THE FUNDAMENTAL LIMITS OF STATISTICAL DATA PRIVACY

BY

PETER KAIROUZ

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2016

Urbana, Illinois

Doctoral Committee:

Professor Pramod Viswanath, Chair
Associate Professor Nikita Borisov
Professor Bruce Hajek
Assistant Professor Sewoong Oh
Professor Rayadurgam Srikant

ABSTRACT

The Internet is shaping our daily lives. On the one hand, social networks like Facebook and Twitter allow people to share their precious moments and opinions with virtually anyone around the world. On the other, services like Google, Netflix, and Amazon allow people to look up information, watch movies, and shop online anytime, anywhere. However, with this unprecedented level of connectivity comes the danger of being monitored. There is an increasing tension between the need to share data and the need to preserve the privacy of Internet users. The need for privacy appears in three main contexts: (1) the global privacy context, as in when private companies and public institutions release personal information about individuals to the public; (2) the local privacy context, as in when individuals disclose their personal information with potentially malicious service providers; (3) the multi-party privacy context, as in when different parties cooperate to interactively compute a function that is defined over all the parties' data.

Differential privacy has recently surfaced as a strong measure of privacy in all three contexts. Under differential privacy, privacy is achieved by randomizing the data before releasing it. This leads to a fundamental tradeoff between privacy and utility. In this thesis, we take a concrete step towards understanding the fundamental structure of privacy mechanisms that achieve the best privacy-utility tradeoff. This tradeoff is formulated as a constrained optimization problem: maximize utility subject to differential privacy constraints. We show, perhaps surprisingly, that in all three privacy contexts, the optimal privacy mechanisms have the same combinatorial staircase structure. This deep result is a direct consequence of the geometry of the constraints imposed by differential privacy on the privatization mechanisms.

“All I am or ever hope to be, I owe to my lovely mother.”– Abraham Lincoln

To my mother, for her love and support.

ACKNOWLEDGMENTS

First and foremost, I would like to express my sincerest appreciation to my mother for her infinite love and support. Without her, I would not have been the successful person I am today.

I would also like to express my gratitude to my advisors, Professors Pramod Viswanath and Sewoong Oh. Their patience, guidance, and support helped me tremendously in the research that I conducted during my PhD. More importantly, their openness to new ideas and continuous engagement helped me become a better researcher and shaped my research vision. I recognize that this work would not have been possible without their unparalleled support. Thank you so much for everything you have done for me.

A very big thank-you goes to all my research mentors. Specifically, I would like to thank my PhD committee members, Professors Bruce Hajek, Rayadurgam Srikant, and Nikita Borisov. Their comments and feedback contributed to the shaping of this work. I would also like to thank my amazing research collaborators: Professor Kannan Ramchandran and Giulia Fanti of UC Berkeley; Daniel Ramage, Brendan McMahan, and Keith Bonawitz of Google Research; Ahmed Sadek, Tamer Kadous, and Kambiz Yazdi of Qualcomm Research. I learned a lot from each and every one of you.

A very special thank-you goes to all my colleagues and friends at UIUC. In particular, I would like to thank Ihab Nahlus, Marc Ghossoub, Shaileshh BV, Charbel Sakr, Tarek Sakakini, Adel Ejjeh, Aolin Xu, Eric Kim, Andrew Bean, and Thomas Riedl. Our stimulating and exciting discussions were balanced between research interests and personal pursuits. Together, we learned a lot about religion, science, philosophy, politics, and food.

Finally, I am thankful to everyone who helped with and contributed to the development of this work: my graduate and undergraduate professors, friends, and family members. Your continuous encouragement and constant support were an everlasting source of inspiration.

TABLE OF CONTENTS

CHAPTER 1 INTRODUCTION 1
1.1 The Fundamental Limits of Differential Privacy 2
1.2 Outline and Contributions 3

CHAPTER 2 GLOBAL DIFFERENTIAL PRIVACY 6
2.1 Introduction 6
2.2 Operational Interpretation of Differential Privacy 6
2.3 Optimal Mechanisms for Differential Privacy 10
2.4 The Composition Theorem in Differential Privacy 18
2.5 Conclusion and Summary 23

CHAPTER 3 LOCAL DIFFERENTIAL PRIVACY 24
3.1 Introduction 24
3.2 Operational Interpretation of Local Differential Privacy 31
3.3 Optimal Mechanisms for Local Differential Privacy 32
3.4 Private Hypothesis Testing 37
3.5 Information Preservation 47
3.6 Approximate Local Differential Privacy 53
3.7 Conclusions and Summary 58

CHAPTER 4 MULTI-PARTY DIFFERENTIAL PRIVACY 59
4.1 Introduction 59
4.2 Private Multi-Party Computation 63
4.3 Differentially Private Multi-Party Computation 66
4.4 Optimal Mechanisms for Multi-Party Differential Privacy . . . 69
4.5 Private Multi-Party XOR Computation 69
4.6 Generalization to Multiple Bits 70
4.7 Conclusions and Summary 72

CHAPTER 5 CONCLUSION 73
5.1 Global Privacy 73
5.2 Local Privacy 73
5.3 Multi-Party Privacy 75

APPENDIX A PROOFS FOR GLOBAL DIFFERENTIAL PRIVACY 77
A.1 Operational Interpretation of Differential Privacy 77
A.2 Optimal Mechanisms for Differential Privacy 79
A.3 Composition Theorem in Differential Privacy 102

APPENDIX B PROOFS FOR LOCAL DIFFERENTIAL PRIVACY 114
B.1 Operational Interpretation of Differential Privacy 114
B.2 Proof of Theorem 3.2.1 114
B.3 Optimal Mechanisms for Differential Privacy 115
B.4 Private Hypothesis Testing 122
B.5 Information Preservation 133
B.6 Approximate Local Differential Privacy 145

APPENDIX C PROOFS FOR MULTI-PARTY DIFFERENTIAL PRIVACY 146
C.1 Proof of Main Result 146
C.2 Proof of Optimal Multi-Party XOR Computation 148

REFERENCES 150

CHAPTER 1

INTRODUCTION

Privacy is a fundamental individual right. Traditionally, individual information access was limited and corresponding privacy violations were essentially local, both temporally and geographically. In the era of big data, massive amounts of data about individuals are collected both voluntarily and involuntarily. With the ready ability to search for information and correlate it across distinct sources, privacy violation takes on an ominous note in this information age.

Classical approaches to providing privacy guarantees involve anonymizing user information. While this seems to be a reasonable approach to protect the privacy of individuals, it is not invulnerable to correlation attacks: by correlating the anonymized database with another (perhaps publicly available) deanonymized database, a user's privacy could still be divulged. Early work in 1997 by Sweeney [1] demonstrated such an attack by correlating anonymized health records released by the state of Massachusetts with voter registration records. Similar deanonymization attacks have been routinely conducted in the ensuing years, despite the adoption of more sophisticated anonymization strategies [2]: AOL search logs (reported by NYTimes in 2006), Netflix collaborative filtering contest [3], Kaggle recommender system contest of Flickr data [4], and surname inference from genome datasets [5] are instances that have received widespread attention. While correlation attacks using currently available databases are already devastating for anonymization techniques, an even larger issue is that anonymization is susceptible to future data releases.

A way out of the limitations of anonymization is to release randomized data. We refer to this privatization method as statistical data privacy. Under statistical data privacy, the introduced randomness guarantees that upon observing the released data, no one should be able to learn any sensitive information about an individual. Indeed, statistical data privacy is robust to adversarial side information: any adversary cannot learn much beyond whatever side information she has access to. This thesis explores a relatively recent notion of statistical data privacy called differential privacy [6, 7, 8, 9]. At a high level, differential privacy imposes mathematical constraints on the shape and amount of noise that is added to the raw data. The differential privacy framework is very general and provides strong plausible deniability guarantees - worst case over present auxiliary information and future discoveries. With the plausible deniability guarantees of differential privacy, data could be made more widely available leading to faster and more accurate data analytics.

1.1 The Fundamental Limits of Differential Privacy

One of the main limitations of differential privacy in practice is that current approaches make the released database “too random”, thus making the data released essentially useless. Despite a decade of research efforts, many fundamental questions in differential privacy are still left open. The lack of theoretical understanding of the fundamental tradeoffs in differential privacy has led to a widespread practice of using coarse methods that are strictly sub-optimal. For instance, the Laplacian noise adding mechanism, featured in the vast majority of the literature on differential privacy, is often used without any clear justification.

It is of fundamental interest to characterize privacy mechanisms that randomize “just enough” to keep the released data as true as possible, providing maximal utility. In this thesis, we address the following important question: for a given application and a fixed privacy level, what is the best privacy preserving mechanism that maximizes the utility of the application while achieving the desired privacy level? In specific, we study the fundamental limits of differential privacy in three main contexts.

The global context

In the global context, trusted service providers or institutions want to release sensitive information about individuals. For instance, the National Institutes of Health (NIH) might be interested in releasing medical records so that researchers can find the causes and cures of certain diseases. This information is clearly sensitive and should be privatized carefully prior to its release. In this context, differential privacy provides a formal guarantee on the anonymity level of an individual user with respect to a data release.

The local context

In the local context, data providers want to share their personal data with a potentially malicious service provider. For instance, Android users might want to share their Android keyboard activities (clicks, swipes, chats, etc.) with Google so that they can benefit from improved services (e.g., auto-completion, next word prediction, etc.). However, the users are worried that Google can learn a lot of personal information about them by analyzing the data that it collects. In this context, differential privacy ensures that the service provider (Google) can only learn aggregate information about individuals.

The multi-party context

In the multi-party context, individuals interact to compute a joint function on their private data. For instance, the individuals might be interested in computing their average salary, height, or weight. In this context, differential privacy allows the users to interactively compute the function while preventing them from learning the each other's information.

1.2 Outline and Contributions

An outline of the thesis is as follows.

Chapter 2: Global Differential Privacy

Chapter 2 studies global differential privacy. We start by providing an operational interpretation of global differential privacy. Specifically, we show that differential privacy guarantees that the probabilities of false alarm and missed detection of any binary hypothesis testing problem involving the presence/absence of a user's data in a released database query cannot be simultaneously small. We then derive the optimal privacy mechanism for one and two dimensional real-valued database queries under a universal utility-maximization framework. Precisely, we show that a simple noise adding mechanism with a staircase distribution achieves the best utility-privacy tradeoff. We conclude Chapter 2 by studying the impact of sequential querying of differentially private mechanisms. In particular, we characterize how the overall privacy level degrades under the composition of differentially private mechanisms. Our solution is fundamental: we prove an upper bound on the overall privacy level and construct a sequence of privatization mechanisms that achieves this bound.

Chapter 3: Local Differential Privacy

Chapter 3 investigates local differential privacy. Similar to Chapter 2, we start by providing an operational definition of local differential privacy. We then uncover the combinatorial structure of the family of optimal privatization mechanisms for a broad class of information theoretic utility functions such as mutual information and f-divergences. Surprisingly, we show that, similar to the global privacy context, the optimal privacy mechanisms in the local privacy context have the same staircase shape. We also prove that for a given utility function and a fixed privacy level, solving the privacy-utility maximization problem is equivalent to solving a finite-dimensional linear program, the outcome of which is the optimal privatization mechanism. However, solving this linear program can be computationally expensive since it has a number of variables that is exponential in the size of the alphabet the data lives in. To account for this, we show that two simple privatization mechanisms are universally optimal in the high and low privacy regimes. We conclude Chapter 3 by proving the universal optimality of a simple privatization mechanism under approximate differential privacy, a popular relaxation to differential privacy.

Chapter 4: Multi-Party Differential Privacy

Chapter 4 studies multi-party differential privacy. We start by studying the problem of interactive function computation by multiple parties, each possessing a bit, in a differential privacy setting. Each party wants to compute a function, which could differ from party to party, and there could be a central observer interested in computing a separate function. Performance at each party is measured via the accuracy of the function to be computed. We allow for an arbitrary cost metric to measure the distortion between the true and the computed function values. Our main result is the optimality of a simple non-interactive protocol: each party randomizes its bit (sufficiently) and shares the privatized version with the other parties. This optimality result is very general: it holds for all types of functions, heterogeneous privacy conditions on the parties, all types of cost metrics, and both average and worst-case (over the inputs) measures of accuracy. We conclude Chapter 4 by showing that interaction can be helpful in settings where parties possess more than just one bit.

Chapter 5: Conclusion and Summary

Chapter 5 concludes this thesis and discusses a few interesting and non-trivial directions for future research.

CHAPTER 2

GLOBAL DIFFERENTIAL PRIVACY

2.1 Introduction

Differential privacy is a formal framework to quantify to what extent individual privacy in a statistical database is preserved while releasing useful aggregate information about the database. It provides strong privacy guarantees by requiring the indistinguishability of whether or not an individual is in a database based on the released information, regardless of the side information on the other aspects of the database the adversary may possess. Denoting the database when the individual is present as D_1 and as D_0 when the individual is not, a differentially private mechanism provides indistinguishability guarantees with respect to the pair (D_0, D_1). The databases D_0 and D_1 are referred to as "neighboring" databases.

Definition 2.1.1 (Differential Privacy [7, 9]) *A randomized mechanism M over a set of databases is* (ε, δ)-differentially private *if for all pairs of neighboring databases D_0 and D_1, and for all sets S in the output space of the mechanism $\mathcal{X}$,*

$$\mathbb{P}(M(D_0) \in S) \;\; \leq \;\; e^{\varepsilon}\, \mathbb{P}(M(D_1) \in S) + \delta \ .$$

2.2 Operational Interpretation of Differential Privacy

Given a random output Y of a database access mechanism M, consider the following hypothesis testing experiment. We choose a null hypothesis as database D_0 and alternative hypothesis as D_1:

$$\begin{aligned} H0 &: Y \text{ came from a database } D_0 \ , \\ H1 &: Y \text{ came from a database } D_1 \ . \end{aligned}$$

For a choice of a rejection region S, the probability of false alarm (type I error), when the null hypothesis is true but rejected, is defined as

$$P_{\mathrm{FA}}(D_0, D_1, M, S) \equiv \mathbb{P}\big(M(D_0) \in S\big),$$

and the probability of missed detection (type II error), when the null hypothesis is false but retained, is defined as

$$P_{\mathrm{MD}}(D_0, D_1, M, S) \equiv \mathbb{P}\big(M(D_1) \in \bar{S}\big),$$

where $\bar{S}$ is the complement of S. It turns out that imposing differential privacy conditions on a mechanism M is equivalent to restricting the probabilities of false alarm and missed detection from being simultaneously small. Wasserman and Zhu proved that $(\varepsilon, 0)$-differential privacy implies the conditions in Equation (2.1) for the special case when $\delta = 0$ [10, Theorem 2.4]. The same proof technique can be used to prove a similar result for a general $\delta \in [0, 1]$, and to prove that the conditions in Equation (2.1) imply (ε, δ)-differential privacy as well. We refer the reader to Section A.1.1 for a proof.

Theorem 2.2.1 *For any $\varepsilon \geq 0$ and $\delta \in [0, 1]$, a database mechanism M is (ε, δ)-differentially private if and only if the following conditions are satisfied for all pairs of neighboring databases D_0 and D_1, and all rejection region $S \subseteq \mathcal{X}$:*

$$\begin{aligned} P_{\mathrm{FA}}(D_0, D_1, M, S) + e^{\varepsilon} P_{\mathrm{MD}}(D_0, D_1, M, S) &\geq 1 - \delta \ , \ \textit{and} \\ e^{\varepsilon} P_{\mathrm{FA}}(D_0, D_1, M, S) + P_{\mathrm{MD}}(D_0, D_1, M, S) &\geq 1 - \delta \ . \end{aligned} \tag{2.1}$$

This operational perspective relates the privacy parameters ε and δ to a set of conditions on probability of false alarm and missed detection. This shows that under differential privacy, it is impossible for both P_{MD} and P_{FA} to be simultaneously small. This operational interpretation of differential privacy

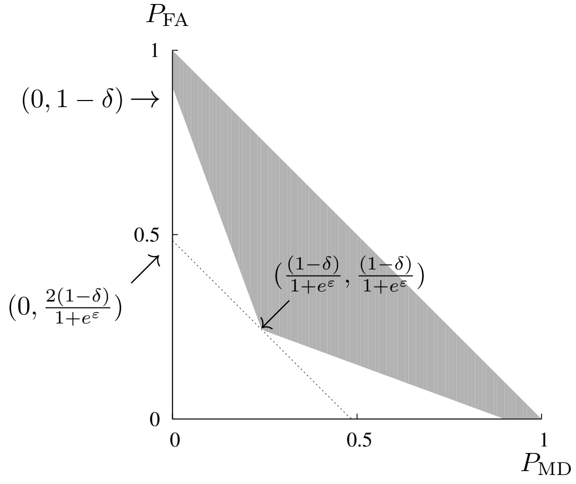


Figure 2.1: Privacy region for (ε, δ)-differential privacy. Dotted line represents the solution of a maximization problem (A.1). For simplicity, we only show the privacy region below the line $P_{\rm FA} + P_{\rm MD} \leq 1$, since the whole region is symmetric w.r.t. the line $P_{\rm FA} + P_{\rm MD} = 1$.

suggests a graphical representation of differential privacy as illustrated in Figure 2.1. We define the *privacy region* for (ε, δ)-differential privacy as

$$\begin{aligned} \mathcal{R}(\varepsilon, \delta) \;\; \equiv \;\; \Big\{ (P_{\rm MD}, P_{\rm FA}) \,\Big|\, & P_{\rm FA} + e^{\varepsilon} P_{\rm MD} \geq 1 - \delta \;, \\ & \text{and} \quad e^{\varepsilon} P_{\rm FA} + P_{\rm MD} \geq 1 - \delta \Big\} \;. \end{aligned} \tag{2.2}$$

Similarly, we define the *privacy region* of a database access mechanism M with respect to two neighboring databases D_0 and D_1 as

$$\begin{aligned} \mathcal{R}(M, D_0, D_1) \;\; \equiv \;\; \text{conv}\Bigg(\Big\{ & (P_{\rm MD}(D_0, D_1, M, S), P_{\rm FA}(D_0, D_1, M, S)) \\ & \Big| \text{for all } S \subseteq \mathcal{X} \Big\} \Bigg), \end{aligned} \tag{2.3}$$

where conv$(\cdot)$ is the convex hull of a set and $\mathcal{X}$ is the alphabet of the privatized output. Operationally, by taking the convex hull, the region includes the pairs of false alarm and missed detection probabilities achieved by soft decisions that might use internal randomness in the hypothesis testing rule. Precisely, let $\gamma : \mathcal{X} \to \{H_0, H_1\}$ be any randomized decision. For example, we can accept the null hypothesis with a certain probability p_1 if the output is in a set S_1 and probability p_2 if it is in another set S_2. In full generality, a decision rule γ can be fully described by a partition $\{S_i\}$ of the output space $\mathcal{X}$, and a corresponding accept probabilities $\{p_i\}$. The probabilities of false alarm and missed detection for a decision rule γ are defined as $P_{\rm FA}(D_0, D_1, M, \gamma) \equiv \mathbb{P}(\gamma(M(D_0)) = H_1)$ and $P_{\rm MD}(D_0, D_1, M, \gamma) \equiv \mathbb{P}(\gamma(M(D_1)) = H_0)$.

Remark 1 *For all neighboring databases D_0 and D_1, and a database access mechanism M, the pair of false alarm and missed detection probabilities achieved by any decision rule γ is included in the privacy region:*

$$(P_{\rm MD}(D_0, D_1, M, \gamma), P_{\rm FA}(D_0, D_1, M, \gamma)) \;\; \in \;\; \mathcal{R}(M, D_0, D_1) \; ,$$

for all decision rules γ.

The proof of Remark 1 is provided in Appendix A.1.2. Let $D_0 \sim D_1$ denote that the two databases are neighbors. The union over all neighboring databases defines the *privacy region of the mechanism.*

$$\mathcal{R}(M) \;\; \equiv \;\; \bigcup_{D_0 \sim D_1} \mathcal{R}(M, D_0, D_1) \; .$$

The following corollary, which follows immediately from Theorem 2.2.1, gives a necessary and sufficient condition on the privacy region for (ε, δ)-differential privacy.

Corollary 2.2.2 *A mechanism M is (ε, δ)-differentially private if and only if $\mathcal{R}(M) \subseteq \mathcal{R}(\varepsilon, \delta)$.*

To illustrate the strengths of the graphical representation of differential privacy, we provide simpler proofs for some well-known results in differential privacy in Appendix A.1.3.

Consider two database access mechanisms $M(\cdot)$ and $M'(\cdot)$. Let X and Y denote the random outputs of mechanisms M and M' respectively. We say that M *dominates* M' if $M'(D)$ is conditionally independent of D given the outcome of $M(D)$. In other words, the database D, $X = M(D)$ and $Y = M'(D)$ form the following Markov chain: D–X–Y.

Theorem 2.2.3 (Data processing inequality for differential privacy) *If a mechanism M dominates a mechanism M', then for all pairs of neigh-*

boring databases D_0 and D_1,

$$\mathcal{R}(M', D_0, D_1) \ \subseteq \ \mathcal{R}(M, D_0, D_1) \ .$$

We refer the reader to Appendix A.1.4 for a proof. Wasserman and Zhu [10, Lemma 2.6] have proved a similar result for the special case when M is $(\varepsilon, 0)$-differentially private, M' is also $(\varepsilon, 0)$-differentially private, which is a corollary to the above theorem. Perhaps surprisingly, the converse is also true.

Theorem 2.2.4 ([11, Corollary of Theorem 10]) *Fix a pair of neighboring databases D_0 and D_1, and let X and Y denote the random outputs of mechanisms M and M', respectively. If M and M' satisfy*

$$\mathcal{R}(M', D_0, D_1) \ \subseteq \ \mathcal{R}(M, D_0, D_1) \ ,$$

then there exists a coupling of the random outputs X and Y such that they form a Markov chain D–X–Y where $D \in \{D_0, D_1\}$.

In other words, when the privacy region of M' is included in M, there exists a stochastic transformation T that operates on X and produces a random output that has the same marginal distribution as Y conditioned on the database D. We can consider this mechanism T as a privatization mechanism that takes a (privatized) output X and provides even further privatization. The above theorem was proved in [11, Corollary of Theorem 10] in the context of comparing two experiments, where a *statistical experiment* denotes a mechanism in the context of differential privacy.

2.3 Optimal Mechanisms for Differential Privacy

In this section, we formulate the utility-maximization (cost-minimization) framework under ϵ-differential privacy as a functional optimization problem and prove that the multi-dimensional (correlated) staircase mechanism achieves the best privacy-utility tradeoff. Our formulation and proof techniques follow those developed by Geng et al. in [12, 13, 14].

2.3.1 Problem formulation

Consider a multidimensional real-valued query function

$$q : \mathcal{D}^n \rightarrow \mathbb{R}^d,$$

where $\mathcal{D}^n$ is the domain of the databases, and d is the dimension of the query output. Given $D \in \mathcal{D}^n$, the query output can be written as

$$q(D) = (q_1(D), q_2(D), \dots, q_d(D)),$$

where $q_i(D) \in R, \forall 1 \leq i \leq d$. The global sensitivity of the query function q is defined as

$$\Delta \triangleq \max_{D_0, D_1 \subseteq \mathcal{D}^n : |D_0 - D_1| \leq 1} \|q(D_0) - q(D_1)\|_1 = \sum_{i=1}^{d} |q_i(D_0) - q_i(D_1)|, \tag{2.4}$$

where the maximum is taken over all possible pairs of neighboring database entries D_0 and D_1 which differ in at most one element, i.e., one is a proper subset of the other and the larger database contains just one additional element [15]. For instance, the global sensitivity of a histogram query function is one, since each element in the dataset can affect only one component of the query output by one.

The standard approach to preserving the differential privacy is to add noise to the output of the query function. Letting $q(D)$ be the value of the query function evaluated at $D \subseteq \mathcal{D}^n$, the noise-adding mechanism M will output

$$M(D) = q(D) + \mathbf{X} = (q_1(D) + X_1, \dots, q_d(D) + X_d),$$

where $\mathbf{X} = (X_1, \dots, X_d) \in R^d$ is the noise added by the mechanism to the output of the query function. Due to the optimality of query-output independent perturbation mechanisms (under a technical condition) in [13], we restrict ourselves to query-output independent noise-adding mechanisms, i.e., we assume that the noise $\mathbf{X}$ is independent of the query output.

Using the definition of differential privacy in Equation (2.1), observe that differential privacy imposes the following constraints on the probability distribution of $\mathbf{X}$:

$$\begin{aligned}
& \Pr[M(D_0) \in S] && \leq e^{\epsilon} \Pr[M(D_1) \in S] \\
\Leftrightarrow\ & \Pr[q(D_0) + \mathbf{X} \in S] && \leq e^{\epsilon} \Pr[q(D_1) + \mathbf{X} \in S] \\
\Leftrightarrow\ & \Pr[\mathbf{X} \in S - q(D_0)] && \leq e^{\epsilon} \Pr[\mathbf{X} \in S - q(D_1)] \\
\Leftrightarrow\ & \Pr[\mathbf{X} \in S'] && \leq e^{\epsilon} \Pr[\mathbf{X} \in S' + q(D_0) - q(D_1)],
\end{aligned} \tag{2.5}$$

where $S' \triangleq S - q(D_0) = \{s - q(D_0) | s \in S\}$. Moreover, since the differential privacy conditions in (2.1) must hold for all measurable sets $S \subseteq R^d$ and $\|q(D_0) - q(D_1)\|_1 \leq \Delta$, from (2.5) we have

$$\Pr[\mathbf{X} \in S'] \leq e^{\epsilon} \Pr[\mathbf{X} \in S' + \mathbf{t}], \tag{2.6}$$

for all measurable sets $S' \subseteq R$ and for all $\mathbf{t} \in R^d$ such that $\|\mathbf{t}\|_1 \leq \Delta$.

Consider a cost function $\mathcal{L}(\cdot) : R^d \to R$ which is a function of the added noise $\mathbf{X}$. Our goal is to minimize the expectation of the cost subject to the ϵ-differential privacy constraint (2.6).

More precisely, let $\mathcal{P}$ denote the probability distribution of $\mathbf{X}$ and use $\mathcal{P}(S)$ to denote the probability $\Pr[\mathbf{X} \in S]$. The optimization problem we study is given by

$$\begin{aligned}
& \underset{\mathcal{P}}{\text{minimize}} \int\int \dots \int_{R^d} \mathcal{L}(x_1, x_2, \dots, x_d) \mathcal{P}(dx_1 dx_2 \dots dx_d) \\
& \text{subject to } \mathcal{P}(S) \leq e^{\epsilon} \mathcal{P}(S + \mathbf{t}), \forall \text{ measurable set } S \subseteq R^d,\ \forall \|\mathbf{t}\|_1 \leq \Delta.
\end{aligned} \tag{2.7}$$

We solve the above functional optimization problem and derive the optimal noise probability distribution for $\mathcal{L}(x_1, \dots, x_d) = \sum_{i=1}^{d} |x_i|$ with $d = 2$.

2.3.2 Main result

In this section, we state our main result: The correlated multi-dimensional staircase mechanism is the optimal solution to the functional optimization problem in (2.7) (see Theorem 2.3.1). The detailed proof is given in Appendix A.2.

In this work we consider the ℓ^1 cost function:

$$\mathcal{L}(x_1, x_2, \dots, x_d) = \sum_{i=1}^{d} |x_i|, \forall (x_1, x_2, \dots, x_d) \in R^d.$$

Consider a class of multidimensional probability distributions with symmetric and staircase-shaped probability density function defined as follows. Given $\gamma \in [0, 1]$, define $\mathcal{P}_\gamma$ as the probability distribution with probability density function $f_\gamma(\cdot)$ defined as

$$f_\gamma(\mathbf{x}) = \begin{cases} e^{-k\epsilon} a(\gamma) & \|\mathbf{x}\|_1 \in [k\Delta, (k+\gamma)\Delta) \text{ for } k \in \mathbb{N} \\ e^{-(k+1)\epsilon} a(\gamma) & \|\mathbf{x}\|_1 \in [(k+\gamma)\Delta, (k+1)\Delta) \text{ for } k \in \mathbb{N}, \end{cases} \tag{2.8}$$

where $a(\gamma)$ is the normalization factor to make

$$\int \int \dots \int_{R^d} f_\gamma(\mathbf{x}) dx_1 dx_2 \dots dx_d = 1.$$

Define $b \triangleq e^{-\epsilon}$, and define

$$c_k \triangleq \sum_{i=0}^{+\infty} i^k b^i, \forall k \in \mathbb{N},$$

where by convention 0^0 is defined as 1. Then the closed-form expression for $a(\gamma)$ is

$$a(\gamma) \triangleq \frac{d!}{2^d \Delta^d \sum_{k=1}^{d} \binom{d}{k} c_{d-k} (b + (1-b)\gamma^k)}.$$

It is straightforward to verify that $f_\gamma(\cdot)$ is a valid probability density function and $\mathcal{P}_\gamma$ satisfies the differential privacy constraint (2.7). Indeed, the probability density function $f_\gamma(x)$ satisfies

$$f_\gamma(\mathbf{x}) \leq e^\epsilon f_\gamma(\mathbf{x} + \mathbf{t}), \forall \mathbf{x} \in R^d, \forall \mathbf{t} \in R^d \text{ s.t. } \|\mathbf{t}\|_1 \leq \Delta,$$

which implies (2.7).

We plot the probability density function $f_\gamma(\mathbf{x})$ in Figure 2.2 for $d = 2$. It is easy to see that $f_\gamma(\mathbf{x})$ is multi-dimensional staircase-shaped.

Let $\mathcal{SP}$ be the set of all probability distributions which satisfy the differ-

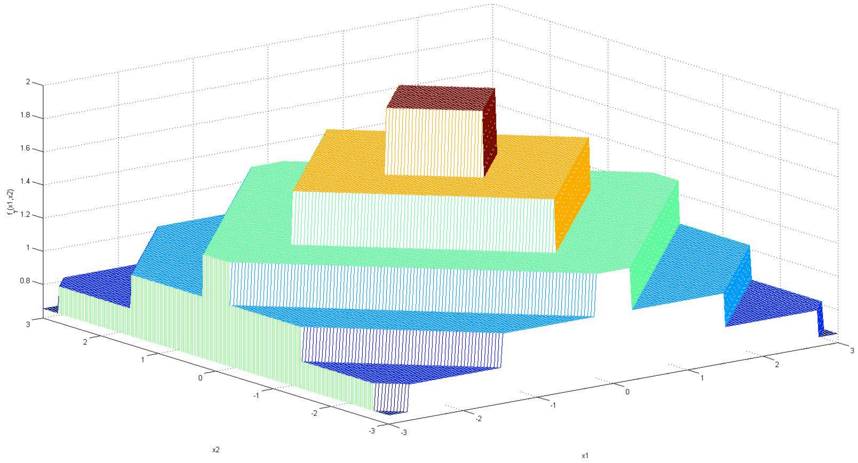


Figure 2.2: Multi-dimensional staircase-shaped probability density function

ential privacy constraint (2.7). Our main result is Theorem 2.3.1.

Theorem 2.3.1 *For $d = 2$ and the cost function $\mathcal{L}(\mathbf{x}) = \|\mathbf{x}\|_1, \forall \mathbf{x} \in R^2$, then*

$$\inf_{\mathcal{P} \in \mathcal{SP}} \int \int_{R^2} \mathcal{L}(\mathbf{x}) \mathcal{P}(dx_1 dx_2) = \inf_{\gamma \in [0,1]} \int \int_{R^2} \mathcal{L}(\mathbf{x}) f_\gamma(\mathbf{x}) dx_1 dx_2.$$

We briefly discuss the main proof idea and technique. For the complete proof, see Appendix A.2. First, by using a combinatorial argument, we show that given any noise probability distribution satisfying the ϵ-differential privacy constraint, we can discretize the probability distribution by averaging it over each ℓ^1 layer without increasing the cost. Therefore, we only need to consider those probability distributions with the probability density function being a piecewise constant function of the ℓ^1-norm of the noise. Second, we show that to minimize the cost, the probability density function as a function of the ℓ^1-norm of the noise should be monotonically and geometrically decaying. Lastly, we show that the optimal probability density function should be staircase-shaped.

Therefore, the optimal noise probability distribution to preserve ϵ-differential privacy for multidimensional real-valued query function has a staircase-shaped probability density function, which is specified by three parameters ϵ, Δ and $\gamma^* = \underset{\gamma \in [0,1]}{\arg\min} \int \int_{R^2} \mathcal{L}(x_1, x_2) f_\gamma(\mathbf{x}) dx_1 dx_2$.

We conjecture that Theorem 2.3.1 holds for arbitrary dimension d. To prove this conjecture, one can reuse the whole proof in Appendix A.2 and only needs to prove that Lemma A.2.1 and Lemma A.2.6 hold for arbitrary d, which we believe are true. Lemma A.2.1 shows that when $d = 2$, we can discretize the probability distribution by averaging it over each ℓ^1 layer without increasing the cost, and the new probability distribution also satisfies the differential privacy constraint. We give a constructive combinatorial argument to prove Lemma A.2.1 for $d = 2$, and believe it holds for arbitrary $d \geq 2$. We prove Lemma A.2.6 for $d = 2$ by studying the monotonicity of the ratio between the cost and volume over each ℓ^1 layer. Indeed, to prove Lemma A.2.6, one only needs to show that h_k, which is defined in (A.12), first decreases and then increases as a function of k, and $h_0 \leq h_{i-1}$. For fixed d, one can derive the explicit formula for d and verify whether h_k satisfies this property (we show it is true for $d = 2$ in our proof).

We also conjecture that Theorem 2.3.1 holds for other cost functions, which may not be a function only depending on the ℓ^1-norm of the noise. Numeric simulations suggest that for $d = 2$, the correlated multidimensional staircase mechanism is optimal for $\mathcal{L}(\mathbf{x}) = \|\mathbf{x}\|_2^2$. To prove this conjecture, one has to use a different proof technique, as Lemma A.2.1 in our proof does not work for the cost functions that do not depend on the ℓ^1-norm of the noise only.

2.3.3 Asymptotic analysis

In this subsection, we study the asymptotic properties and performances of the correlated staircase mechanism for the ℓ^1 cost function.

Note that the closed-form expressions for c_0, c_1 and c_2 are

$$\begin{aligned} c_0 &= \frac{1}{1-b}, \\ c_1 &= \frac{b}{(1-b)^2}, \\ c_2 &= \frac{b^2+b}{(1-b)^3}. \end{aligned}$$

For $d = 2$, we have

$$\begin{aligned} a(\gamma) &= \frac{1}{2\Delta^2\left(2c_1(b+(1-b)\gamma)+c_0(b+(1-b)\gamma^2)\right)} \\ &= \frac{1}{2\Delta^2\left(\gamma^2+\frac{2b}{1-b}\gamma+\frac{b+b^2}{(1-b)^2}\right)}. \end{aligned}$$

Given the two-dimensional staircase-shaped probability density function $f_\gamma(\mathbf{x})$, the cost is

$$\begin{aligned} V(\mathcal{P}_\gamma) &\triangleq \int\int_{\mathbb{R}^2}(|x_1|+|x_2|)f_\gamma(x_1,x_2)\mathcal{P}(dx_1dx_2) \\ &= 4\left(\sum_{i=0}^{+\infty}\int_{i\Delta}^{(i+\gamma)\Delta} tta(\gamma)e^{-i\epsilon}dt + \sum_{i=0}^{+\infty}\int_{(i+\gamma)\Delta}^{(i+1)\Delta} tta(\gamma)e^{-(i+1)\epsilon}dt\right) \\ &= \frac{4a(\gamma)\Delta^3}{3}\left(\sum_{i=0}^{+\infty}b^i(3i^2\gamma+3i\gamma^2+\gamma^3)\right. \\ &\qquad\qquad \left. +b\sum_{i=0}^{+\infty}b^i(3i^2+3i+1-3i^2\gamma-3i\gamma^2-\gamma^3)\right) \\ &= \frac{4a(\gamma)\Delta^3}{3}\left(3c_2\gamma+3c_1\gamma^2+c_0\gamma^3+b\left(3(1-\gamma)c_2+3(1-\gamma^2)c_1\right.\right. \\ &\qquad\qquad \left.\left.+(1-\gamma^3)c_0\right)\right) \\ &= \frac{2\Delta}{3}\frac{3c_2\gamma+3c_1\gamma^2+c_0\gamma^3+b(3(1-\gamma)c_2+3(1-\gamma^2)c_1+(1-\gamma^3)c_0)}{\gamma^2+\frac{2b}{1-b}\gamma+\frac{b+b^2}{(1-b)^2}} \\ &= \frac{2\Delta}{3}\frac{c_0(1-b)\gamma^3+3c_1(1-b)\gamma^2+3c_2(1-b)\gamma+b(c_0+3c_1+3c_2)}{\gamma^2+\frac{2b}{1-b}\gamma+\frac{b+b^2}{(1-b)^2}}. \\ &= \frac{2\Delta}{3}\frac{\gamma^3+\frac{3b}{1-b}\gamma^2+\frac{3(b^2+b)}{(1-b)^2}\gamma+b\frac{1+4b+b^2}{(1-b)^3}}{\gamma^2+\frac{2b}{1-b}\gamma+\frac{b+b^2}{(1-b)^2}}. \end{aligned} \tag{2.9}$$

Therefore, in the two-dimensional setting, the optimal γ^* is

$$\gamma^* = \arg\min_{\gamma\in[0,1]}\frac{\gamma^3+\frac{3b}{1-b}\gamma^2+\frac{3(b^2+b)}{(1-b)^2}\gamma+b\frac{1+4b+b^2}{(1-b)^3}}{\gamma^2+\frac{2b}{1-b}\gamma+\frac{b+b^2}{(1-b)^2}}.$$

By setting the derivative of (2.9) to be zero, we use Mathematica to get a closed-form expression for γ^*, which is too complicated to show here. We plot γ^* as a function of b in Figure 2.3. The optimal cost $V^* = V(\mathcal{P}_{\gamma^*})$. We use Mathematica to analyze the asymptotic behavior of V^* as $\epsilon \to 0$ and

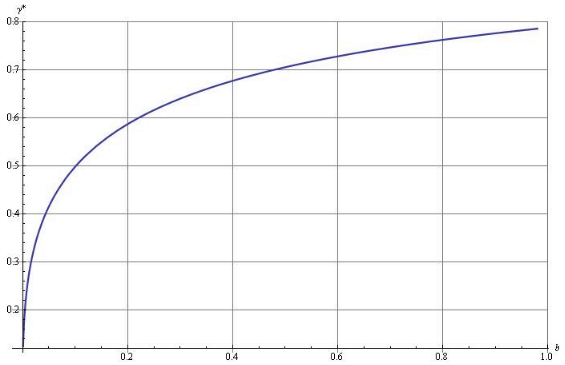


Figure 2.3: The optimal γ^* as a function of b

$\epsilon \to +\infty$.

Corollary 2.3.2 *In the high privacy regime,*

$$V^* = \frac{2\Delta}{\epsilon} - \frac{\Delta\epsilon^2}{36\sqrt{3}} + O(\epsilon^3), \epsilon \to 0,$$

and in the low privacy regime,

$$V^* = \sqrt[3]{2}\Delta e^{-\frac{\epsilon}{3}} + \frac{\Delta e^{-\frac{2\epsilon}{3}}}{\sqrt[3]{2}} + o(e^{-\frac{2\epsilon}{3}}), \epsilon \to +\infty.$$

The Laplacian mechanism adds independent Laplacian noise to each component of the query output, and the cost is $\frac{2\Delta}{\epsilon}$. Therefore, in the high privacy regime, the gap between optimal cost and the cost achieved by the Laplacian mechanism goes to zero, as $\epsilon \to 0$, and we conclude that the Laplacian mechanism is approximately optimal in the high privacy regime. However, in the low privacy regime (as $\epsilon \to +\infty$), the optimal cost is proportional to $e^{-\frac{\epsilon}{3}}$, while the cost of the Laplacian mechanism is proportional to $\frac{1}{\epsilon}$. We conclude that the gap is significant in the low privacy regime.

It is natural to compare the performance of the optimal multi-dimensional staircase mechanism and the composite single-dimensional staircase mechanism which adds independent staircase noise to each component of the query output. If independent staircase noise is added to each component of query output, to satisfy the ϵ-differential privacy constraint, the parameter of the staircase noise is $\frac{\epsilon}{2}$ instead of ϵ, and thus the total cost will be proportional to $e^{-\frac{\epsilon}{4}}$, which is worse than the optimal cost $\Theta(e^{-\frac{\epsilon}{3}})$.

2.4 The Composition Theorem in Differential Privacy

In this section, we address how differential privacy guarantees compose when accessing databases multiple times via differentially private mechanisms, each of which has its own privacy guarantees. Precisely, we address the following fundamental question: How much privacy can be guaranteed after multiple database accesses? To formally define composition, we consider the following scenario known as the 'composition experiment', proposed in [16].

A composition experiment takes as input a parameter $b \in \{0, 1\}$, and an adversary $\mathcal{A}$. From the hypothesis testing perspective proposed in the previous section, b can be interpreted as the hypothesis: null hypothesis for $b = 0$ and alternative hypothesis for $b = 1$. At each time i, a database $D^{i,b}$ is accessed depending on b. For example, one includes a particular individual and another does not. For example, $D^{1,0}$ could be medical records including a particular individual and $D^{1,1}$ does not include the person, and $D^{2,0}$ could be voter registration database with the same person present and $D^{2,1}$ with the person absent. An adversary $\mathcal{A}$ is trying to figure out whether or not a particular individual is in the database by testing the hypotheses on the output of k sequential database accesses via differentially private mechanisms. In full generality, we allow the adversary to have full control over which pair of databases to access, which query to ask, and which mechanism to be used at each repeated access. Further, the adversary is free to make these choices adaptively based on the previous outcomes. The only restrictions are: (a) the differentially private mechanisms belong to a family $\mathcal{M}$ (e.g., the family of all (ε, δ)-differentially private mechanisms), (b) the internal randomness of the mechanisms are independent at each repeated access, and (c) that the hypothesis b is not known to the adversary.

```
COMPOSE($\mathcal{A}, \mathcal{M}, k, b$)

Input: $\mathcal{A}$, $\mathcal{M}$, $k$, $b$
Output: $V^b$
  for $i = 1$ to $k$ do
      $\mathcal{A}$ requests $(D^{i,0}, D^{i,1}, q_i, M_i)$ for some $M_i \in \mathcal{M}$;
      $\mathcal{A}$ receives $y_i = M_i(D^{i,b}, q_i)$;
  end for
  Output the view of the adversary $V^b = (R^b, Y_1^b, \ldots, Y_k^b)$.
```

The outcome of this k-fold composition experiment is the *view of the adversary $\mathcal{A}$*: $V^b \equiv (R, Y_1^b, \ldots, Y_k^b)$, which is the sequence of random outcomes $Y_1^b, \ldots, Y_k^b$, and the outcome R of any internal randomness of $\mathcal{A}$.

2.4.1 Optimal privacy region under composition

We would like to characterize how much privacy degrades after a k-fold composition experiment. It is known that the privacy degrades under composition by at most the 'sum' of the differential privacy parameters of each access.

Theorem 2.4.1 ([7, 9, 8, 16]) *For any $\varepsilon > 0$ and $\delta \in [0,1]$, the class of (ε,δ)-differentially private mechanisms satisfies $(k\varepsilon, k\delta)$-differential privacy under k-fold adaptive composition.*

In general, one can show that if M_i is $(\varepsilon_i, \delta_i)$-differentially private, then the composition satisfies $(\sum_{i\in[k]} \varepsilon_i, \sum_{i\in[k]} \delta_i)$-differential privacy. If we do not allow for any slack in the δ, this bound cannot be tightened. Precisely, there are examples of mechanisms which under k-fold composition violate $(\varepsilon, \sum_{i\in[k]} \delta_i)$-differential privacy for any $\varepsilon < \sum_{i\in[k]} \varepsilon_i$. We can prove this by providing a set S such that the privacy condition is met with equality: $\mathbb{P}(V^0 \in S) = e^{\sum_{i\in[k]} \varepsilon_i} \mathbb{P}(V^1 \in S) + \sum_{i\in[k]} \delta_i$. However, if we allow for a slightly larger value of δ, then Dwork et al. showed in [16] that one can gain a significantly higher privacy guarantee in terms of ε.

Theorem 2.4.2 ([16, Theorem III.3]) *For any $\varepsilon > 0$, $\delta \in [0,1]$, and $\tilde{\delta} \in (0,1]$, the class of (ε,δ)-differentially private mechanisms satisfies $(\tilde{\varepsilon}_{\tilde{\delta}}, k\delta + \tilde{\delta})$-differential privacy under k-fold adaptive composition, for*

$$\tilde{\varepsilon}_{\tilde{\delta}} \;=\; k\varepsilon(e^{\varepsilon}-1) + \varepsilon\sqrt{2k\log(1/\tilde{\delta})}. \qquad (2.10)$$

By allowing a slack of $\tilde{\delta} > 0$, one can get a higher privacy of $\tilde{\varepsilon}_{\tilde{\delta}} = O(k\varepsilon^2 + \sqrt{k\varepsilon^2})$, which is significantly smaller than $k\varepsilon$. This is the best known guarantee so far, and has been used whenever one requires a privacy guarantee under composition (e.g. [16, 17, 18]). However, the important question of optimality has remained open. Namely, is there a composition of mechanisms where the above privacy guarantee is tight? In other words, is it possible to get a tighter bound on differential privacy under composition?

We give a complete answer to this fundamental question in the following theorems. We prove a tighter bound on the privacy guarantee under composition. Further, we also prove the achievability of the privacy guarantee: we provide a set of mechanisms such that the privacy region under k-fold composition is exactly the region defined by the conditions in (2.11). Hence, this bound on the privacy region is tight and cannot be improved upon.

Theorem 2.4.3 *For any $\varepsilon \geq 0$ and $\delta \in [0,1]$, the class of (ε, δ)-differentially private mechanisms satisfies*

$$\big(\,(k-2i)\varepsilon\,,\,1-(1-\delta)^k(1-\delta_i)\,\big)\text{-differential privacy} \tag{2.11}$$

under k-fold adaptive composition, for all $i = \{0, 1, \ldots, \lfloor k/2 \rfloor\}$, where

$$\delta_i \;=\; \frac{\sum_{\ell=0}^{i-1}\binom{k}{\ell}\big(e^{(k-\ell)\varepsilon}-e^{(k-2i+\ell)\varepsilon}\big)}{(1+e^{\varepsilon})^k}\,.$$

Hence, the privacy region of k-fold composition is an intersection of k regions, each of which is $((k-2i)\varepsilon, 1-(1-\delta)^k(1-\delta_i))$-differentially private: $\mathcal{R}(\{(k-2i)\varepsilon, 1-(1-\delta)^k(1-\delta_i)\}_{i\in[k/2]}) \equiv \bigcap_{i=0}^{\lfloor\frac{k}{2}\rfloor}\mathcal{R}((k-2i)\varepsilon, 1-(1-\delta)^k(1-\delta_i))$. We prove this result in Section A.3.1 by constructing an explicit mechanism that achieves this region under composition. Hence, this bound on the privacy region is tight, and gives the exact description of how privacy degrades under k-fold adaptive composition. This settles the question that was left open in [7, 9, 8, 16] by providing, for the first time, the fundamental limit of composition and proving a matching mechanism with the worst-case privacy degradation.

To prove the optimality of our main result in Theorem 2.4.3, namely that it is impossible to have a privacy worse than (2.11), we rely on the operational interpretation of the privacy as hypothesis testing. To this end, we use the new analysis tools (Theorem 2.2.3 and Theorem 2.2.4) provided in the previous section. Figure 2.4 illustrates how much the privacy region of Theorem 2.4.3 degrades as we increase the number of composition k. Figure 2.5 provides a comparison of the three privacy guarantees in Theorems 2.4.1, 2.4.2 and 2.4.3 for 30-fold composition of $(0.1, 0.001)$-differentially private mechanisms. Smaller region gives a tighter bound, since it guarantees the higher privacy.

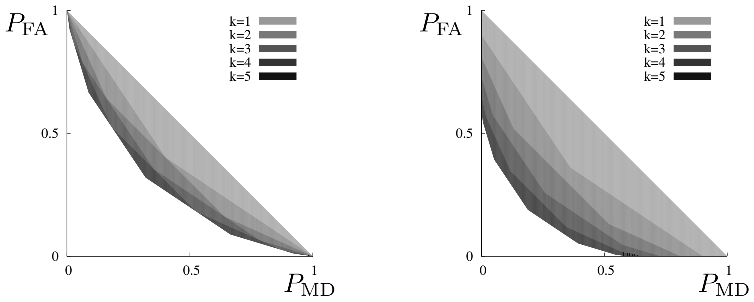


Figure 2.4: Privacy region $\mathcal{R}(\{(k-2i)\varepsilon, \delta_i\})$ for the class of $(\varepsilon, 0)$-differentially private mechanisms (left) and (ε, δ)-differentially private mechanisms (right) under k-fold adaptive composition.

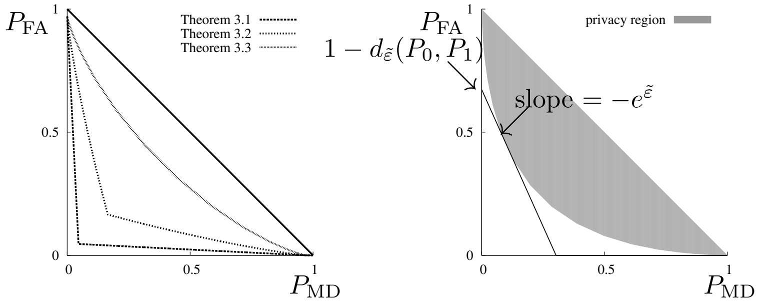


Figure 2.5: Theorem 2.4.3 provides the tightest bound (left). Given a mechanism M, the privacy region can be completely described by its boundary, which is represented by a set of tangent lines of the form $P_{\text{FA}} = -e^{\tilde{\varepsilon}} P_{\text{MD}} + 1 - d_{\tilde{\varepsilon}}(P_0, P_1)$ (right).

2.4.2 Simplified privacy region under composition

In many applications of the composition theorems, a closed form expression of the composition privacy guarantee is required. The privacy guarantee in (2.11) is tight, but can be difficult to evaluate. The next theorem provides a simpler expression which is an outer bound on the exact region described in (2.11). Compared to (2.10), the privacy guarantee is significantly improved from $\tilde{\varepsilon}_{\tilde{\delta}} = O\Big(k\varepsilon^2 + \sqrt{k\varepsilon^2 \log(1/\tilde{\delta})}\Big)$ to $\tilde{\varepsilon}_{\tilde{\delta}} = O\Big(k\varepsilon^2 + \min\big\{\sqrt{k\varepsilon^2 \log(1/\tilde{\delta})}, \varepsilon \log(\varepsilon/\tilde{\delta})\big\}\Big)$, especially when composing a large number k of interactive queries. Further, the δ-approximate differential privacy degradation of $(1-(1-\delta)^k(1-\tilde{\delta}))$ is also strictly smaller than the previous $(k\delta + \tilde{\delta})$. We discuss the significance of this improvement in the next section using examples from the existing differential privacy literature.

Theorem 2.4.4 *For any $\varepsilon > 0$, $\delta \in [0,1]$, and $\tilde{\delta} \in [0,1]$, the class of (ε,δ)-differentially private mechanisms satisfies $\big(\tilde{\varepsilon}_{\tilde{\delta}}, 1-(1-\delta)^k(1-\tilde{\delta})\big)$-differential privacy under k-fold adaptive composition, for*

$$\tilde{\varepsilon}_{\tilde{\delta}} \;=\; \min\Bigg\{ k\varepsilon \;,\; \frac{(e^\varepsilon-1)\varepsilon k}{e^\varepsilon+1} + \varepsilon\sqrt{2k\,\log\Big(e+\frac{\sqrt{k\varepsilon^2}}{\tilde{\delta}}\Big)} \;, \qquad \frac{(e^\varepsilon-1)\varepsilon k}{e^\varepsilon+1} + \varepsilon\sqrt{2k\,\log\Big(\frac{1}{\tilde{\delta}}\Big)} \Bigg\} \;. \tag{2.12}$$

In the high privacy regime, where $\varepsilon \le 0.9$, this bound can be further simplified as

$$\tilde{\varepsilon}_{\tilde{\delta}} \;\le\; \min\Big\{ k\varepsilon, k\varepsilon^2 + \varepsilon\sqrt{2k\log\big(e+(\sqrt{k\varepsilon^2}/\tilde{\delta}\,)\big)}, k\varepsilon^2 + \varepsilon\sqrt{2k\log(1/\tilde{\delta})} \Big\} \;.$$

A proof is provided in Section A.3.2. This privacy guarantee improves over the existing result of Theorem 2.4.2 when $\tilde{\delta} = \Theta(\sqrt{k\varepsilon^2})$. Typical regime of interest is the high-privacy regime for composition privacy guarantee, i.e. when $\sqrt{k\varepsilon^2} \ll 1$. The above theorem suggests that we only need the extra slack of approximate privacy $\tilde{\delta}$ of order $\sqrt{k\varepsilon^2}$.

2.4.3 Composition theorem for heterogeneous mechanisms

So far, we considered homogeneous mechanisms, where all mechanisms are (ε,δ)-differentially private. Our analysis readily extends to heterogeneous mechanisms, where the ℓ-th query satisfies $(\varepsilon_\ell,\delta_\ell)$-differential privacy (we refer to such mechanisms as $(\varepsilon_\ell,\delta_\ell)$-differentially private mechanisms).

Theorem 2.4.5 *For any $\varepsilon_\ell > 0$, $\delta_\ell \in [0,1]$ for $\ell \in \{1,\ldots,k\}$, and $\tilde{\delta} \in [0,1]$, the class of $(\varepsilon_\ell,\delta_\ell)$-differentially private mechanisms satisfies $\big(\tilde{\varepsilon}_{\tilde{\delta}}, 1-(1-\tilde{\delta})\prod_{\ell=1}^{k}(1-\delta_\ell)\big)$-differential privacy under k-fold adaptive composition,* *for* $\tilde{\varepsilon}_{\tilde{\delta}} =$

$$\min\left\{ \sum_{\ell=1}^{k} \varepsilon_\ell \;,\; \sum_{\ell=1}^{k} \frac{(e^{\varepsilon_\ell}-1)\varepsilon_\ell}{e^{\varepsilon_\ell}+1} + \sqrt{\sum_{\ell=1}^{k} 2\,\varepsilon_\ell^2 \,\log\Big(e + \frac{\sqrt{\sum_{\ell=1}^{k}\varepsilon_\ell^2}}{\tilde{\delta}}\Big)} \;, \right.$$
$$\left. \sum_{\ell=1}^{k} \frac{(e^{\varepsilon_\ell}-1)\varepsilon_\ell}{e^{\varepsilon_\ell}+1} + \sqrt{\sum_{\ell=1}^{k} 2\,\varepsilon_\ell^2 \,\log\Big(\frac{1}{\tilde{\delta}}\Big)} \right\} . \quad (2.13)$$

This tells us that the ε_ℓ's *sum up under composition*: whenever we have $k\varepsilon$ or $k\varepsilon^2$ in (2.12) we can replace it by the summation to get the general result for heterogeneous case. We refer the reader to Appendix A.3.3 for the proof of Theorem 2.4.5.

2.5 Conclusion and Summary

In this chapter, we have studied the fundamental limits of global differential privacy. In particular, we showed the following key results:

1. Global differential privacy guarantees that the probabilities of false alarm and missed detection of a binary hypothesis testing problem involving a specific user presence/absence cannot be simultaneously small.

2. The correlated multi-dimensional staircase mechanism achieves the optimal privacy-utility tradeoff under ℓ_1 losses and one/two-dimensional query functions. We also conjectured that the same mechanism is optimal for higher dimensional queries and more general loss function.

3. The composition of k queries, each of which is (ϵ, δ)-differentially private, is at least $(\tilde{\varepsilon}_{\tilde{\delta}}, k\delta + \tilde{\delta})$-differential private. Here $\tilde{\varepsilon}_{\tilde{\delta}} = O\Big(k\varepsilon^2 + \varepsilon\sqrt{k\log(e + (\varepsilon\sqrt{k}/\tilde{\delta})\,)}\Big)$ and $\tilde{\delta}$ is any nonnegative number.

CHAPTER 3

LOCAL DIFFERENTIAL PRIVACY

3.1 Introduction

In statistical analyses involving data from individuals, there is an increasing tension between the need to share the data and the need to protect sensitive information about the individuals. For example, users of social networking sites are increasingly cautious about their privacy, but still inevitably agree to share their personal information in order to benefit from customized services such as recommendations and personalized search [19, 20]. There is a certain utility in sharing data for both data providers and data analysts, but at the same time, individuals want *plausible deniability* when it comes to sensitive information.

For such applications, there is a natural core optimization problem to be solved. Assuming both the data providers and analysts want to maximize the utility of the released data, how can they do so while preserving the privacy of participating individuals? The formulation and study of a framework addressing this fundamental tradeoff is the focus of this chapter.

3.1.1 Local differential privacy

The need for data privacy appears in two different contexts: the *local privacy* context, as in when individuals disclose their personal information (e.g., voluntarily on social network sites), and the *global privacy* context, as in when institutions release databases of information of several people or answer queries on such databases (e.g., US Government releases census data, companies like Netflix release proprietary data for others to test state of the art data analytics). In both contexts, privacy is achieved by *randomizing* the data before releasing it. We study the setting of local privacy, in which data providers do not trust the data collector (analyst). Local privacy dates back to [21], who proposed the *randomized response* method to provide plausible deniability for individuals responding to sensitive surveys.

A natural notion of privacy protection is making inference of information beyond what is released hard. *Differential privacy* has been proposed in the global privacy context to formally capture this notion of privacy [6, 7, 8]. In a nutshell, differential privacy ensures that an adversary should not be able to reliably infer whether or not a particular individual is participating in the database query, even with unbounded computational power and access to every entry in the database except for that particular individual's data. Recently, [22] extended the notion of differential privacy to the local privacy context. Formally, consider a setting where there are n data providers each owning a data X_i defined on an input alphabet $\mathcal{X}$. The X_i's are independently sampled from some distribution P_ν parameterized by ν. A statistical privatization mechanism Q is a conditional distribution that maps $X_i \in \mathcal{X}$ stochastically to $Y_i \in \mathcal{Y}$, where $\mathcal{Y}$ is an output alphabet possibly larger than $\mathcal{X}$. The Y_i's are referred to as the privatized (sanitized) views of X_i's. In a non-interactive setting where all X_i's are independently sampled from the same distribution, the same privatization mechanism Q is used by all individuals. This setting is shown in Figure 3.1 for a special case with $n = 2$. For some non-negative ε, we follow the definition of [22] and say that a mechanism Q is *ε-locally differentially private* if

$$\sup_{S \subset \mathcal{Y}, x, x' \in \mathcal{X}} \frac{Q(S|x)}{Q(S|x')} \leq e^{\varepsilon} , \tag{3.1}$$

where $Q(S|x) = \mathbb{P}(Y_i \in S | X_i = x)$ represents the privatization mechanism. This ensures that for small values of ε, given a privatized data Y_i, it is (almost) equally likely to have come from any data, i.e. x or x'. A small value of ε means that we require a high level of privacy and a large value corresponds to a low level of privacy. At one extreme, for $\varepsilon = 0$, the privatized output must be independent of the private data, and on the other extreme, for $\varepsilon = \infty$, the privatized output can be made equal to the private data.

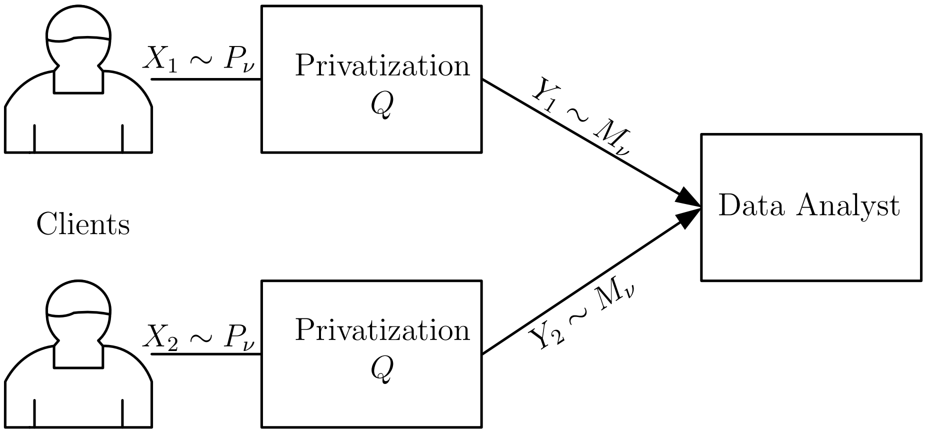


Figure 3.1: Client server model

3.1.2 Information theoretic utilities for statistical analysis

In analyses of statistical databases, the analyst is interested in the *statistics* of the data as opposed to individual records. Naturally, the utility should also be measured in terms of the distribution rather than sample quantities. Concretely, consider a client-server setting, where each client with data X_i sends a privatized version of the data Y_i, via a non-interactive ε-locally differentially private privatization mechanism Q. Assume all the clients use the same privatization mechanism denoted by Q, and each client's data is an i.i.d. sample from a distribution P_ν for some parameter ν. Given the privatized views $\{Y_i\}_{i=1}^n$, the data analyst wants to make inferences based on the induced marginal distribution

$$M_\nu(S) \;\equiv\; \sum_{x\in\mathcal{X}} Q(S|x)P_\nu(x) \;, \qquad (3.2)$$

for $S \subseteq \mathcal{Y}$. We consider a broad class of convex utility functions, and identify the class of optimal mechanisms, which we call *staircase mechanisms*, in Section 3.3. We apply this framework to two specific applications: (a) hypothesis testing where the utility is measured in Kullback-Leibler divergence (Section 3.4) and (b) information preservation where the utility is measured in mutual information (Section 3.5).

In the binary hypothesis testing setting, $\nu \in \{0,1\}$; therefore, X can be generated by one of two possible distributions P_0 and P_1. The power to discriminate data generated from P_0 to data generated from P_1 depends on the 'distance' between the marginals M_0 and M_1. To measure the ability of such statistical discrimination, our choice of utility of a particular privatization mechanism Q is an information theoretic quantity called Csiszár's f-divergence defined as

$$D_f(M_0||M_1) = \sum_{x\in\mathcal{X}} f\Big(\frac{M_0(x)}{M_1(x)}\Big)\, M_1(x)\ , \tag{3.3}$$

for some convex function f such that $f(1) = 0$. The Kullback-Leibler (KL) divergence $D_{\rm kl}(M_0||M_1)$ is a special case with $f(x) = x \log x$, and so is the total variation $\|M_0 - M_1\|_{\rm TV}$ with $f(x) = (1/2)|x - 1|$. Such f-divergences capture the quality of statistical inference, such as minimax rates of statistical estimation or error exponents in hypothesis testing [23]. As a motivating example, suppose a data analyst wants to test whether the data is generated from P_0 or P_1 based on privatized views $Y_1, \ldots, Y_n$. According to Chernoff-Stein's lemma, for a bounded type I error probability, the best type II error probability scales as $e^{-n\, D_{\rm kl}(M_0||M_1)}$. Naturally, we are interested in finding a privatization mechanism Q that minimizes the probability of error by solving the following constraint maximization problem

$$\begin{array}{ll} \underset{Q}{\text{maximize}} & D_{\rm kl}(M_0||M_1) \\ \text{subject to} & Q \in \mathcal{D}_\varepsilon \end{array}\ , \tag{3.4}$$

where $\mathcal{D}_\varepsilon$ is the set of all ε-locally differentially private mechanisms satisfying (3.1).

In the information preservation setting, X is generated from an underlying distribution P. We are interested in quantifying how much information can be preserved when releasing a private view of the data. In other words, the data provider would like to release an ε-locally differentially private view Y of X that preserves the amount of information in X as much as possible. The utility in this case is measured by the mutual information between X and Y

$$I\left(X;Y\right) = \sum_{\mathcal{X}}\sum_{\mathcal{Y}} P\left(x\right) Q\left(y|x\right) \log\left(\frac{Q\left(y|x\right)}{\sum_{l\in\mathcal{X}} P\left(l\right) Q\left(y|l\right)}\right). \tag{3.5}$$

Mutual information, as the name suggests, measures the mutual dependence between two random variables. It has been used as a criterion for feature selection and for determining the similarity between two different clusterings of a dataset, in addition to many other applications in signal processing and machine learning. To characterize the fundamental tradeoff between privacy and information preservation, we solve the following constrained maximization problem

$$\begin{aligned} \underset{Q}{\text{maximize}} \quad & I(X;Y) \\ \text{subject to} \quad & Q \in \mathcal{D}_\varepsilon \end{aligned}, \tag{3.6}$$

where $\mathcal{D}_\varepsilon$ is the set of all ε-locally differentially private mechanisms satisfying (3.1).

Motivated by such applications in statistical analysis, our goal is to provide a general framework for finding optimal privatization mechanisms that maximize information theoretic utilities under local differential privacy. We demonstrate the power of our techniques in a very general setting that includes both hypothesis testing and information preservation.

3.1.3 Our contributions

We study the fundamental tradeoff between local differential privacy and a rich class of convex utility functions. This class of utilities includes several information theoretic quantities such as mutual information and f-divergences. The privacy-utility tradeoff is posed as a constrained maximization problem: maximize utility subject to local differential privacy constraints. This maximization problem is (a) nonlinear: the utility functions we consider are convex in Q; (b) non-standard: we are maximizing instead of minimizing a convex function; and (c) infinite dimensional: the space of all differentially private mechanisms is uncountable. We show, in Theorem 3.3.2, that for all utility functions considered and any privacy level ε, a *finite* family of *extremal* mechanisms (a subset of the corner points of the space of privatization mechanisms), which we call *staircase* mechanisms, contains the optimal privatization mechanism. We further prove, in Theorem 3.3.4, that solving the original problem is equivalent to solving a linear program, the outcome of which is the optimal staircase mechanism. However, solving this linear program can be computationally expensive since it has $2^{|\mathcal{X}|}$ variables. To account for this, we show that two simple staircase mechanisms (the binary and randomized response mechanisms) are optimal in the high and low privacy regimes, respectively, and well approximate the intermediate regime. This contributes an important advance in the differential privacy area, where the privatization mechanisms have been few and almost no exact optimality results are known. As an application, we show that the effective sample size reduces from n to $\varepsilon^2 n$ under local differential privacy in the context of hypothesis testing.

We also study the fundamental tradeoff between utility and approximate differential privacy, a generalized notion of privacy that was first introduced in [9]. The techniques we develop for differential privacy do not generalize to approximate differential privacy. To account for this, we use the operational interpretation of approximate differential privacy (developed in [24]) to prove that a simple mechanism maximizes utility for all levels of privacy when the data is binary.

3.1.4 Related work

Our work is closely related to the recent work of [22] where an upper bound on $D_{\text{kl}}(M_0||M_1)$ was derived under the same local differential privacy setting. Precisely, Duchi et al. proved that the KL-divergence maximization problem in (3.4) is at most $4(e^{\varepsilon}-1)^2\|P_1-P_2\|_{TV}^2$. This bound was further used to provide a minimax bound on statistical estimation using information theoretic converse techniques such as Fano's and Le Cam's inequalities. Such tradeoffs also provide tools for comparing various notions of privacy [25].

In a similar spirit, we are also interested in maximizing information theoretic quantities of the marginals under local differential privacy. We generalize the results of [22], and provide stronger results in the sense that we (a) consider a broader class of information theoretic utilities; (b) provide explicit constructions of the optimal mechanisms; and (c) recover the existing result of [22, Theorem 1] (with a stronger condition on ε).

Our work provides a formal connection to the information-theoretical notion of privacy, where privacy loss is defined as information leakage. Information leakage has been widely studied as a practical notion of privacy [26, 27]. Such a connection to differential privacy has been studied only indirectly through comparisons to how much distortion is incurred under the two notions of privacy [28]. Given a privatization mechanism, mutual information privacy is measured by the mutual information between the data and the released output, i.e. $I(X;Y)$. We show that under ε-locally differentially, mutual information is bounded by $I(X;Y) = 0.5\varepsilon^2 \max_{A\subseteq\mathcal{X}} P(A)P(A^c)+O(\varepsilon^3)$. Moreover, we provide an explicit privatization mechanism that achieves this bound.

While there is a vast literature on differential privacy, exact optimality results are only known for a few cases. The typical recipe is to propose a differentially private mechanism inspired by the work of [6, 7, 29] and [30], and then establish its near-optimality by comparing the achievable utility to a converse, for example in principal component analysis [31, 17, 32, 33], linear queries [34, 35], logistic regression [36] and histogram release [37]. In this work, we take a different route and solve the utility maximization problem *exactly.*

Optimal differentially private mechanisms are known only in a few cases. [38] showed that the geometric noise adding mechanism is optimal (under a Bayesian setting) for monotone utility functions under count queries (sensitivity one). This was generalized by Geng et al. (for a worst-case input setting) who proposed a family of mechanisms and proved its optimality for monotone utility functions under queries with arbitrary sensitivity [12, 13, 14]. The family of optimal mechanisms was called *staircase mechanisms* because for any y and any neighboring x and x', the ratio of $Q(y|x)$ to $Q(y|x')$ takes one of three possible values e^{ε}, $e^{-\varepsilon}$, or 1. Since the optimal mechanisms we develop also have an identical property, we retain the same nomenclature.

3.1.5 Organization

The remainder of this chapter is organized as follows. In Section 3.3, we introduce the family of staircase mechanisms, prove its optimality for a broad class of convex utility functions, and study its combinatorial structure. In Section 3.4, we study the problem of private hypothesis testing and prove that two staircase mechanisms, the binary and randomized response mechanisms, are optimal for KL-divergence in the high and low privacy regimes, respectively, and (nearly) optimal the intermediate regime. We show, in Section 3.5, similar results for mutual information. In Section 3.6, we study approximate local differential privacy, a more general notion of local privacy. Finally, we conclude this chapter with a few interesting and nontrivial extensions in Section 3.7.

3.2 Operational Interpretation of Local Differential Privacy

Given an observation $Y = y$, consider a binary hypothesis test on whether $X \in A$ or $X \in B$ for some $A, B \subset \mathcal{X}$ such that $A \cap B = \emptyset$. Any binary hypothesis test is completely described by a, possibly randomized, decision rule $\hat{X} : Y \to \{A, B\}$. The two types of error associated with $\hat{X}$ are *false alarm:* $\hat{X} = A$ when $X \in B$, and *missed detection:* $\hat{X} = B$ when $X \in A$. The probability of false alarm is given by $P_{\text{FA}} = \mathbb{P}(\hat{X} = A | X \in B)$ while the probability of miss detection is given by $P_{\text{MD}} = \mathbb{P}(\hat{X} = B | X \in A)$. Notice that P_{FA} and P_{MD} are a function of Q, $\hat{X}$, A and B, and not the distribution of X. The next theorem provides an equivalent *operational definition* for local differential privacy.

Theorem 3.2.1 (Operational Definition of Local Differential Privacy) *A conditional distribution Q is ε-locally differentially private if and only if for all $A, B \subset \mathcal{X}$ such that $A \cap B = \emptyset$ and all decision rules $\hat{X} : Y \to \{A, B\}$, we have that*

$$\begin{aligned} P_{\text{FA}} + e^{\varepsilon} P_{\text{MD}} &\geq 1 \\ e^{\varepsilon} P_{\text{FA}} + P_{\text{MD}} &\geq 1. \end{aligned} \tag{3.7}$$

The proof of the above theorem is found in Appendix B.1. As a corollary to the above theorem, if you set $B = A^c$, then local differential privacy guarantees that upon the observation of Y, the adversary cannot figure out whether or not $X \in A$ reliably for any $A \subset \mathcal{X}$.

3.3 Optimal Mechanisms for Local Differential Privacy

In this section, we provide a formal definition for staircase mechanisms and show that they are the optimal solutions to optimization problems of the form (3.9). Using the structure of staircase mechanisms, we propose a combinatorial representation of staircase mechanisms. This allows us to reduce the infinite dimensional nonlinear program of (3.9) to a linear program with $2^{|\mathcal{X}|}$ variables. Potentially, for any instance of the problem, one can solve this linear program to obtain the optimal privatization mechanism, albeit with significant computational challenges since the number of variables scales exponentially in the alphabet size. To address this issue, we prove, in Sections 3.4 and 3.5, that two simple staircase mechanisms, which we call the binary mechanism and the randomized response mechanism, are optimal in the high and low privacy regimes, respectively, and well approximate the intermediate regime.

3.3.1 Optimality of staircase mechanisms

For an input alphabet $\mathcal{X}$ with $|\mathcal{X}| = k$, we represent the set of ε-locally differentially private mechanisms that lead to output alphabets $\mathcal{Y}$ with $|\mathcal{Y}| = \ell$ by

$$\mathcal{D}_{\varepsilon,\ell} = \mathcal{Q}_{k\times\ell} \cap \left\{ Q \ : \ \forall\, x, x' \in \mathcal{X}, S \subseteq \mathcal{Y}, \ \Big|\ \ln \frac{Q\,(S|x)}{Q\,(S|x')} \Big| \leq \varepsilon \right\},$$

where $\mathcal{Q}_{k\times\ell}$ denotes the set of all $k \times \ell$ dimensional conditional distributions. The set of all ε-locally differentially private mechanisms is given by

$$\mathcal{D}_\varepsilon = \cup_{\ell\in\mathbb{N}} \mathcal{D}_{\varepsilon,\ell}. \qquad (3.8)$$

The set of all conditional distributions acting on $\mathcal{X}$ is given by $\mathcal{Q} = \cup_{\ell\in\mathbb{N}} \mathcal{Q}_{k,\ell}$.

We consider two types of utility functions, one for the hypothesis testing setup and another for the mutual information setup. In the hypothesis testing setup, the utility is a function of the privatization mechanism and two priors defined on the input alphabet. Namely, $U\,(P_0, P_1, Q) : \mathbb{S}^k \times \mathbb{S}^k \times \mathcal{Q} \to \mathbb{R}_+$, where P_0 and P_1 are positive priors defined on $\mathcal{X}$ and $\mathbb{S}^k$ is the $(k-1)$-dimensional probability simplex. P_ν is said to be positive if $P_\nu\,(x) > 0$ for all $x \in \mathcal{X}$. In the information preservation setup, the utility is a function of the privatization mechanism and a prior defined on the input alphabet. Namely, $U(P,Q) : \mathbb{S}^k \times \mathcal{Q} \to \mathbb{R}_+$, where P is a positive prior defined on $\mathcal{X}$. For notational convenience, we will use $U(Q)$ to refer to both $U(P,Q)$ and $U(P_0, P_1, Q)$.

Definition 3.3.1 (Sublinear Functions) *A function $\mu(z) : \mathbb{R}^k \to \mathbb{R}$ is said to be sublinear if the following two conditions are met:*

1. *$\mu(\gamma z) = \gamma \mu(z)$ for all $\gamma \in \mathbb{R}_+$.*

2. *$\mu(z_1 + z_2) \leq \mu(z_1) + \mu(z_2)$ for all $z_1, z_2 \in \mathbb{R}$.*

Let Q_y be the column of Q corresponding to $Q(y|\cdot)$ and μ be any sublinear function. We are interested in utilities that can be decomposed as a summation of sublinear functions. We study the *fundamental tradeoff between privacy and utility* by solving the following constrained maximization problem:

$$\begin{array}{ll} \underset{Q}{\text{maximize}} & U(Q) = \sum_{y \in \mathcal{Y}} \mu(Q_y) \\ \text{subject to} & Q \in \mathcal{D}_\varepsilon \end{array} . \tag{3.9}$$

This includes maximization over information theoretic quantities of interest in statistical estimation and hypothesis testing such as mutual information, total variation, KL-divergence, and χ^2-divergence [23]. Since sub-linearity implies convexity, this is in general a complicated nonlinear program: We are maximizing (instead of minimizing) a convex function in Q. Further, the dimension of Q might be unbounded: the optimal privatization mechanism Q^* might produce an infinite output alphabet $\mathcal{Y}$. The following theorem proves that one never needs an output alphabet larger than the input alphabet in order to achieve the maximum utility, and provides a combinatorial representation of the optimal solution.

Theorem 3.3.2 *For any sublinear function μ and any $\varepsilon \geq 0$, there exists an optimal mechanism Q^* maximizing the utility in* (3.9) *over all ε-locally differentially private mechanisms, such that*

(*a*) *the output alphabet size is at most the input alphabet size, i.e. $|\mathcal{Y}| \leq |\mathcal{X}|$; and* (*b*) *for all* $y \in \mathcal{Y}$, *and* $x, x' \in \mathcal{X}$

$$\left| \ln \frac{Q^*(y|x)}{Q^*(y|x')} \right| \in \{0, \varepsilon\} \ . \tag{3.10}$$

The first claim of bounded alphabet size is more generally true for any general utility $U(Q)$ that is convex in Q (not necessarily decomposing into a sum of sublinear functions as in (3.9)). The second claim establishes that there is an optimal mechanism with an extremal structure; the absolute value of the log-likelihood ratios can only take one of the two extremal values: zero or e^ε (see Figure 3.2 for example). We refer to such a mechanism as a staircase mechanism, and define the *family of staircase mechanisms* formally as

$$\mathcal{S}_\varepsilon \ \equiv \ \{Q \,|\ \text{satisfying (3.10)}\} \ .$$

For all choices of $U(Q) = \sum_{\mathcal{Y}} \mu(Q_y)$ and any $\varepsilon \geq 0$, Theorem 3.3.2 implies that the family of staircase mechanisms includes the optimal solutions to maximization problems of the form (3.9). Notice that staircase mechanisms are ε-locally differentially private, since any Q satisfying (3.10) implies that $Q(y|x)/Q(y|x') \leq e^\varepsilon$.

$$Q^T = \tfrac{1}{1+e^\varepsilon} \begin{bmatrix} e^\varepsilon & e^\varepsilon & 1 & e^\varepsilon & 1 \\ 1 & 1 & e^\varepsilon & 1 & e^\varepsilon \end{bmatrix} \qquad Q^T = \tfrac{1}{3+e^\varepsilon} \begin{bmatrix} e^\varepsilon & 1 & 1 & 1 \\ 1 & e^\varepsilon & 1 & 1 \\ 1 & 1 & e^\varepsilon & 1 \\ 1 & 1 & 1 & e^\varepsilon \end{bmatrix}$$

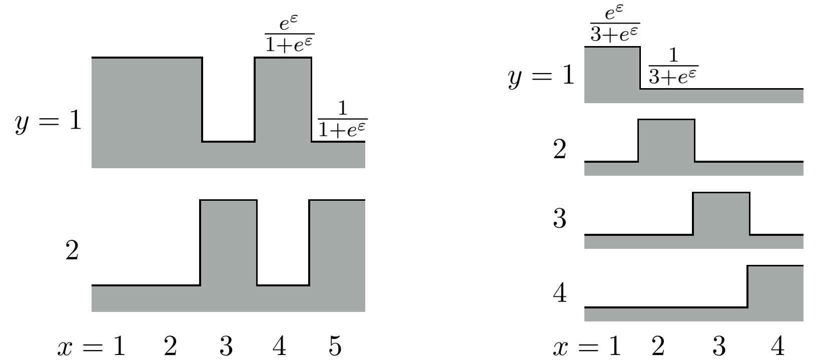

$\frac{e^\varepsilon}{1+e^\varepsilon}$
$\frac{1}{1+e^\varepsilon}$
$y = 1$
2
$x = 1$ 2 3 4 5
$\frac{e^\varepsilon}{3+e^\varepsilon}$
$\frac{1}{3+e^\varepsilon}$
$y = 1$
2
3
4
$x = 1$ 2 3 4

Figure 3.2: Examples of staircase mechanisms: the binary (left) and the randomized response (right) mechanisms.

For global differential privacy, we can generalize the definition of staircase mechanisms to hold for all neighboring database queries x, x' (or equivalently within some sensitivity), and recover all known existing optimal mechanisms. Precisely, the geometric mechanism shown to be optimal in [38], and the mechanisms shown to be optimal in [12, 13] (also called staircase mechanisms) are special cases of the staircase mechanisms defined above. We believe that the characterization of these extremal mechanisms and the analysis techniques developed in this chapter can be of independent interest to researchers interested in optimal mechanisms for global privacy and more general utilities.

3.3.2 Combinatorial representation of staircase mechanisms

Now that we know staircase mechanisms are optimal, we can try to combinatorially search for the best staircase mechanism for an instance of the function μ and a fixed ε. To this end, we give a simple representation of all staircase mechanisms, exploiting the fact that they are scaled copies of a finite number of patterns.

Let $Q \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{Y}|}$ be a staircase mechanism, and $k = |\mathcal{X}|$ denote the input alphabet size. Then, from the definition of staircase mechanisms, $Q(y|x)/Q(y|x') \in \{e^{-\varepsilon}, 1, e^{\varepsilon}\}$ and each column $Q(y|\cdot)$ must be proportional to one of the canonical staircase patterns we define next.

Definition 3.3.3 (Staircase Pattern Matrix) *Let b_j be the k-dimensional binary vector corresponding to the binary representation of j for $j \le 2^k - 1$. A matrix $S^{(k)} \in \{1, e^{\varepsilon}\}^{k\times 2^k}$ is called a staircase pattern matrix if the j-th column of $S^{(k)}$ is $S^{(k)}_j = (e^{\varepsilon} - 1)\, b_{j-1} + \mathbb{1}$, for $j \in \{1, \ldots, 2^k\}$. Each column of $S^{(k)}$ is a staircase pattern.*

When $k = 3$, there are $2^k = 8$ staircase patterns and the staircase pattern matrix is given by

$$S^{(3)} = \begin{bmatrix} 1 & 1 & 1 & 1 & e^{\varepsilon} & e^{\varepsilon} & e^{\varepsilon} & e^{\varepsilon} \\ 1 & 1 & e^{\varepsilon} & e^{\varepsilon} & 1 & 1 & e^{\varepsilon} & e^{\varepsilon} \\ 1 & e^{\varepsilon} & 1 & e^{\varepsilon} & 1 & e^{\varepsilon} & 1 & e^{\varepsilon} \end{bmatrix} .$$

For all values of k, there are exactly 2^k such patterns, and any column $Q(y|\cdot)$ of Q, a staircase mechanism, is a scaled version of one of the columns of $S^{(k)}$.

Using this pattern matrix, we will show that we can represent (an equivalence class of) any staircase mechanism Q as

$$Q \;=\; S^{(k)}\Theta\;, \tag{3.11}$$

where $\Theta = \mathrm{diag}(\theta)$ is a $2^k \times 2^k$ diagonal matrix and θ is a 2^k-dimensional vector representing the scaling of the columns of $S^{(k)}$. We can now formulate the problem of maximizing the utility as a linear program and prove their equivalence.

Theorem 3.3.4 *For any sublinear function μ and any $\varepsilon \geq 0$, the nonlinear program of* (3.9) *and the following linear program have the same optimal value:*

$$\begin{aligned} \underset{\theta\in\mathbb{R}^{2^k}}{\text{maximize}} \quad & \sum_{j=1}^{2^k} \mu(S_j^{(k)})\theta_j = \mu^T\theta \\ \text{subject to} \quad & S^{(k)}\theta \;=\; \mathbb{1} \\ & \theta \geq 0\;, \end{aligned} \tag{3.12}$$

and the optimal solutions are related by (3.11).

The infinite dimensional nonlinear program of (3.9) is now reduced to a finite dimensional linear program. The constraints in (3.12) ensure that we get a valid probability matrix $Q = S^{(k)}\Theta$ with rows that sum to one. One could potentially solve this LP with 2^k variables but its computational complexity scales exponentially in the alphabet size $k = |\mathcal{X}|$. For practical values of k this might not always be possible. However, in the following sections, we prove that in the high privacy regime ($\varepsilon \leq \varepsilon^*$ for some positive ε^*), there is a single optimal mechanism, which we call the *binary mechanism*, which dominates over all other mechanisms in a very strong sense for all utility functions of practical interest.

In order to understand the above theorem, observe that both the objective function and differential privacy constraints are invariant under *permutation* (or relabelling) of the columns of a privatization mechanism Q. Similarly, both the objective function and differential privacy constraints are invariant under *merging/splitting* of outputs with the same pattern. To be specific, consider a privatization mechanism Q and suppose there exist two outputs y

and y' that have the same pattern, i.e. $Q(y|\cdot) = C\,Q(y'|\cdot)$ for some positive constant C. Then, we can consider a new mechanism Q' by merging the two columns corresponding to y and y'. Let y'' denote this new output. It follows that Q' satisfies the differential privacy constraints and the resulting utility is also preserved. Precisely, using the fact that $Q(y|\cdot) = C\,Q(y'|\cdot)$, it follows that

$$\mu(Q_y) + \mu(Q_{y'}) \;=\; \mu((1+C)Q_y) \;=\; \mu(Q'_{y''})\;,$$

by the homogeneity of μ. We can naturally define equivalence classes for staircase mechanisms that are equivalent up to a permutation of columns and merging/splitting of columns with the same pattern:

$$\begin{aligned}[Q] = \{Q' \in \mathcal{S}_\varepsilon \mid \;& \exists \text{a sequence} \\ & \text{of permutations and merge/split of columns from } Q' \text{ to } Q\}\;.\end{aligned}$$

To represent an equivalence class, we use a mechanism in $[Q]$ that is ordered and merged to match the patterns of the pattern matrix $S^{(k)}$. For any staircase mechanism Q, there exists a possibly different staircase mechanism $Q' \in [Q]$ such that $Q' = S^{(k)}\Theta$ for some diagonal matrix Θ with nonnegative entries. Therefore, to solve optimization problems of the form (3.9), we can restrict our attention to such representatives of equivalent classes. Further, for privatization mechanisms of the form $Q = S^{(k)}\Theta$, the objective function takes the form $\sum_j \mu(S^{(k)}_j)\theta_j$, a simple linear function of Θ.

3.4 Private Hypothesis Testing

In this section, we study the fundamental tradeoff between local privacy and hypothesis testing. In this setting, there are n individuals each with data X_i from a distribution P_ν for a fixed $\nu \in \{0,1\}$. Let Q be a non-interactive privatization mechanism guaranteeing ε-local differential privacy. The output of the privatization mechanism Y_i is distributed according to the induced marginal M_ν defined in (3.2). With a slight abuse of notation, we will use M_ν and P_ν to represent both probability distributions and probability mass functions. The power to discriminate data from P_0 to the data from P_1

depends on the 'distance' between the marginals M_0 and M_1. To measure the ability of such statistical discrimination, our choice of utility of a privatization mechanism Q is an information theoretic quantity called Csiszár's f-divergence defined as

$$D_f(M_0||M_1) = \sum_{\mathcal{Y}} M_1(y) f\Big(\frac{M_0(y)}{M_1(y)}\Big) = U\left(P_0, P_1, Q\right) = U\left(Q\right) \ , \quad (3.13)$$

for some convex function f such that $f(1) = 0$. The Kullback-Leibler (KL) divergence $D_{\text{kl}}(M_0||M_1)$ is a special case of f-divergence with $f(x) = x \log x$, and total variation $\|M_0 - M_1\|_{\text{TV}}$ is a special case with $f(x) = (1/2)|x - 1|$. Note that the f-divergence is not a distance since it might not be symmetric or satisfy triangular inequality. We are interested in characterizing the optimal solution to

$$\underset{Q \in \mathcal{D}_\varepsilon}{\text{maximize}} \qquad D_f(M_0||M_1) \ , \qquad (3.14)$$

where $\mathcal{D}_\varepsilon$ is the set of all ε-locally differentially private mechanisms defined in (4.7).

A motivating example for this choice of utility is the Neyman-Pearson hypothesis testing framework [39]. Given the privatized views $\{Y_i\}_{i=1}^n$, the data analyst wants to test whether they are generated from M_0 or M_1. Let the null hypothesis be H_0 : Y_i's are generated from M_0, and the alternative hypothesis H_1 : Y_i's are generated from M_1. For a choice of rejection region $R \subseteq \mathcal{Y}^n$, the probability of false alarm (type I error) is $\alpha = M_0^n(R)$ and the probability of missed detection (type II error) is $\beta = M_1^n(\mathcal{Y}^n \setminus R)$. Let $\beta^\delta = \min_{R \subseteq \mathcal{Y}^n, \alpha < \alpha^*} \beta$ denote the minimum type II error achievable while keeping type I error rate at most α^*. According to Chernoff-Stein lemma [39], we know that

$$\lim_{n \to \infty} \frac{1}{n} \log \beta^{\alpha^*} = -D_{\text{kl}}(M_0||M_1) \ .$$

Suppose the analyst knows P_0, P_1, and Q. Then in order to achieve optimal asymptotic error rate, one would want to maximize the KL divergence of the induced marginals, over all ε-locally differentially private mechanisms Q. The results we present in this section (Theorems 3.4.1 and 3.4.4 to be precise) provide an explicit construction of optimal mechanisms in high and low privacy regimes. Using those optimality results, we prove a fundamental limit on the achievable error rates under differential privacy. Precisely, with data collected from an ε-locally differentially privatization mechanism, one cannot achieve an asymptotic type II error smaller than

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\log\beta^{\alpha^*} &\geq -\frac{(1+\delta)(e^\varepsilon-1)^2}{(e^\varepsilon+1)}\|P_0-P_1\|_{\mathrm{TV}}^2 \\ &\geq -\frac{(1+\delta)(e^\varepsilon-1)^2}{2(e^\varepsilon+1)}D_{\mathrm{kl}}(P_0||P_1)\ ,\end{aligned}$$

whenever $\varepsilon \leq \varepsilon^*$, where ε^* is dictated by Theorem 3.4.1 and $\delta > 0$ is some positive constant. In the equation above, the second inequality follows from Pinsker's inequality. Since $(e^\varepsilon-1)^2 = O(\varepsilon^2)$ for small ε, the effective sample size is now reduced from n to $\varepsilon^2 n$. This is the price of privacy. In the low privacy regime where $\varepsilon \geq \varepsilon^*$, for ε^* dictated by Theorem 3.4.4, one cannot achieve an asymptotic type II error smaller than

$$\lim_{n\to\infty}\frac{1}{n}\log\beta^{\alpha^*} \geq -D_{\mathrm{kl}}(P_0||P_1) + (1-\delta)G(P_0,P_1)e^{-\varepsilon}\ .$$

3.4.1 Optimal staircase mechanisms

From the definition of $D_f(M_0||M_1)$, we have that

$$D_f(M_0||M_1) = \sum_{\mathcal{Y}}(P_1^TQ_y)f(P_0^TQ_y/P_1^TQ_y) = \sum_{\mathcal{Y}}\mu\left(Q_y\right),$$

where $P_\nu^TQ_y = \sum_{\mathcal{X}}P_\nu\left(x\right)Q\left(y|x\right)$ and $\mu\left(Q_y\right) = (P_1^TQ_y)f(P_0^TQ_y/P_1^TQ_y)$. For any $\gamma > 0$,

$$\begin{aligned}\mu\left(\gamma Q_y\right) &= \left(P_1^T(\gamma Q_y)\right)f\left(P_0^T\left(\gamma Q_y\right)/P_1^T\left(\gamma Q_y\right)\right) \\ &= \gamma\left(P_1^TQ_y\right)f\left(P_0^TQ_y/P_1^TQ_y\right) \\ &= \gamma\mu\left(Q_y\right).\end{aligned}$$

Moreover, since the function $\phi(z,t) = tf\left(\frac{z}{t}\right)$ is convex in (z,t) for $0 \leq z,t \leq 1$, then μ is convex in Q_y. Convexity and homogeneity together imply sublinearity. Therefore, Theorems 3.3.2 and 3.3.4 apply to $D_f(M_0||M_1)$ and we have that staircases are optimal.

For a given P_0 and P_1, the *binary mechanism* is defined as a staircase mechanism with only two outputs $y \in \{0,1\}$ satisfying (see Figure 3.2)

$$Q(0|x) \;=\; \begin{cases} \dfrac{e^{\varepsilon}}{1+e^{\varepsilon}} & \text{if } P_0(x) \geq P_1(x) \;, \\ \dfrac{1}{1+e^{\varepsilon}} & \text{if } P_0(x) < P_1(x) \;. \end{cases}$$

$$Q(1|x) \;=\; \begin{cases} \dfrac{e^{\varepsilon}}{1+e^{\varepsilon}} & \text{if } P_0(x) < P_1(x) \;, \\ \dfrac{1}{1+e^{\varepsilon}} & \text{if } P_0(x) \geq P_1(x) \;. \end{cases} \tag{3.15}$$

Although this mechanism is extremely simple, perhaps surprisingly, we will establish that this is the optimal mechanism when a high level of privacy is required. Intuitively, the output is very noisy in the high privacy regime, and we are better off sending just one bit of information that tells you whether your data is more likely to have come from P_0 or P_1.

Theorem 3.4.1 *For any pair of distributions P_0 and P_1, there exists a positive ε^* that depends on P_0 and P_1 such that for any f-divergences and any positive $\varepsilon \leq \varepsilon^*$, the binary mechanism maximizes the f-divergence between the induced marginals over all ε-locally differentially private mechanisms.*

This implies that in the high privacy regime, which is a typical setting studied in much of differential privacy literature, the binary mechanism is a universally optimal solution for all f-divergences in (3.14). In particular this threshold ε^* is *universal*, in that it does not depend on the particular choice of which f-divergence we are maximizing. This is established by proving a very strong statistical dominance using Blackwell's celebrated result on comparisons of statistical experiments [11]. In a nutshell, we prove that any ε-differentially private mechanism for sufficiently small ε can be simulated from the output of the binary mechanism. Hence, the binary mechanism dominates over all other mechanisms and at the same time achieves the maximum divergence. A similar idea has been used previously in [24] to exactly characterize how much privacy degrades under composition.

The optimality of binary mechanisms is not just for high privacy regimes. The next theorem shows that it is *the* optimal solution of (3.14) for all ε, when the objective function is the total variation $D_f(M_0||M_1) = \|M_0 - M_1\|_{\text{TV}}$.

Theorem 3.4.2 *For any pair of distributions P_0 and P_1, and any $\varepsilon \geq 0$, the binary mechanism maximizes total variation of the induced marginals M_0 and M_1 among all ε-locally differentially private mechanisms.*

When maximizing the KL divergence between the induced marginals, we show that the binary mechanism still achieves good performance for $\varepsilon \leq C$ where C now does not depend on P_0 and P_1. For a special case of KL divergence, let OPT denote the maximum value of (3.14) and BIN denote the KL divergence when the binary mechanism is used. The next theorem shows that

$$\text{BIN} \;\; \geq \;\; \frac{1}{2(e^\varepsilon + 1)^2}\text{OPT} \, .$$

Theorem 3.4.3 *For any ε and for any pair of distributions P_0 and P_1, the binary mechanism is an $1/(2(e^\varepsilon + 1)^2)$ approximation of the maximum KL divergence of the induced marginals M_0 and M_1 among all ε-locally differentially private mechanisms.*

Note that $2(e^\varepsilon + 1)^2 \leq 32$ for $\varepsilon \leq 1$, and for any $\varepsilon \leq 1$ which is the typical regime of interest in differential privacy, we can always use the simple binary mechanism and the resulting divergence is at most a constant factor away from the optimal.

The *randomized response mechanism* is defined as a staircase mechanism with the same set of outputs as the input, $\mathcal{Y} = \mathcal{X}$, satisfying (see Figure 3.2)

$$Q(y|x) \;\; = \;\; \begin{cases} \dfrac{e^\varepsilon}{|\mathcal{X}| - 1 + e^\varepsilon} & \text{if } y = x \, , \\ \dfrac{1}{|\mathcal{X}| - 1 + e^\varepsilon} & \text{if } y \neq x \, . \end{cases} \tag{3.16}$$

It is a randomization over the same alphabet, and we are more likely to give an honest response. We view it as a multiple choice generalization of the randomized response method proposed by [21], assuming equal level of sensitivity for all choices. We establish that this is the optimal mechanism when a low level of privacy is required. Intuitively, the noise is small in the low privacy regime, and we want to send as much information about our current data as allowed, but no more. For a special case of maximizing KL divergence, we show that the *randomized response mechanism* is the optimal solution of (3.14) in the low privacy regime ($\varepsilon \geq \varepsilon^*$).

Theorem 3.4.4 *There exists a positive ε^* that depends on P_0 and P_1 such that for any P_0 and P_1, and all $\varepsilon \geq \varepsilon^*$, the randomized response mechanism maximizes the KL divergence between the induced marginals over all ε-locally differentially private mechanisms.*

3.4.2 Numerical experiments

A typical approach for achieving ε-local differential privacy is to add geometric noise with appropriately chosen variance. For an input with alphabet size $|\mathcal{X}| = k$, this amounts to relabelling the input as integers $\{1, \dots, k\}$ and adding geometric noise, i.e., $Q(y|x) = ((1-\varepsilon^{1/(k-1)})/(1+\varepsilon^{1/(k-1)}))\varepsilon^{|y-x|/(k-1)}$. The output is then truncated at 1 and k to preserve the support.

For 100 instances of randomly chosen P_0 and P_1 over input alphabet of size $|\mathcal{X}| = 6$, we compare the average performance of the binary, randomized response, and the geometric mechanisms to the optimal staircase mechanism. The optimal staircase mechanism is computed by solving the linear program in Equation (3.12) for each fixed pair (P_0, P_1) and ε. We plot (in Figure 3.3, left) the average performance measured by the normalized divergence $D_{\text{kl}}(M_0||M_1)/D_{\text{kl}}(P_0||P_1)$ for all 4 mechanisms. The average is taken over the 100 instances of P_0 and P_1. In the low privacy (large ε) regime, the randomized response achieves optimal performance as predicted, which converges to one. In the high privacy regime (small ε), the binary mechanism achieves optimal performance as predicted. In all regimes, both mechanisms significantly improve over the geometric mechanism.

To illustrate how much worse the binary and the randomized response mechanisms can be (relative to the optimal extremal mechanism), we plot (in Figure 3.3, right) the divergence under each mechanism normalized by the divergence under the optimal mechanism. This is done for all 100 instances of P_0 and P_1. In all instances, the binary mechanism is optimal for small ε and the randomized response mechanism is optimal for large ε. However, $D_{\text{kl}}(M_0||M_1)$ under the randomized response mechanism can be as bad as 10% of the optimal one (for small ε). Similarly, $D_{\text{kl}}(M_0||M_1))$ under the binary mechanism can be as bad as 25% of the optimal one (for large ε). To overcome this issue, we propose the following simple strategy: use the better among these two mechanisms. The performance of this strategy is illustrated

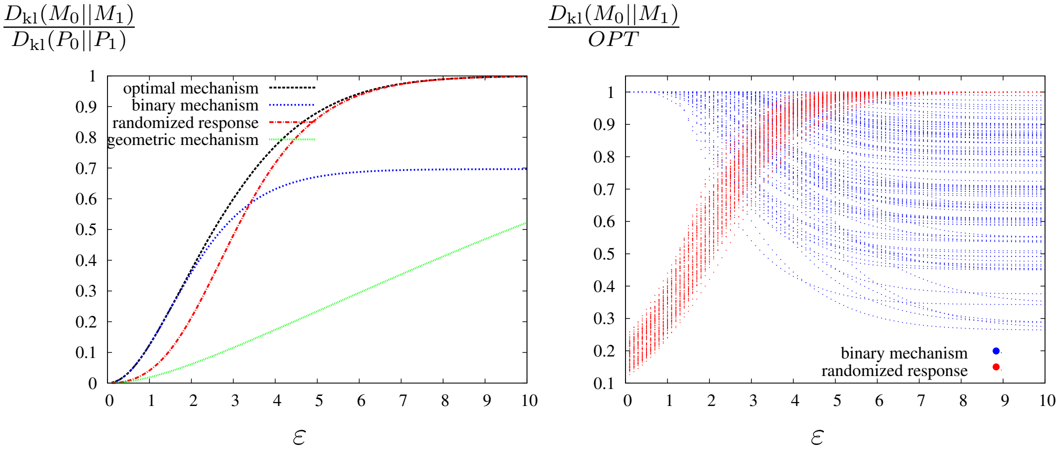


Figure 3.3: The binary and randomized response mechanisms are optimal in the high-privacy (small ε) and low-privacy (large ε) regimes, respectively, and improve over the geometric mechanism significantly (left). When the regimes are mismatched, $D_{kl}(M_0||M_1)$ under these mechanisms can be as bad as 10% of the optimal one (right).

in Figure 3.4. For various input alphabet size $|\mathcal{X}| \in \{3, 4, 5, 6\}$, we plot the performance of this mixed strategy for each value of ε and each of the 100 randomly generated instances of P_0 and P_1. This mixed strategy achieves 60% of the optimal divergence for all instances. Further, it is not sensitive to the size of the alphabet k. This strategy provides a good mechanism that can be readily used in practice for any value of ε.

3.4.3 Lower bounds

In this section, we provide converse results on the fundamental limit of differentially private mechanisms; these results follow from our main theorems and are of independent interest in other applications where lower bounds in statistical analysis are studied [40, 34, 41, 42]. For example, a bound similar to the one we present next was used to provide converse results on the sample complexity for statistical estimation with differentially private data in [22].

Corollary 3.4.5 *For any $\varepsilon \geq 0$, let Q be any conditional distribution that guarantees ε-local differential privacy. Then, for any pair of distributions P_0 and P_1 and any positive $\delta > 0$, there exists a positive ε^* that depends on P_0 and P_1 and δ such that for any $\varepsilon \leq \varepsilon^*$ the induced marginals M_0 and M_1*

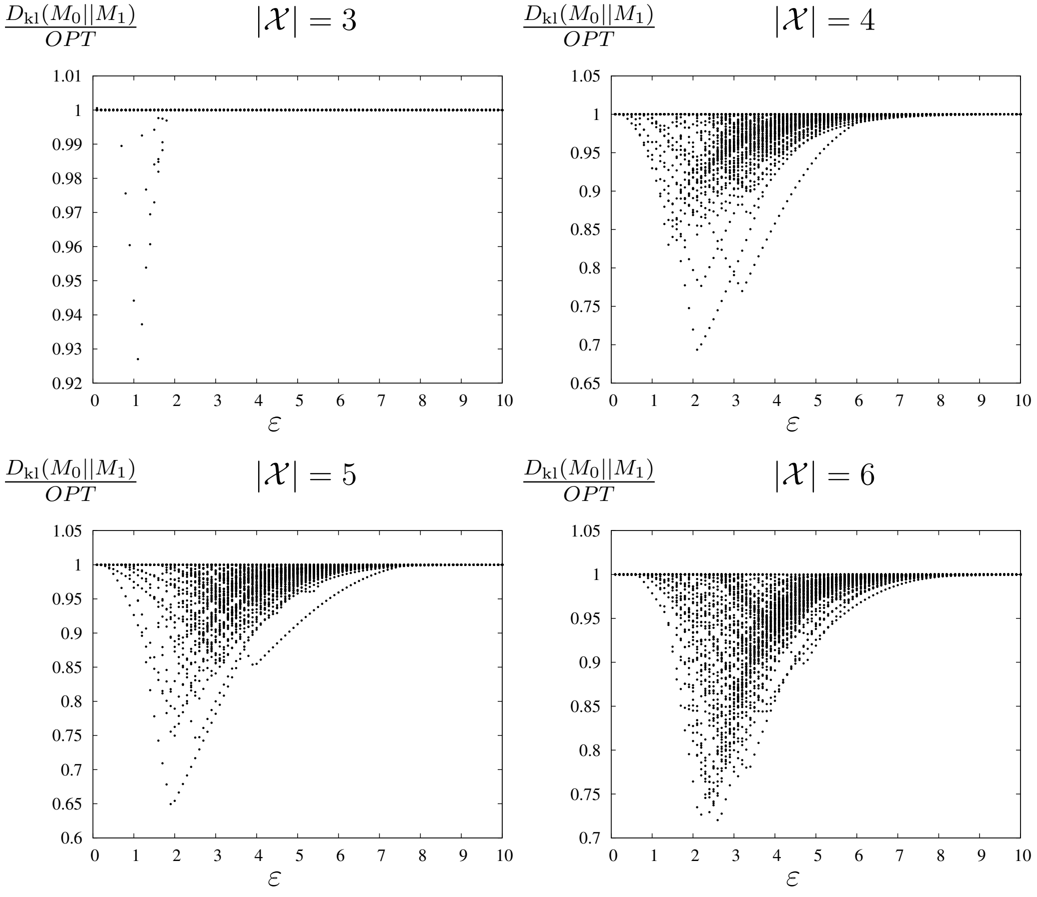


Figure 3.4: For varying input alphabet size $|\mathcal{X}| \in \{3, 4, 5, 6\}$, at least 60% of the optimal divergence can be achieved by taking the better one between the binary and the randomized response mechanisms.

satisfy the bound

$$D_{\text{kl}}\big(M_0||M_1\big) + D_{\text{kl}}\big(M_1||M_0\big) \quad \leq \quad \frac{2(1+\delta)(e^{\varepsilon}-1)^2}{(e^{\varepsilon}+1)} \, \big\|P_0 - P_1\big\|_{\text{TV}}^2 \, .$$

This follows from Theorem 3.4.1 and observing that the binary mechanism achieves

$$\begin{aligned} D_{\rm kl}\big(M_0\|M_1\big) & \\ = \;& \frac{(e^\varepsilon-1)P_0(T)+1}{e^\varepsilon+1}\log\Big(\frac{1+(e^\varepsilon-1)P_0(T)}{1+(e^\varepsilon-1)P_1(T)}\Big) \\ & \qquad\qquad + \frac{(e^\varepsilon-1)P_0(T^c)+1}{e^\varepsilon+1}\log\Big(\frac{1+(e^\varepsilon-1)P_0(T^c)}{1+(e^\varepsilon-1)P_1(T^c)}\Big) \\ = \;& \frac{(e^\varepsilon-1)^2}{e^\varepsilon+1}(P_0(T)-P_1(T))+O(\varepsilon^3) \\ = \;& \frac{(e^\varepsilon-1)^2}{e^\varepsilon+1}\,\big\|P_0-P_1\big\|_{\rm TV}^2+O(\varepsilon^3)\;, \end{aligned} \qquad (3.17)$$

where $T\subseteq\mathcal{X}$ is the set of x such that $P_0(x)\geq P_1(x)$. Compared to [22, Theorem 1], we recover their bound of $4(e^\varepsilon-1)^2\|P_0-P_1\|_{\rm TV}^2$ with a smaller constant. We want to note that Duchi et al.'s bound holds for all values of ε and uses a different technique of bounding the KL divergence directly, but no achievable mechanism has been provided. We instead provide an explicit mechanism that is optimal in the high privacy regime.

Similarly, in the low privacy regime, we can show the following converse result.

Corollary 3.4.6 *For any $\varepsilon\geq 0$, let Q be any conditional distribution that guarantees ε-local differential privacy. Then, for any pair of distributions P_0 and P_1 and any positive $\delta>0$, there exists a positive ε^* that depends on P_0 and P_1 and δ such that for any $\varepsilon\geq\varepsilon^*$ the induced marginals M_0 and M_1 satisfy the bound*

$$D_{\rm kl}\big(M_0\|M_1\big)+D_{\rm kl}\big(M_1\|M_0\big)\;\;\leq\;\; D_{\rm kl}(P_0\|P_1)-(1-\delta)G(P_0,P_1)e^{-\varepsilon}\;,$$

where $G(P_0,P_1)=\sum_{\mathcal{X}}(1-P_0(x))\log(P_1(x)/P_0(x))$.

This follows directly from Theorem 3.4.4 and observing that the randomized response mechanism achieves

$$D_{\rm kl}(M_0\|M_1)=D_{\rm kl}(P_0\|P_1)-G(P_0,P_1)e^{-\varepsilon}+O(e^{-2\varepsilon})\;. \qquad (3.18)$$

Similarly, for total variation, we can get the following converse result. This follows from Theorem 3.4.2 and explicitly computing the total variation achieved by the binary mechanism.

Corollary 3.4.7 *For any $\varepsilon \geq 0$, let Q be any conditional distribution that guarantees ε-local differential privacy. Then, for any pair of distributions P_0 and P_1, the induced marginals M_0 and M_1 satisfy the bound $\left\|M_0 - M_1\right\|_{\mathrm{TV}} \leq ((e^\varepsilon - 1)/(e^\varepsilon + 1)) \left\|P_0 - P_1\right\|_{\mathrm{TV}}$, and equality is achieved by the binary mechanism.*

Figure 3.5 illustrates the gap between the divergence achieved by the geometric mechanism described in the previous section and the optimal mechanisms (the binary mechanism for the high privacy regime and the randomized response mechanism for the low privacy regime). For each instance of the 100 randomly generated P_0 and P_1 over input of size $k = 6$, we plot the resulting divergence $D_{\mathrm{kl}}(M_0||M_1)$ as a function of $\|P_0 - P_1\|_{\mathrm{TV}}$ for $\varepsilon = 0.1$, and as a function of $D_{\mathrm{kl}}(P_0||P_1)$ for $\varepsilon = 10$. The binary and the randomized response mechanisms exhibit the scaling predicted by Equation (3.17) and (3.18), respectively.

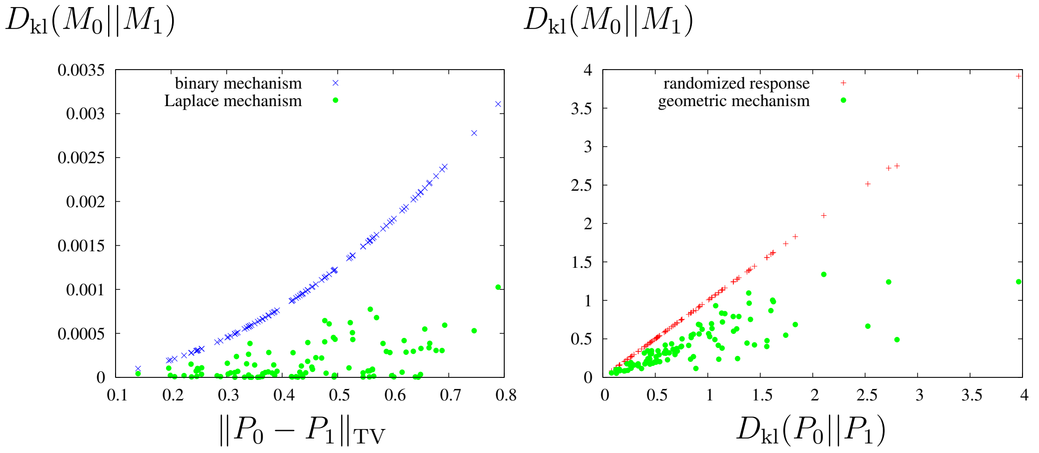


Figure 3.5: For small $\varepsilon = 0.1$ (left) the binary mechanism achieves the optimal KL divergence, which scales as Equation (3.17). For large $\varepsilon = 10$ (right) the randomized response achieves the optimal KL divergence, which scales as Equation (3.18). Both mechanisms improve significantly over the geometric mechanism.

3.5 Information Preservation

In this section, we study the fundamental tradeoff between local privacy and mutual information. Consider a random variable X distributed according to P. The information content in X is captured by entropy

$$H\left(X\right) = -\sum_{\mathcal{X}} P\left(x\right) \log P\left(x\right).$$

We are interested in releasing a differentially private version of X represented by Y. The random variable Y should preserve the information content of X as much as possible while meeting the local differential privacy constraints. Similar to the hypothesis testing setting, we will show that a variant of the binary mechanism is optimal in the high privacy regime and the randomized response mechanism is optimal in the low privacy regime.

Let Q be a non-interactive privatization mechanism guaranteeing ε-local differential privacy. The output of the privatization mechanism Y is distributed according to the induced marginal M given by

$$M(S) \;=\; \sum_{x\in\mathcal{X}} Q(S|x)P(x)\;,$$

for $S \subseteq \mathcal{Y}$. With a slight abuse of notation, we will use M and P to represent both probability distributions and probability mass functions. The information content in Y about X is captured by the well celebrated information theoretic quantity called mutual information. The mutual information between X and Y is given by

$$I\left(X;Y\right) = \sum_{\mathcal{X}}\sum_{\mathcal{Y}} P\left(x\right) Q\left(y|x\right) \log\left(\frac{Q\left(y|x\right)}{\sum_{l\in\mathcal{X}} P\left(l\right) Q\left(y|l\right)}\right) = U\left(Q\right). \quad (3.19)$$

Notice that $I\left(X;Y\right) \le H\left(X\right)$ and $I\left(X;Y\right)$ is convex in Q [39]. To preserve the information context in X, we wish to choose a privatization mechanism Q such that the mutual information between X and Y is maximized subject to differential privacy constraints. In other words, we are interested in characterizing the optimal solution to

$$\begin{array}{ll} \underset{Q}{\text{maximize}} & I\left(X;Y\right) \\ \text{subject to} & Q \in \mathcal{D}_{\varepsilon} \end{array}, \tag{3.20}$$

where $\mathcal{D}_{\varepsilon}$ is the set of all ε-locally differentially private mechanisms defined in (4.7). The above mutual information maximization problem can be thought of as a conditional entropy minimization problem since $I\left(X;Y\right) = H\left(X\right) - H\left(X|Y\right)$.

3.5.1 Optimal staircase mechanisms

From the definition of $I\left(X;Y\right)$, we have that

$$I\left(X;Y\right) = \sum_{\mathcal{Y}} \sum_{\mathcal{X}} P\left(x\right) Q\left(y|x\right) \log\left(\frac{Q\left(y|x\right)}{P^T Q_y}\right) = \sum_{\mathcal{Y}} \mu\left(Q_y\right),$$

where $P^T Q_y = \sum_{\mathcal{X}} P(x) Q\left(y|x\right)$ and

$$\mu\left(Q_y\right) = \sum_{\mathcal{X}} P\left(x\right) Q\left(y|x\right) \log\left(Q\left(y|x\right)/P^T Q_y\right).$$

Notice that $\mu\left(\gamma Q_y\right) = \gamma \mu\left(Q_y\right)$, and by the log-sum inequality, μ is convex. Convexity and homogeneity together imply sublinearity. Therefore, Theorems 3.3.2 and 3.3.4 apply to $I\left(X;Y\right)$ and we have that staircase mechanisms are optimal.

For a given P, the *binary mechanism for mutual information* is defined as a staircase mechanism with only two outputs $y \in \{0,1\}$ (see Figure 3.2). Let $T \subseteq \mathcal{X}$ be the set that partitions $\mathcal{X}$ into two partitions, T and T^c, such that $|P(T) - P(T^c)|$ is minimized. Precisely,

$$T \in \underset{A \subseteq \mathcal{X}}{\arg\min} \left| P(A) - \frac{1}{2} \right|. \tag{3.21}$$

Observe that there are always multiple choices for T. Indeed, for any minimizing set T, T^c is also a minimizing set since $|P(T) - 1/2| = |P(T^c) - 1/2|$. When there is only one such pair, the binary mechanism is uniquely defined as

$$Q(0|x) \;=\; \begin{cases} \frac{e^{\varepsilon}}{1+e^{\varepsilon}} & \text{if } x \in T\;, \\ \frac{1}{1+e^{\varepsilon}} & \text{if } x \notin T\;. \end{cases} \quad Q(1|x) \;=\; \begin{cases} \frac{e^{\varepsilon}}{1+e^{\varepsilon}} & \text{if } x \notin T\;, \\ \frac{1}{1+e^{\varepsilon}} & \text{if } x \in T\;. \end{cases} \tag{3.22}$$

When there are multiple pairs, any pair (T, T^c) can be chosen to define the binary mechanism. All resulting binary mechanisms are equivalent from a utility maximization perspective.

In what follows, we will establish that this simple mechanism is the optimal mechanism in the high privacy regime. Intuitively, in the high privacy regime, we cannot release more than one bit of information, and hence, the input alphabet is reduced to a binary output alphabet. In this case we have to maximize the information contained in the released bit by maximizing its entropy: $T \in \arg\max_{A \subseteq \mathcal{X}} \big(- P(A) \log P(A) - P(A^c) \log P(A^c) \big) = \arg\max_{A \subseteq \mathcal{X}} |P(A) - 1/2|$.

Theorem 3.5.1 *For any distribution P, there exists a positive ε^* that depends on P such that for any positive $\varepsilon \le \varepsilon^*$, the binary mechanism maximizes the mutual information between the input and the output of a privatization mechanism over all ε-locally differentially private mechanisms.*

This implies that in the high privacy regime, the binary mechanism is the optimal solution for (3.20).

Next, we show that the binary mechanism achieves near-optimal performance for all $(\mathcal{X}, P)$ and $\varepsilon \le 1$ even when $\varepsilon^* < 1$. Let OPT denote the maximum value of (3.20) and BIN denote the mutual information achieved by the binary mechanism given in (3.22). The next theorem shows that

$$\text{BIN} \;\; \ge \;\; \frac{1}{1+e^{\varepsilon}} \text{OPT}\;.$$

Theorem 3.5.2 *For any $\varepsilon \le 1$ and any distribution P, the binary mechanism is an $(1 + e^{\varepsilon})$-approximation of the maximum mutual information between the input and the output of a privatization mechanism among all ε-locally differentially private mechanisms.*

Note that $1 + e^{\varepsilon} \le 4$ for $\varepsilon \le 1$ which is a commonly studied regime in differential privacy applications. Therefore, we can always use the simple binary mechanism and the resulting mutual information is at most a constant factor away from the optimal.

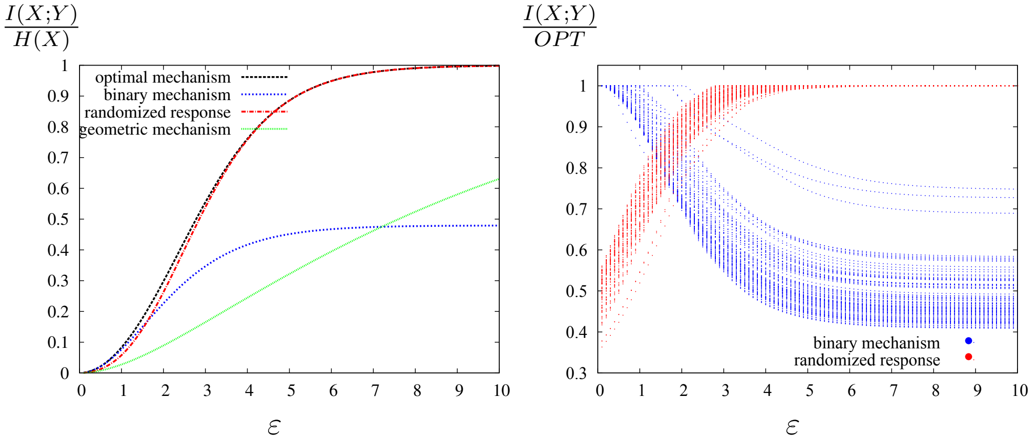


Figure 3.6: The binary and randomized response mechanisms are optimal in the high-privacy (small ε) and low-privacy (large ε) regimes, respectively, and improve over the geometric mechanism significantly (left). When the regimes are mismatched, $I(X;Y)$ under these mechanisms can each be as bad as 35% of the optimal one (right).

In the low privacy regime ($\varepsilon \geq \varepsilon^*$), the *randomized response mechanism* defined in(3.16) is optimal.

Theorem 3.5.3 *There exists a positive ε^* that depends on P such that for any distribution P and all $\varepsilon \geq \varepsilon^*$, the randomized response mechanism maximizes the mutual information between the input and the output of as privatization mechanism over all ε-locally differentially private mechanisms.*

3.5.2 Numerical experiments

For 100 instances of randomly chosen P defined over input alphabet of size $|\mathcal{X}| = 6$, we compare the average performance of the binary, randomized response, and the geometric mechanisms to the optimal mechanism. We plot (in Figure 3.6, left) the average performance measured by the normalized mutual information $I(X;Y)/H(X)$ for all 4 mechanisms. The average is taken over the 100 instances of P. In the low privacy (large ε) regime, the randomized response achieves optimal performance as predicted, which converges to one. In the high privacy regime (small ε), the binary mechanism achieves optimal performance as predicted. In all regimes, both mechanisms significantly improve over the geometric mechanism. To illustrate how much worse the binary and randomized response mechanisms can be (relative to the optimal staircase mechanism), we plot (in Figure 3.6, right) the mutual information under each mechanism normalized by the mutual information under the optimal staircase mechanism. This is done for all 100 instances of P. In all instances, the binary mechanism is optimal for small ε and the randomized response mechanism is optimal for large ε. However, $I(X;Y)$ under the randomized response mechanism can be as bad as 35% of the optimal one (for small ε). Similarly, $I(X;Y)$ under the binary mechanism can be as bad as 40% of the optimal one (for large ε).

For $|\mathcal{X}| \in \{3,4,5,6\}$, we plot (in Figure 3.7) the performance of the better between the binary and randomized response mechanisms normalized by the optimal mechanism for all 100 randomly generated instances of P. This mixed strategy achieves at least 75% of the optimal mutual infirmation for all instances of P. Moreover, it is not sensitive to the size of the alphabet $|\mathcal{X}|$.

3.5.3 Lower bounds

In this section, we provide converse results on the fundamental limit of locally differentially private mechanisms when utility is measured via mutual information.

Corollary 3.5.4 *For any $\varepsilon \geq 0$, let Q be any conditional distribution that guarantees ε-local differential privacy. Then, for any distribution P and any positive $\delta > 0$, there exists a positive ε^* that depends on P and δ such that for any $\varepsilon \leq \varepsilon^*$ the following bound holds:*

$$I(X;Y) \;\leq\; (1+\delta)\frac{1}{2}P(T)\,P(T^c)\,\varepsilon^2,$$

where T is defined in (3.21).

This follows from Theorem 3.5.1 (optimality of the binary mechanism) and observing that the binary mechanism achieves $I(X;Y)$ equal to

$$\begin{aligned}
&\frac{1}{e^\varepsilon+1}\left\{P(T)\,e^\varepsilon\log\frac{e^\varepsilon}{P(T^c)+e^\varepsilon P(T)} + P(T^c)\log\frac{1}{P(T^c)+e^\varepsilon P(T)}\right\}\\
&+\frac{1}{e^\varepsilon+1}\left\{P(T^c)\,e^\varepsilon\log\frac{e^\varepsilon}{P(T)+e^\varepsilon P(T^c)} + P(T)\log\frac{1}{P(T)+e^\varepsilon P(T^c)}\right\}\\
=\;& \frac{1}{2}P(T)\,P(T^c)\,\varepsilon^2 + O\left(\varepsilon^3\right). \qquad (3.23)
\end{aligned}$$

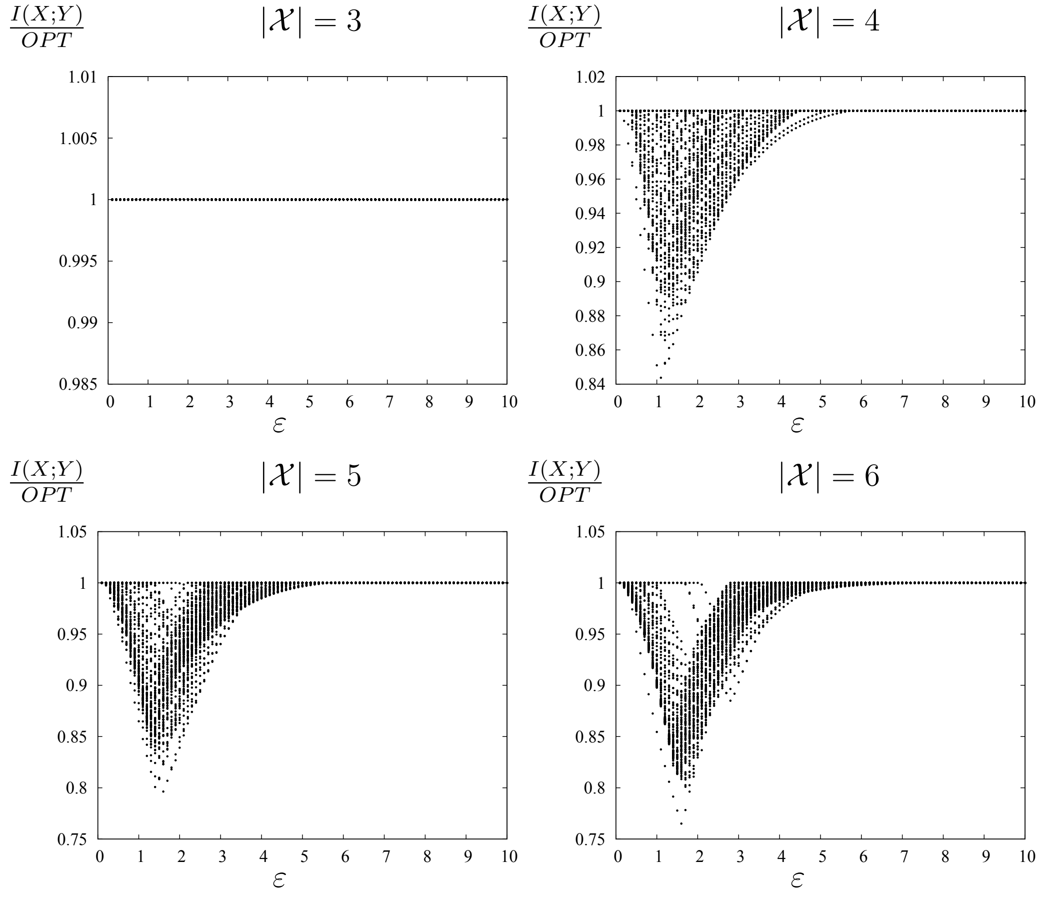


Figure 3.7: For varying input alphabet size $|\mathcal{X}| \in \{3, 4, 5, 6\}$, at least 75% of the maximum mutual information can be achieved by taking the better one between the binary and the randomized response mechanisms.

Similarly, in the low privacy regime, we can show the following converse result.

Corollary 3.5.5 *For any $\varepsilon \geq 0$, let Q be any conditional distribution that guarantees ε-local differential privacy. Then, for any distributions P and any positive $\delta > 0$, there exists a positive ε^* that depends on P and δ such that for any $\varepsilon \geq \varepsilon^*$ the following bound holds*

$$I\left(X;Y\right) \quad \leq \quad H\left(X\right) - \left(1-\delta\right)\left(k-1\right)\varepsilon e^{-\varepsilon}.$$

This follows directly from Theorem 3.5.3 (optimality of the randomized response mechanism) and observing that the randomized response mechanism achieves

$$I\left(X;Y\right) = H\left(X\right) - \left(k-1\right)\varepsilon e^{-\varepsilon} + O(e^{-2\varepsilon}). \tag{3.24}$$

Figure 3.8 illustrates the gap between the mutual information achieved by the geometric mechanism and the optimal mechanisms (the binary mechanism for the high privacy regime and the randomized response mechanism for the low privacy regime). For each instance of the 100 randomly generated P over input of size $k = 6$, we plot the resulting mutual information $I\left(X;Y\right)$ as a function of $P\left(T\right)P\left(T^c\right)$ for $\varepsilon = 0.1$, and as a function of $H\left(X\right)$ for $\varepsilon = 10$. The binary and the randomized response mechanisms exhibit the scaling predicted by Equations (3.23) and (3.24), respectively.

3.6 Approximate Local Differential Privacy

In this section, we generalize the results of the previous sections in the following ways:

1. We consider the class of utility functions that obey the data processing inequality. Consider the composition of two privatization mechanisms $QW = Q \circ W$ where the output of the first mechanism Q is applied to another mechanism W. We say that a utility function $U(\cdot)$ obeys the data processing inequality if the following inequality holds for all

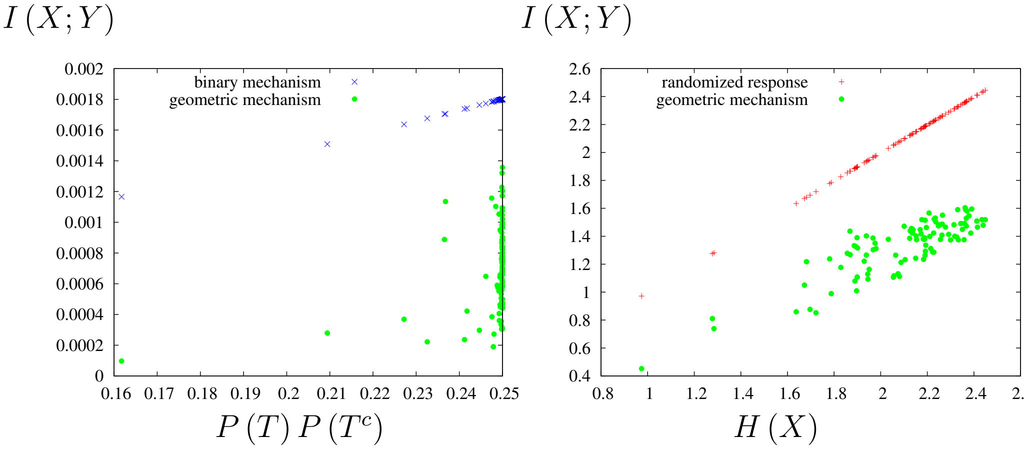


Figure 3.8: For small $\varepsilon = 0.1$ (left) the binary mechanism achieves the optimal mutual information, which scales as Equation (3.23). For large $\varepsilon = 10$ (right) the randomized response mechanism achieves the optimal mutual information, which scales as Equation (3.24). Both mechanisms improve significantly over the geometric mechanism.

Q and W:

$$U(QW) \;\; \leq \;\; U(Q)\,.$$

The following proposition, proved in [43], shows that the class of utilities obeying the data processing inequality includes all the utility functions we studied in Section 3.3.

Proposition 3.6.1 *Any utility function that can be written in the form of $U(Q) = \sum_{\mathcal{Y}} \mu(Q_y)$, where μ is any sublinear function, obeys the data processing inequality.*

2. We consider (ε, δ)-differential privacy which generalizes the notion of ε-differential privacy. (ε, δ)-differential privacy is commonly referred to as approximate differential privacy and it was first introduced in [9]. For the release of a random variable $X \in \mathcal{X}$, we say that a mechanism Q is (ε, δ)-locally differentially private if

$$Q(S|x) \leq e^{\varepsilon} Q(S|x') + \delta, \tag{3.25}$$

for all $S \subseteq \mathcal{Y}$ and all $x, x' \in \mathcal{X}$. Note that ε-local differential privacy is a special case of (ε, δ)-local differential privacy where $\delta = 0$.

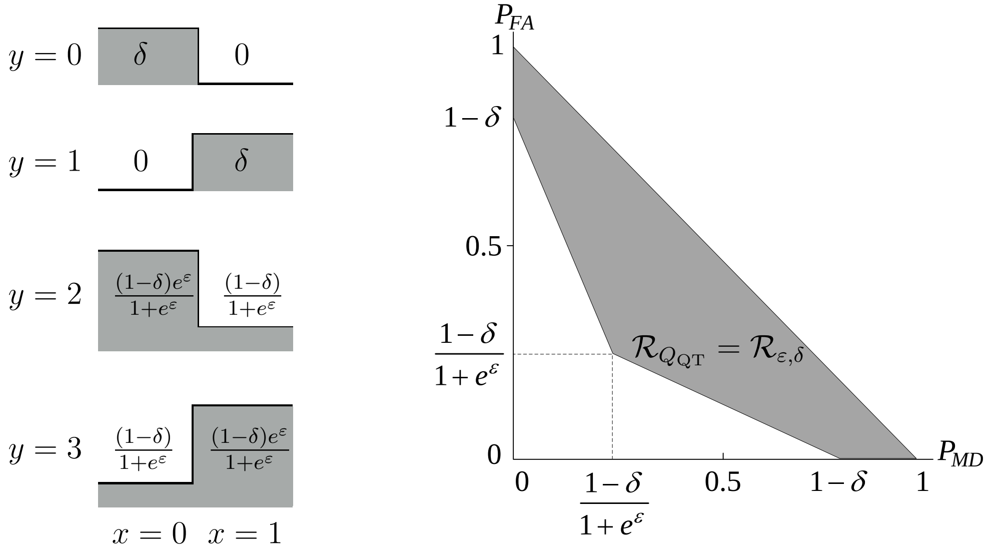
3. We prove that the *quaternary mechanism*, defined in Equation (3.26), is optimal for any ε and any δ. This is different from the treatment conducted in the previous sections where we proved the optimality of the binary (randomized response) mechanism for sufficiently small (large) ε and $\delta = 0$.

The treatment in this section, even though more general than the one in previous sections in the ways described above, holds only for binary alphabets (i.e., $|\mathcal{X}| = 2$). Finding optimal privatization mechanisms under (ε, δ)-local differential privacy for larger input alphabets (i.e., $|\mathcal{X}| > 2$) is an interesting open question. Unlike ε-local differential privacy, the privacy constraints under (ε, δ)-local differential privacy no longer decompose into separate constraints on each output y. This makes it difficult to generalize the techniques developed in previous sections of this chapter. However, for the special case of binary input alphabets, we can prove the optimality of one mechanism for all values of (ε, δ) and all utility functions that obey the data processing inequality.

For a binary random variable $X \in \mathcal{X} = \{0, 1\}$, the *quaternary mechanism* maps X to a quaternary random variable $Y \in \mathcal{Y} = \{0, 1, 2, 3\}$ and is defined as

$$\begin{aligned}
Q_{\text{QT}}(0|x) &= \begin{cases} \delta & \text{if } x = 0\ , \\ 0 & \text{if } x = 1\ , \end{cases} \\
Q_{\text{QT}}(1|x) &= \begin{cases} 0 & \text{if } x = 0\ , \\ \delta & \text{if } x = 1\ , \end{cases} \\
Q_{\text{QT}}(2|x) &= \begin{cases} (1-\delta)\dfrac{1}{1+e^{\varepsilon}} & \text{if } x = 0\ , \\ (1-\delta)\dfrac{e^{\varepsilon}}{1+e^{\varepsilon}} & \text{if } x = 1\ , \end{cases} \\
Q_{\text{QT}}(3|x) &= \begin{cases} (1-\delta)\dfrac{e^{\varepsilon}}{1+e^{\varepsilon}} & \text{if } x = 0\ , \\ (1-\delta)\dfrac{1}{1+e^{\varepsilon}} & \text{if } x = 1\ . \end{cases}
\end{aligned} \tag{3.26}$$

In other words, the quaternary mechanism passes X unchanged with probability δ and applies the binary mechanism (defined in previous sections) with probability $1 - \delta$. The main result of this section can be stated formally as follows.


$y = 0$
δ
0
$y = 1$
0
δ
$y = 2$
$\frac{(1-\delta)e^\varepsilon}{1+e^\varepsilon}$
$\frac{(1-\delta)}{1+e^\varepsilon}$
$y = 3$
$\frac{(1-\delta)}{1+e^\varepsilon}$
$\frac{(1-\delta)e^\varepsilon}{1+e^\varepsilon}$
$x = 0$
$x = 1$
P_{FA}
1
$1-\delta$
0.5
$\frac{1-\delta}{1+e^\varepsilon}$
0
$\mathcal{R}_{Q_{\text{QT}}} = \mathcal{R}_{\varepsilon,\delta}$
P_{MD}
0
$\frac{1-\delta}{1+e^\varepsilon}$
0.5
$1-\delta$
1

(a) Privatization mechanism

(b) Error region

Figure 3.9: The quaternary mechanism

Theorem 3.6.1 *If $|\mathcal{X}| = 2$, then for any ε, any δ, and any $U(Q)$ that obeys the data processing inequality, the quaternary mechanism maximizes $U(Q)$ subject to $Q \in \mathcal{D}_{(\varepsilon,\delta)}$, the set of all (ε,δ)-locally differentially private mechanism.*

The proof of Theorem 3.6.1 depends on an *operational definition* of differential privacy which we describe next. Consider a privatization mechanism Q that maps $X \in \{0,1\}$ stochastically to $Y \in \mathcal{Y}$. Given Y, construct a binary hypothesis test on whether $X = 0$ or $X = 1$. Any binary hypothesis test is completely described by a (possibly randomized) decision rule $\hat{X} : Y \to \{0,1\}$. The two types of error associated with $\hat{X}$ are *false alarm:* $\hat{X} = 1$ when $X = 0$, and *missed detection:* $\hat{X} = 0$ when $X = 1$. The probability of false alarm is given by $P_{\text{FA}} = \mathbb{P}(\hat{X} = 1|X = 0)$ while the probability of missed detection is given by $P_{\text{MD}} = \mathbb{P}(\hat{X} = 0|X = 1)$. For a fixed Q, the convex hull of all pairs $(P_{\text{MD}}, P_{\text{FA}})$ for all decision rules $\hat{X}$ defines a two-dimensional *error region* where P_{MD} is plotted against P_{FA}. For example, the quaternary mechanism given in Figure 3.9a has an error region $\mathcal{R}_{Q_{\text{QT}}}$ shown in Figure 4.1.

It turns out that (ε,δ)-local differential privacy imposes the following conditions on the error region of all (ε,δ)-locally differentially private mechanisms:

$$P_{\rm FA} + e^{\varepsilon} P_{\rm MD} \;\geq\; 1-\delta \;, \quad \text{and} \quad e^{\varepsilon} P_{\rm FA} + P_{\rm MD} \;\geq\; 1-\delta \;,$$

for any decision rule $\hat{X}$. These two conditions define an error region $\mathcal{R}_{\varepsilon,\delta}$ shown in Figure 4.1. Interestingly, the next theorem shows that the converse result is also true.

Theorem 3.6.2 *A mechanism Q is (ε,δ)-locally differentially private if and only if $\mathcal{R}_Q \subseteq \mathcal{R}_{\varepsilon,\delta}$.*

The proof of the above theorem can be found in [24]. Notice that it is no coincidence that $\mathcal{R}_{Q_{\rm QT}} = \mathcal{R}_{\varepsilon,\delta}$. This property will be essential to proving the optimality of the quaternary mechanism.

Theorem 3.6.2 allows us to benefit from the data processing inequality (DPI) and its converse, which follows from a celebrated result by [11]. These inequalities, while simple by themselves, lead to surprisingly strong technical results. Indeed, there is a long line of such a tradition in the information theory literature (see Chapter 17 of [39]). Consider two privatization mechanisms, $Q^{(1)}$ and $Q^{(2)}$. Let Y and Z denote the output of the mechanisms $Q^{(1)}$ and $Q^{(2)}$, respectively. We say that $Q^{(1)}$ dominates $Q^{(2)}$ if there exists a coupling of Y and Z such that X–Y–Z forms a Markov chain. In other words, we say $Q^{(1)}$ dominates $Q^{(2)}$ if there exists a stochastic mapping Q such that $Q^{(2)} = Q^{(1)} \circ Q$.

Theorem 3.6.3 *A mechanism $Q^{(1)}$ dominates a mechanism $Q^{(2)}$ if and only if $\mathcal{R}_{Q^{(2)}} \subseteq \mathcal{R}_{Q^{(1)}}$.*

The proof of the above theorem can be found in [11]. Observe that by Theorems 3.6.3 and 3.6.2, and the fact that $\mathcal{R}_{Q_{\rm QT}} = \mathcal{R}_{\varepsilon,\delta}$, the quaternary mechanism dominates any other differentially private mechanism. In other words, for any differentially private mechanism Q, there exists a stochastic mapping W such that $Q = W \circ Q_{\rm QT}$. Therefore, for any (ε,δ) and any utility function $U(.)$ obeying the data processing inequality, we have that $U(Q) \leq U(Q_{\rm QT})$. This finishes the proof of Theorem 3.6.1.

3.7 Conclusions and Summary

In this chapter, we have considered a broad class of convex utility functions and assumed a setting where individuals cannot collaborate (communicate with each other) before releasing their data. We showed that staircase mechanisms are optimal for a broad class of information theoretic utility functions such as mutual information and f-divergences. We also considered private binary hypothesis testing and information preservation, two canonical problems with a wide range of applications. Binary hypothesis testing and information preservation are two canonical problems with a wide range of applications. However, there are a number of non-trivial and interesting extensions to our work. These extensions are discussed in detail in Chapter 5.

It turns out that the techniques we have developed in this chapter can be generalized to find optimal privatization mechanisms in a setting where different individuals can collaborate interactively and each individual can be an analyst. This is precisely the topic of Chapter 4

CHAPTER 4

MULTI-PARTY DIFFERENTIAL PRIVACY

4.1 Introduction

Multi-party computation (MPC) is a generic framework where multiple parties share their information in an *interactive* fashion towards the goal of computing some functions, potentially different at each of the parties. In many situations of common interest, the key challenge is in computing the functions as *privately* as possible, i.e., without revealing much about one's information to the other (potentially colluding) parties. For instance, an interactive voting system aims to compute the majority of (say, binary) opinions of each of the parties, with each party being averse to declaring their opinion publicly. Another example involves banks sharing financial risk exposures – the banks need to agree on quantities such as the overnight lending rate which depends on each bank's exposure, which is a quantity the banks are naturally loath to truthfully disclose [44]. A central learning theory question involves characterizing the fundamental limits of interactive information exchange such that a strong (and suitably defined) adversary only learns as little as possible while still ensuring that the desired functions can be computed as accurately as possible.

One way to formulate the privacy requirement is to ensure that each party learns nothing more about the others' information than can be learned from the output of the function computed. This topic is studied under the rubric of *secure function evaluation* (SFE); the SFE formulation has been extensively studied with the goal of characterizing which functions can be securely evaluated [45, 46, 47, 48]. One drawback of SFE is that depending on what auxiliary information the adversary might have, disclosing the exact function output might reveal each party's data. For example, consider computing the average of the data owned by all the parties. Even if we use SFE, a party's data can be recovered if all the other parties collaborate. To ensure protection of the private data under such a strong adversary, we want to impose the stronger privacy guarantee of differential privacy. Recent breaches of sensitive information about individuals due to linkage attacks prove the vulnerability of existing ad-hoc privatization schemes, such as anonymization of the records. In linkage attacks, an adversary matches up anonymized records containing sensitive information with public records in a different dataset. Such attacks have revealed the medical record of a former governor of Massachusetts [49], the purchase history of Amazon users [50], genomic information [51], and movie viewing history of Netflix users [52].

An alternative formulation is differential privacy, a relatively recent formulation that has received considerable attention as a formal mathematical notion of privacy that provides protection against such strong adversaries (a recent survey is available at [53]). The basic idea is to introduce enough randomness in the communication so that an adversary possessing arbitrary side information and access to the entire transcript of the communication will still have some residual uncertainty in identifying any of the bits of the parties. This privacy requirement is strong enough that non-trivial functions will be computed only with some error. Thus, there is a great need for understanding the fundamental tradeoff between privacy and accuracy, and for designing privatization mechanisms and communication protocols that achieve the optimal tradeoffs. The formulation and study of an optimal framework addressing this tradeoff is the focus of this chapter.

We study the following problem of multi-party computation under differential privacy: each party possesses a single bit of information and the information bits are statistically independent. Each party is interested in computing a function, which could differ from party to party, and there could be a central observer (observing the entire transcript of the interactive communication protocol) interested in computing a separate function. Performance at each party and the central observer is measured via the accuracy of the function to be computed. We allow an arbitrary cost metric to measure the distortion between the true and the computed function values. Each party imposes a differential privacy constraint on its information bit (the privacy level could be different from party to party) – i.e., there remains an uncertainty in any specific party's bit even to an adversary that has access to the transcript of interactions and all the other parties' bits. The interactive communication is achieved via a broadcast channel that all parties and the central observer can hear (this modeling is without loss of generality since differential privacy protects against an adversary that can listen to the entire transcript, the communication between any two parties might as well be revealed to all the others). It is useful to distinguish between two types of communication protocols: *interactive* and *non-interactive*. We say a communication protocol is non-interactive if a message broadcasted by one party does not depend on the messages broadcasted by other parties. In contrast, interactive protocols allow the messages at any stage of the communication to depend on all the previous messages.

4.1.1 Our contributions

Our main result is the exact optimality of a simple non-interactive protocol in terms of maximizing accuracy for any given privacy levels: each party randomizes (sufficiently) its own bit and broadcasts the noisy version. Each party and the central observer then separately compute their respective decision functions to maximize the appropriate notion of their accuracy measure. The optimality is general: it holds for all types of functions, heterogeneous privacy conditions on the parties, all types of cost metrics, and both average and worst-case (over the inputs) measures of accuracy. Finally, the optimality result is *simultaneous*, in terms of maximizing accuracy at each of the parties and the central observer. Each party only needs to know its own desired level of privacy, its own function to be computed, and its measure of accuracy. Optimal data release and optimal decision making are naturally separated.

4.1.2 Related work

Private MPC was first addressed in [54]. The study of accuracy-privacy tradeoffs in the MPC context was first initiated by [40], which studies a paradigm where differential privacy and secure function evaluation (SFE) coexist: the function to be computed is decided on using differentially private schemes and the method to compute it is decided on using SFE. Specific functions, such as the SUM function, were studied under this setting, but no exact optimality results were provided.

In the context of two parties, privacy-accuracy tradeoffs have been studied in [55, 56] where a single function is computed by a "third-party" observing the transcript of the interactive protocol. [55] constructs natural functions that can only be computed very coarsely (using a natural notion of accuracy) when compared to a client-server model (which is essentially the single party setting). [56] shows that every any non-trivial privacy setting incurs some loss in the accuracy of a non-trivial Boolean function. Further, focusing on the specific scenario where each one of the two parties has a single bit of information, [56] characterizes the *exact* accuracy-privacy tradeoff for AND and XOR functions; the corresponding optimal protocol turns out to be *non-interactive*. However, this result was derived under some assumptions: only two parties are involved, the central observer is the only entity that computes a function, the function has to be either XOR or AND, symmetric privacy conditions are used for both parties, and accuracy is measured only as worst-case over the four possible inputs. Further, their analysis techniques do not generalize to the case when there are more than two parties.

Function approximation has been widely studied in the differential privacy literature under a centralized model where there is a single trusted entity owning a statistical database over a large number of individuals. In the centralized model, an algorithm is called interactive if it involves multiple rounds of communications between the server and the client. Under this centralized model, statistical learning has also been widely studied in differential privacy, e.g., classification [57, 58], k-means clustering [59], and principal component analysis [31, 60, 32, 33]. In particular, it has been shown in [57] that under the centralized setting, there exists a class of concepts that is efficiently learnable by *interactive* algorithms whereas a non-interactive algorithm requires exponential number of samples. In contrast, we consider a multi-party setting where the privacy barrier is place before each individual. In multi-party computation, all communication happens in multiple rounds, and a protocol is called interactive if one party's message depends on other party's previous messages. In this sense, the notion of interaction in multi-party computation is significantly different from what has been previously studied under centralized client-server settings.

4.2 Private Multi-Party Computation

Consider the setting where there are k parties, each with its own private binary data $x_i \in \{0,1\}$ generated independently. The independence assumption here is necessary because without it each party can learn something about others, which violates differential privacy, even without revealing any information. Differential privacy implicitly imposes independence in a multi-party setting. The goal of each party $i \in [k]$ is to compute an arbitrary function $f_i : \{0,1\}^k \to \mathcal{Y}$ of interest by interactively broadcasting messages. There might be a central observer who listens to all the messages being broadcasted, and wants to compute another arbitrary function $f_0 : \{0,1\} \to \mathcal{Y}$. The k parties are honest in the sense that once they agree on what protocol to follow, every party follows the rules. At the same time, they can be curious, and each party needs to ensure that other parties cannot learn its bit with sufficient confidence. This is done by imposing local differential privacy constraints. This setting is similar to the one studied in [22] in the sense that there are multiple privacy barriers, each one separating an individual party from the rest of the world. However, the main difference is that we consider multi-party computation, where there are multiple functions to be computed, and each node might possess a different function to be computed.

Let $x = [x_1, \ldots, x_k] \in \{0,1\}^k$ denote the vector of k bits, and $x_{-i} = [x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k] \in \{0,1\}^{k-1}$ be the vector of bits except for the i^{th} bit. The parties agree on an interactive protocol P to achieve the goal of multi-party computation. A 'transcript' τ is the output of P, and it contains the sequence of messages exchanged between the parties. Let the probability that a transcript τ is broadcasted (via a series of interactive communications) when the data is x be denoted by $P_{x,\tau} = \mathbb{P}(\tau \,|\, x)$ for $x \in \{0,1\}^k$ and for $\tau \in \mathcal{T}$. Then, a protocol can be represented as a matrix denoting the probability distribution over a set of transcripts $\mathcal{T}$ conditioned on x: $P = [P_{x,\tau}] \in [0,1]^{2^k \times |\mathcal{T}|}$.

In the end, each party makes a decision on what the value of function f_i is, based on its own bit x_i and the transcript τ that was broadcasted. A decision rule is a mapping from a transcript $\tau \in \mathcal{T}$ and private bit $x_i \in \{0,1\}$ to a decision $y \in \mathcal{Y}$ represented by a function $\hat{f}_i(\tau, x_i)$. We allow randomized decision rules, in which case $\hat{f}_i(\tau, x_i)$ can be a random variable. For the central observer, a decision rule is a function of just the transcript, denoted by a function $\hat{f}_0(\tau)$.

We consider two notions of accuracy: the average accuracy and the worst-case accuracy. For the i^{th} party, consider an accuracy measure $w_i : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ (or equivalently a negative cost function) such that $w_i(f_i(x), \hat{f}_i(\tau, x_i))$ measures the accuracy when the function to be computed is $f_i(x)$ and the approximation is $\hat{f}_i(\tau, x_i)$. Then the average accuracy for this i^{th} party is defined as

$$\begin{aligned} &\mathrm{ACC}_{\mathrm{ave}}(P, w_i, f_i, \hat{f}_i) \equiv \\ \frac{1}{2^k} \sum_{x \in \{0,1\}^k} &\mathbb{E}_{\hat{f}_i, P_{x,\tau}} [w_i(f_i(x), \hat{f}_i(\tau, x_i))] \ , \end{aligned} \tag{4.1}$$

where the expectation is taken over the random transcript τ and any randomness in the decision function $\hat{f}_i$. For example, if the accuracy measure is an indicator such that $w_i(y, y') = \mathbb{I}_{(y=y')}$, then $\mathrm{ACC}_{\mathrm{ave}}$ measures the average probability of getting the correct function output. For a given protocol P, it takes $(2^k\,|\mathcal{T}|)$ operations to compute the optimal decision rule:

$$f^*_{i,\mathrm{ave}}(\tau, x_i) \;=\; \arg\max_{y \in \mathcal{Y}} \sum_{x_{-i} \in \{0,1\}^{k-1}} P_{x,\tau}\, w_i(f_i(x), y) \ , \tag{4.2}$$

for each $i \in [k]$. The computational cost of $(2^k\,|\mathcal{T}|)$ for computing the optimal decision rule is *unavoidable in general*, since that is the inherent complexity of the problem: describing the distribution of the transcript requires the same cost. We will show that the optimal protocol requires a set of transcripts of size $|\mathcal{T}| = 2^k$, and the computational complexity of the decision rule for a general function is 2^{2k}. However, for a fixed protocol, this decision rule needs to be computed only once before any message is transmitted. Further, it is also possible to find a closed form solution for the decision rule when f has a simple structure. One example is the XOR function where the optimal decision rule is as simple as evaluating the XOR of all the received bits, which requires $O(k)$ operations. When there are multiple maximizers y, we can choose either one of them arbitrarily, and it follows that there is no gain in randomizing the decision rule for average accuracy.

Similarly, the worst-case accuracy is defined as

$$\begin{aligned}\mathrm{ACC}_{\mathrm{wc}}(P, w_i, f_i, \hat{f}_i) \equiv \\ \min_{x\in\{0,1\}^k} \mathbb{E}_{\hat{f}_i, P_{x,\tau}}\left[w_i(f_i(x), \hat{f}_i(\tau, x_i)) \right] .\end{aligned} \tag{4.3}$$

For worst-case accuracy, given a protocol P, the optimal decision rule of the i^{th} party with a bit x_i can be computed by solving the following convex program:

$$\begin{aligned} Q^{(x_i)} = & \\ \arg\max_{Q\in\mathbb{R}^{|\mathcal{T}|\times|\mathcal{Y}|}} & \quad \min_{x_{-i}\in\{0,1\}^{k-1}} \sum_{\tau\in\mathcal{T}} \sum_{y\in\mathcal{Y}} P_{x,\tau}\, w_i(f_i(x), y) Q_{\tau,y} \\ \text{subject to} & \quad \sum_{y\in\mathcal{Y}} Q_{\tau,y} = 1 \ , \ \forall \tau \in \mathcal{T} \text{ and } Q \geq 0. \end{aligned} \tag{4.4}$$

The optimal (random) decision rule $f^*_{i,\mathrm{wc}}(\tau, x_i)$ is to output y given transcript τ according to $\mathbb{P}(y|\tau, x_i) = Q^{(x_i)}_{\tau,y}$. This can be formulated as a linear program with $|\mathcal{T}| \times |\mathcal{Y}|$ variables and $2^k + |\mathcal{T}|$ constraints. Again, it is possible to find a closed form solution for the decision rule when f has a simple structure: for the XOR function, the optimal decision rule is again evaluating the XOR of all the received bits requiring $O(k)$ operations.

For a central observer, the accuracy measures are defined similarly, and the optimal decision rule is now

$$f^*_{0,\mathrm{ave}}(\tau) \quad = \quad \arg\max_{y\in\mathcal{Y}} \sum_{x\in\{0,1\}^k} P_{x,\tau}\, w_0(f_0(x), y) \ , \tag{4.5}$$

and for worst-case accuracy the optimal (random) decision rule $f^*_{0,\mathrm{wc}}(\tau)$ is to output y given transcript τ according to $\mathbb{P}(y|\tau) = Q^{(0)}_{\tau,y}$.

$$\begin{aligned} Q^{(0)} = & \\ \arg\max_{Q\in\mathbb{R}^{|\mathcal{T}|\times|\mathcal{Y}|}} & \quad \min_{x\in\{0,1\}^{k}} \sum_{\tau\in\mathcal{T}} \sum_{y\in\mathcal{Y}} P_{x,\tau}\, w_0(f_0(x), y) Q_{\tau,y} \\ \text{subject to} & \quad \sum_{y\in\mathcal{Y}} Q_{\tau,y} = 1 \ , \ \forall \tau \in \mathcal{T} \text{ and } Q \geq 0, \end{aligned} \tag{4.6}$$

where $w_0 : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is the measure of accuracy for the central observer.

4.3 Differentially Private Multi-Party Computation

Privacy is measured by approximate differential privacy [6, 7]. Since we allow for heterogeneous privacy constraints across parties, we use $(\varepsilon_i, \delta_i)$ to denote the desired privacy level of the i^{th} party. We say that a protocol P is $(\varepsilon_i, \delta_i)$-differentially private for the i^{th} party if for $x_i, x_i' \in \{0,1\}$, $x_{-i} \in \{0,1\}^{k-1}$, and $S \subseteq \mathcal{T}$, we have that

$$\mathbb{P}(\tau \in S | x_i, x_{-i}) \quad \leq \quad e^{\varepsilon_i}\, \mathbb{P}(\tau \in S | x_i', x_{-i}) + \delta_i \; . \tag{4.7}$$

A mechanism P is differentially private if it is $(\varepsilon_i, \delta_i)$-differentially private for all $i \in [k]$. Differential privacy ensures that no adversary can infer the private data x_i with high enough confidence, no matter what auxiliary information or computational power she might have.

Consider the following simple protocol known as the *randomized response*, which is a term first coined by [21] and commonly used in many private communications including the multi-party setting [55]. We will show in Section 4.4 that this is the optimal protocol that simultaneously maximizes the accuracy for all the parties. Each party broadcasts a randomized version of its bit denoted by $\tilde{x}_i$ such that

$$\tilde{x}_i = \begin{cases} 0 & \text{if } x_i = 0 \text{ with probability } \delta_i \; , \\ 1 & \text{if } x_i = 0 \text{ with probability } \dfrac{(1-\delta_i)e^{\varepsilon_i}}{1+e^{\varepsilon_i}} \; , \\ 2 & \text{if } x_i = 0 \text{ with probability } \dfrac{(1-\delta_i)}{1+e^{\varepsilon_i}} \; , \\ 3 & \text{if } x_i = 0 \text{ with probability } 0 \; , \end{cases}$$

$$\tilde{x}_i = \begin{cases} 0 & \text{if } x_i = 1 \text{ with probability } 0 \; , \\ 1 & \text{if } x_i = 1 \text{ with probability } \dfrac{(1-\delta_i)}{1+e^{\varepsilon_i}} \; , \\ 2 & \text{if } x_i = 1 \text{ with probability } \dfrac{(1-\delta_i)e^{\varepsilon_i}}{1+e^{\varepsilon_i}} \; , \\ 3 & \text{if } x_i = 1 \text{ with probability } \delta_i \; . \end{cases} \tag{4.8}$$

The proof of optimality of this randomized response depends on an operational definition of differential privacy which we now present.

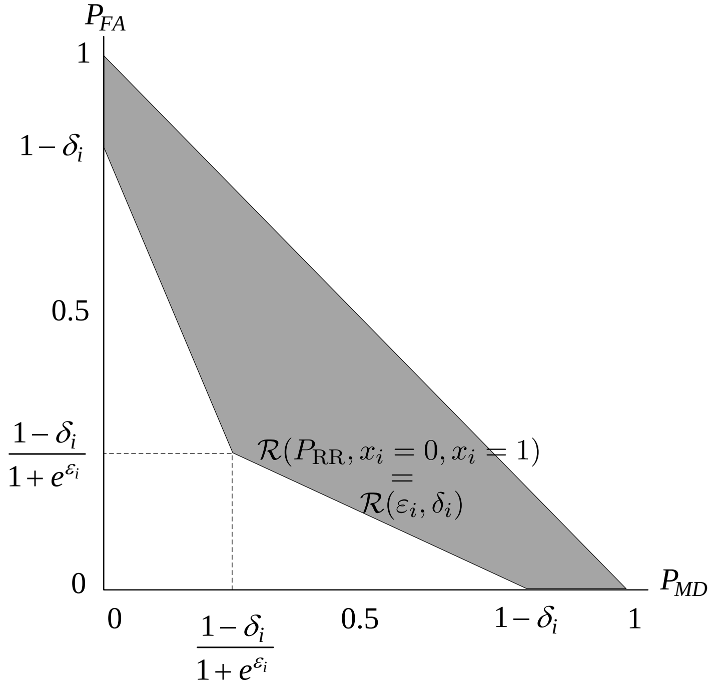


Figure 4.1: Error region dictated by $(\varepsilon_i, \delta_i)$-differential privacy

Given a broadcasted transcript τ and x_{-i} (all private bits except for x_i), construct a binary hypothesis test on whether $x_i = 0$ or $x_i = 1$. A binary hypothesis test is completely characterized by a (possibly randomized) decision rule $\hat{x}_i : (\tau, x_{-i}) \to \{0, 1\}$. The two types of error associated with $\hat{x}_i$ are: (1) false alarm: $\hat{x}_i = 1$ when $x_i = 0$, and (2) missed detection: $\hat{x}_i = 0$ when $x_i = 1$. The probability of false alarm is given by $P_{\mathrm{FA}} = \mathbb{P}(\hat{x}_i = 1 | x_i = 0)$ while the probability of missed detection is given by $P_{\mathrm{MD}} = \mathbb{P}(\hat{x}_i = 0 | x_i = 1)$. For a fixed privacy protocol P, the convex hull of all pairs $(P_{\mathrm{MD}}, P_{\mathrm{FA}})$ for all decision rules $\hat{x}_i$ defines a two-dimensional error region where P_{MD} is plotted against P_{FA}. For example, the randomized response mechanism P_{RR} given in (4.8) has an error region $\mathcal{R}(P_{\mathrm{RR}}, x_i = 0, x_i = 1)$ shown in Figure 4.1.

The differential privacy constraints in Equation (4.7) impose the following conditions on the error regions of all $(\varepsilon_i, \delta_i)$-differentially private protocols:

$$\begin{aligned} P_{\mathrm{FA}} + e^{\varepsilon_i} P_{\mathrm{MD}} &\geq 1 - \delta_i, \\ e^{\varepsilon_i} P_{\mathrm{FA}} + P_{\mathrm{MD}} &\geq 1 - \delta_i, \end{aligned}$$

for any decision rule $\hat{x}_i$ and any $i \in [k]$. The above two conditions define an error region $\mathcal{R}(\varepsilon_i, \delta_i)$ shown in Figure 4.1. Interestingly, the next theorem shows that the converse result is also true.

Lemma 4.3.1 *A mechanism P is differentially private if and only if $\mathcal{R}(P, x_i = 0, x_i = 1) \subseteq \mathcal{R}(\varepsilon_i, \delta_i)$ for all $i \in [k]$.*

The proof of the above lemma can be found in [43] (see Corollary 2.3 on page 4). Notice that it is no coincidence that $\mathcal{R}(P_{\text{RR}}, x_i = 0, x_i = 1) = \mathcal{R}(\varepsilon_i, \delta_i)$ (see Figure 4.1). This property will be essential in proving the optimality of the randomized response.

Lemma 4.3.1 allows us to benefit from the data processing inequality (DPI) and its converse, which follows from a celebrated result by [11]. These inequalities, while simple by themselves, lead to surprisingly strong technical results. Indeed, there is a long line of such a tradition in the information theory literature (see Chapter 17 of [39]).

Recall that τ contains the sequence of messages broadcasted by all k parties. Let $\tau(i)$ represent the messages broadcasted by the i^{th} party and observe that $\tau = \{\tau(1), \cdots, \tau(k)\}$. Consider two privatization protocols, P_1 and P_2, and let τ_1 and τ_2 denote the output transcripts under protocols P_1 and P_2, respectively. We say that P_1 dominates P_2 if there exists a sequence of stochastic transformations $\{W_1, \cdots, W_k\}$ such that for all $i \in [k]$, given x_{-i}, τ_2 can be simulated by applying W_i to $\tau_1(i)$ and x_{-i}. In other words, given x_{-i}, $W_i(\tau_1(i), x_{-i})$ has the same distribution as τ_2 .

Lemma 4.3.2 *A multi-party privacy protocol P_1 dominates a protocol P_2 if and only if $\mathcal{R}(P_2, x_i = 0, x_i = 1) \subseteq \mathcal{R}(P_1, x_i = 0, x_i = 1)$ for all $i \in [k]$.*

The proof of the above lemma can be found in [11]. Lemma 4.3.2 will be critical in proving the optimality of the randomized response.

Corollary 4.3.3 *Any differentially private protocol P is dominated by the randomized response P_{RR} given in Equation* (4.8)*. Therefore, there exists a sequence of stochastic transformations $\{W_1, \cdots, W_k\}$ such that $W_i(\tilde{x}_i, x_{-i})$ has the same distribution as τ for all $i \in [k]$.*

Corollary 4.3.3 follows from Lemma 4.3.1, Lemma 4.3.2, and the fact that $\mathcal{R}(\varepsilon_i, \delta_i) = \mathcal{R}(P_{\mathrm{RR}}, x_i = 0, x_i = 1)$ for all $i \in [k]$.

4.4 Optimal Mechanisms for Multi-Party Differential Privacy

We show, perhaps surprisingly, that the simple randomized response presented in (4.8) is the unique optimal protocol in a very general sense.

Theorem 4.4.1 *Let the optimal decision rule be defined as in* (4.2) *for the average accuracy and* (4.5) *for the worst-case accuracy. Then, for any privacy levels* $(\varepsilon_i, \delta_i)$*, any function* $f_i : \{0,1\}^k \to \mathcal{Y}$*, and any accuracy measure* $w_i : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ *for* $i \in [k]$*, together with the optimal decision rule, the randomized response achieves the maximum accuracy for the* i^{th} *party among all differentially private interactive and non-interactive protocols. For the central observer, the randomized response with the optimal decision rule defined in* (4.5) *and* (4.7) *achieves the maximum accuracy among all* $\{(\varepsilon_i, \delta_i)\}$*-differentially private interactive protocols and all decision rules for any arbitrary function* f_0 *and any measure of accuracy* w_0*.*

This is a strong optimality result. Every party and the central observer can simultaneously achieve the optimal accuracy, using a universal randomized response. Each party only needs to know its own desired level of privacy, its own function to be computed, and its measure of accuracy. Optimal data release and optimal decision making are naturally separated. It does not follow immediately that such a simple non-interactive randomized response mechanism would achieve the maximum accuracy. The proof critically harnesses the data processing inequalities and is provided in Appendix C.1.

4.5 Private Multi-Party XOR Computation

For a given function and a given accuracy measure, analyzing the performance of the optimal protocol provides the exact nature of the privacy-accuracy tradeoff. Consider a scenario where a central observer wants to compute the XOR of all the k-bits, each of which is ε-differentially private. In this special case, we can apply our main theorem to analyze the accuracy exactly in a combinatorial form.

Corollary 4.5.1 *Consider k-party computation for $f_0(x) = x_1 \oplus \cdots \oplus x_k$, and the accuracy measure is one if correct and zero if not, i.e. $w_0(0,0) = w_0(1,1) = 1$ and $w_0(0,1) = w_0(1,0) = 0$. For any $\{\lambda = e^{\varepsilon}\}$-differentially private protocol P and any decision rule $\hat{f}$, the average and worst-case accuracies are bounded by*

$$\begin{aligned} \mathrm{ACC}_{\mathrm{ave}}(P, w_0, f_0, \hat{f}_0) &\leq \frac{\sum_{i=0}^{\lfloor k/2 \rfloor} \binom{k}{2i} \lambda^{k-2i}}{(1+\lambda)^k} \;, \\ \mathrm{ACC}_{\mathrm{wc}}(P, w_0, f_0 \hat{f}_0) &\leq \frac{\sum_{i=0}^{\lfloor k/2 \rfloor} \binom{k}{2i} \lambda^{k-2i}}{(1+\lambda)^k} \;, \end{aligned}$$

and the equality is achieved by the randomized response and optimal decision rules in (4.5) *and* (4.7).

We prove the above corollary in Section C.2. The optimal decision for both accuracies is simply to output the XOR of the received privatized bits. This is a strict generalization of a similar result in [56], where XOR computation was studied but only for a two-party setting. In the high privacy regime, where $\varepsilon \simeq 0$ (equivalently $\lambda = e^{\varepsilon} \simeq 1$), this implies that $\mathrm{ACC}_{\mathrm{ave}} = 0.5 + 2^{-(k+1)}\varepsilon^k + O(\varepsilon^{k+1})$. The leading term is due to the fact that we are considering an accuracy measure of a Boolean function. The second term of $2^{-(k+1)}\varepsilon^k$ captures the effect that we are essentially observing the XOR through k consecutive binary symmetric channels with flipping probability $\lambda/(1+\lambda)$. Hence, the accuracy gets exponentially worse in k. On the other hand, if those k-parties are allowed to collaborate, then they can compute the XOR in advance and only transmit the privatized version of the XOR, achieving accuracy of $\lambda/(1+\lambda) = 0.5 + (1/4)\varepsilon^2 + O(\varepsilon^3)$. This is always better than not collaborating, which is the bound in Corollary 4.5.1.

4.6 Generalization to Multiple Bits

As an example, consider the first party with one bit x and the second party with two bits y_1 and y_2. Each bit needs to be protected as per ε-differential privacy. A central observer wishes to compute the following function:

$$f(x, y_1, y_2) \;=\; \begin{cases} y_1 \oplus\ y_2 & \text{if } x = 0\ , \\ y_1 \wedge\ y_2 & \text{if } x = 1\ . \end{cases}$$

Randomized response would publish privatized versions of x, y_1, and y_2 according to (4.8). In an interactive scheme, looking at $\tilde{x}$, the second party publishes (the privatized version of) either $y_1 \oplus y_2$ (if $\tilde{x} = 0$) or $y_1 \wedge y_2$ (if $\tilde{x} = 1$). Upon receiving the privatized data, the central observer makes optimal decisions in each case. Figure 4.2 illustrates how these two protocols compare in terms of average accuracy, where the accuracy is one if the approximation is correct and zero if the approximation is incorrect. For $\varepsilon = 0$, both protocols cannot do better than the best random guess of zero, which achieves average accuracy of $5/8 = 0.625$. For large ε, both protocols achieve the best accuracy of one.

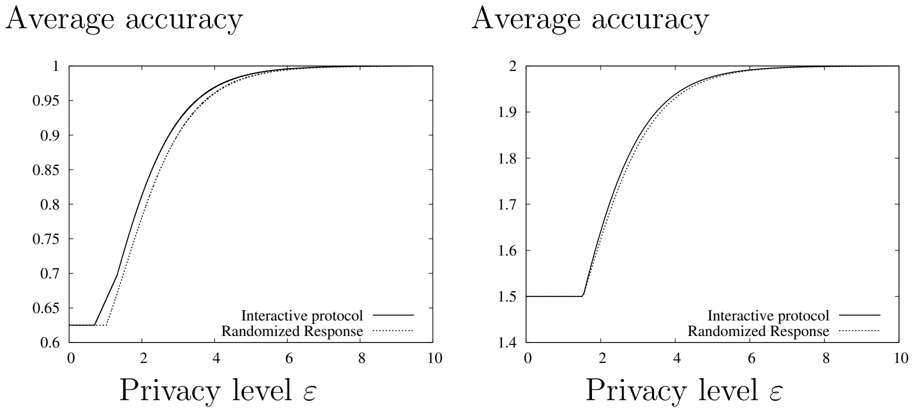


Figure 4.2: Interactive protocols can improve over the randomized response, when each party owns multiple bits, for computing XOR or AND (left) and computing the Hamming distance (right).

Another example of multiple bit multi-party computation is studied in [55]. There are two parties each owning two bits of data $x \in \{0, 1\}^2$ and $y \in \{0, 1\}^2$, and a third party wants to compute the Hamming distance $d_H(x, y) = \sum_{i=1}^{2} |x_i - y_i|$. Assuming each bit needs to be protected, the randomized response would reveal each bit via Equation 4.8. On the other hand, we can design an interactive scheme where one party reveals its two bits via the randomized response, and the other party then outputs its best estimate of the Hamming distance obeying differential privacy guarantees.

Figure 4.2 illustrates how these two protocols compare in terms of average accuracy, where the accuracy is $2 - |d_H(x, y) - \hat{d}|$ where $\hat{d}$ is the optimal decision made by the third party; the Hamming distance d_H is one if the approximation is correct and zero if the approximation is incorrect. For $\varepsilon = 0$, both protocols cannot do better than the best random guess of zero. which achieves average accuracy of $5/8 = 0.625$. For large ε, both protocols achieve the best accuracy of one.

4.7 Conclusions and Summary

In this chapter, we have studied the problem of differentially private multiparty computation. We showed that a simple non-interactive randomized response is optimal for all privacy levels (all values of ε and δ), heterogeneous privacy levels across parties, all types of functions to be computed, all types of cost metrics, and both average and worst-case (over the inputs) measures of accuracy. Though our results are general, they only handle settings where each party possesses a single bit. In the more general scenario where parties can have multiple bits, interaction might be critical to achieving the optimal privacy-utility tradeoffs.

CHAPTER 5

CONCLUSION

In this thesis, we have addressed the fundamental limits of differential privacy in three canonical privacy contexts (global, local, and multi-party). This chapter provides a quick recap of the main results presented in this thesis, and includes a discussion of important future work.

5.1 Global Privacy

In the global privacy context, trusted institutions want to release sensitive information about individuals. Differential privacy provides a formal guarantee on the anonymity level of an individual user with respect to a data release. However, such guarantees come at the expense of utility. The more the privacy demanded, the lesser the utility of the released data. In Chapter 2, we studied the fundamental tradeoff between global differential privacy and utility. Precisely, we showed that the correlated multi-dimensional staircase mechanism achieves the optimal privacy-utility tradeoff under ℓ_1 losses and two-dimensional query functions. We believe that the muti-dimensional staircase mechanism is universally optimal: it achieves the best privacy-utility tradeoff for higher dimensional queries and more general loss function. Even though this conjecture is backed by numerical evidence, it has yet to be proven rigorously.

5.2 Local Privacy

In the local privacy context, data providers trust no one, not even the service providers collecting their data. In this context, privacy is achieved by randomizing the data before releasing it. This leads to a fundamental tradeoff between privacy and utility. In Chapter 3, we studied the aforementioned privacy-utility tradeoff and showed that staircase mechanisms are optimal for a broad class of information theoretic utility functions such as mutual information and f-divergences. We also considered private binary hypothesis testing and information preservation, two canonical problems with a wide range of applications. Despite the generality of our local privacy framework, it can be extended in several important and non-trivial ways.

Correlation among data

In some scenarios the X_i's could be correlated (e.g., when different individuals observe different functions of the same random variable). In this case, the data analyst is interested in inferring whether the data was generated from P_0^n or P_1^n, where P_ν^n is one of two possible joint priors on $X_1, ..., X_n$. This is a challenging problem because knowing X_i reveals information about X_j, $j \neq i$. Therefore, the utility maximization problems for different individuals are coupled in this setting.

Robust and m-ary hypothesis testing

In some cases the data analyst need not have access to P_0 and P_1, but rather two classes of prior distribution P_{θ_0} and P_{θ_1} for $\theta_0 \in \Lambda_0$ and $\theta_1 \in \Lambda_1$. Such problems are studied under the rubric of universal hypothesis testing and robust hypothesis testing. One possible direction is to select the privatization mechanism that maximizes the worst case utility: $Q^* = \arg\max_{Q \in \mathcal{D}_\varepsilon} \min_{\theta_0 \in \Lambda_0, \theta_1 \in \Lambda_1} D_f(M_{\theta_0} || M_{\theta_1})$, where M_{θ_ν} is the induced marginal under P_{θ_ν}.

The more general problem of private m-ary hypothesis testing is also an interesting but challenging one. In this setting, the X_i's can follow one of m distributions P_0, P_1, ..., P_{m-1}. Consequently, the Y_i's can follow one of m distributions M_0, M_1, ..., M_{m-1}. The utility can be defined as the average f-divergence between any two distributions: $1/(m(m-1)) \sum_{i \neq j} D_f(M_i || M_j)$, or the worst case one: $\min_{i \neq j} D_f(M_i || M_j)$.

Non-exchangeable utility functions

The utility studied in this chapter was measured by functions that are exchangeable, i.e. the utility did not depend on the naming (labelling) of the private and privatized data (X and Y). This made sense for statistical learning applications that depend on information theoretic quantities such as f-divergences and mutual information. However, in some other applications, the utility might be defined over $\mathcal{X} \cup \mathcal{Y}$ in a metric space, where there exists a natural measure of distance (or distortion) between the data points. In this case, we can formulate the problem as a distortion minimization one

$$\text{minimize}_{Q \in \mathcal{D}_\varepsilon} \qquad \sum_{x,y} d(x,y) P(x) Q(y|x) \ ,$$

where $d(x,y)$ is some distortion metric. [28] studied this problem, and showed that the mechanism $Q(y|x) \propto e^{\varepsilon(1-d(x,y))}/(k-1+e^{\varepsilon})$ achieves near optimal performance when ε is large enough, which is the low privacy regime. Notice that when Hamming distance is used, $d(x,y) = \mathbb{I}(x \neq y)$, this recovers the randomized response mechanism exactly. This provides a starting point for generalizing the search for optimal mechanisms under non-exchangeable utility functions.

5.3 Multi-Party Privacy

In the multi-party context, different parties interact to compute a joint function on their private data. In this context, differential privacy allows users to interactively compute their functions while preventing them from learning each other's information. In Chapter 4, we studied the privacy-utility tradeoff in context where each individual possesses a single bit. Precisely, we showed the optimality of a simple non-interactive protocol: each party randomizes its bit (sufficiently) and shares the privatized version with the other parties. This optimality result is very general: it holds for all types of functions, heterogeneous privacy conditions on the parties, all types of cost metrics, and both average and worst-case (over the inputs) measures of accuracy.

Generalization to multiple bits

When each party owns multiple bits, it is possible that interactive protocols improve over the randomized response protocol. This issue was briefly discussed with examples in Section 4.6. As argued in Section 4.6, interaction will be useful in settings where parties have more than just one bit. We believe that simple non-interactive mechanisms are not optimal in this more general setting. However, this results is yet to be proven.

Correlated sources

When the data x_i's are correlated (e.g. each party observe a noisy version of the state of the world), knowing x_i reveals some information about other parties' bits. In general, revealing correlated data requires careful coordination between multiple parties. The analysis techniques developed in this thesis do not generalize to correlated data, since the crucial rank-one tensor structure of $S_\tau^{(y)}$ is no longer present.

Extensions to general utility functions

A surprising aspect of the main result is that even though the worst-case accuracy is a concave function over the protocol P, the maximum is achieved at an extremal point of the manifold of rank-1 tensors. This suggests that there is a deeper geometric structure of the problem, leading to possible universal optimality of the randomized response for a broader class of utility functions. It is an interesting task to understand the geometric structure of the problem, and to ask what class of utility functions lead to optimality of the randomized response.

APPENDIX A

PROOFS FOR GLOBAL DIFFERENTIAL PRIVACY

A.1 Operational Interpretation of Differential Privacy

A.1.1 Proof of Theorem 2.2.1

First we prove that (ε, δ)-differential privacy implies (2.1). From the definition of differential privacy, we know that for all rejection set $S \subseteq \mathcal{X}$, $\mathbb{P}(M(D_0) \in \bar{S}) \leq e^{\varepsilon}\mathbb{P}(M(D_1) \in \bar{S}) + \delta$. This implies $1 - P_{\text{FA}}(D_0, D_1, M, S) \leq e^{\varepsilon} P_{\text{MD}}(D_0, D_1, M, S) + \delta$. This implies the first inequality of (2.1), and the second one follows similarly.

The converse follows analogously. For any set S, we assume

$$1 - P_{\text{FA}}(D_0, D_1, M, S) \leq e^{\varepsilon} P_{\text{MD}}(D_0, D_1, M, S) + \delta.$$

Then, it follows that $\mathbb{P}(M(D_0) \in \bar{S}) \leq e^{\varepsilon}\mathbb{P}(M(D_1) \in \bar{S}) + \delta$ for all choices of $S \subseteq \mathcal{X}$. Together with the symmetric condition $\mathbb{P}(M(D_1) \in \bar{S}) \leq e^{\varepsilon}\mathbb{P}(M(D_0) \in \bar{S}) + \delta$, this implies (ε, δ)-differential privacy.

A.1.2 Proof of Remark 1

We have a decision rule γ represented by a partition $\{S_i\}_{i \in \{1,\ldots,N\}}$ and corresponding accept probabilities $\{p_i\}_{i \in \{1,\ldots,N\}}$, such that if the output is in a set S_i, we accept with probability p_i. We assume the subsets are sorted such that $1 \geq p_1 \geq \ldots \geq p_N \geq 0$. Then, the probability of false alarm is

$$\begin{aligned} P_{\rm FA}(D_0, D_1, M, \gamma) &= \sum_{i=1}^{N} p_i \, \mathbb{P}(M(D_0) \in S_i) \\ &= p_N + \sum_{i=2}^{N} (p_{i-1} - p_i) \, \mathbb{P}(M(D_0) \in \cup_{j<i} S_j) \; . \end{aligned}$$

and similarly, $P_{\rm MD}(D_0, D_1, M, \gamma) = (1 - p_1) + \sum_{i=2}^{N} (p_{i-1} - p_i) \, \mathbb{P}(M(D_1) \notin \cup_{j<i} S_j)$. Recall that

$$\begin{aligned} P_{\rm FA}(D_0, D_1, M, S) &= \mathbb{P}(M(D_0) \in S), \\ P_{\rm MD}(D_0, D_1, M, S) &= \mathbb{P}(M(D_1) \in \bar{S}). \end{aligned}$$

So for any decision rule γ, we can represent the pair $(P_{\rm MD}, P_{\rm FA})$ as a convex combination:

$$\begin{aligned} &\big(P_{\rm MD}(D_0, D_1, M, \gamma), P_{\rm FA}(D_0, D_1, M, \gamma) \big) = \\ &\sum_{i=1}^{N+1} (p_{i-1} - p_i) \big(P_{\rm MD}(D_0, D_1, M, \cup_{j<i} S_j), P_{\rm FA}(D_0, D_1, M, \cup_{j<i} S_j) \big) \; , \end{aligned}$$

where we used $p_0 = 1$ and $p_{N+1} = 0$, and hence it is included in the convex hull of the privacy region achieved by decision rules with hard thresholding.

A.1.3 Examples illustrating the strengths of the operational interpretation of differential privacy

Remark 2 *The following statements are true:*

(a) If a mechanism is (ε, δ)-differentially private, then it is $(\tilde{\varepsilon}, \tilde{\delta})$-differentially private for all pairs of $\tilde{\varepsilon}$ and $\tilde{\delta} \geq \delta$ satisfying

$$\frac{1 - \delta}{1 + e^{\varepsilon}} \;\; \geq \;\; \frac{1 - \tilde{\delta}}{1 + e^{\tilde{\varepsilon}}} \; .$$

(b) For a pair of neighboring databases D and D', and all (ε, δ)-differentially private mechanisms, the total variation distance defined as $\|M(D) -$

$M(D')\|_{\mathrm{TV}} = \max_{S\subseteq\mathcal{X}} \mathbb{P}(M(D')\in S) - \mathbb{P}(M(D)\in S)$ is bounded by

$$\sup_{(\varepsilon,\delta)\text{-differentially private } M} \|M(D)-M(D')\|_{\mathrm{TV}} \;\leq\; 1-\frac{2(1-\delta)}{1+e^{\varepsilon}}\ .$$

Proof 1 Proof of (a)**.** *From Figure 2.1, it follows immediately that $\mathcal{R}(\varepsilon,\delta)\subseteq\mathcal{R}(\tilde{\varepsilon},\tilde{\delta})$ when the conditions are satisfied. Then, for a (ε,δ)-private M, it follows from $\mathcal{R}(M)\subseteq\mathcal{R}(\varepsilon,\delta)\subseteq\mathcal{R}(\tilde{\varepsilon},\tilde{\delta})$ that M is $(\tilde{\varepsilon},\tilde{\delta})$-differentially private.*

Proof of (b)**.** *By definition, $\|M(D)-M(D')\|_{\mathrm{TV}} = \max_{S\subseteq\mathcal{X}}\mathbb{P}(M(D')\in S)-\mathbb{P}(M(D)\in S)$. Letting S be the rejection region in our hypothesis testing setting, the total variation distance is defined by the following optimization problem:*

$$\begin{aligned}\max_{S}\quad & 1-P_{\mathrm{MD}}(S)-P_{\mathrm{FA}}(S) \\ \textit{subject to}\quad & (P_{\mathrm{MD}}(S),P_{\mathrm{FA}}(S))\in\mathcal{R}(\varepsilon,\delta),\ \textit{for all } S\subseteq\mathcal{X}\ .\end{aligned}\tag{A.1}$$

From Figure 2.1 it follows immediately that the total variation distance cannot be larger than $\delta+(1-\delta)(e^{\varepsilon}-1)/(e^{\varepsilon}+1)$.

A.1.4 Proof of Theorem 2.2.3

Consider hypothesis testing between D_0 and D_1. If there is a point $(P_{\mathrm{MD}},P_{\mathrm{FA}})$ achieved by M' but not by M, then we claim that this is a contradiction to the assumption that D–X–Y forms a Markov chain. Consider a decision maker who only has access to the output of M. Under the Markov chain assumption, we can simulate the output of M' by generating a random variable Y conditioned on $M(D)$ and achieve every point in the privacy region of M' (see Remark 1 that follows Theorem 2.2.1 in Chapter 2). Hence, the privacy region of M' must be included in the privacy region of M.

A.2 Optimal Mechanisms for Differential Privacy

In this section, we provide a detailed proof for Theorem 2.3.1.

A.2.1 Proof outline

The key idea of the proof is to use a sequence of probability distributions with piecewise constant probability density functions to approximate any probability distribution satisfying the differential privacy constraint (2.7). The proof consists of 4 steps in total, and in each step we narrow down the set of probability distributions in which the optimal probability distribution should lie:

- Step 1 proves that we only need to consider probability distributions which have symmetric piecewise constant probability density functions.
- Step 2 proves that we only need to consider those symmetric piecewise constant probability density functions which are monotonically decreasing.
- Step 3 proves that optimal probability density function should periodically decay.
- Step 4 proves that the optimal probability density function is staircase-shaped in the multidimensional setting, and it concludes the proof of Theorem 2.3.1.

A.2.2 Step 1

Given $\mathcal{P} \in \mathcal{SP}$, define

$$V(\mathcal{P}) \triangleq \int \int \dots \int_{\mathbb{R}^d} \mathcal{L}(\mathbf{x}) \mathcal{P}(dx_1 dx_2 \dots dx_d).$$

Define

$$V^* \triangleq \inf_{\mathcal{P} \in \mathcal{SP}} V(\mathcal{P}). \tag{A.2}$$

Our goal is to prove that $V^* = \inf\limits_{\gamma \in [0,1]} \int \int \dots \int_{\mathbb{R}^d} \mathcal{L}(\mathbf{x}) f_\gamma(\mathbf{x}) dx_1 dx_2 \dots dx_d$.

If $V^* = +\infty$, then due to the definition of V^*, we have

$$\inf_{\gamma \in [0,1]} \int \int \dots \int_{\mathbb{R}^d} \mathcal{L}(\mathbf{x}) f_\gamma(\mathbf{x}) dx_1 dx_2 \dots dx_d \geq V^* = +\infty,$$

and thus $\inf_{\gamma \in [0,1]} \int \int \ldots \int_{\mathbb{R}^d} \mathcal{L}(\mathbf{x}) f_\gamma(\mathbf{x}) dx_1 dx_2 \ldots dx_d = V^* = +\infty$. So we only need to consider the case $V^* < +\infty$, i.e., V^* is finite. Therefore, in the rest of the proof, we assume V^* is finite.

First we show that given any probability measure $\mathcal{P} \in \mathcal{SP}$, we can use a sequence of probability measures with multidimensionally piecewise constant probability density functions to approximate $\mathcal{P}$.

Given $i \in \mathbb{N}$ and $k \in \mathbb{N}$, define

$$A_i(k) = \{\mathbf{x} \in \mathbb{R}^d | k\frac{\Delta}{i} \leq \|\mathbf{x}\|_1 < (k+1)\frac{\Delta}{i}\} \subset \mathbb{R}^d.$$

It is easy to calculate the volume of $A_i(k)$, which is

$$\mathrm{Vol}(A_i(k)) = \frac{2^d}{d!}\left((k+1)^d - k^d\right)\frac{\Delta^d}{i^d}.$$

.

Lemma A.2.1 *Given $\mathcal{P} \in \mathcal{SP}$ with $V(\mathcal{P}) < +\infty$, any positive integer $i \in \mathbb{N}$, define $\mathcal{P}_i$ as the probability distribution with probability density function $f_i(\mathbf{x})$ defined as*

$$f_i(\mathbf{x}) = a_i(k) \triangleq \frac{\mathcal{P}(A_i(k))}{Vol(A_i(k))} \mathbf{x} \in A_i(k) \text{ for } k \in \mathbb{N}. \tag{A.3}$$

Then $\mathcal{P}_i \in \mathcal{SP}$ and

$$\lim_{i \to +\infty} V(\mathcal{P}_i) = V(\mathcal{P}).$$

We conjecture that Lemma A.2.1 holds for arbitrary dimension d, and prove it for the case $d = 2$.

Before proving Lemma A.2.1 for $d = 2$, we prove an auxiliary Lemma which shows that for probability mass function over $\mathbb{Z}^2$ satisfying ϵ-differential privacy constraint, we can construct a new probability mass function by averaging the old probability mass function over each ℓ^1 ball and the new probability mass function still satisfies the ϵ-differential privacy constraint.

Lemma A.2.2 *For any given probability mass function $\mathcal{P}$ defined over the set $\mathbb{Z}^2$ satisfying that*

$$\mathcal{P}(i_1, j_1) \leq e^\epsilon \mathcal{P}(i_2, j_2), \forall |i_1 - i_2| + |j_1 - j_2| \leq \Delta, \tag{A.4}$$

define the probability mass function $\tilde{\mathcal{P}}$ *via*

$$\tilde{\mathcal{P}}(i,j) = \begin{cases} \mathcal{P}(0,0) & (i,j) = (0,0) \\ p_{|i|+|j|} & (i,j) \neq (0,0) \end{cases}$$

where $p_k \triangleq \frac{\sum_{(i',j') \in \mathbb{Z}^2 : |i'|+|j'|=k} \mathcal{P}(i',j')}{4k}, \forall k \geq 1.$

Then $\tilde{\mathcal{P}}$ *is also a probability mass function satisfying the differential privacy constraint, i.e.,*

$$\tilde{\mathcal{P}}(i_1, j_1) \leq e^{\epsilon} \tilde{\mathcal{P}}(i_2, j_2), \forall |i_1 - i_2| + |j_1 - j_2| \leq \Delta. \tag{A.5}$$

Proof 2 *Due to the way how we define* $\tilde{\mathcal{P}}$*, we have*

$$\sum_{(i,j) \in \mathbb{Z}^2} \tilde{\mathcal{P}}(i,j) = \sum_{(i,j) \in \mathbb{Z}^2} \mathcal{P}(i,j) = 1,$$

and thus $\tilde{\mathcal{P}}$ *is a valid probability mass function defined over* $\mathbb{Z}^2$.

Next we prove that $\tilde{\mathcal{P}}$ *satisfies* (A.5). *To simplify notation, define* $p_0 \triangleq \mathcal{P}(0,0)$. *Then we only need to prove that for any* $k_1, k_2 \in \mathbb{N}$ *such that* $|k_1 - k_2| \leq \Delta$*, we have*

$$p_{k_1} \leq e^{\epsilon} p_{k_2}.$$

Due to the symmetry property, without loss of generality, we can assume $k_1 < k_2$.

The easiest case is $k_1 = 0$. *When* $k_1 = 0$*, we have* $k_2 \leq \Delta$ *and*

$$\mathcal{P}(0,0) \leq e^{\epsilon} \mathcal{P}(i,j), \forall |i| + |j| = k_2. \tag{A.6}$$

The number of distinct pairs (i,j) *satisfying* $|i| + |j| = k$ *is* $4k$ *for* $k \geq 1$. *Sum up all inequalities in* (A.6), *and we get*

$$\begin{aligned} &4k_2 \mathcal{P}(0,0) \leq e^{\epsilon} \sum_{(i,j) \in \mathbb{Z}^2 : |i|+|j|=k_2} \mathcal{P}(i,j) \\ \Leftrightarrow &\mathcal{P}(0,0) \leq e^{\epsilon} \frac{\sum_{(i,j) \in \mathbb{Z}^2 : |i|+|j|=k_2} \mathcal{P}(i,j)}{4k_2} \\ \Leftrightarrow &p_0 \leq e^{\epsilon} p_{k_2}. \end{aligned}$$

For general $0 < k_1 < k_2$, *let* $D' \triangleq k_2 - k_1 \le \Delta$. *Define* B_k *via*

$$B_k \triangleq \{(i,j) \in \mathbb{Z}^2 || i| + |j| = k\}, \forall k \in \mathbb{N}.$$

Then the differential privacy constraint (A.4) *implies that*

$$\mathcal{P}(i_1,j_1) \le e^{\epsilon}\mathcal{P}(i_2,j_2), \forall (i_1,j_1) \in B_{k_1}, (i_2,j_2) \in B_{k_2}, |i_1 - i_2| + |j_1 - j_2| = D'. \tag{A.7}$$

The set of points in B_k *forms a rectangle, which has* 4 *corner points and* $4(k-1)$ *interior points on the edges. For each corner point in* B_{k_1}, *which appears in the left side of* (A.7), *there are* $(2D'+1)$ *points in* B_{k_2} *close to it with an* ℓ^1 *distance of* D'. *And for each interior point in* B_{k_1}, *there are* $(D'+1)$ *points in* B_{k_2} *close to it with an* ℓ^1 *distance of* D'. *Therefore, there are in total* $4(2D'+1) + 4(k_1 - 1)(D'+1)$ *distinct inequalities in* (A.7).

If we can find certain nonnegative coefficients such that multiplying each inequality in (A.7) *by these nonnegative coefficients and summing them up gives us*

$$\frac{\sum_{(i',j') \in \mathbb{Z}^2 : |i'|+|j'|=k_1} \mathcal{P}(i',j')}{4k_1} \le e^{\epsilon} \frac{\sum_{(i',j') \in \mathbb{Z}^2 : |i'|+|j'|=k_2} \mathcal{P}(i',j')}{4k_2},$$

then (A.5) *holds. Therefore, our goal is to find the "right" coefficients associated with each inequality in* (A.7). *We formulate it as a matrix filling-in problem in which we need to choose nonnegative coefficients for certain entries in a matrix such that the sum of each row is* $\frac{k_1+D'}{k_1}$, *and the sum of each column is 1.*

More precisely, label the $4k_1$ *points in* B_{k_1} *by* $\{I_1, I_2, I_3, \ldots, I_{4k_1}\}$, *where we label the topmost point by* 1 *and sequentially label other points clockwise. Similarly, we label the* $4k_2$ *points in* B_{k_2} *by* $\{O_1, O_2, O_3, \ldots, O_{4k_2}\}$, *where we label the topmost point by* 1 *and sequentially label other points clockwise.*

Consider the following $4k_1$ *by* $4k_2$ *matrix* M, *where each row corresponds to the point in* B_{k_1} *and each column corresponds to the point in* B_{k_2}, *and the entry* M_{ij} *in the* i*th row and* j*th column is the coefficient corresponds to inequality involved with the points* I_i *and* O_j. *If there is no inequality associated with the points* I_i *and* O_j, *then* $M_{ij} = 0$.

In the case $k_1 = 2$ *and* $D' = 3$, *the zeros/nonzeros pattern of* M *has the* *following form:*

$$\begin{pmatrix}
x & x & x & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & x & x \\
0 & x & x & x & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & x & x & x & x & x & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & x & x & x & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & x & x & x & x & x & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & x & x & x & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & x & x & x & x & x & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & x & x & x
\end{pmatrix},$$

where x denotes an entry which can take any nonnegative coefficient.

For general k_1 and k_2, the pattern of M is that the first, $(k_1+1)th$, $(2k_1+1)th$ and $(3k_1+1)th$ rows can have $2D'+1$ nonzero entries, and all other rows can have $D'+1$ nonzero entries.

We want to show that

$$\frac{\sum_{(i',j')\in\mathbb{Z}^2:|i'|+|j'|=k_1}\mathcal{P}(i',j')}{4k_1} \leq e^{\epsilon}\frac{\sum_{(i',j')\in\mathbb{Z}^2:|i'|+|j'|=k_2}\mathcal{P}(i',j')}{4k_2},$$

or equivalently,

$$(1+\frac{D'}{k_1})\sum_{(i',j')\in\mathbb{Z}^2:|i'|+|j'|=k_1}\mathcal{P}(i',j') \leq e^{\epsilon}\sum_{(i',j')\in\mathbb{Z}^2:|i'|+|j'|=k_2}\mathcal{P}(i',j').$$

Therefore, our goal is to find nonnegative coefficients to substitute each x in the matrix such that the sum of each column is 1 and the sum of each column is $(1+\frac{D'}{k_1})$. We will give explicit formulas on how to choose the coefficients.

The case $k_1 = 1$ is trivial. Indeed, one can set all diagonal entries to be 1, and set all other nonzero entries to be $\frac{1}{2}$. Therefore, we can assume $k_1 > 1$.

Consider two different cases: $k_1 \leq D'$ and $k_1 \geq D'+1$.

We first consider the case $k_1 \leq D'$. Due to the periodic patterns in M, we only need to consider rows from 1 to k_1+1. Set all entries to be zero except that we set

$$\begin{aligned}
&M_{11} = M_{22} = \cdots = M_{k_1k_1} = 1, \\
&M_{2,D'+2} = M_{3,D'+3} = \cdots = M_{k_1+1,k_1+D'+1} = 1 \\
&M_{1,j} = \frac{D'}{2k_1(D'-k_1+1)}, j \in [k_1+1, D'+1] \cup [4k_1 - D' + 1, 4k_1] \\
&M_{k_1+1,j} = \frac{D'}{2k_1(D'-k_1+1)}, j \in [k_1+1, D'+1] \cup [2k_1+1+D', k_1+1+2D'] \\
&M_{i,j} = \frac{1 - \frac{D'}{k_1(D'-k_1+1)}}{k_1 - 1}. \qquad \text{(A.8)}
\end{aligned}$$

It is straightforward to verify that the above matrix M satisfies the properties that the sum of each column is 1 and the sum of each row is $(1 + \frac{D'}{k_1})$. Therefore, we have

$$p_{k_1} \leq e^{\epsilon} p_{k_2}, \forall 0 < k_1 < k_2, k_1 \leq k_2 - k_1 \leq \Delta.$$

Next we solve the case $k_1 \geq D' + 1$. Again due to the periodic patterns in M, we only need to consider the nonzero entries in rows from 1 to $k_1 + 1$. We use the following procedures to construct M:

1. For the first row, set $M_{11} = 1$ and set all other $2D'$ nonzero entries to be $\frac{1}{2k_1}$.

2. For the second row, M_{22} is uniquely determined to be $1 - \frac{1}{2k_1}$. Set the next $D' - 1$ nonzero entries in the second row to be $\frac{1}{k_1}$, i.e., $M_{2j} = \frac{1}{k_1}$ for $j \in [3, D'+1]$. The last nonzero entry $M_{2,D'+2}$ is uniquely determined to be

$$(1 + \frac{D'}{k_1}) - (1 - \frac{1}{2k_1}) - \frac{D'-1}{k_1} = \frac{3}{2k_1}.$$

3. For the third row, the first nonzero entry M_{33} is uniquely determined to be $1 - \frac{1}{2k_1} - \frac{1}{k_1} = 1 - \frac{3}{2k_1}$. Set the next $D' - 1$ nonzero entries to be $\frac{1}{k_1}$, i.e., $M_{3j} = \frac{1}{k_1}$ for $j \in [4, D'+2]$. The last nonzero entry $M_{3,D'+3}$ is uniquely determined to be

$$(1 + \frac{D'}{k_1}) - (1 - \frac{3}{2k_1}) - \frac{D'-1}{k_1} = \frac{5}{2k_1}.$$

4. *In general, for the ith row ($i \in [2, k_1 - 1]$), the first nonzero entry M_{ii} is set to be $M_{ii} = 1 - \frac{2i-3}{2k_1}$, and the next $D' - 1$ nonzero entries are $\frac{1}{k_1}$, and the last nonzero entry $M_{i,i+D'} = \frac{2i-1}{2k_1}$.*

5. *For $(k_1 + 1)$th row, by symmetry, we set $M_{k_1+1,k_1+1} = 1$ and set other $2D'$ nonzero entries to be $\frac{1}{2k_1}$.*

6. *The nonzero entries in the k_1th row are uniquely determined. Indeed, we have*

$$\begin{aligned} M_{k_1,k_1} &= 1 - \frac{2k_1 - 3}{2k_1}, \\ M_{k_1,k_1+D'} &= 1 - \frac{1}{2k_1}, \\ M_{k_1,k_1+j} &= \frac{1}{k_1}, j \in [2, D' - 1]. \end{aligned}$$

It is straightforward to verify that each entry in M is nonnegative and M satisfies the properties that the sum of each column is 1 and the sum of each row is $(1 + \frac{D'}{k_1})$. Therefore, we have

$$p_{k_1} \le e^{\epsilon} p_{k_2}, \forall 0 < k_1 < k_2, k_1 \ge D' + 1 = k_2 - k_1 + 1.$$

Therefore, for all $k_1, k_2 \in \mathbb{N}$ such that $|k_2 - k_1| \le \Delta$, we have

$$p_{k_1} \le e^{\epsilon} p_{k_2}.$$

This completes the proof of Lemma A.2.2.

Proof 3 (Proof of Lemma A.2.1) *First we prove that $\mathcal{P}_i \in \mathcal{SP}$, i.e., $\mathcal{P}_i$ satisfies the differential privacy constraint (2.7).*

By the definition of $f_i(\mathbf{x})$, $f_i(\mathbf{x})$ is a nonnegative function, and

$$\begin{aligned} &\int \int \dots \int_{\mathbb{R}^d} f_i(\mathbf{x}) dx_1 dx_2 \dots dx_d \\ =& \sum_{k=0}^{+\infty} a_i(k) \, Vol(A_i(k)) \\ =& \sum_{k=0}^{+\infty} \mathcal{P}(A_i(k)) \\ =& \mathcal{P}(\mathbb{R}^d) = 1. \end{aligned}$$

So $\mathcal{P}_i$ is a valid probability distribution.

Next we show that $f_i(\mathbf{x})$ satisfies the differential privacy constraint. For fixed i, on the $x_1 - x_2$ plane, we can use the lines $x_2 = x_1 + \frac{k}{i}\Delta$ and $x_2 = -x_1 + \frac{k}{i}\Delta$ for all $k \in \mathbb{Z}$ to divide each $A_i(k)$ into distinct squares with the same size (each $A_i(k)$ will be divided into $8k+4$ squares). By taking the average of the probability density function over each square, we reduce the probability density function to a discrete probability mass function over $\mathbb{Z}^2$ satisfying ϵ-differential privacy constraint. Then apply Lemma A.2.2, and we have

$$a_i(k_1) \le e^{\epsilon} a_i(k_2), \forall k_1, k_2 \in \mathbb{N} \text{ with } |k_1 - k_2| \le i.$$

Given $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ such that $\|\mathbf{x} - \mathbf{y}\|_1 \le \Delta$, let k_1, k_2 be the integers such that

$$\mathbf{x} \in A_i(k_1),$$
$$\mathbf{y} \in A_i(k_2).$$

Then $|k_1 - k_2| \le i$. Therefore,

$$f_i(\mathbf{x}) \le e^{\epsilon} f_i(\mathbf{y}),$$

which implies that the probability distribution $\mathcal{P}_i$ satisfies the differential privacy constraint (2.7).

Therefore, for any integer $i \ge 1$, $\mathcal{P}_i \in \mathcal{SP}$.

Next we show that

$$\lim_{i \to +\infty} V(\mathcal{P}_i) = V(\mathcal{P}).$$

Given $\delta > 0$, since $V(\mathcal{P})$ is finite, there exists $T^* = m\Delta > 1$ for some $m \in \mathbb{N}$ such that

$$\int\int \dots \int_{\{\mathbf{x} \in \mathbb{R}^d \mid \|\mathbf{x}\|_1 \ge T^*\}} \mathcal{L}(\mathbf{x}) \mathcal{P}(dx_1 dx_2 \dots dx_d) < \frac{\delta}{2}.$$

For each $A_i(k)$ we have

$$\begin{aligned}
\int \dots \int_{A_i(k)} \mathcal{L}(\mathbf{x})\mathcal{P}_i(dx_1dx_2\dots dx_d) &= \int \dots \int_{A_i(k)} \|\mathbf{x}\|_1\mathcal{P}_i(dx_1dx_2\dots dx_d) \\
&\leq \mathcal{P}_i(A_i(k))(k+1)\frac{\Delta}{i} \\
&= \mathcal{P}(A_i(k))(k+1)\frac{\Delta}{i} \\
&\leq 2\mathcal{P}(A_i(k))k\frac{\Delta}{i} \\
&\leq 2\int \dots \int_{A_i(k)} \mathcal{L}(\mathbf{x})\mathcal{P}(dx_1dx_2\dots dx_d).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\int\int \dots \int_{\{\mathbf{x}\in\mathbb{R}^d \mid \|\mathbf{x}\|_1 \geq T^*\}} \mathcal{L}(\mathbf{x})\mathcal{P}_i(dx_1dx_2\dots dx_d) \\
&\leq 2\int\int \dots \int_{\{\mathbf{x}\in\mathbb{R}^d \mid \|\mathbf{x}\|_1 \geq T^*\}} \mathcal{L}(\mathbf{x})\mathcal{P}(dx_1dx_2\dots dx_d) \\
&\leq 2\frac{\delta}{2} = \delta.
\end{aligned}$$

$\mathcal{L}(\mathbf{x})$ is a bounded function when $\|\mathbf{x}\|_1 \leq T^*$, and thus by the definition of Riemann-Stieltjes integral, we have

$$\begin{aligned}
&\lim_{i\to\infty}\int\int \dots \int_{\{\mathbf{x}\in\mathbb{R}^d \mid \|\mathbf{x}\|_1 < T^*\}} \mathcal{L}(\mathbf{x})\mathcal{P}_i(dx_1dx_2\dots dx_d) \\
&= \int\int \dots \int_{\{\mathbf{x}\in\mathbb{R}^d \mid \|\mathbf{x}\|_1 < T^*\}} \mathcal{L}(\mathbf{x})\mathcal{P}(dx_1dx_2\dots dx_d).
\end{aligned}$$

So there exists a sufficiently large integer i^* such that for all $i \geq i^*$

$$\begin{aligned}
\Bigg|&\int\int \dots \int_{\{\mathbf{x}\in\mathbb{R}^d \mid \|\mathbf{x}\|_1 < T^*\}} \mathcal{L}(\mathbf{x})\mathcal{P}_i(dx_1dx_2\dots dx_d) \\
&\quad -\int\int \dots \int_{\{\mathbf{x}\in\mathbb{R}^d \mid \|\mathbf{x}\|_1 < T^*\}} \mathcal{L}(\mathbf{x})\mathcal{P}(dx_1dx_2\dots dx_d)\Bigg| \leq \delta.
\end{aligned}$$

To simplify notation, we use $d\mathbf{x}$ to denote $dx_1dx_2\dots dx_d$.

Hence, for all $i \geq i^*$

$$\begin{aligned}
&|V(\mathcal{P}_i) - V(\mathcal{P})| \\
= &\left| \int_{\mathbb{R}^d} \mathcal{L}(\mathbf{x})\mathcal{P}_i(d\mathbf{x}) - \int_{\mathbb{R}^d} \mathcal{L}(\mathbf{x})\mathcal{P}(d\mathbf{x}) \right| \\
= &\left| \int_{\{\mathbf{x}\in\mathbb{R}^d | \|\mathbf{x}\|_1 < T^*\}} \mathcal{L}(\mathbf{x})\mathcal{P}_i(d\mathbf{x}) - \int_{\{\mathbf{x}\in\mathbb{R}^d | \|\mathbf{x}\|_1 < T^*\}} \mathcal{L}(\mathbf{x})\mathcal{P}(d\mathbf{x}) \right. \\
&\qquad \left. + \int_{\{\mathbf{x}\in\mathbb{R}^d | \|\mathbf{x}\|_1 \geq T^*\}} \mathcal{L}(\mathbf{x})\mathcal{P}_i(d\mathbf{x}) - \int_{\{\mathbf{x}\in\mathbb{R}^d | \|\mathbf{x}\|_1 \geq T^*\}} \mathcal{L}(\mathbf{x})\mathcal{P}(d\mathbf{x}) \right| \\
\leq &\left| \int_{\{\mathbf{x}\in\mathbb{R}^d | \|\mathbf{x}\|_1 < T^*\}} \mathcal{L}(\mathbf{x})\mathcal{P}_i(d\mathbf{x}) - \int_{\{\mathbf{x}\in\mathbb{R}^d | \|\mathbf{x}\|_1 < T^*\}} \mathcal{L}(\mathbf{x})\mathcal{P}(d\mathbf{x}) \right| \\
&\qquad + \int_{\{\mathbf{x}\in\mathbb{R}^d | \|\mathbf{x}\|_1 \geq T^*\}} \mathcal{L}(\mathbf{x})\mathcal{P}_i(d\mathbf{x}) + \int_{\{\mathbf{x}\in\mathbb{R}^d | \|\mathbf{x}\|_1 \geq T^*\}} \mathcal{L}(\mathbf{x})\mathcal{P}(d\mathbf{x}) \\
\leq &(\delta + \delta + \frac{\delta}{2}) \\
\leq &\frac{5}{2}\delta.
\end{aligned}$$

Therefore,

$$\lim_{i\to+\infty} V(\mathcal{P}_i) = V(\mathcal{P}).$$

Define $\mathcal{SP}_{i,\text{sym}} \triangleq \{\mathcal{P}_i | \mathcal{P} \in \mathcal{SP}\}$ for $i \geq 1$, i.e., $\mathcal{SP}_{i,\text{sym}}$ is the set of probability distributions satisfying differential privacy constraint (2.7) and having symmetric piecewise constant (over $A_i(k)$ $\forall k \in \mathbb{N}$) probability density functions.

Due to Lemma A.2.1, we have Lemma A.2.3.

Lemma A.2.3

$$V^* = \inf_{\mathcal{P}\in\cup_{i=1}^{\infty}\mathcal{SP}_{i,sym}} V(\mathcal{P}).$$

Therefore, to characterize V^*, we only need to study probability distributions with symmetric and piecewise constant probability density functions.

A.2.3 Step 2

Given $\mathcal{P} \in \mathcal{P}_{\text{sym}}$, we call $\{a_i(0), a_i(1), a_i(2), \dots\}$ the density sequence of $\mathcal{P}_i \in \mathcal{SP}_{i,\text{sym}}$, where $a_i(k)$ is defined in (A.3) $\forall k \in \mathbb{N}$.

Next we show that indeed we only need to consider those probability distributions with symmetric piecewise constant probability density functions the density sequences of which are *monotonically decreasing*.

Define

$$\mathcal{SP}_{i,\text{md}} \triangleq$$
$$\{\mathcal{P}|\mathcal{P} \in \mathcal{SP}_{i,\text{sym}}, \text{ and the density sequence of } \mathcal{P} \text{ is monotonically decreasing}\}.$$

Then we get Lemma A.2.4.

Lemma A.2.4

$$V^* = \inf_{\mathcal{P} \in \cup_{i=1}^{\infty} \mathcal{SP}_{i,md}} V(\mathcal{P}).$$

Proof 4 *We first show that among $\mathcal{SP}_{i,sym}$, to minimize the cost we only need to consider these probability distributions with density sequences $\{a_0, a_1, a_2, \dots\}$ satisfying that $a_0 \geq a_1$. Indeed, given $\mathcal{P}_a \in \mathcal{SP}_{i,sym}$ with density sequence $\{a_0, a_1, a_2, \dots\}$ such that $a_0 < a_1$, there exists $\mathcal{P}_b \in \mathcal{SP}_{i,sym}$ with density sequence $\{b_0, b_1, b_2, \dots\}$ such that $b_0 \geq b_1$ and*

$$V(\mathcal{P}_b) \leq V(\mathcal{P}_a).$$

Consider the probability distribution $\mathcal{P}_b \in \mathcal{SP}_{i,sym}$ with density sequence $\{b_0, b_1, b_2, , \dots\}$ defined as

$$\begin{aligned} b_0 &= (1+\delta)a_0, \\ b_k &= (1-\delta')a_k, \forall k \geq 1, \end{aligned}$$

where we choose $\delta > 0$ and $0 < \delta' < 1$ such that

$$b_0 = b_1, \tag{A.9}$$

$$\sum_{k=0}^{+\infty} b_k \, Vol(A_i(k)) = \sum_{k=0}^{+\infty} a_k \, Vol(A_i(k)) = 1. \tag{A.10}$$

Equation (A.10) *makes $\mathcal{P}_b$ be a valid probability distribution. One can easily solve* (A.9) *and* (A.10)*, and write down the explicit expression for δ, δ'. The density sequence $\{b_0, b_1, b_2, \dots\}$ satisfies $b_0 \geq b_1$ (indeed, we have*

$b_0 = b_1$), and it is easy to check it satisfies the differential privacy constraint, i.e.,

$$b_{k_1} \leq e^{\epsilon} b_{k_2}, \forall k_1, k_2 \in \mathbb{N} \text{ with } |k_1 - k_2| \leq i.$$

Note that $\mathcal{C}(\|\mathbf{x}\|_1)$ is a monotonically increasing function of $\|\mathbf{x}\|_1$, and compared to $\mathcal{P}_a$, $\mathcal{P}_b$ moves some probability of $\mathcal{SP}_{i,md}$ from the (higher cost) area $\{\mathbf{x} | \|bx\| \geq \frac{\Delta}{i}\}$ to the (lower cost) area $\{\mathbf{x} | \|bx\| \leq \frac{\Delta}{i}\}$, and thus we have

$$V(\mathcal{P}_b) \leq V(\mathcal{P}_a).$$

Therefore, among $\mathcal{SP}_{i,sym}$, to minimize the cost we only need to consider these probability distributions with density sequences $\{a_1, a_2, a_3, \dots\}$ satisfying that $a_0 \geq a_1$.

Next we show that among $\mathcal{SP}_{i,sym}$ with density sequences $\{a_1, a_2, a_3, \dots\}$ satisfying $a_0 \geq a_1$, to minimize the cost we only need to consider these probability distributions with density sequences also satisfying that $a_1 \geq a_2$.

Given $\mathcal{P}_a \in \mathcal{SP}_{i,sym}$ with density sequence $\{a_1, a_2, a_3, \dots\}$ such that $a_0 \geq a_1$ and $a_1 < a_2$, there exists $\mathcal{P}_b \in \mathcal{SP}_{i,sym}$ with density sequence $\{b_1, b_2, b_3, \dots\}$ such that $b_0 \geq b_1$ and

$$b_1 \geq b_2.$$

If $i \leq 2$, we can construct $\mathcal{P}_b$ by scaling up a_0, a_1 and scale down a_k for all $k \geq 2$. More precisely, define $\mathcal{P}_b$ with density sequence $\{b_0, b_1, b_2, \dots\}$ via

$$\begin{aligned} b_k &= (1+\delta) a_k, k \leq 1, \\ b_k &= (1-\delta') a_k, k \geq 2, \end{aligned}$$

for some $\delta > 0$ and $0 < \delta' < 1$ such that

$$\begin{aligned} b_2 &= b_1, \\ \sum_{k=0}^{+\infty} b_k \, Vol(A_i(k)) &= \sum_{k=0}^{+\infty} a_k \, Vol(A_i(k)) = 1. \end{aligned}$$

So we have $b_0 \geq b_1 \geq b_2$. It is easy to check that $\mathcal{P}_b$ satisfies the differential privacy constraint, and $V(\mathcal{P}_b) \leq V(\mathcal{P}_a)$ using the fact that $\mathcal{C}(\|\mathbf{x}\|_1)$ is a *monotonically decreasing function in terms of* $\|\mathbf{x}\|_1$.

If $i \geq 3$*, then without loss of generality we can assume* $a_2 \leq a_0$*. Indeed, if* $a_2 > a_0$*, we can scale up* a_0, a_1 *and scale down* a_k *for all* $k \geq 2$ *to make* $a_2 = a_0$*, and this operation will preserve the differential privacy constraint and decrease the cost. Note that in this case we cannot use the same scaling operation to make* $a_2 \leq a_0$*, because it is possible that after the scaling operation* $\frac{a_0}{a_k} > e^\epsilon$ *for some* $3 \leq k \geq i$ *violating the differential privacy constraint. Hence, we can assume* $a_0 \geq a_2 > a_1$*. Let* $a_{k'}$ *be the largest value in* $\{a_3, \ldots, a_{2+i}\}$*. If* $\frac{a_{k'}}{a_2} < e^\epsilon$*, we can scale up* a_1 *and scale down* a_2 *until* $a_1 = a_2$ *or* $\frac{a_{k'}}{a_2} = e^\epsilon$*. It is easy to see this scaling operation will preserve differential privacy and decrease the cost. If after this scaling operation we have* $a_2 = a_1$*, then we are done. Suppose* a_1 *is still bigger than* a_2*. Then* a_2 *is the smallest element in* $\{a_2, a_3, \ldots, a_{2+i}\}$*. Therefore, we have* $\max_{2 \leq k \leq i} \frac{a_0}{a_k} = \frac{a_0}{a_2}$*. Then we can scale up* a_0, a_1 *and scale down* a_k *for* $k \geq 2$ *until* $a_1 = a_2$*. This operation will preserve the differential privacy constraint and decrease the cost. If we call the final probability distribution we obtained* $\mathcal{P}_b$*, we have* $\mathcal{P}_b \in \mathcal{SP}_{i,sym}$*, and the density sequence satisfying* $b_0 \geq b_1 \geq b_2$ *(indeed,* $b_1 = b_2$*), and* $V(\mathcal{P}_b) \leq V(\mathcal{P}_a)$*.*

By induction, we can show that among all probability distributions in $\mathcal{SP}_{i,sym}$*, to minimize the cost we only need to consider probability distributions with monotonically decreasing density sequence.*

Suppose among $\mathcal{SP}_{i,sym}$ *to minimize the cost we only need to consider probability distribution with density sequence* $\{a_0, a_1, a_2, \ldots\}$ *satisfying* $a_0 \geq a_1 \geq a_2 \geq \cdots \geq a_n$*. Then we can show that among* $\mathcal{SP}_{i,sym}$ *to minimize the cost we only need to consider probability distribution with density sequence* $\{a_0, a_1, a_2, \ldots\}$ *satisfying* $a_0 \geq a_1 \geq a_2 \geq \cdots \geq a_n \geq a_{n+1}$*.*

Indeed, given $\mathcal{P}_a \in \mathcal{SP}_{i,sym}$ *with density sequence* $\{a_0, a_1, a_2, \ldots\}$ *satisfying* $a_0 \geq a_1 \geq a_2 \geq \cdots \geq a_n$*, we can construct* $\mathcal{P}_b \in \mathcal{SP}_{i,sym}$ *with density sequence* $\{b_0, b_1, b_2, \ldots\}$ *satisfying*

$$b_0 \geq b_1 \geq b_2 \geq \cdots \geq b_n \geq b_{n+1},$$

and

$$V(\mathcal{P}_b) \leq V(\mathcal{P}_a).$$

If $a_{n+1} \le a_n$, then we can choose $\mathcal{P}_b = \mathcal{P}_a$.

Suppose $a_{n+1} > a_n$. Without loss of generality, we can assume

$$a_{n+1} \le a_k, \text{ for } k \le n+2-i. \tag{A.11}$$

If $a_{n+1} > a_{n+2-i}$, then we can scale up $\{a_0, a_1, \dots, a_n\}$ and scale down $\{a_{n+1}, a_{n+2}, \dots\}$ until $a_{n+1} = a_k$. It is easy to verify that this scaling operation will preserve the differential privacy constraint and decrease the cost.

Let k^* be the smallest integer such that $a_{k^*} < a_{n+1}$. Note that by (A.11) *we have $n+3-i \le k^* \le n$. Let a_j be the biggest element in*

$$\{a_{n+2}, a_{n+3}, \dots, a_{n+1+i}\}.$$

Due to the differential privacy constraint, we have $\frac{a_j}{a_{n+1}} \le e^\epsilon$. Then we can scale up a_{k^*} and scale down a_{n+1} until $a_{k^*} = a_{n+1}$ or $\frac{a_j}{a_{n+1}} = e^\epsilon$. This operation will preserve the differential privacy constraint and decrease the cost. If after this scaling operation a_{k^*} is still bigger than a_{n+1}, then we can scale up $\{a_0, a_1, \dots, a_n\}$ and scale down $\{a_{n+1}, a_{n+2}, \dots\}$ until $a_{k^*} = a_{n+1}$. Due to the fact that a_{n+1} is the smallest element in $\{a_{n+1}, a_{n+2}, \dots, a_{n+1+i}\}$, this scaling operation will preserve the differential privacy constraint and decrease the cost. Therefore, we will have $a_{n+1} \le a_{k^*}$.

Repeat the above steps for each $k \in k^*+1, k^*+2, \dots, n$ such that $a_k < a_{n+1}$. If we call the final probability distribution we obtained $\mathcal{P}_b$, we have $\mathcal{P}_b \in \mathcal{SP}_{i,sym}$, and the density sequence satisfying

$$b_0 \ge b_1 \ge b_2 \ge \cdots \ge b_n,$$

and $V(\mathcal{P}_b) \le V(\mathcal{P}_a)$.

Hence, among $\mathcal{SP}_{i,sym}$ to minimize the cost we only need to consider probability distribution with density sequence $\{a_0, a_1, a_2, \dots\}$ satisfying $a_0 \ge a_1 \ge a_2 \ge \cdots \ge a_n \ge a_{n+1}$.

Therefore, among all probability distributions in $\mathcal{SP}_{i,sym}$, to minimize the cost we only need to consider probability distributions with monotonically decreasing density sequence.

We conclude that

$$V^* = \inf_{\mathcal{P} \in \cup_{i=1}^{\infty} \mathcal{SP}_{i,md}} V(\mathcal{P}).$$

This completes the proof of Lemma A.2.4.

A.2.4 Step 3

Next we show that among all symmetric piecewise constant probability density functions, we only need to consider those which are geometrically decaying.

More precisely, given positive integer i,

$$\mathcal{SP}_{i,\text{pd}} \triangleq \{\mathcal{P} | \mathcal{P} \in \mathcal{SP}_{i,\text{md}}, \text{ and } \mathcal{P} \text{ has density sequence } \{a_0, a_1, \ldots, a_n, \ldots,\} \text{ satisfying} \frac{a_k}{a_{k+i}} = e^{\epsilon}, \forall k \in \mathbb{N}\},$$

then we get Lemma A.2.5

Lemma A.2.5

$$V^* = \inf_{\mathcal{P} \in \cup_{i=1}^{\infty} \mathcal{SP}_{i,pd}} V(\mathcal{P}).$$

Proof 5 *Due to Lemma A.2.4, we only need to consider probability distributions with symmetric and piecewise constant probability density functions which are monotonically decreasing.*

We first show that given $\mathcal{P}_a \in \mathcal{SP}_{i,md}$ with density sequence $\{a_0, a_1, \ldots, a_n, \ldots,\}$, if $\frac{a_0}{a_i} < e^{\epsilon}$, then we can construct a probability distributions $\mathcal{P}_b \in \mathcal{SP}_{i,md}$ with density sequence $\{b_0, b_1, \ldots, b_n, \ldots,\}$ such that $\frac{b_0}{b_i} = e^{\epsilon}$ and

$$V(\mathcal{P}_b) \leq V(\mathcal{P}_a).$$

Define a new sequence $\{b_0, b_1, \ldots, b_n, \ldots\}$ by scaling up a_0 and scaling down

$\{a_1, a_2, \dots\}$. More precisely, define $\{b_0, b_1, \dots, b_n, \dots\}$ via

$$\begin{aligned} b_0 &= a_0(1+\delta), \\ b_k &= a_k(1-\delta'), \forall\, k \geq 1, \end{aligned}$$

for some $\delta > 0$ and $0 < \delta' < 1$ such that

$$\begin{aligned} \frac{b_0}{b_i} &= e^{\epsilon}, \\ \sum_{k=0}^{+\infty} b_k\, Vol(A_i(k)) &= \sum_{k=0}^{+\infty} a_k\, Vol(A_i(k)) = 1. \end{aligned}$$

So $\{b_0, b_1, \dots, b_n, \dots\}$ is a valid probability density sequence. Let $\mathcal{P}_b$ be the corresponding probability distribution. It is easy to check that $\mathcal{P}_b$ satisfies the differential privacy constraint, i.e.,

$$\frac{b_k}{b_{k+i}} \leq e^{\epsilon}, \forall k \geq 0.$$

Hence, $\mathcal{P}_b \in \mathcal{SP}_{i,md}$. Since $\mathcal{C}(\|bx\|_1)$ is a monotonically increasing function of $\|\mathbf{x}\|_1$, we have $V(\mathcal{P}_b) \leq V(\mathcal{P}_a)$.

Therefore, for given $i \in \mathbb{N}$, we only need to consider $\mathcal{P} \in \mathcal{SP}_{i,md}$ with density sequence $\{a_0, a_1, \dots, a_n, \dots\}$ satisfying $\frac{a_0}{a_i} = e^{\epsilon}$.

Next, we argue that among all probability distributions $\mathcal{P} \in \mathcal{SP}_{i,md}$ with density sequence $\{a_0, a_1, \dots, a_n, \dots,\}$ satisfying $\frac{a_0}{a_i} = e^{\epsilon}$, we only need to consider those probability distributions with density sequence also satisfying $\frac{a_1}{a_{i+1}} = e^{\epsilon}$.

Given $\mathcal{P}_a \in \mathcal{SP}_{i,md}$ with density sequence $\{a_0, a_1, \dots, a_n, \dots\}$ satisfying $\frac{a_0}{a_i} = e^{\epsilon}$ and $\frac{a_1}{a_{i+1}} < e^{\epsilon}$, we can construct a new probability distribution $\mathcal{P}_b \in \mathcal{SP}_{i,md}$ with density sequence $\{b_0, b_1, \dots, b_n, \dots\}$ satisfying

$$\begin{aligned} \frac{b_0}{b_i} &= e^{\epsilon}, \\ \frac{b_1}{b_{i+1}} &= e^{\epsilon}, \end{aligned}$$

and $V(\mathcal{P}_a) \geq V(\mathcal{P}_b)$.

First, it is easy to see a_1 is strictly less than a_0, since if $a_0 = a_1$, then $\frac{a_1}{a_{i+1}} = \frac{a_0}{a_{i+1}} \geq \frac{a_0}{a_i} = e^{\epsilon}$. We can construct a new density sequence by increasing

a_1 and decreasing a_{i+1} to make $\frac{a_1}{a_{i+1}}$. More precisely, we define a new sequence $\{b_0, b_1, \dots, b_n, \dots\}$ as

$$\begin{aligned} b_k &= a_k, \forall k \neq 1, k \neq i+1, \\ b_1 &= a_1(1+\delta), \\ b_{i+1} &= a_{i+1}(1-\delta'), \end{aligned}$$

where $\delta > 0$ and $\delta' > 0$ are chosen such that $\frac{b_1}{b_{i+1}} = e^{\epsilon}$ and

$$\sum_{k=0}^{+\infty} b_k \, Vol(A_i(k)) = \sum_{k=0}^{+\infty} a_k \, Vol(A_i(k)) = 1.$$

It is easy to verify that $\{b_0, b_1, \dots, b_n, \dots\}$ is a valid probability density sequence and the corresponding probability distribution $\mathcal{P}_b$ satisfies the differential privacy constraint (2.7). Moreover, $V(\mathcal{P}_b) \leq V(\mathcal{P}_a)$. Therefore, we only need to consider $\mathcal{P} \in \mathcal{SP}_{i,md}$ with density sequences $\{a_0, a_1, \dots, a_n, \dots\}$ satisfying $\frac{a_0}{a_i} = e^{\epsilon}$ and $\frac{a_1}{a_{i+1}} = e^{\epsilon}$.

Use the same argument, we can show that we only need to consider $\mathcal{P} \in \mathcal{SP}_{i,md}$ with density sequences $\{a_0, a_1, \dots, a_n, \dots\}$ satisfying

$$\frac{a_k}{a_{i+k}} = e^{\epsilon}, \forall k \geq 0.$$

Therefore,

$$V^* = \inf_{\mathcal{P} \in \cup_{i=1}^{\infty} \mathcal{SP}_{i,pd}} V(\mathcal{P}).$$

Due to Lemma A.2.5, we only need to consider probability distribution with symmetric, monotonically decreasing, and geometrically decaying piecewise constant probability density function. Because of the properties of symmetry and periodically (geometrically) decaying, for this class of probability distributions, the probability density function over $\mathbb{R}^d$ is completely determined by the probability density function over the set $\{\mathbf{x} \in \mathbb{R}^d | \|\mathbf{x}\|_1 < \Delta\}$.

Next, we study what the optimal probability density function should be over the set $\{\mathbf{x} \in \mathbb{R}^d | \|\mathbf{x}\|_1 < \Delta\}$. It turns out that the optimal probability density function over the set $\{\mathbf{x} \in \mathbb{R}^d | \|\mathbf{x}\|_1 < \Delta\}$ is a step function. We use the following three steps to prove this result.

A.2.5 Step 4

Lemma A.2.6 *Consider a probability distribution $\mathcal{P}_a \in \mathcal{SP}_{i,pd}$ ($i \geq 2$) with density sequence $\{a_0, a_1, \ldots, a_n, \ldots\}$. Then there exists an integer $k(i)$ and a probability distribution $\mathcal{P}_b \in \mathcal{SP}_{i,pd}$ with density sequence $\{b_0, b_1, \ldots, b_n, \ldots\}$ such that*

$$\begin{aligned} b_0 = b_1 = b_2 = \cdots = b_{k(i)}, \\ \frac{b_0}{b_{i-1}} = e^{\epsilon}, \end{aligned}$$

and

$$V(\mathcal{P}_b) \leq V(\mathcal{P}_a).$$

Proof 6 *For $0 \leq k \leq i-1$, define*

$$w_k \triangleq \sum_{j=0}^{+\infty} e^{-j\epsilon} \int \int \cdots \int_{(j+\frac{k}{i})\Delta \leq \|\mathbf{x}\|_1 < (j+\frac{k}{i})\Delta} \mathcal{C}(\mathbf{x}) dx_1 dx_2 \ldots dx_d,$$

and

$$u_k \triangleq \sum_{j=0}^{+\infty} e^{-j\epsilon} Vol(A_i(ji+k)).$$

Then the cost $V(\mathcal{P}_a) = \sum_{k=0}^{i-1} w_k a_k$, and the constraint on a_k is that

$$\begin{aligned} a_0 \geq a_1 \geq \cdots \geq a_{i-1}, \\ a_0 \leq a_{i-1} e^{\epsilon}, \\ \sum_{k=0}^{+\infty} u_k a_k = 1. \end{aligned}$$

Therefore, to minimize $V(\mathcal{P})$ among all probability distributions $\mathcal{P} \in \mathcal{SP}_{i,pd}$, *we need to solve the following linear programming problem*

$$\begin{aligned} \textit{minimize}_{a_0,a_1,\ldots,a_{i-1}} & \sum_{k=0}^{i-1} w_k a_k, \\ \textit{subject to} \quad & a_0 \geq a_1 \geq \cdots \geq a_{i-1}, \\ & a_0 \leq a_{i-1} e^{\epsilon}, \\ & \sum_{k=0}^{+\infty} u_k a_k = 1. \end{aligned}$$

Let

$$h_k \triangleq \frac{w_k}{u_k}. \tag{A.12}$$

In the following we show that when $d = 2$, there exists an integer $k(i)$ such that:

$$h_0 \geq h_1 \geq \cdots \geq h_{k(i)}, \tag{A.13}$$

$$h_{k(i)} \leq h_{k(i)+1} \leq \cdots \leq h_{i-1}, \tag{A.14}$$

$$h_0 \leq h_{i-1}. \tag{A.15}$$

When $d = 2$,

$$\begin{aligned} h_k &= \frac{w_k}{u_k} \\ &= \frac{\frac{4}{3}\frac{\Delta^3}{i^3}\sum_{j=0}^{+\infty} e^{-j\epsilon}(1 + 3(ji+k) + 3(ij+k)^2}{2\frac{\Delta^2}{i^2}\sum_{j=0}^{+\infty} e^{-j\epsilon}(1 + 2(ji+k))} \\ &= \frac{2}{3}\frac{\Delta}{i}\frac{3i^2c_2 + (6ik+3i)c_1 + (1+3k+3k^2)c_0}{(1+2k)c_0 + 2ic_1}. \end{aligned}$$

Let $g(k) =\triangleq \frac{3i^2c_2+(6ik+3i)c_1+(1+3k+3k^2)c_0}{(1+2k)c_0+2ic_1}$. It is easy to compute the derivative of $g(k)$ with respect to k:

$$g'(k) = \frac{6c_0^2k^2 + 6c_0^2k + c_0^2 + 12c_0c_1ik + 6c_0c_1i - 6c_2c_0i^2 + 12c_1^2i^2}{((1+2k)c_0 + 2ic_1)^2}.$$

Note that the numerator of $g'(k)$ is an increasing function of k, and

$$\begin{aligned} g'(0) &= c_0^2 + 6c_0c_1i - 6c_2c_0i^2 + 12c_1^2i^2 \\ &= \frac{b(6i^2-6i+1)-1}{(b-1)^3} < 0, \end{aligned}$$

for sufficiently large i, and

$$g'(i-1) = \frac{6i^2-6i+1-b}{(1-b)^3} > 0.$$

Therefore, h_k first increases as k increases, and then decreases as k increases to $i-1$. Hence, there exists an integer $k(i)$ such that (A.13) *and* (A.14) *hold.*

Next we compare h_{i-1} and h_0:

$$\begin{aligned} h_{i-1} - h_0 &= \frac{w_{i-1}}{u_{i-1}} - \frac{w_0}{u_0} \\ &= \frac{2}{3}\frac{\Delta}{i}\frac{(3i-2)(b-1)^2(i-1)}{(2bi-b+1)(b+2i-1)} > 0. \end{aligned}$$

Hence, (A.15) *also holds.*

We are now ready to prove Lemma A.2.6. Suppose $a_{k(i)} < a_{k(i)-1}$. We can scale up $a_{k(i)}$ and scale down $a_{k(i)-1}$ to make $a_{k(i)} = a_{k(i)-1}$. Since $h_{k(i)} \le h_{k(i)-1}$, i.e., $\frac{w_{k(i)}}{u_{k(i)}} \le \frac{w_{k(i)-1}}{u_{k(i)-1}}$, this scaling operation will not increase the cost $V(\mathcal{P}_a)$. Now we have $a_{k(i)} = a_{k(i)-1}$.

Suppose $a_{k(i)} = a_{k(i)-1} < a_{k(i)-2}$. Then we can scale up $a_{k(i)}$ and $a_{k(i)-1}$, and scale down $a_{k(i)-2}$ to make $a_{k(i)} = a_{k(i)-1} = a_{k(i)-2}$. Since $h_{k(i)} \le h_{k(i)-1} \le h_{k(i)-2}$, this scaling operation will not increase the cost $V(\mathcal{P}_a)$. Now we have $a_{k(i)} = a_{k(i)-1} = a_{k(i)-2}$.

After $k(i)$ steps of these scaling operations, we can make $a_0 = a_1 = \cdots = a_{k(i)}$, and this will not increase the cost $V(\mathcal{P}_a)$.

Finally, if $\frac{a_0}{a_{i-1}} < e^{\epsilon}$, we can scale up $a_0, a_1, \ldots, a_{k(i)}$, and scale down a_{i-1} to make $\frac{a_0}{a_{i-1}} = e^{\epsilon}$. Since $h_{i-1} \ge h_0 \ge h_1 \ge \cdots \ge h_{k(i)}$, this scaling operation will not increase the cost $V(\mathcal{P}_a)$.

Let $\mathcal{P}_b$ be the probability distribution we obtained after the $k(i)+1$ steps of scaling operations. Then $\mathcal{P}_b \in \mathcal{SP}_{i,pd}$, and its density sequence

$$\{b_0, b_1, \ldots, b_n, \ldots\}$$

satisfies

$$\begin{aligned} b_0 &= b_1 = b_2 = \cdots = b_{k(i)}, \\ \frac{b_0}{b_{i-1}} &= e^{\epsilon}, \end{aligned}$$

and

$$V(\mathcal{P}_b) \leq V(\mathcal{P}_a).$$

This completes the proof of Lemma A.2.6.

Therefore, due to Lemma A.2.6, for sufficiently large i, we only need to consider probability distributions $\mathcal{P} \in \mathcal{SP}_{i,\mathrm{pd}}$ with density sequence

$$\{a_0, a_1, \ldots, a_n, \ldots\}$$

satisfying

$$a_0 = a_1 = a_2 = \cdots = a_{k(i)}, \tag{A.16}$$

$$\frac{b_0}{b_{i-1}} = e^{\epsilon}. \tag{A.17}$$

More precisely, define

$$\begin{aligned} \mathcal{SP}_{i,\mathrm{fr}} = \{ & \mathcal{P} \in \mathcal{SP}_{i,\mathrm{pd}} | \mathcal{P} \text{ has density sequence} \\ & \{a_0, a_1, \ldots, a_n, \ldots\} \text{ satisfying (A.16) and (A.17)}\}. \end{aligned}$$

Then due to Lemma A.2.6,

Lemma A.2.7

$$V^* = \inf_{\mathcal{P} \in \cup_{i=3}^{\infty} \mathcal{SP}_{i,fr}} V(\mathcal{P}).$$

Next, we argue that for each probability distribution $\mathcal{P} \in \mathcal{SP}_{i,\mathrm{fr}}$ ($i \geq 3$) with density sequence $\{a_0, a_1, \ldots, a_n, \ldots\}$, we can assume that there exists an integer $k(i)+1 \leq k \leq (i-2)$, such that

$$a_j = a_0, \forall 0 \leq j < k, \tag{A.18}$$

$$a_j = a_{i-1}, \forall k < j < i. \tag{A.19}$$

More precisely,

Lemma A.2.8 *Consider a probability distribution $\mathcal{P}_a \in \mathcal{SP}_{i,fr}$ ($i \geq 3$) with density sequence $\{a_0, a_1, \ldots, a_n, \ldots\}$. Then there exists a probability distribution $\mathcal{P}_b \in \mathcal{SP}_{i,fr}$ with density sequence $\{b_0, b_1, \ldots, b_n, \ldots\}$ such that there exists an integer $k(i)+1 \leq k \leq (i-2)$ with*

$$b_j = a_0, \forall\ 0 \leq j < k, \tag{A.20}$$

$$b_j = a_{i-1}, \forall\ k < j < i, \tag{A.21}$$

and

$$V(\mathcal{P}_b) \leq V(\mathcal{P}_a). \tag{A.22}$$

Proof 7 *If there exists an integer $k(i)+1 \leq k \leq (i-2)$ such that*

$$a_j = a_0, \forall\ 0 \leq j < k,$$
$$a_j = a_{i-1}, \forall\ k < j < i,$$

then we can set $\mathcal{P}_b = \mathcal{P}_a$.

Otherwise, let k_1 be the smallest integer in $\{k(i)+1, k(i)+2, \ldots, i-1\}$ such that

$$a_{k_1} \neq a_0,$$

and let k_2 be the biggest integer in $\{k(i)+1, k(i)+2, \ldots, i-1\}$ such that

$$a_{k_2} \neq a_{i-1}.$$

It is easy to see that $k_1 \neq k_2$. Then we can scale up a_{k_1} and scale down a_{k_2} simultaneously until either $a_{k_1} = a_0$ or $a_{k_2} = a_{i-1}$. Since $h_k \triangleq \frac{w_k}{u_k}$ is an increasing function of k when $k > k(i)$, and $k(i) < k_1 < k_2$, this scaling operation will not increase the cost.

After this scaling operation we can update k_1 and k_2, and either k_1 is increased by one or k_2 is decreased by one.

Therefore, continue in this way, and finally we will obtain a probability distribution $\mathcal{P}_b \in \mathcal{SP}_{i,fr}$ with density sequence $\{b_0, b_1, \ldots, b_n, \ldots\}$ such that (A.20)*,* (A.21) *and* (A.22) *hold.*

This completes the proof.

Define

$$\begin{aligned}\mathcal{SP}_{i,\text{step}} = \{\mathcal{P} \in \mathcal{SP}_{i,\text{fr}} \mid \mathcal{P} \text{ has density sequence } \{a_0, a_1, \ldots, a_n, \ldots\} \\ \text{satisfying(A.20) and (A.21) for some } k(i) < k \leq (i-2)\}.\end{aligned}$$

Then due to Lemma A.2.8, we have Lemma A.2.9.

Lemma A.2.9

$$V^* = \inf_{\mathcal{P} \in \cup_{i=3}^{\infty} \mathcal{SP}_{i,step}} V(\mathcal{P}).$$

As $i \to \infty$, the probability density function of $\mathcal{P} \in \mathcal{SP}_{i,\text{fr}}$ will converge to a multidimensional staircase function. Therefore, for $d = 2$ and the cost function $\mathcal{L}(\mathbf{x}) = \|\mathbf{x}\|_1, \forall \mathbf{x} \in \mathbb{R}^2$, then

$$\inf_{\mathcal{P} \in \mathcal{SP}} \int\int_{\mathbb{R}^2} \mathcal{L}(\mathbf{x}) \mathcal{P}(dx_1 dx_2) = \inf_{\gamma \in [0,1]} \int\int_{\mathbb{R}^2} \mathcal{L}(\mathbf{x}) f_{\gamma}(\mathbf{x}) dx_1 dx_2.$$

This completes the proof of Theorem 2.3.1.

A.3 Composition Theorem in Differential Privacy

A.3.1 Proof of Theorem 2.4.3

We propose a simple mechanism and prove that the proposed mechanism dominates over all (ε, δ)-differentially private mechanisms. Analyzing the privacy region achieved by the k-fold composition of the proposed mechanism, we get a bound on the privacy region under adaptive composition. This gives an exact characterization of privacy under composition, since we show both converse and achievability. We prove that no other family of mechanisms can achieve 'more degraded' privacy (converse), and that there is a mechanism that we propose which achieves the privacy region (achievability).

Achievability

We propose the following simple mechanism $\tilde{M}_i$ at the i-th step in the composition. Null hypothesis ($b = 0$) outcomes $X^{i,0} = M_i(D^{i,0}, q_i)$'s which are independent and identically distributed as a discrete random variable $\tilde{X}_0 \sim \tilde{P}_0(\cdot)$, where

$$\mathbb{P}(\tilde{X}_0 = x) \;=\; \tilde{P}_0(x) \;\equiv\; \begin{cases} \delta & \text{for } x = 0\ , \\ \frac{(1-\delta)\,e^{\varepsilon}}{1+e^{\varepsilon}} & \text{for } x = 1\ , \\ \frac{1-\delta}{1+e^{\varepsilon}} & \text{for } x = 2\ , \\ 0 & \text{for } x = 3\ . \end{cases} \tag{A.23}$$

Alternative hypothesis ($b = 1$) outcomes $X^{i,1} = M_i(D^{i,1}, q_i)$'s are independent and identically distributed as a discrete random variable $\tilde{X}_1 \sim \tilde{P}_1(\cdot)$, where

$$\mathbb{P}(\tilde{X}_1 = x) \;=\; \tilde{P}_1(x) \;\equiv\; \begin{cases} 0 & \text{for } x = 0\ , \\ \frac{1-\delta}{1+e^{\varepsilon}} & \text{for } x = 1\ , \\ \frac{(1-\delta)\,e^{\varepsilon}}{1+e^{\varepsilon}} & \text{for } x = 2\ , \\ \delta & \text{for } x = 3\ . \end{cases} \tag{A.24}$$

In particular, the output of this mechanism does not depend on the database $D^{i,b}$ or the query q_i, and only depends on the hypothesis b. The privacy region of a single access to this mechanism is $\mathcal{R}(\varepsilon, \delta)$ in Figure 2.1. Hence, by Theorem 2.2.4, all (ε, δ)-differentially private mechanisms are dominated by this mechanism.

In general, the privacy region $\mathcal{R}(M, D_0, D_1)$ of any mechanism can be represented as an intersection of multiple $\{(\tilde{\varepsilon}_j, \tilde{\delta}_j)\}$ privacy regions. For a mechanism M, we can compute the $(\tilde{\varepsilon}_j, \tilde{\delta}_j)$ pairs representing the privacy region as follows. Given a null hypothesis database D_0, an alternative hypothesis database D_1, and a mechanism M whose output space is $\mathcal{X}$, let P_0 and P_1 denote the probability density function of the outputs $M(D_0)$ and $M(D_1)$ respectively. To simplify notations we assume that P_0 and P_1 are symmetric, i.e. there exists a permutation π over $\mathcal{X}$ such that $P_0(x) = P_1(\pi(x))$ and

$P_1(x) = P_0(\pi(x))$. This ensures that we get a symmetric privacy region.

The privacy region $\mathcal{R}(M, D_0, D_1)$ can be described by its boundaries. Since it is a convex set, a tangent line on the boundary with slope $-e^{\tilde{\varepsilon}_j}$ can be represented by the smallest $\tilde{\delta}_j$ such that

$$P_{\rm FA} \;\; \geq \;\; -e^{\tilde{\varepsilon}_j} P_{\rm MD} + 1 - \tilde{\delta}_j \;, \tag{A.25}$$

for all rejection sets (cf. Figure 2.5). Letting S denote the complement of a rejection set, such that $P_{\rm FA} = 1 - P_0(S)$ and $P_{\rm MD} = P_1(S)$, the minimum shift $\tilde{\delta}_j$ that still ensures that the privacy region is above the line (A.25) is defined as $\tilde{\delta}_j = d_{\tilde{\varepsilon}_j}(P_0, P_1)$ where

$$d_{\tilde{\varepsilon}}(P_0, P_1) \;\; \equiv \;\; \max_{S \subseteq \mathcal{X}} \Big\{ P_0(S) - e^{\tilde{\varepsilon}} \, P_1(S) \Big\} \;.$$

The privacy region of a mechanism is completely described by the set of slopes and shifts, $\{(\tilde{\varepsilon}_j, \tilde{\delta}_j) \, : \, \tilde{\varepsilon}_j \in E \text{ and } \tilde{\delta}_j = d_{\tilde{\varepsilon}_j}(P_0, P_1)\}$, where

$$E \;\; \equiv \;\; \{\, 0 \leq \tilde{\varepsilon} < \infty \, : \, P_0(x) = e^{\tilde{\varepsilon}} \, P_1(x) \text{ for some } x \in \mathcal{X} \} \;.$$

Any $\tilde{\varepsilon} \notin E$ does not contribute to the boundary of the privacy region. For the above example distributions $\tilde{P}_0$ and $\tilde{P}_1$, $E = \{\varepsilon\}$ and $d_\varepsilon(\tilde{P}_0, \tilde{P}_1) = \delta$.

Remark 3 *For a database access mechanism M over a output space $\mathcal{X}$ and a pair of neighboring databases D_0 and D_1, let P_0 and P_1 denote the probability density function for random variables $M(D_0)$ and $M(D_1)$ respectively. Assume there exists a permutation π over $\mathcal{X}$ such that $P_0(x) = P_1(\pi(x))$. Then, the privacy region is*

$$\mathcal{R}(\, M, D_0, D_1 \,) \;\; = \;\; \bigcap_{\tilde{\varepsilon} \in E} \mathcal{R}\big(\, \tilde{\varepsilon}, d_{\tilde{\varepsilon}}(P_0, P_1) \big) \;,$$

where $\mathcal{R}(M, D, D')$ and $\mathcal{R}(\tilde{\varepsilon}, \tilde{\delta})$ are defined as in (2.3) *and* (2.2)*.*

The symmetry assumption is to simplify notations, and the analysis can be easily generalized to deal with non-symmetric distributions.

Now consider a k-fold composition experiment, where at each sequential access $\tilde{M}_i$, we receive a random output $X^{i,b}$ independent and identically distributed as $\tilde{X}_b$. We can explicitly characterize the distribution of k-fold composition of the outcomes: $\mathbb{P}(X^{1,b} = x_1, \ldots, X^{k,b} = x_k) = \prod_{x=1}^{k} \tilde{P}_b(x_i)$. It follows form the structure of these two discrete distributions that, $E = \{e^{(k-2\lfloor k/2 \rfloor)\varepsilon}, e^{(k+2-2\lfloor k/2 \rfloor)\varepsilon}, \ldots, e^{(k-2)\varepsilon}, e^{k\varepsilon}\}$. After some algebra, it also follows that

$$d_{(k-2i)\varepsilon}\big((\tilde{P}_0)^k, (\tilde{P}_1)^k \big) = \; 1 - (1-\delta)^k + (1-\delta)^k \frac{\sum_{\ell=0}^{i-1} \binom{k}{\ell} \Big(e^{\varepsilon(k-\ell)} - e^{\varepsilon(k-2i+\ell)} \Big)}{(1+e^{\varepsilon})^k}$$

for $i \in \{0, \ldots, \lfloor k/2 \rfloor\}$. From Remark 3, it follows that the privacy region is $\mathcal{R}(\{\varepsilon_i, \delta_i\}) = \bigcap_{i=0}^{\lfloor k/2 \rfloor} \mathcal{R}\big(\varepsilon_i, \delta_i\big)$, where $\varepsilon_i = (k-2i)\varepsilon$ and δ_i's are defined as in (2.12). Figure 2.4 shows this privacy region for $k = 1, \ldots, 5$ and for $\varepsilon = 0.4$ and for two values of $\delta = 0$ and $\delta = 0.1$.

Converse

We will now prove that this region is the largest region achievable under k-fold adaptive composition of any (ε, δ)-differentially private mechanisms.

From Corollary 2.2.2, any mechanism whose privacy region is included in $\mathcal{R}(\{\varepsilon_i, \delta_i\})$ satisfies $(\tilde{\varepsilon}, \tilde{\delta})$-differential privacy. We are left to prove that for the family of all (ε, δ)-differentially private mechanisms, the privacy region of the k-fold composition experiment is included inside $\mathcal{R}(\{\varepsilon_i, \delta_i\})$. To this end, consider the following composition experiment, which reproduces the *view of the adversary* from the original composition experiment.

At each time step i, we generate a random variable $X^{i,b}$ distributed as $\tilde{X}_b$ independent of any other random events, and call this the output of a database access mechanism $\tilde{M}_i$ such that $\tilde{M}_i(D^{i,b}, q_i) = X^{i,b}$. Since, $X^{i,b}$ only depends on b, and is independent of the actual database or the query, we use $\tilde{M}_i(b)$ to denote this outcome.

We know that $\tilde{M}_i(b)$ has privacy region $\mathcal{R}(\varepsilon, \delta)$ for any choices of $D^{i,0}$, $D^{i,1}$ and q_i. Now consider the mechanism M_i from the original experiment. Since it is (ε, δ)-differentially private, we know from Theorem 2.2.1 that $\mathcal{R}(M_i, D^{i,0}, D^{i,1}) \subseteq \mathcal{R}(\varepsilon, \delta)$ for any choice of neighboring databases $D^{i,0}$, $D^{i,1}$. Hence, from the converse of data processing inequality (Theorem 2.2.4), we know that there exists a mechanism T_i that takes as input $X^{i,b}$ and produces an output $Y^{i,b}$ which is distributed as $M_i(D^{i,b}, q_i)$ for all $b \in \{0, 1\}$. Hence, $Y^{i,b}$ is independent of the past conditioned on $X^{i,b}$, $D^{i,0}$, $D^{i,1}$, q_i, M_i. Precisely we have the following Markov chain:

$$(b, R, \{X^{\ell,b}, D^{\ell,0}, D^{\ell,1}, q_\ell, M_\ell\}_{\ell\in[i-1]})\text{–}(X^{i,b}, D^{i,0}, D^{i,1}, q_i, M_i)\text{–}Y^{i,b} \ ,$$

where R is any internal randomness of the adversary $\mathcal{A}$. Since, (X, Y)–Z–W implies X–(Y, Z)–W, we have

$$b\text{–}(R, \{X^{\ell,b}, D^{\ell,0}, D^{\ell,1}, q_\ell, M_\ell\}_{\ell\in[i]})\text{–}Y^{i,b} \ .$$

Notice that if we know R and the outcomes $\{Y^{\ell,b}\}_{\ell\in[i]}$, then we can reproduce the original experiment until time i. This is because the choices of $D^{i,0}, D^{i,1}, q_i, M_i$ are exactly specified by R and $\{Y^{\ell,b}\}_{\ell\in[i]}$. Hence, we can simplify the Markov chain as

$$b\text{–}(R, X^{i,b}, \{X^{\ell,b}, Y^{\ell,b}\}_{\ell\in[i-1]})\text{–}Y^{i,b} \ . \tag{A.26}$$

Further, since $X^{i,b}$ is independent of the past conditioned on b, we have

$$X^{i,b}\text{–}b\text{–}(R, \{X^{\ell,b}, Y^{\ell,b}\}_{\ell\in[i-1]}) \ . \tag{A.27}$$

It follows that

$$\begin{aligned}
&\mathbb{P}(b, r, x_1 \ldots, x_k, y_1, \ldots, y_k) \\
&\quad = \ \mathbb{P}(b, r, x_1, \ldots, x_k, y_1, \ldots, y_{k-1})\mathbb{P}(y_k|r, x_1, \ldots, x_k, y_1, \ldots, y_{k-1}) \\
&\quad = \ \mathbb{P}(b, r, x_1, \ldots, x_{k-1}, y_1, \ldots, y_{k-1})\mathbb{P}(x_k|b)\mathbb{P}(y_k|r, x_1, \ldots, x_k, y_1, \ldots, y_{k-1}) \ ,
\end{aligned}$$

where we used (A.26) in the first equality and (A.27) in the second. By induction, we get a decomposition

$$\begin{aligned}
&\mathbb{P}(b, r, x_1, \ldots, x_k, y_1, \ldots, y_k) \\
&= \mathbb{P}(b, r) \prod_{i=1}^{k} \mathbb{P}(x_i|b) \prod_{i=1}^{k} \mathbb{P}(y_i|r, x_1, \ldots, x_i, y_1, \ldots, y_{i-1}) \\
&= \mathbb{P}(b, r, x_1, \ldots, x_k)\mathbb{P}(y_1, \ldots, y_k|r, x_1, \ldots, x_k) \\
&= \mathbb{P}(b|r, x_1, \ldots, x_k)\, \mathbb{P}(y_1, \ldots, y_k, r, x_1, \ldots, x_k) \ .
\end{aligned}$$

From the construction of the experiment, it also follows that the internal randomness R is independent of the hypothesis b and the outcomes

$X^{i,b}$'s: $\mathbb{P}(b|r, x_1, \ldots, x_k) = \mathbb{P}(b|x_1, \ldots, x_k)$. Then, marginalizing over R, we get $\mathbb{P}(b, x_1, \ldots, x_k, y_1, \ldots, y_k) = \mathbb{P}(b|x_1, \ldots, x_k)\,\mathbb{P}(y_1, \ldots, y_k, x_1, \ldots, x_k)$. This implies the following Markov chain:

$$b\text{–}(\{X^{i,b}\}_{i\in[k]})\text{–}(\{Y^{i,b}\}_{i\in[k]}) \ , \tag{A.28}$$

and it follows that a set of mechanisms $(M_1, \ldots, M_k)$ dominates $(\tilde{M}_1, \ldots, \tilde{M}_k)$ for two databases $\{D^{i,0}\}_{i\in[k]}$ and $\{D^{i,1}\}_{i\in[k]}$. By the data processing inequality for differential privacy (Theorem 2.2.3), this implies that

$$\begin{aligned}
&\mathcal{R}\big(\{M_i\}_{i\in[k]}, \{D^{i,0}\}_{i\in[k]}, \{D^{i,1}\}_{i\in[k]}\big)\\
&\subseteq \mathcal{R}\big(\{\tilde{M}_i\}_{i\in[k]}, \{D^{i,0}\}_{i\in[k]}, \{D^{i,1}\}_{i\in[k]}\big)\\
&= \mathcal{R}\big(\{\varepsilon_i, \delta_i\}\big) \ .
\end{aligned}$$

This finishes the proof of the desired claim.

Alternatively, one can prove (A.28), using a probabilistic graphical model. Precisely, the Bayesian network shown in Figure A.1 describes the dependencies among various random quantities of the experiment described above. Since the set of nodes $(X^{1,b}, X^{2,b}, X^{3,b}, X^{4,b})$ d-separates node b from the rest of the Bayesian network, it follows immediately from the Markov property of this Bayesian network that (A.28) is true (cf. [61]).

A.3.2 Proof of Theorem 2.4.4

We need to provide an outer bound on the privacy region achieved by $\tilde{X}_0$ and $\tilde{X}_1$ defined in (A.23) and (A.24) under k-fold composition. Let P_0 denote the probability mass function of $\tilde{X}_0$ and P_1 denote the PMF of $\tilde{X}_1$. Also, let P_0^k and P_1^k denote the joint PMF of k i.i.d. copies of $\tilde{X}_0$ and $\tilde{X}_1$ respectively. Also, for a set $S \subseteq \mathcal{X}^k$, we let $P_0^k(S) = \sum_{x\in S} P_0^k(x)$. In our example,

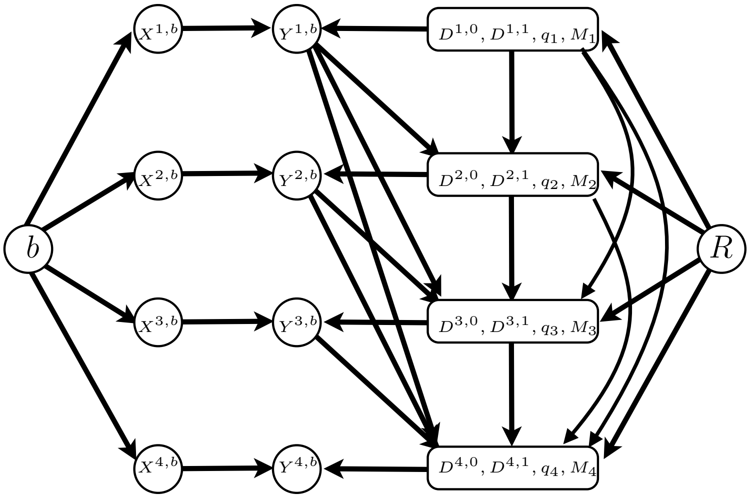

$X^{1,b}$
$Y^{1,b}$
$D^{1,0}, D^{1,1}, q_1, M_1$
$X^{2,b}$
$Y^{2,b}$
$D^{2,0}, D^{2,1}, q_2, M_2$
b
R
$X^{3,b}$
$Y^{3,b}$
$D^{3,0}, D^{3,1}, q_3, M_3$
$X^{4,b}$
$Y^{4,b}$
$D^{4,0}, D^{4,1}, q_4, M_4$

Figure A.1: Bayesian network representation of the composition experiment. The subset of nodes $(X^{1,b}, X^{2,b}, X^{3,b}, X^{4,b})$ d-separates node b from the rest of the network.

$\mathcal{X} = \{1, 2, 3, 4\}$, and

$$\begin{aligned} P_0 &= \begin{bmatrix} \delta & \frac{(1-\delta)e^{\varepsilon}}{1+e^{\varepsilon}} & \frac{1-\delta}{1+e^{\varepsilon}} & 0 \end{bmatrix} , \\ P_1 &= \begin{bmatrix} 0 & \frac{1-\delta}{1+e^{\varepsilon}} & \frac{(1-\delta)e^{\varepsilon}}{1+e^{\varepsilon}} & \delta \end{bmatrix} , \\ P_0^2 &= \begin{bmatrix} \delta^2 & \delta\frac{(1-\delta)e^{\varepsilon}}{1+e^{\varepsilon}} & \delta\frac{(1-\delta)}{1+e^{\varepsilon}} & 0 \\ \delta\frac{(1-\delta)e^{\varepsilon}}{1+e^{\varepsilon}} & \left(\frac{(1-\delta)e^{\varepsilon}}{1+e^{\varepsilon}}\right)^2 & \left(\frac{1-\delta}{1+e^{\varepsilon}}\right)^2 e^{\varepsilon} & 0 \\ \delta\frac{1-\delta}{1+e^{\varepsilon}} & \left(\frac{1-\delta}{1+e^{\varepsilon}}\right)^2 e^{\varepsilon} & \left(\frac{1-\delta}{1+e^{\varepsilon}}\right)^2 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} , \text{etc.} \end{aligned}$$

We can compute the privacy region from P_0^k and P_1^k directly, by computing the line tangent to the boundary. A tangent line with slope $-e^{\tilde{\varepsilon}}$ can be represented as

$$P_{\text{FA}} = -e^{\tilde{\varepsilon}} P_{\text{MD}} + 1 - d_{\tilde{\varepsilon}}(P_0^k, P_1^k) \; . \tag{A.29}$$

To find the tangent line, we need to maximize the shift, which is equivalent to moving the line downward until it is tangent to the boundary of the privacy region (cf. Figure 2.5).

$$d_{\tilde{\varepsilon}}(P_0^k, P_1^k) \;\equiv\; \max_{S \subseteq \mathcal{X}^k} P_0^k(S) - e^{\tilde{\varepsilon}} P_1^k(S) \;.$$

Notice that the maximum is achieved by a set $B \equiv \{x \in \mathcal{X}^k \,|\, P_0^k(x) \geq e^{\tilde{\varepsilon}} P_1^k(x)\}$. Then,

$$d_{\tilde{\varepsilon}}(P_0^k, P_1^k) \;=\; P_0^k(B) - e^{\tilde{\varepsilon}} P_1^k(B) \;.$$

For the purpose of proving the bound of the form (2.12), we separate the analysis of the above formula into two parts: one where either $P_0^k(x)$ or $P_1^k(x)$ is zero and the other when both are positive. Effectively, this separation allows us to treat the effects of $(\varepsilon, 0)$-differential privacy and $(0, \delta)$-differential privacy separately. In previous work [16], they separated the analysis in a similar way. Here we provide a simpler proof technique. Further, all the proof techniques we use naturally generalize to compositions of general (ε, δ)-differentially private mechanisms other than the specific example of $\tilde{X}_0$ and $\tilde{X}_1$ we consider in this section.

Let $\tilde{X}_0^k$ denote a k-dimensional random vector whose entries are independent copies of $\tilde{X}_0$. We partition B into two sets: $B = B_0 \bigcup B_1$ and $B_0 \bigcap B_1 = \emptyset$. Let $B_0 \equiv \{x \in \mathcal{X}^k \;:\; P_0^k(x) \geq e^{\tilde{\varepsilon}} P_1^k(x)$, and $P_1^k(x) = 0\}$ and $B_1 \equiv \{x \in \mathcal{X}^k \;:\; P_0^k(x) \geq e^{\tilde{\varepsilon}} P_1^k(x)$, and $P_1^k(x) > 0\}$. Then, it is not hard to see that $P_0^k(B_0) = 1 - \mathbb{P}(\tilde{X}_0^k \in \{1,2,3\}^k) = 1 - (1-\delta)^k$, $P_1^k(B_0) = 0$, $P_0^k(B_1) = P_0^k(B_1 | \tilde{X}_0^k \in \{1,2\}^k)\mathbb{P}(\tilde{X}_0^k \in \{1,2\}^k) = (1-\delta)^k\, P_0^k(B_1 | \tilde{X}_0^k \in \{1,2\}^k)$, and $P_1^k(B_1) = (1-\delta)^k\, P_1^k(B_1 | \tilde{X}_1^k \in \{1,2\}^k)$. It follows that

$$\begin{aligned}
&P_0^k(B_0) - e^{\tilde{\varepsilon}} P_1^k(B_0) = 1 - (1-\delta)^k \;, \text{ and} \\
&P_0^k(B_1) - e^{\tilde{\varepsilon}} P_1^k(B_1) = (1-\delta)^k \big(P_0^k(B_1 | \tilde{X}_0^k \in \{1,2\}^k) \\
&\qquad\qquad\qquad\qquad\qquad - e^{\tilde{\varepsilon}} P_1^k(B_1 | \tilde{X}_1^k \in \{1,2\}^k) \big) \;.
\end{aligned}$$

Let $\tilde{P}_0^k(x) \equiv P_0^k(x | x \in \{1,2\}^k)$ and $\tilde{P}_1^k(x) \equiv P_1^k(x | x \in \{1,2\}^k)$. Then, we have

$$\begin{aligned}
d_{\tilde{\varepsilon}}(P_0^k, P_1^k) &= P_0^k(B_0) - e^{\tilde{\varepsilon}} P_1^k(B_0) + P_0^k(B_1) - e^{\tilde{\varepsilon}} P_1^k(B_1) \\
&= 1 - (1-\delta)^k + (1-\delta)^k \big(\tilde{P}_0^k(B_1) - e^{\tilde{\varepsilon}} \tilde{P}_1^k(B_1) \big) \;. \quad \text{(A.30)}
\end{aligned}$$

Now, we focus on upper bounding $\tilde{P}_0^k(B_1) - e^{\tilde{\varepsilon}}\tilde{P}_1^k(B_1)$, using a variant of Chernoff's tail bound. Notice that

$$\begin{aligned}\tilde{P}_0^k(B_1) - e^{\tilde{\varepsilon}}\tilde{P}_1^k(B_1) &= \mathbb{E}_{\tilde{P}_0^k}\big[\mathbb{I}_{\left(\log(\tilde{P}_0^k(\tilde{X}^k)/\tilde{P}_1^k(\tilde{X}^k))\geq\tilde{\varepsilon}\right)}\big] \\ &\quad - e^{\tilde{\varepsilon}}\mathbb{E}_{\tilde{P}_0^k}\big[\mathbb{I}_{\left(\log(\tilde{P}_0^k(\tilde{X}^k)/\tilde{P}_1^k(\tilde{X}^k))\geq\tilde{\varepsilon}\right)}\frac{\tilde{P}_1^k(\tilde{X}^k)}{\tilde{P}_0^k(\tilde{X}^k)}\big] \\ &= \mathbb{E}_{\tilde{P}_0^k}\Big[\mathbb{I}_{\left(\log(\tilde{P}_0^k(\tilde{X}^k)/\tilde{P}_1^k(\tilde{X}^k))\geq\tilde{\varepsilon}\right)}\Big(1 - e^{\tilde{\varepsilon}}\frac{\tilde{P}_1^k(\tilde{X}^k)}{\tilde{P}_0^k(\tilde{X}^k)}\Big)\Big] \\ &\leq \mathbb{E}\big[e^{\lambda Z - \lambda\tilde{\varepsilon} + \lambda\log\lambda - (\lambda+1)\log(\lambda+1)}\big]\,, \qquad \text{(A.31)}\end{aligned}$$

where we use a random variable $Z \equiv \log(\tilde{P}_0^k(\tilde{X}_0^k)/\tilde{P}_1^k(\tilde{X}_0^k))$ and the last line follows from $\mathbb{I}_{(x\geq\tilde{\varepsilon})}(1 - e^{\tilde{\varepsilon}-x}) \leq e^{\lambda(x-\tilde{\varepsilon})+\lambda\log\lambda-(\lambda+1)\log(\lambda+1)}$ for any $\lambda \geq 0$. To show this inequality, notice that the right-hand side is always non-negative. So it is sufficient to show that the inequality holds, without the indicator on the left-hand side. Precisely, let $f(x) = e^{\lambda(x-\tilde{\varepsilon})+\lambda\log\lambda-(\lambda+1)\log(\lambda+1)} + e^{\tilde{\varepsilon}-x} - 1$. This is a convex function with $f(x^*) = 0$ and $f'(x^*) = 0$ at $x^* = \tilde{\varepsilon} + \log((\lambda+1)/\lambda)$. It follows that this is a non-negative function.

Next, we give an upper bound on the moment generating function of Z.

$$\begin{aligned}\mathbb{E}_{\tilde{P}_0}\big[e^{\lambda\log(P_0(X)/P_1(X))}\big] &= \frac{e^{\varepsilon}}{e^{\varepsilon}+1}e^{\lambda\varepsilon} + \frac{1}{e^{\varepsilon}+1}e^{-\lambda\varepsilon} \\ &\leq e^{\frac{e^{\varepsilon}-1}{e^{\varepsilon}+1}\lambda\varepsilon + \frac{1}{2}\lambda^2\varepsilon^2}\,,\end{aligned}$$

for any λ, which follows from the fact that $pe^x + (1-p)e^{-x} \leq e^{(2p-1)x+(1/2)x^2}$ for any $x \in \mathbb{R}$ and $p \in [0,1]$ [62, Lemma A.1.5]. Substituting this into (A.31) with a choice of $\lambda = \frac{\tilde{\varepsilon} - k\varepsilon(e^{\varepsilon}-1)/(e^{\varepsilon}+1)}{k\varepsilon^2}$, we get

$$
\begin{aligned}
&\tilde{P}_0^k(B_1) - e^{\tilde{\varepsilon}}\tilde{P}_1^k(B_1) \\
&\qquad \le \exp\Big\{\frac{e^{\varepsilon}-1}{e^{\varepsilon}+1}\lambda\varepsilon k + \frac{1}{2}\lambda^2\varepsilon^2 k - \lambda\tilde{\varepsilon} + \lambda\log\lambda - (\lambda+1)\log(\lambda+1)\Big\} \\
&\qquad = \exp\Big\{-\frac{k\varepsilon^2}{2}\Big(\lambda - \frac{1}{k\varepsilon^2}\Big(\tilde{\varepsilon} - k\varepsilon\frac{e^{\varepsilon}-1}{e^{\varepsilon}+1}\Big)\Big)^2 \\
&\qquad\qquad - \frac{1}{2k\varepsilon^2}\Big(\tilde{\varepsilon} - \frac{k\varepsilon(e^{\varepsilon}-1)}{e^{\varepsilon}+1}\Big)^2 + \lambda\log\frac{\lambda}{\lambda+1} - \log(\lambda+1)\Big\} \\
&\qquad \le \exp\Big\{-\frac{1}{2k\varepsilon^2}\Big(\tilde{\varepsilon} - k\varepsilon\frac{e^{\varepsilon}-1}{e^{\varepsilon}+1}\Big)^2 - \log(\lambda+1)\Big\} \\
&\qquad \le \frac{1}{1+\frac{\tilde{\varepsilon}-k\varepsilon(e^{\varepsilon}-1)/(e^{\varepsilon}+1)}{k\varepsilon^2}}\exp\Big\{-\frac{1}{2k\varepsilon^2}\Big(\tilde{\varepsilon} - k\varepsilon\frac{e^{\varepsilon}-1}{e^{\varepsilon}+1}\Big)^2\Big\} \\
&\qquad = \frac{1}{1+\frac{\sqrt{2k\varepsilon^2\log(e+(\sqrt{k\varepsilon^2}/\tilde{\delta}))}}{k\varepsilon^2}}\,\frac{1}{e+\frac{\sqrt{k\varepsilon^2}}{\tilde{\delta}}} \\
&\qquad \le \frac{1}{\sqrt{k\varepsilon^2}+\sqrt{2\log(e+(\sqrt{k\varepsilon^2}/\tilde{\delta}))}}\,\frac{\tilde{\delta}}{\frac{e\tilde{\delta}}{\sqrt{k\varepsilon^2}}+1}\ ,
\end{aligned}
$$

for our choice of $\tilde{\varepsilon} = k\varepsilon(e^{\varepsilon}-1)/(e^{\varepsilon}+1) + \varepsilon\sqrt{2k\,\log(e+(\sqrt{k\varepsilon^2}/\tilde{\delta}))}$. The right-hand side is always less than $\tilde{\delta}$.

Similarly, one can show that the right-hand side is less than $\tilde{\delta}$ for the choice of $\tilde{\varepsilon} = k\varepsilon(e^{\varepsilon}-1)/(e^{\varepsilon}+1) + \varepsilon\sqrt{2k\,\log(1/\tilde{\delta})}$. We get that the k-fold composition is $(\tilde{\varepsilon}, 1-(1-\delta)^k(1-\tilde{\delta}))$-differentially private.

A.3.3 Proof of Theorem 2.4.5

In this section, we closely follow the proof of Theorem 2.4.4 in Section A.3.2 carefully keeping the dependence on ℓ, the index of the composition step. For brevity, we omit the details which overlap with the proof of Theorem 2.4.4. By the same argument as in the proof of Theorem 2.4.3, we only need to provide an outer bound on the privacy region achieved by $\tilde{X}_0^{(\ell)}$ and $\tilde{X}_1^{(\ell)}$ under k-fold composition, defined as

$$
\mathbb{P}(\tilde{X}_0^{(\ell)} = x) \;=\; \tilde{P}_0^{(\ell)}(x) \;\equiv\; \left\{\begin{array}{rl} \delta_\ell & \text{for } x = 0\ , \\ \frac{(1-\delta_\ell)\,e^{\varepsilon_\ell}}{1+e^{\varepsilon_\ell}} & \text{for } x = 1\ , \\ \frac{1-\delta_\ell}{1+e^{\varepsilon_\ell}} & \text{for } x = 2\ , \\ 0 & \text{for } x = 3\ . \end{array}\right. \ , \text{ and}
$$

$$\mathbb{P}(\tilde{X}_1^{(\ell)} = x) \;=\; \tilde{P}_1^{(\ell)}(x) \;\equiv\; \begin{cases} 0 & \text{for } x = 0\;, \\ \frac{1-\delta_\ell}{1+e^{\varepsilon_\ell}} & \text{for } x = 1\;, \\ \frac{(1-\delta_\ell)\,e^{\varepsilon_\ell}}{1+e^{\varepsilon_\ell}} & \text{for } x = 2\;, \\ \delta_\ell & \text{for } x = 3\;. \end{cases}$$

Using the similar notations as Section A.3.2, it follows that under k-fold composition,

$$\begin{aligned} d_{\tilde{\varepsilon}}(P_0^k, P_1^k) \;&=\; 1 - \prod_{\ell=1}^{k}(1-\delta_\ell) \\ &\qquad + \big(\tilde{P}_0^k(B_1) - e^{\tilde{\varepsilon}}\tilde{P}_1^k(B_1)\big)\prod_{\ell=1}^{k}(1-\delta_\ell)\;. \end{aligned} \tag{A.32}$$

Now, we focus on upper bounding $\tilde{P}_0^k(B_1) - e^{\tilde{\varepsilon}}\tilde{P}_1^k(B_1)$, using a variant of Chernoff's tail bound. We know that

$$\begin{aligned} \tilde{P}_0^k(B_1) - e^{\tilde{\varepsilon}}\tilde{P}_1^k(B_1) \;&=\; \mathbb{E}_{\tilde{P}_0^k}\big[\mathbb{I}_{\left(\log(\tilde{P}_0^k(\tilde{X}^k)/\tilde{P}_1^k(\tilde{X}^k))\geq\tilde{\varepsilon}\right)}\big] \\ &\qquad - e^{\tilde{\varepsilon}}\mathbb{E}_{\tilde{P}_0^k}\big[\mathbb{I}_{\left(\log(\tilde{P}_0^k(\tilde{X}^k)/\tilde{P}_1^k(\tilde{X}^k))\geq\tilde{\varepsilon}\right)}\frac{\tilde{P}_1^k(\tilde{X}^k)}{\tilde{P}_0^k(\tilde{X}^k)}\big] \\ &=\; \mathbb{E}_{\tilde{P}_0^k}\Big[\mathbb{I}_{\left(\log(\tilde{P}_0^k(\tilde{X}^k)/\tilde{P}_1^k(\tilde{X}^k))\geq\tilde{\varepsilon}\right)}\Big(1 - e^{\tilde{\varepsilon}}\frac{\tilde{P}_1^k(\tilde{X}^k)}{\tilde{P}_0^k(\tilde{X}^k)}\Big)\Big] \\ &\leq\; \mathbb{E}\big[e^{\lambda Z - \lambda\tilde{\varepsilon} + \lambda\log\lambda - (\lambda+1)\log(\lambda+1)}\big]\;, \end{aligned} \tag{A.33}$$

where we use a random variable $Z \equiv \log(\tilde{P}_0^k(\tilde{X}_0^k)/\tilde{P}_1^k(\tilde{X}_0^k))$ and the last line follows from the fact that $\mathbb{I}_{(x\geq\tilde{\varepsilon})}(1-e^{\tilde{\varepsilon}-x}) \leq e^{\lambda(x-\tilde{\varepsilon})+\lambda\log\lambda-(\lambda+1)\log(\lambda+1)}$ for any $\lambda \geq 0$.

Next, we give an upper bounds on the moment generating function of Z. From the definition of $\tilde{P}_0^{(\ell)}$ and $\tilde{P}_1^{(\ell)}$, $\mathbb{E}[e^{\lambda Z}] = \Big(\mathbb{E}_{\tilde{P}_0^{(\ell)}}[e^{\lambda\log(\tilde{P}_0^{(\ell)}(\tilde{X}_0^{(\ell)})/\tilde{P}_1^{(\ell)}(\tilde{X}_0^{(\ell)}))}]\Big)^k$. Let $\tilde{\varepsilon} = \sum_{\ell=1}^k(e^{\varepsilon_\ell}-1)\varepsilon_\ell/(e^{\varepsilon_\ell}+1) + \sqrt{2\sum_{\ell=1}^k\varepsilon_\ell^2\log\big(e + (\sqrt{\sum_{\ell=1}^k\varepsilon_\ell^2}/\tilde{\delta})\big)}$. Next we show that the k-fold composition is $(\tilde{\varepsilon}, 1-(1-\tilde{\delta})\prod_{\ell\in[k]}(1-\delta_\ell)\,)$-differentially private.

$$\mathbb{E}_{\tilde{P}_0^{(\ell)}}\big[e^{\lambda\log(P_0^{(\ell)}(X)/P_1^{(\ell)}(X))}\big] \;\leq\; e^{\frac{e^{\varepsilon_\ell}-1}{e^{\varepsilon_\ell}+1}\lambda\varepsilon_\ell+\frac{1}{2}\lambda^2{\varepsilon_\ell}^2}\;,$$

for any λ. Substituting this into (A.33) with a choice of

$$\lambda = \frac{\tilde{\varepsilon} - \sum_{\ell \in [k]} \varepsilon_\ell (e^{\varepsilon_\ell} - 1)/(e^{\varepsilon_\ell} + 1)}{\sum_{\ell \in [k]} \varepsilon_\ell^2},$$

we get

$$\begin{aligned} &\tilde{P}_0^k(B_1) - e^{\tilde{\varepsilon}} \tilde{P}_1^k(B_1) \\ &\leq \frac{1}{1 + \frac{\tilde{\varepsilon} - \sum_{\ell \in [k]} \varepsilon_\ell (e^{\varepsilon_\ell} - 1)/(e^{\varepsilon_\ell} + 1)}{\sum_{\ell \in [k]} \varepsilon_\ell^2}} \exp\Big\{ - \frac{1}{2 \sum_{\ell \in [k]} \varepsilon_\ell^2} \Big(\tilde{\varepsilon} - \sum_{\ell \in [k]} \varepsilon_\ell \frac{e^{\varepsilon_\ell} - 1}{e^{\varepsilon_\ell} + 1} \Big)^2 \Big\} \,. \end{aligned}$$

Substituting $\tilde{\varepsilon}$, we get the desired bound.

Similarly, we can prove that with

$$\tilde{\varepsilon} = \sum_{\ell=1}^{k} (e^{\varepsilon_\ell} - 1)\varepsilon_\ell / (e^{\varepsilon_\ell} + 1) + \sqrt{2 \sum_{\ell=1}^{k} \varepsilon_\ell^2 \log\big(1/\tilde{\delta}\big)},$$

the desired bound also holds.

APPENDIX B

PROOFS FOR LOCAL DIFFERENTIAL PRIVACY

B.1 Operational Interpretation of Differential Privacy

B.2 Proof of Theorem 3.2.1

We start by proving the following equivalent definition for local differential privacy.

Claim 1 *A conditional distribution Q is said to be ε-locally differentially private if and only if for all $A, B \subset \mathcal{X}$, such that $A \cap B = \emptyset$ and all $S \subset \mathcal{Y}$, we have that*

$$Q\left(S|A\right) \leq e^{\varepsilon} Q\left(S|B\right), \tag{B.1}$$

where $Q\left(S|A\right) = \mathbb{P}(Y \in S|X \in A)$ and $\varepsilon \in [0, \infty)$.

Proof 8 *To see that Claim 1 implies local differential privacy, set $A = \{x\}$ and $B = \{x'\}$ for any $x \neq x'$. Observe that $Q\left(S|x\right) \leq e^{\varepsilon} Q\left(S|x'\right)$ holds trivially for $x = x'$. We now show that local differential privacy implies Claim 1. First, observe that*

$$Q\left(S|A\right) = \mathbb{P}(Y \in S|X \in A) = \frac{\sum_{x \in A} Q\left(S|x\right) \mathbb{P}\left(X = x\right)}{\mathbb{P}\left(X \in A\right)}.$$

Therefore,

$$\begin{aligned} \frac{Q\left(S|A\right)}{Q\left(S|B\right)} &= \frac{\mathbb{P}\left(X \in B\right) \sum_{x \in A} Q\left(S|x\right) \mathbb{P}\left(X = x\right)}{\mathbb{P}\left(X \in A\right) \sum_{x \in B} Q\left(S|x\right) \mathbb{P}\left(X = x\right)} \\ &\leq \frac{\max_{x \in A} Q\left(S|x\right)}{\min_{x \in B} Q\left(y|x\right)} \\ &\leq e^{\varepsilon}, \end{aligned}$$

where the first inequality follows from the fact that $\sum_{x\in A} Q(y|x)\,\mathbb{P}(X=x) \leq \max_{x\in A} Q(y|x)\,\mathbb{P}(X\in A)$, $\sum_{x\in B} Q(y|x)\,\mathbb{P}(X=x) \geq \min_{x\in B} Q(y|x)\,\mathbb{P}(X\in B)$, *and the second inequality follows from Claim 1. This finishes the proof.*

Assume that Q is ε-locally differentially private and fix any sets $A, B \subset \mathcal{X}$ such that $A \cap B = \emptyset$. Without loss of generality[1], consider the set of all deterministic decision rules $\hat{X} : Y \to \{A, B\}$. These rules can be described by (a) partitioning the output alphabet $\mathcal{Y}$ into (S, S^c) for some $S \subset \mathcal{Y}$, and (b) assigning $\hat{X} = A$ whenever $Y \in S$ and $\hat{X} = B$ whenever $Y \in S^c$. In this case,

$$\begin{aligned} P_{\mathrm{FA}} = \mathbb{P}(\hat{X} = A|X \in B) = Q(Y \in S|X \in B) \geq e^{-\varepsilon} Q(Y \in S|X \in A) \\ P_{\mathrm{MD}} = \mathbb{P}(\hat{X} = B|X \in A) = Q(Y \in S^c|X \in A) \geq e^{-\varepsilon} Q(Y \in S^c|X \in B), \end{aligned}$$

where both inequalities follow from Claim 1. Replacing $Q(Y \in S|X \in A)$ by $1 - P_{\mathrm{MD}}$ and $Q(Y \in S^c|X \in B)$ by $1 - P_{\mathrm{FA}}$, we get that

$$\begin{aligned} P_{\mathrm{FA}} + e^{\varepsilon} P_{\mathrm{MD}} &\geq 1 \\ e^{\varepsilon} P_{\mathrm{FA}} + P_{\mathrm{MD}} &\geq 1. \end{aligned} \tag{B.2}$$

This proves that local differential privacy implies Theorem 3.2.1. The converse can be shown similarly.

B.3 Optimal Mechanisms for Differential Privacy

We start the proof with a few definitions, a lemma, and a general result that applies to any convex utility function that obeys a mild assumption.

Recall that for an input alphabet $\mathcal{X}$ with $|\mathcal{X}| = k$, we represent the set of ε-locally differentially private mechanisms that lead to output alphabets $\mathcal{Y}$ with $|\mathcal{Y}| = \ell$ by $\mathcal{D}_{\varepsilon,\ell}$. The set of all ε-locally differentially private mechanisms is given by $\mathcal{D}_\varepsilon = \cup_{\ell\in\mathbb{N}} \mathcal{D}_{\varepsilon,\ell}$. A utility function $U(Q)$ is convex in Q if $U\left(\lambda Q^{(1)} + (1-\lambda) Q^{(2)}\right) \leq \lambda U\left(Q^{(1)}\right) + (1-\lambda) U\left(Q^{(2)}\right)$ for any $\lambda \in (0,1)$. Convex utility functions are ubiquitous in information theory and statistics.

[1] Randomized rules can never achieve a $(P_{\mathrm{FA}}, P_{\mathrm{MD}})$ pair outside the convex hull of $(P_{\mathrm{FA}}, P_{\mathrm{MD}})$ pairs achieved by deterministic rules.

Assumption 1 *If a $k \times \ell$ privatization mechanism $Q^{(1)} \in \mathcal{D}_{\varepsilon,\ell}$ is obtained by deleting an all-zero column of a $k \times \ell+1$ privatization mechanism $Q^{(2)} \in \mathcal{D}_{\varepsilon,\ell+1}$, then $U\left(Q^{(1)}\right) = U\left(Q^{(2)}\right)$.*

Naturally, one would expect that if we delete the zero columns of a privatization mechanism $Q^{(2)}$ to obtain a new privatization mechanism $Q^{(1)}$, we would still get the same utility. This is because a "reasonable" utility function should not depend on output alphabets with zero probability.

Theorem B.3.1 *If $U\left(Q\right)$ is a convex utility function that satisfies Assumption 1, then the following holds*

$$\max_{Q\in\mathcal{D}_\varepsilon} U\left(Q\right) = \max_{Q\in\cup_{\ell=1}^{k}\mathcal{D}_{\varepsilon,\ell}} U\left(Q\right). \tag{B.3}$$

This theorem implies that among all ε-locally differentially private mechanisms, we only need to consider those that lead to output alphabets of size $\ell \leq k$. In other words, enlarging the input alphabet cannot further maximize the utility. The proof of Theorem B.3.1 is given in Section B.3.1.

Lemma B.3.2 *A $k \times \ell$ conditional distribution Q is ε-locally differentially private if and only if it can be written as $Q = S\Theta$, where S is a $k \times \ell$ matrix with $S_{ij} \in [1, e^\varepsilon]$ and $\Theta = diag\left(\theta_1, \ldots, \theta_\ell\right)$ with its diagonal entries in $\mathbb{R}_+$.*

The proof of Lemma B.3.2 is provided in Section B.3.2. With the above results, we are now ready to prove Theorems 3.3.2 and 3.3.4. By Lemma B.3.2, for any $Q \in \mathcal{D}_{\varepsilon,\ell}$ we have that $Q_j = \theta_j S_j$. Suppose $U\left(Q\right) = \sum_{j\in[\ell]} \mu(Q_j)$, where μ is a sublinear function. Since μ is sublinear, it is convex and $\mu\left(\theta_j S_j\right) = \theta_j \mu\left(S_j\right)$. $U\left(Q\right)$ is convex in Q because

$$\begin{aligned} U\left(\lambda Q^{(1)} + \left(1-\lambda\right) Q^{(2)}\right) &= \sum_{j\in[\ell]} \mu\left(\lambda \theta_j^{(1)} S_j^{(1)} + \left(1-\lambda\right) \theta_j^{(2)} S_j^{(2)}\right) \\ &\leq \sum_{j\in[\ell]} \lambda \mu\left(\theta_j^{(1)} S_j^{(1)}\right) + \left(1-\lambda\right) \mu\left(\theta_j^{(2)} S_j^{(2)}\right) \\ &= \lambda U\left(Q^{(1)}\right) + \left(1-\lambda\right) U\left(Q^{(2)}\right), \end{aligned}$$

for any $\lambda \in (0,1)$. Furthermore, $U\left(Q\right)$ satisfies Assumption 1 because $\mu\left(Q_j\right) = 0$ whenever $\theta_j = 0$. Let $Q^* = S^*\Theta^* \in \arg\max_{Q\in\cup_{\ell=1}^{k}\mathcal{D}_{\varepsilon,\ell}} U\left(Q\right)$ and note that by Theorem B.3.1, $U\left(Q^*\right) = \max_{Q\in\mathcal{D}_\varepsilon} U\left(Q\right)$. Suppose that

Q^* is of dimensions $k \times \ell$, where $\ell \leq k$. Each of the ℓ columns of Q^* can be expressed as a convex combination of the columns of $S^{(k)}$, the staircase pattern matrix. This is because the 2^k columns of $S^{(k)}$ are the corner points of the cube $[1, e^{\varepsilon}]^k$ and each $S_j^* \in [1, e^{\varepsilon}]^k$. Therefore, $S_j^* = \sum_{i=1}^{2^k} \lambda_{ij} S_i^{(k)}$, where $\lambda_{ij} \geq 0$ for all i and j, and $\sum_{i=1}^{2^k} \lambda_{ij} = 1$ for all j. Create the 2^k-dimensional vector $\tilde{\theta}$ such that $\tilde{\theta}_i = \sum_{j=1}^{\ell} \lambda_{ij} \theta_j^*$ and let $\tilde{Q} = S^{(k)} \tilde{\Theta}$.

$$\begin{aligned}
U\left(Q^*\right) - U(\tilde{Q}) &= \sum_{j=1}^{\ell} \mu\left(S_j^*\right) \theta_j^* - \sum_{i=1}^{2^k} \mu\left(\left(\sum_{j=1}^{\ell} \lambda_{ij} \theta_j^*\right) S_j^{(k)}\right) \\
&= \sum_{j=1}^{\ell} \mu\left(\sum_{i=1}^{2^k} \lambda_{ij} S_i^{(k)}\right) \theta_j^* - \sum_{i=1}^{2^k} \sum_{j=1}^{\ell} \lambda_{ij} \theta_j^* \mu\left(S_j^{(k)}\right) \\
&= \sum_{j=1}^{\ell} \theta_j^* \left\{ \mu\left(\sum_{i=1}^{2^k} \lambda_{ij} S_i^{(k)}\right) - \sum_{i=1}^{2^k} \lambda_{ij} \mu\left(S_j^{(k)}\right) \right\} \\
&\leq 0,
\end{aligned}$$

by the convexity of $\mu(z)$ and the non-negativity of θ_j^*'s. Moreover, observe that since $S^{(k)} \tilde{\theta} = \mathbb{1}$, $\tilde{\theta}$ is a valid choice for the linear program of (3.12). This implies that

$$\max_{S^{(k)}\theta = \mathbb{1}, \theta \geq 0} \sum_{j=1}^{2^k} \mu\left(S_j^{(k)}\right) \theta_j \geq U(\tilde{Q}) \geq U\left(Q^*\right) = \max_{Q \in \mathcal{D}_{\varepsilon}} U(Q) \qquad \text{(B.4)}$$

On the other hand, for any $\tilde{Q} = S^{(k)} \tilde{\Theta}$, where $\tilde{\theta}$ is valid for the linear program of (3.12), we have that $\tilde{Q} \in \mathcal{D}_{\varepsilon, 2^k} \subset \mathcal{D}_{\varepsilon}$ and therefore,

$$\max_{S^{(k)}\theta = \mathbb{1}, \theta \geq 0} \sum_{j=1}^{2^k} \mu\left(S_j^{(k)}\right) \theta_j \leq \max_{Q \in \mathcal{D}} U(Q).$$

Thus, $\max_{S^{(k)}\theta = \mathbb{1}, \theta \geq 0} \sum_{j=1}^{2^k} \mu\left(S_j^{(k)}\right) \theta_j = \max_{Q \in \mathcal{D}} U(Q)$. This proves Theorem 3.3.4.

The polytope given by $S^{(k)}\theta = \mathbb{1}$ and $\theta \geq 0$ is a closed and bounded one. Thus, the linear program of (3.12) is bounded and has a solution, say θ^*, at one of the corner points of the polytope. Since there are k equality constraints given by $S^{(k)}\theta = \mathbb{1}$ and 2^k inequality constraints given by $\theta \geq 0$, any corner point, including θ^*, cannot have more than k non-zero entries. Form $\tilde{S}$ by deleting the columns of $S^{(k)}$ corresponding to zero entries of θ^*. Similarly, form $\tilde{\theta}$ by deleting the zero entries of θ^* and let $\tilde{Q} = \tilde{S}\tilde{\Theta}$, where $\tilde{\Theta} = \text{diag}\tilde{\theta}$. It is easy to verify that $U(\tilde{Q}) = U(Q^*) = \mu^T\theta^*$; hence, $\tilde{Q}$ solves linear program of (3.12). Moreover, $\tilde{Q}$ has at most k columns and $\tilde{S}_{ij} = \{1, e^{\varepsilon}\}$. Therefore, $\tilde{Q}$ is a staircase mechanism of dimension $k \times \ell$, where $\ell \leq k$.

B.3.1 Proof of Theorem B.3.1

We start the proof of Theorem B.3.1 with an important lemma the proof of which is presented in Section B.3.3.

Lemma B.3.3 *The set of all $k \times \ell$, ε-locally differentially private mechanisms $\mathcal{D}_{\varepsilon,\ell}$ forms a closed and bounded polytope in $\mathbb{R}_+^{k\ell}$. Moreover, let Q be a corner point of the polytope formed by $\mathcal{D}_{\varepsilon,\ell}$, then Q has at most k non-zero columns.*

Fix an $\ell > k$. Since $U(Q)$ is convex in Q, it suffices to consider the corner points of $\mathcal{D}_{\varepsilon,\ell}$ when maximizing $U(Q)$ subject to $Q \in \mathcal{D}_{\varepsilon,\ell}$. By Lemma B.3.3, any $Q^{(1)}$, a $k \times \ell$ corner point of $\mathcal{D}_{\varepsilon,\ell}$, has at most k non-zero columns. Therefore, the privatization mechanism $Q^{(2)}$, obtained by deleting the all-zero columns of $Q^{(1)}$, has at most k columns. Notice that $Q^{(2)} \in \cup_{i=1}^{k}\mathcal{D}_{\varepsilon,i}$. Since $U(Q)$ satisfies Assumption 1, we have that $U\left(Q^{(1)}\right) = U\left(Q^{(2)}\right)$ and therefore, it suffices to consider $Q \in \cup_{i=1}^{k}\mathcal{D}_{\varepsilon,i}$ when maximizing $U(Q)$ subject to $Q \in \mathcal{D}_{\varepsilon,\ell}$. Thus,

$$\begin{aligned}
\sup_{Q\in\mathcal{D}_{\varepsilon}} U(Q) &= \sup_{\ell\in\mathbb{N}}\left\{\max_{Q\in\mathcal{D}_{\varepsilon,\ell}} U(Q)\right\} \\
&= \sup_{\ell\in\mathbb{N}}\left\{\max_{Q\in\cup_{i=1}^{k}\mathcal{D}_{\varepsilon,i}} U(Q)\right\} \\
&= \max_{Q\in\cup_{i=1}^{k}\mathcal{D}_{\varepsilon,i}} U(Q),
\end{aligned}$$

which finishes the proof.

B.3.2 Proof of Lemma B.3.2

Claim 2 *Let $Q \in \mathcal{D}_{\varepsilon,\ell}$. If $Q_{ij} = 0$ for some $j \in \{1, ..., \ell\}$ then $Q_{ij} = 0$ for all $i \in \{1, ..., k\}$.*

Proof 9 *Assume that $Q_{i_1 j} = 0$ and $Q_{i_2 j} \neq 0$ for some $i_1, i_2 \in \{1, ..., k\}$. It is obvious that $q\left(y_j | x_{i_2}\right) \leq q\left(y_j | x_{i_1}\right) e^{\varepsilon}$ is not satisfied. Therefore, Q is not a locally differentially private mechanism.*

It is easy to check that any $k \times \ell$ stochastic matrix $Q = S\Theta$, where Θ is a diagonal matrix with non-negative entries and S is a $k \times \ell$ matrix with $S_{ij} \in [1, e^{\varepsilon}]$, satisfies the local differential privacy constraints. Thus, $Q \in \mathcal{D}_{\varepsilon,\ell}$. On the other hand, assume that $Q \in \mathcal{D}_{\varepsilon,\ell}$. If $Q_{ij} = 0$ for some j then by Claim 2 we have that $Q_{ij} = 0$ for all i and therefore, we can set $\theta_j = 0$ and $S_{ij} = 1$ for all i. If $Q_{ij} > 0$ then by Claim 2 we have that $Q_{ij} > 0$ for all i. In this case, let $\theta_j = \min_i Q_{ij}$ and observe that $\theta_j > 0$ since $Q_{ij} > 0$ for all i. Let $S_{ij} = Q_{ij}/\theta_i$, then it is clear (from the definition of θ_i) that $S_{ij} \geq 1$. On the other hand, from the differential privacy constraints, we have that $Q_{ij} \leq Q_{kj} e^{\varepsilon}$ for all k and thus, $Q_{ij} \leq \min_k Q_{kj} e^{\varepsilon}$ which proves that $S_{ij} = Q_{ij}/\min_k Q_{kj} \leq e^{\varepsilon}$. This establishes that any $Q \in \mathcal{D}_{\varepsilon,\ell}$ can be written as $Q = S\Theta$, where Θ is a diagonal matrix with non-negative entries and S is a $k \times \ell$ matrix with $S_{ij} \in [1, e^{\varepsilon}]$.

B.3.3 Proof of Lemma B.3.3

We start by showing that $\mathcal{D}_{\varepsilon,\ell}$ forms a closed and bounded polytope in $\mathbb{R}_+^{k\ell}$. We are interested in studying the corner points of the polytope formed by $\mathcal{D}_{\varepsilon,\ell}$ because convex utility functions are maximized at one of these corner points whenever the space of privatization mechanisms is restricted to $\mathcal{D}_{\varepsilon,\ell}$.

Claim 3 *A privatization mechanism $Q \in \mathcal{D}_{\varepsilon,\ell}$ if and only if for all $x, x' \in \mathcal{X}$ and all $y \in \mathcal{Y}$ we have that $Q\left(y|x\right) \leq Q\left(y|x'\right) e^{\varepsilon}$.*

Proof 10 *By definition, Q is differentially private if for all x, $x' \in \mathcal{X}$ and all $B \subseteq \mathcal{Y}$ we have that $Q\left(B|x\right) \leq Q\left(B|x'\right) e^{\varepsilon}$. By choosing $B = \{y\}$ for some $y \in \mathcal{Y}$ the first direction of the above lemma is proven. In order to prove the other direction, assume that for all $x, x' \in \mathcal{X}$ and all $y \in \mathcal{Y}$ we have that $Q\left(y|x\right) \leq Q\left(y|x'\right) e^{\varepsilon}$. Then for any $B \subseteq \mathcal{Y}$, the following holds:*

$$\begin{aligned} \sum_{y \in B} Q\left(y|x\right) &\leq \sum_{y \in B} Q\left(y|x'\right) e^{\varepsilon} \\ \Leftrightarrow \quad Q\left(B|x\right) &\leq Q\left(B|x'\right) e^{\varepsilon}. \end{aligned}$$

Letting $Q \in \mathcal{D}_{\varepsilon,\ell}$, then by Claim 3, it is easy to see that Q must satisfy $\ell k(k-1)$ inequalities of the form $Q\left(y|x\right) \leq Q\left(y|x'\right)e^{\varepsilon}$. These inequalities can be compactly represented by

$$\tilde{A}q \leq 0, \tag{B.5}$$

where $q = \left[Q\left(y_1|x_1\right), ..., Q\left(y_1|x_k\right), ..., Q\left(y_\ell|x_1\right), ..., Q\left(y_\ell|x_k\right)\right]^T$ and $\tilde{A}$ is a $\ell k(k-1) \times k\ell$ matrix that contains all the local differential privacy linear constraints. Observe that there is a one-to-one mapping between Q and q. Here is an example for the case when $k = \ell = 2$

$$\underbrace{\begin{bmatrix} 1 & -e^{\varepsilon} & 0 & 0 \\ -e^{\varepsilon} & 1 & 0 & 0 \\ 0 & 0 & 1 & -e^{\varepsilon} \\ 0 & 0 & -e^{\varepsilon} & 1 \end{bmatrix}}_{\tilde{A}} \begin{bmatrix} Q\left(y_1|x_1\right) \\ Q\left(y_1|x_2\right) \\ Q\left(y_2|x_1\right) \\ Q\left(y_2|x_2\right) \end{bmatrix} \leq 0. \tag{B.6}$$

Moreover, since Q is a row stochastic matrix (a conditional distribution) it satisfies $Q\mathbb{1} = \mathbb{1}$, where $\mathbb{1}$ represents the all ones vector of appropriate dimensions. This condition can be rewritten as

$$Bq = \mathbb{1}, \tag{B.7}$$

where B is a $k \times k\ell$ binary matrix. For the case when $k = \ell = 2$, we have that

$$\underbrace{\begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}}_{B} \begin{bmatrix} Q\left(y_1|x_1\right) \\ Q\left(y_1|x_2\right) \\ Q\left(y_2|x_1\right) \\ Q\left(y_2|x_2\right) \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}. \tag{B.8}$$

Finally, observe that $Q\left(y|x\right) \geq 0$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. These constraints can be incorporated as follows. Let $A = \left[\tilde{A}^T\ ,\ -I_{\ell k}\right]^T$, where $I_{\ell k}$ is the $\ell k \times \ell k$ identity matrix, then $Aq \leq 0$. To summarize, $Q \in \mathcal{D}_{\varepsilon,\ell}$ if and only if

$$\begin{aligned} Aq &\leq 0 \\ Bq &= \mathbb{1}. \end{aligned} \tag{B.9}$$

Therefore, the set of all $k \times \ell$, ε-locally differentially private mechanisms $\mathcal{D}_{\varepsilon,\ell}$ forms a convex polytope in $\mathbb{R}_{+}^{k\ell}$.

We now proceed to proving that if Q is a corner point of the polytope formed by $\mathcal{D}_{\varepsilon,\ell}$, then Q has at most k non-zero columns. This claim is obvious for all $k \times \ell$ privatization mechanisms with $\ell \leq k$. Therefore, we restrict our attention to the case where $\ell > k$. Let A_j be the matrix including all the inequality constraints imposed on the j^{th} column of Q. Observe that the rows of A_j form a subset of the rows of A, defined in (B.9), and recall that there are $k(k-1)$ differential privacy and k non-negativity inequality constraints imposed on the j^{th} column of Q. Therefore, A_j is a $k^2 \times k$ matrix and we have that $A_j Q_j \leq 0$, where Q_j represents the j^{th} column of Q. By Claim 2, we know that Q_j is either equal to zero or contains non-zero entries.

Claim 4 *In what follows, the term linearly independent inequality constraints refers to linear independent rows of A_j.*

- *If $Q_j = 0$, then k linearly independent inequality constraints are achieved with equality.*
- *If $Q_j \neq 0$, then at most $k-1$ linearly independent inequality constraints can be achieved with equality.*

Proof 11 *In fact, the number of linearly independent inequality constraints (achieved or not) cannot exceed k because A_j has a rank less than or equal to k. If $Q_j = 0$, then the k non-negativity inequality constraints are achieved with equality and it is easy to see that they are all linearly independent (in fact, they form an orthonormal basis to $\mathbb{R}^k$). This proves the first part of the claim. We now establish the second part of the claim by showing that if $Q_j \neq 0$, we cannot have k linearly independent inequality constraints achieved with equality. Assume that $Q_j \neq 0$ and let $\tilde{A}_j$ be the matrix including the largest collection of linearly independent rows of A_j corresponding to the inequality constraints that are achieved with equality. In other words, $\tilde{A}_j Q_j = 0$. If $\tilde{A}_j$ contains k rows, then its rank is equal to k. However, this implies that $Q_j = 0$, a contradiction. Therefore, at most $k-1$ linearly independent inequality constraints can be achieved with equality when $Q_j \neq 0$.*

Suppose that Q is a corner point of $\mathcal{D}_{\varepsilon,\ell}$ and out of its ℓ columns, $\ell_{>0}$ are non-zero and $\ell_{=0}$ ($\ell_{=0} = \ell - \ell_{>0}$) are zero. Moreover, assume that the number of non-zero columns of Q is larger than k (i.e., $\ell_{>0} > k$). In this case, from Claim 4, we can see that Q achieves at most $\ell_{>0}(k-1) + (\ell - \ell_{>0})k$ linearly independent inequality constraints with equality. Furthermore, at most k additional linearly independent equality constraints (linearly independent rows of the matrix B defined in (B.9)) can be met by Q. Therefore, the total number of linearly independent constraints that Q achieves with equality is at most $\ell_{>0}(k-1) + (\ell - \ell_{>0})k + k = -\ell_{>0} + (\ell+1)k < \ell k$, where the last strict inequality follows from the fact that $\ell_{>0} > k$. This implies that Q cannot be a corner point of $\mathcal{D}_{\varepsilon,\ell}$. Therefore, any corner point of $\mathcal{D}_{\varepsilon,\ell}$ must have at most k non-zero columns.

B.4 Private Hypothesis Testing

B.4.1 Proof of Theorem 3.4.1

Let $T = \{x : P_0(x) \geq P_1(x)\}$. In other words,

$$P_0(T) - P_1(T) = \max_{A \subseteq \mathcal{X}} P_0(A) - P_1(A).$$

Recall that for a given P_0 and P_1, the binary mechanism is defined as a staircase mechanism with only two outputs $y \in \{0,1\}$ satisfying

$$\begin{aligned} Q(0|x) &= \begin{cases} \frac{e^\varepsilon}{1+e^\varepsilon} & \text{if } P_0(x) \geq P_1(x) \ , \\ \frac{1}{1+e^\varepsilon} & \text{if } P_0(x) < P_1(x) \ . \end{cases} \\ Q(1|x) &= \begin{cases} \frac{e^\varepsilon}{1+e^\varepsilon} & \text{if } P_0(x) < P_1(x) \ , \\ \frac{1}{1+e^\varepsilon} & \text{if } P_0(x) \geq P_1(x) \ . \end{cases} \end{aligned} \tag{B.10}$$

Lemma B.4.1 *For any pair of distributions P_0 and P_1, there exists a positive ε^* that depends on P_0 and P_1 such that for all $y \in \mathcal{Y}$, all $\ell \in \mathbb{N}$, and all $Q \in \mathcal{D}_{\varepsilon,\ell}$ with $\varepsilon \leq \varepsilon^*$, we have that*

$$\frac{(e^\varepsilon - 1)\, P_0\,(T^c) + 1}{(e^\varepsilon - 1)\, P_1\,(T^c) + 1} \leq \frac{M_0(y)}{M_1(y)} \leq \frac{(e^\varepsilon - 1)\, P_0\,(T) + 1}{(e^\varepsilon - 1)\, P_1\,(T) + 1}. \tag{B.11}$$

Moreover, the above upper and lower bounds are achieved by the binary mechanism given in (B.10).

Observe that because $P_0(T) \geq P_1(T)$ and $P_0(T^c) \leq P_1(T^c)$, the direction of the above inequalities makes sense.

Let $\tilde{M}_\nu$ be the induced marginal for the binary mechanism when P_ν is the original distribution. Following the analysis techniques developed in [43], we define hypothesis testing region $R(\tilde{M}_0, \tilde{M}_1)$ as the convex hull of all achievable probabilities of missed detection and false alarm, when testing whether $\nu = 0$ or $\nu = 1$ based on Y_{bin} distributed as $\tilde{M}_\nu$:

$$R(\tilde{M}_0, \tilde{M}_1) \equiv \text{conv}\Big(\big\{ (\tilde{M}_1(S), \tilde{M}_0(S^c)) \,:\, \forall S \subseteq \mathcal{Y} \big\} \Big) \,,$$

where $S \in \mathcal{Y}$ is the accept region for hypothesis $\nu = 0$. For the binary mechanism, this ends up being a very simple triangular region. The slopes defining the two sides of the triangular region are: $-\max_S \tilde{M}_0(S)/\tilde{M}_1(S) = -((e^\varepsilon - 1)P_0(T) + 1)/((e^\varepsilon - 1)P_1(T) + 1)$ and $-\min_S \tilde{M}_0(S^c)/\tilde{M}_1(S^c) = -((e^\varepsilon - 1)P_0(T^c) + 1)/((e^\varepsilon - 1)P_1(T^c) + 1)$.

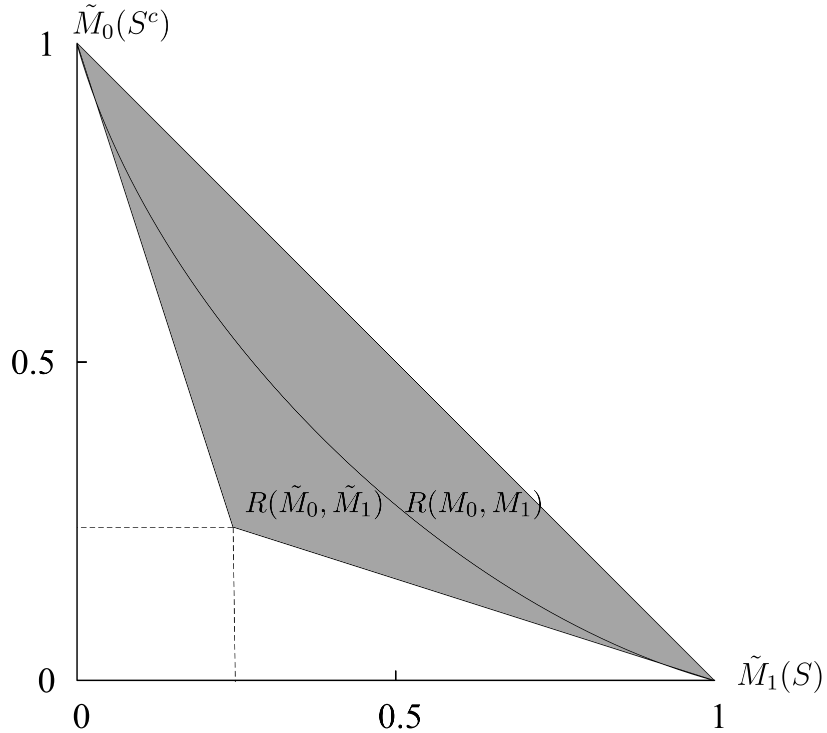


Figure B.1: Hypothesis testing regions for two mechanisms.

For any other mechanism satisfying the ε-local differential privacy for $\varepsilon \leq \varepsilon^*$, the above lemma implies that for any choice of the rejection region S, the slopes satisfy $-M_0(S)/M_1(S) \geq -((e^\varepsilon - 1)P_0(T) + 1)/((e^\varepsilon - 1)P_1(T) + 1)$ and $-M_0(S^c)/M_1(S^c) \leq -((e^\varepsilon - 1)P_0(T^c) + 1)/((e^\varepsilon - 1)P_1(T^c) + 1)$. In the hypothesis testing region, this implies that

$$R(M_0, M_1) \subseteq R(\tilde{M}_0, \tilde{M}_1) \; ,$$

as in the following Figure B.1.

From Theorem 2.5 of [43], we know that this implies a certain Markov property. Precisely, let $Y_{\rm bin}$ denote the output of the binary mechanism, and $Y_{\rm dp}$ denote the output of any ε-local differentially private mechanism. Then, it follows that there exists a coupling of $Y_{\rm bin}$ and $Y_{\rm dp}$ such that they form a Markov chain: ν–$Y_{\rm bin}$–$Y_{\rm dp}$, where ν is the hypothesis on P_ν whether the data was generated from $\nu = 0$ or $\nu = 1$. Then, it follows from the data processing inequality of f-divergences that

$$D_f(\tilde{M}_0, \tilde{M}_1) \geq D_f(M_0, M_1) \; .$$

It follows that there is no algorithm with larger f-divergence than the binary mechanism.

B.4.2 Proof of Lemma B.4.1

We start by showing that the binary mechanism achieves the upper and lower bounds presented in the statement of the lemma. Let M_0^B and M_1^B denote the induced marginals under the binary mechanism given in (B.10). For $\nu \in \{0, 1\}$, we have that

$$\begin{aligned} M_\nu^B(0) &= \sum_{x \in \mathcal{X}} P_0(x)\, Q(0|x) = \frac{1}{e^\varepsilon + 1}\left((e^\varepsilon - 1)\, P_\nu(T) + 1\right) \\ M_\nu^B(1) &= \sum_{x \in \mathcal{X}} P_0(x)\, Q(1|x) = \frac{1}{e^\varepsilon + 1}\left((e^\varepsilon - 1)\, P_\nu(T^c) + 1\right). \end{aligned}$$

Computing $M_0^B(0)/M_1^B(0)$ and $M_0^B(1)/M_1^B(1)$ we see that the binary mechanism achieves the upper and lower bounds for all values of ε.

As in Lemma B.3.2, for any $\ell \in \mathbb{N}$, $Q \in \mathcal{D}_{\varepsilon,\ell}$ can be represented as $Q = S\Theta$, where $S \in [1, e^\varepsilon]^{k\times\ell}$ and $\Theta = \text{diag}(\theta_1, ..., \theta_\ell)$ with its diagonal entries in $\mathbb{R}_+$. We now show that for any $\ell \in \mathbb{N}$ and any $Q \in \mathcal{D}_{\varepsilon,\ell}$, the following upper bound holds:

$$\frac{M_0(y)}{M_1(y)} = \frac{\sum_{i\in[k]} P_0\left(x_i\right) S_{ij}}{\sum_{i\in[k]} P_1\left(x_i\right) S_{ij}} \leq \frac{\left(e^{\varepsilon}-1\right) P_0\left(T\right)+1}{\left(e^{\varepsilon}-1\right) P_1\left(T\right)+1}, \tag{B.12}$$

for all $y \in \mathcal{Y}$ and sufficiently small ε. The above expression can be alternatively written as

$$\begin{aligned}&\left(e^{\varepsilon}-1\right)\left(P_0\left(T\right)-P_1\left(T\right)\right)\\ &\quad+\left(e^{\varepsilon}-1\right)\sum_{i\in[k]}\left(S_{ij}-1\right)\left(P_0\left(T\right)P_1\left(x_i\right)-P_1\left(T\right)P_0\left(x_i\right)\right)\\ &\quad-\sum_{i\in[k]}\left(S_{ij}-1\right)\left(P_0\left(x_i\right)-P_1\left(x_i\right)\right)\geq 0,\end{aligned} \tag{B.13}$$

where $S_j \in [1, e^{\varepsilon}]^k$. Equation (B.13) is linear in S_j and is therefore minimized (and maximized) at the corner points of $[1, e^{\varepsilon}]^{k\times\ell}$, a cube in $\mathbb{R}_+^{k\times\ell}$. The corner points of this cube are given by the staircase patterns: $S_j \in \{1, e^{\varepsilon}\}^k$. To begin with, let S_j be a staircase pattern with $T_j = \{x_i : S_{ij} = e^{\varepsilon}\} \neq T$. Then Equation (B.13) is equivalent to

$$\begin{aligned}&\left(e^{\varepsilon}-1\right)\left\{\left(P_0\left(T\right)-P_1\left(T\right)\right)-\left(P_0\left(T_j\right)-P_1\left(T_j\right)\right)\right\}\\ &\quad+\left(e^{\varepsilon}-1\right)^2\left\{P_0\left(T\right)P_1\left(T_j\right)-P_1\left(T\right)P_0\left(T_j\right)\right\}\geq 0.\end{aligned} \tag{B.14}$$

Using the fact that $P_0\left(T\right)-P_1\left(T\right) > P_0\left(T_j\right)-P_1\left(T_j\right)$ for all $T_j \neq T$, the inequality in (B.13) holds true for all ε whenever $P_0\left(T\right)P_1\left(T_j\right) \geq P_1\left(T\right)P_0\left(T_j\right)$. If $P_0\left(T\right)P_1\left(T_j\right) < P_1\left(T\right)P_0\left(T_j\right)$, then the inequality in (B.13) holds true for all $\varepsilon \leq \varepsilon(j)$, where

$$\varepsilon(j) = \log\left(\frac{\left(P_0\left(T\right)-P_1\left(T\right)\right)-\left(P_0\left(T_j\right)-P_1\left(T_j\right)\right)}{P_1\left(T\right)P_0\left(T_j\right)-P_0\left(T\right)P_1\left(T_j\right)}+1\right) > 0. \tag{B.15}$$

On the other hand, it is easy to verify that when $T_j = T$, we have that

$$\begin{aligned}&\left(e^{\varepsilon}-1\right)\left\{\left(P_0\left(T\right)-P_1\left(T\right)\right)-\left(P_0\left(T_j\right)-P_1\left(T_j\right)\right)\right.\\ &\qquad\qquad\left.+\left(e^{\varepsilon}-1\right)\left(P_0\left(T\right)P_1\left(T_j\right)-P_1\left(T\right)P_0\left(T_j\right)\right)\right\}=0,\end{aligned}$$

for all ε. In this case, set $\varepsilon(j) = 0$ and $\varepsilon_1 = \min_{j\in[2^k]} \varepsilon(j)$. Therefore, for any $\ell \in \mathbb{N}$ and any $Q \in \mathcal{D}_{\varepsilon,\ell}$, the upper bound in the statement of the lemma holds for all $\varepsilon \leq \varepsilon_1$.

We now show that for any $\ell \in \mathbb{N}$ and any $Q \in \mathcal{D}_{\varepsilon,\ell}$, the following lower bound holds

$$\frac{(e^\varepsilon - 1)\, P_0\left(T^c\right) + 1}{(e^\varepsilon - 1)\, P_1\left(T^c\right) + 1} \leq \frac{M_0(y)}{M_1(y)} = \frac{\sum_{i\in[k]} P_0\left(x_i\right) S_{ij}}{\sum_{i\in[k]} P_1\left(x_i\right) S_{ij}}, \tag{B.16}$$

for all $y \in \mathcal{Y}$ and sufficiently small ε. The above expression can be alternatively written as

$$\begin{aligned}(e^\varepsilon - 1)&\left(P_0\left(T\right) - P_1\left(T\right)\right)\\ &+ (e^\varepsilon - 1)\sum_{i\in[k]}\left(S_{ij} - 1\right)\left(P_0\left(T\right) P_1\left(x_i\right) - P_1\left(T\right) P_0\left(x_i\right)\right)\\ &+ e^\varepsilon \sum_{i\in[k]}\left(S_{ij} - 1\right)\left(P_0\left(x_i\right) - P_1\left(x_i\right)\right) \geq 0,\end{aligned} \tag{B.17}$$

where $S_j \in [1, e^\varepsilon]^k$. Equation (B.17) is linear in S_j and is therefore minimized at the corner points of $[1, e^\varepsilon]^k$, a cube in $\mathbb{R}_+^k$. The corner points of this cube are given by staircase patterns: $S_j \in \{1, e^\varepsilon\}^k$. To begin with, let S_j be a staircase pattern with $T_j = \{x_i : S_{ij} = e^\varepsilon\} \neq T^c$, then Equation (B.17) is equivalent to

$$\begin{aligned}(e^\varepsilon - 1)&\left\{\left(P_0\left(T\right) - P_1\left(T\right)\right) + e^\varepsilon\left(P_0\left(T_j\right) - P_1\left(T_j\right)\right)\right\}\\ &+ \left(e^\varepsilon - 1\right)^2\left\{P_0\left(T\right) P_1\left(T_j\right) - P_1\left(T\right) P_0\left(T_j\right)\right\} \geq 0.\end{aligned} \tag{B.18}$$

Using the fact that $P_0\left(T\right) - P_1\left(T\right) > P_1\left(T_j\right) - P_0\left(T_j\right)$ for all $T_j \neq T^c$, then for sufficiently small ε, Equation (B.17) can be written as

$$\varepsilon\left\{\left(P_0\left(T\right) - P_1\left(T\right)\right) - \left(P_1\left(T_j\right) - P_0\left(T_j\right)\right)\right\} + O\left(\varepsilon^2\right) > 0. \tag{B.19}$$

This proves that there exists a positive $\varepsilon(j)$ such that the left hand side of Equation (B.18) is positive for all $\varepsilon \leq \varepsilon(j)$. On the other hand, it is easy to verify that when $T_j = T^c$, we have that

$$\begin{aligned}(e^\varepsilon - 1)&\left\{\left(P_0\left(T\right) - P_1\left(T\right)\right) + e^\varepsilon\left(P_0\left(T_j\right) - P_1\left(T_j\right)\right)\right.\\ &\left.+ (e^\varepsilon - 1)\left(P_0\left(T\right) P_1\left(T_j\right) - P_1\left(T\right) P_0\left(T_j\right)\right)\right\} = 0,\end{aligned}$$

for all ε. In this case, let $\varepsilon(j) = 0$ and let $\varepsilon_2 = \min_{j\in[2^k]} \varepsilon(j)$. Therefore, for any $\ell \in \mathbb{N}$ and any $Q \in \mathcal{D}_{\varepsilon,\ell}$, the lower bound in the statement of the lemma holds for all $\varepsilon \leq \varepsilon_2$. To conclude, let $\varepsilon^* = \min(\varepsilon_1, \varepsilon_2)$. Then both, the upper and lower bounds, hold for all $\varepsilon \leq \varepsilon^*$.

B.4.3 Proof of Theorem 3.4.2

The total variation (TV) distance $\|M_0 - M_1\|_{\text{TV}}$ is a special case of f-divergence $D_f(M_0||M_1)$ with $f(x) = \frac{1}{2}|x-1|$. Therefore, by Theorem 3.3.4, we have that

$$\max_{Q\in\mathcal{D}_\varepsilon} \big\|M_0 - M_1\big\|_{\text{TV}} = \begin{array}{rr} \underset{\theta}{\text{maximize}} & \mu^T\theta \\ \text{subject to} & S^{(k)}\theta = \mathbb{1} \\ & \theta \geq 0, \end{array} \tag{B.20}$$

where $\mu_j = \mu\left(S_j^{(k)}\right) = \frac{1}{2}\big|\sum_{i\in[k]}\left(P_0(x_i) - P_1(x_i)\right) S_{ij}^{(k)}\big|$ for $j \in \{1, \ldots, 2^k\}$ and $S^{(k)}$ is the $k \times 2^k$ staircase pattern matrix given in Definition 3.3.3.

The polytope given by $S^{(k)}\theta = \mathbb{1}$ and $\theta \geq 0$ is a closed and bounded one. Thus, there is no duality gap and solving the above linear program is equivalent to solving its dual

$$\begin{array}{ll} \underset{\alpha}{\text{minimize}} & \mathbb{1}^T\alpha \\ \text{subject to} & {S^{(k)}}^T\alpha \geq \mu. \end{array} \tag{B.21}$$

Note that any satisfiable solution α^* to (B.21) provides an upper bound to (B.20) since $\max \mu^T\theta = \min \mathbb{1}^T\alpha \leq \mathbb{1}^T\alpha^*$. Let $T = \{x : P_0(x) \geq P_1(x)\}$ and $T_j = \{x_i : S_{ij}^{(k)} = e^\varepsilon\}$ for $j \in [2^k]$. Consider the following choice of dual variable

$$\alpha_i^* = \frac{1}{2}\frac{e^\varepsilon - 1}{e^\varepsilon + 1}\big|P_0(x_i) - P_1(x_i)\big|, \tag{B.22}$$

for $i \in [k]$. Observe that

$$\begin{aligned} \mathbb{1}^T\alpha^* &= \frac{1}{2}\frac{e^\varepsilon - 1}{e^\varepsilon + 1}\sum_{i\in[k]}\big|P_0(x_i) - P_1(x_i)\big| \\ &= \frac{1}{2}\frac{e^\varepsilon - 1}{e^\varepsilon + 1}\big\|P_0 - P_1\big\|_1 \\ &= \frac{e^\varepsilon - 1}{e^\varepsilon + 1}\big\|P_0 - P_1\big\|_{\text{TV}}. \end{aligned} \tag{B.23}$$

We claim that α^* is a feasible dual variable for all values of ε. In order to prove that α^* is a feasible dual variable, we show that ${S^{(k)}}_j^T \alpha^* - \mu_j \geq 0$ for all $j \in [2^k]$ and all ε. For all $j \in [2^k]$, we have that

$$\begin{aligned} g_j &= 2\left({S^{(k)}}_j^T \alpha^* - \mu_j\right) \\ &= \frac{e^\varepsilon - 1}{e^\varepsilon + 1} \sum_{i \in [k]} |P_0(x_i) - P_1(x_i)|\, S_{ij}^{(k)} - \left| \sum_{i \in [k]} (P_0(x_i) - P_1(x_i))\, S_{ij}^{(k)} \right| \\ &= \frac{e^\varepsilon - 1}{e^\varepsilon + 1} \left\{ \sum_{x_i \in T} (P_0(x_i) - P_1(x_i))\, S_{ij}^{(k)} + \sum_{x_i \in T^c} (P_1(x_i) - P_0(x_i))\, S_{ij}^{(k)} \right\} \\ &\quad - \left| \sum_{x_i \in T} (P_0(x_i) - P_1(x_i))\, S_{ij}^{(k)} - \sum_{x_i \in T^c} (P_1(x_i) - P_0(x_i))\, S_{ij}^{(k)} \right| \end{aligned} \tag{B.24}$$

Notice that we have arranged the equation such that all the summands are non-negative. Without loss of generality, we will assume that

$$\sum_{x_i \in T} (P_0(x_i) - P_1(x_i))\, S_{ij}^{(k)} \geq \sum_{x_i \in T^c} (P_1(x_i) - P_0(x_i))\, S_{ij}^{(k)} .$$

From the equality $\sum_{x_i \in T} (P_0(x_i) - P_1(x_i)) = \sum_{x_i \in T^c} (P_1(x_i) - P_0(x_i))$ and the fact that $S_{ij}^{(k)} \in \{1, e^\varepsilon\}$ for all i and j, it follows that

$$e^{-\varepsilon} \sum_{x_i \in T} (P_0(x_i) - P_1(x_i))\, S_{ij}^{(k)} \leq \sum_{x_i \in T^c} (P_1(x_i) - P_0(x_i))\, S_{ij}^{(k)} \ . \tag{B.25}$$

This is true because the right-hand side is minimized when the $S_{ij}^{(k)}$'s for $x_i \in T^c$ are all equal to 1 and the left-hand side is maximized when the $S_{ij}^{(k)}$'s for $x_i \in T$ are all equal to e^ε. Now, (B.24) can be written as

$$\begin{aligned} g_j = \frac{1}{e^\varepsilon + 1} &\left\{ -2 \sum_{x_i \in T} (P_0(x_i) - P_1(x_i))\, S_{ij}^{(k)} + 2e^\varepsilon \sum_{x_i \in T^c} (P_1(x_i) - P_0(x_i))\, S_{ij}^{(k)} \right\} \\ \geq 0 \ , & \end{aligned}$$

where the last inequality follows from (B.25).

This establishes the satisfiability of α^* for all ε which, in turn, shows that (B.23) is indeed an upper bound to the primal problem. It remains to show that this upper bound can be achieved via the binary mechanism. To this extent, recall that for a given P_0 and P_1, the binary mechanism is defined as a staircase mechanism with only two outputs $y \in \{0,1\}$ satisfying

$$\begin{aligned} Q(0|x) &= \begin{cases} \frac{e^{\varepsilon}}{1+e^{\varepsilon}} & \text{if } P_0(x) \geq P_1(x)\ , \\ \frac{1}{1+e^{\varepsilon}} & \text{if } P_0(x) < P_1(x)\ . \end{cases} \\ Q(1|x) &= \begin{cases} \frac{e^{\varepsilon}}{1+e^{\varepsilon}} & \text{if } P_0(x) < P_1(x)\ , \\ \frac{1}{1+e^{\varepsilon}} & \text{if } P_0(x) \geq P_1(x)\ . \end{cases} \end{aligned} \tag{B.26}$$

Computing the TV distance between M_0 and M_1 under (B.26), we get that

$$\|M_0 - M_1\|_{\mathrm{TV}} = \frac{e^{\varepsilon}-1}{e^{\varepsilon}+1}\|P_0 - P_1\|_{\mathrm{TV}}. \tag{B.27}$$

Hence, the binary mechanism in (B.26) achieves the upper bound in (B.23). This proves the optimality of the binary mechanism for all ε.

B.4.4 Proof of Theorem 3.4.4

The Kullback-Leibler (KL) divergence $D_{\mathrm{kl}}(M_0||M_1)$ is a special f-divergence $D_f(M_0||M_1)$ with $f(x) = x\log x$. Therefore, by Theorem 3.3.4, we have that

$$\begin{aligned} \max_{Q\in\mathcal{D}_\varepsilon} D_{\mathrm{kl}}(M_0||M_1) = \quad & \underset{\theta}{\text{maximize}} && \mu^T\theta \\ & \text{subject to} && S^{(k)}\theta = \mathbb{1} \\ & && \theta \geq 0, \end{aligned} \tag{B.28}$$

where

$$\mu_j = \mu\left(S_j^{(k)}\right) = \sum_{i\in[k]} P_0(x_i)S_{ij}^{(k)} \log\left(\frac{\sum_{i\in[k]} P_0(x_i)S_{ij}^{(k)}}{\sum_{i\in[k]} P_1(x_i)S_{ij}^{(k)}}\right)$$

for $j \in \{1,\ldots,2^k\}$ and $S^{(k)}$ is the $k \times 2^k$ staircase pattern matrix given in Definition 3.3.3.

The polytope given by $S^{(k)}\theta = \mathbb{1}$ and $\theta \geq 0$ is a closed and bounded one. Thus, there is no duality gap and solving the above linear program is equivalent to solving its dual

$$\begin{aligned} & \underset{\alpha}{\text{minimize}} && \mathbb{1}^T\alpha \\ & \text{subject to} && {S^{(k)}}^T\alpha \geq \mu. \end{aligned} \tag{B.29}$$

Note that any satisfiable solution α^* to (B.29) provides an upper bound to (B.28) since $\max \mu^T\theta = \min \mathbb{1}^T\alpha \leq \mathbb{1}^T\alpha^*$. Let $T = \{x : P_0(x) \geq P_1(x)\}$ and $T_j = \{x_i : S^{(k)}_{ij} = e^\varepsilon\}$ for $j \in [2^k]$. Set $j_i = \{j : T_j = x_i\}$ for $i \in [k]$, and consider the following choice of dual variable

$$\alpha_i^* = \frac{1}{(e^\varepsilon - 1)(e^\varepsilon + k - 1)} \left\{ (e^\varepsilon + k - 2)\, \mu\left(S^{(k)}_{j_i}\right) - \sum_{l\in[k], l\neq i} \mu\left(S^{(k)}_{j_l}\right) \right\}, \tag{B.30}$$

for $i \in [k]$. Observe that since $T_{j_i} = x_i$ we have that $P_\nu\left(T_{j_i}\right) = P_\nu\left(x_i\right)$ and since

$$\begin{aligned} \mu_j &= \sum_{i\in[k]} P_0(x_i) S^{(k)}_{ij} \log\left(\frac{\sum_{i\in[k]} P_0(x_i) S^{(k)}_{ij}}{\sum_{i\in[k]} P_1(x_i) S^{(k)}_{ij}} \right) \\ &= \left(P_0\left(T_j\right)(e^\varepsilon - 1) + 1\right) \log \frac{\left(P_0\left(T_j\right)(e^\varepsilon - 1) + 1\right)}{\left(P_1\left(T_j\right)(e^\varepsilon - 1) + 1\right)} \end{aligned}$$

we have that

$$\begin{aligned} &\mathbb{1}^T\alpha^* \\ &= \frac{1}{(e^\varepsilon - 1)(e^\varepsilon + k - 1)} \sum_{i\in[k]} \left\{ (e^\varepsilon + k - 2)\, \mu\left(S^{(k)}_{j_i}\right) - \sum_{l\in[k], l\neq i} \mu\left(S^{(k)}_{j_l}\right) \right\} \\ &= \frac{1}{(e^\varepsilon - 1)(e^\varepsilon + k - 1)} \left\{ (e^\varepsilon + k - 2) \sum_{i\in[k]} \mu\left(S^{(k)}_{j_i}\right) - \sum_{i\in[k]} \sum_{l\in[k], l\neq i} \mu\left(S^{(k)}_{j_l}\right) \right\} \\ &= \frac{1}{(e^\varepsilon - 1)(e^\varepsilon + k - 1)} \left\{ (e^\varepsilon + k - 2) \sum_{i\in[k]} \mu\left(S^{(k)}_{j_i}\right) - (k-1) \sum_{i\in[k]} \mu\left(S^{(k)}_{j_i}\right) \right\} \\ &= \frac{1}{(e^\varepsilon + k - 1)} \sum_{i\in[k]} \mu\left(S^{(k)}_{j_i}\right) \\ &= \frac{1}{(e^\varepsilon + k - 1)} \sum_{i\in[k]} \left(P_0\left(x_i\right)(e^\varepsilon - 1) + 1\right) \log \frac{\left(P_0\left(x_i\right)(e^\varepsilon - 1) + 1\right)}{\left(P_1\left(x_i\right)(e^\varepsilon - 1) + 1\right)}. \end{aligned} \tag{B.31}$$

We claim that α^* is a feasible dual variable for sufficiently large ε. In order to prove that α^* is a feasible dual variable, we show that ${S^{(k)}}^T_j \alpha^* - \mu_j \geq 0$ for all $j \in [2^k]$ for all $\varepsilon \geq \varepsilon^*$, where ε^* is a positive quantity that depends on the priors P_0 and P_1. Using the facts that

$$\begin{aligned}
\log\left(a+e^{\varepsilon}b\right) &= \varepsilon+\log b+O\left(e^{-\varepsilon}\right)\\
\frac{1}{e^{\varepsilon}+k-1} &= e^{-\varepsilon}+O\left(e^{-2\varepsilon}\right),
\end{aligned}$$

for large ε, we get that

$$\begin{aligned}
\mu_j &= \left(P_0\left(T_j\right)\left(e^{\varepsilon}-1\right)+1\right)\log\frac{\left(P_0\left(T_j\right)\left(e^{\varepsilon}-1\right)+1\right)}{\left(P_1\left(T_j\right)\left(e^{\varepsilon}-1\right)+1\right)}\\
&= \left(P_0\left(T_j\right)\log\frac{P_0\left(T_j\right)}{P_1\left(T_j\right)}\right)e^{\varepsilon}+\left(1-P_0\left(T_j\right)\right)\log\frac{P_0\left(T_j\right)}{P_1\left(T_j\right)}+O\left(e^{-\varepsilon}\right).
\end{aligned}$$

On the other hand,

$$\begin{aligned}
&{S^{(k)}}_j^T\alpha^* =\\
&\frac{1}{\left(e^{\varepsilon}-1\right)\left(e^{\varepsilon}+k-1\right)}\left\{\sum_{i\in[k]}S_{ij}^{(k)}\left(e^{\varepsilon}+k-2\right)\left(P_0\left(x_i\right)\log\frac{P_0\left(x_i\right)}{P_1\left(x_i\right)}e^{\varepsilon}+O\left(1\right)\right)\right\}\\
&\quad-\frac{1}{\left(e^{\varepsilon}-1\right)\left(e^{\varepsilon}+k-1\right)}\left\{\sum_{i\in[k]}\sum_{l\in[k],l\neq i}S_{ij}^{(k)}\left(P_0\left(x_l\right)\log\frac{P_0\left(x_l\right)}{P_1\left(x_l\right)}e^{\varepsilon}+O\left(1\right)\right)\right\}\\
&=\frac{1}{\left(e^{\varepsilon}-1\right)\left(e^{\varepsilon}+k-1\right)}\left(\left(\sum_{x_i\in T_j}P_0\left(x_i\right)\log\frac{P_0\left(x_i\right)}{P_1\left(x_i\right)}\right)e^{3\varepsilon}+O\left(e^{2\varepsilon}\right)\right)\\
&=\left(\sum_{x_i\in T_j}P_0\left(x_i\right)\log\frac{P_0\left(x_i\right)}{P_1\left(x_i\right)}\right)e^{\varepsilon}+O\left(1\right).
\end{aligned}$$

Assume, to begin with, that $j\neq\{j_1,j_2,...,j_k\}$. Then

$$ {S^{(k)}}_j^T\alpha^*-\mu_j=\left(P_0\left(T_j\right)\log\frac{P_0\left(T_j\right)}{P_1\left(T_j\right)}-\sum_{x_i\in T_j}P_0\left(x_i\right)\log\frac{P_0\left(x_i\right)}{P_1\left(x_i\right)}\right)e^{\varepsilon}+O\left(1\right). \tag{B.32}$$

Notice that for $j\neq\{j_1,j_2,...,j_k\}$, $P_0\left(T_j\right)\log\frac{P_0(T_j)}{P_1(T_j)}>\sum_{x_i\in T_j}P_0\left(x_i\right)\log\frac{P_0(x_i)}{P_1(x_i)}$ by the log-sum inequality. Therefore, there exists a $\varepsilon(j)>0$ such that ${S^{(k)}}_j^T\alpha^*-\mu_j\geq 0$ for all $\varepsilon\geq\varepsilon(j)$. If $j\in\{j_1,j_2,...,j_k\}$, it is not hard to check that ${S^{(k)}}_j^T\alpha^*-\mu_j=0$ for all ε. In this case, set $\varepsilon(j)=0$. This establishes the satisfiability of α^* for all $\varepsilon\geq\varepsilon^*=\max_{j\in[2^k]}\varepsilon(j)$. The satisfiability of α^*, in turn, shows that (B.31) is indeed an upper bound to the primal problem. It remains to show that this upper bound can be achieved via the randomized response. To this extent, recall that the randomized response is given by

$$Q(y|x) \;=\; \begin{cases} \frac{e^{\varepsilon}}{|\mathcal{X}|-1+e^{\varepsilon}} & \text{if } y = x \;, \\ \frac{1}{|\mathcal{X}|-1+e^{\varepsilon}} & \text{if } y \neq x \;. \end{cases} \tag{B.33}$$

Computing the KL divergence between M_0 and M_1 under (B.33), we get that

$$D_{\text{kl}}(M_0||M_1) = \frac{1}{(e^{\varepsilon}+k-1)} \sum_{i\in[k]} \left(P_0\left(x_i\right)\left(e^{\varepsilon}-1\right)+1\right) \log \frac{\left(P_0\left(x_i\right)\left(e^{\varepsilon}-1\right)+1\right)}{\left(P_1\left(x_i\right)\left(e^{\varepsilon}-1\right)+1\right)}. \tag{B.34}$$

Hence, the randomized response in (B.33) achieves the upper bound in (B.31). This proves the optimality of the randomized response for all $\varepsilon \geq \varepsilon^*$.

B.4.5 Proof of Theorem 3.4.3

We start the proof with a fundamental bound on the symmetrized KL divergence between the M_0 and M_1.

Lemma B.4.2 *For any $\varepsilon \geq 0$, let Q be any conditional distribution that guarantees ε differential privacy. Then for any pair of distributions P_0 and P_1, the induced marginals M_0 and M_1 must satisfy the bound*

$$D_{\text{kl}}\big(M_0||M_1\big) + D_{\text{kl}}\big(M_1||M_0\big) \leq 4\left(e^{\varepsilon}-1\right)^2 \left\|P_0 - P_1\right\|_{\text{TV}}^2. \tag{B.35}$$

The above lemma appears as Theorem 1 in [22]. By Lemma B.4.2, we have that

$$\text{OPT} = \max_{Q\in\mathcal{D}_{\varepsilon}} D_{\text{kl}}\big(M_0||M_1\big) \leq 4\left(e^{\varepsilon}-1\right)^2 \left\|P_0 - P_1\right\|_{\text{TV}}^2. \tag{B.36}$$

Let M_0^B and M_1^B be the marginals obtained by using the binary mechanism given in (3.15). By Corollary 3.4.7, we have that $\|M_0^B - M_1^B\|_{\text{TV}} = \frac{e^{\varepsilon}-1}{e^{\varepsilon}+1}\|P_0 - P_1\|_{\text{TV}}$. Consequently, by applying Pinsker's inequality to the KL divergence between M_0^B and M_1^B we get that

$$\begin{aligned} \text{BIN} &= D_{\text{kl}}\big(M_0^B || M_1^B\big) \\ &\geq 2\big\|M_0^B - M_1^B\big\|_{\text{TV}}^2 \\ &= 2\left(\frac{e^\varepsilon - 1}{e^\varepsilon + 1}\right)^2 \big\|P_0 - P_1\big\|_{\text{TV}}^2. \end{aligned} \tag{B.37}$$

Combining (B.36) and (B.37) we get that BIN $\geq \frac{1}{2(e^\varepsilon+1)^2}$OPT which was to be shown.

B.5 Information Preservation

B.5.1 Proof of Theorem 3.5.1

By Theorem 3.3.4, we have that

$$\begin{aligned} \max_{Q \in \mathcal{D}_\varepsilon} I(X;Y) = \quad & \underset{\theta}{\text{maximize}} & & \mu^T \theta \\ & \text{subject to} & & S^{(k)}\theta = \mathbb{1} \\ & & & \theta \geq 0, \end{aligned} \tag{B.38}$$

where $\mu_j = \mu\left(S_j^{(k)}\right) = \sum_{i \in [k]} P(x_i)\, S_{ij}^{(k)} \log\left(\frac{S_{ij}^{(k)}}{\sum_{i\in[k]} P(x_i) S_{ij}^{(k)}}\right)$ for $j \in \{1, \dots, 2^k\}$ and $S^{(k)}$ is the $k \times 2^k$ staircase pattern matrix given in Definition 3.3.3. The polytope given by $S^{(k)}\theta = \mathbb{1}$ and $\theta \geq 0$ is a closed and bounded one. Thus, there is no duality gap and solving the above linear program is equivalent to solving its dual

$$\begin{aligned} & \underset{\alpha}{\text{minimize}} & & \mathbb{1}^T \alpha \\ & \text{subject to} & & {S^{(k)}}^T \alpha \geq \mu. \end{aligned} \tag{B.39}$$

Note that any satisfiable solution α^* to (B.39) provides an upper bound to (B.38) since $\max \mu^T\theta = \min \mathbb{1}^T\alpha \leq \mathbb{1}^T\alpha^*$. Let $T_j = \{x_i : S_{ij}^{(k)} = e^\varepsilon\}$ and set $j_1 = \{j : T_j = T\}$ and $j_2 = \{j : T_j = T^c\}$. Consider the following choice of dual variable

$$\alpha_i^* = \frac{1}{\left(e^\varepsilon+1\right)\left(e^\varepsilon-1\right)}\begin{cases} \frac{e^\varepsilon \mu\left(S_{j_1}^{(k)}\right)-\mu\left(S_{j_2}^{(k)}\right)}{|T|} & \forall i \in T \\ \frac{e^\varepsilon \mu\left(S_{j_2}^{(k)}\right)-\mu\left(S_{j_1}^{(k)}\right)}{|T^c|} & \forall i \in T^c \end{cases}. \tag{B.40}$$

Observe that since $T_{j_1} = T$, $T_{j_2} = T^c$, and

$$\begin{aligned} \mu_j \quad = \quad & P\left(T_j\right) e^\varepsilon \log \frac{e^\varepsilon}{P\left({T_j}^c\right) + e^\varepsilon P\left(T_j\right)} \\ & + P\left({T_j}^c\right) \log \frac{1}{P\left({T_j}^c\right) + e^\varepsilon P\left(T_j\right)}, \end{aligned} \tag{B.41}$$

we have that

$$\begin{aligned} & \mathbb{1}^T \alpha^* \\ & = \frac{1}{\left(e^\varepsilon+1\right)\left(e^\varepsilon-1\right)}\left\{\sum_{i\in T}\frac{1}{|T|}\left(e^\varepsilon\mu\left(S_{j_1}^{(k)}\right)-\mu\left(S_{j_2}^{(k)}\right)\right)\right. \\ & \qquad \left. + \sum_{i \in T^c}\frac{1}{|T^c|}\left(e^\varepsilon\mu\left(S_{j_2}^{(k)}\right)-\mu\left(S_{j_1}^{(k)}\right)\right)\right\} \\ & = \frac{1}{\left(e^\varepsilon+1\right)}\left(\mu\left(S_{j_1}^{(k)}\right)+\mu\left(S_{j_1}^{(k)}\right)\right) \\ & = \frac{1}{e^\varepsilon+1}\left\{P\left(T\right)e^\varepsilon\log\frac{e^\varepsilon}{P\left(T^c\right)+e^\varepsilon P\left(T\right)} + P\left(T^c\right)\log\frac{1}{P\left(T^c\right)+e^\varepsilon P\left(T\right)}\right\} + \\ & \qquad \frac{1}{e^\varepsilon+1}\left\{P\left(T^c\right)e^\varepsilon\log\frac{e^\varepsilon}{P\left(T\right)+e^\varepsilon P\left(T^c\right)} + P\left(T\right)\log\frac{1}{P\left(T\right)+e^\varepsilon P\left(T^c\right)}\right\}. \end{aligned} \tag{B.42}$$

We claim that α^* is a feasible dual variable for sufficiently small ε. In order to prove that α^* is a feasible dual variable, we show that $\left({S^{(k)}}^T\alpha^*\right)_j - \mu_j \geq 0$ for all $j \in \{1, \ldots, 2^k\}$ and all $\varepsilon \leq \varepsilon^*$, where ε^* is a positive quantity that depends on P. Using the facts that

$$\begin{aligned} e^\varepsilon &= 1 + \varepsilon + \frac{1}{2}\varepsilon + O\left(\varepsilon^3\right) \\ \log\left(a + e^\varepsilon b\right) &= b\varepsilon + \frac{b(1-b)}{2}\varepsilon^2 + O\left(\varepsilon^3\right) \\ \frac{1}{1+e^\varepsilon} &= \frac{1}{2} - \frac{1}{4}\varepsilon + O\left(\varepsilon^2\right), \end{aligned}$$

for small ε, we get that

$$\begin{aligned}
\mu_j &= P\left(T_j\right) e^{\varepsilon} \log \frac{e^{\varepsilon}}{P\left({T_j}^c\right) + e^{\varepsilon} P\left(T_j\right)} + P\left({T_j}^c\right) \log \frac{1}{P\left({T_j}^c\right) + e^{\varepsilon} P\left(T_j\right)} \\
&= P\left(T_j\right) e^{\varepsilon} \varepsilon - \left(P\left(T_j\right)\left(e^{\varepsilon} - 1\right) + 1\right) \log\left(P\left(T_j\right)\left(e^{\varepsilon} - 1\right) + 1\right) \\
&= \frac{1}{2} P\left(T_j\right) P\left(T_j^c\right) \varepsilon^2 + O\left(\varepsilon^3\right).
\end{aligned}$$

On the other hand,

$$\begin{aligned}
&\left(S^{(k)^T} \alpha^*\right)_j \\
&= S_j^{(k)^T} \alpha^* \\
&= \frac{1}{\left(e^{\varepsilon}+1\right)\left(e^{\varepsilon}-1\right)} \left\{ \sum_{i \in T} \frac{S_{ij}^{(k)}}{|T|} \left(e^{\varepsilon} \mu\left(S_{j_1}^{(k)}\right) - \mu\left(S_{j_2}^{(k)}\right)\right) \right. \\
&\qquad\qquad\qquad\qquad \left. + \sum_{i \in T^c} \frac{S_{ij}^{(k)}}{|T^c|} \left(e^{\varepsilon} \mu\left(S_{j_2}^{(k)}\right) - \mu\left(S_{j_1}^{(k)}\right)\right) \right\} \\
&= \frac{1}{\left(e^{\varepsilon}+1\right)\left(e^{\varepsilon}-1\right)} \left(e^{\varepsilon} \mu\left(S_{j_1}^{(k)}\right) - \mu\left(S_{j_2}^{(k)}\right)\right) \left(\frac{|T_j \cap T|}{|T|} e^{\varepsilon} + \frac{|T_j^c \cap T|}{|T|} \right) \\
&\quad + \frac{1}{\left(e^{\varepsilon}+1\right)\left(e^{\varepsilon}-1\right)} \left(e^{\varepsilon} \mu\left(S_{j_2}^{(k)}\right) - \mu\left(S_{j_1}^{(k)}\right)\right) \left(\frac{|T_j \cap T^c|}{|T^c|} e^{\varepsilon} + \frac{|T_j^c \cap T^c|}{|T^c|} \right) \\
&= \frac{1}{\left(e^{\varepsilon}+1\right)} \left(\frac{1}{2} P(T) P(T^c) \varepsilon^2 + O\left(\varepsilon^3\right) \right) \left\{ \frac{|T_j \cap T^c|}{|T^c|} + \frac{|T_j^c \cap T^c|}{|T^c|} \right. \\
&\qquad\qquad\qquad\qquad \left. + \frac{|T_j \cap T|}{|T|} + \frac{|T_j^c \cap T|}{|T|} + O(\varepsilon) \right\} \\
&= \frac{1}{2} P(T) P(T^c) \varepsilon^2 + O\left(\varepsilon^3\right),
\end{aligned}$$

where we have used the facts that $T_{j_1} = T$, $T_{j_2} = T^c$, and

$$\begin{aligned}
\mu\left(S_{j_1}^{(k)}\right) &= \frac{1}{2} P(T) P(T^c) \varepsilon^2 + O\left(\varepsilon^3\right) \\
\mu\left(S_{j_2}^{(k)}\right) &= \frac{1}{2} P(T) P(T^c) \varepsilon^2 + O\left(\varepsilon^3\right).
\end{aligned}$$

Let $f(z) = |z - \frac{1}{2}|$, $g(z) = -z \log z - (1-z)\log(1-z)$, and $h(z) = z(1-z)$ for $0 \leq z \leq 1$. On the one hand, g and h are monotonically increasing over $0 \leq z \leq \frac{1}{2}$ and monotonically decreasing over $\frac{1}{2} \leq z \leq 1$ but on the other hand, f is monotonically decreasing over $0 \leq z \leq \frac{1}{2}$ and monotonically increasing over $\frac{1}{2} \leq z \leq 1$. Therefore,

$$\begin{aligned}
&T \in \arg\min_{A \subseteq \mathcal{X}} \ \left|P(A) - \frac{1}{2}\right| \\
&\Leftrightarrow T \in \arg\max_{A \subseteq \mathcal{X}} \ -P(A)\log P(A) - P(A^c)\log P(A^c) \\
\\
&\Leftrightarrow T \in \arg\max_{A \subseteq \mathcal{X}} \ P(A)P(A^c).
\end{aligned}$$

Since the set T was chosen so that it maximizes $P(T)\,P(T^c)$, we have that $P(T)\,P(T^c) \geq P(T_j)\,P\left(T_j^c\right)$ for all $j \in \{1, \ldots, 2^k\}$. Assume, to begin with, that $j \neq \{j_1, j_2\}$. Then by the uniqueness of the maximizer assumption stated in the theorem, we have that $P(T)\,P(T^c) > P(T_j)\,P\left(T_j^c\right)$.

$$\left(S^T \alpha^*\right)_j - \mu_j = \frac{1}{2}\left(P(T)\,P(T^c) - P(T_j)\,P\left(T_j^c\right)\right)\varepsilon^2 + O\left(\varepsilon^3\right), \tag{B.43}$$

and thus, there exists an ε^* that depends on P such that $\left(S^{(k)^T}\alpha^*\right)_j - \mu_j \geq 0$ for all $\varepsilon \leq \varepsilon^*$. If $j = \{j_1, j_2\}$, it is not hard to check that $\left(S^{(k)^T}\alpha^*\right)_j - \mu_j = 0$ for all ε. This establishes the satisfiability of α^* for all $\varepsilon \leq \varepsilon^*$ which proves an upper bound on the primal problem (given in (B.42)). It remains to show that the upper bound can be indeed achieved via the binary mechanism. To this extent, recall that the binary mechanism is given by

$$Q(0|x) = \begin{cases} \frac{e^\varepsilon}{1+e^\varepsilon} & \text{if } x \in T\ , \\ \frac{1}{1+e^\varepsilon} & \text{if } x \notin T\ . \end{cases} \quad Q(1|x) = \begin{cases} \frac{e^\varepsilon}{1+e^\varepsilon} & \text{if } x \notin T\ , \\ \frac{1}{1+e^\varepsilon} & \text{if } x \in T\ . \end{cases} \tag{B.44}$$

Computing the $I(X;Y)$ under (B.44), we get that

$$\begin{aligned}
&I(X;Y) \\
&= \frac{1}{e^\varepsilon + 1}\left\{P(T)\,e^\varepsilon \log\frac{e^\varepsilon}{P(T^c) + e^\varepsilon P(T)} + P(T^c)\log\frac{1}{P(T^c) + e^\varepsilon P(T)}\right\} + \\
&\qquad \frac{1}{e^\varepsilon + 1}\left\{P(T^c)\,e^\varepsilon \log\frac{e^\varepsilon}{P(T) + e^\varepsilon P(T^c)} + P(T)\log\frac{1}{P(T) + e^\varepsilon P(T^c)}\right\}.
\end{aligned}$$

Hence, the binary mechanism in (B.44) achieves the upper bound in (B.42). This proves the optimality of the binary mechanism for all $\varepsilon \leq \varepsilon^*$.

B.5.2 Proof of Theorem 3.5.2

We start by proving an upper bound on $\max_{Q\in\mathcal{D}_\varepsilon} I\left(X;Y\right)$ which is tight for $\varepsilon \leq 1$. Recall that by Theorem 3.3.4, we have that

$$\begin{aligned}\mathrm{OPT} = \max_{Q\in\mathcal{D}_\varepsilon} I\left(X;Y\right) = \quad \underset{\theta}{\text{maximize}} \quad & \sum_{j=1}^{2^k} \mu_j \theta_j \\ \text{subject to} \quad & S^{(k)}\theta = \mathbb{1} \\ & \theta \geq 0,\end{aligned}$$

where

$$\begin{aligned}\mu_j &= \mu\left(S_j^{(k)}\right) \\ &= \sum_{i\in[k]} P\left(x_i\right) S_{ij}^{(k)} \log\left(\frac{S_{ij}^{(k)}}{\sum_{i\in[k]} P\left(x_i\right) S_{ij}^{(k)}}\right) \\ &= P\left(T_j\right) e^{\varepsilon}\varepsilon \\ &\qquad - \left(P\left(T_j\right)\left(e^{\varepsilon}-1\right)+1\right)\log\left(P\left(T_j\right)\left(e^{\varepsilon}-1\right)+1\right), \quad \text{(B.45)}\end{aligned}$$

$T_j = \{i : S_{ij}^{(k)} = e^{\varepsilon}\}$, and $S^{(k)}$ is the $k \times 2^k$ staircase pattern matrix given in Definition 3.3.3.

Lemma B.5.1 *For all distributions P and all ε, the following bound holds*

$$\mathrm{OPT} = \max_{Q\in\mathcal{D}_\varepsilon} I\left(X;Y\right) \leq \left(\max_j \mu_j\right)\frac{k}{e^{\varepsilon}+k-1}. \quad \text{(B.46)}$$

The proof of this lemma is given in Section B.5.3. In what follows, we will make the dependency of μ_j on $P\left(T_j\right)$ and ε explicit by writing $\mu_j\left(P\left(T_j\right),\varepsilon\right)$ for μ_j. From the proof of Theorem 3.5.1, we have that the partition set T defined in (3.21) is given by $T \in \arg\max_{A\subseteq\mathcal{X}} P(A)P(A^c)$. It is easy to check that the binary mechanism given in (3.22) achieves the following utility

$$\mathrm{BIN} = \frac{\mu\left(P\left(T\right),\varepsilon\right)+\mu\left(P\left(T^c\right),\varepsilon\right)}{e^{\varepsilon}+1}. \quad \text{(B.47)}$$

Lemma B.5.2 *For all distributions P and all $\varepsilon \leq 1$, the following bound holds:*

$$\frac{\max_j \mu_j}{\mu\left(P\left(T\right),\varepsilon\right)+\mu\left(P\left(T^c\right),\varepsilon\right)} \leq 1. \quad \text{(B.48)}$$

The proof of the above lemma is given in Section B.5.4. Combining the results of lemmas B.5.1 and B.5.2 we get that

$$\begin{aligned}\frac{\text{OPT}}{\text{BIN}} &\leq \frac{\max_j \mu_j}{\mu\left(P\left(T\right),\varepsilon\right)+\mu\left(P\left(T^c\right),\varepsilon\right)}\frac{k}{e^\varepsilon+k-1}\left(e^\varepsilon+1\right)\\ &\leq \frac{k}{e^\varepsilon+k-1}\left(e^\varepsilon+1\right)\\ &\leq e^\varepsilon+1,\end{aligned}$$

for all $\varepsilon \leq 1$. This concludes the proof.

B.5.3 Proof of Lemma B.5.1

To begin with, since $S_1^{(k)}=\mathbb{1}=\frac{1}{e^\varepsilon}S_{2^k}^{(k)}$ and μ is homogeneous, we have that $\theta_1\mu_1+\theta_{2^k}\mu_{2^k}=\left(\frac{1}{e^\varepsilon}\theta_1+\theta_{2^k}\right)\mu_{2^k}$. Therefore, the following two maximization problems are equivalent:

$$\begin{array}{ll}\underset{\theta}{\text{maximize}} & \sum_{j=1}^{2^k}\mu_j\theta_j\\ \text{subject to} & S^{(k)}\theta=\mathbb{1}\\ & \theta\geq 0\end{array} = \begin{array}{ll}\underset{\theta}{\text{maximize}} & \sum_{j=1}^{2^k-1}\tilde{\mu}_j\theta_j\\ \text{subject to} & \tilde{S}^{(k)}\theta=\mathbb{1}\\ & \theta\geq 0,\end{array} \tag{B.49}$$

where $\tilde{\mu}_j=\mu_{j+1}$ and $\tilde{S}^{(k)}$ is obtained by deleting the first column of $S^{(k)}$. Moreover, using the fact that $\max_{j\in[2^k-1]}\tilde{\mu}_j\leq\max_{j\in[2^k]}\mu_j$ and weak duality, we get that

$$\begin{array}{ll}\underset{\theta}{\text{maximize}} & \tilde{\mu}^T\theta\\ \text{subject to} & \tilde{S}^{(k)}\theta=\mathbb{1}\\ & \theta\geq 0\end{array} \leq \left(\max_{j\in[2^k-1]}\tilde{\mu}_j\right)\begin{array}{ll}\underset{\theta}{\text{maximize}} & \mathbb{1}^T\theta\\ \text{subject to} & \tilde{S}^{(k)}\theta=\mathbb{1}\\ & \theta\geq 0\end{array}$$

$$\leq \left(\max_{j\in[2^k]}\mu_j\right)\begin{array}{ll}\underset{\alpha}{\text{minimize}} & \mathbb{1}^T\alpha\\ \text{subject to} & \tilde{S}^{(k)^T}\alpha\geq\mathbb{1}.\end{array}$$

Consider the following choice of dual variable $\alpha_i^* = \frac{1}{e^\varepsilon+k-1}$. We claim that α^* is satisfiable. This can be easily verified by noting that

$$\left(\tilde{S}^{(k)^T}\alpha^*\right)_j = \tilde{S}_j^{(k)^T}\alpha^* = \frac{|T_j|e^\varepsilon + (k-|T_j|)}{e^\varepsilon+k-1} = \frac{|T_j|(e^\varepsilon-1)+k}{e^\varepsilon+k-1} \geq 1,$$

where the last inequality holds since $|T_j| \geq 1$ (this is true because we have deleted the first column of $S^{(k)}$). Therefore, OPT $\leq (\max_j \mu_j)\, \mathbb{1}^T\alpha^* = (\max_j \mu_j)\frac{k}{e^\varepsilon+k-1}$ which was to be shown.

B.5.4 Proof of Lemma B.5.2

Let $\mu\left(z,\varepsilon\right)$ be the function obtained by replacing $P\left(T_j\right)$ by the continuous variable $z \in [0,1]$ in $\mu_j\left(P\left(T_j\right),\varepsilon\right)$. Taking the derivative of $\mu\left(z,\varepsilon\right)$ with respect to z we get

$$\mu'\left(z,\varepsilon\right) = e^\varepsilon\varepsilon - (e^\varepsilon-1) - (e^\varepsilon-1)\log\left(z(e^\varepsilon-1)+1\right). \tag{B.50}$$

Observe that $\mu'\left(z,\varepsilon\right) > 0$ for all

$$z < z^*(\varepsilon) = \frac{1}{e^\varepsilon-1}\left(e^{\left\{\frac{e^\varepsilon\varepsilon}{e^\varepsilon-1}-1\right\}}-1\right), \tag{B.51}$$

$\mu'\left(z,\varepsilon\right) < 0$ for all $z > z^*(\varepsilon)$, and $\mu'\left(z,\varepsilon\right) = 0$ for $z = z^*(\varepsilon)$. Combining this with the fact that $\mu\left(0,\varepsilon\right) = \mu\left(1,\varepsilon\right) = 0$ we get that $\mu\left(z,\varepsilon\right) \geq 0$ for all $z \in [0,1]$ and for any fixed ε, $\mu\left(z,\varepsilon\right)$ is maximized at $z^*(\varepsilon)$.

Set $x^* \in \arg\max_{x\in\mathcal{X}} P\left(x\right)$ and fix an $\varepsilon \leq 1$. We will treat the following three cases separately.

Case 1: $P(x^*) \in [1-z^*(\varepsilon),1]$.

Claim 5 *Let $T = \{x^*\}$. Then*

$$\{T,T^c\} = \arg\max_{A\subseteq\mathcal{X}} P(A)P(A^c)$$

and

$$\max_{A\subseteq\mathcal{X}} \mu(P(A),\varepsilon) = \max\left(\mu(P(T),\varepsilon),\mu(P(T^c),\varepsilon)\right).$$

Proof 12 *Observe that $z^*(\varepsilon) \leq \frac{1}{2}$ for all ε and $T^c = \mathcal{X}\setminus\{x^*\}$. The function $f(z) = z(1-z)$ decreases over the range $[\frac{1}{2},1] \supseteq [1-z^*(\varepsilon),1]$.*

Thus, for all $A \supset T$, $P(T)P(T^c) > P(A)P(A^c)$ because $P(T) \geq 1 - z^*(\varepsilon)$. This proves that $T \in \arg\max_{A\subseteq\mathcal{X}} P(A)P(A^c)$ and for all $A \supset T$, $A \notin \arg\max_{A\subseteq\mathcal{X}} P(A)P(A^c)$. Using a similar approach, we can show that $T^c \in \arg\max_{A\subseteq\mathcal{X}} P(A)P(A^c)$ and for all $A \subset T^c$, $A \notin \arg\max_{A\subseteq\mathcal{X}} P(A)P(A^c)$. Therefore, $\{T, T^c\} = \arg\max_{A\subseteq\mathcal{X}} P(A)P(A^c)$. This proves the first part of the claim. The function $\mu\left(z, \varepsilon\right)$ increases over the range $[0, z^*(\varepsilon)]$. Thus, for all $A \subseteq T^c$, $\mu(P(A), \varepsilon) \leq \mu(P(T^c), \varepsilon)$ because $P(T^c) \leq z^*(\varepsilon)$. On the other hand, note that $\mu\left(z, \varepsilon\right)$ decreases over the range $[z^*(\varepsilon), 1]$ which includes the range $[1-z^*(\varepsilon), 1]$. Thus, for all A such that $A \supseteq T$, $\mu(P(A), \varepsilon) \leq \mu(P(T), \varepsilon)$ because $P(T) \geq 1 - z^*(\varepsilon)$. This proves that $\max\left(\mu(P(T), \varepsilon), \mu(P(T^c), \varepsilon)\right) = \max_{A\subseteq\mathcal{X}} \mu(P(A), \varepsilon)$.

Using the above claim, we can conclude that the partition set T defined in (3.21) is equal to $\{x^*\}$ and

$$\begin{aligned}
\frac{\max_j \mu_j}{\mu\left(P\left(T\right), \varepsilon\right) + \mu\left(P\left(T^c\right), \varepsilon\right)} &= \frac{\max_{A\subseteq\mathcal{X}} \mu(P(A), \varepsilon)}{\mu\left(P\left(T\right), \varepsilon\right) + \mu\left(P\left(T^c\right), \varepsilon\right)} \\
&\leq \frac{\max_{A\subseteq\mathcal{X}} \mu(P(A), \varepsilon)}{\max\left(\mu(P(T), \varepsilon), \mu(P(T^c), \varepsilon)\right)} \\
&= 1.
\end{aligned}$$

Case 2: $P(x^*) \in [\frac{1}{2}, 1 - z^*(\varepsilon)]$. Using the first part of the proof of Claim 5, we can show that if $T = \{x^*\}$, then $\{T, T^c\} = \arg\max_{A\subseteq\mathcal{X}} P(A)P(A^c)$. Therefore, the partition set T defined in (3.21) is equal to $\{x^*\}$ and

$$\begin{aligned}
\frac{\max_j \mu_j}{\mu\left(P\left(T\right), \varepsilon\right) + \mu\left(P\left(T^c\right), \varepsilon\right)} &= \frac{\max_{A\subseteq\mathcal{X}} \mu(P(A), \varepsilon)}{\mu\left(P\left(T\right), \varepsilon\right) + \mu\left(P\left(T^c\right), \varepsilon\right)} \\
&\leq \frac{\mu(z^*(\varepsilon), \varepsilon)}{\mu\left(P\left(x^*\right), \varepsilon\right) + \mu\left(1 - P\left(x^*\right), \varepsilon\right)} \\
&\leq 1.
\end{aligned}$$

Case 3: $P(x^*) \in [0, \frac{1}{2}]$.

Claim 6 *There exists a set $A \subset \mathcal{X}$ such that $\frac{1}{2} - P(x^*) \leq P(A) \leq \frac{1}{2}$.*

Proof 13 *Without loss of generality, assume that the sequence $P(x_i)$, $i \in [k]$, is sorted in increasing order. Let $l^* = \min\{l : \sum_{i=1}^{l} P(x_i) \geq \frac{1}{2}\}$. From the definition of l^*, $P(\{x_1, \ldots, x_{l^*-1}\}) < \frac{1}{2}$ and $P(\{x_1, \ldots, x_{l^*}\}) \geq \frac{1}{2}$. Further,*

$$P(\{x_1, \ldots, x_{l^*-1}\}) = P(\{x_1, \ldots, x_{l^*}\}) - P(x_{l^*})$$

and since $x^* \in \arg\max_{x \in \mathcal{X}} P(x)$*,* $P(x_{l^*}) \leq P(x^*)$*. Therefore, if* $A = \{x_1, \ldots, x_{l^*-1}\}$*, then* $\frac{1}{2} - P(x^*) \leq P(A) \leq \frac{1}{2}$*.*

Let $P(T) = \min\{P(B) : B \in \arg\max_{A \subseteq \mathcal{X}} P(A)P(A^c)\}$. We claim that $\frac{1}{4} \leq P(T) \leq \frac{1}{2}$. The upper bound on $P(T)$ follows immediately from its definition. To prove the lower bound on $P(T)$, consider the set A given in Claim 6 and observe that

$$\begin{aligned} P(T) &\geq \max(P(x^*), P(A)) \\ &\geq \max(P(x^*), \frac{1}{2} - P(x^*)) \\ &\geq \frac{1}{4}. \end{aligned}$$

All the inequalities follow from Claim 6 and the fact that $P(x^*) \in [0, \frac{1}{2}]$.

Since $\frac{1}{4} \leq P(T) \leq \frac{1}{2}$, we have that $\frac{1}{2} \leq P(T^c) \leq \frac{3}{4}$. Moreover, the function $\mu(z, \varepsilon)$ decreases over the range $[z^*(\varepsilon), 1] \supset [\frac{1}{2}, \frac{3}{4}]$ and increases over the range $[\frac{1}{4}, z^*(\varepsilon)]$. Therefore, $\mu(P(T^c), \varepsilon) \geq \mu\left(\frac{3}{4}, \varepsilon\right)$ and $\mu(P(T), \varepsilon) \geq \min\left(\mu\left(\frac{1}{2}, \varepsilon\right), \mu\left(\frac{1}{4}, \varepsilon\right)\right)$. Putting it all together, we have that

$$\begin{aligned} \frac{\max_j \mu_j}{\mu(P(T), \varepsilon) + \mu(P(T^c), \varepsilon)} &= \frac{\max_{A \subseteq \mathcal{X}} \mu(P(A), \varepsilon)}{\mu(P(T), \varepsilon) + \mu(P(T^c), \varepsilon)} \\ &\leq \frac{\mu(z^*(\varepsilon), \varepsilon)}{\min\left(\mu\left(\frac{1}{2}, \varepsilon\right), \mu\left(\frac{1}{4}, \varepsilon\right)\right) + \mu\left(\frac{3}{4}, \varepsilon\right)} \\ &\leq 1. \end{aligned}$$

B.5.5 Proof of Theorem 3.5.3

By Theorem 3.3.4, we have that

$$\max_{Q \in \mathcal{D}_\varepsilon} I(X;Y) = \begin{aligned} &\underset{\theta}{\text{maximize}} && \mu^T \theta \\ &\text{subject to} && S^{(k)}\theta = \mathbb{1} \\ &&& \theta \geq 0, \end{aligned} \tag{B.52}$$

where

$$\mu_j = \mu\left(S_j^{(k)}\right) = \sum_{i \in [k]} P(x_i) S_{ij}^{(k)} \log\left(\frac{S_{ij}^{(k)}}{\sum_{i \in [k]} P(x_i) S_{ij}^{(k)}}\right)$$

for $j \in \{1, \ldots, 2^k\}$ and $S^{(k)}$ is the $k \times 2^k$ staircase pattern matrix given in Definition 3.3.3. The polytope given by $S^{(k)}\theta = \mathbb{1}$ and $\theta \geq 0$ is a closed and bounded one. Thus, there is no duality gap and solving the above linear program is equivalent to solving its dual

$$\begin{aligned} \underset{\alpha}{\text{minimize}} \quad & \mathbb{1}^T \alpha \\ \text{subject to} \quad & {S^{(k)}}^T \alpha \geq \mu. \end{aligned} \tag{B.53}$$

Note that any satisfiable solution α^* to (B.53) provides an upper bound to (B.52) since $\max \mu^T \theta = \min \mathbb{1}^T \alpha \leq \mathbb{1}^T \alpha^*$. Let $T_j = \{x_i : S_{ij}^{(k)} = e^\varepsilon\}$ and set $j_i = \{j : T_j = i\}$ for $i \in \{1, \ldots, k\}$. Consider the following choice of dual variable

$$\alpha_i^* = \frac{1}{(e^\varepsilon - 1)(e^\varepsilon + k - 1)} \left\{ (e^\varepsilon + k - 2)\, \mu\left(S_{j_i}^{(k)}\right) - \sum_{l \in [k], l \neq i} \mu\left(S_{j_l}^{(k)}\right) \right\}, \tag{B.54}$$

for $i \in \{1, \ldots, k\}$. Observe that since $T_{j_i} = i$ we have that $P(T_{j_i}) = P(x_i)$ and since

$$\mu_j \;=\; P(T_j)\, e^\varepsilon \log \frac{e^\varepsilon}{P({T_j}^c) + e^\varepsilon P(T_j)} + P({T_j}^c) \log \frac{1}{P({T_j}^c) + e^\varepsilon P(T_j)},$$

we have that

$$
\begin{aligned}
&\mathbb{1}^T\alpha^* \\
&= \frac{1}{\left(e^\varepsilon-1\right)\left(e^\varepsilon+k-1\right)}\sum_{i\in[k]}\left\{\left(e^\varepsilon+k-2\right)\mu\left(S_{j_i}^{(k)}\right)-\sum_{l\in[k],l\neq i}\mu\left(S_{j_l}^{(k)}\right)\right\} \\
&= \frac{1}{\left(e^\varepsilon-1\right)\left(e^\varepsilon+k-1\right)}\left\{\left(e^\varepsilon+k-2\right)\sum_{i\in[k]}\mu\left(S_{j_i}^{(k)}\right)-\sum_{i\in[k]}\sum_{l\in[k],l\neq i}\mu\left(S_{j_l}^{(k)}\right)\right\} \\
&= \frac{1}{\left(e^\varepsilon-1\right)\left(e^\varepsilon+k-1\right)}\left\{\left(e^\varepsilon+k-2\right)\sum_{i\in[k]}\mu\left(S_{j_i}^{(k)}\right)-(k-1)\sum_{i\in[k]}\mu\left(S_{j_i}^{(k)}\right)\right\} \\
&= \frac{1}{\left(e^\varepsilon+k-1\right)}\sum_{i\in[k]}\mu\left(S_{j_i}^{(k)}\right) \\
&= \frac{1}{\left(e^\varepsilon+k-1\right)}\sum_{i\in[k]}\left\{P\left(x_i\right)e^\varepsilon\log\frac{e^\varepsilon}{P\left(x_i\right)\left(e^\varepsilon-1\right)+1}\right. \\
&\qquad\qquad\left.+\left(1-P\left(x_i\right)\right)\log\frac{1}{P\left(x_i\right)\left(e^\varepsilon-1\right)+1}\right\}.
\end{aligned}
\tag{B.55}
$$

We claim that α^* is a feasible dual variable for sufficiently large ε. In order to prove that α^* is a feasible dual variable, we show that $\left(S^{(k)^T}\alpha^*\right)_j-\mu_j\geq 0$ for all $j\in\{1,\ldots,2^k\}$ and all $\varepsilon\geq\varepsilon^*$, where ε^* is a positive quantity that depends on P. Using the fact that

$$\log\left(a+e^\varepsilon b\right)=\varepsilon+\log b+O\left(e^{-\varepsilon}\right), \tag{B.56}$$

for large ε, we get that

$$
\begin{aligned}
\mu_j &= P\left(T_j\right)e^\varepsilon\log\frac{e^\varepsilon}{P\left({T_j}^c\right)+e^\varepsilon P\left(T_j\right)}+P\left({T_j}^c\right)\log\frac{1}{P\left({T_j}^c\right)+e^\varepsilon P\left(T_j\right)} \\
&= P\left(T_j\right)e^\varepsilon\varepsilon-\left(P\left(T_j\right)\left(e^\varepsilon-1\right)+1\right)\log\left(P\left(T_j\right)\left(e^\varepsilon-1\right)+1\right) \\
&= P\left(T_j\right)e^\varepsilon\varepsilon-\left(P\left(T_j\right)\left(e^\varepsilon-1\right)+1\right)\left(\varepsilon+\log P\left(T_j\right)+O\left(e^{-\varepsilon}\right)\right) \\
&= -\left(P\left(T_j\right)\log P\left(T_j\right)\right)e^\varepsilon+O\left(\varepsilon\right).
\end{aligned}
$$

On the other hand,

$$
\begin{aligned}
&\left(S^{(k)^T}\alpha^*\right)_j \\
&= S_j^{(k)^T}\alpha^* \\
&= e^\varepsilon \sum_{i\in T_j}\alpha_i^* + \sum_{i\in T_j^c}\alpha_i^* \\
&= \frac{-1}{\left(e^\varepsilon-1\right)\left(e^\varepsilon+k-1\right)}\left\{\sum_{i\in[k]} S_{ij}^{(k)}\left(e^\varepsilon+k-2\right)\left(P\left(x_i\right)\log P\left(x_i\right)e^\varepsilon \ +O\left(\varepsilon\right)\right)\right\} \\
&\quad + \frac{1}{\left(e^\varepsilon-1\right)\left(e^\varepsilon+k-1\right)}\left\{\sum_{i\in[k]}\sum_{l\in[k],l\neq i} S_{ij}^{(k)}\left(\left(P\left(x_l\right)\log P\left(x_l\right)\right)e^\varepsilon+O\left(\varepsilon\right)\right)\right\} \\
&= -\frac{1}{\left(e^\varepsilon-1\right)\left(e^\varepsilon+k-1\right)}\left(\left(\sum_{i\in T_j} P\left(x_i\right)\log P\left(x_i\right)\right)e^{3\varepsilon}+O\left(e^{2\varepsilon}\varepsilon\right)\right) \\
&= -\left(\sum_{i\in T_j} P\left(x_i\right)\log P\left(x_i\right)\right)e^\varepsilon+O\left(\varepsilon\right).
\end{aligned}
$$

Assume, to begin with, that $j\neq\{j_1,j_2,...,j_k\}$. Then

$$
\left(S^{(k)^T}\alpha^*\right)_j-\mu_j=\left(P\left(T_j\right)\log P\left(T_j\right)-\sum_{i\in T_j}P\left(x_i\right)\log P\left(x_i\right)\right)e^\varepsilon+O\left(\varepsilon\right). \tag{B.57}
$$

Notice that for $j\neq\{j_1,j_2,...,j_k\}$, $P\left(T_j\right)\log P\left(T_j\right)>\sum_{i\in T_j}P\left(x_i\right)\log P\left(x_i\right)$. Therefore, there exists an $\varepsilon^*>0$ such that $\left(S^{(k)^T}\alpha^*\right)_j-\mu_j\geq 0$ for all $\varepsilon\geq\varepsilon^*$. If $j\in\{j_1,j_2,...,j_k\}$, it is not hard to check that $\left(S^{(k)^T}\alpha^*\right)_j-\mu_j=0$ for all ε. This establishes the satisfiability of α^* for all $\varepsilon\geq\varepsilon^*$. It remains to show that the upper bound can be indeed achieved via the randomized response mechanism. To this extent, recall that the randomized response is given by

$$
Q(y|x) \ = \begin{cases} \dfrac{e^\varepsilon}{|\mathcal{X}|-1+e^\varepsilon} & \text{if } y=x\ , \\ \dfrac{1}{|\mathcal{X}|-1+e^\varepsilon} & \text{if } y\neq x\ . \end{cases} \tag{B.58}
$$

Computing the $I\left(X;Y\right)$ under (B.58), we get that

$$\begin{aligned} I\left(X;Y\right) &= \frac{1}{e^{\varepsilon}+k-1}\sum_{i\in[k]}\left\{P\left(x_i\right)e^{\varepsilon}\log\frac{e^{\varepsilon}}{P\left(x_i\right)\left(e^{\varepsilon}-1\right)+1}\right. \\ &\qquad\qquad \left.+\left(1-P\left(x_i\right)\right)\log\frac{1}{P\left(x_i\right)\left(e^{\varepsilon}-1\right)+1}\right\}. \end{aligned}$$

Hence, the randomized response mechanism achieves the upper bound (B.55). This proves the optimality of the randomized response for all $\varepsilon \geq \varepsilon^*$.

B.6 Approximate Local Differential Privacy

B.6.1 Proof of Proposition 3.6.1

Let $U\left(Q\right)$ be a utility mechanism of the form $U\left(Q\right) = \sum_{\mathcal{Y}}\mu(Q_y)$, where μ is a sublinear function. Consider a stochastic mapping W of dimensions $\ell \times m$ and let QW be the stochastic mapping obtained by first applying Q to $X \in \mathcal{X}$ to obtain $Y \in \mathcal{Y}$ and then applying W to Y to obtain $Z \in \mathcal{Z}$.

$$\begin{aligned} U\left(QW\right) &= \sum_{\mathcal{Z}}\mu\left(\left(QW\right)_z\right) \\ &= \sum_{\mathcal{Z}}\mu\left(\sum_{\mathcal{Y}}Q_yW_{y,z}\right) \\ &\leq \sum_{\mathcal{Y},\mathcal{Z}}W_{y,z}\mu\left(Q_y\right) \\ &= \sum_{\mathcal{Y}}\mu(Q_y) \\ &= U\left(Q\right), \end{aligned}$$

where the inequality follows from sublinearity and the second to last equality follows from the row stochastic property of W. Therefore, $U\left(Q\right)$ obeys the data processing inequality.

APPENDIX C

PROOFS FOR MULTI-PARTY DIFFERENTIAL PRIVACY

C.1 Proof of Main Result

To prove Theorem 4.4.1, it is sufficient to prove Theorem C.1.1 stating that any other protocol can be simulated from the randomized response outputs. Let $\{x_i\}_{i\in[k]}$ and $\tau_{\rm RR} = \{\tilde{x}_i\}_{i\in[k]}$ denote the k private bits and transcript under the randomized response $P_{\rm RR}$ (Equation (4.8)), respectively. We will prove that any differentially private multi-party protocol can be simulated from $\tau_{\rm RR}$. This proves the desired theorem, since the optimal protocol and the optimal decision rules can be simulated by each node (and the central observer) upon receiving the randomized responses. Hence, proving that randomized response is sufficient to achieve optimal performance (on any metric).

Theorem C.1.1 *For any protocol P that generates a random transcript τ, there exists a stochastic transformation T such that the joint distribution of the bits and the transcript can be simulated from the randomized outputs:*

$$(x_1,\ldots,x_k,\tau) \overset{D}{=} (x_1,\ldots,x_k,T(\tilde{x}_1,\ldots,\tilde{x}_k)) \,, \tag{C.1}$$

where $\overset{D}{=}$ denotes equality in distribution, and $\tilde{x}_i$ is a randomized response of x_i.

To prove the above theorem, our strategy is to apply an induction argument over a class of stochastic transformations $\{T_1, T_2, \cdots, T_k\}$, where T_ℓ operates on $\tilde{x}_1^\ell = (\tilde{x}_1,\ldots,\tilde{x}_\ell)$ and $x_{\ell+1}^k = (x_{\ell+1},\ldots,x_k)$. We will prove the following series of equations:

$$\begin{aligned} (x_1,\ldots,x_k,\tau) &\overset{D}{=} (x_1,\ldots,x_k,T_1(\tilde{x}_1,x_2^k)) && \text{(C.2)} \\ &\overset{D}{=} (x_1,\ldots,x_k,T_2(\tilde{x}_1^2,x_3^k)) && \text{(C.3)} \\ &\vdots \\ &\overset{D}{=} (x_1,\ldots,x_k,T_k(\tilde{x}_1^k))\ . && \text{(C.4)} \end{aligned}$$

We first prove Equation (C.2). To do so, we show an equivalent version of this equation, which is $(x_1,\tau) \overset{D}{=} (x_1, T(\tilde{x}_1,x_2^k))$ for all fixed values of x_2^k. Equation (C.2) follows by applying Bayes rule to this equation. First, note that for all fixed x_2^k,

$$\mathcal{R}\big(P, x_1=0, x_1=1\big) \ \subseteq\ \mathcal{R}(\varepsilon_1,\delta_1)\ , \tag{C.5}$$

by the fact that τ is (ε_1,δ_1)-differentially private and Lemma 4.3.1. Next, notice that by construction, the randomized response achieves this outer bound, i.e.

$$\mathcal{R}\big(P_{\mathrm{RR}}, x_1=0, x_1=1\big) \ =\ \mathcal{R}(\varepsilon_1,\delta_1)\ , \tag{C.6}$$

for all values of x_2^k which holds only under the current assumption that x_1^k are independent. Hence from the reverse data processing inequality in Corollary 4.3.3, it follows that for each instance of x_2^k, there exists a stochastic transformation such that τ is simulated from $\tilde{x}_1$, i.e. $(x_1,\tau) \overset{D}{=} (x_1, T(\tilde{x}_1,x_2^k))$. This proves the desired Equation (C.2).

We now prove an inductive step that allows us to recursively show Equations (C.3) and (C.4). We want to prove that there always exists a stochastic transformation $T_{\ell+1}$ such that

$$(x_1^k, T_\ell(\tilde{x}_1^\ell, x_{\ell+1}^k)) \ \overset{D}{=}\ (x_1^k, T_{\ell+1}(\tilde{x}_1^{\ell+1}, x_{\ell+2}^k))\ , \tag{C.7}$$

for any stochastic transformation T_ℓ satisfying $(\varepsilon_{\ell+1},\delta_{\ell+1})$-differential privacy. Again, we prove that $(x_{\ell+1}, T_\ell(\tilde{x}_1^\ell, x_{\ell+1}^k)) \overset{D}{=} (x_{\ell+1}, T_{\ell+1}(\tilde{x}_1^{\ell+1}, x_{\ell+2}^k))$ for all values of $(x_1^\ell, \tilde{x}_1^\ell, x_{\ell+1}^k)$. Then, Equation (C.7) follows from Bayes rule. First note that from the assumption that $T_\ell(\tilde{x}_1^\ell, x_{\ell+1}^k)$ is $(\varepsilon_{\ell+1},\delta_{\ell+1})$-differentially private with respect to $x_{\ell+1}$, we know that for any fixed values of $(x_1^\ell, \tilde{x}_1^\ell, x_{\ell+2}^k)$, binary hypothesis testing on $x_{\ell+1}$ based on the observation $T_\ell(\tilde{x}_1^\ell, x_{\ell+1}^k)$ must obey the differential privacy constraint:

$$\mathbb{P}(T_\ell(\tilde{x}_1^\ell, x_{\ell+1}^k) \in S | x_{\ell+1}, x_1^\ell, \tilde{x}_1^\ell, x_{\ell+2}^k) \leq \\ e^{\varepsilon_{\ell+1}} \mathbb{P}(T_\ell(\tilde{x}_1^\ell, x_{\ell+1}^k) \in S | \overline{x_{\ell+1}}, x_1^\ell, \tilde{x}_1^\ell, x_{\ell+2}^k) + \delta_{\ell+1} \ ,$$

and since $T_\ell(\tilde{x}_1^\ell, x_{\ell+1}^k)$ is conditionally independent of x_1^ℓ given $\tilde{x}_1^\ell$, we get

$$\mathbb{P}(T_\ell(\tilde{x}_1^\ell, x_{\ell+1}^k) \in S | x_{\ell+1}, \tilde{x}_1^\ell, x_{\ell+2}^k) \leq \\ e^{\varepsilon_{\ell+1}} \mathbb{P}(T_\ell(\tilde{x}_1^\ell, x_{\ell+1}^k) \in S | \overline{x_{\ell+1}}, \tilde{x}_1^\ell, x_{\ell+2}^k) + \delta_{\ell+1} \ .$$

This implies that for each value of $(\tilde{x}_1^\ell, x_{\ell+2}^k)$,

$$\mathcal{R}\big(T_\ell, x_{\ell+1} = 0, x_{\ell+1} = 1\big) \ \subseteq \ \mathcal{R}(\varepsilon_{\ell+1}, \delta_{\ell+1}) \ .$$

Next, notice that by construction, the randomized response achieves this outer bound, i.e.

$$\mathcal{R}\big(P_{\mathrm{RR}}, x_{\ell+1} = 0, x_{\ell+1} = 1\big) \ = \ \mathcal{R}(\varepsilon_{\ell+1}, \delta_{\ell+1}) \ , \tag{C.8}$$

for all values of $(\tilde{x}_1^\ell, x_{\ell+2}^k)$ which holds only under the current assumption that x_1^k are independent. Hence from the reverse data processing inequality in Corollary 4.3.3, it follows that for each instance of $(\tilde{x}_1^\ell, x_{\ell+2}^k)$, there exists a stochastic transformation such that T_ℓ is simulated from $\tilde{x}_{\ell+1}$, i.e. $(x_{\ell+1}, T_\ell(\tilde{x}_1^\ell, x_{\ell+1}^k)) \overset{D}{=} (x_{\ell+1}, T_{\ell+1}(\tilde{x}_{\ell+1}, \tilde{x}_1^\ell, x_{\ell+2}^k))$. This proves the desired induction step in Equation (C.7). Consequently, by induction Equation (C.4) holds, and this proves Theorem C.1.1.

C.2 Proof of Optimal Multi-Party XOR Computation

Recall that $\lambda = e^\varepsilon$. Let $\tilde{X}$ denote the random output of the randomized response, and let $f(\tilde{X})$ denote the XOR of all k bits. Notice that $P(X, \tilde{X}) = (\lambda^{k - d_h(X, \tilde{X})})/(1+\lambda)^k$ where $d_h(\cdot, \cdot)$ denotes the Hamming distance. For a given $\tilde{X}$ the decision is either $f(\tilde{X})$ or the complement of it. We will first show that $f(\tilde{X})$ is the optimal decision rule.

It is sufficient to show that $\mathbb{E}[w(f(X), f(\tilde{X}))|\tilde{X}] \geq \mathbb{E}[w(f(X), \bar{f}(\tilde{X}))|\tilde{X}]$.

Since, $\mathbb{E}[w(f(X),f(\tilde{X}))|\tilde{X}] = \sum_{i \text{ even}} \binom{k}{i} \lambda^{k-i}/(1+\lambda)^k$ and $\mathbb{E}[w(f(X),\bar{f}(\tilde{X}))|\tilde{X}] = \sum_{i \text{ odd}} \binom{k}{i} \lambda^{k-i}/(1+\lambda)^k$, it follows that

$$\mathbb{E}[w(f(X),f(\tilde{X}))|\tilde{X}] - \mathbb{E}[w(f(X),\bar{f}(\tilde{X}))|\tilde{X}] = (\lambda-1)^k/(1+\lambda)^k \geq 0 \ ,$$

since $\lambda \geq 1$. By symmetry, the decision rule is the same for all $\tilde{X}$, and also for the worst case accuracy. This finishes the desired characterization of the optimal accuracy.

To get the asymptotic analysis of the accuracy, notice that $\mathbb{E}[w(f(X),f(\tilde{X}))] + \mathbb{E}[w(f(X),\bar{f}(\tilde{X}))] = 1$ and $\mathbb{E}[w(f(X),f(\tilde{X}))] + \mathbb{E}[w(f(X),\bar{f}(\tilde{X}))] = (\lambda - 1)^k/(1+\lambda)^k = (e^\varepsilon - 1)^k/(2+(e^\varepsilon-1))^k = (1/2)^k\varepsilon^k + O(\varepsilon^{k+1})$. It follows that $\mathbb{E}[w(f(X),f(\tilde{X}))] = 1/2 + (1/2)^{k+1}\varepsilon^k + O(\varepsilon^{k+1})$.

REFERENCES

[1] L. Sweeney, “Guaranteeing anonymity when sharing medical data, the datafly system.” in *Proceedings of the AMIA Annual Fall Symposium.* American Medical Informatics Association, 1997, p. 51.

[2] L. Sweeney, “Achieving k-anonymity privacy protection using generalization and suppression,” *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 571–588, 2002.

[3] A. Narayanan and V. Shmatikov, “Robust de-anonymization of large sparse datasets,” in *Security and Privacy, 2008. SP 2008. IEEE Symposium on.* IEEE, 2008, pp. 111–125.

[4] A. Narayanan, E. Shi, and B. I. Rubinstein, “Link prediction by de-anonymization: How we won the kaggle social network challenge,” in *Neural Networks (IJCNN), The 2011 International Joint Conference on.* IEEE, 2011, pp. 1825–1834.

[5] M. Gymrek, A. L. McGuire, D. Golan, E. Halperin, and Y. Erlich, “Identifying personal genomes by surname inference,” *Science*, vol. 339, no. 6117, pp. 321–324, 2013.

[6] C. Dwork, “Differential privacy,” in *Automata, Languages and Programming.* Springer, 2006, pp. 1–12.

[7] C. Dwork, F. McSherry, K. Nissim, and A. Smith, “Calibrating noise to sensitivity in private data analysis,” in *Theory of Cryptography.* Springer, 2006, pp. 265–284.

[8] C. Dwork and J. Lei, “Differential privacy and robust statistics,” in *Proceedings of the 41st Annual ACM Symposium on Theory of Computing.* ACM, 2009, pp. 371–380.

[9] 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.* Springer, 2006, pp. 486–503.

[10] L. Wasserman and S. Zhou, “A statistical framework for differential privacy,” *Journal of the American Statistical Association*, vol. 105, no. 489, pp. 375–389, 2010.

[11] D. Blackwell, "Equivalent comparisons of experiments," *The Annals of Mathematical Statistics*, vol. 24, no. 2, pp. 265–272, 1953.

[12] Q. Geng and P. Viswanath, "The optimal mechanism in differential privacy," *arXiv preprint arXiv:1212.1186*, 2012.

[13] Q. Geng and P. Viswanath, "The optimal mechanism in (ϵ,δ)-differential privacy," *arXiv preprint arXiv:1305.1330*, 2013.

[14] Q. Geng, P. Kairouz, S. Oh, and P. Viswanath, "The staircase mechanism in differential privacy," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 9, no. 7, pp. 1176–1184, 2015.

[15] C. Dwork, "Differential privacy: A survey of results," in *Theory and applications of models of computation.* Springer, 2008, pp. 1–19.

[16] C. Dwork, G. N. Rothblum, and S. Vadhan, "Boosting and differential privacy," in *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on.* IEEE, 2010, pp. 51–60.

[17] J. Blocki, A. Blum, A. Datta, and O. Sheffet, "The Johnson-Lindenstrauss transform itself preserves differential privacy," in *Foundations of Computer Science, 2012 IEEE 53rd Annual Symposium on.* IEEE, 2012, pp. 410–419.

[18] M. Hardt and A. Roth, "Beyond worst-case analysis in private singular vector computation," in *Proceedings of the 45th Annual ACM Symposium on Symposium on Theory of Computing.* ACM, 2013, pp. 331–340.

[19] A. Acquisti, "Privacy in electronic commerce and the economics of immediate gratification," in *Proceedings of the 5th ACM Conference on Electronic Commerce.* ACM, 2004, pp. 21–29.

[20] A. Acquisti and J. Grossklags, "What can behavioral economics teach us about privacy," *Digital Privacy*, p. 329, 2007.

[21] S. L. Warner, "Randomized response: A survey technique for eliminating evasive answer bias," *Journal of the American Statistical Association*, vol. 60, no. 309, pp. 63–69, 1965.

[22] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy and statistical minimax rates," in *Foundations of Computer Science, 2013 IEEE 54th Annual Symposium on.* IEEE, 2013, pp. 429–438.

[23] A. B. Tsybakov and V. Zaiats, *Introduction to Nonparametric Estimation.* Springer, 2009, vol. 11.

[24] S. Oh and P. Viswanath, "The composition theorem for differential privacy," *CoRR*, vol. abs/1311.0776, 2013. [Online]. Available: http://arxiv.org/abs/1311.0776

[25] R. F. Barber and J. C. Duchi, "Privacy and statistical risk: Formalisms and minimax bounds," *arXiv preprint arXiv:1412.4451*, 2014.

[26] K. Chatzikokolakis, T. Chothia, and A. Guha, "Statistical measurement of information leakage," in *Tools and Algorithms for the Construction and Analysis of Systems.* Springer, 2010, pp. 390–404.

[27] L. Sankar, S. R. Rajagopalan, and H. V. Poor, "Utility-privacy tradeoffs in databases: An information-theoretic approach," *Information Forensics and Security, IEEE Transactions on*, vol. 8, no. 6, pp. 838–852, 2013.

[28] W. Wang, L. Ying, and J. Zhang, "On the relation between identifiability, differential privacy and mutual-information privacy," *arXiv preprint arXiv:1402.3757*, 2014.

[29] F. McSherry and K. Talwar, "Mechanism design via differential privacy," in *Foundations of Computer Science, 2007. 48th Annual IEEE Symposium on.* IEEE, 2007, pp. 94–103.

[30] M. Hardt and G. N. Rothblum, "A multiplicative weights mechanism for privacy-preserving data analysis," in *Foundations of Computer Science, 2010 51st Annual IEEE Symposium on.* IEEE, 2010, pp. 61–70.

[31] K. Chaudhuri, A. D. Sarwate, and K. Sinha, "Near-optimal differentially private principal components," in *NIPS*, 2012, pp. 998–1006.

[32] M. Hardt and A. Roth, "Beating randomized response on incoherent matrices," in *Proceedings of the 44th Symposium on Theory of Computing.* ACM, 2012, pp. 1255–1268.

[33] M. Kapralov and K. Talwar, "On differentially private low rank approximation," in *SODA*, vol. 5. SIAM, 2013, p. 1.

[34] M. Hardt and K. Talwar, "On the geometry of differential privacy," in *Proceedings of the 42nd ACM Symposium on Theory of Computing.* ACM, 2010, pp. 705–714.

[35] M. Hardt, K. Ligett, and F. McSherry, "A simple and practical algorithm for differentially private data release," in *NIPS*, 2012, pp. 2348–2356.

[36] K. Chaudhuri and C. Monteleoni, "Privacy-preserving logistic regression," in *NIPS*, vol. 8, 2008, pp. 289–296.

[37] J. Lei, "Differentially private m-estimators," in *NIPS*, 2011, pp. 361–369.

[38] A. Ghosh, T. Roughgarden, and M. Sundararajan, “Universally utility-maximizing privacy mechanisms,” *SIAM Journal on Computing*, vol. 41, no. 6, pp. 1673–1693, 2012.

[39] T. M. Cover and J. A. Thomas, *Elements of Information Theory.* John Wiley & Sons, 2012.

[40] A. Beimel, K. Nissim, and E. Omri, “Distributed private data analysis: Simultaneously solving how and what,” in *Advances in Cryptology–CRYPTO 2008.* Springer, 2008, pp. 451–468.

[41] K. Chaudhuri and D. Hsu, “Convergence rates for differentially private statistical estimation,” *arXiv preprint arXiv:1206.6395*, 2012.

[42] A. De, “Lower bounds in differential privacy,” in *Theory of Cryptography.* Springer, 2012, pp. 321–338.

[43] P. Kairouz, S. Oh, and P. Viswanath, “Extremal mechanisms for local differential privacy,” *arXiv preprint arXiv:1407.1338*, 2014.

[44] E. A. Abbe, A. Khandani, and A. W. Lo, “Privacy-preserving methods for sharing financial risk exposures,” *The American Economic Review*, vol. 102, pp. 65–70, 2011.

[45] A. C. Yao, “Protocols for secure computations,” in *2013 IEEE 54th Annual Symposium on Foundations of Computer Science.* IEEE, 1982, pp. 160–164.

[46] M. Ben-Or, S. Goldwasser, and A. Wigderson, “Completeness theorems for non-cryptographic fault-tolerant distributed computation,” in *Proceedings of the Twentieth Annual ACM Symposium on Theory of Computing.* ACM, 1988, pp. 1–10.

[47] O. Goldreich, S. Micali, and A. Wigderson, “How to play any mental game,” in *Proceedings of the Nineteenth Annual ACM Symposium on Theory of Computing*, ser. STOC ’87. New York, NY, USA: ACM, 1987. [Online]. Available: http://doi.acm.org/10.1145/28395.28420 pp. 218–229.

[48] D. Chaum, “The dining cryptographers problem: Unconditional sender and recipient untraceability,” *Journal of Cryptology*, vol. 1, no. 1, 1988.

[49] L. Sweeney, “Simple demographics often identify people uniquely,” *Health*, vol. 671, pp. 1–34, 2000.

[50] J. A. Calandrino, A. Kilzer, A. Narayanan, E. W. Felten, and V. Shmatikov, “ “You might also like:” privacy risks of collaborative filtering,” in *Security and Privacy (SP), 2011 IEEE Symposium on.* IEEE, 2011, pp. 231–246.

[51] 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 Genetics*, vol. 4, no. 8, p. e1000167, 2008.

[52] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *Security and Privacy, 2008. SP 2008. IEEE Symposium on.* IEEE, 2008, pp. 111–125.

[53] C. Dwork, "Differential privacy: A survey of results," in *Theory and Applications of Models of Computation.* Springer, 2008, pp. 1–19.

[54] 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.* Springer, 2006, pp. 486–503.

[55] A. McGregor, I. Mironov, T. Pitassi, O. Reingold, K. Talwar, and S. Vadhan, "The limits of two-party differential privacy," in *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on.* IEEE, 2010, pp. 81–90.

[56] V. Goyal, I. Mironov, O. Pandey, and A. Sahai, "Accuracy-privacy tradeoffs for two-party differentially private protocols," in *Advances in Cryptology–CRYPTO 2013.* Springer, 2013, pp. 298–315.

[57] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith, "What can we learn privately?" *SIAM Journal on Computing*, vol. 40, no. 3, pp. 793–826, 2011.

[58] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate, "Differentially private empirical risk minimization," *The Journal of Machine Learning Research*, vol. 12, pp. 1069–1109, 2011.

[59] A. Blum, C. Dwork, F. McSherry, and K. Nissim, "Practical privacy: the SuLQ framework," in *Proceedings of the Twenty-Fourth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems.* ACM, 2005, pp. 128–138.

[60] K. Chaudhuri, A. D. Sarwate, and K. Sinha, "A near-optimal algorithm for differentially-private principal components," *Journal of Machine Learning Research*, vol. 14, pp. 2905–2943, 2013.

[61] S. L. Lauritzen, *Graphical Models.* Oxford University Press, 1996.

[62] N. Alon and J. H. Spencer, *The Probabilistic Method.* Wiley, 2004.