ is (ϵ_g, δ_g)-differential privacy where for any $\widehat{\delta} > 0$*

$$\epsilon_g = \sum_{i=1}^k \epsilon_i \left(\frac{e^{\epsilon_i} - 1}{e^{\epsilon_i} + 1} \right) + \sqrt{2 \sum_{i=1}^k \epsilon_i^2 \log(1/\widehat{\delta})}, \qquad \textit{and} \qquad \delta_g = 1 - (1 - \widehat{\delta}) \prod_{i=1}^k (1 - \delta_i).$$

There has also been work (Kairouz et al., 2015; Murtagh and Vadhan, 2016) on obtaining optimal composition bounds for differentially private algorithms, so as a function of the privacy parameters $(\epsilon_1, \delta_1), \cdots, (\epsilon_k, \delta_k)$ and $\delta_g > 0$, find the best possible privacy parameter ϵ_g, so that any adaptive composition of $\mathcal{M}_{1:k}$, where each algorithm is (ϵ_i, δ_i)-DP for $i \in [k]$ is (ϵ_g, δ_g)-DP.

We then define the privacy loss random variable, which quantifies how much the output distributions of an algorithm on two neighboring datasets can differ.

Definition 2.1.6 [Privacy Loss]**.** *Let* $\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}$ *be a randomized algorithm. We then define the privacy loss variable* $\mathrm{PrivLoss}\left(\mathcal{M}(\boldsymbol{x}) || \mathcal{M}(\boldsymbol{x}')\right)$ *for neighboring datasets* $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}^n$ *in the following way: let* $Z(y) = \log\left(\frac{\Pr[\mathcal{M}(\boldsymbol{x})=y]}{\Pr[\mathcal{M}(\boldsymbol{x}')=y]}\right)$ *and then* $\mathrm{PrivLoss}\left(\mathcal{M}(\boldsymbol{x}) || \mathcal{M}(\boldsymbol{x}')\right)$ *is distributed the same as* $Z(\mathcal{M}(\boldsymbol{x}))$*.*

Note that if we can bound the privacy loss random variable with certainty over all neighboring datasets, then the algorithm is pure DP. Otherwise, if we can bound the privacy loss with high probability then it is approximate DP (see Kasiviswanathan and Smith (2014) for a more detailed discussion on this connection). It is worth pointing out that the privacy loss random variable is central to proving Theorem 2.1.5 from Dwork et al. (2010).

2.2. Concentrated Differential Privacy

We will also use in our results a recently proposed definition of privacy called *zero concentrated differential privacy* (zCDP), defined by Bun and Steinke (2016) (Note that Dwork and Rothblum (2016) initially gave a definition of concentrated differential privacy which Bun and Steinke (2016) then modified).

Definition 2.2.1 [zCDP]**.** *An algorithm* $\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}$ *is* (ξ, ρ)*-zero concentrated differentially private (zCDP), if for all neighboring datasets* $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}^n$ *and all* $\lambda > 0$ *we have*

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

Typically, we will write $(0, \rho) - zCDP$ *simply as* ρ*-zCDP.*

Note that the definition of zCDP implies that the privacy loss random variable is *subgaussian.*

Similar to the Laplace mechanism, we can add appropriately scaled Gaussian noise to a particular query to ensure zCDP. Note that the following *Gaussian mechanism* was introduced prior to zCDP and was shown to be approximate differentially private (Dwork et al., 2006b; Nikolov et al., 2013; Dwork and Roth, 2014). However, the following connection to zCDP is attributed to Bun and Steinke (2016).

Theorem 2.2.2 [Gaussian Mechanism (Bun and Steinke, 2016)]**.** *Let* $\phi : \mathcal{X}^n \to \mathbb{R}^d$*. The algorithm* $\mathcal{M}(\boldsymbol{x}) = \phi(\boldsymbol{x}) + \boldsymbol{Z}$ *where* $\boldsymbol{Z} \sim \mathrm{N}\left(\boldsymbol{0}, \frac{\Delta_2(\phi)^2}{2\rho} I_d\right)$*, is* ρ*-zCDP.*

We next relate (pure and approximate) differentially private with zCDP and show that it shares many of the nice properties of differential privacy. In fact zCDP can be thought of as being *in between* pure and approximate differential privacy, see Bun and Steinke (2016) for more details on this.

Theorem 2.2.3 [Bun and Steinke (2016)]**.** *If* $\mathcal{M}$ *is* ϵ*-DP, then* $\mathcal{M}$ *is* $\frac{\epsilon^2}{2}$*-zCDP. Further,*

$\mathcal{M}$ is ϵ-DP if and only if $\mathcal{M}$ is $(\epsilon, 0)$-zCDP.

Theorem 2.2.4 [Bun and Steinke (2016)]**.** *If $\mathcal{M}$ is (ξ, ρ)-zCDP then for any $\delta > 0$, we also have that $\mathcal{M}$ is $\left(\xi + \rho + 2\sqrt{\rho \log(\sqrt{\pi\ \rho}/\delta)}, \delta\right)$-DP.*

Theorem 2.2.5 [Post Processing and Composition (Bun and Steinke, 2016)]**.** *Let $\mathcal{M}_1 : \mathcal{X}^n \to \mathcal{Y}_1$ and $\mathcal{M}_2 : \mathcal{X}^n \times \mathcal{Y}_1 \to \mathcal{Y}_2$ be randomized algorithms where $\mathcal{M}_1$ is (ξ_1, ρ_1)-zCDP and $\mathcal{M}_2(\cdot, y_1)$ is (ξ_2, ρ_2)-zCDP for each $y_1 \in \mathcal{Y}_1$. Then the composition $\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}_2$ where $\mathcal{M}(\boldsymbol{x}) = \mathcal{M}_2(\boldsymbol{x}, \mathcal{M}_1(\boldsymbol{x}))$ is $(\xi_1 + \xi_2, \rho_1 + \rho_2)$-zCDP.*

Note that we can iteratively apply the above result to conclude that if each $\mathcal{M}_i : \mathcal{X}^n \times \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_{i-1}$ is (ξ_i, ρ_i)-zCDP in its first argument for $i \in [k]$, then $\mathcal{M}_{1:k} : \mathcal{X}^n \to \mathcal{Y}_k$ is $(\sum_{i=1}^k \xi_i, \sum_{i=1}^k \rho_i)$-zCDP

One immediate result from the connection between zCDP and DP is that we can get a better bound of the privacy parameter after adaptively selecting k Gaussian mechanisms going through Theorem 2.2.5 rather than Theorem 2.1.5 when we add the same scale of noise to the queries.

Lemma 2.2.6. *Let $\mathcal{M}_i : \mathcal{X}^n \times \mathcal{Y}_1 \times \cdots \mathcal{Y}_{i-1} \to \mathcal{Y}_i$ be a Gaussian mechanism and ρ-zCDP in its first argument for $i \in [k]$. Then $\mathcal{M}_{1:k} : \mathcal{X}^n \to \mathcal{Y}_k$ is (ϵ_g, δ_g)-DP where $\delta_g > 0$ and*

$$\epsilon_g = k\rho + 2\sqrt{k\rho\ \log(\sqrt{\pi\ \rho k}/\delta_g)}$$

Proof. We simply apply Theorem 2.2.5 followed by Theorem 2.2.4 □

Thus, if we want to achieve fixed privacy parameters (ϵ_g, δ_g), we require substantially less Gaussian noise added to each of the k queries using the above result with $\rho = \epsilon^2/2$ than with the results of Theorem 2.1.5. Specifically, we can improve on the standard deviation of the Gaussian noise added to each query. A similar improvement was noted by Abadi et al. (2016) (see Theorem 1).

CHAPTER 3

COMPARISON TO DATA-SPLITTING

We point out that the problem of adaptivity can often be avoided as long as we know how many adaptive analyses k are to be performed up front. If the data has n independent entries, then we can split the data into n/k chunks, and run each analysis on each chunk. This allows us to still apply the classical statistical tools to each split of data because the analysis and data are independent, which gives clear validity guarantees. However, this is inefficient in its use of data and requires $n \gg k$. We have stated some results in Chapter 1 that can do much better than data-splitting so that we can obtain valid results even for $n \ll k$. Further, many of the stated results do not require the data to be independent, in which case it is not even clear how to split the data into independent chunks.

We use this section to show how we can use previous results (Dwork et al., 2015c; Bassily et al., 2016; Russo and Zou, 2016) to obtain confidence intervals for adaptively chosen queries. Although we know that these results will *asymptotically* outperform traditional data-splitting techniques, we show that we can obtain smaller confidence intervals for finite sample sizes n. In fact, we can combine the analysis of Russo and Zou (2016) with the *monitor argument* from Bassily et al. (2016) to get valid confidence intervals that improves over data splitting techniques for reasonably sized data n and number of queries k.

3.1. Preliminaries

Throughout this section, we will write $\phi : \mathcal{X} \to [0,1]$ to denote a statistical query. Recall, that we assume that the data $\boldsymbol{X} = (X_1, \cdots, X_n) \in \mathcal{X}^n$ comes from a product distribution $\boldsymbol{X} \sim \mathcal{D}^n$, in addition we denote the empirical average as $\phi(\boldsymbol{X}) = \frac{1}{n}\sum_{i=1}^n \phi(X_i)$ and the true expected value as $\phi(\mathcal{D}) = \mathbb{E}_{X \sim \mathcal{D}}\left[\phi(X)\right]$.

In our implementation, we are comparing the true average $\phi(\mathcal{D})$ to the answer a, which will be the empirical average on the sample $\phi(\boldsymbol{X})$ with additional noise to ensure each query is selected in a differentially private way. Similar to how we defined accuracy of a sequence of algorithms with respect to the population in Definition 1.1.1, we define accuracy with respect to the sample.

Definition 3.1.1. *A sequence of algorithms $\mathcal{M} = (\mathcal{M}_1, \cdots, \mathcal{M}_k)$ is (τ', β')*-accurate on the sample *if for all analysts $\mathcal{A}$ that select $\phi_i : \mathcal{X}^n \to [0,1]$ which may depend on previous answers $a_j = \mathcal{M}_j(\boldsymbol{X})$ for $j = 1, \cdots, i-1$, we have*

$$\Pr\left[\max_{i \in [k]} |\phi_i(\boldsymbol{X}) - a_i| \leq \tau'\right] \geq 1 - \beta'.$$

We then use the following string of inequalities to find the width τ of the confidence interval,

$$\begin{aligned}\Pr\left[|\phi(\mathcal{D}) - a| \geq \tau\right] &\leq \Pr\left[|\phi(\mathcal{D}) - \phi(\boldsymbol{X})| + |\phi(\boldsymbol{X}) - a| \geq \tau\right] \\ &\leq \underbrace{\Pr\left[|\phi(\mathcal{D}) - \phi(\boldsymbol{X})| \geq \tau/2\right]}_{\text{Population Accuracy}} + \underbrace{\Pr\left[|\phi(\boldsymbol{X}) - a| \geq \tau/2\right]}_{\text{Sample Accuracy}}. \end{aligned} \tag{3.1}$$

Thus, we will bound the accuracy on the sample and the accuracy on the population. Some of the results in this line of work use a *transfer theorem* that states that if a query is selected via a differentially private method, then the query evaluated on the sample is close to the true population answer, thus providing a bound on *population accuracy*. However, we also need to control the *sample accuracy* which is affected by the amount of noise that is added to ensure differential privacy. We then give the accuracy guarantees of the Laplace and Gaussian mechanism from Theorem 2.1.2 and Theorem 2.2.2, respectively.

Theorem 3.1.2. *Let* $\{Y_i : i \in [k]\} \overset{i.i.d.}{\sim} \mathrm{Lap}(b)$ *then for* $\beta \in (0,1]$

$$\Pr\left[|Y_i| \geq \log(1/\beta) b\right] = \beta \implies \Pr\left[\exists i \in [k] \text{ s.t. } |Y_i| \geq b\log(k/\beta)\right] \leq \beta \tag{3.2}$$

Further, if $\{Z_i : i \in [k]\} \overset{i.i.d.}{\sim} \mathrm{N}\left(0, \sigma^2\right)$ *then for* $\beta \in (0,1]$

$$\Pr\left[|Z_i| \geq \sigma\sqrt{2\log(2/\beta)}\right] \leq \beta \implies \Pr\left[\exists i \in [k] \text{ s.t. } |Z_i| \geq \sigma\sqrt{2\log(2k/\beta)}\right] \leq \beta \tag{3.3}$$

We then seek a balance between both the sample and population accuracy, where too much noise will give terrible sample accuracy but great accuracy on the population – due to the noise making the choice of query essentially independent of the data – and too little noise makes for great sample accuracy but bad accuracy to the population. We will consider both Gaussian and Laplace noise and use the composition theorems from Chapter 2 to determine the privacy parameters after k adaptively selected statistical queries.

Given the size of our dataset n, number of adaptively chosen statistical queries k, and confidence level $1-\beta$, we want to find what *confidence width* τ ensures $\mathcal{M} = (\mathcal{M}_1, \cdots, \mathcal{M}_k)$ is (τ, β)-accurate with respect to the population when each algorithm $\mathcal{M}_i$ adds either Laplace or Gaussian noise to the answers computed on the sample.

3.2. Confidence Bounds from Dwork et al. (2015a)

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

Theorem 3.2.1. *If* $\mathcal{M}$ *is* (ϵ, δ)*-DP where* $\phi \leftarrow \mathcal{M}(\boldsymbol{X})$, $\tau \geq \sqrt{\frac{48}{n}\log(4/\beta)}$, $\epsilon \leq \tau/4$ *and* $\delta = \exp\left(\frac{-4\log(8/\beta)}{\tau}\right)$ *then*

$$\Pr\left[|\phi(\mathcal{D}) - \phi(\boldsymbol{X})| \geq \tau\right] \leq \beta$$

We pair this together with the accuracy from either the Gaussian mechanism or the Laplace mechanism with (3.1) to get the following result

Theorem 3.2.2. *Given confidence level* $1-\beta$ *and using the Laplace or Gaussian mechanism for each algorithm* $\mathcal{M}_i$ *for* $i \in [k]$, *then* $(\mathcal{M}_1, \cdots, \mathcal{M}_k)$ *is* $(\tau^{(1)}, \beta)$*-accurate.*

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

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

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

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

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

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

We then want to bound the two terms in (3.1). To bound the sample accuracy, we then use (3.2) so that

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

For the population accuracy, we need to apply Theorem 3.2.1, which requires us to have the following,

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

We then write ϵ in terms of the δ we fixed,

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

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

When can then follow a similar argument when we add Gaussian noise with variance $\frac{1}{2n^2\rho'}$. The only modification we make is using Lemma 2.2.6 to get a composed zCDP algorithm with parameters in terms of ρ', and the accuracy guarantee in (3.3) for Gaussian noise. ☐

We can then use the above result to actually generate valid confidence intervals for adaptively chosen statistical queries.

3.3. Confidence Bounds from Bassily et al. (2016)

We defer the details of the argument to Appendix A.2, which carefully goes through the analysis of Bassily et al. (2016) without using loose inequalities.

Theorem 3.3.1. *Given confidence level* $1-\beta$ *and using the Laplace or Gaussian mechanism for each algorithm* $\mathcal{M}_i$ *for* $i \in [k]$*, then* $(\mathcal{M}_1, \cdots, \mathcal{M}_k)$ *is* $(\tau^{(2)}, \beta)$*-accurate.*

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

$$\frac{1}{1-(1-\beta)^{\lfloor 1/\beta \rfloor}} \cdot \min_{\epsilon'>0, \delta \in (0,1)} \left\{ \exp\left(\frac{e^{\epsilon'}-1}{e^{\epsilon'}+1}\epsilon' k + \epsilon'\sqrt{2k\log(1/\delta)} \right) - 1 + 2\lfloor 1/\beta \rfloor \delta + \frac{\log(k/(2\delta))}{\epsilon' n} \right\}$$

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

$$\frac{1}{1-(1-\beta)^{\lfloor 1/\beta \rfloor}} \cdot \min_{\rho'>0, \delta \in (\sqrt{\pi\rho'},1)} \left\{ \exp\left(k\rho' + 2\sqrt{k\rho'\log(\sqrt{\pi\rho'}/\delta)} \right) - 1 + 2\lfloor 1/\beta \rfloor \delta + \frac{1}{n}\sqrt{1/\rho' \cdot \log(k/\delta)} \right\}$$

3.4. Confidence Bounds combining work from Russo and Zou (2016) and Bassily et al. (2016)

Using the monitor argument from Bassily et al. (2016) along with results from Russo and Zou (2016), we can obtain the following result, which uses a similar analysis from Bassily et al. (2016)).

Theorem 3.4.1. *Given confidence level* $1-\beta$ *and using the Gaussian mechanism for each algorithm* $\mathcal{M}_i$ *for* $i \in [k]$*, then* $(\mathcal{M}_1, \cdots, \mathcal{M}_k)$ *is* $(\tau^{(3)}, \beta)$*-accurate. We define* $\tau^{(3)}$ *to be the* *solution to the following program:*

$$\begin{aligned}
\min \quad & \tau \\
s.t. \quad & \tau \geq \sqrt{\frac{2}{n\beta} \cdot \left(2\rho' kn + \log\left(\rho' kn\right) + \frac{2\ \rho' kn + \log\left(\rho' kn\right)}{\rho' kn - 1} \right)} \\
& \tau \geq \frac{2}{n}\sqrt{\frac{1}{\rho'} \log(4k/\beta)} \\
& \rho' \geq \frac{1}{kn}
\end{aligned}$$

We will now present the argument to Theorem 3.4.1. Note that we will use $I\left(X;Y\right)$ to denote the mutual information (measured in bits) between random variables X and Y (see Definition 4.1.7 for a formal definition). Rather than use the stated result in Russo and Zou (2016), we use a modified version along with its proof. The result stated here and the one in Russo and Zou (2016) are incomparable.

Theorem 3.4.2. *Let $\mathcal{Q}_\sigma$ be the class of queries $q : \mathcal{X}^n \to \mathbb{R}$ such that $q(\boldsymbol{X}) - q(\mathcal{D}^n)$ is σ-subgaussian where $\boldsymbol{X} \sim \mathcal{D}^n$. If $\mathcal{M} : \mathcal{X}^n \to \mathcal{Q}_\sigma$ is a randomized mapping from datasets to queries such that* $\log(2) \cdot I\left(\mathcal{M}(\boldsymbol{X});\boldsymbol{X}\right) \leq B$ *with* $B \geq 1$, *then*

$$\mathop{\mathbb{E}}_{\boldsymbol{X} \sim \mathcal{D}^n, q \sim \mathcal{M}(\boldsymbol{X})} \left[\left(q(\boldsymbol{X}) - q(\mathcal{D}^n)\right)^2 \right] \leq \sigma^2 \cdot \left(2B + \log\left(B\right) + \frac{2\ B + \log\left(B\right)}{B-1} \right).$$

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

$$\begin{aligned}
& I\left(\mathcal{M}(\boldsymbol{X});\boldsymbol{X}\right) \geq I\left(\mathcal{M}(\boldsymbol{X});\boldsymbol{\phi}(\boldsymbol{X})\right) \\
& = \sum_{\boldsymbol{a},\ q \in \mathcal{Q}_\sigma} \Pr\left[(\boldsymbol{\phi}(\boldsymbol{X}), \mathcal{M}(\boldsymbol{X})) = (\boldsymbol{a}, q)\right] \log_2\left(\frac{\Pr\left[(\boldsymbol{\phi}(\boldsymbol{X}), \mathcal{M}(\boldsymbol{X})) = (\boldsymbol{a}, q)\right]}{\Pr\left[\boldsymbol{\phi}(\boldsymbol{X}) = \boldsymbol{a}\right] \Pr\left[\mathcal{M}(\boldsymbol{X}) = q\right]} \right) \\
& = \sum_{\boldsymbol{a},\ q \in \mathcal{Q}_\sigma} \Pr\left[\mathcal{M}(\boldsymbol{X}) = q\right] \Pr\left[\boldsymbol{\phi}(\boldsymbol{X}) = \boldsymbol{a} | \mathcal{M}(\boldsymbol{X}) = q\right] \log_2\left(\frac{\Pr\left[\boldsymbol{\phi}(\boldsymbol{X}) = \boldsymbol{a} | \mathcal{M}(\boldsymbol{X}) = q\right]}{\Pr\left[\boldsymbol{\phi}(\boldsymbol{X}) = \boldsymbol{a}\right]} \right) \\
& \geq \sum_{a,\ q \in \mathcal{Q}_\sigma} \Pr\left[\mathcal{M}(\boldsymbol{X}) = q\right] \Pr\left[q(\boldsymbol{X}) = a | \mathcal{M}(\boldsymbol{X}) = q\right] \log_2\left(\frac{\Pr\left[q(\boldsymbol{X}) = a | \mathcal{M}(\boldsymbol{X}) = q\right]}{\Pr\left[q(\boldsymbol{X}) = a\right]} \right) \\
& = \sum_{q \in \mathcal{Q}_\sigma} \Pr\left[\mathcal{M}(\boldsymbol{X}) = q\right] \mathrm{D}_{KL}\left[(q(\boldsymbol{X}) | \mathcal{M}(\boldsymbol{X}) = q) || q(\boldsymbol{X}) \right] \qquad (3.4)
\end{aligned}$$

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

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

where the first inequality follows from Fact 1 in Russo and Zou (2015), and the second inequality follows from Fact 3 in Russo and Zou (2015).

Therefore, from Eq. (3.4), we have

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

Rearranging terms, we have

$$\begin{aligned}
&\underset{\boldsymbol{X}\sim\mathcal{D}^n,\ q\sim\mathcal{M}(\boldsymbol{X})}{\mathbb{E}}\left[\left(q(\boldsymbol{X})-q(\mathcal{D})\right)^2\right]\leq\frac{2\sigma^2}{\lambda}\left(\log(2)I\left(\mathcal{M}(\boldsymbol{X});\boldsymbol{X}\right)+\log\left(\frac{1}{\sqrt{1-\lambda}}\right)\right)\\
&\quad=\sigma^2\cdot\frac{1}{\lambda}\left(2\log(2)I\left(\mathcal{M}(\boldsymbol{X});\boldsymbol{X}\right)+\log\left(\frac{1}{1-\lambda}\right)\right)\\
&\quad=\sigma^2\cdot\frac{2\log(2)I\left(\mathcal{M}(\boldsymbol{X});\boldsymbol{X}\right)+s}{1-e^{-s}} &&\text{(Substituting by } s=\log\left(\tfrac{1}{1-\lambda}\right)\text{)}\\
&\quad=\sigma^2\cdot\frac{2\log(2)I\left(\mathcal{M}(\boldsymbol{X});\boldsymbol{X}\right)+\log\left(\log(2)I\left(\mathcal{M}(\boldsymbol{X});\boldsymbol{X}\right)\right)}{1-\frac{1}{\log(2)I\left(\mathcal{M}(\boldsymbol{X});\boldsymbol{X}\right)}}\\
&&&\text{(Assigning } s=\log\left(\log(2)I\left(\mathcal{M}(\boldsymbol{X});\boldsymbol{X}\right)\right)\text{)}\\
&\quad=\sigma^2\cdot 2\log(2)I\left(\mathcal{M}(\boldsymbol{X});\boldsymbol{X}\right)+\log\left(\log(2)I\left(\mathcal{M}(\boldsymbol{X});\boldsymbol{X}\right)\right)\cdot\frac{\log(2)I\left(\mathcal{M}(\boldsymbol{X});\boldsymbol{X}\right)}{\log(2)I\left(\mathcal{M}(\boldsymbol{X});\boldsymbol{X}\right)-1}\\
&\quad=\sigma^2\cdot\left(2\log(2)I\left(\mathcal{M}(\boldsymbol{X});\boldsymbol{X}\right)+\log\left(\log(2)I\left(\mathcal{M}(\boldsymbol{X});\boldsymbol{X}\right)\right)\right)\cdot\left(1+\frac{1}{\log(2)I\left(\mathcal{M}(\boldsymbol{X});\boldsymbol{X}\right)-1}\right)\\
&\quad=\sigma^2\cdot\left(2\log(2)I\left(\mathcal{M}(\boldsymbol{X});\boldsymbol{X}\right)+\log\left(\log(2)I\left(\mathcal{M}(\boldsymbol{X});\boldsymbol{X}\right)\right)\right)\\
&\qquad+\sigma^2\cdot\left(\frac{2\log(2)I\left(\mathcal{M}(\boldsymbol{X});\boldsymbol{X}\right)+\log\left(\log(2)I\left(\mathcal{M}(\boldsymbol{X});\boldsymbol{X}\right)\right)}{\log(2)I\left(\mathcal{M}(\boldsymbol{X});\boldsymbol{X}\right)-1}\right)
\end{aligned}$$

□

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

Lemma 3.4.3. *For statistical queries ϕ and $\boldsymbol{X}\sim\mathcal{D}^n$, we have $\phi(\boldsymbol{X})-\phi(\mathcal{D}^n)$ is $\frac{1}{2\sqrt{n}}$-subgaussian.*

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

Lemma 3.4.4 [Bun and Steinke (2016)]**.** *If* $\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}$ *is* ρ*-zCDP and* $\boldsymbol{X} \sim \mathcal{D}^n$*, then*

$$\log(2) I\left(\mathcal{M}(\boldsymbol{X}); \boldsymbol{X}\right) \leq \rho n$$

We define the monitor in Algorithm 1.

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

```
Input: $\boldsymbol{x} \in \mathcal{X}^n$
    As we outlined in Section 1.1, we simulate $\mathcal{M}(\boldsymbol{X})$ and $\mathcal{A}$ interacting. We write $q_1, \cdots, q_k \in \mathcal{Q}_{SQ}$ as the queries chosen by $\mathcal{A}$ and write $a_1, \cdots, a_k \in \mathbb{R}$ as the corresponding answers of $\mathcal{M}$.
    Let
        $j^* = \operatorname*{argmax}_{j \in [k]} \left| q_j(\mathcal{D}) - a_j \right|.$
    $q^* \leftarrow q_{j^*}$
Output: $q^*$
```

We first need to show that the monitor has bounded mutual information as long as $\mathcal{M}$ does, which follows from mutual information being preserved under postprocessing, or the data processing inequality.

Lemma 3.4.5. *If* $I\left(\mathcal{M}(\boldsymbol{X}); \boldsymbol{X}\right) \leq B$ *where* $\boldsymbol{X} \sim \mathcal{D}^n$*, then* $I\left(\mathcal{W}_{\mathcal{D}}[\mathcal{M}, \mathcal{A}](\boldsymbol{X}); \boldsymbol{X}\right) \leq B$*.*

We are now ready to prove our result.

Proof of Theorem 3.4.1. We follow the same analysis for proving Theorem 3.3.1 where we add Gaussian noise with variance $\frac{1}{2\rho' n^2}$ to each query answer so that the algorithm $\mathcal{M}$ is $\rho' k$-zCDP, which (using Lemmas 3.4.4 and 3.4.5) makes the mutual information bound $B = \rho' k n$. We then use the sub-Gaussian parameter for statistical queries in Lemma 3.4.3 to obtain the following bound from Theorem 3.4.2.

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

We can then bound the population accuracy in (3.1) using Chebyshev's inequality to obtain the following high probability bound with the sequence of answers $a_1, \cdots, a_k$ given by $\mathcal{M}$

at each round,

$$
\begin{aligned}
&\Pr_{\boldsymbol{X}\sim\mathcal{D}^n,\mathcal{M},\mathcal{A}}\left[\max_{i\in[k]}\{|q_i(\mathcal{D})-a_i|\geq\tau\}\right]\\
&\qquad\leq\frac{4}{\tau^2}\,\mathbb{E}\left[\max_{i\in[k]}\left\{|q_i(\boldsymbol{X})-q_i(\mathcal{D})|\right\}^2\right]+\Pr\left[\max_{i\in[k]}\{|q_i(\boldsymbol{X})-a_i|\geq\tau/2\}\right]\\
&\qquad\leq\frac{1}{n\tau^2}\cdot\left(2\rho' kn+\log\left(\rho' kn\right)+\frac{2\ \rho' kn+\log\left(\rho' kn\right)}{\rho' kn-1}\right)+\Pr\left[\max_{i\in[k]}\{|q_i(\boldsymbol{X})-a_i|\geq\tau/2\}\right].
\end{aligned}
$$

To ensure this is at most β, we require τ to satisfy both

$$
\tau\geq\sqrt{\frac{2}{n\beta}\cdot\left(2\rho' kn+\log\left(\rho' kn\right)+\frac{2\ \rho' kn+\log\left(\rho' kn\right)}{\rho' kn-1}\right)}\quad\&\quad\tau\geq\frac{2}{n}\sqrt{\frac{1}{\rho'}\log(4k/\beta)}.
$$

We then minimize over $\rho'>0$, which finishes the proof. Note that the condition $\rho'>\frac{1}{kn}$ is needed due to Theorem 3.4.2.

$\square$

3.5. Confidence Bound Results

In Figure 3, we give the widths of the valid confidence intervals for k adaptively selected statistical queries where each answer has noise added to it. We label "DFHPRR" the bound you get from Theorem 3.2.1, "BNSSSU" as the bound we get from Theorem 3.3.1, and "RZ+Monitor" as the bound we get from Theorem 3.4.1. The traditional approach of splitting the data and running each analysis on each chunk is exhibited in the plot called "Data Splitting", where we are bounding the probability distribution of a binomial random variable for each n/k chunk of data and applying a union bound over all k chunks. That is we find the smallest integer $t\in[n/k]$ such that for each $i\in[k]$

$$
1-\frac{1}{2^{n/k}}\sum_{j=1}^{t}\binom{n/k}{j}\leq\frac{\beta}{2k}.
$$

This will ensure that $\Pr_{\boldsymbol{X}\sim\mathcal{D}^{n/k}}\left[\max_{i\in[k]}|\phi_i(\mathcal{D})-\phi_i(\boldsymbol{X})|\geq t\right]\leq\beta$

We also plot the resulting standard deviation of the noise we added to each statistical query in Figure 4 in order to generate the plots in Figure 3. In our experiments, when we use Gaussian noise and combine the results from Russo and Zou (2016) with the monitor argument of Bassily et al. (2016) we get the best confidence bounds with an improvement over datasplitting when $n=6400$ and $k=640$.

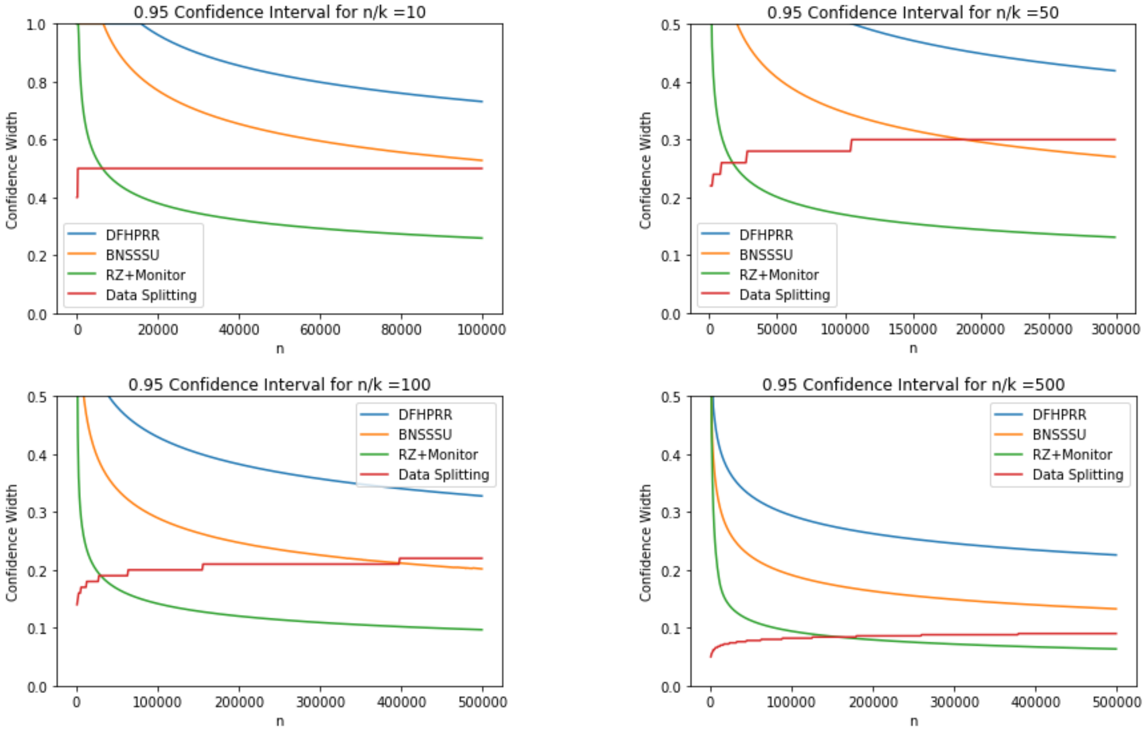


Figure 3: Widths of valid confidence intervals for k adaptively chosen statistical queries via data-splitting techniques or noise addition on the same dataset.

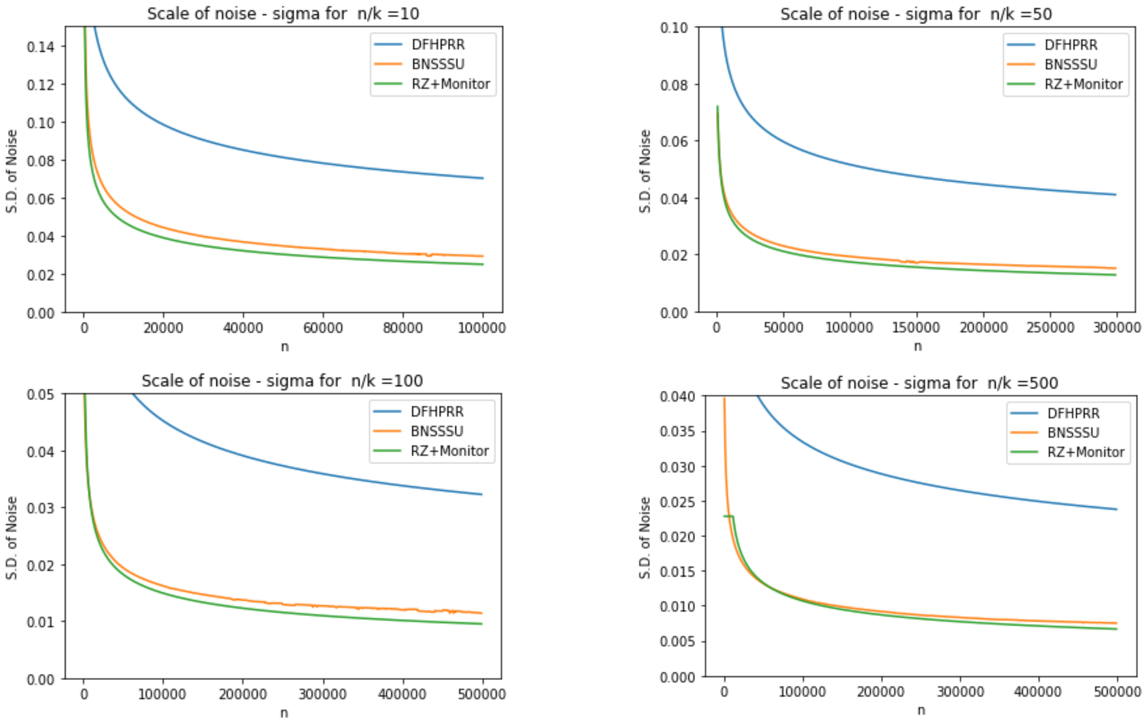


Figure 4: Standard deviations of the Gaussian noise we added to each query to obtain the confidence widths.

Part II

DIFFERENTIAL PRIVACY IN ADAPTIVE DATA ANALYSIS: INFORMATION AND COMPOSITION

We next cover some of the contributions we have made in connecting differential privacy to adaptive data analysis. We unify and improve some of the previous results along this line of research by connecting max-information with (approximate)-differential privacy. Previously, Dwork et al. (2015a) showed that pure-differentially private algorithms had bounded max-information, which we then extend to the much broader class of approximate differentially private algorithms. This connection of max-information with approximate differential privacy is crucial in showing that these methods will outperform data-splitting techniques when stringing together many adaptively chosen analyses. As a special case of our results, we are able to recover the results from Bassily et al. (2016) (with some loss in parameters) for low-sensitivity queries. However, our results extend to much broader analyses, like post-selection hypothesis testing. Further, we improve the connection between p-value corrections and mutual information of the analyses using results from Russo and Zou (2016) via max-information.

Attached to each differentially private algorithm is a privacy parameter, where roughly lower parameter values ensure better privacy. So as we compose several differentially private algorithms, we expect the privacy parameters to get worse, but the privacy parameter only increases sublinearly with the number of analyses that were ran. This sublinear composition is precisely what gives these differentially private methods their power over data-splitting. However, there is a caveat to these composition theorems. In order to apply these results from differential privacy, we need the privacy parameters to all be fixed up front, prior to running any analysis on the data. I will them present a new framework for composition of differential privacy where the privacy parameters and the number of analyses need not be fixed up front. Instead, we want to allow the analyst the freedom to choose a time to quit running analyses and modify his privacy budget based on what he has observed. We then develop novel composition definitions that are better catered to the adaptive data analysis setting. In this new framework, we then show that adaptivity comes at a cost, in that we cannot simply apply the composition bounds as if we knew the parameters upfront. Despite this negative result, we can still obtain sublinear composition bounds.

CHAPTER 4

MAX-INFORMATION, DIFFERENTIAL PRIVACY, AND POST-SELECTION HYPOTHESIS TESTING

Although our presentation in Chapter 1 was motivated by low-sensitivity queries and p-value corrections, an algorithm $\mathcal{M}$ with bounded max-information allows a data analyst to treat *any event* that is a function of the output of the algorithm $\mathcal{M}(\boldsymbol{X})$ "as if" it is independent of the dataset $\boldsymbol{X}$, up to a correction factor determined by the max-information bound. The results presented in this chapter substantially broaden the class of analyses for which approximate differential privacy promises generalization guarantees—this class was previously limited to estimating the values of low-sensitivity numeric valued queries and more generally, the outcomes of low-sensitivity optimization problems (Bassily et al., 2016). This chapter largely follows from Rogers et al. (2016a), where we also introduced the framework for reasoning about adaptive hypothesis testing with p-value corrections that was presented in Section 1.3.

This chapter further develops the extent to which max-information can be viewed as a unifying information theoretic measure controlling the generalization properties of adaptive data analysis. Dwork et al. (2015a) previously showed that algorithms with bounded output description length, and algorithms that satisfy pure-differential privacy (two constraints known individually to imply adaptive generalization guarantees), both have bounded max-information. Because bounded max-information satisfies strong composition properties, this connection implies that algorithms with bounded output description length and pure-differentially private algorithms can be composed in arbitrary order and the resulting composition will still have strong generalization properties. Our result brings approximate-differential privacy partially into this unifying framework. In particular, *when the data is drawn from a product distribution*, if an analysis that starts with an (arbitrary) approximate differentially private computation is followed by an arbitrary composition of algorithms with bounded max-information, then the resulting composition will satisfy a max-information bound. However, unlike with compositions consisting solely of bounded description length mechanisms and pure differentially private mechanisms, which can be composed in arbitrary order, in this case *it is important that the approximate-differentially private computation come first.* This is because, even if the dataset $\boldsymbol{X}$ is initially drawn from a product distribution, the conditional distribution on the data that results after observing the outcome of an initial computation need not be a product distribution any longer. In fact, the lower bound we prove in Section 4.4 is an explicit construction in which the composition of a bounded description length algorithm, followed by an approximate-differentially private algorithm can be used to exactly reconstruct a dataset drawn from a product distribution (which can in turn be used to arbitrarily overfit that dataset).

As was demonstrated in Chapter 1, max-information has some nice properties that are useful in adaptive data analysis. However, it is not the only measure of information for algorithms. Specifically, the measure of Shannon mutual information has been extensively studied, including its connection with differential privacy. In fact, algorithms with bounded mutual information were shown to have low *bias* error in Russo and Zou (2016). These bounds can then obtain high probability guarantees, for the type of generalization guarantees we are concerned with by Markov or Chebeshev's inequality. However, we will demonstrate that max-information provides a tighter connection than the results in Russo and Zou (2016). We then show a general conversion between mutual information and max-information, which improves over existing results of the mutual information for differentially private algorithms from McGregor et al. (2011).

We also connect some of our results with those in Cummings et al. (2016a), specifically with how we can use our lower bound result to show that *robustly generalizing* algorithms do not compose and how *compression schemes* can have large max-information.

4.1. Additional Preliminaries

We first cover some previous results which will prove to be useful in our analysis. In the introduction, we define (approximate-) max-information, and we now give some other measures between distributions. We introduced indistinguishability between two random variables in Definition 2.0.1, but we give a slightly stronger measure of similarity between two random variables, called point-wise indistinguishability.

Definition 4.1.1 [Point-wise indistinguishibility (Kasiviswanathan and Smith, 2014)]**.** *Two random variables X, Z taking values in a set $\mathcal{X}$ are point-wise (ϵ, δ)-indistinguishable if with probability at least $1-\delta$ over $a \sim p(X)$:*

$$e^{-\epsilon} \Pr[Z = a] \leq \Pr[X = a] \leq e^{\epsilon} \Pr[Z = a].$$

We next give several useful connections between indistinguishability, point-wise indistinguishability, and differential privacy along with other more widely known measures between distributions, e.g., KL-divergence, and total-variation distance.

Definition 4.1.2 [KL Divergence]**.** *The KL Divergence between random variables X and Z over domain $\mathcal{X}$, denoted as $\mathrm{D}_{KL}(X||Z)$ is defined as*

$$\mathrm{D}_{KL}(X||Z) = \sum_{x \in \mathcal{X}} \Pr[X = x] \ln\left(\frac{\Pr[X = x]}{\Pr[Z = x]}\right)$$

Definition 4.1.3 [Total Variation Distance]**.** *The* total variation distance *between two random variables X and Z over domain $\mathcal{X}$, denoted as $TV(X; Z)$ is defined as*

$$TV(X, Z) = \frac{1}{2} \cdot \sum_{x \in \mathcal{X}} |\Pr[X = x] - \Pr[Z = x]|.$$

In the following lemma, we state some basic connections between max-information and

(point-wise) indistinguishability:
Lemma 4.1.4. *Let* X, Z *be two random variables over the same domain. We then have:*

1. *(Dwork et al., 2015a)* $I_\infty^\beta(X;Z) \leq m \Leftrightarrow (X,Z) \approx_{(m\log 2),\beta} X \otimes Z$.
2. *(Kasiviswanathan and Smith, 2014) If* $X \approx_{\epsilon,\delta} Y$ *then* X *and* Y *are pointwise* $\left(2\epsilon, \frac{2\delta}{1-e^{-\epsilon}}\right)$*-indistinguishable.*

Another useful result is from Kasiviswanathan and Smith (2014), which we use in the proof of our main result in Theorem 4.2.1:
Lemma 4.1.5 [Conditioning Lemma]**.** *Suppose that* $(X,Z) \approx_{\epsilon,\delta} (X',Z')$*. Then for every* $\widehat{\delta} > 0$*, the following holds:*

$$\Pr_{t\sim Z}\left[X|_{Z=t} \approx_{3\epsilon,\widehat{\delta}} X'|_{Z'=t}\right] \geq 1 - \frac{2\delta}{\widehat{\delta}} - \frac{2\delta}{1-e^{-\epsilon}}.$$

The proof of our main result in Theorem 4.2.1 also makes use of the following standard concentration inequality:
Theorem 4.1.6 [Azuma's Inequality]**.** *Let* $C_1, \cdots, C_n$ *be a sequence of random variables such that for every* $i \in [n]$*, we have*

$$\Pr\left[|C_i| \leq \alpha\right] = 1$$

and for every fixed prefix $c_1, \cdots, c_{i-1}$*, we have*

$$\mathbb{E}\left[C_i|(C_1,\cdots,C_{i-1}) = (c_1,\cdots,c_{i-1})\right] \leq \gamma,$$

then for all $t \geq 0$*, we have*

$$\Pr\left[\sum_{i=1}^n C_i > n\gamma + t\sqrt{n}\alpha\right] \leq e^{-t^2/2}.$$

We will also use Shannon mutual information later in this chapter in order to compare it with max-information.
Definition 4.1.7 [Mutual Information]**.** *Consider two random variables* X *and* Y *and let* $Z(x,y) = \log_2\left(\frac{\Pr[(X,Y)=(x,y)]}{\Pr[X=x]\Pr[Y=y]}\right)$*. We then denote the mutual information as the following, where the expectation is taken over the joint distribution of* (X,Y)*,*

$$I(X;Y) = \mathbb{E}\left[Z(X,Y)\right].$$

4.2. Max-information for (ϵ, δ)-Differentially Private Algorithms

In this section, we prove a bound on approximate max-information for (ϵ, δ)-differentially private algorithms over product distributions.

Theorem 4.2.1. *Let $\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}$ be an (ϵ, δ)-differentially private algorithm for $\epsilon \in (0, 1/2]$ and $\delta \in (0, \epsilon)$. For $\beta = e^{-\epsilon^2 n} + O\left(n\sqrt{\frac{\delta}{\epsilon}}\right)$, we have*

$$I_{\infty,\Pi}^{\beta}\left(\mathcal{M}, n\right) = O\left(\epsilon^2 n + n\sqrt{\frac{\delta}{\epsilon}}\right).$$

We will prove Theorem 4.2.1 over the course of this section, using a number of lemmas. We first set up some notation. We will sometimes abbreviate conditional probabilities of the form $\Pr\left[\boldsymbol{X} = \boldsymbol{x} | \mathcal{M} = a\right]$ as $\Pr\left[\boldsymbol{X} = \boldsymbol{x} | a\right]$ when the random variables are clear from context. We will also abbreviate vectors $\boldsymbol{x}_{<i} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{i-1})$. Further, for any $\boldsymbol{x} \in \mathcal{X}^n$ and $a \in \mathcal{Y}$, we define

$$\begin{aligned} Z(a, \boldsymbol{x}) &\stackrel{\text{def}}{=} \log\left(\frac{\Pr\left[\mathcal{M} = a, \boldsymbol{X} = \boldsymbol{x}\right]}{\Pr\left[\mathcal{M} = a\right] \cdot \Pr\left[\boldsymbol{X} = \boldsymbol{x}\right]}\right) \\ &= \sum_{i=1}^{n} \log\left(\frac{\Pr\left[X_i = x_i | a, \boldsymbol{x}_{<i}\right]}{\Pr\left[X_i = x_i\right]}\right) \end{aligned} \tag{4.1}$$

If we can bound $Z(a, \boldsymbol{x})$ with high probability over $(a, \boldsymbol{x}) \sim \mathcal{M}(\boldsymbol{X}), \boldsymbol{X}$, then we can bound the approximate max-information by using the following lemma:

Lemma 4.2.2 [See Lemma 18 in Dwork et al. (2015a)]**.** *If* $\Pr\left[Z(\mathcal{M}(\boldsymbol{X}), \boldsymbol{X}) \geq k\right] \leq \beta$, *then* $I_{\infty}^{\beta}\left(\boldsymbol{X}; \mathcal{M}(\boldsymbol{X})\right) \leq k$.

We next define each term in the sum of $Z(a, \boldsymbol{x})$ as

$$Z_i(a, \boldsymbol{x}_{\leq i}) \stackrel{\text{def}}{=} \log \frac{\Pr\left[X_i = x_i | a, \boldsymbol{x}_{<i}\right]}{\Pr\left[X_i = x_i\right]}. \tag{4.2}$$

The plan of the proof is simple: our goal is to apply Azuma's inequality (Theorem 4.1.6) to the sum of the Z_i's to achieve a bound on Z with high probability. Applying Azuma's inequality requires both understanding the expectation of each term $Z_i(a, \boldsymbol{x}_{\leq i})$, and being able to argue that each term is bounded. Unfortunately, in our case, the terms are not always bounded – however, we will be able to show that they are bounded with high probability. This plan is somewhat complicated by the conditioning in the definition of $Z_i(a, \boldsymbol{x}_{\leq i})$.

First, we argue that we can bound each Z_i with high probability. This argument takes place over the course of Claims 4.2.3, 4.2.4, 4.2.5 and 4.2.6.

Claim 4.2.3. *If $\mathcal{M}$ is (ϵ, δ)-differentially private and $\boldsymbol{X} \sim \mathcal{D}^n$, then for each $i \in [n]$ and* *each prefix* $\boldsymbol{x}_{<i} \in \mathcal{X}^{i-1}$*, we have:*

$$(\mathcal{M}, X_i)|_{\boldsymbol{x}_{<i}} \approx_{\epsilon,\delta} \mathcal{M}|_{\boldsymbol{x}_{<i}} \otimes X_i.$$

Proof. Fix any set $\mathcal{O} \subseteq \mathcal{Y} \times \mathcal{X}$ and prefix $\boldsymbol{x}_{<i} \in \mathcal{X}^{i-1}$. We then define the set $\mathcal{O}_{x_i} = \{a \in \mathcal{Y} : (a, x_i) \in \mathcal{O}\}$. Now, we have that:

$$\begin{aligned}\Pr\left[(\mathcal{M}(\boldsymbol{X}), X_i) \in \mathcal{O}|\boldsymbol{x}_{<i}\right] &= \sum_{x_i \in \mathcal{X}} \Pr\left[X_i = x_i\right] \Pr\left[\mathcal{M}(\boldsymbol{X}) \in \mathcal{O}_{x_i}|\boldsymbol{x}_{<i}, x_i\right] \\ &\leq \sum_{x_i \in \mathcal{X}} \Pr\left[X_i = x_i\right] \left(e^{\epsilon} \Pr\left[\mathcal{M}(\boldsymbol{X}) \in \mathcal{O}_{x_i}|\boldsymbol{x}_{<i}, t_i\right] + \delta\right) \qquad \forall t_i \in \mathcal{X}\end{aligned}$$

Thus, we can multiply both sides of the inequality by $\Pr\left[X_i = t_i\right]$ and sum over all $t_i \in \mathcal{X}$ to get:

$$\begin{aligned}&\Pr\left[(\mathcal{M}(\boldsymbol{X}), X_i) \in \mathcal{O}|\boldsymbol{x}_{<i}\right] = \sum_{t_i \in \mathcal{X}} \Pr\left[X_i = t_i\right] \Pr\left[(\mathcal{M}(\boldsymbol{X}), X_i) \in \mathcal{O}|\boldsymbol{x}_{<i}\right] \\ &\leq \sum_{x_i \in \mathcal{X}} \sum_{t_i \in \mathcal{X}} \Pr\left[X_i = x_i\right] \Pr\left[X_i = t_i\right] \left(e^{\epsilon} \Pr\left[\mathcal{M}(\boldsymbol{X}) \in \mathcal{O}_{x_i}|\boldsymbol{x}_{<i}, t_i\right] + \delta\right) \\ &\leq e^{\epsilon} \sum_{x_i \in \mathcal{X}} \Pr\left[X_i = x_i\right] \Pr\left[\mathcal{M}(\boldsymbol{X}) \in \mathcal{O}_{x_i}|\boldsymbol{x}_{<i}\right] + \delta = e^{\epsilon} \Pr\left[\mathcal{M}(\boldsymbol{X}) \otimes X_i \in \mathcal{O}|\boldsymbol{x}_{<i}\right] + \delta.\end{aligned}$$

We follow a similar argument to prove:

$$\Pr\left[\mathcal{M}(\boldsymbol{X}) \otimes X_i \in \mathcal{O}|\boldsymbol{x}_{<i}\right] \leq e^{\epsilon} \Pr\left[(\mathcal{M}(\boldsymbol{X}), X_i) \in \mathcal{O}|\boldsymbol{x}_{<i}\right] + \delta.$$

□

We now define the following set of "good outcomes and prefixes" for any $\widehat{\delta} > 0$:

$$\mathrm{E}_i(\widehat{\delta}) \stackrel{\text{defn}}{=} \left\{(a, \boldsymbol{x}_{<i}) : X_i \approx_{3\epsilon,\widehat{\delta}} X_i|_{a,\boldsymbol{x}_{<i}}\right\} \tag{4.3}$$

We use a technical lemma from Kasiviswanathan and Smith (2014) stated in Lemma 4.1.5, and Claim 4.2.3 to derive the following result:

Claim 4.2.4. *If* $\mathcal{M}$ *is* (ϵ, δ)*-differentially private and* $\boldsymbol{X} \sim \mathcal{D}^n$*, then for each* $i \in [n]$ *and each prefix* $\boldsymbol{x}_{<i} \in \mathcal{X}^{i-1}$ *we have for* $\widehat{\delta} > 0$ *and* $\delta' \stackrel{def}{=} \frac{2\delta}{\widehat{\delta}} + \frac{2\delta}{1-e^{-\epsilon}}$*:*

$$\Pr\left[(\mathcal{M}, \boldsymbol{X}_{<i}) \in \mathrm{E}_i(\widehat{\delta})|\boldsymbol{x}_{<i}\right] \geq 1 - \delta'.$$

Proof. This follows directly from Lemma 4.1.5:

$$\begin{aligned}
&\Pr\left[(\mathcal{M}, \boldsymbol{X}_{<i}) \in \mathrm{E}_i(\widehat{\delta}) | \boldsymbol{x}_{<i}\right] \\
&\quad = \Pr\left[X_i \approx_{3\epsilon,\widehat{\delta}} X_i|_{\mathcal{M},\boldsymbol{x}_{<i}} | \boldsymbol{x}_{<i}\right] \\
&\quad = \Pr_{a\sim\mathcal{M}|_{\boldsymbol{x}_{<i}}}\left[X_i \approx_{3\epsilon,\widehat{\delta}} X_i|_{a,\boldsymbol{x}_{<i}}\right] \geq 1-\delta'
\end{aligned}$$

□

We now define the set of outcome/dataset prefix pairs for which the quantities Z_i are not large:

$$\mathrm{F}_i \stackrel{\text{defn}}{=} \left\{(a, \boldsymbol{x}_{\leq i}) : |Z_i(a, \boldsymbol{x}_{\leq i})| \leq 6\epsilon\right\}. \tag{4.4}$$

Using another technical lemma from Kasiviswanathan and Smith (2014) (which we state in Lemma 4.1.4 in the appendix), we prove:

Claim 4.2.5. *Given* $(a, \boldsymbol{x}_{<i}) \in \mathrm{E}_i(\widehat{\delta})$ *and* $\delta'' \stackrel{def}{=} \frac{2\widehat{\delta}}{1-e^{-3\epsilon}}$ *we have:*

$$\Pr\left[(\mathcal{M}, \boldsymbol{X}_{\leq i}) \in \mathrm{F}_i | a, \boldsymbol{x}_{<i}\right] \geq 1-\delta''.$$

Proof. Since $(a, \boldsymbol{x}_{<i}) \in \mathrm{E}_i(\widehat{\delta})$, we know that X_i is $(3\epsilon, \widehat{\delta})$-indistinguishable from $X_i|_{a,\boldsymbol{x}_{<i}}$. Using Lemma 4.1.4, we know that X_i and $X_i|_{a,\boldsymbol{x}_{<i}}$ are point-wise $(6\epsilon, \delta'')$-indistinguishable. Thus, by definition of F_i and Z_i, we have:

$$\begin{aligned}
&\Pr\left[(\mathcal{M}, \boldsymbol{X}_{\leq i}) \in \mathrm{F}_i | a, \boldsymbol{x}_{<i}\right] = \Pr\left[Z_i(\mathcal{M}, \boldsymbol{X}_{\leq i}) \leq 6\epsilon | a, \boldsymbol{x}_{<i}\right] \\
&\quad = \Pr_{x_i \sim X_i|_{a,\boldsymbol{x}_{<i}}}\left[\log\left(\frac{\Pr[X_i = x_i | a, \boldsymbol{x}_{<i}]}{\Pr[X_i = x_i]}\right) \leq 6\epsilon\right] \geq 1-\delta''
\end{aligned}$$

□

We now define the "good" tuples of outcomes and databases as

$$\mathrm{G}_i(\widehat{\delta}) \stackrel{\text{defn}}{=} \left\{(a, \boldsymbol{x}_{\leq i}) : (a, \boldsymbol{x}_{<i}) \in \mathrm{E}_i(\widehat{\delta}) \quad \& \quad (a, \boldsymbol{x}_{\leq i}) \in \mathrm{F}_i\right\}, \tag{4.5}$$

$$\mathrm{G}_{\leq i}(\widehat{\delta}) \stackrel{\text{defn}}{=} \left\{(a, \boldsymbol{x}_{\leq i}) : (a, x_1) \in \mathrm{G}_1(\widehat{\delta}), \cdots, (a, \boldsymbol{x}_{\leq i}) \in \mathrm{G}_i(\widehat{\delta})\right\} \tag{4.6}$$

Claim 4.2.6. *If* $\mathcal{M}$ *is* (ϵ, δ)*-differentially private and* $\boldsymbol{X} \sim \mathcal{D}^n$*, then*

$$\Pr\left[(\mathcal{M}, \boldsymbol{X}_{\leq i}) \in \mathrm{G}_i(\widehat{\delta})\right] \geq 1-\delta'-\delta''$$

for δ' *and* δ'' *given in Claim 4.2.4 and Claim 4.2.5, respectively.*

Proof. We have:

$$\begin{aligned}
\Pr\left[(\mathcal{M},\boldsymbol{X}_{\leq i}) \notin \mathrm{G}_i(\widehat{\delta})\right] &= \Pr\left[(\mathcal{M},\boldsymbol{X}_{<i}) \notin \mathrm{E}_i(\widehat{\delta}) \qquad \text{or} \qquad (\mathcal{M},\boldsymbol{X}_{\leq i}) \notin \mathrm{F}_i)\right] \\
&= 1 - \Pr\left[(\mathcal{M},\boldsymbol{X}_{<i}) \in \mathrm{E}_i(\widehat{\delta}) \qquad \text{and} \qquad (\mathcal{M},\boldsymbol{X}_{\leq i}) \in \mathrm{F}_i)\right] \\
&= 1 - \sum_{(a,\boldsymbol{x}_{<i})\in\mathrm{E}_i(\widehat{\delta})} \Pr\left[(\mathcal{M},\boldsymbol{X}_{<i}) = (a,\boldsymbol{x}_{<i})\right]\Pr\left[(\mathcal{M},\boldsymbol{X}_{\leq i}) \in \mathrm{F}_i | a,\boldsymbol{x}_{<i}\right] \\
&\leq 1 - \sum_{(a,\boldsymbol{x}_{<i})\in\mathrm{E}_i(\widehat{\delta})} \Pr\left[(\mathcal{M},\boldsymbol{X}_{<i}) = (a,\boldsymbol{x}_{<i})\right]\cdot(1-\delta'') \\
&= 1 - (1-\delta'')\Pr\left[(\mathcal{M},\boldsymbol{X}_{<i}) \in \mathrm{E}_i(\widehat{\delta})\right] \\
&\leq 1 - (1-\delta'')(1-\delta') = \delta' + \delta'' - \delta'\delta''
\end{aligned}$$

where the last two inequalities follow from Claim 4.2.5 and Claim 4.2.4, respectively. □

Having shown a high probability bound on the terms Z_i, our next step is to bound their expectation so that we can continue towards our goal of applying Azuma's inequality. Note a complicating factor – throughout the argument, we need to condition on the event $(\mathcal{M},\boldsymbol{X}_{\leq i}) \in \mathrm{F}_i$ to ensure that Z_i has bounded expectation.

We will use the following shorthand notation for conditional expectation:

$$\begin{aligned}
&\mathbb{E}\left[Z_i(\mathcal{M},\boldsymbol{X}_{\leq i})|a,\boldsymbol{x}_{<i},\mathrm{F}_i\right] \\
&\quad\overset{\text{defn}}{=} \mathbb{E}\left[Z_i(\mathcal{M},\boldsymbol{X}_{\leq i})|\mathcal{M}=a,\boldsymbol{X}_{<i}=\boldsymbol{x}_{<i},(\mathcal{M},\boldsymbol{X}_{\leq i}) \in \mathrm{F}_i\right],
\end{aligned}$$

with similar notation for sets $\mathrm{G}_i(\widehat{\delta}), \mathrm{G}_{\leq i}(\widehat{\delta})$.

Lemma 4.2.7. *Let $\mathcal{M}$ be (ϵ,δ)-differentially private and $\boldsymbol{X} \sim \mathcal{D}^n$. Given $(a,\boldsymbol{x}_{<i}) \in \mathrm{E}_i(\widehat{\delta})$, for all $\epsilon \in (0,1/2]$ and $\widehat{\delta} \in (0,\epsilon/15]$,*

$$\mathbb{E}\left[Z_i(\mathcal{M},\boldsymbol{X}_{\leq i})|a,\boldsymbol{x}_{<i},\mathrm{F}_i\right] = O(\epsilon^2 + \widehat{\delta}).$$

More precisely, $\mathbb{E}\left[Z_i(\mathcal{M},\boldsymbol{X}_{\leq i})|a,\boldsymbol{x}_{<i},\mathrm{F}_i\right] \leq \nu(\widehat{\delta})$, where $\nu(\widehat{\delta})$ is defined in (4.7).

Proof. Given an outcome and prefix $(a,\boldsymbol{x}_{<i}) \in \mathrm{E}_i(\widehat{\delta})$, we define the set of data entries

$$\mathcal{X}(a,\boldsymbol{x}_{<i}) \overset{\text{defn}}{=} \{x_i \in \mathcal{X} : (a,\boldsymbol{x}_{\leq i}) \in \mathrm{F}_i\}.$$

We then have:

$$\mathbb{E}\left[Z_i(\mathcal{M},\boldsymbol{X}_{\leq i})|a,\boldsymbol{x}_{<i},\mathrm{F}_i\right] = \sum_{x_i\in\mathcal{X}(a,\boldsymbol{x}_{<i})} \Pr\left[X_i = x_i|a,\boldsymbol{x}_{<i},\mathrm{F}_i\right]\log\left(\frac{\Pr[X_i=x_i|a,\boldsymbol{x}_{<i}]}{\Pr[X_i=x_i]}\right)$$

Here, our goal is to mimic the proof of the "advanced composition theorem" of Dwork et al.

(2010) by adding a term that looks like a KL divergence term (see Definition 4.1.2). In our case, however, the sum is not over the entire set $\mathcal{X}$, and so it is not a KL-divergence, which leads to some additional complications. Consider the following term:

$$\begin{aligned}
&\sum_{x_i \in \mathcal{X}(a,\boldsymbol{x}_{<i})} \Pr\left[X_i = x_i\right] \log\left(\tfrac{\Pr[X_i = x_i | a, \boldsymbol{x}_{<i}]}{\Pr[X_i = x_i]}\right) \\
&\qquad = \Pr\left[X_i \in \mathcal{X}(a,\boldsymbol{x}_{<i})\right] \sum_{x_i \in \mathcal{X}(a,\boldsymbol{x}_{<i})} \tfrac{\Pr[X_i = x_i]}{\Pr[X_i \in \mathcal{X}(a,\boldsymbol{x}_{<i})]} \log\left(\tfrac{\Pr[X_i = x_i | a, \boldsymbol{x}_{<i}]}{\Pr[X_i = x_i]}\right) \\
&\qquad \le \log\left(\tfrac{\Pr[X_i \in \mathcal{X}(a,\boldsymbol{x}_{<i}) | a, \boldsymbol{x}_{<i}]}{\Pr[X_i \in \mathcal{X}(a,\boldsymbol{x}_{<i})]}\right) = \log\left(\tfrac{1 - \Pr[X_i \notin \mathcal{X}(a,\boldsymbol{x}_{<i}) | a, \boldsymbol{x}_{<i}]}{1 - \Pr[X_i \notin \mathcal{X}(a,\boldsymbol{x}_{<i})]}\right)
\end{aligned}$$

where the inequality follows from Jensen's inequality. Note that, because $(a, \boldsymbol{x}_{<i}) \in \mathrm{E}_i(\widehat{\delta})$ for $\widehat{\delta} > 0$:

$$\Pr\left[X_i \notin \mathcal{X}(a,\boldsymbol{x}_{<i})\right] \le e^{3\epsilon} \Pr\left[X_i \notin \mathcal{X}(a,\boldsymbol{x}_{<i}) | a, \boldsymbol{x}_{<i}\right] + \widehat{\delta}.$$

We now focus on the term $\Pr\left[X_i \notin \mathcal{X}(a,\boldsymbol{x}_{<i}) | a, \boldsymbol{x}_{<i}\right]$. Note that $x_i \notin \mathcal{X}(a,\boldsymbol{x}_{<i}) \Leftrightarrow (a, \boldsymbol{x}_{\le i}) \notin \mathrm{F}_i$. Thus,

$$\Pr\left[X_i \notin \mathcal{X}(a,\boldsymbol{x}_{<i}) | a, \boldsymbol{x}_{<i}\right] = \Pr\left[(\mathcal{M}, \boldsymbol{X}_{\le i}) \notin \mathrm{F}_i | a, \boldsymbol{x}_{<i}\right] \overset{\text{def}}{=} q$$

Note that $q \le \delta''$ by Claim 4.2.5. Now, we can bound the following:

$$\begin{aligned}
&\sum_{x_i \in \mathcal{X}(a,\boldsymbol{x}_{<i})} \Pr\left[X_i = x_i\right] \log\left(\tfrac{\Pr[X_i = x_i | a, \boldsymbol{x}_{<i}]}{\Pr[X_i = x_i]}\right) \le \log(1-q) - \log(1 - (e^{3\epsilon} q + \widehat{\delta})) \\
&\qquad \le \log(e) \cdot (-q + e^{3\epsilon} q + \widehat{\delta} + 2(e^{3\epsilon} q + \widehat{\delta})^2) = \log(e) \cdot ((e^{3\epsilon} - 1) q + \widehat{\delta} + 2(e^{3\epsilon} q + \widehat{\delta})^2) \\
&\qquad \le \log(e) \cdot \left((e^{3\epsilon} - 1) \frac{2\widehat{\delta}}{1 - e^{-3\epsilon}} + \widehat{\delta} + 2\widehat{\delta}^2 \cdot \left(\frac{2e^{3\epsilon}}{1 - e^{-3\epsilon}} + 1 \right)^2 \right) \\
&\qquad = \widehat{\delta}(\log(e)(2e^{3\epsilon} + 1)) + \widehat{\delta}^2 \left(2\log(e) \left(\frac{4e^{12\epsilon} + 4e^{9\epsilon} - 3e^{6\epsilon} - 2e^{3\epsilon} + 1}{e^{6\epsilon} - 2e^{3\epsilon} + 1} \right) \right) \overset{\text{defn}}{=} \tau(\widehat{\delta})
\end{aligned}$$

where the second inequality follows by using the inequality $(-x - 2x^2)\log(e) \le \log(1-x) \le -x \log(e)$ for $0 < x \le 1/2$, and as $(e^{3\epsilon} q + \widehat{\delta}) \le 1/2$ for ϵ and $\widehat{\delta}$ bounded as in the lemma statement.

We then use this result to upper bound the expectation we wanted:

$$
\begin{aligned}
&\mathbb{E}\left[Z_i(\mathcal{M},\boldsymbol{X}_{\leq i})|a,\boldsymbol{x}_{<i},\mathrm{F}_i\right] \\
&\quad\leq \sum_{x_i\in\mathcal{X}(a,\boldsymbol{x}_{<i})} \Pr\left[X_i=x_i|a,\boldsymbol{x}_{<i},\mathrm{F}_i\right]\log\left(\tfrac{\Pr[X_i=x_i|a,\boldsymbol{x}_{<i}]}{\Pr[X_i=x_i]}\right) \\
&\qquad\quad - \sum_{x_i\in\mathcal{X}(a,\boldsymbol{x}_{<i})} \Pr\left[X_i=x_i\right]\log\left(\tfrac{\Pr[X_i=x_i|a,\boldsymbol{x}_{<i}]}{\Pr[X_i=x_i]}\right)+\tau(\widehat{\delta}) \\
&\quad= \sum_{x_i\in\mathcal{X}(a,\boldsymbol{x}_{<i})} \left(\Pr\left[X_i=x_i|a,\boldsymbol{x}_{<i},\mathrm{F}_i\right]-\Pr\left[X_i=x_i\right]\right)\log\left(\tfrac{\Pr[X_i=x_i|a,\boldsymbol{x}_{<i}]}{\Pr[X_i=x_i]}\right)+\tau(\widehat{\delta}) \\
&\quad\leq 6\epsilon \sum_{x_i\in\mathcal{X}(a,\boldsymbol{x}_{<i})} \left|\Pr\left[X_i=x_i|a,\boldsymbol{x}_{<i},\mathrm{F}_i\right]-\Pr\left[X_i=x_i\right]\right|+\tau(\widehat{\delta}) \\
&\quad\leq \tau(\widehat{\delta})+6\epsilon \sum_{x_i\in\mathcal{X}(a,\boldsymbol{x}_{<i})} \Pr\left[X_i=x_i\right]\cdot \\
&\qquad\qquad \max\left\{\frac{e^{6\epsilon}}{\Pr\left[(\mathcal{M},\boldsymbol{X}_{\leq i})\in\mathrm{F}_i|a,\boldsymbol{x}_{<i}\right]}-1,1-\frac{e^{-6\epsilon}}{\Pr\left[(\mathcal{M},\boldsymbol{X}_{\leq i})\in\mathrm{F}_i|a,\boldsymbol{x}_{<i}\right]}\right\} \\
&\quad\leq 6\epsilon\left(\frac{e^{6\epsilon}}{1-\frac{2\widehat{\delta}}{1-e^{-3\epsilon}}}-1\right)+\tau(\widehat{\delta})\leq 6\epsilon\left(e^{6\epsilon}\left(1+\tfrac{4\widehat{\delta}}{1-e^{-3\epsilon}}\right)-1\right)+\tau(\widehat{\delta}) \\
&\quad\leq 72\epsilon^2+\widehat{\delta}\left(\tfrac{24e^{6\epsilon}}{1-e^{-3\epsilon}}+\log(e)(2e^{3\epsilon}+1)\right) \\
&\qquad\qquad +\widehat{\delta}^2\left(2\log(e)\left(\frac{4e^{12\epsilon}+4e^{9\epsilon}-3e^{6\epsilon}-2e^{3\epsilon}+1}{e^{6\epsilon}-2e^{3\epsilon}+1}\right)\right) \\
&\quad\stackrel{\text{defn}}{=} \nu(\widehat{\delta}) \qquad\qquad (4.7)
\end{aligned}
$$

where the third inequality follows from the definition of F_i, the fourth inequality follows from Claim 4.2.5 and the last inequality follows by substituting the value of $\tau(\widehat{\delta})$, and using the inequalities $1+y\leq e^y$ and $e^{ky}\leq 1+e^k y$ for $y\in(0,0.5], k>1$. $\square$

Finally, we need to apply Azuma's inequality (stated in Theorem 4.1.6) to a set of variables that are bounded with probability 1, not just with high probability. Towards this end, we define variables T_i that will match Z_i for "good events", and will be zero otherwise—and hence, are always bounded:

$$
T_i(a,\boldsymbol{x}_{\leq i}) \stackrel{\text{defn}}{=} \begin{cases} Z_i(a,\boldsymbol{x}_{\leq i}) & \text{if } (a,\boldsymbol{x}_{\leq i})\in \mathrm{G}_{\leq i}(\widehat{\delta}) \\ 0 & \text{otherwise} \end{cases} \qquad (4.8)
$$

The next lemma verifies that the variables T_i indeed satisfy the requirements of Azuma's inequality:

Lemma 4.2.8. *Let $\mathcal{M}$ be (ϵ,δ)-differentially private and $\boldsymbol{X}\sim\mathcal{D}^n$. The variables T_i defined in* (4.8) *are bounded by 6ϵ with probability* 1, *and for any $(a,\boldsymbol{x}_{<i})\in\mathcal{Y}\times\mathcal{X}^{i-1}$ and $\widehat{\delta}\in$*

$[0, \epsilon/15]$,

$$\mathbb{E}\left[T_i(\mathcal{M}, \boldsymbol{X}_{\leq i}) | a, \boldsymbol{x}_{<i}\right] = O(\epsilon^2 + \widehat{\delta}/\epsilon), \tag{4.9}$$

More precisely, $\mathbb{E}\left[T_i(\mathcal{M}, \boldsymbol{X}_{\leq i}) | a, \boldsymbol{x}_{<i}\right] \leq \nu(\widehat{\delta})$*, where* $\nu(\widehat{\delta})$ *is defined in* (4.7)*.*

We can then apply Azuma's inequality to the sum of $T_i(a, \boldsymbol{x}_{\leq i})$, where each term will match $Z_i(a, \boldsymbol{x}_{\leq i})$ for most $(a, \boldsymbol{x}_{\leq i})$ coming from $(\mathcal{M}(\boldsymbol{X}), \boldsymbol{X}_{\leq i})$ for each $i \in [n]$. Note that, from Lemma 4.2.2, we know that a bound on $\sum_{i=1}^n Z_i(a, \boldsymbol{x}_{\leq i})$ with high probability will give us a bound on approximate max-information. We then give the formal analysis below.

Proof of Lemma 4.2.8. By definition, $T_i(\mathcal{M}, \boldsymbol{X}_{\leq i})$ takes values only in $[-6\epsilon, 6\epsilon]$. Thus,

$$\Pr\left[|T_i(\mathcal{M}, \boldsymbol{X}_{\leq i})| \leq 6\epsilon\right] = 1.$$

Now, given $(a, \boldsymbol{x}_{<i}) \in \mathrm{E}_i(\widehat{\delta}) \cap \mathrm{G}_{<i}(\widehat{\delta})$, we can see that:

$$\mathbb{E}\left[T_i(\mathcal{M}, \boldsymbol{X}_{\leq i}) \middle| a, \boldsymbol{x}_{<i}, \mathrm{G}^{\complement}_{\leq i}(\widehat{\delta})\right] = 0.$$

Further, given $(a, \boldsymbol{x}_{<i}) \in \mathrm{E}_i(\widehat{\delta}) \cap \mathrm{G}_{\leq i-1}(\widehat{\delta})$, we have:

$$\begin{aligned}
\mathbb{E}\left[T_i(\mathcal{M}, \boldsymbol{X}_{\leq i}) | a, \boldsymbol{x}_{<i}, \mathrm{G}_{\leq i}(\widehat{\delta})\right] &= \sum_{x_i : (a, \boldsymbol{x}_{\leq i}) \in \mathrm{F}_i} T_i(a, \boldsymbol{x}_{\leq i}) \Pr\left[X_i = x_i | a, \boldsymbol{x}_{<i}, \mathrm{G}_{\leq i}(\widehat{\delta})\right] \\
&= \sum_{x_i : (a, \boldsymbol{x}_{\leq i}) \in \mathrm{F}_i} Z_i(a, \boldsymbol{x}_{\leq i}) \Pr\left[X_i = x_i | a, \boldsymbol{x}_{<i}, \mathrm{G}_{\leq i}(\widehat{\delta})\right] \\
&= \sum_{x_i : (a, \boldsymbol{x}_{\leq i}) \in \mathrm{F}_i} Z_i(a, \boldsymbol{x}_{\leq i}) \Pr\left[X_i = x_i | a, \boldsymbol{x}_{<i}, \mathrm{F}_i\right] \\
&= \mathbb{E}\left[Z_i(\mathcal{M}, \boldsymbol{X}_{\leq i}) | a, \boldsymbol{x}_{<i}, \mathrm{F}_i\right] = O(\epsilon^2 + \widehat{\delta}/\epsilon)
\end{aligned}$$

where the second equality follows from (4.8), and the last equality follows from Lemma 4.2.7. For any $(a, \boldsymbol{x}_{<i}) \notin \mathrm{E}_i(\widehat{\delta}) \cap \mathrm{G}_{\leq i-1}(\widehat{\delta})$, we have that the conditional expectation is zero. This proves the lemma. □

We are now ready to prove our main theorem.

Proof of Theorem 4.2.1. For any constant ν, we have:

$$\begin{aligned}
&\Pr\left[\sum_{i=1}^{n} Z_i(\mathcal{M}, \boldsymbol{X}_{\le i}) > n\nu + 6t\epsilon\sqrt{n}\right] \\
&\le \Pr\left[\sum_{i=1}^{n} Z_i(\mathcal{M}, \boldsymbol{X}_{\le i}) > n\nu + 6t\epsilon\sqrt{n} \cap (\mathcal{M}, \boldsymbol{X}) \in \mathrm{G}_{\le n}(\widehat{\delta})\right] + \Pr\left[(\mathcal{M}, \boldsymbol{X}) \notin \mathrm{G}_{\le n}(\widehat{\delta})\right] \\
&= \Pr\left[\sum_{i=1}^{n} T_i(\mathcal{M}, \boldsymbol{X}_{\le i}) > n\nu + 6t\epsilon\sqrt{n} \cap (\mathcal{M}, \boldsymbol{X}) \in \mathrm{G}_{\le n}(\widehat{\delta})\right] + \Pr\left[(\mathcal{M}, \boldsymbol{X}) \notin \mathrm{G}_{\le n}(\widehat{\delta})\right]
\end{aligned}$$

We then substitute ν by $\nu(\widehat{\delta})$ as defined in Equation (4.7), and apply a union bound on $\Pr\left[(\mathcal{M}, \boldsymbol{X}) \notin \mathrm{G}_{\le n}(\widehat{\delta})\right]$ using Claim 4.2.6 to get

$$\begin{aligned}
\Pr\left[\sum_{i=1}^{n} Z_i(\mathcal{M}, \boldsymbol{X}_{\le i}) > n\nu(\widehat{\delta}) + 6t\epsilon\sqrt{n}\right] &\le \Pr\left[\sum_{i=1}^{n} T_i(\mathcal{M}, \boldsymbol{X}_{\le i}) > n\nu(\widehat{\delta}) + 6t\epsilon\sqrt{n}\right] + n(\delta' + \delta'') \\
&\le e^{-t^2/2} + n(\delta' + \delta'')
\end{aligned}$$

where the two inequalities follow from Claim 4.2.6 and Theorem 4.1.6, respectively. Therefore,

$$\Pr\left[Z(\mathcal{M}(\boldsymbol{X}), \boldsymbol{X}) > n\nu(\widehat{\delta}) + 6t\epsilon\sqrt{n}\right] \le e^{-t^2/2} + n(\delta' + \delta'') \overset{\text{defn}}{=} \beta(t, \widehat{\delta})$$

From Lemma 4.2.2, we have $I_\infty^{\beta(t,\widehat{\delta})}(\boldsymbol{X}; \mathcal{M}(\boldsymbol{X})) \le n\nu(\widehat{\delta}) + 6t\epsilon\sqrt{n}$. We set the parameters $t = \epsilon\sqrt{2n}$ and $\widehat{\delta} = \sqrt{\epsilon\delta}/15$ to obtain our result. Note that setting $\widehat{\delta} = \sqrt{\epsilon\delta}/15$ does not violate the bounds on it stated in the statement of Lemma 4.2.7. $\square$

4.3. Comparison with Results from Bassily et al. (2016)

In this section, we use the bound from our main theorem of this chapter (Theorem 4.2.1) to rederive known results for the generalization properties of differentially private algorithms which select *low sensitivity queries*. Our bounds nearly – but not exactly – match the tight bounds for this problem, given in Bassily et al. (2016). This implies a limit on the extent to which our main theorem can be quantitatively improved, despite its generality.

The goal of generalization bounds for low sensitivity queries is to argue that with high probability, if a low sensitivity function $f : \mathcal{X}^n \to \mathcal{R} = \mathcal{M}(\boldsymbol{x})$ is chosen in a data-dependent way when $\boldsymbol{X}$ is sampled from a product distribution $\mathcal{D}^n$, then the value of the function on the realized data $f(\boldsymbol{x})$ is close to its expectation $f(\mathcal{D}^n) \overset{\text{def}}{=} \mathbb{E}_{\boldsymbol{X}\sim\mathcal{D}^n}[f(\boldsymbol{X})]$. If f were selected in a data-independent manner, this would follow from McDiarmid's inequality. No bound like this is true for arbitrary selection procedures $\mathcal{M}$, but a tight bound by Bassily et al. (2016) is known when $\mathcal{M}$ is (ϵ, δ)-differentially private – which we present in Theorem A.1.3.

Using our main theorem (Theorem 4.2.1), together with McDiarmid's inequality (Theorem A.1.1), we can derive a comparable statement to Theorem A.1.3:

Theorem 4.3.1. *Let* $\epsilon \in (0,1)$, $\delta = O(\epsilon^5)$, *and* $n = \Omega\left(\frac{\log(\epsilon/\delta)}{\epsilon^2}\right)$. *Let* $\mathcal{Y}$ *denote the class of* Δ*-sensitive functions* $f : \mathcal{X}^n \to \mathbb{R}$, *and let* $\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}$ *be an algorithm that is* (ϵ, δ)*-differentially private. Let* $\boldsymbol{X} \sim \mathcal{D}^n$ *for some distribution* $\mathcal{D}$ *over* $\mathcal{X}$, *and let* $\phi = \mathcal{M}(\boldsymbol{X})$. *Then there exists a constant* c *such that:*

$$\Pr_{\boldsymbol{X} \sim \mathcal{D}^n, \mathcal{M}} \left[|\phi(\mathcal{D}^n) - \phi(\boldsymbol{X})| \geq c\ \epsilon \Delta n\right] < n\sqrt{\frac{\delta}{\epsilon}}$$

Proof. If $\mathcal{M}$ satisfied $I_\infty^\beta(\boldsymbol{X}; \mathcal{M}(\boldsymbol{X})) \leq m$, then McDiarmid's inequality (Theorem A.1.1) paired with Definition 1.4.1 would imply:

$$\Pr_{\boldsymbol{X}, \mathcal{M}} \left[|\phi(\mathcal{D}^n) - \phi(\boldsymbol{X})| \geq c\ \epsilon \Delta n\right] < 2^k \exp\left(-2c^2\ \epsilon^2 n\right) + \beta$$

Because $\mathcal{M}$ is (ϵ, δ)-differentially private, Theorem 4.2.1 implies that indeed $I_\infty^\beta(\boldsymbol{X}; \mathcal{M}(\boldsymbol{X})) \leq m$ for $m = O\left(n\epsilon^2 + n\sqrt{\frac{\delta}{\epsilon}}\right) = O\left(\epsilon^2 n\right)$ and $\beta = O\left(n\sqrt{\frac{\delta}{\epsilon}}\right) + e^{-\epsilon^2 n} = O\left(n\sqrt{\frac{\delta}{\epsilon}}\right)$. Hence, the claimed bound follows. ☐

Because Theorem A.1.3 is asymptotically tight, this implies a limit on the extent to which Theorem 4.2.1 can be quantitatively improved.

For comparison, note that the generalization bound for (ϵ, δ)-differentially private mechanisms given in Theorem A.1.3 differs by only constants from the generalization bound proven via the max-information approach for (ϵ, δ')-differentially private mechanisms, where:

$$\delta' = \frac{\delta^2}{\epsilon n^2}$$

Note that in most applications (including the best known mechanism for answering large numbers of low sensitivity queries privately— the median mechanism Roth and Roughgarden (2010) as analyzed in Dwork and Roth (2014) Theorem 5.10), the accuracy of a differentially private algorithm scales with $\sqrt{\frac{\log(1/\delta)}{n}}$ (ignoring other relevant parameters). In such cases, using the bound derived from the max-information approach yields an accuracy that is worse than the bound from Theorem A.1.3 by an additive term that is $O\left(\sqrt{\frac{\log(\epsilon n)}{n}}\right)$.

4.4. A Counterexample to Nontrivial Composition and a Lower Bound for Non-Product Distributions

It is known that algorithms with bounded description length have bounded approximate max-information (Dwork et al., 2015a). In Section 4.2, we showed that (ϵ, δ)-differentially private algorithms have bounded approximate max-information when the dataset is drawn from a product distribution. In this section, we show that although approximate max-information composes adaptively (Dwork et al., 2015a), one cannot always run a bounded description length algorithm, followed by an approximately differentially private algorithm, and expect the resulting composition to have strong generalization guarantees. In particular, this implies that (ϵ, δ)-differentially private algorithms cannot have any nontrivial bounded max-information guarantee over non-product distributions.

Specifically, we give an example of a pair of algorithms $\mathcal{M}_1$ and $\mathcal{M}_2$ such that $\mathcal{M}_1$ has output description length $o(n)$ for inputs of length n, and $\mathcal{M}_2$ is (ϵ, δ)-differentially private, but the adaptive composition of $\mathcal{M}_1$ followed by $\mathcal{M}_2$ can be used to exactly reconstruct the input database with high probability. In particular, it is easy to overfit to the input $\boldsymbol{X}$ given $\mathcal{M}_2(\boldsymbol{X}; \mathcal{M}_1(\boldsymbol{X}))$, and hence, no nontrivial generalization guarantees are possible. Note that this does not contradict our results on the max-information of differentially private algorithms for *product distributions*: even if the database used as input to $\mathcal{M}$ is drawn from a product distribution, the distribution on the database is no longer a product distribution *once conditioned on the output of* $\mathcal{M}$. The distribution of $\mathcal{M}$'s input violates the hypothesis that is used to prove a bound on the max-information of $\mathcal{M}$.

Our construction adapts De's idea (De, 2012). Given as input a *uniformly random* dataset $\boldsymbol{X}$, we show a mechanism $\mathcal{M}_1$ which outputs a short description of a code that contains $\boldsymbol{X}$. Because this description is short, $\mathcal{M}_1$ has small max-information. The mechanism $\mathcal{M}_2$ is then parameterized by this short description of a code. Given the description of a code and the dataset $\boldsymbol{X}$, $\mathcal{M}_2$ approximates (privately) the distance from $\boldsymbol{X}$ to the nearest *codeword*, and outputs that codeword when the distance is small. When $\mathcal{M}_2$ is composed with $\mathcal{M}_1$, we show that it outputs the dataset $\boldsymbol{X}$ with high probability.

Theorem 4.4.1. *Let* $\mathcal{X} = \{0,1\}$ *and* $\mathcal{Y} = \{\mathcal{X}^n \cup \{\perp\}\}$. *Let* $\boldsymbol{X}$ *be a uniformly distributed random variable over* $\mathcal{X}^n$. *For* $n > 64e$, *for every* $\epsilon \in \left(0, \frac{1}{2}\right], \delta \in \left(0, \frac{1}{4}\right]$, *there exists an integer* $r > 0$ *and randomized algorithms* $\mathcal{M}_1 : \mathcal{X}^n \to \{0,1\}^r$, *and* $\mathcal{M}_2 : \mathcal{X}^n \times \{0,1\}^r \to \mathcal{Y}$, *such that:*

1. $r = O\left(\dfrac{\log(1/\delta)\log n}{\epsilon}\right)$ *and* $I_\infty^\beta\left(\boldsymbol{X}; \mathcal{M}_1(\boldsymbol{X})\right) \leq r + \log(\frac{1}{\beta})$ *for all* $\beta > 0$*;*

2. *for every* $\boldsymbol{a} \in \{0,1\}^r$, $\mathcal{M}_2(\boldsymbol{X}, \boldsymbol{a})$ *is* (ϵ, δ)*-differentially private and* $I_\infty^\beta\left(\boldsymbol{X}; \mathcal{M}_2(\boldsymbol{X}, \boldsymbol{a})\right) \leq 1$ *for all* $\beta \geq 2\delta$*;*

3. *for every* $\boldsymbol{x} \in \mathcal{X}^n$, *with probability at least* $1 - \delta$, *we have that* $\mathcal{M}_2(\boldsymbol{x}; \mathcal{M}_1(\boldsymbol{x}))) = \boldsymbol{x}$. *In particular,* $I_\infty^\beta\left(\boldsymbol{X}; \mathcal{M}_2(\boldsymbol{X}; \mathcal{M}_1(\boldsymbol{X}))\right) \geq n - 1$ *for all* $0 < \beta \leq \frac{1}{2} - \delta$.

De (2012) showed that the *mutual information* of (ϵ, δ)-differentially private protocols can be large: if $\frac{1}{\epsilon}\log\left(\frac{1}{\delta}\right) = O(n)$, then there exists an (ϵ, δ)-differentially private algorithm $\mathcal{M}$ and a distribution $\mathcal{S}$ such that for $\boldsymbol{X} \sim \mathcal{S}$, $I\left(\boldsymbol{X}; \mathcal{M}(\boldsymbol{X})\right) = \Omega(n)$, where I denotes mutual information. De's construction also has large approximate max-information.

By the composition theorem for approximate max-information (given in Theorem 1.4.2), our construction implies a similar bound:

Corollary 4.4.2. *There exists an* (ϵ, δ)*-differentially private mechanism* $\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}$ *such* *that* $I_{\infty}^{\beta_2}(\mathcal{M}, n) \geq n - 1 - r - \log(1/\beta_1)$ *for all* $\beta_1 \in (0, 1/2 - \delta)$ *and* $\beta_2 \in (0, 1/2 - \delta - \beta_1)$, *where* $r = O\left(\frac{\log(1/\delta)\log(n)}{\epsilon}\right)$.

Proof. We use the same algorithms $\mathcal{M}_1$ and $\mathcal{M}_2$ from Theorem 4.4.1. Suppose that for all $\boldsymbol{a} \in \{0,1\}^r$ and $0 < \beta_2 < 1/2 - \delta - \beta_1$ for any $\beta_1 \in (0, 1/2 - \delta)$, we have:

$$I_{\infty}^{\beta_2}(\mathcal{M}_2(\cdot, \boldsymbol{a}), n) < n - 1 - r - \log(1/\beta_1).$$

Note that because $\mathcal{M}_1$ has bounded description length r, we can bound $I_{\infty}^{\beta_1}(\mathcal{M}_1, n) \leq r + \log(1/\beta_1)$ for any $\beta_1 > 0$ (Dwork et al., 2015a). We then apply the composition theorem for max-information mechanisms, Theorem 1.4.2, to obtain:

$$I_{\infty}^{\beta_1+\beta_2}(\mathcal{M}_2 \circ \mathcal{M}_1, n) < n - 1.$$

However, this contradicts Theorem 4.4.1, because for any $\beta < 1/2 - \delta$,

$$I_{\infty}^{\beta}(\mathcal{M}_2 \circ \mathcal{M}_1, n) \geq I_{\infty,\Pi}^{\beta}(\mathcal{M}_2 \circ \mathcal{M}_1, n) \geq n - 1.$$

Thus, we know that there exists some $\boldsymbol{a}^* \in \{0,1\}^r$ and (non-product) distribution $\boldsymbol{X} \sim \mathcal{S}$ such that:

$$I_{\infty}^{\beta_2}(\boldsymbol{X}; \mathcal{M}_2(\boldsymbol{X}, \boldsymbol{a}^*)) \geq n - 1 - r - \log(1/\beta_1).$$

We then define $\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}$ to be $\mathcal{M}(\boldsymbol{x}) = \mathcal{M}_2(\boldsymbol{x}, \boldsymbol{a}^*)$. Hence,

$$I_{\infty}^{\beta_2}(\mathcal{M}, n) \geq I_{\infty}^{\beta_2}(\boldsymbol{X}; \mathcal{M}(\boldsymbol{X})) \geq n - 1 - r - \log(1/\beta_1)$$

which completes the proof. $\square$

We adapt ideas from De's construction (De, 2012) in order to prove Theorem 4.4.1. In De's construction, the input is not drawn from a product distribution—instead, the support of the input distribution is an error-correcting code, meaning that all points in the support are far from each other in Hamming distance. For such a distribution, De showed that adding the level of noise required for differential privacy does not add enough distortion to prevent decoding of the dataset.

Our construction adapts De's idea. Given as input a *uniformly random* dataset $\boldsymbol{X}$, we show a mechanism $\mathcal{M}_1$ which outputs a short description of a code that contains $\boldsymbol{X}$. Because this description is short, $\mathcal{M}_1$ has small max-information. The mechanism $\mathcal{M}_2$ is then parameterized by this short description of a code. Given the description of a code and the dataset $\boldsymbol{X}$, $\mathcal{M}_2$ approximates (privately) the distance from $\boldsymbol{X}$ to the nearest *codeword*, and outputs that codeword when the distance is small. When $\mathcal{M}_2$ is composed with $\mathcal{M}_1$, we show that it outputs the dataset $\boldsymbol{X}$ with high probability. We present the formal analysis in Appendix A.3.

4.5. Consequences of Lower Bound Result - Robust Generalization

We use this section to give a consequence of Theorem 4.4.1. We start by showing that it is possible to also output a dataset of dimensionality $d \geq 1$ when an algorithm with bounded description length is used first, followed by an approximate differentially private algorithm.

Corollary 4.5.1. *Let $\mathcal{X} = \{0,1\}^d$ and $\epsilon, \delta = O(1)$. There exists an algorithm $\mathcal{M}_1 : \mathcal{X}^n \to \{0,1\}^r$ where $r = O\left(\frac{d^{3/2}\log(d/\delta)\sqrt{\log(1/\delta)}\log(n)}{\epsilon}\right)$ and $\mathcal{M}_2 : \mathcal{X}^n \times \{0,1\}^r \to \{\mathcal{X}^n \cup \{\perp\}\}$ that is (ϵ, δ)-DP in its first argument such that for each $\boldsymbol{x} \in \mathcal{X}^n$, the mapping $\mathcal{M}_2(\boldsymbol{x}, \mathcal{M}_1(\boldsymbol{x}))$ outputs $\boldsymbol{x}$ with probability at least $1 - \delta$.* [1]

Proof. We will write $\boldsymbol{x}^{(j)} \in \{0,1\}^n$ to denote the vector of n data points in the jth entry of universe $\mathcal{X}$, so that dataset $\boldsymbol{x} = (\boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}^{(d)}) \in \mathcal{X}^n$. From Theorem 4.4.1, we know that for each $j \in [d]$, we have for every $\epsilon' \in (0, 1/2]$, $\delta' \in (0, 1/4]$ and $r' = O\left(\log(1/\delta')\log(n)/\epsilon'\right)$, there exists an algorithm $\mathcal{M}_1^{(j)} : \{0,1\}^n \to \{0,1\}^{r'}$ and $\mathcal{M}_2^{(j)} : \mathcal{X}^n \times \{0,1\}^{r'} \to \{\{0,1\}^n \cup \{\perp\}\}$ such that $\mathcal{M}_2^{(j)}$ is (ϵ', δ')-DP in its first argument and for each $\boldsymbol{y} \in \{0,1\}^n$, with probability at least $1 - \delta'$, $\mathcal{M}_2^{(j)}(\boldsymbol{y}, \mathcal{M}_1^{(j)}(\boldsymbol{y})) = \boldsymbol{y}$. We can apply such an argument for each entry $j \in [d]$, and define the algorithm $\mathcal{M}_1 : \mathcal{X}^n \to \{0,1\}^{dr'}$ as

$$\mathcal{M}_1(\boldsymbol{x}) = \left(\mathcal{M}_1^{(1)}(\boldsymbol{x}^{(1)}), \cdots, \mathcal{M}_1^{(d)}(\boldsymbol{x}^{(d)})\right).$$

Further, by Theorem 2.1.5, we know that for each $\boldsymbol{a} = (\boldsymbol{a}^{(1)}, \cdots, \boldsymbol{a}^{(d)}) \in \{0,1\}^{dr'}$ where $\boldsymbol{a}^{(j)} = \mathcal{M}_1^{(j)}(\boldsymbol{x}_j)$, the algorithm $\mathcal{M}_2(\cdot, \boldsymbol{a})$ is $(d\epsilon'^2/2 + \epsilon'\sqrt{2d\log(1/\widehat{\delta})}, \widehat{\delta} + d\delta')$-DP for $\widehat{\delta} > 0$, where

$$\mathcal{M}_2(\boldsymbol{x}, \boldsymbol{a}) = \left(\mathcal{M}_2^{(1)}(\boldsymbol{x}^{(1)}, \boldsymbol{a}^{(1)}), \cdots, \mathcal{M}_2^{(d)}(\boldsymbol{x}^{(d)}, \boldsymbol{a}^{(d)})\right).$$

Further, from Theorem 4.4.1, we know that for each $\boldsymbol{x} \in \mathcal{X}^n$, we have $\mathcal{M}_2(\boldsymbol{x}, \mathcal{M}_1(\boldsymbol{x})) = \boldsymbol{x}$ with probability at least $1 - d\delta'$. We then set $\widehat{\delta} = \delta/2$, $\delta' = \delta/(2d)$, and $\epsilon' = O\left(\frac{\epsilon}{\sqrt{d\log(1/\delta)}}\right)$ $\square$

Cummings et al. (2016a) also consider validity in adaptive data analysis, but from the context of generalization in learning. Hence, to fit our results into their setting, we present some concepts from learning theory. We will denote a *hypothesis* $h : \mathcal{X} \to \{0,1\}$ to be a boolean map from the data universe where $h(\boldsymbol{x}) = 1$ denotes a positive instance and $h(\boldsymbol{x}) = 0$ is a negative instance. We will consider the data universe to be $\mathcal{X} = \{0,1\}^d$ where d is the dimensionality. We then write the labeled data $\mathcal{X}_L = \mathcal{X} \times \{0,1\}$ and the corresponding labeled data distribution as $\mathcal{D}_L^n$. Thus, a labeled dataset $\boldsymbol{X}_L$ is sampled from $\mathcal{D}_L^n$. A learning algorithm then takes a labeled sample $\boldsymbol{X}_L$ as input and outputs a hypothesis h. Our goal is to ultimately have a learning algorithm that finds a hypothesis whose empirical error matches the true, underlying error on the data distribution. We then define the *empirical error* to be,

---

[1] Thanks to Kobbi Nissim to pointing out this implication.

$$\text{error}(h, \boldsymbol{X}_L) = \frac{1}{n} \sum_{(x,y) \in \boldsymbol{X}_L} \mathbb{1}\{h(x) \neq y\}.$$

We also define the true error to be,

$$\text{error}(h, \mathcal{D}_L) = \Pr_{(x,y) \in \mathcal{D}_L} \left[h(x) \neq y \right].$$

We can then model the traditional notion of generalization.

Definition 4.5.2. *An algorithm $\mathcal{M} : \mathcal{X}_L^n \to \{\mathcal{X} \to \{0,1\}\}$ is (τ, β)-generalizing if for all distributions $\mathcal{D}_L$, given a labeled sample $\boldsymbol{X}_L \sim \mathcal{D}_L^n$,*

$$\Pr\left[\mathcal{M}(\boldsymbol{X}_L) = h \quad s.t. \ |\text{error}(h, \boldsymbol{X}_L) - \text{error}(h, \mathcal{D}_L)| \leq \tau\right] \geq 1 - \beta.$$

where the probability is over the randomness in $\mathcal{M}$ and the data generation.

As Cummings et al. (2016a) point out, this notion of generalization allows for the algorithm to provide information that would help overfit the sample it is given, which we want to prevent. Hence, they present robust generalization which guarantees that no adversary (or analyst) can take the output hypothesis of the learning algorithm as input and find a new hypothesis whose empirical error is entirely different from the true error. Note that they extend the error guarantee presented above to be in terms of the difference between $h(\boldsymbol{X}_L) = \frac{1}{n}\sum_{(x,y) \in \boldsymbol{X}_L} h(x,y)$ and $h(\mathcal{D}_L) = \mathbb{E}_{(x,y) \sim \mathcal{D}_L^n}\left[h(x,y)\right]$ where $h(x,y) \in \{0,1\}$, e.g. $h(x,y) = \mathbb{1}\{h(x) = y\}$ which recovers the guarantee in traditional generalization.

Definition 4.5.3. *An algorithm $\mathcal{M} : \mathcal{X}_L^n \to \mathcal{Y}$ for arbitrary range $\mathcal{Y}$ is (τ, β)-robustly generalizing (RG) if for all distributions $\mathcal{D}_L$ over $\mathcal{X}_L$ and any adversary $\mathcal{A}$ with probability at least $1 - \zeta$ over choice of $\boldsymbol{X}_L \sim \mathcal{D}_L^n$,*

$$\Pr_{\mathcal{A},\mathcal{M}}\left[\mathcal{A}\left(\mathcal{M}(\boldsymbol{X}_L)\right) = h \quad s.t. \ |h(\boldsymbol{X}_L) - h(\mathcal{D}_L)| \leq \tau\right] \geq 1 - \gamma.$$

for some ζ, γ where $\beta = \zeta + \gamma$.

We then give the robust generalization guarantees of differentially private and bounded description length algorithms. The stated results were taken from Cummings et al. (2016a), but follow directly from Bassily et al. (2016); Dwork et al. (2015a), respectively.

Theorem 4.5.4 [Cummings et al. (2016a)]**.** *Let $\mathcal{M} : \mathcal{X}_L^n \to \mathcal{Y}$ be a (ϵ, δ)-DP algorithm and $n \geq O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$ for $\delta > 0$. Then $\mathcal{M}$ is $(O(\epsilon), O(\delta/\epsilon)))$-RG.*

Theorem 4.5.5 [Cummings et al. (2016a)]**.** *Let $\mathcal{M} : \mathcal{X}_L^n \to \mathcal{Y}$ be an algorithm where $|\mathcal{Y}| < \infty$. Then $\mathcal{M}$ is $\left(\sqrt{\frac{\log(|\mathcal{Y}|/\beta)}{2n}}, \beta\right)$-RG for any $\beta > 0$*

We then show that two RG algorithms do not compose in general. Note that the following result requires that the dimensionality of the data d grow with n, so that $n \ll 2^d$. Otherwise, we would have enough samples so that the empirical average for *any* hypothesis $h : \mathcal{X} \to [0,1]$ would be close to the true mean $h(\mathcal{D})$.

Theorem 4.5.6. *Let* $d = o\left(\left(\frac{n}{\log(n)}\right)^{2/3}\right)$*. There exists a pair of algorithms* $\mathcal{M}_1 : \mathcal{X}_L^n \to \mathcal{Y}_1$*, and* $\mathcal{M}_2 : \mathcal{X}_L^n \times \mathcal{Y}_1 \to \mathcal{Y}_2$ *such that* $\mathcal{M}_1$ *is* $(o(1), o(1))$*-RG and* $\mathcal{M}_2(\cdot, y)$ *is* $(o(1), o(1))$*-RG for each* $y \in \mathcal{Y}_1$*, but the algorithm* $\mathcal{M} = \mathcal{M}_2 \circ \mathcal{M}_1 : \mathcal{X}_L^n \times \{0,1\} \to \mathcal{Y}_2$ *is not* $(1 - n2^{-d}, o(1))$*-RG.*

Proof. We use Corollary 4.5.1 to prove our result with $\mathcal{X}_L = \{0,1\}^{d+1}$. So we let $\mathcal{M}_1 : \mathcal{X}_L^n \to \{0,1\}^r$ and $\mathcal{M}_2 : \mathcal{X}_L^n \times \{0,1\}^r \to \mathcal{X}^n$ from Corollary 4.5.1 so that

$$r = O\left(\frac{d^{3/2}\log(d/\delta)\sqrt{\log(1/\delta)}\log(n)}{\epsilon}\right)$$

and $\mathcal{M}_2$ is (ϵ, δ)-DP in its first argument (including the label). We then pick

$$\epsilon = \Theta\left(\left(\frac{d^{3/2}\log(d/\delta)\sqrt{\log(1/\delta)}\log(n)}{n}\right)^{1/3}\right)$$

It is then clear that for such an ϵ we have $n \geq O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$. Thus, we have that $\mathcal{M}_1$ is $(O(\epsilon), \beta)$-RG and $\mathcal{M}_2(\cdot, y)$ is $(O(\epsilon), O(\delta/\epsilon))$-RG for each value y in its last component. We then set $1/\beta = \text{poly}(n)$ and $\delta = o(\epsilon)$. As long as $d = o\left(\left(\frac{n}{\log(n)}\right)^{2/3}\right)$ then we have $\epsilon = o(1)$.

Further, we know that for each $\boldsymbol{x}_L \in \mathcal{X}_L^n$, $\mathcal{M}_2(\boldsymbol{x}_L, \mathcal{M}_1(\boldsymbol{x}_L)) = \boldsymbol{x}_L$ with probability $1 - \delta$. Note that if $\mathcal{A}$ knows the entire dataset $\boldsymbol{x}_L^*$, then an overfitting hypothesis would be $h^* : \mathcal{X} \to \{0,1\}$ where $h^*(x) = 1$ if and only if $x \in \boldsymbol{x}^*$. We then define $h^*(\boldsymbol{x}_L^*) \overset{\text{defn}}{=} \frac{1}{n}\sum_{x \in \boldsymbol{x}^*} h^*(x) = 1$ and our data distribution to be uniform over $\mathcal{X}$. Thus, $h^*(\mathcal{D}_L^n) = \Pr_{X\sim\mathcal{D}}[X \in \boldsymbol{x}^*] = \frac{n}{2^d}$.

This gives us the following condition where for each $\boldsymbol{x}_L^* \in \mathcal{X}_L^n$,

$$\Pr\left[\mathcal{A}(\mathcal{M}(\boldsymbol{x}_L^*)) = h^* \quad \text{s.t. } |h^*(\boldsymbol{x}_L^*) - h^*(\mathcal{D}_L)| \geq 1 - \frac{n}{2^d}\right] \geq 1 - \delta.$$

Thus, $\mathcal{M}$ cannot be (τ, δ)-RG where $\tau \leq 1 - \frac{n}{2^d}$. ☐

4.6. Conversion between Mutual and Max-Information

In this section we present the general conversion between mutual information and max-information, along with some of its consequences.

Theorem 4.6.1. *Let* X, Y *be a pair of discrete random variables defined on the same probability space, where* X *takes values in a finite set* $\mathcal{X}$*.*

- *If* $I(X;Y) \leq t$ *then, for every* $m > 0$*, we have* $I_\infty^{\beta(m)}(X;Y) \leq m$ *where* $\beta(m) = \frac{t+0.54}{m}$*.*
- *If* $I_\infty^{\beta}(X;Y) \leq m$ *for* $m > 0$ *and* $0 \leq \beta \leq \frac{3(1-2^{-m})}{20}$*, then* $I(X;Y) \leq 2m\log(2) + \dfrac{2\beta\log_2(|\mathcal{X}|/2\beta)}{1-2^{-m}}$*.*

Proof. We first prove the first direction of the theorem. We define a "good" set $\mathcal{G}(k) \stackrel{\text{defn}}{=} \{(x,y) : Z(x,y) \leq k\}$ where $Z(x,y) = \log_2\left(\frac{\Pr[(X,Y)=(x,y)]}{\Pr[X \otimes Y=(x,y)]}\right)$, and $\beta(k)$ to be the quantity such that $\Pr\left[(X,Y) \in \mathcal{G}(k)\right] = 1 - \beta(k)$. We then have:

$$\begin{aligned} m &\geq \sum_{(x,y)\in\mathcal{G}(k)} \Pr\left[(X,Y)=(x,y)\right] Z(x,y) + \sum_{(x,y)\notin\mathcal{G}(k)} \Pr\left[(X,Y)=(x,y)\right] Z(x,y) \\ &\geq \sum_{(x,y)\in\mathcal{G}(k)} \Pr\left[(X,Y)=(x,y)\right] Z(x,y) + k\beta(k) \qquad (4.10) \end{aligned}$$

Also, we have:

$$-\sum_{(x,y)\in\mathcal{G}(k)} \frac{\Pr\left[(X,Y)=(x,y)\right]}{\Pr\left[(X,Y)\in\mathcal{G}(k)\right]} Z(x,y) \leq \log_2\left(\frac{\Pr\left[X\otimes Y \in \mathcal{G}(k)\right]}{\Pr\left[(X,Y)\in\mathcal{G}(k)\right]}\right) \leq \log_2\left(\frac{1}{1-\beta(k)}\right)$$

where the first inequality follows from applying Jensen's inequality.

By rearranging terms, we obtain:

$$\sum_{(x,y)\in\mathcal{G}(k)} \Pr\left[(X,Y)=(x,y)\right] \cdot Z(x,y) \geq -(1-\beta(k)) \cdot \log_2\left(\frac{1}{1-\beta(k)}\right).$$

Plugging the above in (4.10), we obtain:

$$m \geq -(1-\beta(k)) \log_2\left(\frac{1}{1-\beta(k)}\right) + k\beta(k).$$

Thus,

$$k \leq \frac{m + (1-\beta(k)) \log_2\left(\frac{1}{1-\beta(k)}\right)}{\beta(k)} \leq \frac{m + 0.54}{\beta(k)}$$

where the last inequality follows from the fact that the function $(1-w)\log_2(1/(1-w))$ is maximized at $w = (e-1)/e$ (and takes value < 0.54). Solving for $\beta(k)$ gives the claimed bound.

We now prove the second direction of the lemma. Note that we can use Lemma 4.1.4 to say that (X,Y) and $(X \otimes Y)$ are point-wise $(2k\log(2), \beta')$ indistinguishable, for $\beta' = \frac{2\beta}{1-2^{-k}}$. Note that $\beta' \leq 0.3$ as $\beta \in \left[0, \frac{3(1-2^{-k})}{20}\right]$. We now define the following "bad" set $\mathcal{B}$:

$$\mathcal{B} = \left\{(x,y) \in \mathcal{X} \times \mathcal{X} : \log_2\left(\frac{(X,Y)=(x,y)}{X\otimes Y=(x,y)}\right) > 2k\right\}.$$

From the definition of point-wise indistinguishability, we have:

$$I(X;Y) \leq 2k + \sum_{(x,y)\in\mathcal{B}} \Pr\left[(X,Y)=(x,y)\right]\log_2\left(\frac{\Pr\left[(X,Y)=(x,y)\right]}{\Pr\left[X=x\right]\Pr\left[Y=y\right]}\right). \tag{4.11}$$

Now, if set $\mathcal{B}$ is empty, we get from (4.11) that:

$$I(X;Y) \leq 2k$$

which trivially gives us the claimed bound.

However, if set $\mathcal{B}$ is non-empty, we then consider the mutual information conditioned on the event $(X,Y)\in\mathcal{B}$. We will write X' for the random variable distributed the same as X conditioned on $(X,Y)\in\mathcal{B}$, and Y' for the random variable Y conditioned on the same event. We can then obtain the following bound, where we write $H(W)$ to denote the entropy of random variable W:

$$I(X';Y') \leq H(X') \leq \log_2|\mathcal{X}|.$$

We can then make a relation between the mutual information $I(X',Y')$ and the sum of terms over $\mathcal{B}$ in (4.11). Note that, for $(x,y)\in\mathcal{B}$, we have from Bayes' rule:

$$\Pr\left[(X,Y)\in\mathcal{B}\right]\Pr\left[(X,Y)=(x,y)|\mathcal{B}\right] = \Pr\left[(X,Y)=(x,y)\right].$$

This then gives us the following bound:

$$\begin{aligned}\sum_{(x,y)\in\mathcal{B}} &\frac{\Pr\left[(X,Y)=(x,y)\right]}{\Pr\left[(X,Y)\in\mathcal{B}\right]}\log_2\left(\frac{\Pr\left[(X,Y)=(x,y)\right]}{\Pr\left[(X,\boldsymbol{Y})\in\mathcal{B}\right]\Pr\left[X=x|\mathcal{B}\right]\Pr\left[Y=y|\mathcal{B}\right]}\right)\\ &\leq \log_2|\mathcal{X}|.\end{aligned} \tag{4.12}$$

Note that $\Pr\left[X=x|\mathcal{B}\right] \leq \frac{\Pr[X=x]}{\Pr[(X,Y)\in\mathcal{B}]}$, and similarly, $\Pr\left[Y=y|\mathcal{B}\right] \leq \frac{\Pr[Y=y]}{\Pr[(X,Y)\in\mathcal{B}]}$. From (4.12), we have:

$$\begin{aligned}\sum_{(x,y)\in\mathcal{B}} &\Pr\left[(X,Y)=(x,y)\right]\log_2\left(\frac{\Pr\left[(X,Y)=(x,y)\right]}{\Pr\left[X=x\right]\Pr\left[Y=y\right]}\right)\\ &\leq \Pr\left[(X,Y)\in\mathcal{B}\right]\log_2|\mathcal{X}| + \Pr\left[(X,Y)\in\mathcal{B}\right]\log_2\left(1/\Pr\left[(X,Y)\in\mathcal{B}\right]\right)\\ &\leq \beta'(\log_2|\mathcal{X}| + \log_2(1/\beta'))\\ &= \frac{2\beta}{1-2^{-k}}\left(\log_2|\mathcal{X}| + \log_2\left(\frac{1-2^{-k}}{2\beta}\right)\right)\\ &\leq \frac{2\beta\log_2(|\mathcal{X}|/2\beta)}{1-2^{-k}}\end{aligned}$$

where the last inequality follows from the fact that $\Pr\left[(X,Y)\in\mathcal{B}\right] \leq \beta'$ and $w\log_2(1/w) \leq \beta'\log_2(1/\beta')$ when $0 \leq w \leq \beta' \leq 0.3$. Substituting the sum of terms over $\mathcal{B}$ in (4.11) by the above gives us the claimed bound. $\square$

We now turn to pointing out two consequences of this result. The first deals with obtaining tighter p-value corrections than was possible using methods from Russo and Zou (2016) and the second deals with improving bounds on mutual information between an approximately differentially private algorithm and a dataset sampled i.i.d. from a distribution, thus improving results from McGregor et al. (2011).

In Section 1.4, we proved a simple theorem about how to correct p-values (using a *valid p-value correction function* – Definition 1.3.1) given a bound on the *max-information* between the input dataset and the test-statistic selection procedure $\mathcal{M}$ (Theorem 1.4.4). We note that we can easily extend the definition of null hypotheses given in the introduction (and hence p-values and correction functions), to allow for distributions $\mathcal{S}$ over $\mathcal{X}^n$ that need not be product distributions. In fact, we can restate Theorem 1.4.4 in terms of non-product distributions:

Theorem 4.6.2. *Let $\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}$ be a data-dependent algorithm for selecting a test statistic such that $I_\infty^\beta(\mathcal{M}, n) \leq m$. Then the following function γ is a valid p-value correction function for $\mathcal{M}$:*

$$\gamma(\alpha) = \max\left(\frac{\alpha - \beta}{2^m}, 0\right).$$

Proof. The proof is exactly the same as the proof of Theorem 1.4.4, except we fix an arbitrary (perhaps non-product) distribution $\mathcal{S}$ from which the dataset $\boldsymbol{X}$ is drawn. ☐

Previously, Russo and Zou (2016) gave a method to correct p-values given a bound on the *mutual information* between the input data and the test-statistic selection procedure.[2] In Theorem 4.6.1, we observe that if we had a bound on the mutual information between the input data and the test-statistic procedure, this would imply a bound on the *max-information* that would be sufficiently strong so that our Theorem 4.6.2 would give us the following corollary:

Corollary 4.6.3. *Let $\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}$ be a test-statistic selection procedure such that $I(\boldsymbol{X}; \mathcal{M}(\boldsymbol{X})) \leq t$. Then $\gamma(\alpha)$ is a valid p-value correction function, where:*

$$\gamma(\alpha) = \frac{\alpha}{2} \cdot 2^{\frac{-2}{\alpha}(t+0.54)}. \tag{4.13}$$

Proof. From Theorem 4.6.1, we know that for any $m > 0$, $I_\infty^{\beta(m)}(\mathcal{M}, n) \leq m$, where $\beta(m) \leq \frac{t+0.54}{m}$. Hence, from Theorem 4.6.2, we know that for any choice of $m > 0$, $\gamma(\alpha)$ is a valid p-value correction function where:

$$\gamma(\alpha) = \frac{\alpha - \frac{t+0.54}{m}}{2^m}.$$

Choosing $m = \frac{2(t+0.54)}{\alpha}$ gives our claimed bound. ☐

---

[2] Actually, Russo and Zou (2016) do not explicitly model the dataset, and instead give a bound in terms of the mutual information between the test-statistics themselves and the test-statistic selection procedure. We could also prove bounds with this dependence, by viewing our input data to be the value of the given test-statistics, however for consistency, we will discuss all bounds in terms of the mutual information between the data and the selection procedure.

We now show that this gives us a strictly improved p-value correction function than the bound given by Russo and Zou (2016), which we state here using our terminology.
Theorem 4.6.4 [Russo and Zou (2016) Proposition 7]**.** *Let $\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}$ be a test-statistic selection procedure such that $I\left(\boldsymbol{X}; \mathcal{M}(\boldsymbol{X})\right) \leq t$. If we define $\phi_i = \mathcal{M}(\boldsymbol{X})$, then for every $\gamma \in [0,1]$:*

$$\Pr\left[p_i(\phi_i(\boldsymbol{X})) \leq \gamma\right] \leq \gamma + \sqrt{\frac{t}{\log(1/2\gamma)}}.$$

If we want to set parameters so that the probability of a false discovery is at most α, then in particular, we must pick γ such that $\sqrt{\frac{t}{\ln(1/2\gamma)}} \leq \alpha$. Equivalently, solving for α, the best valid p-value correction function implied by the bound of Russo and Zou (2016) must satisfy:

$$\gamma_{RZ}(\alpha) \leq \min\left\{\frac{\alpha}{2}, \frac{1}{2} \cdot 2^{-\log_2(e) t/\alpha^2}\right\}.$$

Comparing the p-value correction function $\gamma(\alpha)$ from (4.13), with the function $\gamma_{RZ}(\alpha)$ above, we see that the version above has an exponentially improved dependence on $1/\alpha$. Moreover, it almost always gives a better correction factor in practice: for any value of $\alpha \leq 0.05$, the function $\gamma(\alpha)$ derived above improves over $\gamma_{RZ}(\alpha)$ whenever the mutual information bound $t \geq 0.05$ (whereas, we would naturally expect the mutual information to be $m \gg 1$, and to scale with n).

Our main theorem Theorem 4.2.1 combined with Theorem 4.6.1 also obtains an improved bound on the mutual information of approximate differentially private mechanisms from Proposition 4.4 in McGregor et al. (2011)[3]. The following corollary improves the bound from McGregor et al. (2011) in its dependence on $|\mathcal{X}|$ from $|\mathcal{X}|^2 \cdot \log_2(1/|\mathcal{X}|)$ to $\log_2 |\mathcal{X}|$.
Corollary 4.6.5. *Let $\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}$ be (ϵ, δ)-differentially private and $\boldsymbol{X} \sim \mathcal{D}^n$. If $\epsilon \in (0, 1/2], \epsilon = \Omega\left(\frac{1}{\sqrt{n}}\right)$, and $\delta = O(\frac{\epsilon}{n^2})$, we then have:*

$$I\left(\boldsymbol{X}; \mathcal{M}(\boldsymbol{X})\right) = O\left(n\epsilon^2 + n\sqrt{\tfrac{\delta}{\epsilon}}\left(1 + \log\left(\tfrac{1}{n}\sqrt{\tfrac{\epsilon}{\delta}}\right) + n\log_2|\mathcal{X}|\right)\right).$$

4.7. Max-Information and Compression Schemes

In this section we show that the approximate max-information of a compression scheme can be unbounded. We note that compression schemes were previously known to provide generalization guarantees for statistical queries, (Cummings et al., 2016a). A consequence of this is that a procedure is not required to have bounded max-information in order for it to have strong generalization guarantees.

We begin by defining compression schemes. Intuitively, a compression scheme uses only a small fraction of the dataset to obtain an output.
Definition 4.7.1 [Littlestone and Warmuth (1986)]**.** *An algorithm $\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}$ has a*

---

[3] Thanks to Salil Vadhan for pointing out this implication.

compression scheme of size k *if there exists an algorithm* $\mathcal{B}_1 : \mathcal{X}^n \to \mathcal{X}^k$ *where* $\mathcal{B}_1(\boldsymbol{x}) = (x_i : i \in S_{\boldsymbol{x}} \subseteq [n],\ |S_{\boldsymbol{x}}| \leq k)$ *and an algorithm* $\mathcal{B}_2 : \mathcal{X}^k \to \mathcal{Y}$ *such that* $\mathcal{M}(\boldsymbol{x}) = \mathcal{B}_2(\mathcal{B}_1(\boldsymbol{x}))$.

The following result shows that compression schemes generalize for statistical queries.

Theorem 4.7.2 [Cummings et al. (2016a)]**.** *Let* $\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}$ *have a compression scheme of size* k *where* $\mathcal{Y} = \{h : \mathcal{X} \to [0,1]\}$. *If* $n \geq 8k \log(2n/\beta)$, *then for* $\tau = O\left(\sqrt{\frac{k\log(n/\beta)}{n}}\right)$

$$\Pr_{\boldsymbol{X}\sim\mathcal{D}^n, h\sim\mathcal{M}(\boldsymbol{X})}\left[|h(\boldsymbol{X}) - h(\mathcal{D})| \leq \tau\right] \geq 1-\beta.$$

However, it turns out that compression schemes can have large max-information.

Theorem 4.7.3. *There exists a compression scheme* $\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}$ *of fixed size such that* $I^{\beta}_{\infty,\Pi}(\mathcal{M}, n) \geq \log_2\left(\frac{|\mathcal{X}|\beta}{2\beta+1/|\mathcal{X}|}\right)$ *for* $\beta > 0$ *and* $I_{\infty,\Pi}(\mathcal{M}, n) \geq \log_2(|\mathcal{X}|)$.

Proof. Consider the distribution $\mathcal{D}$ which is uniform over the integers $\mathcal{X} = [H]$ for some large integer $H > 0$. Consider the simple mapping $\mathcal{M} : \mathcal{X}^n \to \mathcal{X}$ where $\mathcal{M}(\boldsymbol{x}) = x_1$, i.e. it just outputs the first data entry. For any $\beta > 0$ define the interval $\mathcal{I} = \{1, 2, \cdots, \lceil 2\beta H\rceil\}$ and then we define the outcome $\mathcal{S} = \{(\boldsymbol{x}, x) \in \mathcal{X}^n \times \mathcal{X} : \mathcal{M}(\boldsymbol{x}) = x \in \mathcal{I}\}$. We will sample a dataset $\boldsymbol{X} \sim \mathcal{D}^n$. We then have

$$\Pr\left[(\boldsymbol{X}, \mathcal{M}(\boldsymbol{X})) \in \mathcal{S}\right] = \Pr\left[\mathcal{M}(\boldsymbol{X}) \in \mathcal{I}\right] \geq 2\beta.$$

We define $\boldsymbol{X'}$ to be an identical, independent copy of $\boldsymbol{X}$, which gives us

$$\begin{aligned}\Pr\left[(\boldsymbol{X}, \mathcal{M}(\boldsymbol{X'})) \in \mathcal{S}\right] &= \Pr\left[\mathcal{M}(\boldsymbol{X}) = \mathcal{M}(\boldsymbol{X'}) \quad \& \quad \mathcal{M}(\boldsymbol{X'}) \in \mathcal{I}\right] \\ &= \sum_{x_1\in\mathcal{I}} \Pr\left[X_1 = x_1 \quad \& \quad X_1' = x_1\right] = \sum_{x_1\in\mathcal{I}} \Pr\left[X_1 = x_1\right]^2 \\ &= \lceil 2\beta H\rceil / H^2 \leq \frac{2\beta}{H} + \frac{1}{H^2}\end{aligned}$$

We can then bound the approximate max-information for $\mathcal{M}$

$$I^{\beta}_{\infty}(\mathcal{M}, n) \geq \log_2\left(\frac{\Pr\left[(\boldsymbol{X}, \mathcal{M}(\boldsymbol{X})) \in \mathcal{S}\right] - \beta}{\Pr\left[(\boldsymbol{X}, \mathcal{M}(\boldsymbol{X'})) \in \mathcal{S}\right]}\right) \geq \log_2\left(\frac{H\beta}{2\beta + 1/H}\right).$$

When $\beta = 0$, we can then set $\mathcal{I} = 1$, so that $\Pr\left[(\boldsymbol{X}, \mathcal{M}(\boldsymbol{X})) \in \mathcal{S}\right] = 1/H$ and $\Pr\left[(\boldsymbol{X}, \mathcal{M}(\boldsymbol{X'})) \in \mathcal{S}\right] = 1/H^2$, thus $I_{\infty}(\mathcal{M}, n) \geq \log_2(H)$ ☐

The above example shows us that the max-information grows with the domain size H, but the robust generalization guarantee from Cummings et al. (2016a) for compression schemes of fixed size does not grow with the domain size. This suggests that max-information is not the right measure for obtaining generalization bounds when dealing with compression schemes.

4.8. Conclusion and Future Work

We have shown that approximately differentially private algorithms have bounded max-information under product distributions but can have large max-information under arbitrary data distributions. Thus, when data is sampled i.i.d., we can compose differentially private algorithms and bounded description length algorithms, but the private algorithms should come first. This will ensure that we can correct for valid p-values in post-selection hypothesis testing as long as the hypothesis tests are differentially private (see later in Chapters 6 and 7 for examples of private hypothesis tests). Our results also give us a stronger connection between mutual information and p-value corrections given in Russo and Zou (2016) and improves on the mutual information bound of approximately differentially private algorithms from McGregor et al. (2011). Further, we applied our lower bound result to show that robustly generalizing algorithms (Cummings et al., 2016a) do not compose in general. Although max-information partially unifies different procedures with strong generalization guarantees in the adaptive setting, it is not the case that every procedure with small max-information must generalize.

Some directions for future work include obtaining a strong bound on max-information for zCDP algorithms. We can easily obtain a bound on max-information due to the fact that zCDP implies approximate DP (see Theorem 2.2.4), so we would like to improve on this bound. Also, we would like to improve on the constants in our max-information bound so that we can obtain valid p-value corrections for practically sized datasets, thus further demonstrating the power of these techniques over datasplitting.

CHAPTER 5

PRIVACY ODOMETERS AND FILTERS: PAY-AS-YOU-GO COMPOSITION

The composition theorems for differential privacy are very strong, and hold even if the choice of which differentially private subroutine to run may depend on the output of previous algorithms. This property is essential in algorithm design, but also more generally in modeling unstructured sequences of data analyses that might be run by a human data analyst, or even by many data analysts on the same data set, while only loosely coordinating with one another. Additionally, we have already pointed out the power of these composition theorems to the application of adaptive data analysis

However, all the known composition theorems for differential privacy (Dwork et al., 2006b,a, 2010; Kairouz et al., 2015; Murtagh and Vadhan, 2016) have an important and generally overlooked caveat. Although the choice of the next subroutine in the composition may be adaptive, the number of subroutines called and choice of the privacy parameters ϵ and δ for each subroutine must be fixed in advance. Indeed, it is not even clear how to define differential privacy if the privacy parameters are not fixed in advance. This is generally acceptable when designing a single algorithm (that has a worst-case analysis), since worst-case eventualities need to be anticipated and budgeted for in order to prove a theorem. However, it is *not* acceptable when modeling the unstructured adaptivity of a data analyst, who may not know ahead of time (before seeing the results of intermediate analyses) what he wants to do with the data. When controlling privacy loss across multiple data analysts, the problem is even worse.

The contents of this chapter are taken largely from Rogers et al. (2016b). We study the composition properties of differential privacy when *everything*—the choice of algorithms, the number of rounds, and the privacy parameters in each round—may be adaptively chosen. We show that this setting is much more delicate than the settings covered by previously known composition theorems, but that these sorts of *ex post* privacy bounds do hold with only a small (but in some cases unavoidable) loss over the standard setting. We note that the conceptual discussion of differential privacy focuses a lot on the idea of arbitrary composition and our results give more support for this conceptual interpretation.

5.1. Results

We give a formal framework for reasoning about the adaptive composition of differentially private algorithms when the privacy parameters themselves can be chosen adaptively. When the parameters are chosen non-adaptively, a *composition theorem* gives a high probability bound on the worst case *privacy loss* that results from the output of an algorithm. In the adaptive parameter setting, it no longer makes sense to have fixed bounds on the privacy loss. Instead, we propose two kinds of primitives capturing two natural use cases for composition theorems:

1. A *privacy odometer* takes as input a global failure parameter δ_g. After every round i in the composition of differentially private algorithms, the odometer outputs a number τ_i that may depend on the *realized* privacy parameters ϵ_i, δ_i in the previous rounds. The privacy odometer guarantees that with probability $1 - \delta_g$, for every round i, τ_i is an upper bound on the privacy loss in round i.

2. A *privacy filter* is a way to cut off access to the dataset when the privacy loss is too large. It takes as input a global privacy "budget" (ϵ_g, δ_g). After every round, it either outputs `CONT` ("continue") or `HALT` depending on the privacy parameters from the previous rounds. The privacy filter guarantees that with probability $1 - \delta_g$, it will output `HALT` before the privacy loss exceeds ϵ_g. When used, it guarantees that the resulting interaction is (ϵ_g, δ_g)-DP.

A tempting heuristic is to take the *realized* privacy parameters $\epsilon_1, \delta_1, \ldots, \epsilon_i, \delta_i$ and apply one of the existing composition theorems to those parameters, using that value as a privacy odometer or implementing a privacy filter by halting when getting a value that exceeds the global budget. However this heuristic *does not* necessarily give valid bounds.

We first prove that the heuristic *does* work for the basic composition theorem from Dwork et al. (2006b) in which the parameters ϵ_i and δ_i add up. We prove that summing the realized privacy parameters yields both a valid privacy odometer and filter. The idea of a privacy filter was also considered in Ebadi and Sands (2015), who show that basic composition works in the privacy filter application. The main contribution of this work is obtaining an *advanced composition* theorem when the privacy parameters can be chosen adaptively, where the overall privacy loss degrades sublinearly with the number of private analyses.

We then show that the heuristic breaks for the advanced composition theorem from Dwork et al. (2010). However, we give a valid privacy filter that gives the same asymptotic bound as the advanced composition theorem, albeit with worse constants. On the other hand, we show that, in some parameter regimes, the asymptotic bounds given by our privacy filter *cannot* be achieved by a privacy odometer. This result gives a formal separation between the two models when the parameters may be chosen adaptively, which does not exist when the privacy parameters are fixed. Finally, we give a valid privacy odometer with a bound that is only slightly worse asymptotically than the bound that the advanced composition theorem would give if it were used (improperly) as a heuristic. Our bound is worse by a factor that is never larger than $\sqrt{\log\log(n)}$ (here, n is the size of the dataset) and for some parameter regimes is only a constant.

5.2. Additional Preliminaries

In addition to the preliminaries that were presented in Chapter 2, we will go over a few quantities which we will need in this chapter and we will cover the composition theorems of DP presented in Chapter 2 in a slightly different way.

We have the following useful characterization from Kairouz et al. (2015): any DP algorithm can be written as the post-processing of a simple, canonical algorithm which is a generalization of *randomized response*.

Definition 5.2.1. *For any* $\epsilon, \delta \geq 0$*, we define the* randomized response *algorithm* $\texttt{RR}_{\epsilon,\delta} : \{0,1\} \to \{0, \top, \bot, 1\}$ *as the following (Note that if* $\delta = 0$*, we will simply write the algorithm* $\texttt{RR}_{\epsilon,\delta}$ *as* $\texttt{RR}_{\epsilon}$*.)*

$$\begin{array}{ll} \Pr\left[\texttt{RR}_{\epsilon,\delta}(0) = 0\right] = \delta & \Pr\left[\texttt{RR}_{\epsilon,\delta}(1) = 0\right] = 0 \\ \Pr\left[\texttt{RR}_{\epsilon,\delta}(0) = \top\right] = (1-\delta)\frac{e^{\epsilon}}{1+e^{\epsilon}} & \Pr\left[\texttt{RR}_{\epsilon,\delta}(1) = \top\right] = (1-\delta)\frac{1}{1+e^{\epsilon}} \\ \Pr\left[\texttt{RR}_{\epsilon,\delta}(0) = \bot\right] = (1-\delta)\frac{1}{1+e^{\epsilon}} & \Pr\left[\texttt{RR}_{\epsilon,\delta}(1) = \bot\right] = (1-\delta)\frac{e^{\epsilon}}{1+e^{\epsilon}} \\ \Pr\left[\texttt{RR}_{\epsilon,\delta}(0) = 1\right] = 0 & \Pr\left[\texttt{RR}_{\epsilon,\delta}(1) = 1\right] = \delta \end{array}$$

Kairouz et al. (2015) show that any (ϵ, δ)–DP algorithm can be viewed as a post-processing of the output of $\texttt{RR}_{\epsilon,\delta}$ for an appropriately chosen input.

Theorem 5.2.2 [Kairouz et al. (2015); Murtagh and Vadhan (2016)]**.** *For every* (ϵ, δ)*-DP algorithm* $\mathcal{M}$ *and for all neighboring databases* $\boldsymbol{x}^0$ *and* $\boldsymbol{x}^1$*, there exists a randomized mapping* T *where* $T(\texttt{RR}_{\epsilon,\delta}(b))$ *is identically distributed to* $\mathcal{M}(\boldsymbol{x}^b)$ *for* $b \in \{0,1\}$*.*

This theorem will be useful in our analyses, because it allows us to without loss of generality analyze compositions of these simple algorithms $\texttt{RR}_{\epsilon,\delta}$ with varying privacy parameters.

We now define the adaptive composition of differentially private algorithms in the setting introduced by Dwork et al. (2010) and then extended to *heterogenous* privacy parameters in Murtagh and Vadhan (2016), in which all of the privacy parameters are fixed prior to the start of the computation. The following "composition game" is an abstract model of composition in which an adversary can adaptively select between neighboring datasets at each round, as well as a differentially private algorithm to run at each round – both choices can be a function of the realized outcomes of all previous rounds. However, crucially, the adversary must select at each round an algorithm that satisfies the privacy parameters which have been fixed ahead of time – the choice of parameters cannot itself be a function of the realized outcomes of previous rounds. We define this model of interaction formally in Algorithm 2 where the output is the *view* of the adversary $\mathcal{A}$ which includes any random coins she uses $R_{\mathcal{A}}$ and the outcomes of algorithms $\mathcal{M}_1, \cdots, \mathcal{M}_k$ of every round.

Algorithm 2 $\texttt{FixedParamComp}(\mathcal{A}, \mathcal{E}, b)$

```
Input: $\mathcal{A}$ is a randomized algorithm, $\mathcal{E} = (\mathcal{E}_1, \cdots, \mathcal{E}_k)$ are classes of randomized algorithms,
  and $b \in \{0,1\}$.
  Select coin tosses $R^b_{\mathcal{A}}$ for $\mathcal{A}$ uniformly at random.
  for $i = 1, \cdots, k$ do
    $\mathcal{A} = \mathcal{A}(R^b_{\mathcal{A}}, \mathcal{M}^b_1, \cdots, \mathcal{M}^b_{i-1})$ gives neighboring datasets $\boldsymbol{x}^{i,0}, \boldsymbol{x}^{i,1}$, and $\mathcal{M}_i \in \mathcal{E}_i$
    $\mathcal{A}$ receives $\mathcal{M}^b_i = \mathcal{M}_i(\boldsymbol{x}^{i,b})$
Output: view $V^b = (R^b_{\mathcal{A}}, \mathcal{M}^b_1, \cdots, \mathcal{M}^b_k)$
```

Definition 5.2.3 [Adaptive Composition (Dwork et al., 2010; Murtagh and Vadhan, 2016)]**.** *We say that the sequence of parameters* $\epsilon_1, \cdots, \epsilon_k \geq 0$*,* $\delta_1, \cdots, \delta_k \in [0,1)$ *satisfies* (ϵ_g, δ_g)*-*differential privacy under adaptive composition *if for every adversary* $\mathcal{M}$*, and* $\mathcal{E} = (\mathcal{E}_1, \cdots, \mathcal{E}_k)$ *where* $\mathcal{E}_i$ *is the class of* (ϵ_i, δ_i)*-DP algorithms, we have* $\texttt{FixedParamComp}(\mathcal{M}, \mathcal{E}, \cdot)$ *is* (ϵ_g, δ_g)*-DP in its last argument, i.e.* $V^0 \approx_{\epsilon_g,\delta_g} V^1$*.*

We then present the composition theorems in terms of this adaptive composition. We first state the basic composition theorem (compare to Theorem 2.1.4) which shows that the adaptive composition satisfies differential privacy where "the parameters just add up."

Theorem 5.2.4 [Basic Composition (Dwork et al., 2006b,a)]**.** *The sequence* $\epsilon_1, \cdots, \epsilon_k$ *and* $\delta_1, \cdots \delta_k$ *satisfies* (ϵ_g, δ_g)*-differential privacy under adaptive composition where* (ϵ_g, δ_g) *are given in Theorem 2.1.4.*

We now state the advanced composition bound from Dwork et al. (2010) and then improved by Kairouz et al. (2015) which gives a quadratic improvement to the basic composition bound (compare to Theorem 2.1.5).

Theorem 5.2.5 [Advanced Composition (Dwork et al., 2010; Kairouz et al., 2015)]**.** *For any* $\widehat{\delta} > 0$*, the sequence* $\epsilon_1, \cdots, \epsilon_k$ *and* $\delta_1, \cdots \delta_k$ *satisfies* (ϵ_g, δ_g)*-differential privacy under adaptive composition where* (ϵ_g, δ_g) *are given in Theorem 2.1.5.*

The remainder of this paper is devoted to laying out a framework for sensibly talking about the privacy parameters ϵ_i and δ_i being chosen adaptively by the data analyst, and to prove composition theorems (including an analogue of Theorem 5.2.5) in this model.

5.3. Composition with Adaptively Chosen Parameters

We now introduce the model of composition with adaptive parameter selection, and define privacy in this setting.

5.3.1. Definition

We want to model composition as in the previous section, but allow the adversary the ability to also choose the privacy parameters (ϵ_i, δ_i) as a function of previous rounds of interaction. We will define the view of the interaction, similar to the view in `FixedParamComp`, to be the tuple that includes $\mathcal{A}$'s random coin tosses $R_{\mathcal{A}}$ and the outcomes $\mathcal{M} = (\mathcal{M}_1, \cdots, \mathcal{M}_k)$ of the algorithms she chose. Formally, we define an *adaptively chosen privacy parameter composition game* in Algorithm 3 which takes as input an adversary $\mathcal{A}$, a number of rounds of interaction k,[1] and an experiment parameter $b \in \{0, 1\}$.

We then define the privacy loss with respect to `AdaptParamComp`$(\mathcal{A}, k, b)$ in the following way for a fixed view $\boldsymbol{v} = (r, \boldsymbol{a})$ where r represents the random coin tosses of $\mathcal{A}$ and we write

---

[1] Note that in the adaptive parameter composition game, the adversary has the option of effectively stopping the composition early at some round $k' < k$ by simply setting $\epsilon_i = \delta_i = 0$ for all rounds $i > k'$. Hence, the parameter k will not appear in our composition theorems the way it does when privacy parameters are fixed. This means that we can effectively take k to be infinite. For technical reasons, it is simpler to have a finite parameter k, but the reader should imagine it as being an enormous number(say the number of atoms in the universe) so as not to put any constraint at all on the number of rounds of interaction with the adversary.

Algorithm 3 `AdaptParamComp`$(\mathcal{A}, k, b)$

```
Input: $\mathcal{A}$ is a randomized algorithm, upper bound $k$, and $b \in \{0,1\}$.
  Select coin tosses $R^b_{\mathcal{A}}$ for $\mathcal{A}$ uniformly at random.
  for $i = 1, \cdots, k$ do
    $\mathcal{A} = \mathcal{A}(R^b_{\mathcal{A}}, \mathcal{M}^b_1, \cdots, \mathcal{M}^b_{i-1})$ gives neighboring $\boldsymbol{x}^{i,0}, \boldsymbol{x}^{i,1}$, parameters $(\epsilon_i, \delta_i)$, $\mathcal{M}_i$ that is $(\epsilon_i, \delta_i)$-DP
    $\mathcal{A}$ receives $\mathcal{M}^b_i = \mathcal{M}_i(\boldsymbol{x}^{i,b})$
Output: view $V^b = (R^b_{\mathcal{A}}, \mathcal{M}^b_1, \cdots, \mathcal{M}^b_k)$
```

$\boldsymbol{v}_{<i} = (r, a_1, \cdots, a_{i-1})$:

$$\begin{aligned}\text{PrivLoss}(\boldsymbol{v}) &= \log\left(\frac{\Pr\left[V^0 = \boldsymbol{v}\right]}{\Pr\left[V^1 = \boldsymbol{v}\right]}\right) \\ &= \sum_{i=1}^{k} \log\left(\frac{\Pr\left[\mathcal{M}_i(\boldsymbol{x}^{i,0}) = v_i | \boldsymbol{v}_{<i}\right]}{\Pr\left[\mathcal{M}_i(\boldsymbol{x}^{i,1}) = v_i | \boldsymbol{v}_{<i}\right]}\right) \\ &\overset{\text{defn}}{=} \sum_{i=1}^{k} \text{PrivLoss}_i(\boldsymbol{v}_{\leq i}). \qquad (5.1)\end{aligned}$$

Note that the privacy parameters (ϵ_i, δ_i) depend on the previous outcomes that $\mathcal{M}$ receives. We will frequently shorten our notation $\epsilon_t = \epsilon_t(\boldsymbol{v}_{<t})$ and $\delta_t = \delta_t(\boldsymbol{v}_{<t})$ when the outcome is understood.

It no longer makes sense to claim that the privacy loss of the adaptive parameter composition experiment is bounded by any fixed constant, because the privacy parameters (with which we would presumably want to use to bound the privacy loss) are themselves random variables. Instead, we define two objects which can be used by a data analyst to control the privacy loss of an adaptive composition of algorithms.

The first object, which we call a *privacy odometer* will be parameterized by one global parameter δ_g and will provide a running real valued output that will, with probability $1 - \delta_g$, upper bound the privacy loss at each round of any adaptive composition in terms of the *realized* values of ϵ_i and δ_i selected at each round.

Definition 5.3.1 [Privacy Odometer]**.** *A function* $\texttt{COMP}_{\delta_g} : \mathbb{R}^{2k}_{\geq 0} \to \mathbb{R} \cup \{\infty\}$ *is a* valid privacy odometer *if for all adversaries* $\mathcal{A}$ *in* `AdaptParamComp`$(\mathcal{A}, k, b)$*, with probability at most* δ_g *over* $\boldsymbol{v} \sim V^0$*:*

$$|\text{PrivLoss}(\boldsymbol{v})| > \texttt{COMP}_{\delta_g}\left(\epsilon_1, \delta_1, \cdots, \epsilon_k, \delta_k\right).$$

The second object, which we call a *privacy filter*, is a stopping time rule. It takes two global parameters (ϵ_g, δ_g) and will at each round either output `CONT` or `HALT`. Its guarantee is that with probability $1 - \delta_g$, it will output `HALT` if the privacy loss has exceeded ϵ_g.

Definition 5.3.2 [Privacy Filter]**.** *A function* $\texttt{COMP}_{\epsilon_g,\delta_g} : \mathbb{R}^{2k}_{\geq 0} \to \{\texttt{HALT}, \texttt{CONT}\}$ *is a* valid privacy filter *for* $\epsilon_g, \delta_g \geq 0$ *if for all adversaries* $\mathcal{A}$ *in* $\texttt{AdaptParamComp}(\mathcal{A}, k, b)$*, the following "bad event" occurs with probability at most* δ_g *when* $\boldsymbol{v} \sim V^0$*:*

$$|\mathrm{PrivLoss}(\boldsymbol{v})| > \epsilon_g \quad \textit{and} \quad \texttt{COMP}_{\epsilon_g,\delta_g}(\epsilon_1, \delta_1, \cdots, \epsilon_k, \delta_k) = \texttt{CONT}.$$

We note two things about the usage of these objects. First, a valid privacy odometer can be used to provide a running upper bound on the privacy loss at each intermediate round: the privacy loss at round $k' < k$ must with high probability be upper bounded by $\texttt{COMP}_{\delta_g}(\epsilon_1, \delta_1, \ldots, \epsilon_{k'}, \delta_{k'}, 0, 0, \ldots, 0, 0)$ – i.e. the bound that results by setting all future privacy parameters to 0. This is because setting all future privacy parameters to zero is equivalent to stopping the computation at round k', and is a feasible choice for the adaptive adversary $\mathcal{A}$. Second, a privacy filter can be used to guarantee that with high probability, the stated privacy budget ϵ_g is never exceeded – the data analyst at each round k' simply queries $\texttt{COMP}_{\epsilon_g,\delta_g}(\epsilon_1, \delta_1, \ldots, \epsilon_{k'}, \delta_{k'}, 0, 0, \ldots, 0, 0)$ *before* she runs algorithm k', and runs it only if the filter returns CONT. Again, this is guaranteed because the continuation is a feasible choice of the adversary, and the guarantees of both a filter and an odometer are quantified over all adversaries. Further, if we have a privacy odometer then it can be used as a filter, giving us the following result.

Lemma 5.3.3. *If* $\texttt{COMP}_{\delta_g}$ *is a valid privacy odometer then the following function* $\texttt{COMP}_{\epsilon_g,\delta_g}$ *is a valid privacy filter:* $\texttt{COMP}_{\epsilon_g,\delta_g}(\epsilon_1, \delta_1, \cdots, \epsilon_k, \delta_k) = \texttt{CONT}$ *if*

$$\texttt{COMP}_{\delta_g}(\epsilon_1, \delta_1, \cdots, \epsilon_k, \delta_k) \leq \epsilon_g$$

and otherwise $\texttt{COMP}_{\epsilon_g,\delta_g}(\epsilon_1, \delta_1, \cdots, \epsilon_k, \delta_k) = \texttt{HALT}$*.*

Proof. With the privacy filter $\texttt{COMP}_{\epsilon_g,\delta_g}$ defined in the lemma statement, we consider the event that $\texttt{COMP}_{\epsilon_g,\delta_g}(\epsilon_1, \delta_1, \cdots, \epsilon_k, \delta_k) = \texttt{CONT}$ which implies that the valid privacy odometer $\texttt{COMP}_{\delta_g}(\epsilon_1, \delta_1, \cdots, \epsilon_k, \delta_k) \leq \epsilon_g$. By definition of a valid privacy odometer, we then know that with probability at most δ_g over $\boldsymbol{v} \sim V^0$ that $|\mathrm{PrivLoss}(\boldsymbol{v})| \leq \texttt{COMP}_{\delta_g}(\epsilon_1, \delta_1, \cdots, \epsilon_k, \delta_k) \leq \epsilon_g$. Hence $\texttt{COMP}_{\epsilon_g,\delta_g}$ is a valid privacy filter. □

We give the formal description of this interaction where $\mathcal{A}$ uses the privacy filter in Algorithm 4.

From the way we have defined a valid privacy filter, we have the following proposition:

Proposition 5.3.4. *If* $\texttt{COMP}_{\epsilon_g,\delta_g}$ *is a valid privacy filter then the views* $V^0_{\texttt{F}}$ *and* $V^1_{\texttt{F}}$ *of the adversary from* $\texttt{PrivacyFilterComp}\left(\mathcal{A}, k, b; \texttt{COMP}_{\epsilon_g,\delta_g}\right)$ *with* $b = 0$ *and* $b = 1$ *respectively, are* (ϵ_g, δ_g)*-point-wise indistinguishable and hence* $V^0_{\texttt{F}} \approx_{\epsilon_g,\delta_g} V^1_{\texttt{F}}$*.*

5.3.2. Focusing on Randomized Response

Theorem 5.2.2 was used to show that for ordinary composition (Definition 5.2.3), it suffices to analyze the composition of randomized response. In this section, we show something similar for privacy odometers and filters. Specifically, we show that we can simulate $\texttt{AdaptParamComp}(\mathcal{A}, k, b)$ by defining a new adversary that chooses the differentially private algorithm $\mathcal{M}_i$ of adversary $\mathcal{A}$, but uses the randomized response algorithm from Defini-

Algorithm 4 $\texttt{PrivacyFilterComp}(\mathcal{A}, k, b; \texttt{COMP}_{\epsilon_g,\delta_g})$

```
Input: $\mathcal{A}$ is a randomized algorithm, upper bound $k$, $b \in \{0,1\}$, and filter $\texttt{COMP}_{\epsilon_g,\delta_g})$.
  Select coin tosses $R^b_{\mathcal{A}}$ for $\mathcal{A}$ uniformly at random.
  for $i = 1, \cdots, k$ do
    $\mathcal{A} = \mathcal{A}(R^b_{\mathcal{A}}, \mathcal{M}^b_1, \cdots, \mathcal{M}^b_{i-1})$ gives neighboring $\boldsymbol{x}^{i,0}, \boldsymbol{x}^{i,1}$, $(\epsilon_i, \delta_i)$, and $\mathcal{M}_i$ that is $(\epsilon_i, \delta_i)$-DP
    if $\texttt{COMP}_{\epsilon_g,\delta_g}(\epsilon_1, \delta_1, \cdots, \epsilon_i, \delta_i, 0, 0, \cdots, 0, 0) = \texttt{HALT}$ then
      $\mathcal{M}_i, \cdots, \mathcal{M}_k = \perp$
      BREAK
    else
      $\mathcal{A}$ receives $\mathcal{M}^b_i = \mathcal{M}_i(\boldsymbol{x}^{i,b})$
Output: view $V^b_{\texttt{F}} = (R^b_{\mathcal{A}}, \mathcal{M}^b_1, \cdots, \mathcal{M}^b_k)$
```

tion 5.2.1 each round along with a post-processing function, which together determine the distribution for $\mathcal{M}_i$.

In Algorithm 5, we define the new composition game $\texttt{SimulatedComp}(\mathcal{A}, k, b)$ with adversary $\mathcal{A}$ that outputs the view W^b, which includes the internal randomness $R^b_{\mathcal{A}}$ of $\mathcal{A}$ with the randomized response outcomes $Z^b = (Z^b_1, \cdots, Z^b_k)$. From Theorem 5.2.2, we know that we can simulate any (ϵ, δ)-DP algorithm as a randomized post-processing function T on top of $\texttt{RR}_{\epsilon,\delta}$. Thus given the outcomes prior to round i, $\mathcal{A}$ selects $\mathcal{M}_i$, which is equivalent to selecting a post-processing function T_i. Note that we can simulate T_i as a deterministic function P_i with access to random coins $R^b_{\texttt{SIM}_i}$, i.e. $P_i\left(\texttt{RR}_{\epsilon_i,\delta_i}(b); R^b_{\texttt{SIM}_i}\right) \sim T_i\left(\texttt{RR}_{\epsilon_i,\delta_i}(b)\right)$. We then include the random coins $R^b_{\texttt{SIM}} = \left(R^b_{\texttt{SIM}_1}, \cdots, R^b_{\texttt{SIM}_k}\right)$ in the view of adversary $\mathcal{A}$ in $\texttt{SimulatedComp}(\mathcal{A}, k, b)$. From the view $W^b = \left(R^b_{\mathcal{A}}, R_{\texttt{SIM}}, Z^b_1, \cdots, Z^b_k\right)$, $\mathcal{A}$ would be able to reconstruct the privacy parameters selected each round along with algorithms $\mathcal{M}_1, \cdots, \mathcal{M}_k$ used, which would also determine the post-processing functions $P_1, \cdots, P_k$.

Algorithm 5 $\texttt{SimulatedComp}(\mathcal{A}, k, b)$

```
Input: $\mathcal{A}$ is a randomized algorithm, upper bound $k$, and $b \in \{0,1\}$.
  Select coin tosses $R^b_{\mathcal{A}}$ for $\mathcal{A}$ uniformly at random.
  for $i = 1, \cdots, k$ do
    $\mathcal{A} = \mathcal{A}\left(R^b_{\mathcal{A}}, Y^b_1, \cdots, Y^b_{i-1}\right)$ gives neighboring $\boldsymbol{x}^{i,0}, \boldsymbol{x}^{i,1}$, parameters $(\epsilon_i, \delta_i)$, $\mathcal{M}_i$ that is $(\epsilon_i, \delta_i)$-DP.
    Let $P_i$ be a deterministic post-processing function, such that

      $$P_i\left(\texttt{RR}_{\epsilon_i,\delta_i}(b); R^b_{\texttt{SIM}_i}\right) \sim \mathcal{M}_i\left(\boldsymbol{x}^{i,b}\right) \quad (5.2)$$

    for uniformly random $R^b_{\texttt{SIM}_i}$.
    Compute $Z^b_i = \texttt{RR}_{\epsilon_i,\delta_i}(b)$ and $Y^b_i = P_i(Z^b_i; R^b_{\texttt{SIM}_i})$.
    $\mathcal{A}$ receives $Y^b_i$.
Output: view $W^b = (R^b_{\mathcal{A}}, R^b_{\texttt{SIM}}, Z^b_1, \cdots, Z^b_k)$, where $R^b_{\texttt{SIM}} = \left(R^b_{\texttt{SIM}_1}, \cdots, R^b_{\texttt{SIM}_k}\right)$.
```

From the way that we have defined P_i in (5.2), for each fixed value of the internal randomness of $\mathcal{A}$, the view of $\texttt{AdaptParamComp}(\mathcal{A}, k, b)$ is distributed identically to a post-processing of the view W^b from $\texttt{SimulatedComp}(\mathcal{A}, k, b)$.

Lemma 5.3.5. *For every adversary $\mathcal{A}$, the deterministic function P defined as*

$$P\left(R^b_{\mathcal{A}}, R^b_{\texttt{SIM}}, Z^b_1, \cdots, Z^b_k\right) = \left(R^b_{\mathcal{A}}, P_1(Z^b_1; R^b_{\texttt{SIM}_1}), \cdots, P_k(Z^b_k; R^b_{\texttt{SIM}_k})\right) \tag{5.3}$$

ensures $P\left(\texttt{SimulatedComp}(\mathcal{A}, k, b)\right)$ and $\texttt{AdaptParamComp}\left(\mathcal{A}, k, b\right)$ are identically distributed.

Since $R^b_{\mathcal{A}}$ is the first argument of both random variables, they are also identically distributed conditioned on any fixed value of $R^b_{\mathcal{A}}$. This point-wise equivalence for every value of the internal randomness allows us to without loss of generality analyze *deterministic* adversaries and post-processing functions of $\texttt{SimulatedComp}(\mathcal{A}, k, b)$ in order to reason about the view of $\texttt{AdaptParamComp}(\mathcal{A}, k, b)$. Because the randomness is fixed, for clarity, we will omit the random coins $R^b_{\mathcal{A}}$ from the view of both composition games for the rest of the analysis.

We will now show that it is sufficient to prove bounds in which ϵ_i may be adaptively chosen at each round, and in which $\{\delta_i\} \equiv 0$ uniformly. We do this by giving a generic way to extend a bound in the $\delta_i = 0$ case to a bound that holds when the δ_i may be non-zero. Define a slight modification of Algorithm 5 called $\widetilde{\texttt{SimulatedComp}}(\mathcal{A}, k, b)$ which is the same as $\texttt{SimulatedComp}(\mathcal{A}, k, b)$ except that it computes $\widetilde{Z}^b_i = \texttt{RR}_{\epsilon_i}(b)$ (where $\delta_i = 0$) and sets $\widetilde{Y}^b_i = P_i\left(\widetilde{Z}^b_i; R^b_{\texttt{SIM}_i}\right)$. We then define the final view of the adversary $\mathcal{A}$ in $\widetilde{\texttt{SimulatedComp}}(\mathcal{A}, k, b)$ as $\widetilde{W}^b$ where

$$\widetilde{W}^b = \left(R^b_{\texttt{SIM}}, \widetilde{Z}^b_1, \cdots, \widetilde{Z}^b_k\right) \qquad \text{and} \qquad \widetilde{V}^b = \left(\widetilde{Y}^b_1, \cdots, \widetilde{Y}^b_k\right) = P\left(\widetilde{W}^b\right) \tag{5.4}$$

for $P(\cdot)$ given in (5.3). We then say that $\widetilde{\texttt{COMP}}_{\delta_g}$ (also $\widetilde{\texttt{COMP}}_{\epsilon_g,\delta_g}$) is a valid privacy odometer (filter) when $\{\delta_i\} \equiv 0$ if over all deterministic adversaries $\mathcal{A}$ in $\widetilde{\texttt{SimulatedComp}}(\mathcal{A}, k, b)$ the condition in Definition 5.3.1 (Definition 5.3.2) holds with probability at most δ_g over $\widetilde{\boldsymbol{v}} \sim P\left(\widetilde{W}^b\right)$ except now the privacy loss is given as

$$\begin{aligned}\widetilde{\text{PrivLoss}}(\widetilde{\boldsymbol{v}}) = \log\left(\frac{\Pr\left[\widetilde{V}^0 = \widetilde{\boldsymbol{v}}\right]}{\Pr\left[\widetilde{V}^1 = \widetilde{\boldsymbol{v}}\right]}\right) &= \sum_{i=1}^k \log\left(\frac{\Pr\left[P_i\left(\texttt{RR}_{\epsilon_i}(0); R^0_{\texttt{SIM}_i}\right) = \widetilde{v}_i | \widetilde{\boldsymbol{v}}_{<i}\right]}{\Pr\left[P_i\left(\texttt{RR}_{\epsilon_i}(1); R^1_{\texttt{SIM}_i}\right) = \widetilde{v}_i | \widetilde{\boldsymbol{v}}_{<i}\right]}\right) \\ &\overset{\text{defn}}{=} \sum_{i=1}^k \widetilde{\text{PrivLoss}}_i(\widetilde{\boldsymbol{v}}_{\leq i}).\end{aligned} \tag{5.5}$$

The following result gives the connection between valid privacy odometers and filters in the modified game $\widetilde{\texttt{SimulatedComp}}(\mathcal{A}, k, b)$ with the original definitions given in Definitions 5.3.1 and 5.3.2.

Lemma 5.3.6. *If $\widetilde{\texttt{COMP}}_{\delta_g}$ is a valid privacy odometer when $\{\delta_i\} \equiv 0$, then for every $\delta'_g \geq 0$,*

$\texttt{COMP}_{\delta_g+\delta'_g}$ *is a valid privacy odometer where*

$$\texttt{COMP}_{\delta_g+\delta'_g}(\epsilon_1,\delta_1,\cdots,\epsilon_k,\delta_k) = \begin{cases} \infty & \text{if } \sum_{i=1}^k \delta_i > \delta'_g \\ \widetilde{\texttt{COMP}}_{\delta_g}(\epsilon_1,0,\cdots,\epsilon_k,0) & \text{otherwise} \end{cases}.$$

If $\widetilde{\texttt{COMP}}_{\epsilon_g,\delta_g}$ *is a valid privacy filter when* $\{\delta_i\} \equiv 0$*, then for every* $\delta'_g \geq 0$*,* $\texttt{COMP}_{\epsilon_g,\delta_g+\delta'_g}$ *is a valid privacy filter where*

$$\texttt{COMP}_{\epsilon_g,\delta_g+\delta'_g}(\epsilon_1,\delta_1,\cdots,\epsilon_k,\delta_k) = \begin{cases} \texttt{HALT} & \text{if } \sum_{i=1}^k \delta_i > \delta'_g \\ \widetilde{\texttt{COMP}}_{\epsilon_g,\delta_g}(\epsilon_1,0,\cdots,\epsilon_k,0) & \text{otherwise} \end{cases}.$$

Proof. Let $W = (R_{\texttt{SIM}}, Z_1, \cdots, Z_k)$ be the view of $\mathcal{A}$ in $\texttt{SimulatedComp}(\mathcal{A},k,0)$ and $\widetilde{W} = (R_{\texttt{SIM}}, \widetilde{Z}_1, \cdots, \widetilde{Z}_k)$ be her view in $\widetilde{\texttt{SimulatedComp}}(\mathcal{A},k,0)$ (where $\{\delta_i\} \equiv 0$). We will also write the view of $\texttt{AdaptParamComp}(\mathcal{A},k,0)$ as $V = (\mathcal{M}_1, \cdots, \mathcal{M}_k)$ and the post-processing functions of $\mathcal{A}$ as P_i from (5.2).

As in (5.3), we will use the notation $P(W) = (P_1(Z_1; R_{\texttt{SIM}_i}), \cdots, P_k(Z_k; R_{\texttt{SIM}_i}))$ and similarly for $\widetilde{V} = P(\widetilde{W})$. Recall that from Lemma 5.3.5 that we know $V \sim P(W)$, even if $\mathcal{A}$ were randomized.

Consider the following method of sampling from $\texttt{RR}_{\epsilon,\delta}$: first select outcome $\widetilde{z}$ from $\texttt{RR}_\epsilon(0)$, then with probability $1-\delta$ set $z = \widetilde{z}$ – otherwise set $z = 0$. Note that this samples from the correct distribution for $\texttt{RR}_{\epsilon,\delta}(0)$. We can thus couple draws from $\texttt{RR}_\epsilon(0)$ and $\texttt{RR}_{\epsilon,\delta}(0)$, so for our setting we write the coupled random variable as: $\boldsymbol{V} = (V, \widetilde{V})$.

We then define the following sets:

$$\begin{aligned}
\mathcal{F} &\overset{\text{defn}}{=} \{(w = (r,\boldsymbol{z}), \widetilde{w} = (r,\widetilde{\boldsymbol{z}})) : \exists t \in [k] \text{ s.t. } z_t \neq \widetilde{z}_t\}, \\
\mathcal{G}_t &\overset{\text{defn}}{=} \left\{ v : \sum_{i=1}^t \delta_i(\boldsymbol{v}_{<i}) \leq \delta'_g \right\}, \\
\mathcal{F}_t &\overset{\text{defn}}{=} \{(w = (r,\boldsymbol{z}), \widetilde{w} = (r,\widetilde{\boldsymbol{z}})) : z_t \neq \widetilde{z}_t \text{ and } z_i = \widetilde{z}_i \quad \forall i < t\}, \\
\mathcal{H} &\overset{\text{defn}}{=} \left\{ \boldsymbol{v} : |\widetilde{\text{PrivLoss}}(\boldsymbol{v})| \geq \widetilde{\texttt{COMP}}_{\delta_g}(\epsilon_1, 0, \cdots, \epsilon_k, 0) \right\}.
\end{aligned}$$

We then want to show that we can bound the privacy loss with high probability. Specifically,

$$\Pr_{\boldsymbol{V}}\left[|\text{PrivLoss}(V)| \geq \widetilde{\texttt{COMP}}_{\delta_g}(\epsilon_1,0,\cdots,\epsilon_k,0) \quad \wedge \quad \sum_{t=1}^k \delta_t \leq \delta'_g \right] \leq \delta_g + \delta'_g.$$

where each ϵ_i is a function of the outputs of the prefix $\widetilde{V}_{<i}$ of the full view $\widetilde{V}$ from $\widetilde{\texttt{SimulatedComp}}(\mathcal{A},k,0)$. We now show that the quantity that we want to bound can be written as the probability of the coupled random variables $\boldsymbol{V}$ and $\boldsymbol{W} = \left(W, \widetilde{W}\right)$ being

contained in the sets that we defined above.

$$\begin{aligned}
&\Pr_{\boldsymbol{V}\sim\left(P(W),P(\widetilde{W})\right)}\left[|\mathrm{PrivLoss}(V)| \geq \widetilde{\texttt{COMP}}_{\delta_g}\left(\epsilon_1, 0, \cdots, \epsilon_k, 0\right) \quad \wedge \quad \sum_{t=1}^{k}\delta_t \leq \delta'_g\right] \\
&\leq \Pr\left[(\boldsymbol{W} \in \mathcal{F} \quad \wedge \quad V \in \mathcal{G}_k) \quad \vee \quad (V \in \mathcal{H} \quad \wedge \quad \boldsymbol{W} \notin \mathcal{F})\right] \\
&\leq \Pr\left[\boldsymbol{W} \in \mathcal{F} \quad \wedge \quad V \in \mathcal{G}_k\right] + \Pr\left[\widetilde{V} \in \mathcal{H}\right] \\
&\leq \Pr\left[\boldsymbol{W} \in \mathcal{F} \quad \wedge \quad V \in \mathcal{G}_k\right] + \delta_g \qquad (5.6)
\end{aligned}$$

Note, that if $\sum_{i=1}^{k}\delta_i(\boldsymbol{v}_{<i}) \leq \delta_g$ then we must have $\sum_{i=1}^{t}\delta_i(\boldsymbol{v}_{<i}) \leq \delta_g$ for each $t < k$, so that $\mathcal{G}_k \subseteq \mathcal{G}_t$. We then use the fact that $\{\mathcal{F}_t : t \in [k]\}$ forms a partition of $\mathcal{F}$, i.e. $\mathcal{F} = \bigcup_{t=1}^{k}\mathcal{F}_t$ and $\mathcal{F}_i \cap \mathcal{F}_j = \emptyset$ for $i \neq j$, to obtain the following:

$$\Pr\left[\boldsymbol{W} \in \mathcal{F} \quad \wedge \quad V \in \mathcal{G}_k\right] = \sum_{t=1}^{k}\Pr\left[\boldsymbol{W} \in \mathcal{F}_t \quad \wedge \quad V \in \mathcal{G}_k\right] \leq \sum_{t=1}^{k}\Pr\left[\boldsymbol{W} \in \mathcal{F}_t \quad \wedge \quad V \in \mathcal{G}_t\right].$$

Focusing on each term in the summation, we get

$$\begin{aligned}
&\sum_{t=1}^{k}\Pr\left[\boldsymbol{W} \in \mathcal{F}_t \quad \wedge \quad V \in \mathcal{G}_t\right] \\
&\qquad \leq \sum_{t=1}^{k}\Pr\left[\boldsymbol{W} \in \mathcal{F}_t \quad \wedge \quad \widetilde{V} \in \mathcal{G}_t\right] \\
&\qquad = \sum_{t=1}^{k}\sum_{\widetilde{\boldsymbol{v}} \in \mathcal{G}_t}\Pr\left[\widetilde{V} = \widetilde{\boldsymbol{v}}\right]\Pr\left[\boldsymbol{W} \in \mathcal{F}_t | \widetilde{\boldsymbol{v}}\right] \\
&\qquad \leq \sum_{t=1}^{k}\sum_{\widetilde{\boldsymbol{v}} \in \mathcal{G}_t}\Pr\left[\widetilde{V} = \widetilde{\boldsymbol{v}}\right]\delta_t(\widetilde{\boldsymbol{v}}_{<t}).
\end{aligned}$$

We now switch the order of summation to obtain our result

$$\begin{aligned}
&\sum_{t=1}^{k}\sum_{\widetilde{\boldsymbol{v}} \in \mathcal{G}_t}\Pr\left[\widetilde{V} = \widetilde{\boldsymbol{v}}\right]\delta_t(\widetilde{\boldsymbol{v}}_{<t}) \\
&\qquad = \sum_{\widetilde{\boldsymbol{v}}}\Pr\left[\widetilde{V} = \widetilde{\boldsymbol{v}}\right]\sum_{t:\sum_{i=1}^{t}\delta_i(\widetilde{\boldsymbol{v}}_{<i}) \leq \delta'_g}\delta_t(\widetilde{\boldsymbol{v}}_{<t}) \leq \sum_{\widetilde{\boldsymbol{v}}}\Pr\left[\widetilde{V} = \widetilde{\boldsymbol{v}}\right]\delta'_g = \delta'_g. \qquad (5.7)
\end{aligned}$$

We then combine this with (5.6) to prove our first statement for the privacy odometer.

Using the same notation as above, we now move to proving the statement for the privacy filter. It suffices to prove the following where the randomness is over $\boldsymbol{V} \sim (P(W), P(\widetilde{W}))$:

$$\Pr\left[|\mathrm{PrivLoss}(V)| \geq \epsilon_g \quad \wedge \quad V \in \mathcal{G}_k \quad \wedge \quad \widetilde{\texttt{COMP}}_{\epsilon_g,\delta_g}\left(\epsilon_1, 0, \cdots, \epsilon_k, 0\right) = \texttt{CONT}\right] \leq \delta_g + \delta'_g.$$

We now define a slight variant of $\mathcal{H}$ from above:

$$\mathcal{H}_{\epsilon_g} \overset{\text{defn}}{=} \left\{\boldsymbol{v} : \left|\widetilde{\mathrm{PrivLoss}}(\boldsymbol{v})\right| \geq \epsilon_g\right\}.$$

Similar to what we showed in (5.6) for the privacy odometer, we have

$$\begin{aligned}
&\Pr_{\boldsymbol{V}}\left[|\mathrm{PrivLoss}(V)| \geq \epsilon_g \quad \wedge \quad V \in \mathcal{G}_k \quad \wedge \quad \widetilde{\texttt{COMP}}_{\epsilon_g,\delta_g}(\epsilon_1, 0, \cdots, \epsilon_k, 0) = \texttt{CONT}\right] \\
&\quad \leq \Pr\left[\left((\boldsymbol{W} \in \mathcal{F} \wedge V \in \mathcal{G}_k) \vee (V \in \mathcal{H}_{\epsilon_g} \wedge \boldsymbol{W} \notin \mathcal{F})\right) \quad \wedge \quad \widetilde{\texttt{COMP}}_{\epsilon_g,\delta_g}(\epsilon_1, 0, \cdots, \epsilon_k, 0) = \texttt{CONT}\right] \\
&\quad \leq \Pr\left[\boldsymbol{W} \in \mathcal{F} \quad \wedge \quad V \in \mathcal{G}_k\right] + \Pr\left[\widetilde{V} \in \mathcal{H}_{\epsilon_g} \quad \wedge \quad \widetilde{\texttt{COMP}}_{\epsilon_g,\delta_g}(\epsilon_1, 0, \cdots, \epsilon_k, 0) = \texttt{CONT}\right] \\
&\quad \leq \Pr\left[\boldsymbol{W} \in \mathcal{F} \quad \wedge \quad V \in \mathcal{G}_k\right] + \delta_g \\
&\quad \leq \delta_g' + \delta_g
\end{aligned}$$

where the last inequality follows from (5.6) and (5.7). □

5.3.3. Basic Composition

We first give an adaptive parameter version of the basic composition in Theorem 5.2.4.
Theorem 5.3.7. *For each $\delta_g \geq 0$, $\texttt{COMP}_{\delta_g}$ is a valid privacy odometer where*

$$\texttt{COMP}_{\delta_g+\delta_g'}(\epsilon_1, \delta_1, \cdots, \epsilon_k, \delta_k) = \begin{cases} \infty & \text{if } \sum_{i=1}^k \delta_i > \delta_g' \\ \sum_{i=1}^k \epsilon_i & \text{otherwise} \end{cases}.$$

Additionally, for any $\epsilon_g, \delta_g \geq 0$, $\texttt{COMP}_{\epsilon_g,\delta_g}$ is a valid privacy filter where

$$\texttt{COMP}_{\epsilon_g,\delta_g}(\epsilon_1, \delta_1, \cdots, \epsilon_k, \delta_k) = \begin{cases} \texttt{HALT} & \text{if } \sum_{i=1}^k \delta_i > \delta_g' \text{ or } \sum_{i=1}^k \epsilon_i > \epsilon_g \\ \texttt{CONT} & \text{otherwise} \end{cases}.$$

Proof. We use Lemmas 5.3.5 and 5.3.6 so that we need to only reason about any deterministic adversary in $\widetilde{\texttt{SimulatedComp}}(\mathcal{A}, k, b)$. We know that $(\epsilon, 0)$-DP is closed under post-processing from Theorem 2.1.3, so that for any (randomized) post-processing function T, we have $T(\texttt{RR}_\epsilon(0)) \approx_{\epsilon,0} T(\texttt{RR}_\epsilon(1))$ and thus we know that $T(\texttt{RR}_\epsilon(0))$ and $T(\texttt{RR}_\epsilon(1))$ are $(\epsilon, 0)$-point-wise indistinguishable for any post-processing function T. The proof then follows simply from the definition of (pure) differential privacy, so for all possible views $\widetilde{\boldsymbol{v}}$ of the adversary in $P\left(\widetilde{\texttt{SimulatedComp}}(\mathcal{A}, k, b)\right)$:

$$\left|\widetilde{\mathrm{PrivLoss}}(\widetilde{\boldsymbol{v}})\right| \leq \sum_{i=1}^k \left|\log\left(\frac{\Pr\left[P_i\left(\texttt{RR}_{\epsilon_i}(0); R^0_{\texttt{SIM}_i}\right) = \widetilde{v}_i | \widetilde{\boldsymbol{v}}_{<i}\right]}{\Pr\left[P_i\left(\texttt{RR}_{\epsilon_i}(1); R^1_{\texttt{SIM}_i}\right) = \widetilde{v}_i | \widetilde{\boldsymbol{v}}_{<i}\right]}\right)\right| \leq \sum_{i=1}^k \epsilon_i(\widetilde{\boldsymbol{v}}_{<i})$$

where we explicitly write the dependence of the choice of ϵ_i by $\mathcal{M}$ at round i on the view from the previous rounds as $\epsilon_i(\widetilde{\boldsymbol{v}}_{<i})$ □

We give a useful concentration bound that will be pivotal in proving an improved valid privacy odometer and filter from that given in Theorem 5.3.7. We first present a concentration bound for *self normalized processes.*

Theorem 5.4.1 [See Corollary 2.2 in de la Peña et al. (2004)]**.** *If B and $C > 0$ are two random variables such that*

$$\mathbb{E}\left[\exp\left(\lambda B - \frac{\lambda^2}{2}C^2\right)\right] \leq 1 \tag{5.8}$$

for all $\lambda \in \mathbb{R}$, then for all $\delta \leq 1/e$, $x > 0$ we have

$$\Pr\left[|B| \geq \sqrt{(C^2 + x)\left(2 + \log\left(\frac{C^2}{x} + 1\right)\right)\log(1/\delta)}\right] \leq \delta.$$

To put this bound into context, suppose that C is a constant and we apply the bound with $x = C^2$. Then the bound simplifies to

$$\Pr\left[|B| \geq O\big(C\sqrt{\log(1/\delta)}\big)\right] \leq \delta,$$

which is just a standard concentration inequality for any subgaussian random variable B with standard deviation C.

Another bound which will be useful for our results is the following.

Theorem 5.4.2 [See Theorem 2.4 in Chen et al. (2014)]**.** *If B and $C > 0$ are two random variables that satisfy* (5.8) *for all $\lambda \in \mathbb{R}$, then for all $c > 0$ and $s, t \geq 1$ we have*

$$\Pr\left[|B| \geq s\; C \quad c \leq C \leq t \cdot c\right] \leq 2\sqrt{e}\left(1 + 2s\log(t)\right)e^{-s^2/2}.$$

We will apply both Theorems 5.4.1 and 5.4.2 to random variables coming from martingales defined from the privacy loss functions.

To set this up, we present some notation: let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple where $\emptyset = \mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}$ is an increasing sequence of σ-algebras. Let X_i be a real-valued $\mathcal{F}_i$-measurable random variable, such that $\mathbb{E}\left[X_i | \mathcal{F}_{i-1}\right] = 0$ a.s. for each i. We then consider the martingale where

$$M_0 = 0 \qquad M_k = \sum_{i=1}^{k} X_i, \qquad \forall k \geq 1. \tag{5.9}$$

We then use the following result which gives us a pair of random variables to which we can apply either Theorem 5.4.1 or Theorem 5.4.2.

Lemma 5.4.3 [See Lemma 2.4 in van de Geer (2002)]**.** *For M_k defined in* (5.9)*, if there exists two random variables $D_i < D_i'$ that are $\mathcal{F}_{i-1}$-measurable for $i \geq 1$*

$$D_i \leq X_i \leq D_i' \qquad a.s. \quad \forall i \geq 1.$$

and we define U_k *as*

$$U_0^2 = 0, \qquad U_k^2 = \sum_{i=1}^k \left(D_i' - D_i\right)^2, \quad \forall k \geq 1 \tag{5.10}$$

then

$$\exp\left[\lambda M_k - \frac{\lambda^2}{8} U_k^2\right]$$

is a supermartingale for all $\lambda \in \mathbb{R}$.

We then obtain the following result from combining Theorem 5.4.1 with Lemma 5.4.3.
Theorem 5.4.4. *Let* M_k *be defined as in* (5.9) *and satisfy the hypotheses of Lemma 5.4.3. Then for every fixed* $k \geq 1$, $x > 0$ *and* $\delta \leq 1/e$, *we have*

$$\Pr\left[|M_k| \geq \sqrt{\left(\frac{U_k^2}{4} + x\right)\left(2 + \log\left(\frac{U_k^2}{4x} + 1\right)\right)\log(1/\delta)}\right] \leq \delta$$

Similarly, we can obtain the following result by combining Theorem 5.4.2 with Lemma 5.4.3.
Theorem 5.4.5. *Let* M_k *be defined as in* (5.9) *and satisfy the hypotheses of Lemma 5.4.3. Then for every fixed* $k \geq 1$, $0 < \beta < 1$ $c > 0$ *and* $t \geq 1$, *we have*

$$\Pr\left[|M_k| \geq \sqrt{\frac{U_k^2}{2}\log(1/\beta)} \quad \textit{and} \quad c \leq \sqrt{\frac{U_k^2}{4}} \leq t \cdot c\right] \leq 2\sqrt{e}\left(\beta + 2\log(t)\sqrt{0.74\ \beta}\right)$$

Proof. We use the fact that $x\log(1/x)$ is maximized at $x = 1/e$ and has value no more than 0.37. ☐

Given Lemma 5.3.6, we will focus on finding a valid privacy odometer and filter when $\{\delta_i\} \equiv 0$. Our analysis will then depend on the privacy loss $\widetilde{\mathrm{PrivLoss}}(\widetilde{V})$ from (5.5) where $\widetilde{V}$ is the view of the adversary in $\widetilde{\texttt{SimulatedComp}}(\mathcal{A}, k, 0)$. We then focus on the following martingale in our analysis:

$$\widetilde{M}_k = \sum_{i=1}^k \left(\widetilde{\mathrm{PrivLoss}}_i(\widetilde{V}_{\leq i}) - \widetilde{\mu}_i\right) \qquad \text{where} \qquad \widetilde{\mu}_i = \mathbb{E}\left[\widetilde{\mathrm{PrivLoss}}_i(\widetilde{V}_{\leq i}) \middle| \widetilde{V}_{<i}\right]. \tag{5.11}$$

We can then bound the conditional expectation $\widetilde{\mu}_i$ with the following result from Dwork and Rothblum (2016) that improves on an earlier result from Dwork et al. (2010) by a factor of 2.
Lemma 5.4.6 [Dwork and Rothblum (2016)]. *For* $\widetilde{\mu}_i$ *defined in* (5.11), *we have* $\widetilde{\mu}_i \leq \epsilon_i\left(e^{\epsilon_i} - 1\right)/2$.

We next show that we can essentially get the same asymptotic bound as Theorem 2.1.5 for the privacy filter setting using the bound in Theorem 5.4.4 for the martingale given in (5.11).

Theorem 5.5.1. *We define $\mathcal{K}$ as the following where we set $x = \frac{\epsilon_g^2}{28.04 \cdot \log(1/\delta_g)}$ and,*[2]

$$\mathcal{K} \stackrel{\text{defn}}{=} \sum_{j=1}^{k} \epsilon_j \left(\frac{e^{\epsilon_j} - 1}{2} \right) + \sqrt{2 \left(\sum_{i=1}^{k} \epsilon_i^2 + x \right) \left(1 + \frac{1}{2} \log \left(\frac{\sum_{i=1}^{k} \epsilon_i^2}{x} + 1 \right) \right) \log(2/\delta_g)}. \quad (5.12)$$

$\texttt{COMP}_{\epsilon_g,\delta_g}$ *is a valid privacy filter for $\delta_g \in (0, 1/e)$ and $\epsilon_g > 0$ where*

$$\texttt{COMP}_{\epsilon_g,\delta_g}(\epsilon_1, \delta_1, \cdots, \epsilon_k, \delta_k) = \begin{cases} \texttt{HALT} & \text{if } \sum_{i=1}^{k} \delta_i > \delta_g/2 \text{ or } \mathcal{K} > \epsilon_g \\ \texttt{CONT} & \text{otherwise} \end{cases}.$$

Note that if we have $\sum_{i=1}^{k} \epsilon_i^2 = O\left(1/\log(1/\delta_g)\right)$ and set $\epsilon_g = \Theta\left(\sqrt{\sum_{i=1}^{k} \epsilon_i^2 \log(1/\delta_g)}\right)$ in (5.12), we are then getting the same asymptotic bound on the privacy loss as in Kairouz et al. (2015) and in Theorem 2.1.5 for the case when $\epsilon_i = \epsilon$ for $i \in [k]$. If $k\epsilon^2 \leq \frac{1}{8\log(1/\delta_g)}$, then Theorem 2.1.5 gives a bound on the privacy loss of $\epsilon\sqrt{8k\log(1/\delta_g)}$. Note that there may be better choices for the constant 28.04 that we divide ϵ_g^2 by in (5.12), but for the case when $\epsilon_g = \epsilon\sqrt{8k\log(1/\delta_g)}$ and $\epsilon_i = \epsilon$ for every $i \in [n]$, it is nearly optimal.

Proof of Theorem 5.5.1. Note that Lemma 5.3.6 allows us to concentrate on showing that we can find an optimal privacy filter when $\{\delta_i\} \equiv 0$. We then focus on the martingale $\widetilde{M}_k$ given in (5.11). In order to apply Theorem 5.4.4 we set the lower bound for $\widetilde{M}_i$ to be $D_i = (-\epsilon_i - \widetilde{\mu}_i)$ and upper bound to be $D_i' = (\epsilon_i - \widetilde{\mu}_i)$ in order to compute U_k^2 from (5.10). We then have for the martingale in (5.11) that

$$U_k^2 = 4 \sum_{i=1}^{k} \epsilon_i^2.$$

We can then directly apply Theorem 5.4.4 to get the following for $x = \frac{\epsilon_g^2}{28.04 \cdot \log(1/\delta_g)} > 0$ with probability at least $1 - \delta_g/2$

$$|\widetilde{M}_k| \leq \sqrt{2 \left(\sum_{i=1}^{k} \epsilon_i^2 + x \right) \left(1 + \frac{1}{2} \log \left(\frac{\sum_{i=1}^{k} \epsilon_i^2}{x} + 1 \right) \right) \log(2/\delta_g)}.$$

We can then obtain a bound on the privacy loss with probability at least $1 - \delta_g/2$ over

---

[2] We thank Daniel Winograd-Cort for catching an incorrectly set constant in an earlier version of this theorem.

$\widetilde{\boldsymbol{v}} \sim \widetilde{V}^0$

$$\begin{aligned}\left|\widetilde{\mathrm{PrivLoss}}(\widetilde{\boldsymbol{v}})\right| &\leq \sum_{i=1}^k \widetilde{\mu}_i + \sqrt{2\left(\sum_{i=1}^k \epsilon_i^2 + x\right)\left(1 + \frac{1}{2}\log\left(\frac{\sum_{i=1}^k \epsilon_i^2}{x} + 1\right)\right)\log(2/\delta_g)} \\ &\leq \sum_{i=1}^k \epsilon_i\left(e^{\epsilon_i} - 1\right)/2 + \sqrt{2\left(\sum_{i=1}^k \epsilon_i^2 + x\right)\left(1 + \frac{1}{2}\log\left(\frac{\sum_{i=1}^k \epsilon_i^2}{x} + 1\right)\right)\log(2/\delta_g)}.\end{aligned}$$

□

5.6. Advanced Composition for Privacy Odometers

One might hope to achieve the same sort of bound on the privacy loss from Theorem 5.2.5 when the privacy parameters may be chosen adversarially. However we show that this cannot be the case for any valid privacy odometer. In particular, even if an adversary selects the same privacy parameter $\epsilon = o(\sqrt{\log(\log(n)/\delta_g)/k})$ each round but can adaptively select a time to stop interacting with `AdaptParamComp` (which is a restricted special case of the power of the general adversary – stopping is equivalent to setting all future $\epsilon_i, \delta_i = 0$), then we show that there can be no valid privacy odometer achieving a bound of $o(\epsilon\sqrt{k\log\left(\log(n)/\delta_g\right)})$. This gives a separation between the achievable bounds for a valid privacy odometers and filters. But for privacy applications, it is worth noting that δ_g is typically set to be (much) smaller than $1/n$, in which case this gap disappears (since $\log(\log(n)/\delta_g) = (1 + o(1))\log(1/\delta_g)$).

Theorem 5.6.1. *For any $\delta_g \in (0, O(1))$ there is no valid $\mathtt{COMP}_{\delta_g}$ privacy odometer where*

$$\mathtt{COMP}_{\delta_g}(\epsilon_1, 0, \cdots, \epsilon_k, 0) = \sum_{i=1}^k \epsilon_i\left(\frac{e^{\epsilon_i} - 1}{e^{\epsilon_i} + 1}\right) + o\left(\sqrt{\sum_{i=1}^k \epsilon_i^2 \log(\log(n)/\delta_g)}\right) \tag{5.13}$$

In order to prove Theorem 5.6.1, we use the following anti-concentration bound for a sum of random variables.

Lemma 5.6.2 [See Lemma 8.1 in Ledoux and Talagrand (1991)]**.** *Let $X_1, \cdots, X_k$ be a sequence of mean zero i.i.d. random variables such that $|X_1| < a$ and $\sigma^2 = \mathbb{E}\left[X_1^2\right]$. For every $\alpha > 0$ there exists two positive constants c_α and c'_α such that for every x satisfying $\sqrt{k}\sigma c_\alpha \leq x \leq c'_\alpha \frac{k\sigma^2}{a}$ we have*

$$\Pr\left[\sum_{i=1}^k X_i \geq x\right] \geq \exp\left[-(1+\alpha)\frac{x^2}{2k\sigma^2}\right]$$

For $\gamma \in [1/2, 1)$, we define the random variables $\xi_i \in \{-1, 1\}$ where

$$\Pr\left[\xi_i = 1\right] = \gamma \quad \Pr\left[\xi_i = -1\right] = 1 - \gamma. \tag{5.14}$$

Note that $\mathbb{E}\left[\xi_i\right] \overset{\text{defn}}{=} \mu = 2\gamma - 1$ and $\mathbb{V}\left[\xi_i\right] \overset{\text{defn}}{=} \sigma^2 = 1 - \mu^2$. We then consider the sequence of i.i.d. random variables $X_1, \cdots, X_n$ where $X_i = (\xi_i - \mathbb{E}\left[\xi_i\right])$. We denote the sum of X_i as

$$M_n = \sum_{i=1}^{n} X_i. \tag{5.15}$$

We then apply Lemma 5.6.2 to prove an anti-concentration bound for the martingale given above.

Lemma 5.6.3 [Anti-Concentration]**.** *Consider the partial sums M_t defined in* (5.15) *for $t \in [n]$. There exists a constant c such that for all $\delta \in (0, O(1))$ and $n > \Omega\left(\log(1/\delta) \cdot \left(\frac{1+\mu}{\sigma}\right)^2\right)$ we have*

$$\Pr\left[\exists t \in [n] \text{ s.t. } M_t \geq c \cdot \sigma\sqrt{t \log(\log(n)/\delta)}\right] \geq \delta.$$

Proof. By Lemma 5.6.2, we know that there exists constants c_1, c_2, c_3 and large N such that for all $m > N \cdot \left(\frac{1+\mu}{\sigma}\right)^2$ and $x \in \left[1, c_3\sqrt{m}\frac{\sigma}{1+\mu}\right]$, we have

$$\Pr\left[\sum_{i=1}^{m} X_i \geq c_1\sqrt{m}\sigma x\right] \geq e^{-c_2 x^2}.$$

Rather than consider every possible $t \in [n]$, we consider $j \in \left\{m_\delta, m_\delta^2, \cdots, m_\delta^{\lfloor \log_{m_\delta}(n) \rfloor}\right\}$ where $m_\delta \in \mathbb{N}$ and $m_\delta > m\log(1/\delta)$. We then have for a constant c that

$$\begin{aligned}
\Pr\left[\exists t \in [n] \text{ s.t. } M_t \geq c\sigma\sqrt{t\log(1/\delta)}\right] &\geq \Pr\left[\exists j \in \left[\lfloor \log_{m_\delta}(n) \rfloor\right] \text{ s.t. } M_{m_\delta^j} \geq c\sigma\sqrt{m_\delta^j \log(1/\delta)}\right] \\
&= \sum_{j=1}^{\lfloor \log_{m_\delta}(n) \rfloor} \Pr\left[M_{m_\delta^j} \geq c\sigma\sqrt{m_\delta^j \log(1/\delta)} \,\middle|\, M_{m_\delta^\ell} \leq c\sigma\sqrt{m_\delta^\ell \log(1/\delta)} \quad \forall \ell < j\right] \\
&\geq \sum_{j=1}^{\lfloor \log_{m_\delta}(n) \rfloor} \Pr\left[M_{m_\delta^j} \geq c\sigma\left(\sqrt{m_\delta^j \log(1/\delta)} + \sqrt{m_\delta^{j-1}\log(1/\delta)}\right)\right] \\
&= \sum_{j=1}^{\lfloor \log_{m_\delta}(n) \rfloor} \Pr\left[M_{m_\delta^j} \geq c\sigma\left(1 + 1/\sqrt{m_\delta}\right)\sqrt{m_\delta^j \log(1/\delta)}\right] \\
&\geq \sum_{j=1}^{\lfloor \log_{m_\delta}(n) \rfloor} \Pr\left[M_{m_\delta^j} \geq 2c\sigma\sqrt{m_\delta^j \log(1/\delta)}\right]
\end{aligned}$$

Thus, we set $c = \frac{c_1}{2\sqrt{c_2}}$ and then for any δ such that $\sqrt{c_2} < \sqrt{\log(1/\delta)} < c_3\sqrt{c_2}\sqrt{m_\delta}\frac{\sigma}{1+\mu}$, we have

$$\Pr\left[\exists t \in [n] \text{ s.t. } M_t \geq c\sigma\sqrt{t\log(1/\delta)}\right] \geq \lfloor \log_{m_\delta}(n) \rfloor \delta.$$

□

Algorithm 6 Stopping Time Adversary: $\mathcal{A}_{\epsilon,\delta}$

```
Input: privacy parameters $(\epsilon, \delta)$ and constant $c$
  for $i = 1, \cdots, k$ do
    $\mathcal{A}_{\epsilon,\delta} = \mathcal{A}_{\epsilon,\delta}(c, Y_1, \cdots, Y_{i-1})$ gives datasets $\{0,1\}$, parameter $(\epsilon, 0)$ and $\texttt{RR}_\epsilon$ to AdaptParamComp.
    $\mathcal{A}_{\epsilon,\delta}$ receives $Y_i \in \{\top, \bot\}$.
    if $Y_i = \top$ then
      $X_i = \epsilon$
    else
      $X_i = -\epsilon$
    if $\sum_{j=1}^{i} \left(X_j - \epsilon\frac{e^\epsilon - 1}{e^\epsilon + 1}\right) \geq c \cdot \left(\epsilon\sqrt{i \log(\log(n)/\delta)}\right)$, then
      $\epsilon_{i+1}, \cdots \epsilon_k = 0$
      BREAK
```

We next use Lemma 5.6.3 to prove that we cannot have a bound like Theorem 5.2.5 in the adaptive privacy parameter setting, which uses the *stopping time adversary* given in Algorithm 6.

Proof of Theorem 5.6.1. Consider the stopping time adversary $\mathcal{A}_{\epsilon,\delta_g}$ from Algorithm 6 for a constant c that we will determine in the proof. Let the number of rounds $k = n$ and $\epsilon = 1/n$. In order to use Lemma 5.6.3 we define $\gamma = \frac{e^\epsilon}{1+e^\epsilon}$ from (5.14). Because we let ϵ depend on n, we have $\mu \equiv \mu_n = \frac{e^{1/n}-1}{e^{1/n}+1} = O(1/n)$ and $\sigma \equiv \sigma_n = 1 - \mu_n^2 = 1 - O(1/n^2)$ which gives $\frac{1+\mu_n}{\sigma_n} = \Theta(1)$. We then relate the martingale in (5.15) with the privacy loss for this particular adversary in $\texttt{AdaptParamComp}(\mathcal{A}_{\epsilon,\delta_g}, n, 0)$ with view V who sets $X_t = \pm\epsilon$ each round,

$$\sum_{j=1}^{t} \left(X_j - \frac{\mu_n}{n}\right) = \frac{1}{n} M_t \qquad \forall t \in [n].$$

Hence, at any round t if $\mathcal{A}_{\epsilon,\delta_g}$ finds that

$$\frac{1}{n} M_t \geq c \left(\frac{1}{n}\sqrt{t \log(\log(n)/\delta_g)}\right) \tag{5.16}$$

then she will set all future $\epsilon_i = 0$ for $i > t$. To find the probability that (5.16) holds in any round $t \in [n]$ we use Lemma 5.6.3 with the constant c from the lemma statement to say that (5.16) occurs with probability at least δ_g.

Assume that $\texttt{COMP}_{\epsilon_g}$ is a valid privacy odometer and (5.13) holds. We then know that with probability at least $1-\delta_g$ over $\boldsymbol{v} \sim V^b$ where V^b is the view for $\texttt{AdaptParamComp}(\mathcal{A}_{1/n,\delta_g}, n, b)$

$$|\text{PrivLoss}(\boldsymbol{v})| \leq \texttt{COMP}_{\delta_g}\left(\epsilon_1, 0, \cdots, \epsilon_k, 0\right)$$

$$\implies \left|\sum_{i=1}^{t} \text{PrivLoss}_i(\boldsymbol{v}_{\leq i})\right| = t \cdot \frac{\mu_n}{n} + o\left(\frac{1}{n}\sqrt{t \log\left(\frac{\log(n)}{\delta_g}\right)}\right) \qquad \forall t \in [n]$$

But this is a contradiction given that the bound in (5.16) at any round $t \in [n]$ occurs with probability at least δ_g. □

We now utilize the bound from Theorem 5.4.4 to obtain a concentration bound on the privacy loss.

Lemma 5.6.4. *For $\delta_g \in (0, 1/e)$, $\texttt{COMP}_{\delta_g}$ is a valid privacy odometer where $\texttt{COMP}_{\delta_g}(\epsilon_1, \delta_1, \cdots, \epsilon_k, \delta_k) = \infty$ if $\sum_{i=1}^k \delta_i > \delta_g/2$ and otherwise for any $x > 0$,*

$$\begin{aligned}&\texttt{COMP}_{\delta_g}(\epsilon_1, \delta_1, \cdots, \epsilon_k, \delta_k)\\ &\quad = \sum_{j=1}^{k} \epsilon_j\left(e^{\epsilon_j} - 1\right)/2 + \sqrt{2\left(\sum_{i=1}^{k}\epsilon_i^2 + x\right)\left(1 + \frac{1}{2}\log\left(\frac{\sum_{i=1}^k \epsilon_i^2}{x} + 1\right)\right)\log(2/\delta_g)}.\end{aligned}$$

Proof. We will follow a similar argument as in Theorem 5.5.1 where we use the same martingale $\widetilde{M}_k$ from (5.11). We can then directly apply Theorem 5.4.4 to get the following for any $x > 0$ with probability at least $1 - \delta_g/2$

$$\left|\widetilde{M}_k\right| \leq \sqrt{2\left(\sum_{i=1}^{k}\epsilon_i^2 + x\right)\left(1 + \frac{1}{2}\log\left(\frac{\sum_{i=1}^k \epsilon_i^2}{x} + 1\right)\right)\log(2/\delta_g)}$$

□

This above result is only useful if we can plug in a constant $x > 0$. If we were to set $x = \sum_{i=1}^k \epsilon_i^2$, then we would get asymptotically close to the same bound as in Theorem 2.1.5, however, the ϵ_i are random variables, and their realizations cannot be used in setting x; further, we know from Theorem 5.6.1 that such a bound cannot hold in this setting. One particular setting for x might be ϵ_1^2 because this parameter must be a constant – it is chosen prior to any interaction with the data.

We now give our main positive result for privacy odometers, which is similar to our privacy filter in Theorem 5.5.1 except that δ_g is replaced by $\delta_g/\log(n)$, as is necessary from Theorem 5.6.1. Note that the bound incurs an additive $1/n^2$ loss to the $\sum_i \epsilon_i^2$ term that is present without privacy. In any reasonable setting of parameters, this translates to at most a constant-factor multiplicative loss, because there is no utility running any differentially private algorithm with $\epsilon_i < \frac{1}{10n}$ (we know that if $\mathcal{M}$ is $(\epsilon_i, 0)$-DP then $\mathcal{M}(\boldsymbol{x})$ and $\mathcal{M}(\boldsymbol{x}')$ for any pair of inputs have statistical distance at most $e^{\epsilon_i n} - 1 < 0.1$, and hence the output is essentially independent of the input - note that a similar statement holds for (ϵ_i, δ_i)-DP.)

Theorem 5.6.5 [Advanced Privacy Odometer [3]]**.** *$\texttt{COMP}_{\delta_g}$ is a valid privacy odometer for $\delta_g \in (0, 1/e)$ where $\texttt{COMP}_{\delta_g}(\epsilon_1, \delta_1, \cdots, \epsilon_k, \delta_k) = \infty$ if $\sum_{i=1}^k \delta_i > \delta_g/2$, otherwise if $\sum_{i=1}^k \epsilon_i^2 \in$*

---

[3] This bound is different from what appeared in Rogers et al. (2016b), which had

$[1/n^2, 1]$ *then*

$$\texttt{COMP}_{\delta_g}\left(\epsilon_1, \delta_1, \cdots, \epsilon_k, \delta_k\right) = \sum_{i=1}^k \epsilon_i \left(\frac{e^{\epsilon_i} - 1}{2}\right) + \sqrt{2 \sum_{i=1}^k \epsilon_i^2 \left(\log(48e) + 2\log\left(\frac{\log(n)}{\delta_g}\right)\right)} \tag{5.17}$$

and if $\sum_{i=1}^k \epsilon_i^2 \notin [1/n^2, 1]$ *then* $\texttt{COMP}_{\delta_g}\left(\epsilon_1, \delta_1, \cdots, \epsilon_k, \delta_k\right)$ *is equal to*

$$\sum_{i=1}^k \epsilon_i \left(\frac{e^{\epsilon_i} - 1}{2}\right) + \sqrt{2\left(1/n^2 + \sum_{i=1}^k \epsilon_i^2\right)\left(1 + \frac{1}{2}\log\left(1 + n^2 \sum_{i=1}^k \epsilon_i^2\right)\right)\log(4\log_2(n)/\delta_g)}. \tag{5.18}$$

Proof. We again focus on a valid privacy odometer for $\{\delta_i\} \equiv 0$ and the martingale $\widetilde{M}_k$ from (5.11). We first focus on proving the bound in (5.17). For the martingale $\widetilde{M}_k$ in (5.11), we can use Theorem 5.4.5 to get the following for any $\beta > 0$

$$\begin{aligned}\Pr&\left[|\widetilde{M}_k| \ge \sum_{i=1}^k \epsilon_i \left(\frac{e^{\epsilon_i} - 1}{2}\right) + \sqrt{2\sum_{i=1}^k \epsilon_i^2 \log\left(1/\beta\right)} \quad \text{and} \quad \frac{1}{n} \le \sqrt{\sum_{i=1}^k \epsilon_i^2} \le 1\right] \\ &\le 2\sqrt{e}\left(\beta + 2\log(n)\sqrt{0.74\ \beta}\right).\end{aligned}$$

We then solve for β so that $\delta_g/2 = 2\sqrt{e}\left(\beta + 2\log(n)\sqrt{0.74\ \beta}\right)$, which yields,

$$\beta = 0.74\ \log^2(n)\left(\sqrt{1 + \frac{\delta_g}{2.96\sqrt{e}\log^2(n)}} - 1\right)^2 \le \frac{3}{4}\ \log^2(n)\left(\frac{\delta_g}{6\sqrt{e}\log^2(n)}\right)^2.$$

This gives the stated bound in (5.17).

For the bound given in (5.18), we set $x = 1/n^2$ in Lemma 5.6.4. Hence, we would have with probability at least $1 - \delta_g/2$ when $\sum_{i=1}^k \epsilon_i^2 \notin [1/n^2, 1]$,

$$|\widetilde{M}_k| \le \sum_{j=1}^k \sqrt{2\left(1/n^2 + \sum_{i=1}^k \epsilon_i^2\right)\left(1 + \frac{1}{2}\log\left(1 + n^2\sum_{i=1}^k \epsilon_i^2\right)\right)\log(4\log_2(n)/\delta_g)}.$$

□

In the above theorem, we only allow privacy parameters such that $\sum_{i=1}^k \epsilon_i^2 \in [1/n^2, 1]$. This assumption is not too restrictive, since the output of a single $(\ll 1/n)$-differentially private

---

$2\sqrt{\sum_{i=1}^k \epsilon_i^2\left(1 + \log(\sqrt{3})\right)\log\left(\frac{4\log_2(n)}{\delta_g}\right)}$ instead of the term $\sqrt{2\sum_{i=1}^k \epsilon_i^2\left(\log(48e) + 2\log\left(\frac{\log(n)}{\delta_g}\right)\right)}$ in (5.17), which is an improvement when $\delta_g < 1/4$ and $n \ge 40$.

algorithm is nearly independent of its input. More generally, we can replace $1/n^2$ with an arbitrary "granularity parameter" γ and require that $\sum_{i=1}^k \epsilon_i^2 \in [\gamma, 1]$. When doing so, $\log(n)$ in (5.17) will be replaced with $\frac{1}{2}\log(1/\gamma)$. For example, we could require that $\epsilon_1 \geq \delta_g$, in which case we can choose $\gamma = \delta_g^2$, which would not affect our bound substantially.

5.7. zCDP Filters and Odometers

We now extend the idea of privacy odometers to ρ-zCDP when the parameters can be chosen adaptively. We return to the definition of zCDP, presented in Definition 2.2.1, except we write it in terms of Rényi divergence (Renyi, 1961).

Definition 5.7.1. *Let P and Q be two probability distributions over the same universe Ω. For $\alpha \in (1, \infty)$ the Rényi divergence of order α between Q and P is*

$$\mathrm{D}_\alpha\left(P||Q\right) = \frac{1}{\alpha - 1} \mathop{\mathbb{E}}_{x \sim P}\left[\left(\frac{P(x)}{Q(x)}\right)^{\alpha-1}\right] \tag{5.19}$$

We then redefine zCDP in terms of Rényi Divergence, which is equivalent to the earlier definition.

Definition 5.7.2. *An algorithm $\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}$ is ρ-zCDP, if for all neighboring datasets $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}^n$ and all $\alpha > 1$ we have*

$$\mathrm{D}_\alpha\left(\mathcal{M}(\boldsymbol{x})||\mathcal{M}(\boldsymbol{x}')\right) \leq \rho\alpha.$$

Note that the Rényi divergence can be defined for $\alpha \in [0, 1]$ as well. In fact van Erven and Harremos (2014) define *simple orders* to be when $\alpha \in (0, 1) \cup (1, \infty)$, meaning that the definition can be stated with the formula given in (5.19), and then explicitly gives the Rényi divergence for *extended orders* as

$$\mathrm{D}_1\left(P||Q\right) = \mathrm{D}_{KL}(P||Q), \qquad \mathrm{D}_0\left(P||Q\right) = -\log\left(Q(\{x : P(x) > 0\})\right).$$

We then define the extension of zCDP to allow for $\alpha \geq 0$.

Definition 5.7.3. *An algorithm $\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}$ is ρ-zCDP$^+$, if for all neighboring datasets $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}^n$ and all $\alpha \geq 0$ we have*

$$\mathrm{D}_\alpha\left(\mathcal{M}(\boldsymbol{x})||\mathcal{M}(\boldsymbol{x}')\right) \leq \rho\alpha.$$

From van Erven and Harremos (2014), we know that D_α is continuous for $\alpha \in [0, 1] \cup \{(1, \infty] | \mathrm{D}_\alpha < \infty\}$. Further, $(1 - \alpha)\mathrm{D}_\alpha$ is concave for $\alpha \in [0, \infty]$ which allows for all the nice properties used in Bun and Steinke (2016) to extend to the case when $\alpha \in [0, 1]$.[4] Thus, all the properties of zCDP that we presented in Chapter 2 extends to zCDP$^+$.

We can also define zCDP$^+$ in terms of the privacy loss random variable, so we can replace the bound on Rényi divergence with the following condition for all $\lambda \in \mathbb{R}$,

---

[4] Thanks to Thomas Steinke and Mark Bun for pointing out that the results in their paper work when $\alpha \in [0, 1]$.

$$\mathbb{E}\left[\exp\left(\lambda\left(\mathrm{PrivLoss}\left(\mathcal{M}(\boldsymbol{x})||\mathcal{M}(\boldsymbol{x}')\right) - \rho\right)\right)\right] \leq e^{\lambda^2\rho}. \tag{5.20}$$

Note that ensuring $\mathrm{D}_\alpha\left(\mathcal{M}(\boldsymbol{x})||\mathcal{M}(\boldsymbol{x}')\right) \leq \rho\alpha$ for $\alpha \geq 0$ tells us that (5.20) holds for $\lambda \geq -1$, but we can extend it to $\lambda < -1$ by noting that zCDP$^+$ is symmetric, that is for neighboring $\boldsymbol{x}, \boldsymbol{x}'$,

$$\begin{aligned}\mathbb{E}\left[\exp\left(\lambda\left(\mathrm{PrivLoss}\left(\mathcal{M}(\boldsymbol{x})||\mathcal{M}(\boldsymbol{x}')\right)\right)\right)\right] &= \mathop{\mathbb{E}}_{y\sim\mathcal{M}(\boldsymbol{x})}\left[\left(\frac{\Pr\left[\mathcal{M}(\boldsymbol{x})=y\right]}{\Pr\left[\mathcal{M}(\boldsymbol{x}')=y\right]}\right)^{\lambda}\right] \\ &= \mathop{\mathbb{E}}_{y\sim\mathcal{M}(\boldsymbol{x}')}\left[\left(\frac{\Pr\left[\mathcal{M}(\boldsymbol{x}')=y\right]}{\Pr\left[\mathcal{M}(\boldsymbol{x})=y\right]}\right)^{-\lambda-1}\right] \\ &= \exp\left[(-\lambda-1)\mathrm{D}_{-\lambda}\left(\mathcal{M}(\boldsymbol{x}')||\mathcal{M}(\boldsymbol{x})\right)\right] \\ &\leq e^{\lambda(\lambda+1)\rho}\end{aligned}$$

where the last inequality follows from the fact that $-\lambda > 1$.

Before we define odometers and filters in this setting, we define the game `AdaptParamComp`$^+(\mathcal{A}, k, b)$ to be the same as `AdaptParamComp`$(\mathcal{A}, k, b)$ in Algorithm 3, except at each round $\mathcal{A}$ gives parameters ρ_i which depend on the previous outcomes in the interaction and $\mathcal{M}_i$ is ρ_i-zCDP$^+$. We can now define zCDP$^+$ odometers.

Definition 5.7.4. *A pair of functions* $(\mathbb{E}\_\mathtt{COMP}, \mathbb{V}\_\mathtt{COMP})$ *is a* valid zCDP$^+$ odometer *if for all adversaries* $\mathcal{A}$ *in* `AdaptParamComp`$^+(\mathcal{A}, k, b)$*, we have for all* $\lambda \in \mathbb{R}$

$$\mathop{\mathbb{E}}_{\boldsymbol{v}\sim V^0}\left[\exp\left(\lambda\left(\mathrm{PrivLoss}(\boldsymbol{v}) - \mathbb{E}\_\mathtt{COMP}\left(\rho_1, \cdots, \rho_k\right)\right) - \frac{\lambda^2}{2}\cdot\mathbb{V}\_\mathtt{COMP}(\rho_1, \cdots, \rho_k)\right)\right] \leq 1$$

Note that we are expecting the privacy loss random variable to be subgaussian with mean $\mathbb{E}\_\mathtt{COMP}(\rho_1, \rho_k)$ and variance $\mathbb{V}\_\mathtt{COMP}(\rho_1, \cdots, \rho_k)$. We presented the definition of zCDP$^+$ in order for us to link together privacy odometers presented in the earlier sections with composition theorems for zCDP when the parameters can be chosen adaptively. The condition in the zCDP$^+$ definition should be compared with the inequality in (5.8) that holds for all $\lambda \in \mathbb{R}$, which we used to prove all of the privacy odometer and filter bounds in the previous sections. Thus, once we have a zCDP$^+$ odometer, we can easily convert it into a valid privacy odometer via the concentration bounds given in Section 5.4. In fact, this is the primary reason to introduce zCDP$^+$ rather than use the original zCDP definition.

Theorem 5.7.5. *If* $(\mathbb{E}\_\mathtt{COMP}, \mathbb{V}\_\mathtt{COMP})$ *is a valid zCDP*$^+$ *odometer, then* $\widetilde{\mathtt{COMP}}_{\delta_g}$ *is a valid privacy odometer when* $\{\delta_i\} \equiv 0$ *for every* $\delta_g > 0$ *and when* $\mathtt{COMP}\left(\epsilon_1^2/2, \cdots, \epsilon_k^2/2\right) \in [1/n^2, 1]$

$$\widetilde{\mathtt{COMP}}_{\delta_g}\left(\epsilon_1, 0, \cdots, \epsilon_k, 0\right) = \mathbb{E}\_\mathtt{COMP}\left(\epsilon_1^2/2, \cdots, \epsilon_k^2/2\right) + \sqrt{2\cdot\mathtt{COMP}\left(\epsilon_1^2/2, \cdots, \epsilon_k^2/2\right)\left(\log(48e) + 2\log\left(\frac{\log(n)}{\delta_g}\right)\right)}.$$

Proof. As we did in Theorem 5.6.5, we use the definition of a valid zCDP$^+$ odometer and apply Theorem 5.4.5 $\square$

We now prove that a valid zCDP$^+$ odometer is to just sum up all the ρ_i for $i \in [k]$ that the analyst has selected for both the mean and variance functions.

Theorem 5.7.6. ($\mathbb{E}\_\texttt{COMP}, \mathbb{V}\_\texttt{COMP}$) *is a valid zCDP$^+$ odometer where*

$$\mathbb{E}\_\texttt{COMP}(\rho_1, \cdots, \rho_k) = \sum_{i=1}^{k} \rho_i \qquad \& \qquad \mathbb{V}\_\texttt{COMP}(\rho_1, \cdots, \rho_k) = 2\sum_{i=1}^{k} \rho_i$$

Proof. We will use the notation $V_i^0 = \mathcal{M}_i^0$ for $i \in [k]$, $V_{i:j}^0 = (\mathcal{M}_i^0, \cdots, \mathcal{M}_j^0)$, and $V_{<j}^0 = (V_1^0, \cdots, V_{j-1}^0)$. Recall that we can decompose the privacy loss as a summation of individual losses $\mathrm{PrivLoss}(\boldsymbol{v}) = \sum_i \mathrm{PrivLoss}_i(\boldsymbol{v}_{\leq i})$. At round 1, the analyst $\mathcal{A}$ selects an algorithm which is ρ_1-zCDP$^+$, thus we have for all $\lambda \in \mathbb{R}$

$$\mathop{\mathbb{E}}_{v_1 \sim V_1^0} \left[\exp\left[\lambda\left(\mathrm{PrivLoss}_1(v_1) - \rho_1\right)\right]\right] \leq \lambda^2 \rho_1.$$

Then, conditioning on round $1, \cdots, j-1$, $\mathcal{A}$ selects parameter ρ_j and an algorithm $\mathcal{M}_j$ that is ρ_j-zCDP$^+$, which again tells us that for all $\lambda \in \mathbb{R}$,

$$\mathop{\mathbb{E}}_{v_j \sim V_j^0} \left[\exp\left[\lambda\left(\mathrm{PrivLoss}_j(\boldsymbol{v}_{\leq j}) - \rho_j\right)\right] | V_{<j}^0 = \boldsymbol{v}_{<j}\right] \leq \lambda^2 \rho_j.$$

Putting this all together, and noting that $\mathrm{PrivLoss}(\boldsymbol{v}) = \sum_{i=1}^{k} \mathrm{PrivLoss}_i(\boldsymbol{v}_{\leq i})$, we have for all $\lambda \in \mathbb{R}$

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\boldsymbol{v} \sim V^0} \left[\exp\left[\lambda\left(\mathrm{PrivLoss}(\boldsymbol{v}) - \sum_{i=1}^{k}\rho_i\right) - \lambda^2 \sum_{i=1}^{k}\rho_i\right]\right] \\
&\quad = \mathop{\mathbb{E}}_{\boldsymbol{v}_{2:k} \sim V_{2:k}^0} \left[\exp\left[\lambda\left(\sum_{i=2}^{k}\mathrm{PrivLoss}_i(\boldsymbol{v}_{\leq i}) - \sum_{i=2}^{k}\rho_i\right) - \lambda^2 \sum_{i=2}^{k}\rho_i\right] | V_1^0 = v_1\right] \\
&\qquad \cdot \mathop{\mathbb{E}}_{v_1 \sim V_1^0}\left[\exp\left[\lambda\left(\mathrm{PrivLoss}_1(v_1) - \rho_1\right) - \lambda^2\rho_1\right]\right] \\
&\quad \leq \mathop{\mathbb{E}}_{\boldsymbol{v}_{2:k} \sim V_{2:k}^0} \left[\exp\left[\lambda\left(\sum_{i=2}^{k}\mathrm{PrivLoss}_i(\boldsymbol{v}_{\leq i}) - \sum_{i=2}^{k}\rho_i\right) - \lambda^2 \sum_{i=2}^{k}\rho_i\right] | V_1^0 = v_1\right] \\
&\quad \vdots \\
&\quad \leq \mathop{\mathbb{E}}_{v_k \sim V_k^0}\left[\exp\left[\lambda\left(\mathrm{PrivLoss}_k(\boldsymbol{v}_{\leq k}) - \rho_k\right) - \lambda^2\rho_k\right] | V_{<k}^0 = \boldsymbol{v}_{<k}\right] \\
&\quad \leq 1
\end{aligned}$$

$\square$

This shows that we can use the same composition theorems from zCDP$^+$ despite adaptively selecting the parameters.

5.8. Conclusion and Future Work

We provided a framework for how to handle composition for differentially private algorithms when the privacy parameters can be chosen adaptively. This is a natural way to think of composition, because the choice of DP algorithm is highly correlated with what privacy parameter the analyst would like to set, due to different DP algorithms having vastly different utility guarantees. We showed that we can achieve a sublinear composition bound on the privacy loss random variable, but we cannot use the existing composition theorems for the realized set of privacy parameters that were chosen adaptively. We then extended the framework to include zCDP (actually zCDP$^+$).

There are many things left to be understood with composition when the parameters do not need to be chosen upfront. One potential direction for future work is to obtain the optimal privacy odometers/filters as was done for differential privacy composition (Kairouz et al., 2015; Murtagh and Vadhan, 2016). Another direction is to link together privacy odometers with being able to obtain valid p-value corrections as the analyst interacts with the data. We were able to obtain valid p-value corrections before by linking differential privacy with max-information. Thus, we would like to be able to connect privacy odometers with max-information, however the analysis we presented in Chapter 4 totally breaks down when the parameters are random variables themselves – we cannot rely on (pointwise) indistinguishability.

Part III

PRIVATE HYPOTHESIS TESTS

Thus far, we have presented the theoretical connection between privacy and adaptive data analysis. However, these results are only useful insofar as there are differentially private analyses that an analyst would actually want to implement. We then turn to modifying traditional hypothesis tests by including the constraint that they be private. Note that these tests are useful beyond adaptive data analysis, where we might want to perform inference on sensitive data. In fact, some of this work is part of the broader effort of the "Privacy Tools for Sharing Research Data"[5] project that aims at developing differentially private tools that can be used for studies, specifically in the social sciences. Social scientists often deal with various sensitive data that contains individual's private information, e.g. voting behavior (Greenwald et al., 1987), attitude toward abortion (Ebaugh and Haney, 1978) and medical records (David and Beards, 1985). The framework of hypothesis testing is frequently used by social scientists to confirm or reject their belief to how a population is modeled, e.g. goodness of fit tests have been used by David and Beards (1985); Gill et al. (1987); Blair et al. (1979); Glaser (1959) and independence tests have been used by Kuklinski and West (1981); Ebaugh and Haney (1978); Berry (1961); Krain and Myers (1997); Greenwald et al. (1987); Mitchell and McCormick (1988).

Homer et al. (2008) published a proof-of-concept attack showing that participation of individuals in scientific studies can be inferred from aggregate data typically published in genome-wide association studies (GWAS). Since then, there has been renewed interest in protecting confidentiality of participants in scientific data (Johnson and Shmatikov, 2013; Uhler et al., 2013; Yu et al., 2014; Simmons et al., 2016) using privacy definitions such as differential privacy and its variations (Dwork et al., 2006b,a; Bun and Steinke, 2016; Dwork and Rothblum, 2016).

As we discussed in Chapter 1, an important tool in statistical inference is *hypothesis testing*, a general framework for determining whether a given model – called the null hypothesis H_0 – of a population should be rejected based on a sample from the population. One of the main benefits of hypothesis testing is that it gives a way to control the probability of false discovery or Type I error – falsely concluding that a model should be rejected when it is indeed true. Type II error is the probability of failing to reject H_0 when it is false. Typically, scientists want a test that guarantees a pre-specified Type I error (say 0.05) and has high *power* – complement of Type II error (recall Table 1).

The standard approach to hypothesis testing can be outlined as follows:

(1) estimate the model parameters,

(2) compute a *test statistic* T (a function of the data and the model parameters),

(3) determine the (asymptotic) distribution of T under the assumption that the model generated the data,

(4) compute the *p-value* (Type I error) as the probability of T being more extreme than the realized value computed from the data.[6]

---

[5] `http://privacytools.seas.harvard.edu`

[6] For one-sided tests, the p-value is the probability of seeing the computed statistic or anything larger

The differentially private tests presented here achieve a target level $1-\alpha$ *significance*, i.e. they reject with probability at most α when the null hypothesis holds (in some cases, we provide a rigorous proof of this fact and in others, it is experimentally verified). This guarantees limited Type I errors. However, all of our tests do lose *power*; that is when the null hypothesis is false, they correctly reject with lower probability than the classical hypothesis tests. This corresponds to an increase in Type II errors. We empirically show that we can recover a level of power similar to the one achieved by the classical versions by adding more samples.

Additional Related Work One of the first works to study the asymptotic distributions of statistics that use differential privacy came from Wasserman and Zhou (2010). Smith (2011) then showed that for a large family of statistics, there is a corresponding differentially private statistic that shares the same asymptotic distribution as the original statistic. However, these results do not ensure that statistically valid conclusions are made for finite samples. It is then the goal of a recent line of work to develop statistical inference tools that give valid conclusions for even reasonably sized datasets.

The previous work on private statistical inference for categorical data can be roughly combined into two main groups (with most primarily dealing with GWAS specific applications). The first group adds appropriately scaled noise to the sampled data (or histogram of data) to ensure differential privacy and uses existing classical hypothesis tests, disregarding the additional noise distribution (Johnson and Shmatikov, 2013). This is warranted in that the impact of the noise becomes small as the sample size grows large. In fact, Vu and Slavković (2009) studies how many more samples would be needed before the test with additional noise recovers the same level of power as the original test on the actual data. However, as we will show and was pointed out in Fienberg et al. (2010); Karwa and Slavković (2012); Karwa and Slavković (2016), this can lead to misleading and statistically invalid results, specifically with much higher Type I error than the prescribed amount.

The second group of work consists of tests that focus on adjusting step (3) in the standard approach to hypothesis testing given in above. That is, these tests use the same statistic in the classical hypothesis tests (without noise) and after making the statistic differentially private, determine the resulting modified asymptotic distribution of the private statistic (Uhler et al., 2013; Yu et al., 2014; Wang et al., 2015). Unfortunately, the resulting asymptotic distribution cannot be written analytically, and so Monte Carlo (MC) simulations or numerical approximations are commonly used to determine at what point to reject the null hypothesis. We will develop tests in Chapter 6 that follows this line of work.

In Chapter 7, we focus on a different technique from these two different approaches, namely modifying step (2) in our outline of hypothesis testing. Thus, we consider transforming the test statistic itself so that the resulting distribution is close to the original asymptotic distribution. The idea of modifying the test statistic for *regression coefficients* to obtain a t-statistic in ordinary least squares has also been considered by Sheffet (2015a).

Due to nice composition properties of Gaussian random variables, our results will mainly

---

under H_0.

focus on adding Gaussian noise, thus ensuring our hypothesis tests are zCDP (Bun and Steinke, 2016) or approximate DP. However, we will also provide a Monte Carlo (MC) approach to obtaining private hypothesis tests with arbitrary noise distributions beyond Gaussian (e.g. Laplace noise for pure differential privacy).

Independent of our work, Wang et al. (2015) also look at hypothesis testing with categorical data subject to differential privacy. They mainly consider adding Laplace noise to the data but point out that their method also generalizes to arbitrary noise distributions. However, in order to compute critical values, they resort to Monte Carlo methods to sample from the asymptotic distribution. Our Monte Carlo approach samples from the *exact* distribution from the underlying null hypothesis, which, unlike sampling from the asymptotic distribution, guarantees significance at least $1 - \alpha$ in goodness of fit tests at finite sample sizes. We only focus on Gaussian noise in our asymptotic analysis due to there being existing methods for finding tail probabilities (and hence critical values) for the resulting distributions, but our approaches can be generalized for arbitrary noise distributions. Further, we also consider the power of each of our differentially private tests.

CHAPTER 6

PRIVATE CHI-SQUARE TESTS: GOODNESS OF FIT AND INDEPENDENCE TESTING

This chapter follows from the work of Gaboardi et al. (2016). We focus here on two classical tests for data drawn from a multinomial distribution: *goodness of fit test*, which determines whether the data was in fact drawn from a multinomial distribution with probability vector $\boldsymbol{p}^0$; and *independence test*, which tests whether two categorical random variables are independent of each other. Both tests depend on the *chi-square* statistic, which is used to determine whether the data is likely or not under the given model.

For this work we will be considering categorical data. That is, we assume the domain $\mathcal{X}$ has been partitioned into d buckets or outcomes and the function $f : \mathcal{X}^n \to \mathbb{R}^d$ returns a histogram counting how many records are in each bucket. Our test statistics will only depend on this histogram. Since neighboring datasets $\boldsymbol{x}, \boldsymbol{x}'$ of size n differ on only one entry, their corresponding histograms differ by ±1 in exactly two buckets. Hence, we will say that two histograms are neighboring if they differ in at most two entries by at most 1. In this case, we can write the sensitivity of the histogram function as $\Delta_1(f) = 2$ and $\Delta_2(f) = \sqrt{2}$. To preserve privacy, we will add noise to the corresponding histogram $\boldsymbol{H} = (H_1, \cdots, H_d)$ of our original dataset to get $\widetilde{\boldsymbol{H}} = (\widetilde{H}_1, \ldots, \widetilde{H}_d)$. We perform hypothesis testing on this noisy histogram $\widetilde{\boldsymbol{H}}$. By Theorem 2.2.5, we know that each of our hypothesis tests will be ρ-zCDP as long as we add Gaussian noise with variance $1/\rho$ to each count in $\boldsymbol{H}$ (see Theorem 2.2.2). Similarly, we could add Laplace noise with scale parameter $2/\epsilon$ to ensure our hypothesis tests will be ϵ-DP (see Theorem 2.1.2).

Note that for DP we can use either Laplace or Gaussian noise, but the variance of the noise we add is typically smaller when we use Laplace noise. For a comparison, to preserve (ϵ, δ)-DP, the Laplace mechanism adds noise with variance $8/\epsilon^2$ whereas the Gaussian mechanism adds noise with variance nearly $\log(1/\delta)/\epsilon^2$ when $\epsilon < 1$. For any nontrivial privacy guarantee, we will have $\delta < e^{-8}$ (typically, one might set $\delta = 10^{-6}$), which forces the Gaussian mechanism to add more noise. The noise affects the probability of Type II error of our tests because it makes it harder to reject a sample if the data with noise has large variance. Thus, Laplace noise is better to use for the privacy benchmark of DP.

Alternatively, for zCDP we can again use either Laplace or Gaussian noise. To preserve ρ-zCDP, the Laplace mechanism incorporates noise with variance $4/\rho$, whereas the Gaussian mechanism adds noise with variance $1/\rho$. Thus, the test with better power (lower probability of Type II error) will typically use Gaussian noise over Laplace. Thus, depending on the privacy benchmark, it may be beneficial to use Laplace noise over Gaussian noise or vice versa.

We consider $\boldsymbol{H} = (H_1, \cdots, H_d)^\intercal \sim \text{Multinomial}(n, \boldsymbol{p})$ where $\boldsymbol{p} = (p_1, \cdots, p_d)^\intercal$ and $\sum_{i=1}^d p_i = 1$. Note that the multinomial distribution is the generalization of a binomial distribution where there are d outcomes as opposed to two – success or failure. For a *goodness of fit test*, we want to test the null hypothesis $\mathrm{H}_0 : \boldsymbol{p} = \boldsymbol{p}^0$. A common way to test this is based on the *chi-square* statistic $\mathrm{T}^{(n)}$ where

$$\mathrm{T}^{(n)} = \sum_{i=1}^d \frac{\left(H_i - np_i^0\right)^2}{np_i^0} \tag{6.1}$$

We present the classical chi-square *goodness of fit test* in Algorithm 7, which compares the chi-square statistic $\mathrm{T}^{(n)}$ to a threshold $\chi^2_{d-1,1-\alpha}$ that depends on a desired level of significance $1-\alpha$ as well as the dimension of the data. The threshold $\chi^2_{d-1,1-\alpha}$ satisfies the following relationship:

$$\Pr\left[\chi^2_{d-1} \geq \chi^2_{d-1,1-\alpha}\right] = \alpha.$$

where χ^2_{d-1} is a chi-square random variable with $d-1$ degrees of freedom, which is the distribution of the random variable $\mathbf{N}^\intercal\mathbf{N}$ where $\mathbf{N} \sim \mathrm{N}\left(\mathbf{0}, I_{d-1}\right)$.

Algorithm 7 Classical Goodness of Fit Test for Multinomial Data: `GOF`

```
Input: $\boldsymbol{h}, \alpha, \mathrm{H}_0 : \boldsymbol{p} = \boldsymbol{p}^0$
  Compute $\mathrm{T}^{(n)}$.
  if $\mathrm{T}^{(n)} > \chi^2_{d-1,1-\alpha}$ then
    Decision ← Reject
  else
    Decision ← Fail to Reject
Output: Decision.
```

The reason why we compare $\mathrm{T}^{(n)}$ with the chi-square distribution is because of the following classical result.

Theorem 6.1.1 [Bishop et al. (1975)]**.** *Assuming* $\mathrm{H}_0 : \boldsymbol{p} = \boldsymbol{p}^0$ *holds, the statistic* $\mathrm{T}^{(n)}$ *converges in distribution to a chi-square with* $d-1$ *degrees of freedom, i.e.*

$$\mathrm{T}^{(n)} \xrightarrow{D} \chi^2_{d-1}.$$

Note that this does not guarantee that $\Pr\left[\mathrm{T}^{(n)} > \chi^2_{d-1,1-\alpha}\right] \leq \alpha$ for finite samples, nevertheless the test works well and is widely used in practice.

It will be useful for our purposes to understand why the asymptotic result holds in Theorem 6.1.1. We present the following classical analysis (Bishop et al., 1975) of Theorem 6.1.1 so that we can understand what adjustments need to be made to find an approximate distribution for a differentially private statistic. Consider the random vector $\boldsymbol{U} = (U_1, \cdots, U_d)^\intercal$

where

$$U_i = \frac{H_i - np_i^0}{\sqrt{np_i^0}} \qquad \forall i \in [d]. \tag{6.2}$$

We write the covariance matrix for $\boldsymbol{U}$ as Σ where

$$\Sigma \overset{\text{defn}}{=} I_d - \sqrt{\boldsymbol{p}^0}\sqrt{\boldsymbol{p}^0}^\intercal \tag{6.3}$$

and $\sqrt{\boldsymbol{p}^0} = (\sqrt{p_1^0}, \cdots, \sqrt{p_d^0})^\intercal$. Note the the covariance matrix Σ is singular and positive semi-definite. By the *central limit theorem* we know that $\boldsymbol{U}$ converges in distribution to a multivariate normal

$$\boldsymbol{U} \overset{D}{\to} \mathrm{N}\left(\boldsymbol{0}, \Sigma\right) \qquad \text{as } n \to \infty.$$

Thus, when we make the assumption that $\boldsymbol{U}$ is multivariate normal, then the significance of `GOF` given in Algorithm 7 is exactly $1 - \alpha$.

We show in the following lemma that if a random vector $\boldsymbol{U}$ is exactly distributed as multivariate normal then we get that $\mathrm{T}^{(n)} = \boldsymbol{U}^\intercal\boldsymbol{U} \sim \chi^2_{d-1}$.

Lemma 6.1.2 [Bishop et al. (1975)]**.** *If* $\boldsymbol{U} \sim \mathrm{N}\left(\boldsymbol{0}, \Sigma\right)$ *for* Σ *given in* (6.3) *then* $\boldsymbol{U}^\intercal\boldsymbol{U} \sim \chi^2_{d-1}$.

Proof. The eigenvalues of Σ must be either 0 or 1 because Σ is idempotent. Thus, the number of eigenvalues that are 1 equals the trace of Σ, which is $d-1$. We then know that there exists a matrix $D \in R^{d \times d-1}$ where $\Sigma = DD^T$ and $D^TD = I_{d-1}$. Define the random variable $\boldsymbol{Y} \sim \mathrm{N}\left(\boldsymbol{0}, I_{d-1}\right)$. Note that $D\boldsymbol{Y}$ is equal in distribution to $\boldsymbol{U}$. We then have

$$\boldsymbol{U}^\intercal\boldsymbol{U} \sim \boldsymbol{Y}^\intercal D^\intercal D\boldsymbol{Y} \sim \boldsymbol{Y}^\intercal\boldsymbol{Y} \sim \chi^2_{d-1}$$

□

6.1.1. Private Chi-Square Statistic

To ensure privacy (either DP or zCDP), we add independent noise to each component of $\boldsymbol{H}$. For the time being, we will consider arbitrary noise distribution $\mathcal{Z}$ which are mean zero and have variance σ. We then form the *private chi-square* statistic $\mathrm{T}^{(n)}\left(\mathcal{Z}\right)$ based on the noisy counts,

$$\mathrm{T}^{(n)}\left(\mathcal{Z}\right) = \sum_{i=1}^{d} \frac{\left(H_i + Z_i - np_i^0\right)^2}{np_i^0}, \qquad \{Z_i\} \overset{i.i.d.}{\sim} \mathcal{Z} \tag{6.4}$$

Some examples of $\mathcal{Z}$ would be $\mathcal{Z} = \mathrm{N}\left(0, 1/\rho\right)$ to ensure ρ-zCDP, or $\mathcal{Z} = \mathrm{Lap}(2/\epsilon)$ to ensure ϵ-DP. Recall that in the original goodness of fit test without privacy in Algorithm 7 we compare the distribution of $\mathrm{T}^{(n)}$ with that of a chi-squared random variable with $d-1$ degrees of freedom. The following result shows that adding noise to each cell count does not affect this asymptotic distribution.

Lemma 6.1.3. *Fixing* $\boldsymbol{p}^0 > \boldsymbol{0}$,[1] *and having noise distribution* $\mathcal{Z}$ *with mean zero and variance* σ_n *where* $\sigma_n/n \to 0$*, then the private chi-squared statistic* $\mathrm{T}^{(n)}\left(\mathcal{Z}\right)$ *given in* (6.4)

---

[1] We use the notation $\boldsymbol{p} > \boldsymbol{0}$ to denote that each coordinate of $\boldsymbol{p}$ is positive.

converges in distribution to χ^2_{d-1} *as* $n \to \infty$.

Proof. We first expand (6.4) to get

$$\mathrm{T}^{(n)}\left(\mathcal{Z}\right) = \sum_{i=1}^{d} \left(\frac{H_i - np_i^0}{\sqrt{np_i^0}} \right)^2 + 2 \sum_{i=1}^{d} \left(\frac{Z_i}{\sqrt{np_i^0}} \right) \left(\frac{H_i - np_i^0}{\sqrt{np_i^0}} \right) + \sum_{i=1}^{d} \left(\frac{Z_i}{\sqrt{np_i^0}} \right)^2$$

We define the two random vectors $\boldsymbol{Z}^{(n)} = \left(\frac{Z_i}{\sqrt{np_i^0}} \right)_{i=1}^{n}$ and $\boldsymbol{H}^{(n)} = \left(\frac{H_i - np_i^0}{\sqrt{np_i^0}} \right)_{i=1}^{n}$. We have that $\mathbb{V}\left[Z_i^{(n)}\right] = \frac{\sigma_n^2}{np_i^0}$ which goes to zero by hypothesis. Additionally $\mathbb{E}\left[\boldsymbol{Z}^{(n)}\right] = \boldsymbol{0}$, so we know that $\boldsymbol{Z}^{(n)} \xrightarrow{P} \boldsymbol{0}$ (meaning convergence in probability as $n \to \infty$). We also know that $\boldsymbol{H}^{(n)} \xrightarrow{D} \mathrm{N}\left(\boldsymbol{0}, \Sigma\right)$, so that $\boldsymbol{Z}^{(n)} \cdot \boldsymbol{H}^{(n)} \xrightarrow{D} 0$ by Slutsky's Theorem[2] and thus $\boldsymbol{Z}^{(n)} \cdot \boldsymbol{H}^{(n)} \xrightarrow{P} 0$ (because 0 is constant). Another application of Slutsky's Theorem tells us that $\mathrm{T}^{(n)}\left(\mathcal{Z}\right) \xrightarrow{D} \chi^2_{d-1}$, since $\mathrm{T}^{(n)}\left(\mathcal{Z}\right) - \mathrm{T}^{(n)} \xrightarrow{P} 0$ and $\mathrm{T}^{(n)} \xrightarrow{D} \chi^2_{d-1}$ from Theorem 6.1.1. $\square$

It then seems natural to use `GOF` on the private chi-squared statistic as if we had the actual chi-squared statistic that did not introduce noise to each count since both private and nonprivate statistics have the same asymptotic distribution.

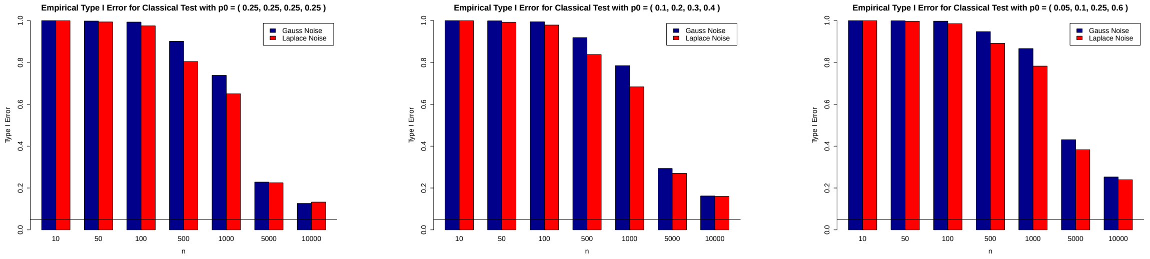


Figure 5: Empirical Type I Error in 10,000 trials when using the classical `GOF` test without modification after incorporating noise due to privacy for ($\epsilon = 0.1$)-DP and ($\rho = \epsilon^2/8$)-zCDP.

We show in Figure 5 that if we were to simply compare the private statistic to the critical value $\chi^2_{d-1,1-\alpha}$, we will typically not get a good significance level even for relatively large n which we need in order for it to be practical tool for data analysts. In the figure we equalize the variance of the Laplace noise and the Gaussian noise we add, hence the choice of $\rho = \epsilon^2/8$. In the following lemma we show that for every realization of data, the statistic $\mathrm{T}^{(n)}\left(\mathcal{Z}\right)$ is expected to be larger than the actual chi-squared statistic $\mathrm{T}^{(n)}$.

Lemma 6.1.4. *For each realization* $\boldsymbol{H} = \boldsymbol{h}$, *we have* $\mathbb{E}_{\mathcal{Z}}\left[\mathrm{T}^{(n)}\left(\mathcal{Z}\right)|\boldsymbol{h}\right] \geq \mathrm{T}^{(n)}$, *where* $\mathcal{Z}$ *has mean zero.*

Proof. Consider the convex function $f(y) = y^2$. Applying Jensen's inequality, we have $f(y) \leq \mathbb{E}_{Z_i}\left[f(y + Z_i)\right]$ for all $i = 1, \cdots, d$ where Z_i is sampled i.i.d. from $\mathcal{Z}$ which has mean zero. We then have for $\boldsymbol{H} = \boldsymbol{h}$

---

[2] Slutsky's Theorem states that if $H^{(n)} \xrightarrow{D} X$ and $Z^{(n)} \xrightarrow{P} c$ then $H^{(n)} \cdot Z^{(n)} \xrightarrow{D} cX$ and $H^{(n)} + Z^{(n)} \xrightarrow{D} X + c$.

$$\begin{aligned}
\mathrm{T}^{(n)} &= \sum_{i=1}^{d} \frac{f\left(h_i - np_i^0\right)}{np_i^0} = \sum_{i=1}^{d} \frac{f\left(\mathbb{E}\left[h_i - np_i^0 + Z_i\right]\right)}{np_i^0} \\
&\leq \mathop{\mathbb{E}}_{\{Z_i\} \overset{i.i.d.}{\sim} \mathcal{Z}} \left[\sum_{i=1}^{d} \frac{f\left(h_i - np_i^0 + Z_i\right)}{np_i^0}\right] = \mathop{\mathbb{E}}_{\{Z_i\} \overset{i.i.d.}{\sim} \mathcal{Z}} \left[\mathrm{T}^{(n)}\left(\mathcal{Z}\right) | \boldsymbol{h}\right]
\end{aligned}$$

□

This result suggests that the significance threshold for the private version of the chi-squared statistic $\mathrm{T}^{(n)}\left(\mathcal{Z}\right)$ should be higher than the standard one. Otherwise, we would reject H_0 too easily using the classical test, which we show in our experimental results. This motivates the need to develop new tests that account for the distribution of the noise.

6.1.2. Monte Carlo Test: `MCGOF`

Given some null hypothesis $\boldsymbol{p}^0$ and statistic $\mathrm{T}^{(n)}\left(\mathcal{Z}\right)$, we want to determine a threshold τ^α such that $\mathrm{T}^{(n)}\left(\mathcal{Z}\right) > \tau^\alpha$ at most an α fraction of the time when the null hypothesis is true. As a first approach, we determine threshold τ^α using a Monte Carlo (MC) approach by sampling from the distribution of $\mathrm{T}^{(n)}\left(\mathcal{Z}\right)$, where $\boldsymbol{H} \sim \mathrm{Multinomial}(n, \boldsymbol{p}^0)$ and $\{Z_i\} \overset{i.i.d.}{\sim} \mathcal{Z}$ for arbitrary noise distributions with mean zero (e.g. Laplace or Gaussian noise).

Let $M_1, \cdots, M_m$ be m continuous random variables that are i.i.d. from the distribution of $\mathrm{T}^{(n)}\left(\mathcal{Z}\right)$ assuming H_0 holds. Further let M be a fresh sample from the distribution of $\mathrm{T}^{(n)}\left(\mathcal{Z}\right)$ assuming H_0. We will write the density and distribution of $\mathrm{T}^{(n)}\left(\mathcal{Z}\right)$ as $f(\cdot)$ and $F(\cdot)$, respectively. Our test will reject M if it falls above some threshold, i.e. critical value, which we will take to be the t-th order statistic of $\{M_i\}$, also written as $M_{(t)}$, so that with probability at most α, M is above this threshold. This will guarantee significance at least $1-\alpha$. We then find the smallest $t \in [m]$ such that $\alpha \geq \Pr\left[M > M_{(t)}\right]$, or

$$\begin{aligned}
\alpha &\geq \int_{-\infty}^{\infty} f(m) \sum_{j=t}^{m} \binom{m}{j} F(m)^j (1-F(m))^{m-j} dm = \int_0^1 \sum_{j=t}^{m} \binom{m}{j} p^j (1-p)^{m-j} dp \\
&= \sum_{j=t}^{m} \frac{1}{m+1} \implies t \geq (m+1)(1-\alpha).
\end{aligned}$$

We then set our threshold based on the $\lceil (m+1)(1-\alpha) \rceil$ ordered statistic of our m samples. By construction, this will ensure that we achieve the significance level we want. Our test then is to sample m points from the distribution of $\mathrm{T}^{(n)}\left(\mathcal{Z}\right)$ and then take the $\lceil (m+1)(1-\alpha) \rceil$- percentile as our cutoff, i.e. if our statistic falls above this value, then we reject H_0. Note that we require $m \geq 1/\alpha$, otherwise there would not be a $\lceil (m+1)(1-\alpha) \rceil$ ordered statistic in m samples. We give the resulting test in Algorithm 8.

Theorem 6.1.5. *The test `MCGOF`$(\cdot, \mathcal{Z}, \alpha, \boldsymbol{p}^0)$ has significance at least $1-\alpha$, also written as $\Pr\left[\texttt{MCGOF}(\boldsymbol{H}, \mathcal{Z}, \alpha, \boldsymbol{p}^0) = \text{ Reject } | \mathrm{H}_0\right] \leq \alpha$.*

Algorithm 8 MC Private Goodness of Fit: `MCGOF`

```
Input: $\boldsymbol{h}, \mathcal{Z}, \alpha, \mathrm{H}_0 : \boldsymbol{p} = \boldsymbol{p}^0$
  Compute $q = \mathrm{T}^{(n)}(\mathcal{Z})$ (6.4).
  Select $m > 1/\alpha$.
  Sample $q_1, \cdots, q_m$ i.i.d. from the distribution of $\mathrm{T}^{(n)}(\mathcal{Z})$.
  Sort the samples $q_{(1)} \leq \cdots \leq q_{(m)}$.
  Compute threshold $q_{(t)}$ where $t = \lceil (m+1)(1-\alpha) \rceil$.
  if $q > q_{(t)}$ then
    Decision ← Reject
  else
    Decision ← Fail to Reject
Output: Decision
```

In Section 6.4, we present the empirical power results for `MCGOF` (along with all our other tests) when we fix an alternative hypothesis.

6.1.3. Asymptotic Approach: Gaussian Noise

In this section we attempt to determine an analytical approximation to the distribution of $\mathrm{T}^{(n)}(\rho) \stackrel{\text{defn}}{=} \mathrm{T}^{(n)}(\mathrm{N}(0, 1/\rho))$, which will ensure our tests are ρ-zCDP. We focus on Gaussian noise because it is more compatible with the asymptotic analysis of `GOF`, which depends on the central limit theorem, as opposed to say Laplace noise.

Recall the random vector $\boldsymbol{U}$ given in (6.2). We then introduce the Gaussian noise random vector as $\boldsymbol{V} = \rho \cdot (Z_1, \cdots, Z_d)^\intercal \sim \mathrm{N}(\boldsymbol{0}, I_d)$. Let $\boldsymbol{W} \in \mathbb{R}^{2d}$ be the concatenated vector defined as

$$\boldsymbol{W} \stackrel{\text{defn}}{=} \begin{pmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{pmatrix}. \tag{6.5}$$

Note that $\boldsymbol{W} \xrightarrow{D} \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ where the covariance matrix is the $2d$ by $2d$ block matrix

$$\boldsymbol{\Sigma} \stackrel{\text{defn}}{=} \begin{bmatrix} \Sigma & 0 \\ 0 & I_d \end{bmatrix} \tag{6.6}$$

where Σ is given in (6.3). Since Σ is idempotent, so is $\boldsymbol{\Sigma}$. We next define the $2d$ x $2d$ positive semi-definite matrix $\boldsymbol{\Lambda}_\rho$ (composed of four d by d block matrices) as

$$\boldsymbol{\Lambda}_\rho \stackrel{\text{defn}}{=} \begin{bmatrix} I_d & \Lambda_\rho \\ \Lambda_\rho & \Lambda_\rho^2 \end{bmatrix} \qquad \text{where} \qquad \Lambda_\rho = \frac{1}{\sqrt{\rho}} \cdot \mathrm{Diag}\left(\boldsymbol{p}^0\right)^{-1/2} \tag{6.7}$$

We can then rewrite our private chi-square statistic as a quadratic form of the random vectors $\boldsymbol{W}$.

$$\mathrm{T}^{(n)}(\rho) = \boldsymbol{W}^\intercal \boldsymbol{\Lambda}_{n\rho} \boldsymbol{W}. \tag{6.8}$$

Remark 6.1.6. *If we have $n\rho_n \to \rho^\star > 0$ then the asymptotic distribution of $\mathrm{T}^{(n)}(\rho_n)$*

would be a quadratic form of multivariate normals.

Similar to the classical goodness of fit test we consider the limiting case that the random vector $\boldsymbol{U}$ is actually a multivariate normal, which will result in $\boldsymbol{W}$ being multivariate normal as well. We next want to be able to calculate the distribution of the quadratic form of normals $\boldsymbol{W}^\intercal \boldsymbol{\Lambda}_{\rho^\star} \boldsymbol{W}$ for $\rho^\star > 0$. Note that we will write $\{\chi_1^{2,i}\}_{i=1}^r$ as a set of r independent chi-square random variables with one degree of freedom.
Theorem 6.1.7. *Let* $\boldsymbol{W} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ *where* $\boldsymbol{\Sigma}$ *is idempotent and has rank* $r \leq 2d$. *Then the distribution of* $\boldsymbol{W}^\intercal \boldsymbol{\Lambda}_{\rho^\star} \boldsymbol{W}$ *where* $\boldsymbol{\Lambda}_{\rho^\star}$ *is positive semi-definite is*

$$\sum_{i=1}^r \lambda_i \chi_1^{2,i}$$

where $\{\lambda_i\}_{i=1}^r$ *are the eigenvalues of* $B^T \boldsymbol{\Lambda}_{\rho^\star} B$ *where* $B \in \mathbb{R}^{2d \times r}$ *such that* $BB^T = \boldsymbol{\Sigma}$ *and* $B^T B = I_r$.

Proof. Let $\boldsymbol{N}^{(1)} \sim \mathrm{N}(\boldsymbol{0}, I_r)$. Because $\boldsymbol{\Sigma}$ is idempotent, we know that there exists a matrix $B \in \mathbb{R}^{2d \times r}$ as in the statement of the lemma. Then $B\boldsymbol{N}^{(1)}$ has the same distribution as $\boldsymbol{W}$. Also note that because $B^T \boldsymbol{\Lambda}_{\rho^\star} B$ is symmetric, then it is diagonalizable and hence there exists an orthogonal matrix $D \in \mathbb{R}^{r \times r}$ such that

$$D^\intercal (B^\intercal \boldsymbol{\Lambda}_{\rho^\star} B) D = \mathrm{Diag}(\lambda_1, \cdots, \lambda_r) \qquad \text{where} \qquad D^\intercal D = DD^\intercal = I_r$$

Let $\boldsymbol{N}^{(1)} = D\boldsymbol{N}^{(2)}$ where $\boldsymbol{N}^{(2)} \sim \mathrm{N}(\boldsymbol{0}, I_r)$. We then have

$$\begin{aligned} \boldsymbol{W}^\intercal \boldsymbol{\Lambda}_{\rho^\star} \boldsymbol{W} &\sim \left(B\boldsymbol{N}^{(1)}\right)^\intercal \boldsymbol{\Lambda}_{\rho^\star} \left(B\boldsymbol{N}^{(1)}\right) \sim \left(BD\boldsymbol{N}^{(2)}\right)^\intercal \boldsymbol{\Lambda}_{\rho^\star} \left(BD\boldsymbol{N}^{(2)}\right) \\ &\sim \left(\boldsymbol{N}^{(2)}\right)^\intercal \mathrm{Diag}(\lambda_1, \cdots \lambda_r) \boldsymbol{N}^{(2)} \end{aligned}$$

Now we know that $(\boldsymbol{N}^{(2)})^\intercal \mathrm{Diag}(\lambda_1, \cdots, \lambda_r) \boldsymbol{N}^{(2)} \sim \sum_{j=1}^r \lambda_j \chi_1^{2,j}$, which gives us our result. □

Note that in the non-private case, the coefficients $\{\lambda_i\}$ in Theorem 6.1.7 become the eigenvalues for the rank $d-1$ idempotent matrix $\boldsymbol{\Sigma}$, thus resulting in a χ^2_{d-1} distribution. We use the result of Theorem 6.1.7 in order to find a threshold that will achieve the desired significance level $1-\alpha$, as in the classical chi-square goodness of fit test. We then set the threshold τ^α to satisfy the following:

$$\Pr\left[\sum_{i=1}^r \lambda_i \chi_1^{2,i} \geq \tau^\alpha\right] = \alpha \tag{6.9}$$

for $\{\lambda_i\}$ found in Theorem 6.1.7. Note, the threshold τ^α is a function of n, ρ, α and $\boldsymbol{p}^0$, but not the data.

We present our modified goodness of fit test when we are dealing with differentially private counts in Algorithm 9.

Algorithm 9 Private Chi-Squared Goodness of Fit Test: `AsymptGOF`

```
Input: $\boldsymbol{h}, \rho, \alpha, \mathrm{H}_0 : \boldsymbol{p} = \boldsymbol{p}^0$
  Compute $\mathrm{T}^{(n)}(\rho) \overset{\text{defn}}{=} \mathrm{T}^{(n)}(\mathrm{N}(0, 1/\rho))$ from (6.4) and $\tau^\alpha$ that satisfies (6.9).
  if $\mathrm{T}^{(n)}(\rho) > \tau^\alpha$ then
    Decision ← Reject
  else
    Decision ← Fail to Reject
Output: Decision
```

6.1.4. Power Analysis of `AsymptGOF`

To determine the power of our new goodness of fit test `AsymptGOF`, we need to specify an alternate hypothesis $\mathrm{H}_1 : \boldsymbol{p} = \boldsymbol{p}^1$ for $\boldsymbol{p}^1 \neq \boldsymbol{p}^0$. Similar to past works (Mitra, 1958; Meng and Chapman, 1966; Guenther, 1977), we will consider alternatives where

$$\boldsymbol{p}_n^1 \overset{\text{defn}}{=} \boldsymbol{p}^0 + \frac{1}{\sqrt{n}} \cdot \boldsymbol{\Delta} \qquad \text{where} \qquad \sum_{i=1}^d \Delta_i = 0. \tag{6.10}$$

Note that $\mathrm{T}^{(n)}(\rho)$ uses the probability vector given in H_0 but data is generated by $\mathrm{Multinomial}(n, \boldsymbol{p}_n^1)$. In fact, the nonprivate statistic $\mathrm{T}^{(n)}$ when the data is drawn from H_1 no longer converges to a chi-square distribution. Instead, $\mathrm{T}^{(n)}$ converges in distribution to a noncentral chi-square when H_1 holds.[3]

Lemma 6.1.8 [Bishop et al. (1975); Ferguson (1996)]**.** *Under the alternate hypothesis* $\mathrm{H}_1 : \boldsymbol{p} = \boldsymbol{p}_n^1$ *given in* (6.10)*, the chi-square statistic* $\mathrm{T}^{(n)}$ *converges in distribution to a noncentral* $\chi^2_{d-1}(\nu)$ *where* $\nu = \boldsymbol{\Delta}^\intercal \mathrm{Diag}(\boldsymbol{p}^0)^{-1} \boldsymbol{\Delta}$*, i.e. given* $\mathrm{H}_1 : \boldsymbol{p} = \boldsymbol{p}_n^1$ *we have*

$$\mathrm{T}^{(n)} \xrightarrow{D} \chi^2_{d-1}(\nu) \qquad \textit{as } n \to \infty.$$

Another classical result tells us that the vector $\boldsymbol{U}$ from (6.2) converges in distribution to a multivariate normal under the alternate hypothesis.

Lemma 6.1.9 [Mitra (1955, 1958)]**.** *Assume* $\boldsymbol{H} \sim \mathrm{Multinomial}(n, \boldsymbol{p}_n^1)$ *where* $\boldsymbol{p}_n^1$ *is given in* (6.10)*. Then* $\boldsymbol{U} \xrightarrow{D} \mathrm{N}(\boldsymbol{\mu}, \Sigma)$ *where* Σ *is given in* (6.3) *and*

$$\boldsymbol{\mu} = \mathrm{Diag}(\boldsymbol{p}^0)^{-1/2} \boldsymbol{\Delta} \tag{6.11}$$

Corollary 6.1.10. *Under the alternate hypothesis* $\mathrm{H}_1 : \boldsymbol{p} = \boldsymbol{p}_n^1$*, then the random vector* $\boldsymbol{W} \xrightarrow{D} \mathrm{N}(\boldsymbol{\mu}', \boldsymbol{\Sigma})$ *for* $\boldsymbol{W}$ *given in* (6.5) *where* $\boldsymbol{\mu}' = (\boldsymbol{\mu}, \boldsymbol{0})^\intercal$ *and* $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ *given in* (6.11) *and* (6.6)*, respectively.*

---

[3] Note that a noncentral chi-square with noncentral parameter θ and ν degrees of freedom is the distribution of $\boldsymbol{Z}^T\boldsymbol{Z}$ where each $\boldsymbol{Z} \sim \mathrm{N}(\boldsymbol{\mu}, I_\nu)$ and $\theta = \boldsymbol{\mu}^\intercal \boldsymbol{\mu}$.

We then write our private statistic as $\mathrm{T}^{(n)}(\rho) = \boldsymbol{W}^\intercal \boldsymbol{\Lambda}_{n\rho} \boldsymbol{W}$. Similar to the previous section we will write $\{\chi_1^{2,j}(\nu_j)\}_{j=1}^r$ as a set of r independent noncentral chi-squares with noncentral parameter ν_j and one degree of freedom.

Theorem 6.1.11. *Let $\boldsymbol{W} \sim \mathrm{N}(\boldsymbol{\mu}', \boldsymbol{\Sigma})$ where $\boldsymbol{\mu}'$ and $\boldsymbol{\Sigma}$ are given in Corollary 6.1.10. We will write $\boldsymbol{\Sigma} = BB^\intercal$ where $B \in \mathbb{R}^{2d \times (2d-1)}$ has rank $2d-1$ and $B^\intercal B = I_{2d-1}$. We define $\boldsymbol{b}^\intercal = (\boldsymbol{\mu}')^\intercal \boldsymbol{\Lambda}_{\rho^\star} BD$ where D is an orthogonal matrix such that $D^\intercal B^\intercal \boldsymbol{\Lambda}_{\rho^\star} BD = \mathrm{Diag}(\lambda_1, \cdots, \lambda_{2d-1})$ and $\boldsymbol{\Lambda}_{\rho^\star}$ is given in* (6.7)*. Then we have*

$$\boldsymbol{W}^\intercal \boldsymbol{\Lambda}_{\rho^\star} \boldsymbol{W} \sim \sum_{j=1}^r \lambda_j \chi_1^{2,j}(\nu_j) + N\left(\kappa, \sum_{j=r+1}^d 4b_j^2\right), \tag{6.12}$$

where $(\lambda_j)_{j=1}^{2d-1}$ are the eigen-values of $B^\intercal \boldsymbol{\Lambda}_{\rho^\star} B$ such that $\lambda_1 \geq \lambda_2 \geq \lambda_r > 0 = \lambda_{r+1} = \cdots = \lambda_{2d-1}$ and

$$\nu_j = \left(\frac{b_j}{\lambda_j}\right)^2 \qquad \text{for } j \in [r] \quad \& \quad \kappa = \boldsymbol{\Delta}^\intercal \mathrm{Diag}(\boldsymbol{p}^0)^{-1} \boldsymbol{\Delta} - \sum_{j=1}^r \frac{b_j^2}{\lambda_j}.$$

Proof. We follow a similar analysis as Mohsenipour (2012) for finding the distribution of a quadratic form of normals. Consider the random variable $\boldsymbol{N}^{(2)} = BD\boldsymbol{N}^{(1)} + \boldsymbol{\mu}'$ where $\boldsymbol{N}^{(1)} \sim \mathrm{N}(\boldsymbol{0}, I_{2d-1})$. Note that $\boldsymbol{N}^{(2)}$ has the same distribution as $\boldsymbol{W}$. We then have for $t \geq 0$

$$\begin{aligned}
\Pr\left[\boldsymbol{W}^\intercal \boldsymbol{\Lambda}_\rho \boldsymbol{W} \geq t\right] &= \Pr\left[(\boldsymbol{N}^{(1)})^\intercal D^\intercal B^\intercal \boldsymbol{\Lambda}_{\rho^\star} BD\boldsymbol{N}^{(1)} + 2(\boldsymbol{\mu}')^\intercal \boldsymbol{\Lambda}_{\rho^\star} BD\boldsymbol{N}^{(1)} + (\boldsymbol{\mu}')^\intercal \boldsymbol{\Lambda}_{\rho^\star} \boldsymbol{\mu}' \geq t\right] \\
&= \Pr\left[(\boldsymbol{N}^{(1)})^\intercal \mathrm{Diag}(\lambda_1, \cdots, \lambda_{d+1}) \boldsymbol{N}^{(1)} + 2\boldsymbol{b}^\intercal \boldsymbol{N}^{(1)} + (\boldsymbol{\mu}')^\intercal \boldsymbol{\Lambda}_{\rho^\star} \boldsymbol{\mu}' \geq t\right] \\
&= \Pr\left[\sum_{j=1}^r \lambda_j \cdot \left(N_j^{(1)} + b_j/\lambda_j\right)^2 + \sum_{j=r+1}^d 2b_j N_j^{(1)} + \kappa \geq t\right] \\
&= \Pr\left[\sum_{j=1}^d \lambda_j \cdot \chi_1^{2,j}\left(\left(\frac{b_j}{\lambda_j}\right)^2\right) + N\left(0, \sum_{j=r+1}^d 4b_j^2\right) + \kappa \geq t\right]
\end{aligned}$$

□

Remark 6.1.12. *Again, if we have $n\rho_n \to \rho^\star > 0$ then the asymptotic distribution of $\mathrm{T}^{(n)}(\rho_n)$ converges in distribution to the random variable of the form given in* (6.12) *when* H_1 *from* (6.10) *is true.*

Obtaining the asymptotic distribution for $\mathrm{T}^{(n)}(\rho)$ when the alternate hypothesis holds may allow for future results on *effective sample size*, i.e. how large a sample size needs to be in order for `AsymptGOF` to have Type II error at most β against $\mathrm{H}_1 : \boldsymbol{p} = \boldsymbol{p}_n^1$. We see this as an important direction for future work.

6.2. Independence Testing

We now consider the problem of testing whether two random variables $\boldsymbol{Y}^1 \sim \text{Multinomial}(1, \boldsymbol{\pi}^{(1)})$ and $\boldsymbol{Y}^2 \sim \text{Multinomial}(1, \boldsymbol{\pi}^{(2)})$ are independent of each other. Note that $\sum_{i=1}^{r} \boldsymbol{\pi}_i^{(1)} = 1$ and $\sum_{j=1}^{c} \pi_j^{(2)} = 1$, so we can write $\pi_r^{(1)} = 1 - \sum_{i<r} \pi_i^{(1)}$ and $\pi_c^{(2)} = 1 - \sum_{j<c} \pi_j^{(2)}$.

We then form the null hypothesis $\mathrm{H}_0 : \boldsymbol{Y}^{(1)} \perp \boldsymbol{Y}^{(2)}$, i.e. they are independent. One approach to testing H_0 is to sample n joint outcomes of $\boldsymbol{Y}^{(1)}$ and $\boldsymbol{Y}^{(2)}$ and count the number of observed outcomes, $H_{i,j}$ which is the number of times $Y_i^{(1)} = 1$ and $Y_j^{(2)} = 1$ in the n trials, so that we can summarize all joint outcomes as a contingency table $\boldsymbol{H} = (H_{i,j}) \sim \text{Multinomial}(n, \boldsymbol{p} = (p_{i,j} : i \in [r],\ j \in [c]))$, where $p_{i,j}$ is the probability that $Y_i^{(1)} = 1$ and $Y_j^{(2)} = 1$.

In Table 2 we give a $r \times c$ contingency table giving the number of joint outcomes for the variables $\boldsymbol{Y}^{(1)}$ and $\boldsymbol{Y}^{(2)}$ from n independent trials. We will write the full contingency table of counts $\boldsymbol{H} = (H_{i,j})$ as a vector with the ordering convention that we start from the top row and move from left to right across the contingency table.

Table 2: Contingency Table with Marginals.

$\boldsymbol{Y}^{(1)} \backslash \boldsymbol{Y}^{(2)}$	1	2	$\cdots$	c	Marginals
1	$H_{1,1}$	$H_{1,2}$	$\cdots$	$H_{1,c}$	$H_{1,\cdot}$
2	$H_{2,1}$	$H_{2,2}$	$\cdots$	$H_{2,c}$	$H_{2,\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
r	$H_{r,1}$	$H_{r,2}$	$\cdots$	$H_{r,c}$	$H_{r,\cdot}$
Marginals	$H_{\cdot,1}$	$H_{\cdot,2}$	$\cdots$	$H_{\cdot,c}$	n

We want to calculate the chi-square statistic as in (6.1) (where now the summation is over all joint outcomes i and j), but now we do not know the true proportion $\boldsymbol{p} = (p_{i,j})$ which depends on $\boldsymbol{\pi}^{(1)}$ and $\boldsymbol{\pi}^{(2)}$. However, we can use the *maximum likelihood estimator* (MLE) $\widehat{\boldsymbol{p}}$ for the probability vector $\boldsymbol{p}$ subject to H_0 to form the statistic $\widehat{\mathrm{T}}^{(n)}$

$$\widehat{\mathrm{T}}^{(n)} = \sum_{i,j} \frac{\left(H_{i,j} - n\widehat{p}_{i,j}\right)^2}{n\widehat{p}_{i,j}}. \tag{6.13}$$

The intuition is that if the test rejects even when the most likely probability vector that satisfies the null hypothesis was chosen, then the test should reject against all others.

Note that under the null hypothesis we can write $\boldsymbol{p}$ as a function of $\boldsymbol{\pi}^{(1)}$ and $\boldsymbol{\pi}^{(2)}$,

$$\boldsymbol{p} = \left(f_{i,j}(\boldsymbol{\pi}^{(1)}, \boldsymbol{\pi}^{(2)}) : i \in [r],\ j \in [c]\right) \quad \text{where} \quad f_{i,j}(\boldsymbol{\pi}^{(1)}, \boldsymbol{\pi}^{(2)}) = \boldsymbol{\pi}^{(1)} \left(\boldsymbol{\pi}^{(2)}\right)^{\intercal}. \tag{6.14}$$

Further, we can write the MLE $\widehat{\boldsymbol{p}}$ as described below.

Lemma 6.2.1 [Bishop et al. (1975)]**.** *Given $\boldsymbol{H}$, which is n samples of joint outcomes of $\boldsymbol{Y}^{(1)} \sim \text{Multinomial}(1, \boldsymbol{\pi}^{(1)})$ and $\boldsymbol{Y}^{(2)} \sim \text{Multinomial}(1, \boldsymbol{\pi}^{(2)})$, if $\boldsymbol{Y}^{(1)} \perp \boldsymbol{Y}^{(2)}$, then the*

MLE for $\boldsymbol{p} = \boldsymbol{f}(\boldsymbol{\pi}^{(1)}, \boldsymbol{\pi}^{(2)})$ *for* $\boldsymbol{f} = (f_{i,j} : i \in [r], j \in [c])$ *given in* (6.14) *is the following:* $\widehat{\boldsymbol{p}} = \boldsymbol{f}(\widehat{\boldsymbol{\pi}}^{(1)}, \widehat{\boldsymbol{\pi}}^{(1)})$ *where*

$$\widehat{\pi}_i^{(1)} = H_{i,\cdot}/n, \ \widehat{\pi}_j^{(1)} = H_{\cdot,j}/n \qquad \text{for } i \in [r], j \in [c] \tag{6.15}$$

and $H_{i,\cdot} = \sum_{j=1}^{r} H_{i,j}$ *and* $H_{\cdot,j} = \sum_{i=1}^{c} H_{i,j}$.

We then state another classical result that gives the asymptotic distribution of $\widehat{\mathrm{T}}^{(n)}$ given H_0.

Theorem 6.2.2. *(Bishop et al., 1975) Given the assumptions in Lemma 6.2.1, the statistic*

$$\widehat{\mathrm{T}}^{(n)} \xrightarrow{D} \chi^2_{\nu}$$

for $\nu = (r-1)(c-1)$.

Algorithm 10 Pearson Chi-Squared Independence Test: `Indep`

```
Input: $\boldsymbol{h}, \alpha$
  $\widehat{\boldsymbol{p}} \leftarrow$ MLE calculation in (6.15)
  Compute $\widehat{\mathrm{T}}^{(n)}$ from (6.13) and set $\nu = (r-1)(c-1)$.
  if $\widehat{\mathrm{T}}^{(n)} > \chi^2_{\nu,1-\alpha}$ and all entries of $\boldsymbol{h}$ are at least 5 then
    Decision $\leftarrow$ Reject
  else
    Decision $\leftarrow$ Fail to Reject
Output: Decision.
```

The chi-square independence test is then to compare the statistic $\widehat{\mathrm{T}}^{(n)}$, with the value $\chi^2_{(r-1)(c-1),1-\alpha}$ for a $1-\alpha$ significance test. We formally give the Pearson chi-square test in Algorithm 10. An often used "rule of thumb" (Triola, 2014) with this test is that it can only be used if all the cell counts are at least 5, otherwise the test Fails to Reject H_0. We will follow this rule of thumb in our tests.

Similar to our prior analysis for goodness of fit, we aim to understand the asymptotic distribution from Theorem 6.2.2. First, we can define $\widehat{\boldsymbol{U}}$ in terms of the MLE $\widehat{\boldsymbol{p}}$ given in (6.15):

$$\widehat{U}_{i,j} = (H_{i,j} - n\widehat{p}_{i,j})/\sqrt{n\widehat{p}_{i,j}}. \tag{6.16}$$

The following classical result gives the asymptotic distribution of $\widehat{\boldsymbol{U}}$ under H_0, which also proves Theorem 6.2.2.

Lemma 6.2.3. *(Bishop et al., 1975) With the same hypotheses as Lemma 6.2.1, the random vector* $\widehat{\boldsymbol{U}}$ *given in* (6.16) *converges in distribution to a multivariate normal,*

$$\widehat{\boldsymbol{U}} \xrightarrow{D} N\left(0, \Sigma_{ind}\right)$$

where $\Sigma_{ind} \overset{\text{defn}}{=} I_{rc} - \sqrt{\boldsymbol{p}} \cdot \sqrt{\boldsymbol{p}}^{\intercal} - \Gamma(\Gamma^{\intercal}\Gamma)^{-1}\Gamma^{\intercal}$ *with* $\boldsymbol{f}$ *given in* (6.14)*, and*

$$\Gamma \overset{\text{defn}}{=} \mathrm{Diag}(\boldsymbol{p})^{-1/2} \cdot \nabla \boldsymbol{f}(\boldsymbol{\pi}^{(1)}, \boldsymbol{\pi}^{(2)}),$$

$$\nabla \boldsymbol{f}(\boldsymbol{\pi}^{(1)}, \boldsymbol{\pi}^{(2)}) = \begin{bmatrix} \frac{\partial f_{1,1}}{\partial \pi_1^{(1)}} & \cdots & \frac{\partial f_{1,1}}{\partial \pi_{r-1}^{(1)}} & \frac{\partial f_{1,1}}{\partial \pi_1^{(2)}} & \cdots & \frac{\partial f_{1,1}}{\partial \pi_{c-1}^{(2)}} \\ \frac{\partial f_{1,2}}{\partial \pi_1^{(1)}} & \cdots & \frac{\partial f_{1,2}}{\partial \pi_{r-1}^{(1)}} & \frac{\partial f_{1,2}}{\partial \pi_1^{(2)}} & \cdots & \frac{\partial f_{1,2}}{\partial \pi_{c-1}^{(2)}} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_{r,c}}{\partial \pi_1^{(1)}} & \cdots & \frac{\partial f_{r,c}}{\partial \pi_{r-1}^{(1)}} & \frac{\partial f_{r,c}}{\partial \pi_1^{(2)}} & \cdots & \frac{\partial f_{r,c}}{\partial \pi_{c-1}^{(2)}} \end{bmatrix}_{rc,r+c-2}.$$

In order to do a test that is similar to `Indep` given in Algorithm 10, we need to determine an estimate for $\boldsymbol{\pi}^{(1)}$ and $\boldsymbol{\pi}^{(2)}$ where we are only given access to the noisy cell counts.

6.2.1. Estimating Parameters with Private Counts

We now assume that we do not have access to the counts $H_{i,j}$ from Table 2 but instead we have $W_{i,j} = H_{i,j} + Z_{i,j}$ where $Z_{i,j} \sim \mathcal{Z}$ for any type of noise distribution, and we want to perform a test for independence. Here we will consider both Laplace and Gaussian noise distributions. We consider the full likelihood of the noisy $r \times c$ contingency table

$$\begin{aligned} &\Pr\left[\boldsymbol{H} + \boldsymbol{Z} = \widetilde{\boldsymbol{h}} | \mathrm{H}_0, \boldsymbol{\pi}^{(1)}, \boldsymbol{\pi}^{(2)}\right] \\ &\quad = \sum_{\substack{\boldsymbol{h}: \sum_{i,j} h_{i,j} = n \\ h_{i,j} \in \mathbb{N}}} \Pr\left[\boldsymbol{H} = \boldsymbol{h} | \mathrm{H}_0, \boldsymbol{\pi}^{(1)}, \boldsymbol{\pi}^{(2)}\right] \cdot \Pr\left[\boldsymbol{Z} = \widetilde{\boldsymbol{h}} - \boldsymbol{h} | \mathrm{H}_0, \boldsymbol{H} = \boldsymbol{h}\right] \\ &\quad = \sum_{\substack{\boldsymbol{h}: \sum_{i,j} h_{i,j} = n \\ h_{i,j} \in \mathbb{N}}} \underbrace{\Pr\left[\boldsymbol{H} = \boldsymbol{h} | \mathrm{H}_0, \boldsymbol{\pi}^{(1)}, \boldsymbol{\pi}^{(2)}\right]}_{\text{Multinomial}} \prod_{i,j} \underbrace{\Pr\left[Z_{i,j} = \widetilde{h}_{i,j} - H_{i,j} | \mathrm{H}_0, \boldsymbol{H} = \boldsymbol{h}\right]}_{\text{Noise}} \end{aligned}$$

to find the best estimates for $\{\boldsymbol{\pi}^{(i)}\}$ given the noisy counts.

Maximizing this quantity is computationally very expensive for values of $n > 100$ even for 2×2 tables,[4] so we instead follow a two step procedure similar to the work of Karwa and Slavković (2016), where they "denoise" a private degree sequence for a synthetic graph and then use the denoised estimator to approximate the parameters of the β-model of random graphs. We will first find the most likely contingency table given the noisy data $\widetilde{\boldsymbol{h}}$ and then find the most likely probability vectors under the null hypothesis that could have generated that denoised contingency table (this is not equivalent to maximizing the full likelihood, but it seems to work well as our experiments later show). For the latter step, we use (6.15) to get the MLE for $\boldsymbol{\pi}^{(1)}$ and $\boldsymbol{\pi}^{(2)}$ given a vector of counts $\boldsymbol{h}$. For the first step, we need to minimize $||\widetilde{\boldsymbol{h}} - \boldsymbol{h}||$ subject to $\sum_{i,j} h_{i,j} = n$ and $h_{i,j} \geq 0$ where the norm in the objective is either ℓ_1 for Laplace noise or ℓ_2 for Gaussian noise.

Note that for Laplace noise, the above optimization problem does not give a unique solution and it is not clear which minimizing contingency table $\boldsymbol{h}$ to use. One solution to overcome this is to add a *regularizer* to the objective value. We will follow the work of Lee et al.

---

[4] Note that there is a poly(n) time algorithm to solve this, but the coefficients in each term of the sum can be very large numbers, with poly(n) bits, which makes it difficult for numeric solvers.

(2015) to overcome this problem by using an *elastic net* regularizer (Zou and Hastie, 2005):

$$\begin{aligned} \underset{\boldsymbol{h}}{\text{argmin}} \qquad & (1-\gamma)\cdot ||\widetilde{\boldsymbol{h}} - \boldsymbol{h}||_1 + \gamma \cdot ||\widetilde{\boldsymbol{h}} - \boldsymbol{h}||_2^2 \\ s.t. \qquad & \sum_{i,j} h_{i,j} = n, \qquad h_{i,j} \geq 0. \end{aligned} \tag{6.17}$$

where if we use Gaussian noise, we set $\gamma = 1$ and if we use Laplace noise then we pick a small $\gamma > 0$ and then solve the resulting program. Our two step procedure for finding an approximate MLE for $\boldsymbol{\pi}^{(1)}$ and $\boldsymbol{\pi}^{(2)}$ based on our noisy vector of counts $\widetilde{\boldsymbol{h}}$ is given in Algorithm 11, where we take into account the rule of thumb from `Indep` and return NULL if any computed table has counts less than 5.

Algorithm 11 Two Step MLE Calculation: `2MLE`

```
Input: $\widetilde{\boldsymbol{h}} = \boldsymbol{H} + \boldsymbol{Z}$
  if $\mathcal{Z}$ = Gauss then
    set $\gamma = 1$
  if $\mathcal{Z}$ = Lap then
    set $0 < \gamma \ll 1$.
  $\widetilde{\boldsymbol{h}} \leftarrow$ Solution to (6.17).
  if Any cell of $\widetilde{\boldsymbol{h}}$ is less than 5 then
    $\widetilde{\boldsymbol{\pi}}^{(1)}, \widetilde{\boldsymbol{\pi}}^{(2)} \leftarrow$ NULL
  else
    $\widetilde{\boldsymbol{\pi}}^{(1)}, \widetilde{\boldsymbol{\pi}}^{(2)} \leftarrow$ MLE with $\widetilde{\boldsymbol{h}}$ given in (6.15).
Output: $\widetilde{\boldsymbol{\pi}}^{(1)}$ and $\widetilde{\boldsymbol{\pi}}^{(2)}$.
```

We will denote $\widetilde{\boldsymbol{p}}$ to be the probability vector of function $\boldsymbol{f}$ from (6.14) applied to the result of `2MLE`($\boldsymbol{H} + \boldsymbol{Z}$). We now write down the private chi-squared statistic when we use the estimate $\widetilde{\boldsymbol{p}}$ in place of the actual (unknown) probability vector $\boldsymbol{p}$:

$$\widetilde{\mathrm{T}}^{(n)}\left(\mathcal{Z}\right) = \sum_{i,j} \frac{\left(H_{i,j} + Z_{i,j} - n\widetilde{p}_{i,j}\right)^2}{n\widetilde{p}_{i,j}}, \qquad \{Z_{i,j}\} \overset{i.i.d.}{\sim} \mathcal{Z}. \tag{6.18}$$

6.2.2. Monte Carlo Test: `MCIndep`

We first follow a similar procedure as in Section 6.1.2 but using the parameter estimates from `2MLE` instead of the actual (unknown) probabilities. Our procedure `MCIndep` (given in Algorithm 12) works as follows: given a dataset $\boldsymbol{h}$, we will add the appropriately scaled Laplace or Gaussian noise to ensure differential privacy to get the noisy table $\widetilde{\boldsymbol{h}}$. Then we use `2MLE` on the private data to get approximates to the parameters $\boldsymbol{\pi}^{(i)}$, which we denote as $\widetilde{\boldsymbol{\pi}}^{(i)}$ for $i = 1, 2$. Using these probability estimates, we sample $m > 1/\alpha$ many contingency tables and noise terms to get m different values for $\widetilde{\mathrm{T}}^{(n)}\left(\mathcal{Z}\right)$ and choose the $\lceil (m+1)(1-\alpha) \rceil$ ranked statistic as our threshold $\widetilde{\tau}^{\alpha}$. If at any stage `2MLE` returns NULL, then the test Fails to Reject H_0. We formally give our test `MCIndep` in Algorithm 12.

Algorithm 12 MC Independence Testing `MCIndep`

```
Input: $\boldsymbol{h}, \mathcal{Z}, \alpha$
  $\widetilde{\boldsymbol{h}} \leftarrow \boldsymbol{h} + \boldsymbol{Z}$, where $\{Z_{i,j}\} \overset{i.i.d.}{\sim} \mathcal{Z}$.
  $(\widetilde{\boldsymbol{\pi}}^{(1)}, \widetilde{\boldsymbol{\pi}}^{(2)}) \leftarrow \texttt{2MLE}(\widetilde{\boldsymbol{h}})$ and $\widetilde{\boldsymbol{p}} \leftarrow \boldsymbol{f}(\widetilde{\boldsymbol{\pi}}^{(1)}, \widetilde{\boldsymbol{\pi}}^{(2)})$.
  if $(\widetilde{\boldsymbol{\pi}}^{(1)}, \widetilde{\boldsymbol{\pi}}^{(2)}) ==$ NULL then
    Decision $\leftarrow$ Fail to Reject.
  else
    $\widetilde{q} \leftarrow \widetilde{\mathrm{T}}^{(n)}(\mathcal{Z})$ using $\widetilde{\boldsymbol{h}}$ and $\widetilde{\boldsymbol{p}}$.
    Set $m > 1/\alpha$ and $q \leftarrow$ NULL.
    for $t \in [m]$ do
      Generate contingency table $\widetilde{\boldsymbol{h}}$ using $\left(\widetilde{\boldsymbol{\pi}}^{(1)}, \widetilde{\boldsymbol{\pi}}^{(2)}\right)$.
      $\widetilde{\widetilde{\boldsymbol{h}}} \leftarrow \widetilde{\boldsymbol{h}} + \boldsymbol{Z}$, where $\{Z_{i,j}\} \overset{i.i.d.}{\sim} \mathcal{Z}$.
      $\left(\widetilde{\widetilde{\boldsymbol{\pi}}}^{(1)}, \widetilde{\widetilde{\boldsymbol{\pi}}}^{(2)}\right) \leftarrow \texttt{2MLE}\left(\widetilde{\widetilde{\boldsymbol{h}}}\right)$.
      if $\left(\widetilde{\widetilde{\boldsymbol{\pi}}}^{(1)}, \widetilde{\widetilde{\boldsymbol{\pi}}}^{(2)}\right) ==$ NULL then
        Decision $\leftarrow$ Fail to Reject.
      else
        Compute $\widetilde{\mathrm{T}}^{(n)}(\mathcal{Z})$ from (6.18), add it to array $q$.
    $\widetilde{\tau}^{\alpha} \leftarrow$ the $\lceil (m+1)(1-\alpha) \rceil$ ranked statistic in $q$.
    if $\widetilde{q} > \widetilde{\tau}^{\alpha}$ then
      Decision $\leftarrow$ Reject.
    else
      Decision $\leftarrow$ Fail to Reject.
Output: Decision
```

6.2.3. Asymptotic Approach: `AsymptIndep`

We will now focus on the analytical form of our private statistic when Guassian noise is added. We can then write $\widetilde{\mathrm{T}}^{(n)}(\rho) \overset{\text{defn}}{=} \widetilde{\mathrm{T}}^{(n)}(\mathrm{N}(0, 1/\rho))$ in its quadratic form, which is similar to the form of $\mathrm{T}^{(n)}(\rho)$ from (6.8),

$$\widetilde{\mathrm{T}}^{(n)}(\rho) \overset{\text{defn}}{=} \widetilde{\boldsymbol{W}}^{\intercal} \widetilde{\boldsymbol{\Lambda}_{n\rho}} \widetilde{\boldsymbol{W}} \tag{6.19}$$

where $\widetilde{\boldsymbol{W}} = \binom{\widetilde{\boldsymbol{U}}}{\boldsymbol{V}}$ with $\widetilde{\boldsymbol{U}}$ set in (6.16) except with $\widetilde{\boldsymbol{p}}$ used instead of the given $\boldsymbol{p}^0$ in the goodness of fit testing and $\boldsymbol{V}$ set as in (6.5). Further, we denote $\widetilde{\boldsymbol{\Lambda}_\rho}$ as $\boldsymbol{\Lambda}_\rho$ in (6.7) but with estimate $\widetilde{\boldsymbol{p}}$ instead of $\boldsymbol{p}^0$. We will use the $2rc$ by $2rc$ block matrix $\widetilde{\boldsymbol{\Sigma}}_{\text{ind}}$ to estimate the covariance of $\widetilde{\boldsymbol{W}}$, where

$$\widetilde{\boldsymbol{\Sigma}}_{\text{ind}} \overset{\text{defn}}{=} \begin{bmatrix} \widetilde{\Sigma}_{\text{ind}} & 0 \\ 0 & I_{rc} \end{bmatrix} \tag{6.20}$$

and $\widetilde{\Sigma}_{\text{ind}}$ is the matrix Σ_{ind} in Lemma 6.2.3, except we use our estimates $\widetilde{\boldsymbol{\pi}}^{(1)}$, $\widetilde{\boldsymbol{\pi}}^{(2)}$, or $\widetilde{\boldsymbol{p}}$ whenever we need to use the actual (unknown) parameters.

Thus, if we are given a differentially private version of a contingency table where each cell has added independent Gaussian noise with variance $1/\rho$, we calculate $\widetilde{\mathrm{T}}^{(n)}(\rho)$ and compare it to the threshold $\widetilde{\tau}^{\alpha}$ where

$$\Pr\left[\sum_{i=1}^{rc} \widetilde{\lambda}_i \chi_1^{2,i} \geq \widetilde{\tau}^{\alpha}\right] = \alpha \tag{6.21}$$

with $\{\widetilde{\lambda}_i\}$ being the eigenvalues of $\widetilde{B}^T \widetilde{\boldsymbol{\Lambda}_\rho} \widetilde{B}$ with rank $\nu = rc + (r-1)(c-1)$ matrix $\widetilde{B} \in \mathbb{R}^{2rc,\nu}$ where $\widetilde{B}\widetilde{B}^T = \widetilde{\boldsymbol{\Sigma}}_{\text{ind}}$. Our new independence test `AsymptIndep` is given in Algorithm 13, where `2MLE` estimates $\boldsymbol{\pi}^{(i)}$ for $i = 1, 2$ and `AsymptIndep` Fails to Reject if `2MLE` returns NULL.

Algorithm 13 Private Independence Test for $r \times c$ tables: `AsymptIndep`

```
Input: $\boldsymbol{h}, \rho, 1-\alpha$
  $\widetilde{\boldsymbol{h}} \leftarrow \boldsymbol{h} + \boldsymbol{Z}$ where $\boldsymbol{Z} \sim \mathrm{N}(0, 1/\rho \; I_{r \cdot c})$.
  $\left(\widetilde{\boldsymbol{\pi}}^{(1)}, \widetilde{\boldsymbol{\pi}}^{(2)}\right) \leftarrow$ 2MLE($\widetilde{\boldsymbol{h}}$).
  if $\left(\widetilde{\boldsymbol{\pi}}^{(1)}, \widetilde{\boldsymbol{\pi}}^{(2)}\right)$ == NULL then
    Decision ← Fail to Reject
  else
    $\widetilde{\boldsymbol{p}} \leftarrow \boldsymbol{f}\left(\widetilde{\boldsymbol{\pi}}^{(1)}, \widetilde{\boldsymbol{\pi}}^{(2)}\right)$ for $\boldsymbol{f}$ given in (6.14).
    Set $\widetilde{\mathrm{T}}^{(n)}(\rho) = \widetilde{\mathrm{T}}^{(n)}(\mathrm{N}(0, 1/\rho))$ from (6.18) with $\widetilde{\boldsymbol{h}}$ and $\widetilde{\tau}^{\alpha}$ from (6.21).
    if $\widetilde{\mathrm{T}}^{(n)}(\rho) > \widetilde{\tau}^{\alpha}$ then
      Decision ← Reject
    else
      Decision ← Fail to Reject
Output: Decision
```

6.3. Significance Results

We now show how each of our tests perform on simulated data when H_0 holds in goodness of fit and independence testing. We fix our desired significance $1-\alpha = 0.95$. If our privacy benchmark is differential privacy, then we set the privacy parameter $\epsilon \in \{0.02, 0.05, 0.1\}$ and use Laplace noise – thus using our MC approach. If we use zCDP as our benchmark then we set $\rho = \epsilon^2/8 \in \{0.00005, 0.0003125, 0.00125\}$. Note that with these parameters, the variance for both Laplace and Gaussian noise are the same.[5] Although zCDP provides a weaker privacy guarantee than pure differential privacy (see Theorem 2.2.3), we know that even for ($\rho = 0.00125$)-zCDP, we get ($\epsilon \approx 0.24, \delta = 10^{-6}$)-DP from Theorem 2.2.4, which still provides a strong privacy guarantee.

6.3.1. GOF Testing

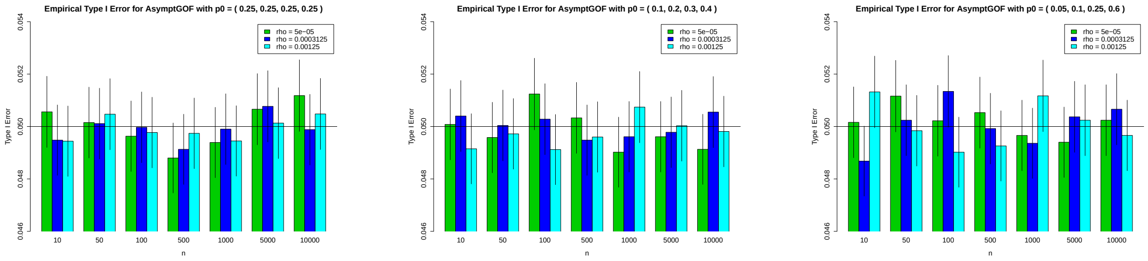


Figure 6: Empirical Type I Error of `AsymptGOF` with error bars corresponding to 1.96 times the standard error in 100,000 trials.

By Theorem 6.1.5, we know that `MCGOF` will have significance at least $1-\alpha$. We then turn to our test `AsymptGOF` to compute the proportion of trials that failed to reject $\mathrm{H}_0 : \boldsymbol{p} = \boldsymbol{p}^0$ when it holds. In Figure 6 we give several different null hypotheses $\boldsymbol{p}^0$ and sample sizes n to show that `AsymptGOF` achieves near 0.95 significance in all our tested cases. We compare this to the results if we did not modify our test due to the additional noise from Figure 5.

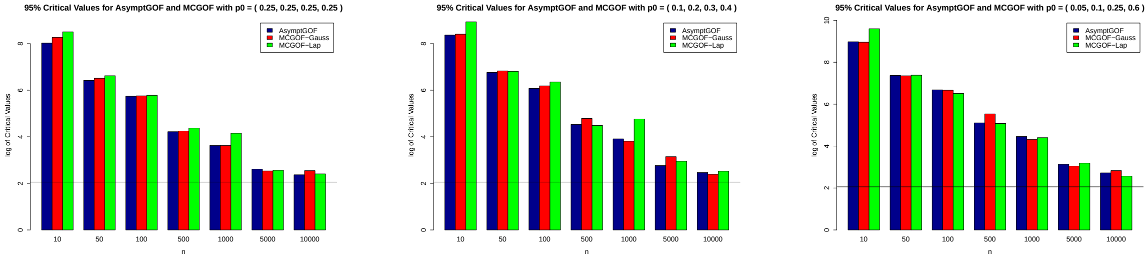


Figure 7: (Log of) Critical values of `AsymptGOF` and `MCGOF` (with $m = 59$) with both Gaussian ($\rho = 0.00125$) and Laplace noise ($\epsilon = 0.1$) along with the classical critical value as the black line.

We also plot the critical values of `AsymptGOF` in Figure 7. For `AsymptGOF` we used the package in R "CompQuadForm" that has various methods for finding estimates to the tail probabilities for quadratic forms of normals, of which we used the "imhof" method (Imhof, 1961) to approximate the threshold for each test. Note that for reasonably sized data, the critical values must be much larger than in the classical test, given as the horizontal line in the plots.

---

[5] The variance of Lap(b) is $2b^2$.

To show that `AsymptGOF` works beyond $d = 4$ multinomial data, we give a table of results in Table 3 for $d = 100$ data and null hypothesis $p_i^0 = 1/100$ for $i \in [100]$. We give the proportion of $10,000$ trials that were rejected by `AsymptGOF` in the "`AsymptGOF` Type I" column and those that were rejected after adding Gaussian noise for privacy by the classical test `GOF` in the "`GOF` Type I" column. Note that the critical value that `GOF` uses is 123.23 for every test in this case, whereas `AsymptGOF`'s critical value changes for each test, given in the column labeled "τ_α".

Table 3: GOF testing with $\alpha = 0.05$ and 0.00125-zCDP for $d = 100$.

$\boldsymbol{p}^0$			n	$\chi^2_{d-1,1-\alpha}$	`GOF` Type I	τ^α	`AsymptGOF` Type I
0.01	$\cdots$	0.01	1,000	123.23	1.00	10,070.47	0.0503
0.01	$\cdots$	0.01	10,000	123.23	1.00	1,117.85	0.0494
0.01	$\cdots$	0.01	100,000	123.23	0.9923	222.64	0.0506
0.01	$\cdots$	0.01	1,000,000	123.23	0.1441	133.16	0.0491

6.3.2. Independence Testing

We then turn to independence testing for 2×2 contingency tables using `AsymptIndep` and `MCIndep` with both Laplace and Gaussian Noise. Note that our methods do apply to arbitrary $r \times c$ tables and run in time $\text{poly}(r, c, \log(n))$ plus the time for the iterative Imhof method to find the critical values. For `MCIndep` and `AsymptIndep` we sample 1,000 trials for various parameters $\boldsymbol{\pi}^{(1)} = (\pi^{(1)}, 1 - \pi^{(1)})$, $\boldsymbol{\pi}^{(2)} = (\pi^{(2)}, 1 - \pi^{(2)})$, and n that could have generated the contingency tables. We set the number of samples $m = 59$ in `MCIndep` regardless of the noise we added and when we use Laplace noise, we set $\gamma = 0.01$ as the parameter in `2MLE`. In Figure 8 we compute the empirical Type I Error of `AsymptIndep`. Further, in Figure 9 and Figure 10 we give the empirical Type I Error on `MCIndep` with Gaussian and Laplace noise, respectively.

Note that when n is small, we get that our private independence tests almost always fail to reject. In fact, when $n = 500$ all of our tests in 1,000 trials fail to reject when $\rho = 0.00005$. This is due to `2MLE` releasing a contingency table based on the private counts with small cell counts. When the cell counts in `2MLE` are small we follow the "rule of thumb" from the classical test `Indep` and output NULL, which results in `AsymptIndep` failing to reject. This will ensure good significance but makes no promises on power for small n, as does the classical test `Indep`. Further, another consequence of this "rule of thumb" is that when we use `Indep` on private counts, with either Laplace or Gaussian noise, it tends to have lower Type I error than for larger n. In fact, it seems like the `AsymptIndep` test is more conservative, meaning the empirical Type I error is smaller than the threshold α, than the MC approach. We can then expect `AsymptIndep` to have worse power.

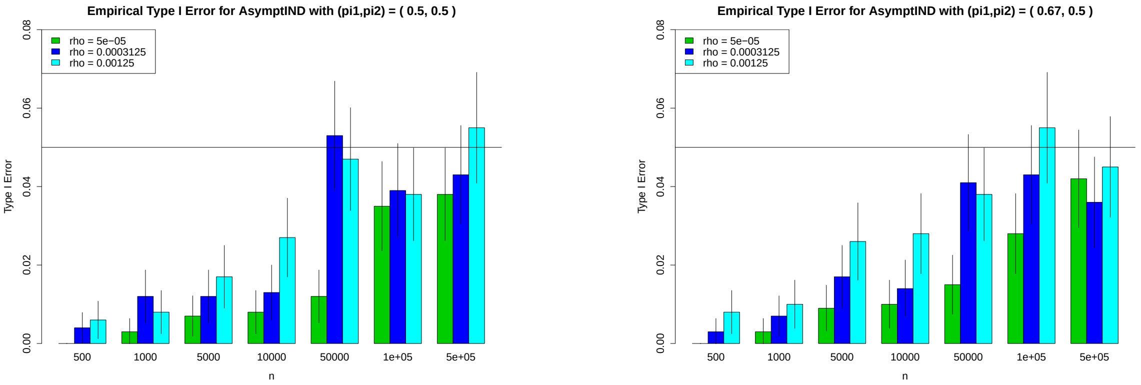


Figure 8: Empirical Type I Error of `AsymptIndep` with error bars corresponding to 1.96 times the standard error in 1,000 trials.

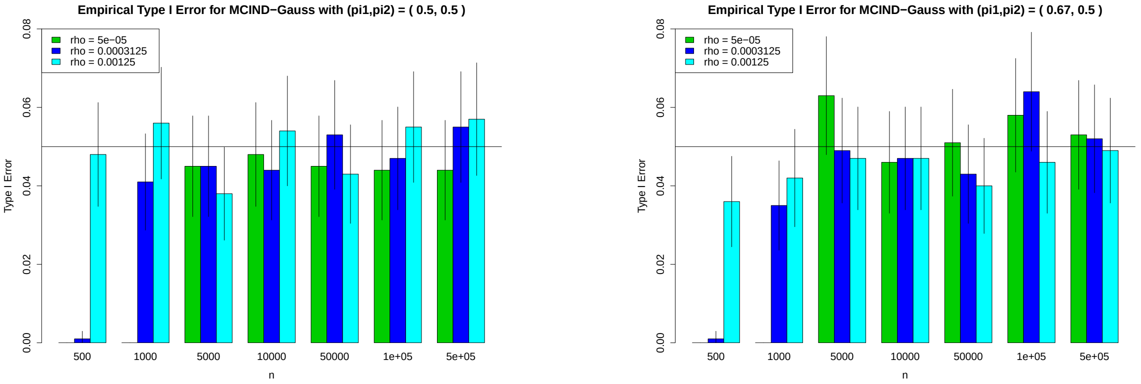


Figure 9: Empirical Type I Error of `MCIndep` using Gaussian noise with error bars corresponding to 1.96 times the standard error in 1,000 trials.

6.4. Power Results

6.4.1. GOF Testing

We now want to show that our tests can correctly reject H_0 when it is false. For our two goodness of fit tests, $\texttt{MCGOF}(\mathcal{Z})$ (with $m = 59$ and either Laplace or Gaussian noise) and `AsymptGOF` we test whether the multinomial data came from $\boldsymbol{p}^0$ when it was actually sampled from $\boldsymbol{p}^1 = \boldsymbol{p}^0 + \boldsymbol{\Delta}$. We compare our zCDP goodness of fit tests `AsymptGOF` and $\texttt{MCGOF}(\mathrm{N}\left(0, 1/\rho\right))$ with the non-private `GOF` test that uses the unaltered data in Figure 11. Then, we compare our differentially private goodness of fit tests $\texttt{MCGOF}(\mathrm{Lap}(2/\epsilon))$ with the non-private test in Figure 12. We then find the proportion of 1,000 trials that each of our tests rejected $\mathrm{H}_0 : \boldsymbol{p} = \boldsymbol{p}^0$ for various n. Note that `GOF` has difficulty distinguishing $\boldsymbol{p}^0$ and $\boldsymbol{p}^1$ for reasonable sample sizes. From the plots, it appears that the MC methods have empirically less power than the asymptotic methods. However the MC methods do provide the guarantee of significance at least $1 - \alpha$.

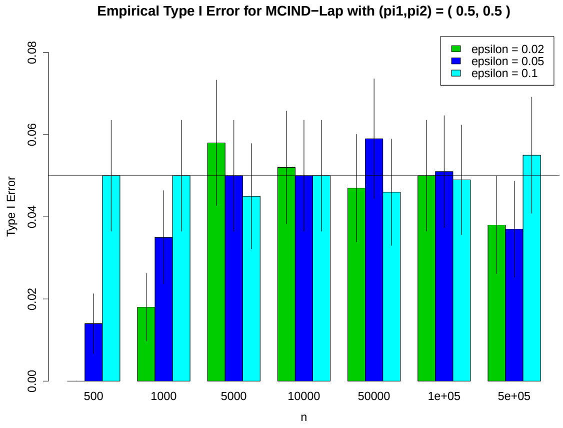


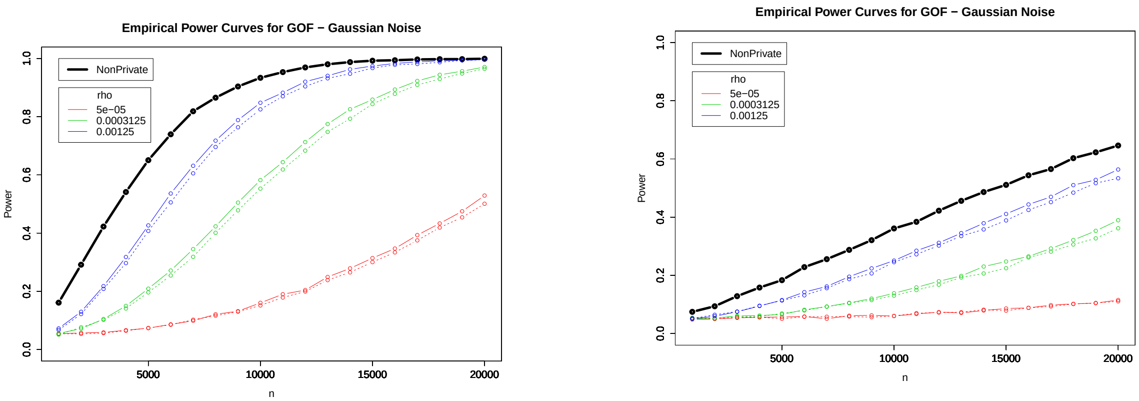
Figure 10: Empirical Type I Error of `MCIndep` using Laplace noise with error bars corresponding to 1.96 times the standard error in $1,000$ trials.

6.4.2. *Independence Testing*

We then turn to independence testing for 2×2 tables with our two differentially private tests `MCIndep` and `AsymptIndep`. We fix the alternate H_1 so that $\boldsymbol{Y}^1 \sim \text{Multinomial}(1, \boldsymbol{\pi}^{(1)} = (\pi^{(1)}, 1 - \pi^{(1)}))$ and $\boldsymbol{Y}^2 \sim \text{Multinomial}(1, \boldsymbol{\pi}^{(1)} = (\pi^{(2)}, 1 - \pi^{(2)}))$ are not independent. We then sample contingency tables from a multinomial distribution with probability $\boldsymbol{\pi}^{(1)} \left(\boldsymbol{\pi}^{(2)}\right)^{\intercal} + \boldsymbol{\Delta}$ and various sizes n. We compute the proportion of $1,000$ trials that `MCIndep` and `AsymptIndep` rejected $\mathrm{H}_0 : \boldsymbol{Y}^1 \perp \boldsymbol{Y}^2$ in Figure 13 and Figure 14 for Gaussian and Laplace noise, respectively. For `MCIndep` we set the number of samples $m = 59$ and when we use Laplace noise, we set $\gamma = 0.01$ in `2MLE`.

6.5. Conclusion

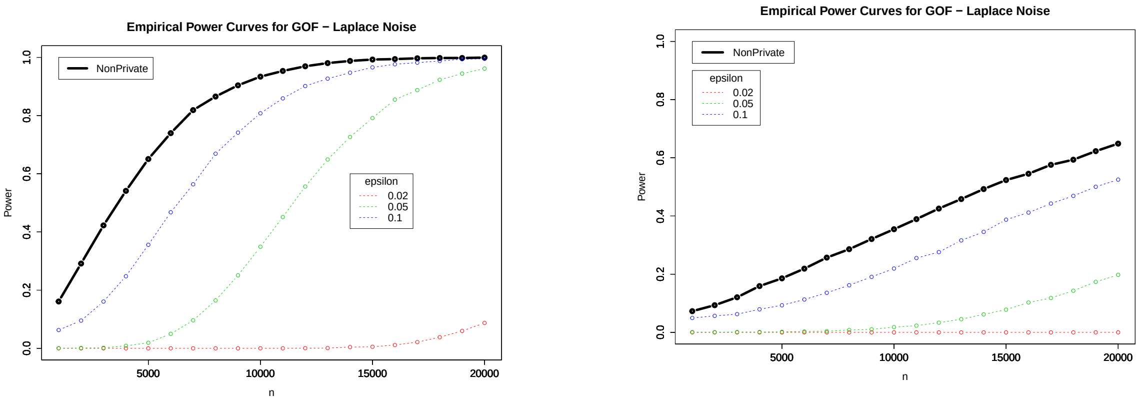
We proposed new hypothesis tests based on a private version of the chi-square statistic for goodness of fit and independence tests. For each test, we showed analytically or experimentally that we can achieve significance close to the target $1 - \alpha$ level similar to the nonprivate tests. We also showed that all the tests have a loss in power with respect to the non-private classical tests. Depending on the privacy benchmark, we would add Laplace or Gaussian noise. If Gaussian noise was used, we computed the asymptotic distribution of the chi-square statistic, so that we could bypass an MC approach. Typically, one would expect differential privacy to require the sample size to blow up by a multiplicative $1/\epsilon$ factor. However, we see a better performance because the noise is dominated by the sampling error for certain privacy levels.



(a) $H_0 : \boldsymbol{p} = \boldsymbol{p}^0 = (1/4, 1/4, 1/4, 1/4)^\intercal$ and $H_1 : \boldsymbol{p}^0 + \boldsymbol{\Delta}$ where $\boldsymbol{\Delta} = 0.01 \cdot (1, -1, -1, 1)^\intercal$.

(b) $H_0 : \boldsymbol{p}^0 = (1/2, 1/6, 1/6, 1/6)^\intercal$ and $H_1 : \boldsymbol{p}^0 + \boldsymbol{\Delta}$ where $\boldsymbol{\Delta} = 0.01 \cdot (1, -1/3, -1/3, -1/3)^\intercal$.

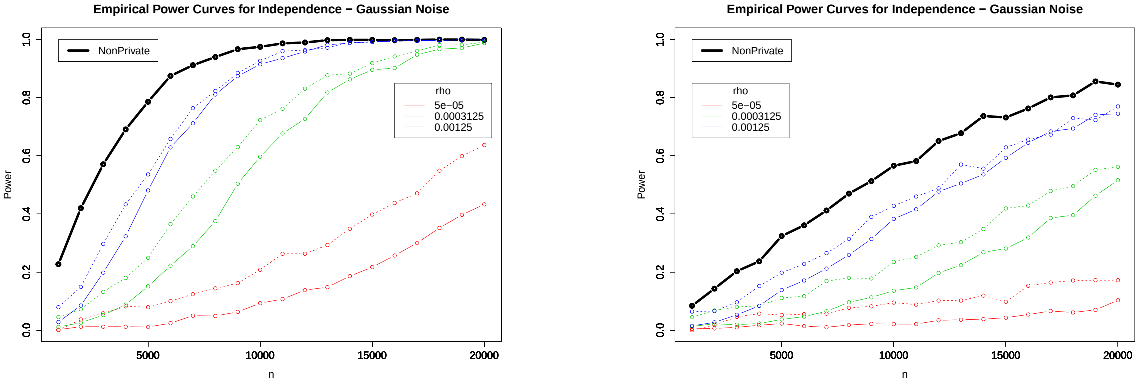
Figure 11: Comparison of empirical power of classical non-private test versus `AsymptGOF` (solid line) and `MCGOF` (dashed line) with Gaussian noise for alternate $H_1 : \boldsymbol{p}^1 = \boldsymbol{p}^0 + \boldsymbol{\Delta}$ in 10,000 trials.



(a) $H_0 : \boldsymbol{p} = \boldsymbol{p}^0 = (1/4, 1/4, 1/4, 1/4)^\intercal$ and $H_1 : \boldsymbol{p}^0 + \boldsymbol{\Delta}$ where $\boldsymbol{\Delta} = 0.01 \cdot (1, -1, -1, 1)^\intercal$.

(b) $H_0 : \boldsymbol{p}^0 = (1/2, 1/6, 1/6, 1/6)^\intercal$ and $H_1 : \boldsymbol{p}^0 + \boldsymbol{\Delta}$ where $\boldsymbol{\Delta} = 0.01 \cdot (1, -1/3, -1/3, -1/3)^\intercal$.

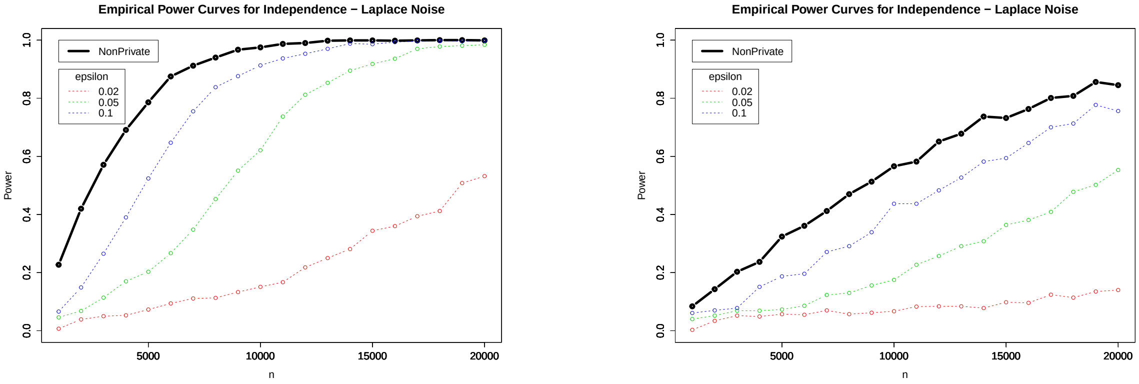
Figure 12: Comparison of empirical power of classical non-private test versus `MCGOF` with Laplace noise for alternate $H_1 : \boldsymbol{p}^1 = \boldsymbol{p}^0 + \boldsymbol{\Delta}$ in 10,000 trials.



(a) We set $\pi^{(1)} = \pi^{(2)} = 1/2$ and $\mathbf{\Delta} = 0.01 \cdot (1, -1)^\intercal(1, -1)$.

(b) We set $\pi^{(1)} = 2/3, \pi^{(2)} = 1/2$ and $\mathbf{\Delta} = 0.01 \cdot (1, -1)^\intercal(1, 0)$.

Figure 13: Comparison of empirical power of classical non-private test versus `AsymptIndep` (solid line) and `MCIndep` (dashed line) with Gaussian noise in $1,000$ trials.



(a) We set $\pi^{(1)} = \pi^{(2)} = 1/2$ and $\mathbf{\Delta} = 0.01 \cdot (1, -1)^\intercal(1, -1)$.

(b) We set $\pi^{(1)} = 2/3, \pi^{(2)} = 1/2$ and $\mathbf{\Delta} = 0.01 \cdot (1, -1)^\intercal(1, 0)$.

Figure 14: Comparison of empirical power of classical non-private test versus `MCIndep` with Laplace noise in $1,000$ trials.

CHAPTER 7

PRIVATE GENERAL CHI-SQUARE TESTS

This chapter largely follows the work in Kifer and Rogers (2016). Our main contribution here is a general template for creating test statistics involving categorical data. Empirically, they improve on the power of previous work on differentially private hypothesis testing, namely the tests given in Chapter 6 and Wang et al. (2015), while maintaining at most some given Type I error α. Our approach is to select certain properties of non-private hypothesis tests (e.g., their asymptotic distributions) and then build new test statistics that match these properties when Gaussian noise is added. Although the test statistics are designed with Gaussian noise in mind, other noise distributions can be applied, e.g. Laplace.[1]

7.1. General Chi-Square Tests

We start by providing a general framework for hypothesis tests involving categorical data that will include goodness of fit and independence testing as special cases, which we covered in Chapter 6. In the non-private setting, a chi-square test involves a histogram $\boldsymbol{H}$ and a model H_0 that produces expected counts $\bar{\boldsymbol{H}}$ over the d buckets. In general, H_0 will have fewer than d parameters and will estimate the parameters from $\boldsymbol{H}$. The chi-square test statistic is defined as the following (compare this statistic to the statistic $\mathrm{T}^{(n)}$ from (6.1) for GOF testing)

$$\mathrm{T}^{(\mathrm{n})}_{\mathrm{chi}} = \sum_{i=1}^{d} (H_i - \bar{H}_i)^2 / \bar{H}_i.$$

If the data were generated from H_0 and if s parameters had to be estimated, then the asymptotic distribution of $\mathrm{T}^{(\mathrm{n})}_{\mathrm{chi}}$ is χ^2_{d-s-1}, a chi-square random variable with $d - s - 1$ degrees of freedom. This is the property we want our statistics to have when they are computed from the noisy histogram $\widetilde{\boldsymbol{H}}$ instead of $\boldsymbol{H}$. Note that in the classical chi-square tests (e.g. Pearson independence test `GOF` in Algorithm 7), the statistic $\mathrm{T}^{(\mathrm{n})}_{\mathrm{chi}}$ is computed and if it is larger than the $1 - \alpha$ percentile of χ^2_{d-s-1}, then the model is rejected.

The above facts are part of a more general *minimum chi-square asymptotic theory* (Ferguson, 1996), which we overview in Section 7.1.2. However, we first explain the differences between private and non-private asymptotics.

7.1.1. Private Asymptotics

In non-private statistics, a function of n data records is considered a random variable, and non-private asymptotics considers this distribution as $n \to \infty$. In private asymptotics, there is another quantity σ_n^2, the variance of the added noise.

---

[1] If we use Laplace noise instead, we cannot match properties like the asymptotic distribution of the non-private statistics, but the new test statistics still empirically improve the power of the tests from previous works.

In the *classical private regime*, one studies what happens as $n/\sigma_n^2 \to \infty$; i.e., when the variance due to privacy is insignificant compared to sampling variance in the data (i.e. $O(n)$), as we assumed in Lemma 6.1.3. In practice, asymptotic distributions derived under this regime result in unreliable hypothesis tests because privacy noise is significant; e.g. see Figure 5 and Uhler et al. (2013).

In the *variance-aware private regime*, one studies what happens as $n/\sigma_n^2 \to$ constant as $n \to \infty$; that is, when the variance due to privacy is proportional to sampling variance. In practice, asymptotic distributions derived under this regime result in hypothesis tests with reliable Type I error (i.e. the p-values they generate are accurate); see Chapter 6 and Wang et al. (2015). [2]

7.1.2. Minimum Chi-Square Theory

In this section, we present important results about *minimum chi-square theory*. The discussion is based largely on Ferguson (1996) (see Chapter 23 there). Our work relies on this theory to construct new private test statistics in Section 7.2 and Section 7.3 whose asymptotic behavior matches the non-private asymptotic behavior of the classical chi-square test.

We consider a sequence of d-dimensional random vectors $\mathbf{V}^{(n)}$ for $n \geq 1$ (e.g. the data histogram). The parameter space Θ is a non-empty open subset of $\mathbb{R}^s$, where $s \leq d$. The model A maps a s-dimensional parameter $\boldsymbol{\theta} \in \Theta$ into a d-dimensional vector (e.g., the expected value of $\mathbf{V}^{(n)}$), hence it maps Θ to a subset of a s-dimensional manifold in d-dimensional space.

In this abstract setting, the null hypothesis H_0 is that there exists a $\boldsymbol{\theta}^0 \in \Theta$ such that:

$$\sqrt{n}\left(\mathbf{V}^{(n)} - A(\boldsymbol{\theta}^0)\right) \xrightarrow{D} \mathrm{N}\left(0, C(\boldsymbol{\theta}^0)\right) \tag{7.1}$$

where $C(\boldsymbol{\theta}) \in \mathbb{R}^{d\times d}$ is an invertible covariance matrix. Intuitively, Equation 7.1 says that the Central Limit Theorem can be applied for $\boldsymbol{\theta}^0$.

We measure the distance between $\mathbf{V}^{(n)}$ and $A(\boldsymbol{\theta})$ with a test statistic given by the following quadratic form:

$$\mathrm{D}^{(n)}(\boldsymbol{\theta}) \overset{\text{defn}}{=} n\left(\mathbf{V}^{(n)} - A(\boldsymbol{\theta})\right)^{\intercal} M(\boldsymbol{\theta})\left(\mathbf{V}^{(n)} - A(\boldsymbol{\theta})\right) \tag{7.2}$$

where $M(\boldsymbol{\theta}) \in \mathbb{R}^{d\times d}$ is a symmetric positive-semidefinite matrix; different choices of M will result in different test statistics. We make the following standard assumptions about $A(\boldsymbol{\theta})$ and $M(\boldsymbol{\theta})$.

Assumption 7.1.1. *For all $\boldsymbol{\theta} \in \Theta$, we have:*

---

[2] Note that taking n and σ_n^2 to infinity is just a mathematical tool for simplifying expressions while mathematically keeping privacy noise variance proportional to the data variance; it does not mean that the amount of actual noise added to the data depends on the data size.

- *$A(\boldsymbol{\theta})$ is bicontinuous,*[3]
- *$A(\boldsymbol{\theta})$ has continuous first partial derivatives, denoted as $\dot{A}(\boldsymbol{\theta})$ with full rank s,*
- *$M(\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$ and there exists an $\eta > 0$ such that $M(\boldsymbol{\theta}) - \eta I_d$ is positive definite in an open neighborhood of $\boldsymbol{\theta}^0$.*

The following theorem will be useful in determining the distribution for the quadratic form $\mathrm{D}^{(n)}(\boldsymbol{\theta})$.
Theorem 7.1.2 [Ferguson (1996)]**.** *Let $\boldsymbol{W} \sim \mathrm{N}(0, \Lambda)$. $\boldsymbol{W}^\intercal \boldsymbol{W} \sim \chi^2_r$ if and only if Λ is a projection of rank r. If $\Lambda \in \mathbb{R}^{d\times d}$ is invertible, $\boldsymbol{W}^\intercal \Lambda^{-1} \boldsymbol{W} \sim \chi^2_d$.*

If $\boldsymbol{\theta}^0$ is known, setting $M(\boldsymbol{\theta}) = C(\boldsymbol{\theta})^{-1}$ in (7.2) and applying Theorem 7.1.2 shows that then $\mathrm{D}^{(n)}(\boldsymbol{\theta}^0)$ converges in distribution to χ^2_d. However, as we show in Section 7.2, this can be a sub-optimal choice of middle matrix M.

When $\boldsymbol{\theta}^0$ is not known, we need to estimate a good parameter $\widehat{\boldsymbol{\theta}}^{(n)}$ to plug into (7.2). One approach is to set $\widehat{\boldsymbol{\theta}}^{(n)} = \arg\min_{\boldsymbol{\theta}\in\Theta} \mathrm{D}^{(n)}(\boldsymbol{\theta})$. However, this can be a difficult optimization. If there is a rough estimate of $\boldsymbol{\theta}^0$ based on the data, call it $\phi(\mathbf{V}^{(n)})$, and if it converges in probability to $\boldsymbol{\theta}^0$ (i.e. $\phi(\mathbf{V}^{(n)}) \xrightarrow{P} \boldsymbol{\theta}^0$ as $n \to \infty$), then we can plug it into the middle matrix to get:

$$\widehat{\mathrm{D}}^{(n)}(\boldsymbol{\theta}) \overset{\text{defn}}{=} n \left(\mathbf{V}^{(n)} - A(\boldsymbol{\theta})\right)^\intercal M(\phi(\mathbf{V}^{(n)})) \left(\mathbf{V}^{(n)} - A(\boldsymbol{\theta})\right). \tag{7.3}$$

and then set our estimator $\widehat{\boldsymbol{\theta}}^{(n)} = \arg\min_{\boldsymbol{\theta}\in\Theta} \widehat{\mathrm{D}}^{(n)}(\boldsymbol{\theta})$. The test statistic becomes $\widehat{\mathrm{D}}^{(n)}(\widehat{\boldsymbol{\theta}}^{(n)})$ and the following theorems describe its asymptotic properties under the null hypothesis.

We use the shorthand $A = A(\boldsymbol{\theta}^0)$, $M = M(\boldsymbol{\theta}^0)$, and $C = C(\boldsymbol{\theta}^0)$. The following result follows a similar argument as Theorem 23 in Ferguson (1996).
Theorem 7.1.3. *Let $\widehat{\boldsymbol{\theta}}^{(n)} = \mathrm{argmin}_{\boldsymbol{\theta}\in\Theta} \widehat{\mathrm{D}}^{(n)}(\boldsymbol{\theta})$. Given Assumption 7.1.1 and (7.1), we have $\sqrt{n}(\widehat{\boldsymbol{\theta}}^{(n)} - \boldsymbol{\theta}^0) \xrightarrow{D} \mathrm{N}(\mathbf{0}, \Psi)$ where $\boldsymbol{\theta}^0$ is the true parameter and*

$$\Psi = \left(\dot{A}^\intercal M \dot{A}\right)^{-1} \dot{A}^\intercal M C M \dot{A} \left(\dot{A}^\intercal M \dot{A}\right)^{-1}.$$

Proof. Since $\phi(\mathbf{V}^{(n)})$ converges in probability to $\boldsymbol{\theta}^0$ and $M(\cdot)$ is a continuous mapping, then for any $b > 0, c > 0$ there exists an n_0 such that when $n \geq n_0$ then $M(\phi(\mathbf{V}^{(n)}))$ is within a distance b from $M(\boldsymbol{\theta}^0)$ with probability at least $1 - c$, which makes $M(\phi(\mathbf{V}^{(n)}))$ positive definite with high probability for sufficiently large n. Furthermore, for any $g > 0$, we can choose n large enough so that the smallest eigenvalue of $M(\phi(\mathbf{V}^{(n)}))$ is at least $\eta - g$, by assumption.

---

[3] i.e. $\boldsymbol{\theta}_j \to \boldsymbol{\theta} \Leftrightarrow A(\boldsymbol{\theta}_j) \to A(\boldsymbol{\theta})$.

Since the parameter space is compact, we know a minimizer exists for $\widehat{\mathrm{D}}^{(n)}(\boldsymbol{\theta})$ Together, this implies that for sufficiently large n and with high probability $\widehat{\mathrm{D}}^{(n)}(\widehat{\boldsymbol{\theta}}^{(n)}) \geq 0$.

Also, $\widehat{\mathrm{D}}^{(n)}(\widehat{\boldsymbol{\theta}}^{(n)}) \leq \widehat{\mathrm{D}}^{(n)}(\boldsymbol{\theta}^0)$ but $\widehat{\mathrm{D}}^{(n)}(\boldsymbol{\theta}^0)/n \xrightarrow{P} 0$ since $M(\phi(\mathbf{V}^{(n)})) \xrightarrow{P} M$ and $\mathbf{V}^{(n)} \xrightarrow{P} A$. Thus $\widehat{\mathrm{D}}^{(n)}(\widehat{\boldsymbol{\theta}}^{(n)})/n \xrightarrow{P} 0$ which means $\mathbf{V}^{(n)} - A(\widehat{\boldsymbol{\theta}}^{(n)}) \xrightarrow{P} 0$ (since $M(\phi(\mathbf{V}^{(n)}))$ is positive definite with high probability and uniformly bounded away from 0 in a neighborhood of $\boldsymbol{\theta}^0$). This implies that $A(\widehat{\boldsymbol{\theta}}^{(n)}) \xrightarrow{P} A$ and so $\widehat{\boldsymbol{\theta}}^{(n)} \xrightarrow{P} \boldsymbol{\theta}^0$ since $A(\boldsymbol{\theta})$ is bicontinuous by assumption.

Thus, with high probability (e.g., $\geq 1-c$ for large enough n), $\widehat{\boldsymbol{\theta}}^{(n)}$ satisfies the first order optimality condition $\nabla\widehat{\mathrm{D}}^{(n)}(\widehat{\boldsymbol{\theta}}^{(n)}) = 0$. This is the same as

$$\dot{A}(\widehat{\boldsymbol{\theta}}^{(n)})^\intercal M(\phi(\mathbf{V}^{(n)})(\mathbf{V}^{(n)} - A(\widehat{\boldsymbol{\theta}}^{(n)})) = 0 \tag{7.4}$$

Expanding $A(\widehat{\boldsymbol{\theta}}^{(n)})$ around $\boldsymbol{\theta}^0$.

$$A(\widehat{\boldsymbol{\theta}}^{(n)}) = A(\boldsymbol{\theta}^0) + \underbrace{\int_0^1 \dot{A}(\boldsymbol{\theta}^0 + t(\widehat{\boldsymbol{\theta}}^{(n)} - \boldsymbol{\theta}^0))\, dt}_{\overset{\text{defn}}{=} B(\widehat{\boldsymbol{\theta}}^{(n)})}(\widehat{\boldsymbol{\theta}}^{(n)} - \boldsymbol{\theta}^0) \tag{7.5}$$

Substituting (7.5) into (7.4), we get:

$$\dot{A}(\widehat{\boldsymbol{\theta}}^{(n)})^\intercal M(\phi(\mathbf{V}^{(n)}))\Big(\mathbf{V}^{(n)} - A(\boldsymbol{\theta}^0) - B(\widehat{\boldsymbol{\theta}}^{(n)})(\widehat{\boldsymbol{\theta}}^{(n)} - \boldsymbol{\theta}^0)\Big) = 0 \tag{7.6}$$

$$\dot{A}(\widehat{\boldsymbol{\theta}}^{(n)})^\intercal M(\phi(\mathbf{V}^{(n)}))B(\widehat{\boldsymbol{\theta}}^{(n)})\sqrt{n}(\widehat{\boldsymbol{\theta}}^{(n)} - \boldsymbol{\theta}^0) = \dot{A}(\widehat{\boldsymbol{\theta}}^{(n)})^\intercal M(\phi(\mathbf{V}^{(n)}))\sqrt{n}(\mathbf{V}^{(n)} - p(\boldsymbol{\theta}^0)) \tag{7.7}$$

Now, by the continuity of $\dot{A}(\cdot)$ and the definition of $B(\cdot)$ and the convergence in probability of $\widehat{\boldsymbol{\theta}}^{(n)}$ to $\boldsymbol{\theta}^0$, we have $B(\widehat{\boldsymbol{\theta}}^{(n)}) \xrightarrow{P} \dot{A}(\boldsymbol{\theta}^0)$. Since $\dot{A}(\boldsymbol{\theta})$ has full rank by assumption, then for sufficiently large n, $B(\widehat{\boldsymbol{\theta}}^{(n)})$ has full rank with high probability. This leads to the following expression with high probability for sufficiently large n,

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}^{(n)} - \boldsymbol{\theta}^0) = \Big(\dot{A}(\widehat{\boldsymbol{\theta}}^{(n)})^\intercal M(\phi(\mathbf{V}^{(n)}))B(\widehat{\boldsymbol{\theta}}^{(n)})\Big)^{-1}\dot{A}(\widehat{\boldsymbol{\theta}}^{(n)})^\intercal M(\phi(\mathbf{V}^{(n)}))\sqrt{n}(\mathbf{V}^{(n)} - A) \tag{7.8}$$

Since $M(\phi(\mathbf{V}^{(n)}))$ has smallest eigenvalue at least $\eta - g > 0$ with high probability for n large enough, and since $\phi(\mathbf{V}^{(n)}) \xrightarrow{P} \boldsymbol{\theta}^0$, $\widehat{\boldsymbol{\theta}}^{(n)} \xrightarrow{P} \boldsymbol{\theta}^0$, $B(\widehat{\boldsymbol{\theta}}^{(n)}) \to \dot{A}(\boldsymbol{\theta}^0)$ in probability, using continuity in all of the above functions, and the assumption that $\sqrt{n}(\mathbf{V}^{(n)} - A) \to \mathrm{N}(0, C)$ in distribution (and Slutsky's theorem) we get:

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}^{(n)} - \boldsymbol{\theta}^0) \xrightarrow{D} N(0, \Psi) \qquad \text{as} \qquad n \to \infty. \tag{7.9}$$

□

We then prove the following result using a slight modification of Theorem 24 in Ferguson

(1996).

Theorem 7.1.4. *Let ν be the rank of $C(\boldsymbol{\theta}_0)$. If Assumption 7.1.1 and* (7.1) *hold, and, for all $\boldsymbol{\theta} \in \Theta$,*

$$C(\boldsymbol{\theta})M(\boldsymbol{\theta})C(\boldsymbol{\theta}) = C(\boldsymbol{\theta})$$

and

$$C(\boldsymbol{\theta})M(\boldsymbol{\theta})\dot{A}(\boldsymbol{\theta}) = \dot{A}(\boldsymbol{\theta})$$

then for $\widehat{\boldsymbol{\theta}}^{(n)}$ given in Theorem 7.1.3 and $\widehat{\mathrm{D}}^{(n)}(\boldsymbol{\theta})$ given in (7.3) *we have:*

$$\widehat{\mathrm{D}}^{(n)}\left(\widehat{\boldsymbol{\theta}}^{(n)}\right) \xrightarrow{D} \chi^2_{\nu-s}.$$

Proof. Note that Theorem 24 in Ferguson (1996) shows that if the hypotheses hold then

$$n\left(\mathbf{V}^{(n)} - A(\widehat{\boldsymbol{\theta}}^{(n)})\right)^{\intercal} M(\widehat{\boldsymbol{\theta}}^{(n)})\left(\mathbf{V}^{(n)} - A(\widehat{\boldsymbol{\theta}}^{(n)})\right) \xrightarrow{D} \chi^2_{\nu-s}.$$

Note that we have $\phi(\mathbf{V}^{(n)}) \xrightarrow{P} \boldsymbol{\theta}^0$ and $\widehat{\boldsymbol{\theta}}^{(n)} \xrightarrow{P} \boldsymbol{\theta}^0$ for the true parameter $\boldsymbol{\theta}^0 \in \Theta$. We can then apply Slutsky's Theorem due to $M(\cdot)$ being continuous, to obtain the result for $\widehat{\mathrm{D}}^{(n)}(\widehat{\boldsymbol{\theta}}^{(n)})$. □

7.2. Private Goodness of Fit Tests

As we did in Chapter 6, we will first cover goodness of fit testing where the null hypothesis is simply testing whether the underlying unknown parameter is equal to a particular probability vector. This will be crucial in introducing our new statistics for testing in this chapter. Once again, we consider categorical data $\boldsymbol{H} = (H_1, \cdots, H_d)^{\intercal} \sim \text{Multinomial}(n, \boldsymbol{p})$ where $\boldsymbol{p} = (p_1, \cdots, p_d)$ is some probability vector over the d outcomes. We want to test the null hypothesis $\mathrm{H}_0 : \boldsymbol{p} = \boldsymbol{p}^0$, where each component of $\boldsymbol{p}^0$ is positive, but we want to do so in a private way. We then have the following classical result (Bishop et al., 1975).

Lemma 7.2.1. *Under the null hypothesis $\mathrm{H}_0 : \boldsymbol{p} = \boldsymbol{p}^0$, $\boldsymbol{H}/n$ is asymptotically normal*

$$\sqrt{n}\left(\frac{\boldsymbol{H}}{n} - \boldsymbol{p}^0\right) \xrightarrow{D} \mathrm{N}\left(\mathbf{0}, \Sigma^0\right)$$

where Σ^0 has rank $d-1$ and can be written as,[4]

$$\Sigma^0 \overset{\text{defn}}{=} \mathrm{Diag}(\boldsymbol{p}^0) - \boldsymbol{p}^0(\boldsymbol{p}^0)^{\intercal}. \tag{7.10}$$

---

[4] Compare Σ^0 with Σ from (6.3).

7.2.1. Nonprojected Private Test Statistic

As we did in Chapter 6, in order to preserve ρ-zCDP, we will add appropriately scaled Gaussian noise to each component of the histogram $\boldsymbol{H}$. We then define the zCDP statistic $\mathbf{V}_\rho^{(n)} = \left(\mathrm{V}_{\rho,1}^{(n)}, \cdots, \mathrm{V}_{\rho,d}^{(n)}\right)^\intercal$ where we write $\boldsymbol{Z} \sim \mathrm{N}\left(\mathbf{0}, 1/\rho \cdot I_d\right)$ and

$$\mathbf{V}_\rho^{(n)} \stackrel{\text{defn}}{=} \sqrt{n}\left(\frac{\boldsymbol{H}+\boldsymbol{Z}}{n} - \boldsymbol{p}^0\right). \tag{7.11}$$

We next derive the asymptotic distribution of $\mathbf{V}_\rho^{(n)}$ under both private asymptotic regimes defined in Section 7.1.1 (note that $\sigma^2 = 1/\rho$).

Lemma 7.2.2. *The random vector* $\mathbf{V}_{\rho_n}^{(n)}$ *from* (7.11) *under the null hypothesis* $\mathrm{H}_0 : \boldsymbol{p} = \boldsymbol{p}^0$ *has the following asymptotic distribution. If* $n\rho_n \to \infty$ *then* $\mathbf{V}_{\rho_n}^{(n)} \xrightarrow{D} \mathrm{N}\left(\mathbf{0}, \Sigma^0\right)$*. Further, if* $n\rho_n \to \rho^\star > 0$ *then* $\mathbf{V}_{\rho_n}^{(n)} \xrightarrow{D} \mathrm{N}\left(\mathbf{0}, \Sigma_{\rho^\star}^0\right)$ *where* $\Sigma_{\rho^\star}^0$ *has full rank and*

$$\Sigma_{\rho^\star}^0 \stackrel{\text{defn}}{=} \Sigma^0 + 1/\rho^\star \cdot I_d. \tag{7.12}$$

Proof. We know from the central limit theorem that $\mathbf{V}_{\rho^\star}^{(n)}$ will converge in distribution to a multivariate normal with covariance matrix given in (7.12). We now show that $\Sigma_{\rho^\star}^0$ is full rank. From (7.10) we know that Σ^0 is positive-semidefinite because it is a covariance matrix, hence it has all nonnegative eigenvalues. We then consider the eigenvalues of $\Sigma_{\rho^\star}^0$. Let $\boldsymbol{v} \in \mathbb{R}^d$ be an eigenvector of $\Sigma_{\rho^\star}^0$ with eigenvalue $\lambda \in \mathbb{R}$, i.e.

$$\Sigma_{\rho^\star}^0 \, \boldsymbol{v} = \lambda \, \boldsymbol{v} \implies \Sigma^0 \, \boldsymbol{v} = \left(\lambda - 1/\rho^\star\right) \boldsymbol{b}.$$

We then must have that $\boldsymbol{v}$ is also an eigenvector of Σ^0. Because Σ^0 is positive-semidefinite we have the following inequality

$$\lambda - 1/\rho^\star \geq 0 \implies \lambda \geq 1/\rho^\star > 0.$$

Thus, all the eigenvalues of $\Sigma_{\rho^\star}^0$ are positive, which results in $\Sigma_{\rho^\star}^0$ being nonsingular. □

Because $\Sigma_{n\rho}^0$ is invertible when the privacy parameter $\rho > 0$, we can create a new statistic based on $\mathbf{V}_\rho^{(n)}$ that has a chi-square asymptotic distribution under variance-aware privacy asymptotics.

Theorem 7.2.3. *Let* $\mathbf{V}_{\rho_n}^{(n)}$ *be given in* (7.11) *for* $n\rho_n \to \rho^\star > 0$*. If the null hypothesis* $\mathrm{H}_0 : \boldsymbol{p} = \boldsymbol{p}^0$ *holds, then for* $\Sigma_{n\rho_n}^0$ *given in* (7.12)*, we have*

$$\mathrm{Q}_{\rho_n}^{(n)} \stackrel{\text{defn}}{=} \left(\mathbf{V}_{\rho_n}^{(n)}\right)^\intercal \left(\Sigma_{n\rho_n}^0\right)^{-1} \mathbf{V}_{\rho_n}^{(n)} \xrightarrow{D} \chi_d^2. \tag{7.13}$$

Proof. We directly apply Theorem 7.1.2 with $\boldsymbol{W}^{(n)} = \left(\Sigma_{n\rho_n}^0\right)^{-1/2} \mathbf{V}_{\rho_n}^{(n)}$ which is asymptotically multivariate normal with mean zero and covariance $\left(\Sigma^0_{\rho^\star}\right)^{-1/2} \Sigma^0_{\rho^\star} \left(\Sigma^0_{\rho^\star}\right)^{-1/2} = I_d$. □

By computing the inverse of $\Sigma^0_{n\rho_n}$ we can simplify the statistic $\mathrm{Q}^{(n)}_{\rho_n}$.
Lemma 7.2.4. *We can rewrite the statistic in* (7.13) *as*

$$\mathrm{Q}^{(n)}_{\rho} = \sum_{i=1}^{d} \frac{\left(\mathrm{V}^{(n)}_{\rho,i}\right)^2}{p^0_i + \frac{1}{n\rho}} + \frac{n\rho}{\sum_{\ell=1}^{d} \frac{p^0_\ell}{p^0_\ell + \frac{1}{n\rho}}} \left(\sum_{j=1}^{d} \frac{p^0_j}{p^0_j + \frac{1}{n\rho}} \cdot \mathrm{V}^{(n)}_{\rho,j}\right)^2. \tag{7.14}$$

Proof. We begin by writing the inverse of the covariance matrix Σ^0_ρ from (7.12) by applying Woodbury's formula (Woodbury, 1950) which gives the inverse of a modified rank deficient matrix,

$$\left(\Sigma^0_\rho\right)^{-1} = \mathrm{Diag}(\boldsymbol{p}^0 + 1/\rho \cdot \mathbf{1})^{-1} + \frac{1}{1 - \boldsymbol{p}^0 \cdot \omega(\rho)} \omega(\rho)\omega(\rho)^\intercal \tag{7.15}$$

where $\omega(\rho) \overset{\text{defn}}{=} \left(\frac{p^0_1}{p^0_1 + 1/\rho}, \cdots, \frac{p^0_d}{p^0_d + 1/\rho}\right)^\intercal = \frac{\boldsymbol{p}^0}{\boldsymbol{p}^0 + 1/\rho \cdot \mathbf{1}}$.

We note that the vector $\mathbf{1}$ is an eigenvector of Σ_ρ and Σ^{-1}_ρ with eigenvalue $1/\rho$ and ρ, respectively. Letting $\widetilde{H}_i = H_i + Z_i$ leads to the test statistic

$$\begin{aligned}\left(\mathbf{V}^{(n)}_\rho\right)^\intercal \left(\Sigma^0_{n\rho}\right)^{-1} \mathbf{V}^{(n)}_\rho &= \sum_{i=1}^{d} \frac{(\widetilde{H}_i - np^0_i)^2}{np^0_i + 1/\rho} + \frac{1}{1 - \sum_i \frac{(p^0_i)^2}{p^0_i + \frac{1}{n\rho}}} \left(\sum_{i=1}^{d} \frac{(\widetilde{H}_i - np^0_i)}{\sqrt{n}} \frac{p^0_i}{p^0_i + \frac{1}{n\rho}}\right)^2 \\ &= \sum_{i=1}^{d} \frac{(\widetilde{H}_i - np^0_i)^2}{np^0_i + 1/\rho} + \frac{1}{1 - \sum_i \frac{(p^0_i)^2}{p^0_i + \frac{1}{n\rho}}} \left(\sum_{i=1}^{d} \frac{(\widetilde{H}_i - np^0_i)}{\sqrt{n}} \frac{p^0_i}{p^0_i + \frac{1}{n\rho}}\right)^2\end{aligned}$$

We can then rewrite the denominator of the coefficient of the second term,

$$1 - \sum_{i=1}^{d} \frac{(p^0_i)^2}{p^0_i + \frac{1}{n\rho}} = \sum_{i=1}^{d} \left(\frac{p^0_i(p^0_i + \frac{1}{n\rho})}{p^0_i + \frac{1}{n\rho}} - \frac{(p^0_i)^2}{p^0_i + \frac{1}{n\rho}}\right) = \frac{1}{n\rho} \cdot \sum_{i=1}^{d} \frac{p^0_i}{p^0_i + \frac{1}{n\rho}}.$$

Recalling the form of $\mathbf{V}^{(n)}_\rho$ from (7.11) concludes the proof. □

Note that the coefficient on the second term of (7.14) grows large as $n\rho \to \infty$, so this test statistic does not approach the nonprivate test for a fixed ρ. This is not surprising since $\Sigma^0_{n\rho}$ must converge to a singular matrix as $n\rho \to \infty$.

Further, the additional noise adds a degree of freedom to the asymptotic distribution of the original statistic. This additional degree of freedom results in increasing the point in which we reject the null hypothesis, i.e. the critical value. Thus, rejecting an incorrect model becomes harder as we increase the degrees of freedom, and hence decreases power.

7.2.2. Projected Private Test Statistic

Given that the test statistic in the previous section depends on a nearly singular matrix, we now derive a new test statistic for the private goodness of fit test. It has the remarkable property that its asymptotic distribution is χ^2_{d-1} under both private asymptotics.

We start with the following observation. In the classical chi-square test, the random variable $\left(\frac{(H_i - np_i^0)}{\sqrt{np_i^0}}\right)_{i=1}^d$ has covariance matrix $\Sigma^0 = I_d - \sqrt{\boldsymbol{p}^0}\sqrt{\boldsymbol{p}^0}^\intercal$ under the null hypothesis $\mathrm{H}_0 : \boldsymbol{p} = \boldsymbol{p}^0$. The classical test essentially uncorrelates these random variables and projects them onto the subspace orthogonal to $\sqrt{\boldsymbol{p}^0}$. We will use a similar intuition for the privacy-preserving random vector $\mathbf{V}_\rho^{(n)}$.

The matrix Σ_ρ^0 in (7.12) has eigenvector $\mathbf{1}$ with eigenvalue $1/\rho$ – regardless of the true parameters of the data-generating distribution. Hence we think of this direction as pure noise. We therefore project $\mathbf{V}_\rho^{(n)}$ onto the space orthogonal to $\mathbf{1}$ (i.e. enforce the constraint that the entries in $\mathbf{V}_\rho^{(n)}$ add up to 0, as they would in the noiseless case). We then define the *projected statistic* $\mathcal{Q}_\rho^{(n)}$ as the following where we write the projection matrix $\Pi \overset{\text{defn}}{=} I_d - \frac{1}{d}\mathbf{1}\mathbf{1}^\intercal$

$$\mathcal{Q}_\rho^{(n)} \overset{\text{defn}}{=} \left(\mathbf{V}_\rho^{(n)}\right)^\intercal \; \Pi \; \left(\Sigma_{n\rho}^0\right)^{-1} \; \Pi \, \mathbf{V}_\rho^{(n)}. \tag{7.16}$$

It will be useful to write out the middle matrix in $\mathcal{Q}_{\rho_n}^{(n)}$ for analyzing its asymptotic distribution.

Lemma 7.2.5. *For the covariance matrix $\Sigma_{n\rho_n}^0$ given in* (7.12)*, we have the following identity when $n\rho_n \to \rho^\star > 0$*

$$\Pi \; \left(\Sigma_{n\rho_n}^0\right)^{-1} \; \Pi \to \left(\Sigma_{\rho^\star}^0\right)^{-1} - \frac{\rho^\star}{d} \cdot \mathbf{1}\mathbf{1}^\intercal \tag{7.17}$$

Further, when $n\rho_n \to \infty$, we have the following

$$\Pi \; \left(\Sigma_{n\rho_n}^0\right)^{-1} \; \Pi \to \Pi \; \left(\mathrm{Diag}\left(\boldsymbol{p}^0\right)\right)^{-1} \; \Pi \tag{7.18}$$

Proof. To prove (7.17), we use the fact that $\left(\Sigma_{\rho^\star}^0\right)^{-1}$ has eigenvalue $\rho^\star$ with eigenvector $\mathbf{1}$. We then focus on proving (7.18). We use the identity for the inverse of $\left(\Sigma_{n\rho_n}^0\right)^{-1}$ from (7.15).

$$\begin{aligned}
&\Pi \left(\Sigma_{n\rho_n}^0\right)^{-1} \Pi \\
&\quad = \Pi \left(\mathrm{Diag}(\boldsymbol{p}^0 + \frac{1}{n\rho_n} \cdot \mathbf{1})\right)^{-1} \Pi \\
&\qquad + \frac{n\rho_n}{\sum_{i=1}^d \frac{p_i^0}{p_i^0 + \frac{1}{n\rho_n}}} \cdot \Pi \left(\frac{\boldsymbol{p}^0}{\boldsymbol{p}^0 + \frac{1}{n\rho_n}\mathbf{1}}\right) \left(\frac{\boldsymbol{p}^0}{\boldsymbol{p}^0 + \frac{1}{n\rho_n}\mathbf{1}}\right)^\intercal \Pi
\end{aligned}$$

We then focus on the second term in the sum and write $\lambda_n = \sum_{i=1}^d \frac{p_i^0}{p_i^0 + \frac{1}{n\rho_n}}$.

$$
\begin{aligned}
&\frac{n\rho_n}{\lambda_n} \cdot \Pi \left(\frac{\boldsymbol{p}^0}{\boldsymbol{p}^0 + \frac{1}{n\rho_n}\mathbf{1}} \right) \left(\frac{\boldsymbol{p}^0}{\boldsymbol{p}^0 + \frac{1}{n\rho_n}\mathbf{1}} \right)^\intercal \Pi \\
&\quad = \frac{n\rho_n}{\lambda_n} \cdot \left(\frac{\boldsymbol{p}^0}{\boldsymbol{p}^0 + \frac{1}{n\rho_n}\mathbf{1}} - \frac{\lambda_n}{d} \cdot \mathbf{1} \right) \left(\frac{\boldsymbol{p}^0}{\boldsymbol{p}^0 + \frac{1}{n\rho_n} \cdot \mathbf{1}} - \frac{\lambda_n}{d} \cdot \mathbf{1} \right)^\intercal \\
&\quad = \frac{n\rho_n}{\lambda_n} \cdot \left(\frac{\boldsymbol{p}^0}{\boldsymbol{p}^0 + \frac{1}{n\rho_n} \cdot \mathbf{1}} \right) \left(\frac{\boldsymbol{p}^0}{\boldsymbol{p}^0 + \frac{1}{n\rho_n} \cdot \mathbf{1}} \right)^\intercal \\
&\qquad - \frac{n\rho_n}{d} \cdot \left(\frac{\boldsymbol{p}^0}{\boldsymbol{p}^0 + \frac{1}{n\rho_n} \cdot \mathbf{1}} \right) \mathbf{1}^\intercal - \frac{n\rho_n}{d} \cdot \mathbf{1} \left(\frac{\boldsymbol{p}^0}{\boldsymbol{p}^0 + \frac{1}{n\rho_n} \cdot \mathbf{1}} \right)^\intercal + \frac{n\rho_n \lambda_n}{d^2} \cdot \mathbf{1}\mathbf{1}^\intercal
\end{aligned}
$$

We consider entry (i, j) of the above matrix, which we can write as

$$
\begin{aligned}
&\frac{n\rho_n}{\lambda_n} \cdot \left(\frac{p_i^0}{p_i^0 + \frac{1}{n\rho_n}} \right) \cdot \left(\frac{p_j^0}{p_j^0 + \frac{1}{n\rho_n}} \right) - \frac{n\rho_n}{d} \left(\frac{p_i^0}{p_i^0 + \frac{1}{n\rho_n}} + \frac{p_j^0}{p_j^0 + \frac{1}{n\rho_n}} \right) + \frac{n\rho_n \lambda_n}{d^2} \\
&\quad = \frac{n\rho_n}{d\lambda_n} \left(\frac{\lambda_n^2}{d} - \frac{1}{(p_i^0 + \frac{1}{n\rho_n})(p_j^0 + \frac{1}{n\rho_n})} \left(\frac{\lambda_n}{n\rho_n}(p_i^0 + p_j^0) - p_i^0 p_j^0 (d - 2\lambda_n) \right) \right) \\
&\quad = n\rho_n \left(\frac{\lambda_n}{d^2} - \frac{(2\lambda_n - d) p_i^0 p_j^0}{d(p_i^0 + \frac{1}{n\rho_n})(p_j^0 + \frac{1}{n\rho_n})} \right) - \frac{p_i^0 + p_j^0}{d\lambda_n (p_i^0 + \frac{1}{n\rho_n})(p_j^0 + \frac{1}{n\rho_n})}.
\end{aligned}
$$

We then let $n \to \infty$ to get

$$
\begin{aligned}
&\frac{n\rho_n}{d} \left(\frac{\lambda_n}{d} - \frac{(2\lambda_n - d) p_i^0 p_j^0}{\lambda_n (p_i^0 + \frac{1}{n\rho_n})(p_j^0 + \frac{1}{n\rho_n})} \right) - \frac{p_i^0 + p_j^0}{d\lambda_n (p_i^0 + \frac{1}{n\rho_n})(p_j^0 + \frac{1}{n\rho_n})} \\
&\quad \to \frac{1}{p_i^0} + \frac{1}{p_j^0} - \frac{1}{p_i^0} - \frac{1}{p_j^0} = 0.
\end{aligned}
$$

Thus, we have shown that for $n\rho_n \to \infty$,

$$
\Pi \, \left(\Sigma_{n\rho_n}^0 \right)^{-1} \, \Pi \to \Pi \, \left(\mathrm{Diag}\left(\boldsymbol{p}^0 \right) \right)^{-1} \, \Pi.
$$

$\square$

We now show that the projected statistic is asymptotically chi-square distributed in both private asymptotic regimes, in fact we can prove it more generally for $\rho_n = \Omega(1/n)$.

Theorem 7.2.6. *Let* $\mathbf{V}_\rho^{(n)}$ *be given in* (7.11). *For null hypothesis* $\mathrm{H}_0 : \boldsymbol{p} = \boldsymbol{p}^0$, *we can write* *the projected statistic* $\mathcal{Q}_\rho^{(n)}$ *in the following way for* $\widetilde{n} = \sum_{i=1}^d (H_i + Z_i)$

$$\mathcal{Q}_\rho^{(n)} = \sum_{i=1}^d \frac{\left(\mathrm{V}_{\rho,i}^{(n)}\right)^2}{p_i^0 + \frac{1}{n\rho}} - \frac{\rho}{d}\left(\widetilde{n} - n\right)^2$$
$$+ \frac{n\rho}{\sum_{\ell=1}^d \frac{p_\ell^0}{p_\ell^0 + \frac{1}{n\rho}}} \left(\sum_{j=1}^d \frac{p_j^0}{p_j^0 + \frac{1}{n\rho}} \cdot \mathrm{V}_{\rho,j}^{(n)} \right)^2 . \tag{7.19}$$

For $\rho_n = \Omega(1/n)$*, if the null hypothesis holds then* $\mathcal{Q}_{\rho_n}^{(n)} \xrightarrow{D} \chi^2_{d-1}$*. Further, for* $n\rho_n \to \infty$*, the difference between* $\mathcal{Q}_{\rho_n}^{(n)}$ *and the classical chi-square statistic* $\sum_{i=1}^d \frac{\left(H_i - np_i^0\right)^2}{np_i^0}$ *converges in probability to 0.*

Proof. We first show that we can write the projected statistic in (7.16) in the proposed way. Using (7.17), we can write the projected statistic in terms of the nonprojected statistic in (7.14), which will give the expression in (7.19)

$$\mathcal{Q}_{\rho_n}^{(n)} = \left(\mathbf{V}_{\rho_n}^{(n)}\right)^\intercal \left(\left(\Sigma_{n\rho_n}^0\right)^{-1} - \frac{n\rho_n}{d} \cdot \mathbf{1}\mathbf{1}^\intercal\right) \mathbf{V}_{\rho_n}^{(n)} = \mathrm{Q}_{\rho_n}^{(n)} - \frac{n\rho_n}{d} \cdot \left(\mathbf{V}_{\rho_n}^{(n)}\right)^\intercal \mathbf{1}\mathbf{1}^\intercal \mathbf{V}_{\rho_n}^{(n)}.$$

Recall that $\mathbf{1}$ is an eigenvector of $\Sigma_{n\rho_n}^0$ for $n\rho_n > 0$, otherwise the matrix is not defined. Note that $\Sigma_{n\rho_n}^0$ is diagonalizable, i.e. $\Sigma_{n\rho_n}^0 = BDB^\intercal$ where D is a diagonal matrix and B is an orthogonal matrix with one column being $1/d \cdot \mathbf{1}$. For the following matrix Λ, we can write it as a $d \times d$ identity matrix except one of the entries on the diagonal is zero.

$$\Lambda = \left(\Sigma_{n\rho_n}^0\right)^{-1/2} \; \Pi \; BDB^\intercal \; \Pi \; \left(\Sigma_{n\rho_n}^0\right)^{-1/2}.$$

Thus, Λ is idempotent and has rank $d-1$ for each n where $n\rho_n > 0$. We define $\boldsymbol{W} \sim \mathrm{N}\left(\mathbf{0}, I_{d-1}\right)$. We then know that $\mathcal{Q}_{\rho_n}^{(n)}$ has the same asymptotic distribution as $\boldsymbol{W}^\intercal \boldsymbol{W}$ and so we can apply Theorem 7.1.2.

When $n\rho_n \to \infty$, we also have that $\mathbf{V}_{\rho_n}^{(n)} \xrightarrow{D} \mathrm{N}\left(\mathbf{0}, \Sigma^0\right)$ from Lemma 7.2.2. We then analyze the asymptotic distribution of the projected statistic, where we write $\mathbf{V} \sim \mathrm{N}\left(\mathbf{0}, \Sigma^0\right)$ and study the distribution of $\mathbf{V}^\intercal \; \Pi \; \left(\mathrm{Diag}(\boldsymbol{p}^0)\right)^{-1} \; \Pi \; \mathbf{V}$. We note that we have $\mathbf{V}^\intercal \mathbf{1} = 0$, which simplifies the asymptotic distribution of the projected statistic.

$$\mathbf{V}^\intercal \; \Pi \; \left(\mathrm{Diag}(\boldsymbol{p}^0)\right)^{-1} \; \Pi \; \mathbf{V} = \sum_{i=1}^d \frac{\mathrm{V}_i^2}{p_i^0}$$

Note that this last final form is exactly the original chi-square statistic used in the classical test, which is known to converge to χ^2_{d-1}. □

7.2.3. Comparison of Statistics

We now want to compare the two private chi-square statistics in (7.13) and (7.16) to see which may lead to a larger *power* (i.e. smaller Type II error). The following theorem shows that we can write the nonprojected statistic (7.13) as a combination of both the projected statistic (7.16) and squared independent Gaussian noise.

Theorem 7.2.7. *Consider histogram data* $\boldsymbol{H}$ *that has Gaussian noise* $\boldsymbol{Z} \sim \mathrm{N}\left(\mathbf{0}, 1/\rho \cdot I_d\right)$ *added to it. For the statistics* $\mathrm{Q}_\rho^{(n)}$ *and* $\mathcal{Q}_\rho^{(n)}$ *based on the noisy counts given in* (7.13) *and* (7.16) *respectively, we have*

$$\mathrm{Q}_\rho^{(n)} = \mathcal{Q}_\rho^{(n)} + \frac{\rho}{d}\left(\sum_{i=1}^{d} Z_i\right)^2 .$$

Further, for any fixed data $\boldsymbol{H}$*,* $\mathcal{Q}_\rho^{(n)}$ *is independent of* $\left(\sum_{i=1}^d Z_i\right)^2$*.*

To prove this we will use the noncentral version of Craig's Theorem.

Theorem 7.2.8 [Craig's Theorem (John G. Reid, 1988)]**.** *Let* $\boldsymbol{Y} \sim \mathrm{N}\left(\boldsymbol{\mu}, C\right)$*. Then the quadratic forms* $\boldsymbol{Y}^\intercal A \boldsymbol{Y}$ *and* $\boldsymbol{Y}^\intercal B \boldsymbol{Y}$ *are independent if* $ACB = 0$*.*

We are now ready to prove our theorem.

Proof of Theorem 7.2.7. We first show that we can write $\mathrm{Q}_\rho^{(n)} - \mathcal{Q}_\rho^{(n)} = \frac{\rho}{d}\left(\sum_{i=1}^d Z_i\right)^2$. Note that $\left(\mathbf{V}_\rho^{(n)}\right)^\intercal \mathbf{1} = \sum_{i=1}^d Z_i/\sqrt{n}$ and $\left(\Sigma_\rho^0\right)^{-1}$ has eigenvalue ρ with eigenvector $\mathbf{1}$. We then have

$$\begin{aligned}
\mathrm{Q}_\rho^{(n)} &= \left(\mathbf{V}_\rho^{(n)}\right)^\intercal \left(\Sigma_{n\rho}^0\right)^{-1} \mathbf{V}_\rho^{(n)} \\
&= \left(\mathbf{V}_\rho^{(n)}\right)^\intercal \left(I_d - \frac{1}{d}\mathbf{1}\mathbf{1}^\intercal + \frac{1}{d}\mathbf{1}\mathbf{1}^\intercal\right)^\intercal \left(\Sigma_{n\rho}^0\right)^{-1} \left(I_d - \frac{1}{d}\mathbf{1}\mathbf{1}^\intercal + \frac{1}{d}\mathbf{1}\mathbf{1}^\intercal\right) \mathbf{V}_\rho^{(n)} \\
&= \mathcal{Q}_\rho^{(n)} + \frac{2}{d}(\mathbf{V}_\rho^{(n)})^\intercal \mathbf{1}\mathbf{1}^\intercal \left(\Sigma_{n\rho}^0\right)^{-1} \left(I_d - \frac{1}{d}\mathbf{1}\mathbf{1}^\intercal\right) \mathbf{V}_\rho^{(n)} + \frac{1}{d^2}(\mathbf{V}_\rho^{(n)})^\intercal \mathbf{1}\mathbf{1}^\intercal \left(\Sigma_{n\rho}^0\right)^{-1} \mathbf{1}\mathbf{1}^\intercal \mathbf{V}_\rho^{(n)} \\
&= \mathcal{Q}_\rho^{(n)} + \frac{2}{d}(\mathbf{V}_\rho^{(n)})^\intercal \mathbf{1}\mathbf{1}^\intercal \left(\Sigma_{n\rho}^0\right)^{-1} \Pi \, \mathbf{V}_\rho^{(n)} + \frac{\rho}{d}\left(\sum_{i=1}^d Z_i\right)^2 \\
&= \mathcal{Q}_\rho^{(n)} + \frac{2n\rho}{d}\left(\sum_{i=1}^d Z_i/\sqrt{n}\right)\mathbf{1}^\intercal \mathbf{V}_\rho^{(n)} - \frac{2\rho}{d}\left(\sum_{i=1}^d Z_i\right)^2 + \frac{\rho}{d}\left(\sum_{i=1}^d Z_i\right)^2 \\
&= \mathcal{Q}_\rho^{(n)} + \frac{\rho}{d}\left(\sum_{i=1}^d Z_i\right)^2
\end{aligned}$$

We now apply Craig's Theorem to show that for a fixed histogram $\boldsymbol{H}$, we have $\mathcal{Q}_\rho^{(n)}$ is independent of $\left(\sum_{i=1}^d Z_i\right)^2$. When $\boldsymbol{H}$ is fixed, we can define the random variable $\boldsymbol{Y} \sim N\left(\boldsymbol{\mu}, 1/\rho I_d\right)$ where $\boldsymbol{\mu} = (\boldsymbol{H} - n\boldsymbol{p}^0)/\sqrt{n}$. If we set $A = \Pi \ \left(\Sigma_{n\rho}^0\right)^{-1} \Pi$, then our projected statistic can be rewritten as $\boldsymbol{Y}^\intercal A \boldsymbol{Y}$. Further, if we define $B = \mathbf{1}\mathbf{1}^\intercal$, then $\left(\sum_{i=1}^d Y_i\right)^2 = \boldsymbol{Y}^\intercal B \boldsymbol{Y}$. We then have $A\left(1/\rho \cdot I_d\right) B = 0$, so that the projected statistic is independent of $\left(\sum_{i=1}^d Y_i\right)^2$ by Theorem 7.2.8. We next note that $\boldsymbol{Y} = \boldsymbol{Z} + \boldsymbol{\mu}$ and that $\mathbf{1}^\intercal \mu = 0$. Hence,

$$\boldsymbol{Y}^\intercal B \boldsymbol{Y} = (\boldsymbol{Z} + \boldsymbol{\mu})^\intercal B (\boldsymbol{Z} + \boldsymbol{\mu}) = \boldsymbol{Z}^\intercal B \boldsymbol{Z} + 2\boldsymbol{\mu}^\intercal B \boldsymbol{Z} + \boldsymbol{\mu}^\intercal B \boldsymbol{\mu} = \boldsymbol{Z}^\intercal B \boldsymbol{Z} = \left(\sum_{i=1}^d Z_i\right)^2.$$

□

Algorithm 14 with procedure `NewStatAsymptGOF` shows how to perform goodness of fit testing with either of these two test statistics, i.e. nonprojected (7.13) or projected (7.16).

We note that our test is zCDP for neighboring histogram datasets due to it being an application of the Gaussian mechanism and Theorem 2.2.5. Hence:

Theorem 7.2.9. *`NewStatAsymptGOF`$(\cdot; \rho, \alpha, \boldsymbol{p}^0)$ is ρ-zCDP.*

Algorithm 14 New Private Statistic Goodness of Fit Test: `NewStatAsymptGOF`

```
Input: $\boldsymbol{h}$ $\rho$, $\alpha$, $\mathrm{H}_0 : \boldsymbol{p} = \boldsymbol{p}^0$
  Set $\widetilde{\boldsymbol{h}} \leftarrow \boldsymbol{h} + \boldsymbol{Z}$ where $\boldsymbol{Z} \sim \mathrm{N}(\mathbf{0}, 1/\rho \cdot I_d)$.
  For the nonprojected statistic:
  $\mathrm{T} \leftarrow \frac{1}{n}\left(\widetilde{\boldsymbol{h}} - n\boldsymbol{p}^0\right)^\intercal \left(\Sigma^0_{n\rho}\right)^{-1} \left(\widetilde{\boldsymbol{h}} - n\boldsymbol{p}^0\right)$
  $\tau \leftarrow (1-\alpha)$ quantile of $\chi^2_d$
  For the projected statistic:
  $\mathrm{T} \leftarrow \frac{1}{n}\left(\widetilde{\boldsymbol{h}} - n\boldsymbol{p}^0\right)^\intercal \Pi \left(\Sigma^0_{n\rho}\right)^{-1} \Pi \left(\widetilde{\boldsymbol{h}} - n\boldsymbol{p}^0\right)$
  $\tau \leftarrow (1-\alpha)$ quantile of $\chi^2_{d-1}$
  if $\mathrm{T} > \tau$ then
    Decision ← Reject
  else
    Decision ← Fail to Reject
Output: Decision
```

7.2.4. Power Analysis

From Theorem 7.2.7 we see that the difference between $\mathrm{Q}^{(n)}_\rho$ and $\mathcal{Q}^{(n)}_\rho$ is the addition of squared independent noise. This additional noise can only hurt *power*, because for the same data the statistic $\mathrm{Q}^{(n)}_\rho$ has larger variance than $\mathcal{Q}^{(n)}_\rho$ and does not depend on the underlying data. If we fix an alternate hypothesis – as we did in the previous chapter – we can obtain asymptotic distributions for our two test statistics.

Theorem 7.2.10. *Consider the null hypothesis $\mathrm{H}_0 : \boldsymbol{p} = \boldsymbol{p}^0$ and the alternate hypothesis $H_1 : \boldsymbol{p} = \boldsymbol{p}^0 + \frac{1}{\sqrt{n}}\boldsymbol{\Delta}$ where $\sum_{i=1}^d \Delta_i = 0$. Assuming the data $\boldsymbol{H}$ comes from the alternate H_1, the two statistics $\mathcal{Q}^{(n)}_{\rho_n}$, and $\mathrm{Q}^{(n)}_{\rho_n}$ have noncentral chi-square distributions when $n\rho_n \to$*

$\rho^\star > 0$, *i.e.*

$$\mathrm{Q}_{\rho_n}^{(n)} \xrightarrow{D} \chi_d^2\left(\boldsymbol{\Delta}^\intercal \left(\Sigma_{\rho^\star}^0\right)^{-1}\boldsymbol{\Delta}\right) \quad \& \quad \mathcal{Q}_{\rho_n}^{(n)} \xrightarrow{D} \chi_{d-1}^2\left(\boldsymbol{\Delta}^\intercal \left(\Sigma_{\rho^\star}^0\right)^{-1}\boldsymbol{\Delta}\right).$$

Further, if $n\rho_n \to \infty$ then

$$\mathcal{Q}_{\rho_n}^{(n)} \xrightarrow{D} \chi_{d-1}^2\left(\sum_i \frac{\Delta_i^2}{p_i^0}\right)$$

We point out that in the case where $n\rho_n \to \infty$, the projected statistic has the same asymptotic distribution as the classical (nonprivate) chi-square test under the same alternate hypothesis, given in Lemma 6.1.8.

We will use the following result to prove this theorem.

Lemma 7.2.11 [Ferguson (1996)]**.** *Suppose $\boldsymbol{W} \sim \mathrm{N}(\boldsymbol{\mu}, C)$. If C is a projection of rank ν and $C\boldsymbol{\mu} = \boldsymbol{\mu}$ then $\boldsymbol{W}^\intercal\boldsymbol{W} \sim \chi_\nu^2(\boldsymbol{\mu}^\intercal\boldsymbol{\mu})$.*

Proof of Theorem 7.2.10. In this case we have the random vector $\mathbf{V}_{\rho_n}^{(n)}$ from (7.11) converging in distribution to $\mathrm{N}\left(\boldsymbol{\Delta}, \Sigma_{\rho^\star}^0\right)$ if $n\rho_n \to \rho^\star > 0$ or $\mathrm{N}\left(\boldsymbol{\Delta}, \Sigma^0\right)$ if $n\rho_n \to \infty$ by Lemma 7.2.2. We first consider the case when $n\rho_n \to \rho^\star > 0$. Consider $\mathbf{V} \sim \mathrm{N}\left(\boldsymbol{\Delta}, \Sigma_{\rho^\star}^0\right)$ and $\boldsymbol{W} = \left(\Sigma_{\rho^\star}^0\right)^{-1/2}\mathbf{V} \sim \mathrm{N}\left(\left(\Sigma_{\rho^\star}^0\right)^{-1/2}\boldsymbol{\Delta}, I_d\right)$. We then know that $\boldsymbol{W}^\intercal\boldsymbol{W}$ and the nonprojected statistic $\mathrm{Q}_{\rho_n}^{(n)}$ have the same asymptotic distribution. In order to use Lemma 7.2.11, we need to verify that $\left(\Sigma_\rho^0\right)^{-1/2}\Sigma_{\rho^\star}^0\left(\Sigma_{\rho^\star}^0\right)^{-1/2}\left(\left(\Sigma_\rho^0\right)^{-1/2}\boldsymbol{\Delta}\right) = \left(\Sigma_{\rho^\star}^0\right)^{-1/2}\boldsymbol{\Delta}$, which indeed holds.

We then consider the projected statistic $\mathcal{Q}_{\rho_n}^{(n)}$ where $n\rho_n \to \rho^\star > 0$. Similar to the proof of Theorem 7.2.6, we diagonalize $\Sigma_{\rho^\star}^0 = BDB^\intercal$ where B is an orthogonal matrix with one column being $1/d \cdot \mathbf{1}$ and D is a diagonal matrix. We then let

$$\boldsymbol{Y} = \left(\Sigma_{\rho^\star}^0\right)^{-1/2}\Pi\,\mathbf{V}$$

We then know that $\boldsymbol{Y}^\intercal\boldsymbol{Y}$ and $\mathcal{Q}_{\rho_n}^{(n)}$ will have the same asymptotic distribution. Recall that $\Lambda = \left(\Sigma_{\rho^\star}^0\right)^{-1/2}\ \Pi\ \Sigma_{\rho^\star}^0\ \Pi\ \left(\Sigma_{\rho^\star}^0\right)^{-1/2}$ is idempotent with rank $d-1$. Lastly, to apply Lemma 7.2.11 we need to show the following

$$\Lambda\ \left(\left(\Sigma_{\rho^\star}^0\right)^{-1/2}\ \Pi\,\boldsymbol{\Delta}\right) = \left(\Sigma_{\rho^\star}^0\right)^{-1/2}\ \Pi\,\boldsymbol{\Delta}.$$

Let $\widehat{B} \in \mathbb{R}^{d\times(d-1)}$ be the same as matrix B whose corresponding column for $1/d{\cdot}\mathbf{1}$ is missing, which we assume to be the last column of B. Further, we define $\widehat{D} \in \mathbb{R}^{(d-1)\times(d-1)}$ to be the same as D without the last row and column. We can then write $\Pi\ \Sigma_{\rho^\star}^0\ \Pi = \widehat{B}\widehat{D}\widehat{B}^\intercal$ to simplify $\Lambda\left(\Sigma^0_{\rho^\star}\right)^{-1/2}\ \Pi$, i.e.

$$\begin{aligned}
\left(\Sigma^0_{\rho^\star}\right)^{-1/2}\ &\Pi\ \Sigma^0_{\rho^\star}\ \Pi\ \left(\Sigma^0_{\rho^\star}\right)^{-1}\ \Pi \\
&= BD^{-1/2}B^\intercal\ \Pi\ \Sigma^0_{\rho^\star}\widehat{B}\widehat{D}^{-1}\widehat{B}^\intercal \\
&= BD^{-1/2}B^\intercal\widehat{B}DB^\intercal\widehat{B}\widehat{D}^{-1}\widehat{B}^\intercal \\
&= BD^{-1/2}B^\intercal\widehat{B}D\widehat{D}^{-1}\widehat{B}^\intercal \\
&= BD^{-1/2}\widehat{B}^\intercal \\
&= BD^{-1/2}B^\intercal\Pi \\
&= \left(\Sigma^0_{\rho^\star}\right)^{-1/2}\ \Pi
\end{aligned}$$

The noncentral parameter is then

$$\mathbf{\Delta}^\intercal\Pi\left(\Sigma^0_{\rho^\star}\right)^{-1}\Pi\mathbf{\Delta}$$

We then note that $\sum_i \Delta_i = 0$.

For the case when $n\rho_n \to \infty$. From (7.18), we have $\Pi\ \Sigma^0_{n\rho_n}\ \Pi \to \Pi\left(\mathrm{Diag}(\boldsymbol{p}^0)\right)^{-1}\Pi$, which can be diagonalized. As we showed in Theorem 7.2.6, we have

$$\left(\mathbf{V}^{(n)}_{\rho_n}\right)^\intercal\Pi\left(\mathrm{Diag}(\boldsymbol{p}^0)\right)^{-1}\Pi\mathbf{V}^{(n)}_{\rho_n} = \left(\mathbf{V}^{(n)}_{\rho_n}\right)^\intercal\mathrm{Diag}(\boldsymbol{p}^0)^{-1}\mathbf{V}^{(n)}_{\rho_n}$$

From Lemma 7.2.2, we know that $\mathbf{V}^{(n)}_{\rho_n} \xrightarrow{D} \mathrm{N}\left(\mathbf{\Delta}, \Sigma^0\right)$. We then write $\boldsymbol{U} \sim \mathrm{N}\left(\mathbf{\Delta}, \Sigma^0\right)$ so that our projected chi-square statistic has the same asymptotic distribution as

$$\boldsymbol{U}^\intercal\left(\mathrm{Diag}(\boldsymbol{p}^0)\right)^{-1}\boldsymbol{U}$$

which has a $\chi^2_{d-1}(\mathbf{\Delta}^\intercal\left(\mathrm{Diag}(\boldsymbol{p}^0)\right)^{-1}\mathbf{\Delta})$ distribution. □

Note that the noncentral parameters in the previous theorem are the same for both statistics and only the degrees of freedom are different when $n\rho_n \to \rho^\star > 0$.

We now compare the variance of the projected and nonprojected statistics with the classical statistic $\mathrm{T}^{(n)}\left(\mathrm{N}\left(0, 1/\rho\right)\right)$ in (6.4) under the alternate hypothesis.

Theorem 7.2.12. *Let* $n\rho_n \to \rho^\star > 0$ *and let* $\mathrm{H}_0 : \boldsymbol{p} = \boldsymbol{p}^0$ *but the data is actually drawn from* $H_1 : \boldsymbol{p} = \boldsymbol{p}^0 + 1/\sqrt{n}\cdot\mathbf{\Delta}$. *We then have the following as* $n \to \infty$

$$\begin{aligned}
&\mathbb{V}\left[\mathrm{T}^{(n)}\left(N\left(0, 1/\rho_n\right)\right)\right] \\
&\qquad \to 2\cdot\left(d - 1 + \frac{2}{\rho^\star}\left(\sum_{i=1}^d \frac{1}{p^0_i} - d\right) + \frac{1}{(\rho^\star)^2}\left(\sum_{i=1}^d \frac{1}{(p^0_i)^2}\right) + 2\cdot\left(\sum_{i=1}^d \frac{\Delta_i^2}{p_i^0} + \frac{1}{\rho^\star}\sum_{i=1}^d \frac{\Delta_i^2}{(p_i^0)^2}\right)\right)
\end{aligned}$$

$$\mathbb{V}\left[\mathrm{Q}^{(n)}_{\rho_n}\right] \to 2\left(d + 2\cdot\mathbf{\Delta}^\intercal\left(\Sigma^0_{\rho^\star}\right)^{-1}\mathbf{\Delta}\right), \qquad \mathbb{V}\left[\mathcal{Q}^{(n)}_{\rho_n}\right] \to 2(d - 1 + 2\cdot\mathbf{\Delta}^\intercal\left(\Sigma^0_{\rho^\star}\right)^{-1}\mathbf{\Delta})$$

Further if $n\rho_n \to \infty$, *then*

$$\mathbb{V}\left[\mathrm{T}^{(n)}\left(N\left(0, 1/\rho_n\right)\right)\right] \to 2\left(d - 1 + 2\sum_{i=1}^{d} \frac{\Delta_i^2}{p_i^0}\right), \qquad \mathbb{V}\left[\mathcal{Q}_{\rho_n}^{(n)}\right] \to 2\left(d - 1 + 2\sum_{i=1}^{d} \frac{\Delta_i^2}{p_i^0}\right)$$

We prove this result by using the following.

Lemma 7.2.13 [Petersen and Pedersen (2012)]**.** *Let* $\boldsymbol{W} \sim \mathrm{N}\left(\boldsymbol{m}, C\right)$*, then*

$$\mathbb{V}\left[\boldsymbol{W}^{\intercal} A \boldsymbol{W}\right] = \mathtt{Trace}\left(AC(A + A^{\intercal})C\right) + \boldsymbol{m}^{\intercal}\left((A + A^{\intercal})C(A + A^{\intercal})\right)\boldsymbol{m}.$$

We are now ready to prove our result about the variance of our statistics.

Proof of Theorem 7.2.12. For the projected and nonprojected statistics, we can obtain the variance by using Theorem 7.2.10. We then focus on the statistic $\mathrm{T}^{(n)}\left(\rho_n\right) \overset{\text{defn}}{=} \mathrm{T}^{(n)}\left(N\left(0, 1/\rho_n\right)\right)$. We first point out that when $n\rho_n \to \infty$, the noise in the statistic becomes insignificant and we end up with a noncentral chi-square statistic. Hence, we consider $n\rho_n \to \rho^\star$. We can then directly apply Lemma 7.2.13, where $\mathrm{T}^{(n)}\left(\rho_n\right)$ has the same asymptotic distribution as $\boldsymbol{W}^{\intercal}\boldsymbol{\Lambda}_{\rho^\star}\boldsymbol{W}$ where $\boldsymbol{\Lambda}_{\rho^\star}$ is given in (6.7) and $\boldsymbol{W} \sim \mathrm{N}\left(\boldsymbol{\mu}', \boldsymbol{\Sigma}\right)$ for $\boldsymbol{\mu'}$ given in Corollary 6.1.10 and $\boldsymbol{\Sigma}$ given in (6.6). ☐

With this result, we can compare which statistic has higher variance which we would expect would lead to worse power. Note that the projected statistic always has smaller asymptotic variance than the nonprojected statistic. As an example, consider $\boldsymbol{p}^0 = (1/d, \cdots, 1/d)^{\intercal}$ and $n\rho_n \to \rho^\star > 0$. We can then write the inverse of the covariance in this case,

$$\left(\Sigma_{\rho^\star}^0\right)^{-1} = \frac{\rho^\star}{\rho^\star + d}\left(d \cdot I_d + \frac{(\rho^\star)^2}{\rho^\star + d} \cdot \mathbf{1}\mathbf{1}^{\intercal}\right). \tag{7.20}$$

The projected statistic then has asymptotic variance

$$2 \cdot (d - 1) + 4d \cdot \left(\frac{\rho^\star}{\rho^\star + d}\right)\boldsymbol{\Delta}^{\intercal}\boldsymbol{\Delta}.$$

Now the variance of the original chi-square statistic which uses the noisy histogram from Chapter 6 is then,

$$2(d - 1)\left(1 + \frac{2d}{\rho^\star}\right) + \frac{d^3}{(\rho^\star)^2} + 4d \cdot \left(\frac{\rho^\star + d}{\rho^\star}\right)\boldsymbol{\Delta}^{\intercal}\boldsymbol{\Delta}.$$

Comparing the variance of these two statistics shows that we would expect the projected statistic to have better power.

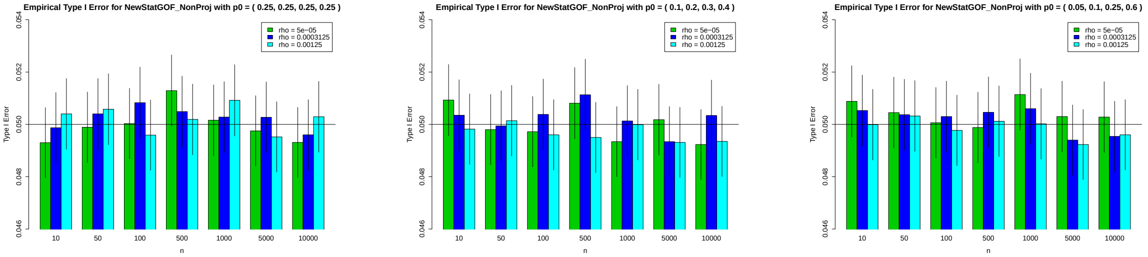


Figure 15: Empirical Type I Error for our goodness of fit tests in `NewStatAsymptGOF` with the nonprojected statistic $\mathrm{Q}_{\rho}^{(n)}$.

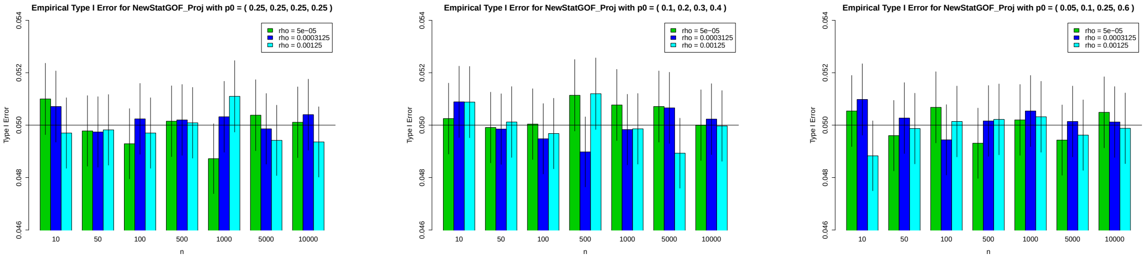


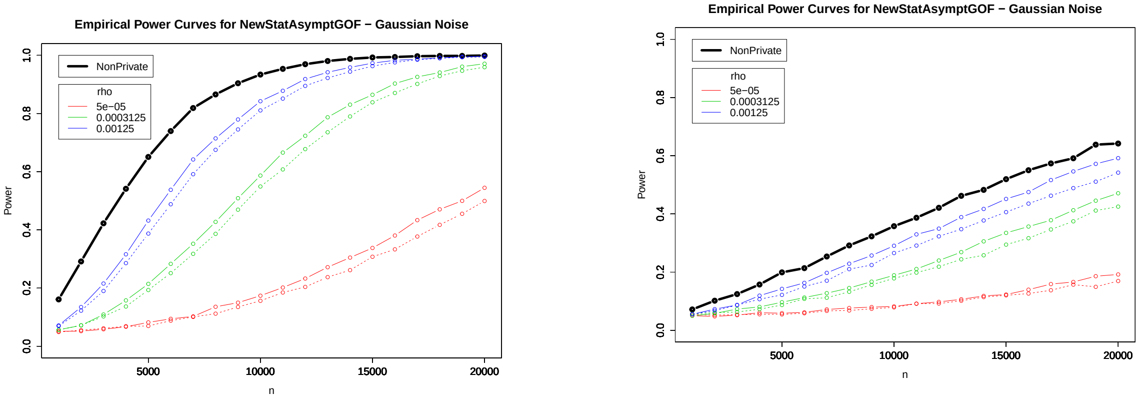
Figure 16: Empirical Type I Error for our goodness of fit tests in `NewStatAsymptGOF` with the projected statistic $\mathcal{Q}_{\rho}^{(n)}$.

7.2.5. Experiments for Goodness of Fit Testing

As we did in Chapter 6, we will fix $\alpha = 0.05$ throughout all of our experiments. All of our tests are designed to achieve Type I error at most α as we empirically show for different null hypotheses $\boldsymbol{p}^0$, privacy parameters ρ, and sample size n in Figure 15 and Figure 16 for the nonprojected and projected statistic, respectively. We include 1.96 times the standard error of our $100,000$ independent trials (giving a 95% confidence interval) as well as the target $\alpha = 0.05$ Type I Error as a horizontal line.

We then empirically check the power of our new tests in `NewStatAsymptGOF` for both the projected and nonprojected statistic. Subject to the constraint that our tests achieve Type I error at most α, we seek to maximize *power*, or the probability of rejecting the null hypothesis when a distribution $\boldsymbol{p}^1 \neq \boldsymbol{p}^0$, called the *alternate hypothesis*, is true. We expect to see the projected statistic achieve higher power than the nonprojected statistic due to Theorem 7.2.7. Further, the fact that the critical value we use for the projected statistic is smaller than the critical value for the nonprojected statistic might lead to the projected statistic having higher power.

For demonstrating the power of our tests, we want to find "bad" alternate hypotheses, which would be hard for our test to reject. The way in which we choose the alternate then is by finding vectors $\boldsymbol{\Delta}$ that make the variance of the statistics that we computed in Theorem 7.2.12 as large as possible. Intuitively, the larger the variance the harder it will be for our test to distinguish between fluctuations in the data due to sampling or due to the



(a) $\mathrm{H}_0 : \boldsymbol{p} = \boldsymbol{p}^0 = (1/4, 1/4, 1/4, 1/4)^\intercal$ and $H_1 : \boldsymbol{p}^0 + \boldsymbol{\Delta}$ where $\boldsymbol{\Delta} = 0.01 \cdot (1, -1, -1, 1)^\intercal$.

(b) $\mathrm{H}_0 : \boldsymbol{p} = \boldsymbol{p}^0 = (1/2, 1/6, 1/6, 1/6)^\intercal$ and $H_1 : \boldsymbol{p}^0 + \boldsymbol{\Delta}$ where $\boldsymbol{\Delta} = 0.01 \cdot (1, -1/3, -1/3, -1/3)^\intercal$.

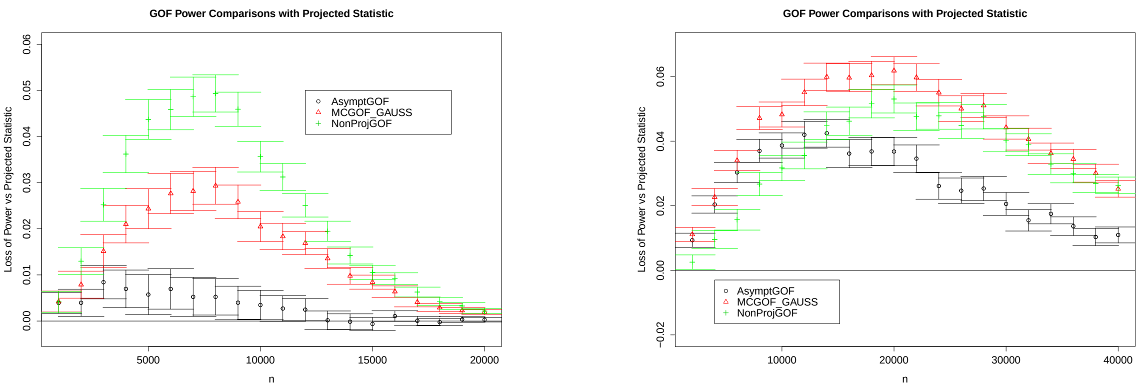
Figure 17: Comparison of empirical power of classical non-private test versus `NewStatAsymptGOF` with both projected (solid line) and nonprojected statistics (dashed line).

null hypothesis being incorrect. We ultimately want the distribution of the test statistic under the null and alternate hypotheses to be far apart.

We then set a null hypothesis $\boldsymbol{p}^0$ and then find the eigenvector $\boldsymbol{\Delta}$ of $\Sigma^0_{n\rho}$ with the largest eigenvalue. Ultimately, for a fixed $\rho > 0$ we want to find a probability vector $\boldsymbol{p}^0$ so that the max eigenvalue of $\left(\Sigma^0_{n\rho}\right)^{-1} = \left(\mathrm{Diag}(\boldsymbol{p}^0) - \boldsymbol{p}^0(\boldsymbol{p}^0)^\intercal + \frac{1}{n\rho}\, I_d\right)^{-1}$ is as large as possible.

As we did in our power experiments in Section 6.4, we set the null hypothesis $\boldsymbol{p}^0$ and alternate hypothesis $\boldsymbol{p}^1 = \boldsymbol{p}^0 + \boldsymbol{\Delta}$ for various sample sizes. For each sample size n, we sample $10,000$ independent datasets from the alternate hypothesis and test $\mathrm{H}_0 : \boldsymbol{p} = \boldsymbol{p}^0$ in `NewStatAsymptGOF`. We present the resulting power plots in Figure 17 for `NewStatAsymptGOF` from Algorithm 14.

We then compare the projected and nonprojected statistic in `NewStatAsymptGOF` to the classical statistic used in Chapter 6 for Type I Error level $\alpha = 0.05$ and $\rho = 0.00125$. Since the projected statistic outperforms the other tests, we plot the difference in power between the projected statistic and the other tests for $\boldsymbol{p}^0 = (1/2, 1/6, 1/6, 1/6)^\intercal$ but the data is drawn from $\boldsymbol{p}^1 = \boldsymbol{p}^0 + \boldsymbol{\Delta}$ where $\boldsymbol{\Delta} = 0.01 \cdot (1, -1/3, -1/3, -1/3)^\intercal$. Note that the error bars show 1.96 times the standard error in the difference of proportions from $100,000$ trials, giving a 95% confidence interval. We give the corresponding plot in Figure 18. Recall that `MCGOF` with Gaussian noise is the test in Algorithm 8, `AsymptGOF` is the test in Algorithm 9, and `NewStatAsymptGOF` with the non-projected statistic is labeled as "NonProjGOF" in the plot.



(a) $\boldsymbol{p}^0 = (1/4, 1/4, 1/4, 1/4)^\intercal$ and $H_1 : \boldsymbol{p}^0 + \boldsymbol{\Delta}$ where $\boldsymbol{\Delta} = 0.01 \cdot (1, -1, -1, 1)^\intercal$.

(b) $\boldsymbol{p}^0 = (1/2, 1/6, 1/6, 1/6)^\intercal$ and $H_1 : \boldsymbol{p}^0 + \boldsymbol{\Delta}$ where $\boldsymbol{\Delta} = 0.01 \cdot (1, -1/3, -1/3, -1/3)^\intercal$.

Figure 18: Comparison of empirical power between all zCDP hypothesis tests for goodness of fit and `NewStatAsymptGOF` with projected statistic.

7.3. General Chi-Square Private Tests

We now consider the case where the null hypothesis contains a subset of data distributions, so that the best fitting distribution must be estimated and used in the test statistics. The data is multinomial $\boldsymbol{H} \sim \text{Multinomial}(n, \boldsymbol{p}(\boldsymbol{\theta}^0))$ and $\boldsymbol{p}$ is a function that converts parameters into a s-dimensional multinomial probability vector. The null hypothesis is $\mathrm{H}_0 : \boldsymbol{\theta}^0 \in \Theta$; i.e. $\boldsymbol{p}(\boldsymbol{\theta}^0)$ belongs to a subset of a lower-dimensional manifold. We again use Gaussian noise $\boldsymbol{Z} \sim \mathrm{N}\left(\mathbf{0}, 1/\rho \cdot I_d\right)$ to ensure ρ-zCDP, and we define

$$\mathbf{V}_\rho^{(n)}(\boldsymbol{\theta}) \overset{\text{defn}}{=} \sqrt{n}\left(\frac{\boldsymbol{H}+\boldsymbol{Z}}{n} - \boldsymbol{p}\left(\boldsymbol{\theta}\right)\right). \tag{7.21}$$

With $\boldsymbol{\theta}^0$ being the unknown true parameter, we are now ready to define our two test statistics in terms of some function $\phi : \mathbb{R}^d \to \mathbb{R}$, such that $\phi(\boldsymbol{H}+\boldsymbol{Z}) \xrightarrow{P} \boldsymbol{\theta}^0$ (recall from Section 7.1.2 that ϕ is a simple but possibly a suboptimal estimate of the true parameter $\boldsymbol{\theta}^0$ based on the noisy data) and the covariance matrix

$$\Sigma_\rho^0(\boldsymbol{\theta}) \overset{\text{defn}}{=} \mathrm{Diag}\left(\boldsymbol{p}(\boldsymbol{\theta})\right) - \boldsymbol{p}(\boldsymbol{\theta})\boldsymbol{p}(\boldsymbol{\theta})^\intercal + 1/\rho \cdot I_d.$$

We define the *nonprojected* statistic $\mathrm{R}_\rho^{(n)}(\boldsymbol{\theta})$ as follows:

$$\begin{aligned} \widehat{M}_{n\rho} &\overset{\text{defn}}{=} \left(\Sigma_{n\rho}^0\left(\phi(\boldsymbol{H}+\boldsymbol{Z})\right)\right)^{-1} \\ \mathrm{R}_\rho^{(n)}(\boldsymbol{\theta}) &\overset{\text{defn}}{=} \mathbf{V}_\rho^{(n)}(\boldsymbol{\theta})^\intercal \widehat{M}_{n\rho} \mathbf{V}_\rho^{(n)}(\boldsymbol{\theta}). \end{aligned} \tag{7.22}$$

This is a specialization of (7.3) in Section 7.1.2 with the following substitutions: $\mathbf{V}^{(n)} = \frac{\boldsymbol{H}+\boldsymbol{Z}}{n}$, $A(\boldsymbol{\theta}) = \boldsymbol{p}(\boldsymbol{\theta})$, and $M(\boldsymbol{\theta}) = \left(\Sigma_{n\rho}^0(\boldsymbol{\theta})\right)^{-1}$.

For the *projected* statistic $\mathcal{R}_\rho^{(n)}(\theta)$, the corresponding substitutions are

$\Pi = I_d - \frac{1}{d}\mathbf{1}\mathbf{1}^\intercal$, $\mathbf{V}^{(n)} = \Pi \cdot \left(\frac{\boldsymbol{H}+\boldsymbol{Z}}{n}\right)$, $A(\boldsymbol{\theta}) = \Pi \cdot \boldsymbol{p}(\boldsymbol{\theta})$, and again $M(\boldsymbol{\theta}) = \left(\Sigma_{n\rho}^0(\boldsymbol{\theta})\right)^{-1}$ giving:

$$\mathcal{R}_\rho^{(n)}(\boldsymbol{\theta}) \overset{\text{defn}}{=} \mathbf{V}_\rho^{(n)}(\boldsymbol{\theta})^\intercal \; \Pi \; \widehat{M}_{n\rho} \; \Pi \; \mathbf{V}_\rho^{(n)}(\boldsymbol{\theta}). \tag{7.23}$$

We then assume that for both the projected and nonprojected statistic Assumption 7.1.1 holds using their relative vectors $\mathbf{V}^{(n)}$, $A(\boldsymbol{\theta})$, and matrix $M(\boldsymbol{\theta})$. We now present the asymptotic distribution of both statistics, which is proved using the result in Theorem 7.1.4.

Theorem 7.3.1. *Under* $\mathrm{H}_0 : \boldsymbol{\theta}^0 \in \Theta$, *the following are true as* $n \to \infty$. *Setting* $\widehat{\boldsymbol{\theta}}^{(n)} = \arg\min_{\boldsymbol{\theta}\in\Theta} \mathrm{R}_{\rho_n}^{(n)}(\boldsymbol{\theta})$ *we have* $\mathrm{R}_{\rho_n}^{(n)}(\widehat{\boldsymbol{\theta}}^{(n)}) \xrightarrow{D} \chi^2_{d-s}$ *if* $n\rho_n \to \rho^\star > 0$. *Furthermore, setting* $\widehat{\boldsymbol{\theta}}^{(n)} = \arg\min_{\theta\in\Theta} \mathcal{R}_{\rho_n}^{(n)}(\theta)$ *we have* $\mathcal{R}_{\rho_n}^{(n)}(\widehat{\boldsymbol{\theta}}^{(n)}) \xrightarrow{D} \chi^2_{d-s-1}$ *if* $\rho_n = \Omega(1/n)$.

Proof. To prove this result, we appeal to Theorem 7.1.4. For the nonprojected statistic $\mathrm{R}_{\rho_n}^{(n)}(\cdot)$ we have that $C(\boldsymbol{\theta}) = \Sigma_{n\rho_n}^0(\boldsymbol{\theta})$ and the middle matrix $M(\boldsymbol{\theta})$ is simply the inverse of it, which satisfies the hypotheses of Theorem 7.1.4.

For the projected statistic $\mathcal{R}_{n\rho_n}^{(n)}(\cdot)$, we will write $C(\boldsymbol{\theta}) = \Pi\, \Sigma_{n\rho_n}^0(\boldsymbol{\theta})\, \Pi$, $M(\boldsymbol{\theta}) = \left(\Sigma_{n\rho_n}^0\right)^{-1}(\boldsymbol{\theta})$, and $\dot{A}(\boldsymbol{\theta}) = \Pi \cdot \nabla \boldsymbol{p}(\boldsymbol{\theta}) \in \mathbb{R}^{d\times m}$. Note that $C(\boldsymbol{\theta})$ has rank $d-1$ for all $\boldsymbol{\theta} \in \Theta$ in a neighborhood of $\boldsymbol{\theta}^0$ and all n. We will now show that we can satisfy the hypotheses in Theorem 7.1.4 with these matrices, i.e. we show the following two equalities hold for all $\boldsymbol{\theta} \in \Theta$

$$C(\boldsymbol{\theta}) \; M(\boldsymbol{\theta}) \; C(\boldsymbol{\theta}) = C(\boldsymbol{\theta}) \qquad \& \qquad C(\boldsymbol{\theta}) \; M(\boldsymbol{\theta}) \; \dot{A}(\boldsymbol{\theta}) = \dot{A}(\boldsymbol{\theta}).$$

We first focus on proving the first equality $C(\boldsymbol{\theta})\; M(\boldsymbol{\theta})\; C(\boldsymbol{\theta}) = C(\boldsymbol{\theta})$. From (7.17), we can simplify the left hand side of the equality significantly by rewriting it as

$$\Pi \; \Sigma_{n\rho_n}^0(\boldsymbol{\theta}) \; \Pi - \frac{n\rho_n}{d} \cdot \Pi \; \Sigma_{n\rho_n}^0(\boldsymbol{\theta}) \; \mathbf{1}\mathbf{1}^\intercal \; \Sigma_{n\rho_n}^0(\boldsymbol{\theta}) \; \Pi$$

We now show that $\Pi \; \Sigma_{n\rho_n}^0(\boldsymbol{\theta}) \; \mathbf{1}\mathbf{1}^\intercal \; \Sigma_{n\rho_n}^0(\boldsymbol{\theta}) = 0$ for all n, which would prove this equality. Note that $\Sigma_{n\rho_n}^0(\boldsymbol{\theta})$ is symmetric and has eigenvector $\mathbf{1}$ with eigenvalue $\frac{1}{n\rho_n}$. Thus,

$$\Pi \; \Sigma_{n\rho_n}^0(\boldsymbol{\theta}) \; \mathbf{1}\mathbf{1}^\intercal \; \Sigma_{n\rho_n}^0(\boldsymbol{\theta}) = \frac{1}{n^2\rho_n^2} \cdot \Pi \; \mathbf{1}\mathbf{1}^\intercal = 0 \quad \forall n.$$

We now prove the second equality $C(\boldsymbol{\theta}) \cdot M(\boldsymbol{\theta}) \cdot \dot{A}(\boldsymbol{\theta}) = \dot{A}(\boldsymbol{\theta})$. We again use (7.17) to simplify the left hand side of the equality:

$$\begin{aligned}
&\Pi \; \Sigma_{n\rho_n}^0(\boldsymbol{\theta}) \left[\left(\Sigma_{n\rho_n}^0(\boldsymbol{\theta})\right)^{-1} - \frac{n\rho_n}{d} \cdot \mathbf{1}\mathbf{1}^\intercal\right] \nabla \boldsymbol{p}(\boldsymbol{\theta}) \\
&= \Pi \left[I_d - \frac{n\rho_n}{d} \cdot \Sigma_{n\rho_n}^0(\boldsymbol{\theta}) \; \mathbf{1}\mathbf{1}^\intercal\right] \nabla \boldsymbol{p}(\theta) \\
&= \Pi \; \Pi \nabla \boldsymbol{p}(\boldsymbol{\theta}) \\
&= \Pi \; \nabla \boldsymbol{p}(\boldsymbol{\theta}).
\end{aligned}$$

This completes the proof for $\rho_n = \Omega(1/n)$. $\square$

Again, the projected statistic has the same distribution under both private asymptotic regimes and matches the non-private chi-square test asymptotics. We present our more general test `GenChiTest` in Algorithm 15. The quick-and-dirty estimator $\phi(\cdot)$ is application-specific (Section 7.3.1 gives independence testing as an example).[5] Further, for neighboring histogram data, we have the following privacy guarantee.

Theorem 7.3.2. *GenChiTest*$(\cdot; \rho, \alpha, \phi, \Theta)$ *is* ρ*-zCDP.*

Algorithm 15 Private General Chi-Square Test: `GenChiTest`

```
Input: $\boldsymbol{h}$; $\rho$, $\alpha$, $\phi$, $\mathrm{H}_0 : \boldsymbol{\theta}^0 \in \Theta$
  Set $\widetilde{\boldsymbol{h}} \leftarrow \boldsymbol{h} + \boldsymbol{Z}$ where $\boldsymbol{Z} \sim \mathrm{N}\left(\boldsymbol{0}, 1/\rho \cdot I_d\right)$.
  Set $\widehat{M} = \left(\Sigma^0_{n\rho}\left(\phi(\widetilde{\boldsymbol{h}})\right)\right)^{-1}$
  For the nonprojected statistic:
      $$\mathrm{T}(\boldsymbol{\theta}) = \frac{1}{n}\left(\widetilde{\boldsymbol{h}} - n\boldsymbol{p}(\boldsymbol{\theta})\right)^{\intercal} \widehat{M}_{n\rho} \left(\widetilde{\boldsymbol{h}} - n\boldsymbol{p}(\boldsymbol{\theta})\right)$$
  Set $\widehat{\boldsymbol{\theta}}^{(n)} = \mathrm{argmin}_{\boldsymbol{\theta}\in\Theta}\,\mathrm{T}(\boldsymbol{\theta})$, $t \leftarrow (1-\alpha)$ quantile of $\chi^2_{d-m}$
  For the projected statistic:
      $$\mathrm{T}(\boldsymbol{\theta}) = \frac{1}{n}\left(\widetilde{\boldsymbol{h}} - n\boldsymbol{p}(\boldsymbol{\theta})\right)^{\intercal} \;\Pi\; \widehat{M}_{n\rho} \;\Pi\; \left(\widetilde{\boldsymbol{h}} - n\boldsymbol{p}(\boldsymbol{\theta})\right)$$
  Set $\widehat{\boldsymbol{\theta}}^{(n)} = \mathrm{argmin}_{\boldsymbol{\theta}\in\Theta}\,\mathrm{T}(\boldsymbol{\theta})$, $t \leftarrow (1-\alpha)$ quantile of $\chi^2_{d-m-1}$
  if $\mathrm{T}(\widehat{\boldsymbol{\theta}}^{(n)}) > t$ then
    Decision ← Reject.
  else
    Decision ← Fail to Reject.
Output: Decision.
```

7.3.1. Application - Independence Test

We showcase our general chi-square test `GenChiTest` by giving results for independence testing. Conceptually, it is convenient to think of the data histogram as an $r \times c$ table, with $p_{i,j}$ being the probability a person is in the bucket in row i and column j. We then consider two multinomial random variables $\boldsymbol{Y}^{(1)} \sim \mathrm{Multinomial}(1, \boldsymbol{\pi}^{(1)})$ for $\boldsymbol{\pi}^{(1)} \in \mathbb{R}^r$ (the marginal row probability vector) and $\boldsymbol{Y}^{(2)} \sim \mathrm{Multinomial}(1, \boldsymbol{\pi}^{(2)})$ for $\boldsymbol{\pi}^{(2)} \in \mathbb{R}^c$ (the marginal column probability vector). Under the null hypothesis of independence between $\boldsymbol{Y}^{(1)}$ and $\boldsymbol{Y}^{(2)}$, $p_{i,j} = \pi^{(1)}_i \pi^{(2)}_j$. Generally, we write the probabilities as $\boldsymbol{p}(\boldsymbol{\pi}^{(1)}, \boldsymbol{\pi}^{(2)}) = \boldsymbol{\pi}^{(1)} \left(\boldsymbol{\pi}^{(2)}\right)^{\intercal}$ so that

$$\boldsymbol{H} \sim \mathrm{Multinomial}\left(n, \boldsymbol{p}(\boldsymbol{\pi}^{(1)}, \boldsymbol{\pi}^{(2)})\right).$$

---

[5] For goodness-of-fit testing, ϕ always returns $\boldsymbol{p}^0$ and $s = 0$ so `GenChiTest` is a generalization of `NewStatAsymptGOF`.

Thus we have the underlying parameter vector $\boldsymbol{\theta}^0 = \left(\pi_1^{(1)}, \cdots, \pi_{r-1}^{(1)}, \pi_1^{(2)}, \cdots, \pi_{c-1}^{(2)}\right)$ - we do not need the last component of $\boldsymbol{\pi}^{(1)}$ or $\boldsymbol{\pi}^{(2)}$ because we know that each must sum to 1. Also, we have $d = rc$ and $s = (r-1) + (c-1)$ in this case. We want to test whether $\boldsymbol{Y}^{(1)}$ is independent of $\boldsymbol{Y}^{(2)}$. For our data, we are given a collection of n independent trials of $\boldsymbol{Y}^{(1)}$ and $\boldsymbol{Y}^{(2)}$. We then count the number of joint outcomes in a contingency table given in Table 2. Each cell in the contingency table contains element $H_{i,j}$ that gives the number of occurrences of $Y_i^{(1)} = 1$ and $Y_j^{(2)} = 1$. Since our test statistics notationally treat the data as a vector, when needed, we convert $\boldsymbol{H}$ to a vector that goes from left to right along each row of the table.

In order to compute the statistic $\mathrm{R}_\rho^{(n)}\left(\widehat{\boldsymbol{\theta}}^{(n)}\right)$ or $\mathcal{R}_\rho^{(n)}\left(\widehat{\boldsymbol{\theta}}^{(n)}\right)$ in `GenChiTest`, we need to find a quick-and-dirty estimator $\phi(\boldsymbol{H}+\boldsymbol{Z})$ that converges in probability to $\boldsymbol{p}\left(\boldsymbol{\pi}^{(1)}, \boldsymbol{\pi}^{(2)}\right)$ as $n \to \infty$. We will use the estimator for the unknown probability vector based on the marginals of the table with noisy counts, so that for naïve estimates $\widetilde{\pi}_i^{(1)} = \frac{H_{i,\cdot}+Z_{i,\cdot}}{\widetilde{n}}$, $\widetilde{\pi}_j^{(2)} = \frac{H_{\cdot,j}+Z_{\cdot,j}}{\widetilde{n}}$ where $\widetilde{n} = n + \sum_{i,j} Z_{i,j}$ we have,[6]

$$\phi\left(\boldsymbol{H}+\boldsymbol{Z}\right) = \left(\widetilde{\pi}_1^{(1)}, \cdots, \widetilde{\pi}_{r-1}^{(1)}, \widetilde{\pi}_1^{(2)}, \cdots, \widetilde{\pi}_{c-1}^{(2)}\right). \tag{7.24}$$

Note that as $n \to \infty$, the marginals converge in probability to the true probabilities even for $\boldsymbol{Z} \sim \mathrm{N}\left(\boldsymbol{0}, 1/\rho_n \cdot I_{rc}\right)$ with $\rho_n = \omega(1/n^2)$, i.e. we have that $\widetilde{\pi}_i^{(1)} \xrightarrow{P} \pi_i^{(1)}$ and $\widetilde{\pi}_j^{(2)} \xrightarrow{P} \pi_j^{(2)}$ for all $i \in [r]$ and $j \in [c]$. Recall that in Theorem 7.3.1, in order to guarantee the correct asymptotic distribution we require the $n\rho_n \to \rho^\star > 0$, or in the case of the projected statistic, we need $\rho_n = \Omega(1/n)$. Thus, Theorem 7.3.1 imposes more restrictive settings of ρ_n for the nonprojected statistic than what we need in order for the naïve estimate to converge to the true underlying probability. For the projected statistic, we only need $\rho_n = \Omega(1/n)$ to satisfy the conditions in Theorem 7.3.1 and for $\phi(\boldsymbol{H}+\boldsymbol{Z}) \xrightarrow{P} \boldsymbol{p}\left(\boldsymbol{\pi}^{(1)}, \boldsymbol{\pi}^{(2)}\right)$.

We then use this statistic $\phi(\boldsymbol{H}+\boldsymbol{Z})$ in our nonprojected and projected statistic in `GenChiTest` to have a ρ-zCDP hypothesis test for independence between two categorical variables. Note that in this setting, the projected statistic has a $\chi^2_{(r-1)(c-1)}$ distribution, which is exactly the same asymptotic distribution used in the classical Pearson chi-square independence test.

For our results we will again fix $\alpha = 0.05$, which we give as a horizontal line in our plots. For our data distributions, we will again consider 2×2 contingency tables where $\boldsymbol{\pi}^{(1)} = (\pi^{(1)}, 1-\pi^{(1)})$ and $\boldsymbol{\pi}^{(2)} = (\pi^{(2)}, 1-\pi^{(2)})$. In Figure 19 and Figure 20 we give the empirical Type I error for our independence tests given in `GenChiTest` for both the nonprojected and projected statistics for various n, data distributions, and zCDP parameter ρ. We also include error bars denoting 1.96 times the standard error over the $10,000$ trials. We note that for small sample sizes we are achieving much smaller Type I Errors than the target α due to the fact that sometimes the noise forces us to have small expected counts (< 5 in any cell) in the contingency table based on the noisy counts, in which case our tests are

---

[6] We note that in the case of small sample sizes, we follow a common rule of thumb where if any of the expected cell counts are less than 5, i.e. if $n\ \widetilde{\pi}_i^{(1)}\widetilde{\pi}_j^{(2)} \leq 5$ for any $(i,j) \in [r] \times [c]$, then we do not make any conclusion.

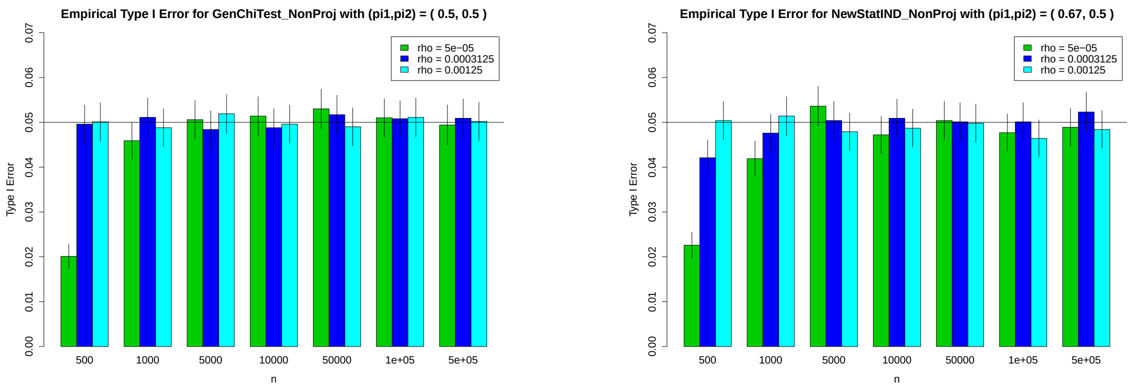


Figure 19: Empirical Type I Error for our new independence tests in `GenChiTest` with the nonprojected statistic.

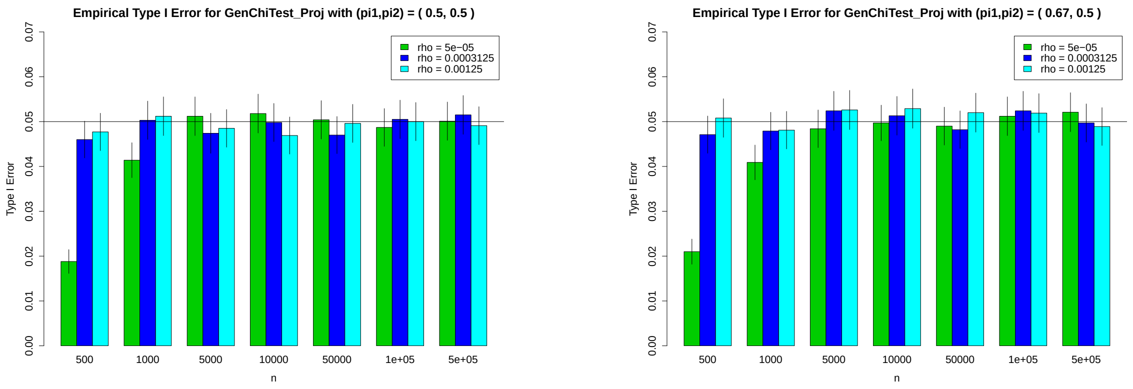
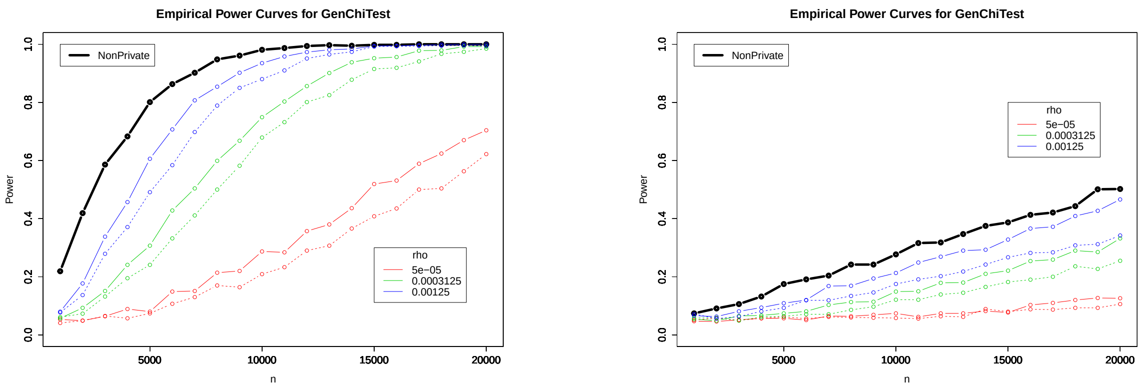


Figure 20: Empirical Type I Error for our new independence tests in `GenChiTest` with the projected statistic.

inconclusive.

We then compare the power that `GenChiTest` achieves for both of our test statistics. We then sample our contingency table $\boldsymbol{H}$ from Multinomial$(n, \boldsymbol{p}(\boldsymbol{\pi}^{(1)}, \boldsymbol{\pi}^{(2)}) + \boldsymbol{\Delta})$ for various $\boldsymbol{\Delta}$, so that the null hypothesis is indeed false and should be rejected. We give the empirical power in $1,000$ trials of `GenChiTest` in Figure 21 using both the nonprojected $\mathrm{R}_\rho^{(n)}\left(\widehat{\boldsymbol{\theta}}^{(n)}\right)$ from (7.22) and projected statistic $\mathcal{R}_\rho^{(n)}\left(\widehat{\boldsymbol{\theta}}^{(n)}\right)$ from (7.23). Note that again we pick $\widehat{\boldsymbol{\theta}}^{(n)}$ from Theorem 7.1.3 relative to the statistic we use. We label the test from `GenChiTest` with the nonprojected statistic with a dashed line whereas the solid line uses the projected statistic.

As we did for our goodness of fit tests, we compare the projected and nonprojected statistic in `GenChiTest` to the classical Pearson chi-square statistic used for independence testing in Chapter 6 for Type I Error level $\alpha = 0.05$ and $\rho = 0.00125$ to ensure our hypothesis



(a) We set $\pi^{(1)} = \pi^{(2)} = 1/2$ and $\mathbf{\Delta} = 0.01 \cdot (1, -1)^\intercal (1, -1)$.

(b) We set $\pi^{(1)} = 2/3, \pi^{(2)} = 1/2$ and $\mathbf{\Delta} = 0.01 \cdot (1, -1)^\intercal (1, 0)$.

Figure 21: Comparison of empirical power of classical non-private test versus `GenChiTest` in $1,000$ trials. The solid line is with the projected statistic and the dashed line is with the nonprojected statistic.

tests for independence are 0.00125-zCDP. Once again, the projected statistic outperforms the other tests, so we plot the difference in power between the projected statistic and the other tests for $\boldsymbol{\pi}^{(1)} = (\pi^{(1)}, 1 - \pi^{(1)})$ and $\boldsymbol{\pi}^{(2)} = (\pi^{(2)}, 1 - \pi^{(2)})$ where $\pi^{(1)} = 2/3$ and $\pi^{(2)} = 1/2$ but the data is not independent and is then drawn from the following table of probabilities $\boldsymbol{\pi}^{(1)} \left(\boldsymbol{\pi}^{(2)}\right)^\intercal + 0.01 \cdot (1, -1)^\intercal (1, 0)$. Note that the error bars show 1.96 times the standard error in the difference of proportions from $10,000$ trials, giving a 95% confidence interval. We give the corresponding plot in Figure 22. Recall that `MCIndep` with Gaussian noise is the test in Algorithm 12, `AsymptIndep` is the test in Algorithm 13, and we label "NonProjIndep" as the test `GenChiTest` with the nonprojected statistic.

7.3.2. Application - GWAS Testing

We next turn to demonstrating that our new general class of private hypothesis tests for categorical data significantly improves over existing private hypothesis tests even when extra structure is assumed about the dataset. Specifically, we will be interested in GWAS data, which was the primary application for private hypothesis tests due to the attack from Homer et al. (2008). We will then assume that $r = 3$ and $c = 2$ and the data is evenly split between the two columns - as is the case in a trial with both a control and case group. For such tables, we can directly compute the sensitivity of the classical chi-square statistic:

$$\sum_{i=1}^{3} \sum_{j=1}^{2} \frac{n \cdot \left(H_{i,j} - \frac{H_{\cdot,j} \cdot H_{i,\cdot}}{n}\right)^2}{H_{\cdot,j} \cdot H_{i,\cdot}}$$

Lemma 7.3.3 [Uhler et al. (2013); Yu et al. (2014)]. *The ℓ_1 and ℓ_2 global sensitivity of the chi-square statistic based on a 3×2 contingency table with positive margins and $n/2$ cases and $n/2$ controls is $\Delta_\chi = 4n/(n+2)$.*

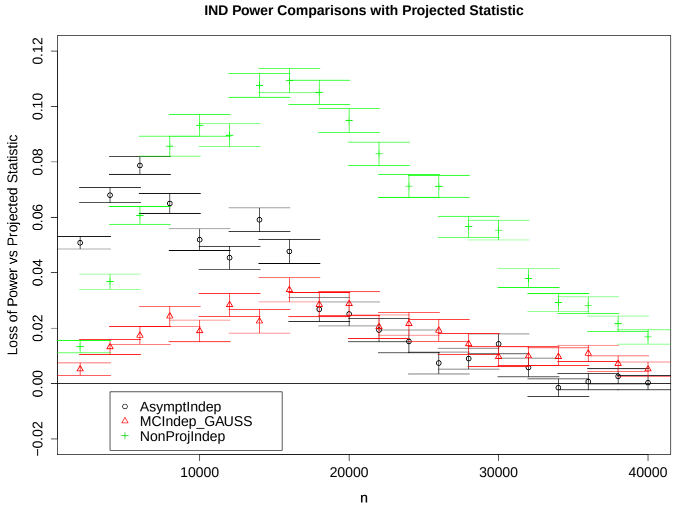


Figure 22: Comparison of empirical power between all zCDP hypothesis tests for independence and `GenChiTest` with projected statistic.

Hence, a different approach for a private independence test is to add Gaussian noise with variance $\frac{\Delta_\chi^2}{2\rho}$ to the chi-square statistic itself, which we call *output perturbation*. Our statistic is then simply the Gaussian mechanism for the chi-square statistic. We then compare the private statistic value with the distribution of $\mathcal{T} = \chi_2^2 + N\left(0, \frac{\Delta_\chi^2}{2\rho}\right)$ where the degrees of freedom is 2 because we have $(r-1)\cdot(c-1) = 2$. Thus, given a Type I error of at most α, we then set our critical value as $\tau_{\text{Gauss}}(\alpha; n, \rho)$ where

$$\Pr\left[\mathcal{T} > \tau_{\text{Gauss}}(\alpha; n, \rho)\right] = \alpha$$

We note that $\mathbb{V}\left[\mathcal{T}\right] = 4 + \frac{16n^2}{2\rho(n+2)^2}$, whereas $\mathcal{R}_\rho^{(n)}\left(\widehat{\boldsymbol{\theta}}^{(n)}\right)$ has asymptotic (as $n \to \infty$) variance 4. As we will see in our simulations, this additional term in the variance turns out to hurt power substantially. Thus, output perturbation does not seem to be a useful approach in hypothesis testing even if extra conditions on the data are assumed – as in the even split between case and control groups.

For our experiments, we again set Type I error threshold $\alpha = 0.05$ and consider the empirical power in $10,000$ trials. We fix the probability vector $(1/3, 1/3, 1/3)$ over the 3 rows in the first column whereas in the second column we set $(1/2, 1/4, 1/4)$, therefore the case and control groups do not produce the same outcomes and thus not independent of the disease. In Figure 23, we show a comparison in the power between our test with the projected statistic, which assumes no structure on the data, and the output perturbation test, which crucially relies on the fact that the data is evenly split between the case and control groups. Note that our new proposed test does not require evenly split data and significantly improves on the output perturbation test – we can get comparable power to the output perturbation test when the privacy parameter ρ is 25 times smaller.

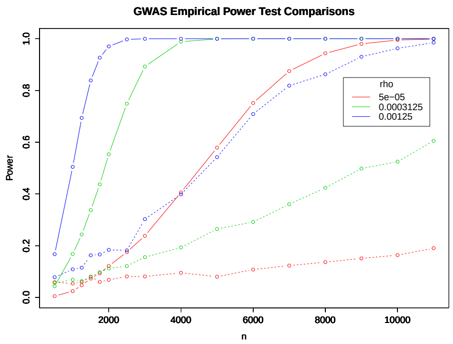


Figure 23: A comparison of empirical power between `GenChiTest` with projected statistic and output perturbation from Yu et al. (2014) for independence testing for GWAS type datasets.

7.4. General Chi-Square Tests with Arbitrary Noise Distributions

We next show that we can apply our testing framework in Algorithm 15 for any type of noise distribution we want to include for privacy concerns. For example, we consider adding Laplace noise rather than Gaussian noise if our privacy benchmark were (pure) differential privacy. In this case, we add Laplace noise with variance $8/\epsilon^2$ when computing the two statistics $\mathrm{R}^{(n)}_{\epsilon^2/8}\left(\widehat{\boldsymbol{\theta}}^{(n)}\right)$ from (7.22) and $\mathcal{R}^{(n)}_{\epsilon^2/8}\left(\widehat{\boldsymbol{\theta}}^{(n)}\right)$ from (7.23) so that the resulting tests will be ϵ-DP and hence $\frac{\epsilon^2}{2}$-zCDP from Theorem 2.2.3. Note that the resulting asymptotic distribution in this case will not be chi-square when we use noise other than Gaussian. We will then rely on Monte Carlo (MC) sampling to find the critical value in which to reject the null hypothesis. We give the MC based test which adds independent Laplace noise with variance $8/\epsilon^2$ in Algorithm 16 and is thus ϵ-DP, but any noise distribution can be used where we replace the parameter $1/\rho$ in the two statistics to be the variance of the noise that is added to each count. In fact, Gaussian noise can be used in this framework although the asymptotic distribution seems to do well in practice even for small sample sizes.

We show that we can use the general chi-square test `MC-GenChiTest` with ϵ-DP which uses Laplace noise in Algorithm 16 for goodness of fit testing $\mathrm{H}_0 : \boldsymbol{p} = \boldsymbol{p}^0$. In this case we select $\boldsymbol{p}(\widehat{\boldsymbol{\theta}}^{(n)}) = \boldsymbol{p}^0$ and $\phi(\boldsymbol{H} + \boldsymbol{Z}) = \boldsymbol{p}^0$ in both the nonprojected and projected statistics. From the way that we have selected the critical value $\tau(\alpha, \epsilon^2/8)$ in Algorithm 16, we have the following result on Type I error, which follows directly from Theorem 6.1.5.

Theorem 7.4.1. *When the number of independent samples m we choose for our MC sampling is larger than $1/\alpha$, testing $\mathrm{H}_0 : \boldsymbol{p} = \boldsymbol{p}^0$ in Algorithm 16 guarantees Type I error at most α.*

Note, that we are not guaranteed to have Type I error at most α when we have composite tests, e.g. independence testing, in `MC-GenChiTest` because we are not sampling from the exact data distribution.

Algorithm 16 Private Minimum Chi-Square Test using MC `MC-GenChiTest`

```
Input: Histogram data $\boldsymbol{h}$; $\epsilon$, $\alpha$, $\mathrm{H}_0 : \boldsymbol{\theta}^0 \in \Theta$, $m$ trials.
  Set $\widetilde{\boldsymbol{h}} \leftarrow \boldsymbol{h} + \boldsymbol{Z}$ where $\boldsymbol{Z} = (Z_1, \cdots, Z_d)$, where $\{Z_i\} \overset{i.i.d.}{\sim} \mathcal{P}$ where $\mathbb{V}[Z_i] = \sigma^2$.
  Set $\widehat{M} = \left(\Sigma^0_{n/\sigma^2}\left(\phi(\widetilde{\boldsymbol{h}})\right)\right)^{-1}$
  For the nonprojected statistic:
      $\mathrm{T}(\theta) = \frac{1}{n}\left(\widetilde{\boldsymbol{h}} - n\boldsymbol{p}(\boldsymbol{\theta})\right)^\intercal \widehat{M} \left(\widetilde{\boldsymbol{h}} - n\boldsymbol{p}(\boldsymbol{\theta})\right)$
  Set $\widehat{\boldsymbol{\theta}}^{(n)} = \mathrm{argmin}_{\boldsymbol{\theta} \in \Theta} \mathrm{T}(\boldsymbol{\theta})$
  Sample $\{r_1, \cdots, r_m\}$ as $m$ samples from the distribution of $\mathrm{T}(\widehat{\boldsymbol{\theta}}^{(n)})$.
  Set $\tau(\alpha, \sigma^2)$ to be the $\lceil (m+1)(1-\alpha) \rceil$-largest value of $\{r_1, \cdots, r_m\}$.
  For the projected statistic:
      $\mathrm{T}(\boldsymbol{\theta}) = \frac{1}{n}\left(\widetilde{\boldsymbol{h}} - n\boldsymbol{p}(\boldsymbol{\theta})\right)^\intercal \Pi \; \widehat{M} \; \Pi \; \left(\widetilde{\boldsymbol{h}} - n\boldsymbol{p}(\boldsymbol{\theta})\right)$
  Set $\widehat{\boldsymbol{\theta}}^{(n)} = \mathrm{argmin}_{\boldsymbol{\theta} \in \Theta} \mathrm{T}(\boldsymbol{\theta})$
  Sample $\{r_1, \cdots, r_m\}$ as $m$ samples from the distribution of $\mathrm{T}(\widehat{\boldsymbol{\theta}}^{(n)})$.
  Set $\tau(\alpha, \sigma^2)$ to be the $\lceil (m+1)(1-\alpha) \rceil$-largest value of $\{r_1, \cdots, r_m\}$.
  if $\mathrm{T}(\widehat{\boldsymbol{\theta}}^{(n)}) > \tau(\alpha, \sigma^2)$ then
    Decision ← Reject.
  else
    Decision ← Fail to Reject.
Output: Decision
```

7.5. Conclusion

We have demonstrated a new broad class of private hypothesis tests `GenChiTest` for categorical data based on the minimum chi-square theory. We gave two statistics (*nonprojected* and *projected*) that converge to a chi-square distribution when we use Gaussian noise and thus lead to zCDP hypothesis tests. Unlike prior work, these statistics have the same asymptotic distributions in the private asymptotic regime as the classical chi-square tests have in the classical asymptotic regime.

Our simulations show that with either the nonprojected or projected statistic our tests achieve at most α Type I error. We then empirically showed that our tests using the projected statistic significantly improves the Type II error when compared to the nonprojected statistic and previous private hypothesis tests from Chapter 6 which used the traditional chi-square statistic. Further, our new tests give comparable power to the classical (nonprivate) chi-square tests. We then gave further applications of our new statistics to GWAS data and how we can incorporate other noise distributions (e.g. Laplace) using an MC sampling approach.

CHAPTER 8

LOCAL PRIVATE HYPOTHESIS TESTS

8.1. Introduction

We now explore some differentially private hypothesis tests in the *local* model, i.e. there is no trusted curator that collects the data of all the individuals. In this model, each person injects their own independent noise and releases only the sanitized version of their data. The first differentially private algorithm called *randomized response* – in fact it predates the definition of differential privacy by more than 40 years – guarantees differential privacy in the local model (Warner, 1965). Recall that we have already presented randomized response in Definition 5.2.1.

Most of the work in differential privacy has been in the *curator model* so that the raw data of all the individuals is stored in some centralized location. One of the main reasons for this is that we can achieve much greater accuracy in our differentially private statistics when used in the curator setting, where accuracy is measured as the difference between the true value of the statistic on the data and the differentially private version.

We now define *local* differential privacy, which was formalized by Raskhodnikova et al. (2008) and gives the strongest level of privacy presented thus far.

Definition 8.1.1 [LR Oracle]**.** *An* LR Oracle $LR_{\boldsymbol{x}}(\cdot,\cdot)$ *takes an input an index* $i \in [n]$ *and* ϵ*-DP algorithm* R *and outputs* $y \in \mathcal{Y}$ *chosen according to the distribution of* $R(x_i)$*, i.e.* $LR_{\boldsymbol{x}}(i, R) = R(x_i)$*.*

Definition 8.1.2. *An algorithm* $\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}$ *is* ϵ-local differentially private *(LDP) if it accesses input database* $\boldsymbol{x}$ *via the LR oracle* $LR_{\boldsymbol{x}}$ *with the following restriction: if* $LR(i, R_j)$ *for* $j \in [k]$ *are the* $\mathcal{M}$*'s invocations of* $LR_{\boldsymbol{x}}$ *on index* i*, where each* R_j *for* $j \in [k]$ *is an* ϵ_j*-DP and* $\sum_{j=1}^{k} \epsilon_j \leq \epsilon$*.*

An easy consequence of this definition is that an algorithm which is ϵ-LDP is also ϵ-DP. Note that we can easily extend these definitions to include versions of (ϵ, δ)-LDP and (ξ, ρ)-local zCDP (LzCDP).

The utility guarantees that we are after in hypothesis testing is to fix a bound on the probability of Type I error and then minimize the probability of Type II error. We now explore some differentially private hypothesis tests in the local model.

8.2. Local Private Chi-Square Tests

8.2.1. Local zCDP – Gaussian Noise

The two previous chapters have looked at private chi-square tests in the curator model and so here we look at private chi-square tests in the local model. We start by considering adding Gaussian noise to each person's data. Recall that before in the curator model we started with a histogram of everyone's data $\boldsymbol{H} \sim \text{Multinomial}(n,\boldsymbol{p})$ and then added independent Gaussian noise $\mathrm{N}\left(0, 1/\rho\right)$ to each of the d bins to ensure ρ-zCDP. We can write $\boldsymbol{H} = \sum_{i=1}^n \boldsymbol{X}_i$ where $\boldsymbol{X}_i = (X_{i,1}, \cdots, X_{i,d}) \sim \text{Multinomial}(1,\boldsymbol{p})$. Each individual i's data is represented as a vector $\boldsymbol{X}_i$ which is a standard basis element of $\mathbb{R}^d$. Hence, in the local model we can add $\boldsymbol{Z}_i \sim \mathrm{N}\left(\boldsymbol{0}, \frac{1}{\rho}\, I_d\right)$ independent noise to $\boldsymbol{X}_i$ to ensure ρ-LzCDP.

The resulting noisy histogram that we get after accumulating all the noise terms from each individual is then

$$\widetilde{\boldsymbol{H}} = \boldsymbol{H} + \boldsymbol{Z} \qquad \text{where } \boldsymbol{Z} \sim \mathrm{N}\left(\boldsymbol{0}, \frac{n}{\rho}\, I_d\right). \tag{8.1}$$

We compare this to the noisy histogram we used in the previous two chapters, which was $\boldsymbol{H} + \boldsymbol{Z}$ where $\boldsymbol{Z} \sim N(0, 1/\rho\; I_d)$. Thus, we have increased the variance of the noise by a factor of n by moving to the local model. Recall that in the *variance aware privacy regime*, we think of $\rho \equiv \rho_n$ and require $n\rho_n \to \rho^\star > 0$. This is the asymptotic rate we required for the modified asymptotic distribution of the statistic in Chapter 6 and for the asymptotic distribution of the *non-projected* statistic in Chapter 7. However, in the local model the variance of the noise we are adding is n/ρ_n, thus the variance aware privacy regime in the local model requires $\rho_n \to \rho^\star > 0$, or simply put we could just have $\rho_n > 0$ fixed for each n. We then restate the asymptotic distributions of the statistics in Chapters 6 and 7 in the local model.

Theorem 8.2.1. *Under the null hypothesis* $\mathrm{H}_0 : \boldsymbol{p} = \boldsymbol{p}^0$*, the statistic* $\mathrm{T}^{(n)}\left(N\left(0, \frac{n}{\rho}\right)\right) = \boldsymbol{W}^\intercal \boldsymbol{\Lambda}_\rho \boldsymbol{W}$ *(compare with* (6.8)*) for* $\rho > 0$ *converges in distribution to the linear combination of independent chi-squared random variables each with one degree of freedom given in Theorem 6.1.7, where we replace* $\rho^\star$ *with* ρ*.*

Theorem 8.2.2. *Under the null hypothesis* $\mathrm{H}_0 : \boldsymbol{p} = \boldsymbol{p}^0$*, the statistic* $\mathrm{Q}^{(n)}_{\rho/n}$ *given in* (7.13) *for* $\rho > 0$ *converges in distribution to* χ^2_d *(compare with Theorem 7.2.3). Further, the statistic* $\mathcal{Q}^{(n)}_{\rho/n}$ *from* (7.16) *converges in distribution to* χ^2_{d-1} *(compare with Theorem 7.2.6).*

We also point out that all of the zCDP tests in the two previous chapters are already $n\rho$-LzCDP. Thus, we have already considered the power of these locally private hypothesis tests. However, for a fixed level ρ-LzCDP, our loss in power will follow the power curves in our experiments for ρ/n-zCDP. Hence, there is a cost in power of our tests in the local setting versus the curator setting.

For our composite tests, we again can say that `AsymptIndep` and `GenChiTest` from Algorithm 13 and Algorithm 15, respectively, are $n\rho$-LzCDP.

8.2.2. Local DP

Note that the sum of two Laplace random variables is not Laplace. Thus, our MC based tests are not automatically LDP. We can modify the MC tests where we now add Laplace noise with parameter $2/\epsilon$ to each component of each person's input vector so that the histogram has the added random vector $\boldsymbol{Z} \overset{i.i.d.}{\sim} \mathcal{L}$ where $\mathcal{L}$ is distributed as the sum of n independent $\mathrm{Lap}(2/\epsilon)$.

Rather than having to add noise to each component of the original data histogram with variance $n\frac{8}{\epsilon^2}$, we consider applying randomized response to obtain a LDP hypothesis test. We will use a form of the *exponential mechanism* (McSherry and Talwar, 2007) given in Algorithm 17 which takes a single data entry from the set $\{\boldsymbol{e}_1, \cdots, \boldsymbol{e}_d\}$, where $\boldsymbol{e}_j \in \mathbb{R}^d$ is the standard basis element with a 1 in the jth coordinate and is zero elsewhere, and reports the original entry with probability slightly more than uniform and otherwise reports a different element. Note that $\mathcal{M}_{\texttt{EXP}}$ takes a single data entry and is ϵ-DP.[1]

Algorithm 17 Exponential Mechanism $\mathcal{M}_{\texttt{EXP}}$

```
Input: Data $\boldsymbol{x} \in \{\boldsymbol{e}_1, \cdots, \boldsymbol{e}_d\}$, privacy parameter $\epsilon$.
  Let $q(\boldsymbol{x}, \boldsymbol{z}) = \mathbb{1}\{\boldsymbol{x} = \boldsymbol{z}\}$
  Select $\widetilde{\boldsymbol{x}}$ with probability $\frac{\exp[\epsilon\ q(\boldsymbol{x}, \widetilde{\boldsymbol{x}})]}{e^\epsilon - 1 + d}$
Output: $\widetilde{\boldsymbol{x}}$
```

We have the following result when we use $\mathcal{M}_{\texttt{EXP}}$ on each individual's data to obtain a private histogram.

Lemma 8.2.3. *If we have histogram* $\boldsymbol{H} = \sum_{i=1}^n \boldsymbol{X}_i$*, where* $\{\boldsymbol{X}_i\} \overset{i.i.d.}{\sim} \text{Multinomial}(1, \boldsymbol{p})$ *and we write* $\check{\boldsymbol{H}} = \sum_{i=1}^n \check{\boldsymbol{H}}_i$*, where* $\check{\boldsymbol{H}}_i = \mathcal{M}_{\texttt{EXP}}(\boldsymbol{X}_i, \epsilon)$ *for each* $i \in [n]$*, then*

$$\check{\boldsymbol{H}} \sim \text{Multinomial}(n, \check{\boldsymbol{p}}) \qquad \textit{where} \qquad \check{\boldsymbol{p}} = \boldsymbol{p}\left(\frac{e^\epsilon}{e^\epsilon + d - 1}\right) + (1 - \boldsymbol{p})\left(\frac{1}{e^\epsilon + d - 1}\right). \quad (8.2)$$

We can then form a chi-square statistic using the private histogram $\check{\boldsymbol{H}}$, which gives us the following result.

Theorem 8.2.4. *Let* $\boldsymbol{H} \sim \text{Multinomial}(n, \boldsymbol{p})$ *and* $\check{\boldsymbol{H}}$ *be given in Lemma 8.2.3 with privacy parameter* $\epsilon > 0$*. Under the null hypothesis* $\mathrm{H}_0 : \boldsymbol{p} = \boldsymbol{p}^0$*, we have for* $\check{\boldsymbol{p}}^0 = \frac{1}{e^\epsilon + d - 1}\left(e^\epsilon \boldsymbol{p}^0 + (1 - \boldsymbol{p}^0)\right)$

$$\check{\mathrm{T}}_\epsilon^{(n)} \stackrel{\text{defn}}{=} \sum_{j=1}^d \frac{(\check{H}_j - n\check{p}_j^0)^2}{n\check{p}_j^0} \xrightarrow{D} \chi^2_{d-1}. \quad (8.3)$$

We then base our LDP goodness of fit test on this result, which is presented in Algorithm 18

Theorem 8.2.5. *The test* `LocalGOF`$(\cdot, \epsilon, \alpha, \boldsymbol{p}^0)$ *is* ϵ*-LDP.*

Proof. The proof follows simply from the fact that we use $\mathcal{M}_{\texttt{EXP}}$ for each individual's data and then `LocalGOF` aggregates the privatized data, which is just a post-processing function. □

Although we cannot guarantee the probability of a Type I error at most α due to the fact

---

[1] We point out that $\mathcal{M}_{\texttt{EXP}}$ is ϵ-DP, whereas the traditional exponential mechanism tells us that it is 2ϵ-DP. The savings in the factor of 2 results in that the normalizing constant is not affected by the input data, it is always $\frac{1}{e^\epsilon - 1 + d}$.

Algorithm 18 Local DP GOF Test `LocalGOF`

```
Input: $\boldsymbol{x} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)$, $\epsilon$, $\alpha$, $\mathrm{H}_0 : \boldsymbol{p} = \boldsymbol{p}^0$.
  Let $\check{\boldsymbol{p}}^0 = \frac{1}{e^\epsilon + d - 1}\left(e^\epsilon \boldsymbol{p}^0 + (1 - \boldsymbol{p}^0)\right)$.
  Let $\check{\boldsymbol{h}} = \sum_{i=1}^n \check{\boldsymbol{x}}_i$ where $\check{\boldsymbol{x}}_i = \mathcal{M}_{\texttt{EXP}}(\boldsymbol{x}_i, \epsilon)$.
  Set $q = \sum_{j=1}^d \frac{\check{h}_j - n\check{p}^0_j}{n\check{p}^0_j}$
  if $q > \chi^2_{d-1,1-\alpha}$ then
    Decision ← Reject.
  else
    Decision ← Fail to Reject.
Output: Decision
```

that we use the asymptotic distribution (as in the tests from the previous chapters), we can show experimentally that we can bound this error. We expect the Type I errors to be similar to those from the nonprivate test.

We now turn to the power of our test, `LocalGOF`. Recall from Equation (6.10) that we consider the alternate $\mathrm{H}_1 : \boldsymbol{p} = \boldsymbol{p}^1_n$ where $\boldsymbol{p}^1_n = \boldsymbol{p}^0 + \boldsymbol{\Delta}/\sqrt{n}$ where $\sum_{j=1}^d \Delta_j = 0$.

Theorem 8.2.6. *Assume* $\boldsymbol{H} \sim \mathrm{Multinomial}(n, \boldsymbol{p}^1_n)$ *where* $\boldsymbol{p}^1_n$ *is given in* (6.10). *Then we have*

$$\check{\mathrm{T}}^{(n)}_\epsilon \xrightarrow{D} \chi^2_{d-1}\left(\left(\frac{e^\epsilon - 1}{e^\epsilon + d - 1}\right)^2 \sum_{j=1}^d \frac{\Delta_j^2}{\check{p}^0_j}\right).$$

Proof. This follows the same analysis as in Lemma 6.1.8 □

We point out here that the noncentral parameter has been reduced by roughly a multiplicative factor of $\left(\frac{e^\epsilon - 1}{e^\epsilon + d - 1}\right)^2 \approx \epsilon^2/d^2$ from the noncentral term in Lemma 6.1.8. Thus, if we consider an alternate $\mathrm{H}_1 : \boldsymbol{p} = \boldsymbol{p}^0 + \boldsymbol{\Delta}$ (without the factor of $\sqrt{n}$) for a fixed significance $1 - \alpha$, and null hypothesis $\boldsymbol{p}^0$, we would expect `LocalGOF` to have similar power with $\frac{d}{\epsilon} n$ many samples when compared to the nonprivate classical test `GOF` with n samples.

We can also compare `LocalGOF` with the addition of Gaussian noise from the previous section. For a fair comparison, we use the test which has empirically the best power, `NewStatAsymptGOF` with the projected statistic $\mathcal{Q}^{(n)}_{\rho/n}$, which ensures ρ-LzCDP. Further, for a fixed $\epsilon > 0$ we will set $\rho = \epsilon^2/2$, so that both `LocalGOF` and `NewStatAsymptGOF` are ρ-LzCDP, by Theorem 2.2.3 . We summarize this in the following result.

Theorem 8.2.7. *For* $\epsilon > 0$*, both tests* `LocalGOF` *and* `NewStatAsymptGOF` *with projected statistic* $\mathcal{Q}^{(n)}_{\epsilon^2/(2n)}$ *are* $\epsilon^2/2$*-LzCDP. Further, assuming that* $\boldsymbol{H} \sim \mathrm{Multinomial}(n, \boldsymbol{p}^1_n)$ *where* $\boldsymbol{p}^1_n$ *is given in* (6.10)*, we have*

$$\mathcal{Q}^{(n)}_{\epsilon^2/(2n)} \xrightarrow{D} \chi^2_{d-1}\left(\boldsymbol{\Delta}^\intercal \left(\Sigma^0_{\epsilon^2/2}\right)^{-1} \boldsymbol{\Delta}\right)$$

Proof. The second statement follows directly from Theorem 7.2.10 □

To compare the two noncentral parameters $\left(\frac{e^\epsilon-1}{e^\epsilon+d-1}\right)^2 \sum_{j=1}^d \frac{\Delta_j^2}{\tilde{p}_j^0}$ and $\boldsymbol{\Delta}^\intercal \left(\Sigma^0_{\epsilon^2/2}\right)^{-1} \boldsymbol{\Delta}$, we use the form of the$\left(\Sigma^0_\rho\right)^{-1}$ given in (7.15). As an example, let $\boldsymbol{p}^0 = (1/d, \cdots, 1/d)^\intercal$. We can then directly compare the two noncentral parameters

$$\boldsymbol{\Delta}^\intercal \left(\Sigma^0_{\epsilon^2/2}\right)^{-1} \boldsymbol{\Delta} = \frac{d\epsilon^2 \; \boldsymbol{\Delta}^\intercal \boldsymbol{\Delta}}{\epsilon^2 + 2d} \qquad \& \qquad \left(\frac{e^\epsilon - 1}{e^\epsilon + d - 1}\right)^2 \sum_{j=1}^d \frac{\Delta_j^2}{\tilde{p}_j^0} = d \left(\frac{e^\epsilon - 1}{e^\epsilon + d - 1}\right)^2 \boldsymbol{\Delta}^\intercal \; \boldsymbol{\Delta}$$

Thus, for large d we would have a smaller noncentral parameter for the test `LocalGOF` than for the other tests, thus `LocalGOF` would be expected to have worse power. However, for small values of d, we might expect `LocalGOF` to have better power.

We then empirically check the power comparison between `LocalGOF` and `MC-GenChiTest` in Algorithm 16 with the projected statistic using $\boldsymbol{Z} \overset{i.i.d.}{\sim} \mathcal{L}$ to ensure ϵ-LDP. We consider various null hypotheses $\boldsymbol{p}^0$ in Figure 24.

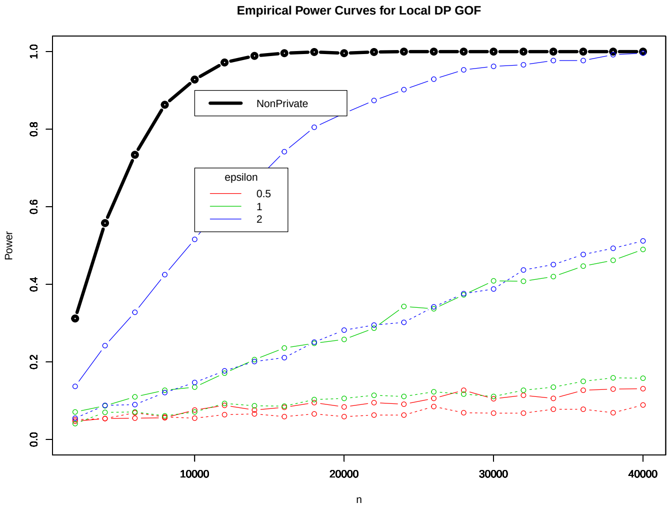


Figure 24: Comparison of empirical power of classical non-private test versus local private tests `LocalGOF` (solid line) and `MC-GenChiTest` with projected statistic and Laplace noise (dashed line) for alternate $\mathrm{H}_1 : \boldsymbol{p}^1 = \boldsymbol{p}^0 + \boldsymbol{\Delta}$ in 1,000 trials. We set $\boldsymbol{p}^0 = (1/4, 1/4, 1/4, 1/4)^\intercal$ and $\mathrm{H}_1 : \boldsymbol{p}^0 + \boldsymbol{\Delta}$ where $\boldsymbol{\Delta} = 0.01 \cdot (1, -1, -1, 1)^\intercal$.

We also point out that we can do more general chi-square tests where the null hypothesis represents a family of distributions $\mathrm{H}_0 : \boldsymbol{\theta}^0 \in \Theta \subseteq \mathbb{R}^s$ after we have applied randomized response to each individual's data. We just need to keep in mind that for $\boldsymbol{p}(\boldsymbol{\theta})$ where $\boldsymbol{\theta} \in \mathbb{R}^s$, the privatized data $\check{\boldsymbol{H}} \sim \text{Multinomial}(n, \check{\boldsymbol{p}}(\boldsymbol{\theta}))$ where

$$\check{\boldsymbol{p}}(\boldsymbol{\theta}) = \left(\frac{1}{e^{\epsilon} + d - 1}\right)\left(e^{\epsilon}\boldsymbol{p}(\boldsymbol{\theta}) + (1 - \boldsymbol{p}(\boldsymbol{\theta}))\right).$$

In order to find a good estimate for the unknown true parameter $\boldsymbol{\theta}^0$, we find a function ϕ such that $\phi\left(\frac{\check{\boldsymbol{H}}}{n}\right) \xrightarrow{P} \boldsymbol{\theta}^0$ and then find $\widehat{\boldsymbol{\theta}}^{(n)} = \arg\min_{\boldsymbol{\theta}\in\Theta} \widehat{\mathrm{D}}^{(n)}(\boldsymbol{\theta})$, for the statistic $\widehat{\mathrm{D}}^{(n)}(\boldsymbol{\theta})$ given in (7.3) with the following substitutions: $\mathbf{V}^{(n)} = \check{\boldsymbol{H}}/n$, $A(\boldsymbol{\theta}) = \check{\boldsymbol{p}}(\boldsymbol{\theta})$, $M(\boldsymbol{\theta}) = \mathrm{Diag}\left(\check{\boldsymbol{p}}(\boldsymbol{\theta})\right)^{-1}$, and $C(\boldsymbol{\theta}) = M(\boldsymbol{\theta}) - A(\boldsymbol{\theta})A(\boldsymbol{\theta})^{\intercal}$. Note that in this case that $C(\boldsymbol{\theta})$ is the covariance matrix of a multinomial, which has rank $d-1$. The following result follows from Theorem 7.1.4.

Theorem 8.2.8. *Under the null hypothesis* $\mathrm{H}_0 : \boldsymbol{\theta} \in \Theta$*, let conditions given in Assumption 7.1.1 and* (7.1) *hold where* $\mathbf{V}^{(n)} = \check{\boldsymbol{H}}/n$*,* $A(\boldsymbol{\theta}) = \check{\boldsymbol{p}}(\boldsymbol{\theta})$*,* $M(\boldsymbol{\theta}) = \mathrm{Diag}\left(\check{\boldsymbol{p}}(\boldsymbol{\theta})\right)^{-1}$*,* $C(\boldsymbol{\theta}) = \check{\boldsymbol{p}}(\boldsymbol{\theta}) - \check{\boldsymbol{p}}(\boldsymbol{\theta})\check{\boldsymbol{p}}(\boldsymbol{\theta})^{\intercal}$*, and true parameter* $\boldsymbol{\theta}^0 \in \Theta$*. We then have,*

$$\sum_{j=1}^{d} \frac{\left(\check{H}_j - n\check{p}_j\left(\widehat{\boldsymbol{\theta}}^{(n)}\right)\right)^2}{n\check{p}_j\left(\widehat{\boldsymbol{\theta}}^{(n)}\right)} \xrightarrow{D} \chi^2_{d-1-s}$$

We then present our more general chi-square test in Algorithm 19.

Algorithm 19 Local DP General Chi-Square Test `LocalGeneral`

```
Input: $\boldsymbol{x} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)$, $\phi$, $\epsilon$, $\alpha$, $\mathrm{H}_0 : \boldsymbol{\theta} \in \Theta$.
  Let $\check{\boldsymbol{p}}^{\mathrm{MIN}} = \check{\boldsymbol{p}}(\check{\widehat{\boldsymbol{\theta}}}^{(n)})$.
  Let $\check{\boldsymbol{h}} = \sum_{i=1}^{n} \check{\boldsymbol{x}}_i$ where $\check{\boldsymbol{x}}_i = \mathcal{M}_{\mathtt{EXP}}(\boldsymbol{x}_i, \epsilon)$.
  Set $q = \sum_{j=1}^{d} \frac{\check{h}_j - n\check{p}_j^{\mathrm{MIN}}}{n\check{p}_j^{\mathrm{MIN}}}$
  if $q > \chi^2_{d-1-s,1-\alpha}$ then
    Decision ← Reject.
  else
    Decision ← Fail to Reject.
Output: Decision
```

As an example of our more general test, we again consider independence testing with null hypothesis $\boldsymbol{p}(\boldsymbol{\pi}^{(1)}, \boldsymbol{\pi}^{(2)}) = \boldsymbol{\pi}^{(1)}\left(\boldsymbol{\pi}^{(2)}\right)^{\intercal}$. Thus from a contingency table $\check{\boldsymbol{H}}$ with r rows and c columns and each person's data is sanitized by $\mathcal{M}_{\mathtt{EXP}}$, we can form the estimate where

$$\begin{aligned}
\phi\left(\frac{\check{\boldsymbol{H}}}{n}\right) &= \left(\check{\boldsymbol{\pi}}^{(1)}, \check{\boldsymbol{\pi}}^{(2)}\right) \\
\text{where } \check{\pi}_i^{(1)} &= \left(\frac{e^{\epsilon} + d - 1}{e^{\epsilon} - 1}\right)\left(\frac{\check{H}_{i,\cdot}}{n} - \frac{c}{e^{\epsilon} + d - 1}\right) \quad \text{for } i \in [r] \\
\check{\pi}_i^{(2)} &= \left(\frac{e^{\epsilon} + d - 1}{e^{\epsilon} - 1}\right)\left(\frac{\check{H}_{\cdot,j}}{n} - \frac{r}{e^{\epsilon} + d - 1}\right) \quad \text{for } i \in [c]
\end{aligned}$$

which does converge in probability to the true parameters $(\boldsymbol{\pi}^{(1)}, \boldsymbol{\pi}^{(2)})$. We then form the chi-square statistic

$$\begin{aligned}
\mathrm{T}\left(\boldsymbol{\pi}^{(1)}, \boldsymbol{\pi}^{(2)}\right) &= n \sum_{i,j} \frac{\left(\frac{\check{H}_{i,j}}{n} - \check{p}_{i,j}\left(\boldsymbol{\pi}^{(1)}, \boldsymbol{\pi}^{(2)}\right)\right)^2}{\check{p}_{i,j}\left(\check{\boldsymbol{\pi}}^{(1)}, \check{\boldsymbol{\pi}}^{(2)}\right)} \\
&= n(e^\epsilon + d - 1) \sum_{i,j} \frac{\left[\frac{\check{H}_{i,j}}{n} - \left(\frac{1}{e^\epsilon + d - 1}\right)\left(e^\epsilon \ \pi_i^{(1)}\pi_j^{(2)} + (1 - \pi_i^{(1)}\pi_j^{(2)})\right)\right]^2}{e^\epsilon \ \check{\pi}^{(1)}\check{\pi}^{(2)} + \left(1 - \check{\pi}^{(1)}\check{\pi}^{(2)}\right)}
\end{aligned}$$

We then maximize $\mathrm{T}\left(\boldsymbol{\pi}^{(1)}, \boldsymbol{\pi}^{(2)}\right)$ over the region $\boldsymbol{\pi}^{(t)} \geq \mathbf{0}$ for $t \in \{1, 2\}$, $\sum_{i=1}^r \pi_i^{(1)} = 1$, and $\sum_{j=1}^c \pi_j^{(2)} = 1$.

We then turn to empirical results demonstrating the power of our local private independence test using randomized response. We consider the same setting as in the previous two chapters: 2×2 contingency table where $\boldsymbol{\pi}^{(i)} = (\pi^{(i)}, 1 - \pi^{(i)})$ for $i \in \{1, 2\}$. We present our results in Figure 25.

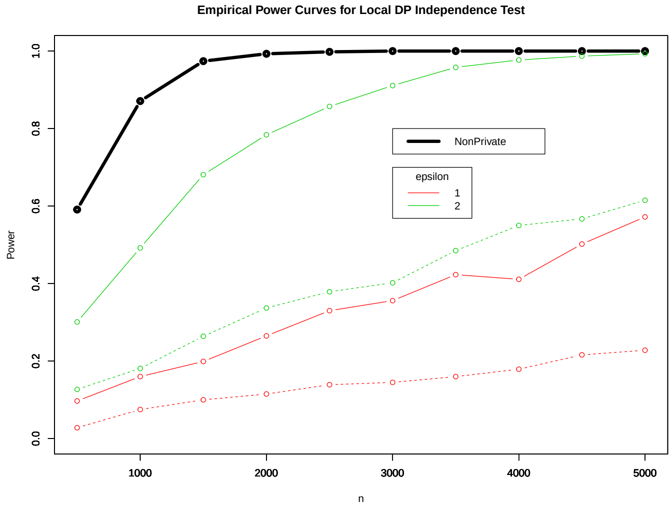


Figure 25: Comparison of empirical power of classical non-private test versus local private tests `LocalGOF` (solid line) and `MC-GenChiTest` with Laplace noise and projected statistic (dashed line) in 1,000 trials. We set $\pi^{(1)} = \pi^{(2)} = 1/2$ and $\boldsymbol{\Delta} = 0.025 \cdot (1, -1)^\intercal (1, -1)$.

8.2.3. Local DP - Flip Components

We next consider another differentially private algorithm $\mathcal{M} : \{\boldsymbol{e}_1, \cdots, \boldsymbol{e}_d\} \to \{0, 1\}^d$, given in Algorithm 20.[2]

Theorem 8.2.9. *The algorithm $\mathcal{M}_{\texttt{FC}}$ is ϵ-DP.*

We then want to form a statistic based on the output $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$, (where $\boldsymbol{x}_i = (x_{i,j} : j \in [d])$)

---

[2] Special thanks to Adam Smith for recommending to use this particular algorithm.

Algorithm 20 Local Randomizer $\mathcal{M}_{\texttt{FC}}$

```
Input: Data $\boldsymbol{x} \in \{\boldsymbol{e}_1, \cdots, \boldsymbol{e}_d\}$, privacy parameter $\epsilon$.
  for $j \in [d]$ do
    Set $z_j = x_j$ with probability $\frac{e^{\epsilon/2}}{e^{\epsilon/2}+1}$, otherwise $z_j = (1 - x_j)$.
Output: $\boldsymbol{z}$
```

that is asymptotically distributed as a chi-square under the null hypothesis.

Theorem 8.2.10. *Let* $\mathrm{H}_0 : \boldsymbol{p} = \boldsymbol{p}^0$. *We consider data* $\boldsymbol{X}_i \sim \mathrm{Multinomial}(1, \boldsymbol{p})$ *for each* $i \in [n]$. *We define the following covariance matrix*

$$\Sigma^{FC} = \left(\frac{e^{\epsilon/2} - 1}{e^{\epsilon/2} + 1}\right)^2 \left[\mathrm{Diag}\left(\boldsymbol{p}^0\right) - \boldsymbol{p}^0 \left(\boldsymbol{p}^0\right)^\intercal\right] + \frac{e^{\epsilon/2}}{\left(e^{\epsilon/2} + 1\right)^2} I_d.$$

If H_0 *holds, then the empirical average* $\boldsymbol{H}^{FC} = \frac{1}{n}\sum_{i=1}^n \mathcal{M}_{\texttt{FC}}(\boldsymbol{X}_i)$ *has the following asymptotic distribution*

$$\sqrt{n}\left(\boldsymbol{H}^{FC} - \frac{1}{1 + e^{\epsilon/2}}\left[\left(e^{\epsilon/2} - 1\right)\boldsymbol{p}^0 + 1\right]\right) \xrightarrow{D} \mathrm{N}\left(\boldsymbol{0}, \Sigma^{FC}\right).$$

Further, Σ^{FC} *is invertible for any* $\epsilon \geq 0$.

Proof. Follows from the central limit theorem. We first compute the expected value

$$\mathbb{E}\left[\boldsymbol{H}^{\texttt{FC}}\right] = \frac{e^{\epsilon/2}}{e^{\epsilon/2} + 1}\boldsymbol{p}^0 + \frac{1}{e^{\epsilon/2} + 1}(1 - \boldsymbol{p}^0).$$

In order to compute the covariance matrix, we consider the diagonal term (j, j)

$$\mathbb{E}\left[(H_j^{\texttt{FC}})^2\right] = \frac{e^{\epsilon/2}}{e^{\epsilon/2} + 1}p_j^0 + \frac{1}{e^{\epsilon/2} + 1}(1 - p_j^0)$$

Next we compute the off diagonal term (j, k)

$$\begin{aligned}
\mathbb{E}\left[H_j^{\texttt{FC}}\ H_k^{\texttt{FC}}\right] = &\left(\frac{e^{\epsilon/2}}{e^{\epsilon/2} + 1}\right)^2 \Pr\left[X_j = 1 \ \&\ X_k = 1\right] \\
&+ \frac{e^{\epsilon/2}}{\left(e^{\epsilon/2} + 1\right)^2}\left(\Pr\left[X_j = 1 \ \&\ X_k = 0\right] + \Pr\left[X_j = 0 \ \&\ X_k = 1\right]\right) \\
&+ \frac{1}{\left(e^{\epsilon/2} + 1\right)^2}\Pr\left[X_j = 0 \ \&\ X_k = 0\right] \\
= &\left(\frac{1}{e^{\epsilon/2} + 1}\right)^2 \left[e^{\epsilon/2}\left(p_j^0 + p_k^0\right) + (1 - p_j^0 - p_k^0)\right]
\end{aligned}$$

Before we construct the covariance matrix, we simplify a few terms

$$\begin{aligned}
&\mathbb{E}\left[(H_j^{\mathtt{FC}})^2\right] - \mathbb{E}\left[H_j^{\mathtt{FC}}\right]^2 \\
&\quad = \left(\frac{1}{e^{\epsilon/2}+1}\right)^2 \left[\left(e^{\epsilon/2}-1\right)\left(e^{\epsilon/2}+1\right)p_j^0 + 1 + e^{\epsilon/2} - \left(\left(e^{\epsilon/2}-1\right)p_j^0+1\right)^2\right] \\
&\quad = \left(\frac{1}{e^{\epsilon/2}+1}\right)^2 \left[\left(e^{\epsilon/2}-1\right)\left(e^{\epsilon/2}+1-2\right)p_j^0 - (e^{\epsilon/2}-1)^2\left(p_j^0\right)^2 + e^{\epsilon/2}\right] \\
&\quad = \frac{1}{\left(e^{\epsilon/2}+1\right)^2}\left[\left(e^{\epsilon/2}-1\right)^2 p_j^0\left(1-p_j^0\right) + e^{\epsilon/2}\right]
\end{aligned}$$

Further, we have

$$\begin{aligned}
&\mathbb{E}\left[(H_j^{\mathtt{FC}}H_k^{\mathtt{FC}}\right] - \mathbb{E}\left[H_j^{\mathtt{FC}}\right]\mathbb{E}\left[H_k^{\mathtt{FC}}\right] \\
&\quad = \left(\frac{1}{e^{\epsilon/2}+1}\right)^2\left[\left(e^{\epsilon/2}-1\right)\left(p_j^0+p_k^0\right) + 1 - \left((e^{\epsilon/2}-1)p_j^0+1\right)\left((e^{\epsilon/2}-1)p_k^0+1\right)\right] \\
&\quad = -\left(\frac{e^{\epsilon/2}-1}{e^{\epsilon/2}+1}\right)^2 p_j^0 p_k^0
\end{aligned}$$

Putting this together, the covariance matrix can then be written as

$$\begin{aligned}
\Sigma^{\mathtt{FC}} &= \mathbb{E}\left[\boldsymbol{H}^{\mathtt{FC}}\left(\boldsymbol{H}^{\mathtt{FC}}\right)^{\intercal}\right] - \mathbb{E}\left[\boldsymbol{H}^{\mathtt{FC}}\right]\left(\mathbb{E}\left[\boldsymbol{H}^{\mathtt{FC}}\right]\right)^{\intercal} \\
&= \left(\frac{e^{\epsilon/2}-1}{e^{\epsilon/2}+1}\right)^2\left[\mathrm{Diag}\left(\boldsymbol{p}^0\right) - \boldsymbol{p}^0\left(\boldsymbol{p}^0\right)^{\intercal}\right] + \frac{e^{\epsilon/2}}{\left(e^{\epsilon/2}+1\right)^2} I_d
\end{aligned}$$

□

As we did in Chapter 7, we can form the projected statistic in terms of the histogram $\boldsymbol{H}^{\mathtt{FC}}$.
Theorem 8.2.11. *We define the following quantities*

$$\boldsymbol{\mu} = \frac{1}{1+e^{\epsilon/2}}\left[\left(e^{\epsilon/2}-1\right)\boldsymbol{p}^0 + 1\right] \quad \& \quad \Pi = I_d - \frac{1}{d}\mathbf{1}\mathbf{1}^{\intercal}$$

$$\boldsymbol{q} = (e^{\epsilon/2}-1)^2\boldsymbol{p}^0 + e^{\epsilon/2}\mathbf{1}$$

If the null hypothesis $\mathrm{H}_0 : \boldsymbol{p} = \boldsymbol{p}^0$ *holds, then the following statistic*

$$\begin{aligned}
\mathrm{T}^{\mathit{FC}} &= n\left(e^{\epsilon/2}+1\right)^2\left(\boldsymbol{H}^{\mathit{FC}} - \boldsymbol{\mu}\right)^{\intercal} \\
&\qquad \cdot \Pi\left[\mathrm{Diag}\left(\boldsymbol{q}\right)^{-1} + \left(\frac{1}{\left(e^{\epsilon/2}-1\right)^2} - \sum_{j=1}^{d}\frac{(p_j^0)^2}{q_j}\right)^{-1}\left(\boldsymbol{p}^0/\boldsymbol{q}\right)\left(\boldsymbol{p}^0/\boldsymbol{q}\right)^{\intercal}\right]\Pi\left(\boldsymbol{H}^{\mathit{FC}} - \boldsymbol{\mu}\right)
\end{aligned}$$

is asymptotically distributed as a chi-square, $\mathrm{T}^{\mathit{FC}} \xrightarrow{D} \chi^2_{d-1}$.

Proof. We first note that $\Sigma^{\mathtt{FC}}$ has eigenvector $\mathbf{1}$ with eigenvalue $\frac{e^{\epsilon/2}}{\left(e^{\epsilon/2}+1\right)^2}$. Hence, we can project out this eigenvector with $\Pi = I_d - 1/d\mathbf{1}\mathbf{1}^\intercal$, as we did in Theorem 7.2.6 to get that the following statistic is asymptotically distributed as χ^2_{d-1},

$$n\left(\boldsymbol{H}^{\mathtt{FC}} - \frac{1}{1+e^{\epsilon/2}}\left[\left(e^{\epsilon/2}-1\right)\boldsymbol{p}^0 + 1\right]\right)^\intercal \Pi\left(\Sigma^{\mathtt{FC}}\right)^{-1}\Pi\left(\boldsymbol{H}^{\mathtt{FC}} - \frac{1}{1+e^{\epsilon/2}}\left[\left(e^{\epsilon/2}-1\right)\boldsymbol{p}^0 + 1\right]\right).$$

We then compute $\left(\Sigma^{\mathtt{FC}}\right)^{-1}$ using the Woodbury formula (Woodbury, 1950).

$$\begin{aligned}\left(\Sigma^{\mathtt{FC}}\right)^{-1} &= \left(e^{\epsilon/2}+1\right)^2\left[\mathrm{Diag}\left(\boldsymbol{q}\right) - \left(e^{\epsilon/2}-1\right)^2\boldsymbol{p}^0\left(\boldsymbol{p}^0\right)^\intercal\right]^{-1} \\ &= \left(e^{\epsilon/2}+1\right)^2\left[\mathrm{Diag}\left(\boldsymbol{q}\right)^{-1} + \left(\frac{1}{\left(e^{\epsilon/2}-1\right)^2} - \sum_{j=1}^d \frac{(p^0_j)^2}{q_j}\right)^{-1}\left(\boldsymbol{p}^0/\boldsymbol{q}\right)\left(\boldsymbol{p}^0/\boldsymbol{q}\right)^\intercal\right]\end{aligned}$$

□

8.3. Ongoing Work

As part of ongoing work, we compare the various local DP hypothesis tests to see which performs best as a function of the dimensionality of the data d and the privacy parameters. We expect that as d gets large enough that the statistic $\mathrm{T}^{\mathtt{FC}}$ will get lower Type II error than using statistic $\check{\mathrm{T}}^{(n)}_\epsilon$. Further, we hope to analyze other hypothesis tests, e.g. *analysis of variance* (ANOVA), in the local model.

Part IV

CONCLUSION

This dissertation presented some of the recent work in understanding adaptivity in data analysis. We discussed why traditional statistical inference methods break down when an analyst can select an analysis based on previous outcomes from the same dataset, perhaps as part of some exploratory study. As a potential fix, we showed how if each analysis is differentially private – or more generally has bounded max-information – then we can still correct for the adaptivity. It is then a tradeoff between the amount of noise that we incorporate to ensure validity over an entire sequence of analyses and the usefulness of each analysis: with too much noise, the analyses are essentially independent but makes the results of each analysis worthless, alternatively with no noise we run the risk of overfitting. By optimizing for the tradeoff between these two sources of error, we showed how we can take the previous results along this line of work to develop confidence intervals for a large number of adaptively selected statistical queries that outperforms traditional data-splitting techniques.

We then presented work that extended the connection of differential privacy to adaptive data analysis. Specifically, we proved that approximately differentially private algorithms have bounded max-information (when data comes from a product distributions). This allowed us to correct for adaptivity in more general types of analyses, like post-selection hypothesis testing. Further, previous results for specific types of analyses, like low-sensitivity queries, could then be proven as special cases of this connection between max-information and approximate-differential privacy. One of the main benefits of approximate differential privacy is that we can apply nice composition theorems which tells us that the privacy parameters degrade gracefully, even sublinearly with the number of analyses that are conducted – a feature that is necessary in order to show these methods outperform traditional data-splitting techniques. However, there is a caveat to applying the composition theorems, and that is that the parameters and the number of analyses need to all be fixed up front. We then provided a new framework for differential privacy composition where the parameters and number of analyses to be conducted do not need to be known upfront, prior to running any computation on the dataset.

These results linking differential privacy and adaptive data analysis are only practical if we actually have differentially private analyses that an analyst would want to use. We then presented some differentially private hypothesis tests for categorical data, specifically chi-square tests. We developed tests that ensured (theoretically or empirically) probability at most α of a false discovery with the probability of Type II error being comparable to the classical, non-private tests. In ongoing work, we hope to develop an entire suite of differentially private tools that can be used in adaptive data analysis.

There are many unanswered questions in this line of work, showing that we do not fully understand adaptivity. Some fundamental questions left include, which types of analyses can be composed without causing problems? We showed that we cannot use an algorithm with bounded description length followed by an approximately differentially private algorithm because we could output the entire dataset, thus causing us to potentially overfit on the next query. We also showed that algorithms that *robustly generalize* do not compose. If a new type of algorithm is shown to have good generalization guarantees, can we use it in sequence with other types of algorithms? It would be nice to have a general theory for why composition breaks down for different types of analyses.

Further, is there some unifying measure in adaptive data analysis? We argued that max-information partially unified some existing techniques, but then showed that compression schemes do not have bounded max-information. Is differential privacy even the right approach? Currently, there is a gap between the best lower-bounds for estimating adaptively chosen statistical queries and what we can achieve with differentially private methods from Bassily et al. (2016). We know that the connection between differential privacy and generalization is optimal, so any improvement would have to come a different method.

APPENDIX

A.1. Sensitivity of p-Values

We use this section to demonstrate that hypothesis testing is beyond the setting of statistical or low-sensitivity queries. It will be useful in our argument to use McDiarmid's inequality

Theorem A.1.1 [McDiarmid's Inequality]**.** *Let* $X_1, \cdots, X_n$ *be independent random variables with domain* $\mathcal{X}$*. Further, let* $f : \mathcal{X}^n \to \mathbb{R}$ *be a function of sensitivity* $\Delta_f > 0$*. Then for every* $\tau > 0$ *and* $\mu = \mathbb{E}\left[f(X_1, \cdots, X_n)\right]$ *we have*

$$\Pr\left[f(X_1, \cdots, X_n) - \mu \geq \tau\right] \leq \exp\left(\frac{-2\tau^2}{n\Delta^2}\right).$$

We then show that the sensitivity of the p-values for a hypothesis test is not low sensitivity enough for the existing results from Bassily et al. (2016) to give meaningful p-value corrections.

Lemma A.1.2. *Let* $\phi : \mathcal{X}^n \to \mathbb{R}$ *be a test statistic with null hypothesis* H_0*, and* $p : \mathbb{R} \to [0,1]$*, where* $p(a) = \Pr_{\boldsymbol{x} \sim \mathcal{D}^n}\left[\phi(\boldsymbol{x}) \geq a\right]$*, and* $\mathcal{D} \in \mathrm{H}_0$*. The sensitivity of* $p \circ \phi$ *must be larger than* $0.37/\sqrt{n}$*.*

Proof. Note that if $\boldsymbol{X} \sim \mathcal{D}^n$, then $p \circ \phi(\boldsymbol{X})$ is uniform on $[0,1]$, and thus, has mean $1/2$. From Theorem A.1.1, we know that if $p \circ \phi$ has sensitivity Δ, then for any $0 < \delta < 1/2$, we have:

$$\Pr\left[p \circ \phi(\boldsymbol{X}) \geq 1/2 + \Delta\sqrt{\frac{n}{2}\ln(1/\delta)}\right] \leq \delta.$$

However, we also know that $p \circ \phi(\boldsymbol{X})$ is uniform, so that

$$\Pr\left[p \circ \phi(\boldsymbol{X}) \geq 1 - \delta\right] = \delta.$$

Hence, if $\Delta < \frac{1/2 - \delta}{\sqrt{\frac{n}{2}\ln(1/\delta)}}$, we obtain a contradiction:

$$\delta \geq \Pr\left[p \circ \phi(\boldsymbol{X}) \geq 1/2 + \Delta\sqrt{\frac{n}{2}\ln(1/\delta)}\right] > \Pr\left[p \circ \phi(\boldsymbol{X}) \geq 1 - \delta\right] = \delta.$$

We then set $\delta = 0.08$ to get our stated bound on sensitivity. $\square$

Thus, the sensitivity Δ for the p-value for any test statistic and any null hypothesis must be at least $0.37/\sqrt{n}$. This is too large for the following theorem, proven in Bassily et al. (2016), to give a nontrivial guarantee:

Theorem A.1.3 [Bassily et al. (2016)]**.** *Let* $\epsilon \in (0, 1/3)$*,* $\delta \in (0, \epsilon/4)$*, and* $n \geq \frac{\log(4\epsilon/\delta)}{\epsilon^2}$*. Let* $\mathcal{Y}$ *denote the class of* Δ*-sensitive functions. Let* $\boldsymbol{X} \sim \mathcal{D}^n$ *for some distribution* $\mathcal{D}$ *over* $\mathcal{X}$*.*

There exists an algorithm $\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}$ where $\phi = \mathcal{M}(\boldsymbol{X})$ with the following guarantee:

$$\Pr_{\boldsymbol{X},\mathcal{M}} \left[|\phi(\mathcal{D}^n) - \phi(\boldsymbol{X})| \geq 18\epsilon\Delta n\right] < \frac{\delta}{\epsilon}.$$

When we attempt to apply this theorem to a p-value, we see that the error it guarantees, by Lemma A.1.2, is at least: $18\epsilon\Delta n \geq 18\epsilon(.37)\sqrt{n}$. However, the theorem is only valid for $n \geq \frac{1}{\epsilon^2} \ln\left(\frac{4\epsilon}{\delta}\right)$. Plugging this in, we see that $18\epsilon(.37)\sqrt{n} \geq 1$, which is a trivial error guarantee for p-values (which take values in $[0, 1]$).

A.2. Omitted Proofs from Chapter 3

We now present the proofs for Theorem 3.3.1 and Theorem 3.4.1. This requires going through the argument of Bassily et al. (2016) to improve the constants as much as we can via their analysis to get a decent confidence bound on k adaptively chosen statistical queries. We then present their *monitoring argument*, which amplifies the guarantee that a single query has good generalization error with constant probability so that the generalization error holds with high probability and the guarantee holds simultaneously over all queries in an adaptively chosen sequence. We replicate the analysis here in order to improve constants that appear in Theorem 3.3.1 while also applying the analysis to results in Russo and Zou (2016) to get a new accuracy guarantee via mutual information. We begin with a technical lemma which considers an algorithm $\mathcal{W}$ that takes as input a collection of s samples and outputs both an index in $[s]$ and a statistical query, where we denote $\mathcal{Q}_{SQ}$ as the set of all statistical queries $q : \mathcal{X} \to [0, 1]$ and their negation.

Lemma A.2.1 [Bassily et al. (2016)]**.** *Let $\mathcal{W} : (\mathcal{X}^n)^s \to \mathcal{Q}_{SQ} \times [s]$ be (ϵ, δ)-DP. If $\vec{\boldsymbol{X}} = (\boldsymbol{X}^{(1)}, \cdots, \boldsymbol{X}^{(s)}) \sim (\mathcal{D}^n)^s$ then*

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

The particular algorithm $\mathcal{W}$, called the *monitor*, that we use is given in Algorithm 21, which is a more general version that what appeared in Algorithm 1. We then present a series of lemmas that leads to an accuracy bound from Bassily et al. (2016).

Lemma A.2.2 [Bassily et al. (2016)]**.** *For each $\epsilon, \delta \geq 0$, if $\mathcal{M}$ is (ϵ, δ)-DP for k adaptively chosen queries from $\mathcal{Q}_{SQ}$, then for every data distribution $\mathcal{D}$ and analyst $\mathcal{A}$, the monitor $\mathcal{W}_{\mathcal{D}}[\mathcal{M}, \mathcal{A}]$ is (ϵ, δ)-DP.*

Lemma A.2.3 [Bassily et al. (2016)]**.** *If $\mathcal{M}$ fails to be (τ, β)-accurate, and $q^*(\mathcal{D}) - a^* > 0$, where a^* is the answer to q^* during the simulation ($\mathcal{A}$ can determine a^* from output (q^*, t^*)) then*

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

The following result is not stated exactly the same as in Bassily et al. (2016), but it follows

Algorithm 21 Extended Monitor $\mathcal{W}_{\mathcal{D}}[\mathcal{M}, \mathcal{A}](\vec{\boldsymbol{X}})$

```
Input: $\vec{\boldsymbol{x}} = (\boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}^{(s)}) \in (\mathcal{X}^n)^s$
  for $t \in [s]$ do
    As we outlined in Section 1.1, we simulate $\mathcal{M}(\boldsymbol{X}^{(t)})$ and $\mathcal{A}$ interacting. We write $q_{t,1}, \cdots, q_{t,k} \in \mathcal{Q}_{SQ}$ as the queries chosen by $\mathcal{A}$ and write $a_{t,1}, \cdots, a_{t,k} \in \mathbb{R}$ as the corresponding answers of $\mathcal{M}$.
  Let
    $$(j^*, t^*) = \operatorname*{argmax}_{j \in [k], t \in [s]} |q_{t,j}(\mathcal{D}) - a_{t,j}|.$$
  if $a_{t^*,j^*} - q_{t^*,j^*}(\mathcal{D}) \geq 0$ then
    $q^* \leftarrow q_{t^*,j^*}$
  else
    $q^* \leftarrow -q_{t^*,j^*}$
Output: $(q^*, t^*)$
```

the same analysis – we just do not simplify the expressions in the inequalities as they did.

Lemma A.2.4. *If $\mathcal{M}$ is (τ', β') accurate on the sample but not (τ, β)-accurate for the population, then*

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

We are now put everything together to get our result.

Proof of Theorem 3.3.1. We ultimately want a contradiction between the result given in Lemma A.2.1 and Lemma A.2.4. Thus, we want to find the parameter values that minimizes τ but satisfies the following inequality

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

We first analyze the case when we add noise $\mathrm{Lap}\left(\frac{1}{n\epsilon'}\right)$ to each query answer on the sample to preserve ϵ'-DP of each query and then use advanced composition Theorem 2.1.5 to get total privacy parameter ϵ.

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

Further, we obtain (τ', β')-accuracy on the sample from (3.2) where for $\beta' > 0$ we have

$$\tau' = \frac{\log(k/\beta')}{\epsilon' n}.$$

We then plug these values into (A.1) to get the following bound on τ

$$\tau \geq \left(\frac{1}{1-(1-\beta)^s}\right)\left(\frac{\log(k/\beta')}{\epsilon' n} + 2s\beta' + \exp\left[\left(\frac{e^{\epsilon'}-1}{e^{\epsilon'}+1}\right)\epsilon' k + \epsilon'\sqrt{2k\log(1/\delta)}\right] - 1 + s\delta\right)$$

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

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

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

$$\frac{1}{1-(1-\beta)^{\lfloor 1/\beta \rfloor}} \cdot \inf_{\rho'>0, \delta\in(0,1)} \left\{\frac{1}{n}\sqrt{1/\rho' \cdot \log(4k/\delta)} + 2\lfloor 1/\beta \rfloor \delta + \exp\left(k\rho' + 2\sqrt{k\rho'\log(\sqrt{\pi\rho' k}/\delta)}\right) - 1\right\}$$

$\square$

A.3. Omitted Proofs from Chapter 4

Before getting into the details of our analysis of Theorem 4.4.1, we present some preliminaries for linear codes.

A.3.1. Preliminaries for Linear Codes

For the current and the next subsections, we limit our scope to $\mathbb{F}_2$, i.e., the finite field with 2 elements. First, we define linear codes:
Definition A.3.1 [Linear Code]. *A code $C \subseteq \{0,1\}^n$ of length n and rank k is called linear iff it is a k dimensional linear subspace of the vector space $\mathbb{F}_2^n$. The vectors in C are called codewords.*

The minimum distance t of a linear code C is $t = \min_{\boldsymbol{c}_1, \boldsymbol{c}_2 \in C} dist_{Hamm}(\boldsymbol{c}_1, \boldsymbol{c}_2)$, where, $dist_{Hamm}(\boldsymbol{p}, \boldsymbol{q})$ denotes the Hamming distance between binary vectors $\boldsymbol{p}$ and $\boldsymbol{q}$.

We now define parity check matrices, which can be used to construct linear codes. Every linear code has a parity-check matrix corresponding to it. Thus, given a parity-check matrix, one can reconstruct the corresponding linear code.

Definition A.3.2 [Parity-check matrix]**.** *For a linear code* $C \subseteq \{0,1\}^n$ *of length* n *and rank* k*,* $H \in \{0,1\}^{(n-k)\times n}$ *is a parity-check matrix of* C *iff* H *is a matrix whose null space is* C*, i.e.,* $\boldsymbol{c} \in C$ *iff* $H\boldsymbol{c} = \boldsymbol{0}$*, where* $\boldsymbol{0}$ *represents the zero vector.*

Now, we state a theorem which shows the existence of high-rank linear codes when the minimum distance is less than half the code length:

Theorem A.3.3 [From Theorem 5.1.8 in Lint (1999)]**.** *For every* $t \in (0, \frac{n}{2})$*, there exists a linear code of rank* k *such that* $k \geq n - 3t\log(n)$*.*

Next, we will define an affine code, which is a translation of a linear code by a fixed vector in the vector space of the linear code:

Definition A.3.4 [Affine Code]**.** *Let* $C \subseteq \{0,1\}^n$ *be a linear code of length* n*, rank* k *and minimum distance* t*. For any vector* $\boldsymbol{b} \in \{0,1\}^n$*, the code defined by* $C_{\boldsymbol{a}} = \{\boldsymbol{c} + \boldsymbol{b} : \boldsymbol{c} \in C\}$*, where* $\boldsymbol{a} = H\boldsymbol{b}$*, is called an affine code.*

Lemma A.3.5. *If* C *is a linear code with parity check matrix* H *and minimum distance* t*, then the affine code* $C_{\boldsymbol{a}}$ *also has minimum distance* t*. Further, for all* $\boldsymbol{c}' \in C_{\boldsymbol{a}}$*, we have* $H\boldsymbol{c}' = \boldsymbol{a}$*.*

Proof. Let $\boldsymbol{c}' \in C_{\boldsymbol{a}}$. We know that there exists a $\boldsymbol{c} \in C$ such that

$$H\boldsymbol{c}' = H(\boldsymbol{c} + \boldsymbol{b}) = \boldsymbol{0} + H\boldsymbol{b} = \boldsymbol{a}.$$

□

Lastly, we define the concept of a Hamming ball around a point, which is helpful in understanding the point's neighborhood – i.e., the points close to it with respect to Hamming distance.

Definition A.3.6 [Hamming ball]**.** *A Hamming ball of radius* r *around a point* $\boldsymbol{p} \in \{0,1\}^n$*, denoted by* $B_r(\boldsymbol{p})$*, is the set of strings* $\boldsymbol{x} \in \{0,1\}^n$ *such that* $dist_{Hamm}(\boldsymbol{x}, \boldsymbol{p}) \leq r$*.*

The volume of a Hamming ball, denoted by $Vol(B_r)$, is independent of the point around which the ball is centered, i.e., for any point $\boldsymbol{p} \in \{0,1\}^n$:

$$Vol(B_r) = \left|B_r(\boldsymbol{p})\right| = \sum_{i=0}^{r} \left|\{\boldsymbol{x} \in \{0,1\}^n : dist_{Hamm}(\boldsymbol{x}, \boldsymbol{p}) = i\}\right| = \sum_{i=0}^{r} \binom{n}{i}. \tag{A.2}$$

A.3.2. Proof of Theorem 4.4.1

In this section, we define the mechanisms $\mathcal{M}_1$ and $\mathcal{M}_2$ from the theorem statement, and then prove our result in three parts: First, we show that the first bullet in the theorem statement directly follows from setting the parameters appropriately and from Dwork et al. (2015a). Next, we show the proof of the second bullet in two pieces. We start by showing that the algorithm $\mathcal{M}_2$ that we define is differentially private, and then, we show that the approximate max-information of $\mathcal{M}_2$ is small when its inputs are chosen independently.

Lastly, we prove the third bullet by first showing that the adaptive composition of $\mathcal{M}_1$ followed by $\mathcal{M}_2$ results in the reconstruction of the input with high probability. Subsequently, we show that such a composition has large approximate max-information.

Before we define the mechanisms $\mathcal{M}_1$ and $\mathcal{M}_2$, we must set up some notation. We fix t such that $t = \frac{8\log(1/\delta)}{\epsilon} + 1$. We know that $t \geq 33$ because $\epsilon \in (0, 1/2]$ and $\delta \in (0, 1/4]$. Now, fix an $((n-k) \times n)$ parity-check matrix H for a linear code $C \subseteq \{0,1\}^n$ of rank k over $\mathbb{F}_2$ where t is the minimum distance of C and $k = n - 3t\log n$, and let $r = n - k = 3t \log n$. We can ensure the existence of C from Theorem A.3.3.

We define the mechanisms $\mathcal{M}_1$ and $\mathcal{M}_2$ from the theorem statement in Algorithm 22 and Algorithm 23, respectively.

Brief description of $\mathcal{M}_1$*:* For any input $\boldsymbol{x} \in \mathcal{X}^n$, mechanism $\mathcal{M}_1$ returns a vector $\boldsymbol{a_x} \in \{0,1\}^r$ such that $\boldsymbol{x} \in C_{\boldsymbol{a_x}}$, where $C_{\boldsymbol{a_x}}$ is an affine code with minimum distance t. This follows as $\boldsymbol{a_x} = \mathcal{M}_1(\boldsymbol{x}) = H\boldsymbol{x}$, and from Lemma A.3.5, as $C_{\boldsymbol{a_x}} = \{\boldsymbol{c} \in \mathcal{X}^n : H\boldsymbol{c} = \boldsymbol{a_x}\}$.

Algorithm 22 First Algorithm in Lower Bound Construction: $\mathcal{M}_1$

```
Input: $\boldsymbol{x} \in \{0,1\}^n$
  Let $\boldsymbol{a_x} \leftarrow H\boldsymbol{x} \in \{0,1\}^r$ (multiplication in $\mathbb{F}_2$)
Output: $\boldsymbol{a_x}$
```

Brief description of $\mathcal{M}_2$*:* For any input $\boldsymbol{x} \in \mathcal{X}^n$ and $\boldsymbol{a} \in \{0,1\}^r$, mechanism $\mathcal{M}_2$ first computes $d_{\boldsymbol{x}}$, which is the distance of $\boldsymbol{x}$ from $f(\boldsymbol{x})$, i.e., the nearest codeword to $\boldsymbol{x}$ in code $C_{\boldsymbol{a}}$. Next, it sets $\widehat{d_{\boldsymbol{x}}}$ to be $d_{\boldsymbol{x}}$ perturbed with Laplace noise $L \sim \text{Lap}(1/\epsilon)$. It returns $f(\boldsymbol{x})$ if $\widehat{d_{\boldsymbol{x}}}$ is below a threshold $w \overset{\text{defn}}{=} \left(\frac{t-1}{4} - \frac{\log(1/\delta)}{\epsilon}\right)$, and $\perp$ otherwise.

Algorithm 23 Second Algorithm in Lower Bound Construction: $\mathcal{M}_2$

```
Input: $\boldsymbol{x} \in \{0,1\}^n$ (private) and $\boldsymbol{a} \in \{0,1\}^r$(public)
  Compute the distance of $\boldsymbol{x}$ to the nearest codeword in code $C_{\boldsymbol{a}}$.
  Let $d_{\boldsymbol{x}} = \min_{\boldsymbol{c} \in C_{\boldsymbol{a}}}(dist_{Hamm}(\boldsymbol{x}, \boldsymbol{c}))$.
  Let $f(\boldsymbol{x}) = \arg\min_{\boldsymbol{c} \in C_{\boldsymbol{a}}}(dist_{Hamm}(\boldsymbol{x}, \boldsymbol{c}))$ (breaking ties arbitrarily).
  Let $\widehat{d_{\boldsymbol{x}}} = d_{\boldsymbol{x}} + L$, where $L \sim \text{Lap}(1/\epsilon)$.
  if $\widehat{d_{\boldsymbol{x}}} < \left(\frac{t-1}{4} - \frac{\log(1/\delta)}{\epsilon}\right)$ then
    $\boldsymbol{b} \leftarrow f(\boldsymbol{x})$
  else
    $\boldsymbol{b} \leftarrow \perp$
Output: $\boldsymbol{b} \in \mathcal{Y}$
```

Now, we present the proof of our theorem.

Proof of Theorem 4.4.1, part 1. Observe that $r = O\left(\frac{\log(1/\delta)\log n}{\epsilon}\right)$ from the value assigned to t. We know that the second statement holds by the max-information bound for mechanisms with bounded description length from Dwork et al. (2015a). $\square$

Proof of Theorem 4.4.1, part 2. First, we show that $\mathcal{M}_2$ is indeed differentially private.

Lemma A.3.7. *$\mathcal{M}_2(\cdot, \boldsymbol{a})$ is (ϵ, δ)-differentially private for every $\boldsymbol{a} \in \{0,1\}^r$.*

Proof. We will prove this lemma by following the proof of Proposition 3 in Smith and Thakurta (2013). Fix any $\boldsymbol{a} \in \{0,1\}^r$. Firstly, observe that for every $\boldsymbol{x} \in \{0,1\}^n$, there are only 2 possible outputs for $\mathcal{M}_2(\boldsymbol{x}, \boldsymbol{a})$: $\perp$ or $f(\boldsymbol{x}) = \arg\min_{\boldsymbol{c}\in C_{\boldsymbol{a}}}(dist_{Hamm}(\boldsymbol{x}, \boldsymbol{c}))$. Also, $\mathcal{M}_2(\boldsymbol{x}, \boldsymbol{a}) = f(\boldsymbol{x})$ iff $\widehat{d_{\boldsymbol{x}}} = d_{\boldsymbol{x}} + L < w$ in Algorithm 23, where $d_{\boldsymbol{x}} = \min_{\boldsymbol{c}\in C_{\boldsymbol{a}}}(dist_{Hamm}(\boldsymbol{x}, \boldsymbol{c}))$ and $L \sim \text{Lap}(1/\epsilon)$.

Now, for any pair of points $\boldsymbol{x}$ and $\boldsymbol{x}'$ such that $dist_{Hamm}(\boldsymbol{x}, \boldsymbol{x}') = 1$, there are two possible cases:

1. $f(\boldsymbol{x}) \neq f(\boldsymbol{x}')$:

   In this case,

$$\pi \overset{\text{defn}}{=} \Pr\left[\mathcal{M}_2(\boldsymbol{x}, \boldsymbol{a}) = f(\boldsymbol{x})\right] = \Pr\left[\widehat{d_{\boldsymbol{x}}} < w\right] = \Pr\left[d_{\boldsymbol{x}} + L < \frac{t-1}{4} - \frac{\log(1/\delta)}{\epsilon}\right]$$
$$\leq \Pr\left[\frac{t-1}{2} + L < \frac{t-1}{4} - \frac{\log(1/\delta)}{\epsilon}\right] \leq \Pr\left[L < -\frac{\log(1/\delta)}{\epsilon}\right] \leq \delta$$

   where the first inequality follows as $f(\boldsymbol{x}) \neq f(\boldsymbol{x}')$ implies $d_{\boldsymbol{x}} > \frac{t-1}{2}$, and the last inequality follows from the tail property of the Laplace distribution. Therefore, $\Pr\left[\mathcal{M}(\boldsymbol{x}, \boldsymbol{a}) = \perp\right] = 1 - \pi$.

   Similarly, $\pi' \overset{\text{defn}}{=} \Pr\left[\mathcal{M}_2(\boldsymbol{x}', \boldsymbol{a}) = f(\boldsymbol{x}')\right] \leq \delta$, and consequently, $\Pr\left[\mathcal{M}_2(\boldsymbol{x}', \boldsymbol{a}) = \perp\right] = 1 - \pi'$.

   Thus, for any set $\mathcal{O} \subseteq \mathcal{Y}$, we can bound the following difference in terms of the total variation distance $TV(\mathcal{M}_2(\boldsymbol{x}, \boldsymbol{a}), \mathcal{M}_2(\boldsymbol{x}', \boldsymbol{a}))$

$$\begin{aligned}\left|\Pr\left[\mathcal{M}_2(\boldsymbol{x}, \boldsymbol{a}) \in \mathcal{O}\right] - \Pr\left[\mathcal{M}_2(\boldsymbol{x}', \boldsymbol{a}) \in \mathcal{O}\right]\right| &\leq TV(\mathcal{M}_2(\boldsymbol{x}, \boldsymbol{a}), \mathcal{M}_2(\boldsymbol{x}', \boldsymbol{a})) \\ &= \frac{(\pi - 0) + (\pi' - 0) + |(1-\pi) - (1-\pi')|}{2} \\ &= \frac{\pi + \pi' + |\pi' - \pi|}{2} = \max\{\pi, \pi'\} \leq \delta\end{aligned}$$

2. $f(\boldsymbol{x}) = f(\boldsymbol{x}')$:

Observe that for every $\boldsymbol{x}'' \in \{0,1\}^n$, the value of $d_{\boldsymbol{x}''}$ can change by at most 1 if exactly one coordinate is changed in $\boldsymbol{x}''$. Computing $\widehat{d_{\boldsymbol{x}''}}$ is then just an instantiation of the Laplace mechanism, given in Theorem 2.1.2. Therefore, $\widehat{d_{\boldsymbol{x}''}}$ satisfies $(\epsilon, 0)$-differential privacy. Notice that determining whether to output $f(\boldsymbol{x}) = f(\boldsymbol{x}')$ or $\perp$ is a post-processing function of the $(\epsilon, 0)$-differentially private $\widehat{d_{\boldsymbol{x}}}$, and thus, by Theorem 2.1.3, $\mathcal{M}_2(\cdot, \boldsymbol{a})$ is $(\epsilon, 0)$-differentially private for such inputs.

Therefore, from the above two cases, for any set $\mathcal{O} \subseteq \mathcal{Y}$, we have that:

$$\Pr\left[\mathcal{M}_2(\boldsymbol{x}, \boldsymbol{a}) \in \mathcal{O}\right] \leq e^{\epsilon} \Pr\left[\mathcal{M}_2(\boldsymbol{x}', \boldsymbol{a}) \in \mathcal{O}\right] + \delta.$$

Thus, we can conclude that $\mathcal{M}_2(\cdot, \boldsymbol{a})$ is (ϵ, δ)-differentially private for every $\boldsymbol{a} \in \{0,1\}^r$. $\square$

Next, we look at the outcome of $\mathcal{M}_2(\boldsymbol{X}, \boldsymbol{a})$ when $\boldsymbol{X}$ is drawn uniformly over $\mathcal{X}^n$ and $\boldsymbol{a}$ is a fixed r-bit string. Note that $\mathcal{M}_2(\boldsymbol{X}, \boldsymbol{a})$ outputs either $\perp$ or a codeword of $C_{\boldsymbol{a}}$. Thus,

$$\Pr\left[\mathcal{M}_2(\boldsymbol{X}, \boldsymbol{a}) \neq \perp\right] = \Pr\left[\widehat{d_{\boldsymbol{X}}} < w\right] = \Pr\left[d_{\boldsymbol{X}} + L < \left(\frac{t-1}{4} - \frac{\log(1/\delta)}{\epsilon}\right)\right] \tag{A.3}$$

Now, let us define the set $\mathcal{R} = \left\{\boldsymbol{x} \in \mathcal{X}^n : d_{\boldsymbol{x}} < \left(\frac{t-1}{4}\right)\right\}$. If $\left(d_{\boldsymbol{X}} + L < \left(\frac{t-1}{4} - \frac{\log(1/\delta)}{\epsilon}\right)\right)$, then either $\boldsymbol{X} \in \mathcal{R}$, or $L < -\frac{\log(1/\delta)}{\epsilon}$, or both. Thus,

$$\Pr\left[\mathcal{M}_2(\boldsymbol{X}, \boldsymbol{a}) \neq \perp\right] \leq \Pr\left[\boldsymbol{X} \in \mathcal{R}\right] + \Pr\left[L < -\frac{\log(1/\delta)}{\epsilon}\right] \tag{A.4}$$

From the tail bound of the Laplace distribution,

$$\Pr\left[L < -\frac{\log(1/\delta)}{\epsilon}\right] \leq \delta \tag{A.5}$$

Next, we will calculate the probability of $\boldsymbol{X} \in \mathcal{R}$. We then assign $s \overset{\text{defn}}{=} \frac{t-1}{4}$. Notice that as the minimum distance of $C_{\boldsymbol{a}}$ is t, the Hamming balls B_{2s} of radius $2s$ around the codewords of $C_{\boldsymbol{a}}$ are disjoint and thus we can bound the volume (defined in (A.2)) of each,

$$|C_{\boldsymbol{a}}| \cdot Vol(B_{2s}) \leq 2^n \tag{A.6}$$

Therefore,

$$\begin{aligned} |C_{\boldsymbol{a}}| \cdot Vol(B_s) &\leq \frac{2^n \cdot Vol(B_s)}{Vol(B_{2s})} = 2^n \cdot \frac{\sum\limits_{i=0}^{s} \binom{n}{i}}{\sum\limits_{j=0}^{2s} \binom{n}{j}} \\ &\leq 2^n \cdot \frac{s \cdot \binom{n}{s}}{\binom{n}{2s}} \leq 2^n \cdot \frac{s \cdot \left(\frac{ne}{s}\right)^s}{\left(\frac{n}{2s}\right)^{2s}} = 2^n s \left(\frac{4es}{n}\right)^s \end{aligned} \tag{A.7}$$

where the first inequality follows from equation (A.6), and the last inequality follows as $\left(\frac{n}{k}\right)^k \leq \binom{n}{k} \leq \left(\frac{ne}{k}\right)^k$ for $k \geq 1$ (from Appendix C.1 in Cormen et al. (2009)).

Thus,

$$\Pr\left[\boldsymbol{X} \in \mathcal{R}\right] = \frac{|C_{\boldsymbol{a}}| \cdot Vol(B_s)}{2^n} \leq s \left(\frac{4s\ e}{n}\right)^s \tag{A.8}$$

where the inequality follows from equation (A.7).

Hence,

$$\Pr\left[\mathcal{M}_2(\boldsymbol{X}, \boldsymbol{a}) \neq \perp\right] \leq s \left(\frac{4s\ e}{n}\right)^s + \delta < s \cdot 2^{-s} + \delta \tag{A.9}$$

where the first inequality follows from equations (A.4),(A.5) and (A.8), and the last inequality follows from the fact that $n > 8s\ e = 2(t-1)\ e$.

Bounding the term $s \cdot 2^{-s}$ from above, we have

$$\begin{aligned} s \cdot 2^{-s} &= \frac{t-1}{4} \cdot 2^{(1-t)/4} = \frac{2\log(1/\delta)}{\epsilon} \cdot 2^{-2\log(1/\delta)/\epsilon} \\ &= \frac{2\log(1/\delta)}{\epsilon} \cdot \delta^{2/\epsilon} = \left(\delta \log(1/\delta)\right) \left(\frac{2}{\epsilon} \cdot \delta^{(2/\epsilon)-2}\right) \delta \leq \delta \end{aligned} \tag{A.10}$$

where the inequality follows as $\delta \log(1/\delta) \leq 1$ for $\delta \in \left(0, \frac{1}{4}\right]$, and $\frac{2}{\epsilon} \cdot \delta^{(2/\epsilon)-2} \leq 1$ for $\epsilon \in \left(0, \frac{1}{2}\right], \delta \in \left(0, \frac{1}{4}\right]$. From equations (A.9) and (A.10),

$$\Pr\left[\mathcal{M}_2(\boldsymbol{X}, \boldsymbol{a}) = \perp\right] > 1 - 2\delta \tag{A.11}$$

Now, for any $\boldsymbol{x} \in \mathcal{X}^n$,

$$\log\left(\frac{\Pr\left[(\boldsymbol{X}, \mathcal{M}_2(\boldsymbol{X}, \boldsymbol{a})) = (\boldsymbol{x}, \perp)\right]}{\Pr\left[\boldsymbol{X} \otimes \mathcal{M}_2(\boldsymbol{X}, \boldsymbol{a}) = (\boldsymbol{x}, \perp)\right]}\right) = \log\left(\frac{\Pr\left[\mathcal{M}_2(\boldsymbol{X}, \boldsymbol{a}) = \perp | \boldsymbol{X} = \boldsymbol{x}\right]}{\Pr\left[\mathcal{M}_2(\boldsymbol{X}, \boldsymbol{a}) = \perp\right]}\right)$$
$$< \log\left(\frac{1}{1-2\delta}\right) \leq \log\left(\frac{1}{1-0.5}\right) = 1 \tag{A.12}$$

where the first inequality follows from equation (A.11), and the second inequality follows from the fact that $\delta \leq \frac{1}{4}$.

We then apply Lemma 4.2.2 using (A.11) and (A.12) to get,

$$I_\infty^\beta\left(\boldsymbol{X}; \mathcal{M}_2(\boldsymbol{X}, \boldsymbol{a})\right) \leq 1, \text{ for } \beta \geq 2\delta.$$

□

Proof of Theorem 4.4.1, part 3. Let us look at the outcome of $\mathcal{M}_2(\boldsymbol{x}, \mathcal{M}_1(\boldsymbol{x}))$. First, as $\boldsymbol{x} \in C_{\mathcal{M}_1(\boldsymbol{x})}$, $f(\boldsymbol{x}) = \boldsymbol{x}$ and $d_{\boldsymbol{x}} = 0$. Thus, $\mathcal{M}_2(\boldsymbol{x}, \mathcal{M}_1(\boldsymbol{x}))$ will either return $\boldsymbol{x}$ or $\perp$. Furthermore, we can show the probability of outputting $\boldsymbol{x}$ is high:

$$\Pr_{\substack{\text{coins} \\ \text{of } \mathcal{M}_2}}\left[\mathcal{M}_2(\boldsymbol{x}, \mathcal{M}_1(\boldsymbol{x})) = \boldsymbol{x}\right] = \Pr\left[\widehat{d_{\boldsymbol{x}}} < w\right]$$
$$\geq \Pr\left[\widehat{d_{\boldsymbol{x}}} < \frac{\log(1/\delta)}{\epsilon}\right] = \Pr\left[\mathrm{Lap}(1/\epsilon) < \frac{\log(1/\delta)}{\epsilon}\right] \geq 1 - \delta$$

where the first inequality follows from the fact that $\left(\frac{t-1}{4} - \frac{\log(1/\delta)}{\epsilon}\right) \geq \frac{\log(1/\delta)}{\epsilon}$, the equality after it follows since $d_{\boldsymbol{x}} = 0$, and the last inequality follows from a tail bound of the Laplace distribution. Thus, for every $\boldsymbol{x} \in \mathcal{X}^n$,

$$\Pr\left[\mathcal{M}_2(\boldsymbol{x}, \mathcal{M}_1(\boldsymbol{x})) = \boldsymbol{x}\right] \geq 1 - \delta. \tag{A.13}$$

Consider the event $\mathcal{E} \overset{\text{defn}}{=} \{(\boldsymbol{x}, \boldsymbol{x}) : \boldsymbol{x} \in \mathcal{X}^n\}$. From equation (A.13),

$$\Pr\left[(\boldsymbol{X}, \mathcal{M}_2(\boldsymbol{X}, \mathcal{M}_1(\boldsymbol{X}))) \in \mathcal{E}\right] \geq 1 - \delta. \tag{A.14}$$

Also, for $\boldsymbol{b} \in \mathcal{Y}$, if $\boldsymbol{b} = \perp$, then $\Pr\left[\boldsymbol{X} = \boldsymbol{b}\right] = 0$, and if $\boldsymbol{b} \in \mathcal{X}^n$, then $\Pr\left[\boldsymbol{X} = \boldsymbol{b}\right] = 2^{-n}$ as $\boldsymbol{X}$ is drawn uniformly over $\mathcal{X}^n$. Thus, for all $\boldsymbol{b} \in \mathcal{Y}$,

$$\Pr\left[(\boldsymbol{X}, \boldsymbol{b}) \in \mathcal{E}\right] \leq 2^{-n}.$$

Hence,

$$\begin{aligned}
&\Pr\left[(\boldsymbol{X} \otimes \mathcal{M}_2(\boldsymbol{X}, \mathcal{M}_1(\boldsymbol{X}))) \in \mathcal{E}\right] \\
&\quad = \sum_{\boldsymbol{b} \in \mathcal{Y}} \Pr\left[(\boldsymbol{X}, \boldsymbol{b}) \in \mathcal{E}\right] \Pr\left[\mathcal{M}(\boldsymbol{X}, \mathcal{M}(\boldsymbol{X})) = \boldsymbol{b}\right] \leq 2^{-n} \qquad \text{(A.15)}
\end{aligned}$$

Therefore, for $\beta \leq \frac{1}{2} - \delta$,

$$\begin{aligned}
I_{\infty}^{\beta}\left(\boldsymbol{X}; \mathcal{M}_2(\boldsymbol{X}, \mathcal{M}_1(\boldsymbol{X}))\right) &= \log\left(\max_{\substack{\mathcal{O} \subseteq (\mathcal{X} \times \mathcal{Y}), \\ \Pr[(\boldsymbol{X}, \mathcal{M}_2(\boldsymbol{X}, \mathcal{M}_1(\boldsymbol{X}))) \in \mathcal{O}] > \beta}} \frac{\Pr\left[(\boldsymbol{X}, \mathcal{M}_2(\boldsymbol{X}, \mathcal{M}_1(\boldsymbol{X}))) \in \mathcal{O}\right] - \beta}{\Pr\left[(\boldsymbol{X} \otimes \mathcal{M}_2(\boldsymbol{X}, \mathcal{M}_1(\boldsymbol{X}))) \in \mathcal{O}\right]} \right) \\
&\geq \log\left(\frac{\Pr\left[(\boldsymbol{X}, \mathcal{M}_2(\boldsymbol{X}, \mathcal{M}_1(\boldsymbol{X}))) \in \mathcal{E}\right] - \beta}{\Pr\left[(\boldsymbol{X} \otimes \mathcal{M}_2(\boldsymbol{X}, \mathcal{M}_1(\boldsymbol{X}))) \in \mathcal{E}\right]} \right) \\
&\geq \log\left(\frac{1 - \delta - \beta}{2^{-n}} \right) \\
&= n + \log(1 - \delta - \beta) \geq n - 1
\end{aligned}$$

where the first inequality follows from equation (A.14) and as $(1 - \delta) > \beta$, the second inequality follows from equations (A.14) and (A.15), and the last inequality follows from the fact that $\beta \leq \frac{1}{2} - \delta$. $\square$

BIBLIOGRAPHY

M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, pages 308–318, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4139-4. doi: 10.1145/2976749.2978318. URL `http://doi.acm.org/10.1145/2976749.2978318`.

R. Bassily and A. Smith. Local, private, efficient protocols for succinct histograms. In *Proceedings of the Forty-seventh Annual ACM Symposium on Theory of Computing*, STOC '15, pages 127–135, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3536-2. doi: 10.1145/2746539.2746632. URL `http://doi.acm.org/10.1145/2746539.2746632`.

R. Bassily, K. Nissim, A. D. Smith, T. Steinke, U. Stemmer, and J. Ullman. Algorithmic stability for adaptive data analysis. In *Proceedings of the 48th Annual ACM on Symposium on Theory of Computing, STOC*, 2016.

Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995. ISSN 00359246. doi: 10.2307/2346101. URL `http://dx.doi.org/10.2307/2346101`.

B. J. Berry. City size distributions and economic development. *Economic development and cultural change*, pages 573–588, 1961.

Y. M. M. Bishop, S. E. Fienberg, and P. W. Holland. Discrete multivariate analysis: Theory and practice, 1975.

A. Blair, P. Decoufle, and D. Grauman. Causes of death among laundry and dry cleaning workers. *American journal of public health*, 69(5):508–511, 1979.

A. Blum and M. Hardt. The ladder: A reliable leaderboard for machine learning competitions. In D. Blei and F. Bach, editors, *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 1006–1014. JMLR Workshop and Conference Proceedings, 2015. URL `http://jmlr.org/proceedings/papers/v37/blum15.pdf`.

C. E. Bonferroni. Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8:3–62, 1936.

O. Bousquet and A. Elisseeff. Stability and generalization. *J. Mach. Learn. Res.*, 2:499–526, Mar. 2002. ISSN 1532-4435. doi: 10.1162/153244302760200704. URL `http://dx.doi.org/10.1162/153244302760200704`.

M. Bun and T. Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography - 14th International Conference, TCC 2016-B, Beijing, China, October 31 - November 3, 2016, Proceedings, Part I*, pages

635–658, 2016. doi: 10.1007/978-3-662-53641-4_24. URL `http://dx.doi.org/10.1007/978-3-662-53641-4_24`.

S. Chen, Z. Wang, W. Xu, and Y. Miao. Exponential inequalities for self-normalized martingales. *Journal of Inequalities and Applications*, 2014(1):289, 2014. ISSN 1029-242X. doi: 10.1186/1029-242X-2014-289. URL `http://dx.doi.org/10.1186/1029-242X-2014-289`.

T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms, Third Edition*. The MIT Press, 3rd edition, 2009. ISBN 0262033844, 9780262033848.

R. Cummings, M. Kearns, A. Roth, and Z. S. Wu. Privacy and truthful equilibrium selection for aggregative games. In *Proceedings of the 11th International Conference on Web and Internet Economics - Volume 9470*, WINE 2015, pages 286–299, New York, NY, USA, 2015. Springer-Verlag New York, Inc. ISBN 978-3-662-48994-9. doi: 10.1007/978-3-662-48995-6_21. URL `http://dx.doi.org/10.1007/978-3-662-48995-6_21`.

R. Cummings, K. Ligett, K. Nissim, A. Roth, and Z. S. Wu. Adaptive learning with robust generalization guarantees. *arXiv preprint arXiv:1602.07726*, 2016a.

R. Cummings, K. Ligett, J. Radhakrishnan, A. Roth, and Z. S. Wu. Coordination complexity: Small information coordinating large populations. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, ITCS '16, pages 281–290, New York, NY, USA, 2016b. ACM. ISBN 978-1-4503-4057-1. doi: 10.1145/2840728.2840767. URL `http://doi.acm.org/10.1145/2840728.2840767`.

T. David and S. Beards. Asthma and the month of birth. *Clinical & Experimental Allergy*, 15(4):391–395, 1985.

A. De. Lower bounds in differential privacy. In *Proceedings of the 9th International Conference on Theory of Cryptography*, TCC'12, pages 321–338, Berlin, Heidelberg, 2012. Springer-Verlag. ISBN 978-3-642-28913-2. doi: 10.1007/978-3-642-28914-9_18.

V. H. de la Peña, M. J. Klass, and T. Leung Lai. Self-normalized processes: exponential inequalities, moment bounds and iterated logarithm laws. *Ann. Probab.*, 32(3):1902–1933, 07 2004. doi: 10.1214/009117904000000397. URL `http://dx.doi.org/10.1214/009117904000000397`.

I. Dinur and K. Nissim. Revealing information while preserving privacy. In *Proceedings of the Twenty-second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '03, pages 202–210, New York, NY, USA, 2003. ACM. ISBN 1-58113-670-6. doi: 10.1145/773153.773173. URL `http://doi.acm.org/10.1145/773153.773173`.

O. J. Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64, 1961. ISSN 01621459. URL `http://www.jstor.org/stable/2282330`.

C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

C. Dwork and G. N. Rothblum. Concentrated differential privacy. *CoRR*, abs/1603.01887, 2016. URL `http://arxiv.org/abs/1603.01887`.

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

C. Dwork, F. Mcsherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *In Proceedings of the 3rd Theory of Cryptography Conference*, pages 265–284. Springer, 2006b.

C. Dwork, G. N. Rothblum, and S. P. Vadhan. Boosting and differential privacy. In *51th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2010, October 23-26, 2010, Las Vegas, Nevada, USA*, pages 51–60, 2010. doi: 10.1109/FOCS.2010.12. URL `http://dx.doi.org/10.1109/FOCS.2010.12`.

C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth. Generalization in adaptive data analysis and holdout reuse. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, R. Garnett, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2341–2349. Curran Associates, Inc., 2015a. URL `http://papers.nips.cc/paper/5993-generalization-in-adaptive-data-analysis-and-holdout-reuse.pdf`.

C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015b. doi: 10.1126/science.aaa9375. URL `http://www.sciencemag.org/content/349/6248/636.abstract`.

C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. L. Roth. Preserving statistical validity in adaptive data analysis. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, STOC '15, pages 117–126, New York, NY, USA, 2015c. ACM. ISBN 978-1-4503-3536-2. doi: 10.1145/2746539.2746580. URL `http://doi.acm.org/10.1145/2746539.2746580`.

C. Dwork, W. Su, and L. Zhang. Private false discovery rate control. *arXiv preprint arXiv:1511.03803*, 2015d.

H. Ebadi and D. Sands. Featherweight PINQ. *CoRR*, abs/1505.02642, 2015. URL `http://arxiv.org/abs/1505.02642`.

H. R. F. Ebaugh and C. A. Haney. Church attendance and attitudes toward abortion:

Differentials in liberal and conservative churches. *Journal for the Scientific Study of Religion*, pages 407–413, 1978.

T. Ferguson. *A Course in Large Sample Theory*. Chapman & Hall Texts in Statistical Science Series. Taylor & Francis, 1996. ISBN 9780412043710.

S. E. Fienberg, A. Rinaldo, and X. Yang. Differential privacy and the risk-utility tradeoff for multi-dimensional contingency tables. In *Proceedings of the 2010 International Conference on Privacy in Statistical Databases*, PSD'10, pages 187–199, Berlin, Heidelberg, 2010. Springer-Verlag.

W. Fithian, D. Sun, and J. Taylor. Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*, 2014.

M. Gaboardi, H. Lim, R. M. Rogers, and S. P. Vadhan. Differentially private chi-squared hypothesis testing: Goodness of fit and independence testing. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pages 2111–2120, 2016. URL `http://jmlr.org/proceedings/papers/v48/rogers16.html`.

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

J. C. Gill, J. Endres-Brooks, P. J. Bauer, W. J. Marks Jr, and R. R. Montgomery. The effect of abo blood group on the diagnosis of von willebrand disease. *Blood*, 69(6):1691–1695, 1987.

W. A. Glaser. The family and voting turnout. *Public Opinion Quarterly*, 23(4):563–570, 1959.

A. G. Greenwald, C. G. Carnot, R. Beach, and B. Young. Increasing voting behavior by asking people if they expect to vote. *Journal of Applied Psychology*, 72(2):315, 1987.

W. C. Guenther. Power and sample size for approximate chi-square tests. *The American Statistician*, 31(2):83–85, 1977.

M. Hardt and G. N. Rothblum. A multiplicative weights mechanism for privacy-preserving data analysis. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 61–70, Oct 2010. doi: 10.1109/FOCS.2010.85.

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

N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D. A. Stephan, S. F. Nelson, and D. W. Craig. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS Genet*, 4(8), 08 2008.

J. Hsu, J. Morgenstern, R. Rogers, A. Roth, and R. Vohra. Do prices coordinate markets? In *Proceedings of the Forty-eighth Annual ACM Symposium on Theory of Computing*, STOC '16, pages 440–453, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4132-5. doi: 10.1145/2897518.2897559. URL `http://doi.acm.org/10.1145/2897518.2897559`.

J. P. Imhof. Computing the distribution of quadratic forms in normal variables. *Biometrika*, 48(3-4):419–426, 1961.

J. P. A. Ioannidis. Why most published research findings are false. *PLOS Medicine*, 2 (8), 08 2005. doi: 10.1371/journal.pmed.0020124. URL `http://dx.doi.org/10.1371%2Fjournal.pmed.0020124`.

M. F. D. John G. Reid. An accessible proof of craig's theorem in the noncentral case. *The American Statistician*, 42(2):139–142, 1988. ISSN 00031305. URL `http://www.jstor.org/stable/2684489`.

A. Johnson and V. Shmatikov. Privacy-preserving data exploration in genome-wide association studies. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 1079–1087, New York, NY, USA, 2013. ACM.

P. Kairouz, S. Oh, and P. Viswanath. The composition theorem for differential privacy. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 1376–1385, 2015.

S. Kannan, J. Morgenstern, A. Roth, and Z. S. Wu. Approximately stable, school optimal, and student-truthful many-to-one matchings (via differential privacy). In *Proceedings of the Twenty-sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '15, pages 1890–1903, Philadelphia, PA, USA, 2015. Society for Industrial and Applied Mathematics. URL `http://dl.acm.org/citation.cfm?id=2722129.2722255`.

V. Karwa and A. Slavković. Differentially private graphical degree sequences and synthetic graphs. In J. Domingo-Ferrer and I. Tinnirello, editors, *Privacy in Statistical Databases*, volume 7556 of *Lecture Notes in Computer Science*, pages 273–285. Springer Berlin Heidelberg, 2012.

V. Karwa and A. Slavković. Inference using noisy degrees: Differentially private beta-model and synthetic graphs. *The Annals of Statistics*, 44(1):87–112, 2016.

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

M. Kearns. Efficient noise-tolerant learning from statistical queries. In *Proceedings of the Twenty-fifth Annual ACM Symposium on Theory of Computing*, STOC '93, pages 392–401, New York, NY, USA, 1993. ACM. ISBN 0-89791-591-7. doi: 10.1145/167088.167200. URL `http://doi.acm.org/10.1145/167088.167200`.

M. Kearns, M. M. Pai, R. M. Rogers, A. Roth, and J. Ullman. Robust mediators in large games. *CoRR*, abs/1512.02698, 2015. URL `http://arxiv.org/abs/1512.02698`.

D. Kifer and R. Rogers. A New Class of Private Chi-Square Tests. *ArXiv e-prints*, Oct. 2016.

M. Krain and M. E. Myers. Democracy and civil war: A note on the democratic peace proposition. *International Interactions*, 23(1):109–118, 1997.

J. H. Kuklinski and D. M. West. Economic expectations and voting behavior in united states house and senate elections. *American Political Science Review*, 75(02):436–447, 1981.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. A Series of Modern Surveys in Mathematics Series. Springer, 1991. ISBN 9783540520139. URL `https://books.google.com/books?id=cyKYDfvxRjsC`.

J. Lee, Y. Wang, and D. Kifer. Maximum likelihood postprocessing for differential privacy under consistency constraints. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 635–644, New York, NY, USA, 2015. ACM.

J. D. Lee, D. L. Sun, Y. Sun, and J. E. Taylor. Exact post-selection inference, with application to the lasso. *arXiv preprint arXiv:1311.6238*, 2013.

J. H. v. Lint. *Introduction to Coding Theory*. Springer, Berlin, 3rd edition, 1999. ISBN 9783540641339.

N. Littlestone and M. K. Warmuth. Relating data compression and learnability. Technical report, 1986.

T. Lykouris, V. Syrgkanis, and E. Tardos. Learning and efficiency in games with dynamic population. In *Proceedings of the Twenty-seventh Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '16, pages 120–129, Philadelphia, PA, USA, 2016. Society for Industrial and Applied Mathematics. ISBN 978-1-611974-33-1. URL `http://dl.acm.org/citation.cfm?id=2884435.2884444`.

A. McGregor, I. Mironov, T. Pitassi, O. Reingold, K. Talwar, and S. P. Vadhan. The limits of two-party differential privacy. *Electronic Colloquium on Computational Complexity (ECCC)*, 18:106, 2011. URL `http://eccc.hpi-web.de/report/2011/106`.

F. McSherry and K. Talwar. Mechanism design via differential privacy. In *Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, Providence, RI, October 2007. IEEE. URL `https://www.microsoft.com/en-us/research/publication/mechanism-design-via-differential-privacy/`.

R. C. Meng and D. G. Chapman. The power of chi square tests for contingency tables. *Journal of the American Statistical Association*, 61(316):965–975, 1966.

N. J. Mitchell and J. M. McCormick. Economic and political explanations of human rights violations. *World Politics*, 40(04):476–498, 1988.

S. Mitra. *Contributions to the Statistical Analysis of Categorical Data.* Institute of Statistics mimeo series. 1955.

S. Mitra. On the limiting power function of the frequency chi-square test. *Ann. Math. Statist.*, 29(4):1221–1233, 12 1958.

A. A. Mohsenipour. *On the Distribution of Quadratic Expressions in Various Types of Random Vectors.* PhD thesis, The University of Western Ontario, Electronic Thesis and Dissertation Repository, 12 2012. Paper 955.

S. Mukherjee, P. Niyogi, T. Poggio, and R. Rifkin. Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics*, 25(1):161–193, 2006. ISSN 1572-9044. doi: 10.1007/s10444-004-7634-z. URL `http://dx.doi.org/10.1007/s10444-004-7634-z`.

J. Murtagh and S. P. Vadhan. The complexity of computing the optimal composition of differential privacy. In *Theory of Cryptography - 13th International Conference, TCC 2016-A, Tel Aviv, Israel, January 10-13, 2016, Proceedings, Part I*, pages 157–175, 2016.

A. Nikolov, K. Talwar, and L. Zhang. The geometry of differential privacy: The sparse and approximate cases. In *Proceedings of the Forty-fifth Annual ACM Symposium on Theory of Computing*, STOC '13, pages 351–360, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2029-0. doi: 10.1145/2488608.2488652. URL `http://doi.acm.org/10.1145/2488608.2488652`.

K. B. Petersen and M. S. Pedersen. The matrix cookbook, nov 2012. URL `http://www2.imm.dtu.dk/pubdb/p.php?3274`. Version 20121115.

T. Poggio, R. Rifkin, S. Mukherjee, and P. Niyogi. General conditions for predictivity in learning theory. *Nature*, 428(6981):419–422, 03 2004. URL `http://dx.doi.org/10.1038/nature02341`.

S. Raskhodnikova, A. Smith, H. K. Lee, K. Nissim, and S. P. Kasiviswanathan. What can we learn privately? *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, 00:531–540, 2008. ISSN 0272-5428. doi: doi.ieeecomputersociety.org/10.1109/FOCS.2008.27.

A. Renyi. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 547–561, Berkeley, Calif., 1961. University of California Press. URL `http://projecteuclid.org/euclid.bsmsp/1200512181`.

R. Rogers, A. Roth, J. Ullman, and Z. S. Wu. Inducing approximately optimal flow using truthful mediators. In *Proceedings of the Sixteenth ACM Conference on Economics and*

Computation, EC '15, pages 471–488, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3410-5. doi: 10.1145/2764468.2764509. URL `http://doi.acm.org/10.1145/2764468.2764509`.

R. Rogers, A. Roth, A. Smith, and O. Thakkar. Max-information, differential privacy, and post-selection hypothesis testing. In *Foundations of Computer Science (FOCS)*, 2016a. arXiv:1604.03924 [cs.LG].

R. M. Rogers, S. P. Vadhan, A. Roth, and J. Ullman. Privacy odometers and filters: Pay-as-you-go composition. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 1921–1929, 2016b. URL `http://papers.nips.cc/paper/6170-privacy-odometers-and-filters-pay-as-you-go-composition`.

A. Roth and T. Roughgarden. Interactive privacy via the median mechanism. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 765–774. ACM, 2010.

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

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

S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Learnability, stability and uniform convergence. *J. Mach. Learn. Res.*, 11:2635–2670, Dec. 2010. ISSN 1532-4435. URL `http://dl.acm.org/citation.cfm?id=1756006.1953019`.

O. Sheffet. Differentially private least squares: Estimation, confidence and rejecting the null hypothesis. *arXiv preprint arXiv:1507.02482*, 2015a.

O. Sheffet. Differentially private least squares: Estimation, confidence and rejecting the null hypothesis. *arXiv preprint arXiv:1507.02482*, 2015b.

J. P. Simmons, L. D. Nelson, and U. Simonsohn. False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, Oct. 2011. doi: 10.1177/0956797611417632. URL `http://pss.sagepub.com/lookup/doi/10.1177/0956797611417632`.

S. Simmons, C. Sahinalp, and B. Berger. Enabling privacy-preserving {GWASs} in heterogeneous human populations. *Cell Systems*, 3(1):54 – 61, 2016.

A. Smith. Privacy-preserving statistical estimation with optimal convergence rates. In *Proceedings of the Forty-third Annual ACM Symposium on Theory of Computing*, STOC '11, pages 813–822, New York, NY, USA, 2011. ACM.

A. Smith and A. Thakurta. Differentially private feature selection via stability arguments, and the robustness of the lasso. In *COLT 2013 - The 26th Annual Conference on Learning*

Theory, June 12-14, 2013, Princeton University, NJ, USA, pages 819–850, 2013. URL `http://jmlr.org/proceedings/papers/v30/Guha13.html`.

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

M. Triola. *Essentials of Statistics*. Pearson Education, 2014. ISBN 9780321924636. URL `https://books.google.com/books?id=QZN-AgAAQBAJ`.

C. Uhler, A. Slavkovic, and S. E. Fienberg. Privacy-preserving data sharing for genome-wide association studies. *Journal of Privacy and Confidentiality*, 5(1), 2013.

S. A. van de Geer. *On Hoeffding's inequality for dependent random variables*. Springer, 2002.

T. van Erven and P. Harremos. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, July 2014. ISSN 0018-9448. doi: 10.1109/TIT.2014.2320500.

D. Vu and A. Slavković. Differential privacy for clinical trial data: Preliminary evaluations. In *Proceedings of the 2009 IEEE International Conference on Data Mining Workshops*, ICDMW '09, pages 138–143, Washington, DC, USA, 2009. IEEE Computer Society.

Y. Wang, J. Lee, and D. Kifer. Differentially private hypothesis testing, revisited. *arXiv preprint arXiv:1511.03376*, 2015.

Y. Wang, J. Lei, and S. E. Fienberg. A minimax theory for adaptive data analysis. *CoRR*, abs/1602.04287, 2016. URL `http://arxiv.org/abs/1602.04287`.

S. L. Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60:63–69, 1965.

L. Wasserman and S. Zhou. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389, 2010.

R. L. Wasserstein and N. A. Lazar. The asa's statement on p-values: context, process, and purpose. *The American Statistician*, 0(ja):00–00, 2016. doi: 10.1080/00031305.2016.1154108. URL `http://dx.doi.org/10.1080/00031305.2016.1154108`.

M. A. Woodbury. *Inverting Modified Matrices*. Number 42 in Statistical Research Group Memorandum Reports. Princeton University, Princeton, NJ, 1950.

F. Yu, S. E. Fienberg, A. B. Slavkovic, and C. Uhler. Scalable privacy-preserving data sharing methodology for genome-wide association studies. *Journal of Biomedical Informatics*, 50:133–141, 2014.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.