Quantifying Information Leakage via Adversarial Loss Functions: Theory and Practice

by

Jiachun Liao

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved April 2020 by the
Graduate Supervisory Committee:

Lalitha Sankar, Chair
Oliver Kosut
Junshan Zhang
Gautam Dasarathy

ARIZONA STATE UNIVERSITY

May 2020

ABSTRACT

Modern digital applications have significantly increased the leakage of private and sensitive personal data. While worst-case measures of leakage such as Differential Privacy (DP) provide the strongest guarantees, when utility matters, average-case information-theoretic measures can be more relevant. However, most such information-theoretic measures do not have clear operational meanings. This dissertation addresses this challenge.

This work introduces a tunable leakage measure called maximal α-leakage which quantifies the maximal gain of an adversary in inferring any function of a data set. The inferential capability of the adversary is modeled by a class of loss functions, namely, α-loss. The choice of α determines specific adversarial actions ranging from refining a belief for $\alpha = 1$ to guessing the best posterior for $\alpha = \infty$, and for the two specific values maximal α-leakage simplifies to mutual information and maximal leakage, respectively. Maximal α-leakage is proved to have a composition property and be robust to side information.

There is a fundamental disjoint between theoretical measures of information leakages and their applications in practice. This issue is addressed in the second part of this dissertation by proposing a data-driven framework for learning Censored and Fair Universal Representations (CFUR) of data. This framework is formulated as a constrained minimax optimization of the expected α-loss where the constraint ensures a measure of the usefulness of the representation. The performance of the CFUR framework with $\alpha = 1$ is evaluated on publicly accessible data sets; it is shown that multiple sensitive features can be effectively censored to achieve group fairness via demographic parity while ensuring accuracy for several *a priori* unknown downstream tasks.

Finally, focusing on worst-case measures, novel information-theoretic tools are

used to refine the existing relationship between two such measures, (ϵ, δ)-DP and Rényi-DP. Applying these tools to the moments accountant framework, one can track the privacy guarantee achieved by adding Gaussian noise to Stochastic Gradient Descent (SGD) algorithms. Relative to state-of-the-art, for the same privacy budget, this method allows about 100 more SGD rounds for training deep learning models.

ACKNOWLEDGEMENTS

I would like to thank my advisor Dr. Lalitha Sankar for supporting me to successfully study and do research at Arizona State University (ASU) and bringing the interesting topic of data privacy to me. Without her guiding, I might not have the chance to attend high-level international conferences on information theory and learn from fantastic researchers all over the world. I would like to acknowledge support for my research projects from the National Science Foundation via grants CCF-1422358, CCF-1350914, CIF-1815361 and CIF-1901243.

I also would like to thank Dr. Oliver Kosut, Dr. Flavor Clamon (of Harvard University) and Dr. Vincent Tan (of National University of Singapore) for their precise and efficient instructions on my research for joint projects. I appreciate my other two committee members Dr. Junshan Zhang and Dr.Gautam Dasarathy for their insightful questions and meaningful suggestions, and appreciate my collaborators Dr. Shahab Asoodeh (of Harvard), Dr. Peter Kairouz (of Google), Maunil Vyas and Mit Patel for their help and kindness. Let me also say thanks to all the ASU staffs who have helped me with various problems during my Ph.D. life.

Last but not least, I would like to thank my dear parents for trying their best to support all my decisions, and thank my lovely friends and labmates: Jiazi Zhang, Ruolei Ji, Zhigang Chu, Chong Huang, Mario Diaz and Tyler Sypherd for their valuable friendship. Without their encouragement and accompany, I would not be able to survive from all difficulties.

TABLE OF CONTENTS

CHAPTER Page

CHAPTER Page

CHAPTER Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Chapter 1

INTRODUCTION

1.1 Background and Contributions

The use of deep learning algorithms for data analytics has recently seen unprecedented success for a variety of problems such as image classification, natural language processing, and prediction of consumer behavior, electricity use, political preferences, to name a few. The success of these algorithms hinges on the availability of large data sets that may include patterns of societal bias and discrimination as well as sensitive personal information. It has been shown that models learned from such data sets can potentially inherit such biases [1, 2] as well as glean sensitive features even when such features are not explicitly used during training [3]. As a result, concerns about the fairness, censoring, and privacy of learning algorithms has led to a growing body of research focused on both defining meaningful measures and designing learning models with a desired fairness/censoring/privacy guarantee.

The measure and control of private/sensitive information leakage is a recognized objective in communications, information theory, and computer science. Modern cryptography [4, 5, 6], for example, aims at designing and analyzing security systems that are believed to be impervious to computationally bounded adversaries. Alternatively, information-theoretic security studies settings where an asymmetry of information between an adversary and the legitimate parties (e.g., the wiretap channel [7, 8, 9]) can be exploited to guarantee that no private information is leaked regardless of computational assumptions. An adversary that *only* observes the output of a (computationally) secure cipher or cannot overcome the information asymmetry in a

wiretap-like setting does not, for all practical purposes, pose a privacy risk.

However, modern applications such as online data sharing, social networks, cloud-based services, and mobile computing have significantly increased the number of interfaces through which sensitive/private information can leak. Services that require a user to disclose data in order to receive utility inevitably incur a privacy risk through unwanted inferences. For example, political preference can be reliably estimated from movie ratings [10], an online store can infer a medical condition by observing your shopping history [11], or social network users can be deanonymized by tracking their interaction with peers [12, 13]. Moreover, practical implementations of cryptographic schemes are susceptible to so-called "side-channel attacks," where sensitive information leaks through unexpected channels. For example, a malicious application may get timing characteristics [14, 15]. In these examples, an adversary that observes information leaking through a side channel can more reliably infer sensitive data, such as a key or a plaintext.

Despite the array of (often overlapping) privacy/censoring, various measures are proposed over the past decade. The most well-known measure is differential privacy (DP) [16, 17], which captures privacy in the context of querying databases. Recently, targeting a guessing adversary, Asoodeh et al. use the probability of correctly guessing to measure censoring [18]; and Issa et al. introduce maximal leakage (MaxL), which is essentially the maximal logarithmic gain in the probability of correctly guessing any arbitrary function of original data from released data [19]. Many other measures of information leakages are derived from information theory, such as mutual information (MI). Rassouli et al. use a total variation distance between the prior and posterior distributions as the leakage measure [20], and Mironov introduce Rényi differential privacy based on Rényi divergence [21].

In this report, we provide two new metrics called *α-leakage* and *maximal α-leakage*

that quantify information leakages through the lens of adversarial inference capabilities. Specifically, α-leakage captures an adversary's ability in inferring a *specific* sensitive attribute in the data set, and in contrast, maximal α-leakage is for any *arbitrary* attribute of the data set. These metrics can be applied to the aforementioned privacy/censoring and side-channel settings, and directly capture an adversary's ability to infer (ranging from the most likely realization to the posterior distribution) for *any information* of original data from the released version or the one leaked via side-channels. We show that α-leakage is Arimoto mutual information (A-MI) and maximal α-leakage is MI for $\alpha = 1$ and Arimoto channel capacity for $\alpha > 1$. Therefore, the proposed maximal α-leakage captures MI and MaxL at extrema.

In privacy protection, one essential problem is that an adversary's side information could increase the amount of private information leaked to this adversary from released data. One of the advantages of DP is that it is robust to arbitrary external knowledge (side information). This robustness is formalized in [22], wherein the authors model side information by a prior probability distribution on the support of the original data set. Differently, in this work we model side information as a random variable possessed by an adversary that is interested to learn an arbitrary function of the original data from the released data. An adversarial inference involving side information is, therefore, modeled as a conditional Markov chain, which is also used by Issa *et al.* to study the effect of side information on maximal leakage [23, Def. 6]. We justify the reasonability of the conditional Markov chain in exploring impacts of side information on privacy problems. Making use of the conditional Markov chain, we introduce conditional maximal α-leakage, which is an extended version of maximal α-leakage involving side information, and show that maximal α-leakage upper bounds maximal conditional α-leakage if the side information is conditionally independent of the released data given the original data. That is, maximal α-leakage is robust to

arbitrary side information that is not used in generating the released data from the original data.

In this work, we also evaluate the proposed measures of information leakage by using them as privacy metrics in privacy-guaranteed data publishing settings. In most non-trivial settings of data publishing, there is a fundamental tradeoff between privacy and utility, called privacy-utility tradeoff (PUT): on the one hand, releasing data "as is" can lead to unwanted inferences of private information (an arbitrary or a specific function of original data). On the other hand, perturbing or limiting the released data reduces its quality. We concern general statistical inference applications and guarantee a general utility via preserving the fidelity of the released data to the original data. The fidelity is measured by an arbitrarily chosen distortion function. In contrast to statistical utilities in most information-theoretic PUTs, which capture utility as a *statistical average* of desired measures [24, 25, 26, 18, 20], we introduce a new *hard distortion* metric to measure utility, which constrains the privacy mechanism so that the distortion between original and released data is bounded with probability 1. The concept of deterministic/hard utility has been considered in the form of ρ-recoverable functions in [27]. Differently, we bound the distortion of data itself instead of data functions, which naturally guarantees some recoverability of any arbitrary data functions. In addition, compared to average-case distortion constraints [28], the hard distortion metric is quite stringent but allows the data curator to make specific, deterministic guarantees on the fidelity of the released data set to the original one. The deterministic guarantee can lead to more accurate statistical estimators, e.g., the empirical distribution estimation. Using the aforementioned tunable measures of information leakage and hard distortion as privacy and utility measures, respectively, we precisely quantify the PUT for data sets that are entirely sensitive or contain both non-sensitive and sensitive private data, respectively.

While the fundamental question of how to formally define algorithmic fairness continues to be open, most algorithmic fairness measures have been motivated by legal systems that evaluate the fairness of a decision-making process using two distinct notions [29]: *disparate treatment* and *disparate impact*. The process of making decisions suffers from disparate treatment if the decisions are (partly) based on the subject's sensitive information/attribute, and it has disparate impact if its outcomes disproportionately hurt (or benefit) people with certain sensitive attribute values. These two legal definitions have led to many distinct interpretations of algorithmic fairness, and therefore, to many quantitative measures. Furthermore, whether fairness should be enforced at a group or individual level has also led to different quantitative definitions (see [30, 31, 32] and the references therein) and two broad approaches: group fairness and individual fairness. Group fairness ensures statistical/demographic parity by seeking similar outcomes for all groups [33]. In contrast, individual fairness requires treating similar individuals, perceived as such in some measurable space, similarly [34, 35].

For specific learning tasks, fairness guarantees can be achieved either via pre-processing, or in-processing, or post-processing the data. In-processing approaches are most commonly used in the supervised setting where the learning objective (e.g., target labels for classification) are known, and have been explored in the context of classification [34, 36, 37, 38, 39], regression [40, 41] and ranking [42, 43, 44, 45, 46, 47]. In this setting, knowledge of the learning objective/task is required in the training phrase and the resulting trained model gives fair results only for the specified learning objective. Therefore, in-processing approaches are not applicable for data sets with limited or no labels. Pre-processing approaches generally produce fair representations of data at hand and post-processing approaches provide fairness by properly altering decision outputs [48, 49, 50, 51]. Both these two approaches do not require the

knowledge of learning objectives in the training phase.

In a variety of data collection settings, the learning tasks may not be known *a priori* or the data may be collected for learning multiple tasks. In this context, the pre-processing approach of learning fair representations of data sets is effective. Learning fair representations using information-theoretic objective functions and constrained optimization have been proposed in [48, 52, 53]. However, these approaches, focusing on information-theoretic formulations, require knowledge of the statistics of the data sets. To circumvent the lack of statistical knowledge for real data sets, *data-driven approach* have been considered wherein fair representations are learned directly from the data via adversarial models. Adversarial learning models have been developed [54, 55, 56] and applied to semi-supervised learning [57, 58], domain adaptation [59] and segmentation [60]. Recently, such methods have also been applied to context-aware censoring and fairness [61, 62, 63, 64, 65, 66, 67, 37].

In this work, we apply the α-loss to the generative adversary model proposed in [64] and develop a framework that outputs a censored and fair universal representation (CFUR) of the data at hand. The resulting representation can then be used to learn a variety of task-appropriate models that are information-theoretically guaranteed to be fair. These representations are universally fair in that the representations can be used for a variety of downstream learning tasks and the fairness guarantees are independent of such tasks. Since fairness assurances for CFURs are achieved via pre-processing the data, a reasonably appropriate measure of fairness is the group fairness measure of demographic parity (DemP) which ensures the same proportion of outcomes to all groups. As an immediate consequence of such pre-processing, the CFUR problem becomes a tradeoff between ensuring sufficient fidelity of the representation to ensure high accuracy on downstream learning tasks while guaranteeing a desired measure of demographic parity.

The key idea of our CFUR approach is that of actively decorrelating the sensitive data from the other features, thereby ensuring that the downstream tasks are learned independent of such sensitive features. As a result, the CFUR model itself serves broader purpose: it can be used to create decorrelated representations for new data (from the same distribution or via transfer learning for other data sets) prior to learning any task on it. One can view this effort as a *censoring* approach, i.e., all sensitive features are not just stripped from the data but its correlation with other features are actively damped to censor/restrict its inference from the representation while ensuring some measure of usefulness of the data.

More generally, when learning data representations, censoring and fairness are very similar in that both can be ensured by perturbing the data set to decorrelate the sensitive variables from the rest of the data set. Depending on the context and problem requirement, decorrelating operations can be designed as privacy preserving mechanisms to hide the sensitive variables from inference or as a fairness enforcing algorithm that prevents a machine learning model from discriminating based on the sensitive variables. We now detail our specific contributions to designing fair universal representations via a censoring approach.

To showcase the power of our approach, we conduct 2 sets of extensive experiments on publicly accessible data sets: UCI Adult [68] and UTKFace [69]. For relevant data sets, our visual results show that our data driven training methods succeed at creating high quality representations that increasingly erase the sensitive attributes with decreasing fidelity requirements. Both our theoretical framework and our experiments consider non-binary sensitive attributes and data sets with multiple attributes. Our experimental results show that one can still learn high quality classifiers even when the downstream ML task is not known *a priori*. In particular, we consider the UCI data set that is often used in fair ML analyses to showcase the advantage of our

approach relative to related approaches (for example, by Edwards and Storkey [61] and Madras *et al.* [67]). Moreover, our results straddle a wide range of values for the chosen fairness measure (DemP) including perfect fairness, in contrast to the above cited works.

Several methods have recently been proposed to ensure differentially private training of ML models [70, 71, 72, 73, 74, 75]. Here, the parameters of the model determined by a learning algorithm (e.g., weights of a neural network) are sought to be differentially private with respect to the data used for fitting the model (i.e. the *training* data). When the model parameters are computed by applying stochastic gradient descent (SGD) algorithm to minimize a given loss function, DP can be ensured by directly adding noise to the gradient. The empirical and theoretical flexibility of this noise-adding procedure for ensuring DP was demonstrated, for example, in [71, 70]. This method is currently being used for privacy-preserving training of large-scale ML models in industry, see e.g., the implementation of [76] in the Google's open-source TensorFlow Privacy framework [77].

Not surprisingly, for a fixed training data set, privacy deteriorates with each SGD iteration. In practice, the DP constraints are set *a priori*, and then mapped to a permissible number of SGD iterations for fitting the model parameters. Thus, a key question is: *given a DP constraint, how many iterations are allowed before the SGD algorithm is no longer private?* The main challenge in determining the DP guarantees provided by noise-added SGD is keeping track of the evolution of the privacy loss random variable during subsequent gradient descent iterations. This can be done, for example, by invoking advanced composition theorems for DP, such as [78, 79]. Such composition results, while theoretically significant, may be difficult to apply to the SGD setting due to their generality (e.g., they do not take into account the noise distribution used by the privacy mechanism).

Recently, Abadi et al. [70] circumvented the use of DP composition results by developing a method called *moments accountant* (MA). Instead of dealing with DP directly, the MA approach provides privacy guarantees in terms of *Rényi differential privacy* (RDP) [80] for which composition has a simple linear form. Once the privacy guarantees of the SGD execution are determined in terms of RDP, they are mapped back to DP guarantees via a conversion result between (ϵ, δ)-DP and RDP [70, Theorem 2]. This approach renders tighter DP guarantees than those obtained from advanced composition theorems (see [70, Figure 2]). We provide a framework which settles the *optimal* conversion from RDP to (ϵ, δ)-DP, and thus further enhances the privacy guarantee obtained by the MA approach. Our technique relies on the information-theoretic study of joint range of f-divergences [81, 82]. It is known that both (ϵ, δ)-DP and RDP can be expressed via two certain types of the f-divergences [83, 84], namely *hockey-stick* [85] divergence and χ^α-divergence (also called *Hellinger divergence*[86]), respectively. Based on this result, we apply [87, Theorem 8] to characterize the joint range of these two f-divergences which, in turn, leads to the "optimal" conversion from RDP to DP . Specifically, this optimal conversion allows us to derive bounds on the number of SGD iterations for a given DP constraint in the context of Gaussian perturbation of the gradient. Our result improves upon the state-of-the-art [70] by allowing more training iterations (often hundreds more) for the same privacy budget, and thus providing higher utility for free.

1.2 Outline of the Report

The outline of this paper is as follows. In Chapter 2, two operational measures of information leakage: α-leakage and maximal α-leakage, are introduced based on a novel tunable loss function, namely α-loss ($\alpha > 0$). The properties of the loss function and two leakage measures are also presented in Chapter 2. Chapter 3 shows

the robustness of maximal α-leakage to arbitrary side information. Chapter 4 gives the optimal privacy mechanisms that achieves the optimal PUTs using either maximal α-leakage or α-leakage as the privacy measure subject to a hard distortion constraint. In Chapter 5, the α-loss is applied to a generative adversary model and the resulting framework is proved to be able to produce censored and fair representations of data for multiple downstream tasks not known *a priori*. The performance of the framework is evaluated on publicly accessible data sets: UCI Adult and UTK Face. In Chapter 6, an information-theoretical study of the joint region of two f-divergences is used to derived the optimal conversion from RDP to (ϵ, δ)-DP, which can be used to improve the tracking of privacy protection provided by a noisy gradient descent algorithm in machine learning and deep learning. The conclusion and future work are in Chapter 7.

Chapter 2

TUNABLE INFORMATION LEAKAGE MEASURES AND PROPERTIES

In this chapter, we introduce a tunable loss function, namely *α-loss* for $\alpha \in (0, \infty]$, which simplifies to the logarithmic loss (log-loss) and soft 0-1 loss at the two extrema. Viewing information leakage through the lens of adversarial inference capabilities, we quantify the leakage via α-loss, which the adversary intends to minimize, and define two tunable measures of information leakage, called *α-leakage* and *maximal α-leakage*, respectively. Note that for the sake of concise expressions, we first present the definitions and properties of α-loss and (maximal) α-leakage for $\alpha \in [1, \infty]$ and summarize the extension to $0 < \alpha < 1$ in Section 2.5.

2.1 Preliminaries

We use capital letters to represent *discrete* random variables, and the corresponding capital calligraphic and lower-case letters represent their *finite* supports and the elements of the supports, respectively. For example, for a random variable X, its support is $\mathcal{X}$ with any possible realization $x \in \mathcal{X}$. In addition, we use log to represent the natural logarithm, and $[a, b]$ to indicate the set of integers from a to b. We use $|\cdot|$ to indicate the cardinality of a set, e.g., $|\mathcal{X}|$, and $\|\cdot\|_p$ to represent the p-norm of a vector, e.g., for $\alpha \geq 1$, $\|P_X\|_\alpha \triangleq (\sum_{x \in \mathcal{X}} P_X(x)^\alpha)^{\frac{1}{\alpha}}$.

We begin by reviewing Rényi entropy and divergence [88, 89].

Definition 2.1.1. *Given a distribution P_X, the Rényi entropy of order $\alpha \in (0, 1) \cup$*

$(1, \infty)$ *is defined as*

$$H_\alpha(P_X) = \frac{1}{1-\alpha} \log \sum_{x \in \mathcal{X}} P_X(x)^\alpha, \tag{2.1}$$

$$= \frac{\alpha}{1-\alpha} \log \|P_X\|_\alpha, \quad (\alpha \geq 1). \tag{2.2}$$

Let Q_X *be a distribution over the support of* P_X*. The Rényi divergence (between* P_X *and* Q_X*) of order* $\alpha \in (0,1) \cup (1,\infty)$ *is defined as*

$$D_\alpha(P_X \| Q_X) = \frac{1}{\alpha - 1} \log \left(\sum_{x \in \mathcal{X}} \frac{P_X(x)^\alpha}{Q_X(x)^{\alpha-1}} \right). \tag{2.3}$$

Both of the two quantities are defined by their continuous extensions for $\alpha = 1$ *and* ∞*. Specifically, for* $\alpha = \infty$*, the two quantities are given by*

$$H_\infty(P_X) = \min_x \log \frac{1}{P_X(x)}, \tag{2.4}$$

which is called min-entropy, and

$$D_\infty(P_X \| Q_X) = \log \max_x \frac{P_X(x)}{Q_X(x)}. \tag{2.5}$$

For $\alpha = 1$*, the Rényi entropy and divergence reduce to Shannon entropy and Kullback-Leibler divergence, respectively [90].*

The α-leakage and maximal α-leakage metrics can be expressed in terms of Sibson MI [91] and Arimoto MI [92]. These quantities generalize the usual notion of MI. We review these definitions next.

Definition 2.1.2. *Let discrete random variables* $(X, Y) \sim P_{X,Y}$ *with* P_X *and* $P_{Y|X}$ *as the marginal and conditional distributions, respectively, and* Q_Y *be an arbitrary distribution over the finite support* $\mathcal{Y}$*. The Sibson mutual information of order* $\alpha \in$

$(0,1)\cup(1,\infty)$ *is defined as*

$$I_\alpha^S(X;Y) \triangleq \inf_{Q_Y} D_\alpha(P_{X,Y}\|P_X\times Q_Y) = \frac{\alpha}{\alpha-1}\log\sum_y\left(\sum_x P_X(x)P_{Y|X}(y|x)^\alpha\right)^{\frac{1}{\alpha}}. \tag{2.6}$$

The Arimoto mutual information of order $\alpha\in(0,1)\cup(1,\infty)$ *is defined as*

$$I_\alpha^A(X;Y) \triangleq H_\alpha(X) - H_\alpha^A(X|Y) = \frac{\alpha}{\alpha-1}\log\frac{\sum_y\left(\sum_x P_{X,Y}(x,y)^\alpha\right)^{\frac{1}{\alpha}}}{\left(\sum_x P_X(x)^\alpha\right)^{\frac{1}{\alpha}}}, \tag{2.7}$$

$$= \frac{\alpha}{\alpha-1}\log\frac{\sum_y\|P_{X,Y}(\cdot,y)\|_\alpha}{\|P_X\|_\alpha}, \quad (\alpha\geq 1) \tag{2.8}$$

where $H_\alpha^A(X|Y)$ *is Arimoto conditional entropy of* X *given* Y *defined as*

$$H_\alpha^A(X|Y) = \frac{\alpha}{1-\alpha}\log\sum_y\left(\sum_x P_{X,Y}(x,y)^\alpha\right)^{\frac{1}{\alpha}}. \tag{2.9}$$

All of these quantities are defined by their continuous extension for $\alpha=1$ *or* ∞.

Note that for $\alpha=1$, both Sibson and Arimoto MIs reduce to Shannon's MI; however, for $\alpha=\infty$, the Sibson MI is

$$I_\infty^{\mathrm{S}}(X;Y) = \log\sum_y\max_x P_{Y|X}(y|x), \tag{2.10}$$

and the Arimoto MI is given by

$$I_\infty^{\mathrm{A}}(X;Y) = \log\frac{\sum_y\max_x P_{X,Y}(x,y)}{\max_x P_X(x)}. \tag{2.11}$$

The two metrics of information generalize Shannon's MI and have a number of interesting and useful properties in various problems [91, 92, 90, 93].

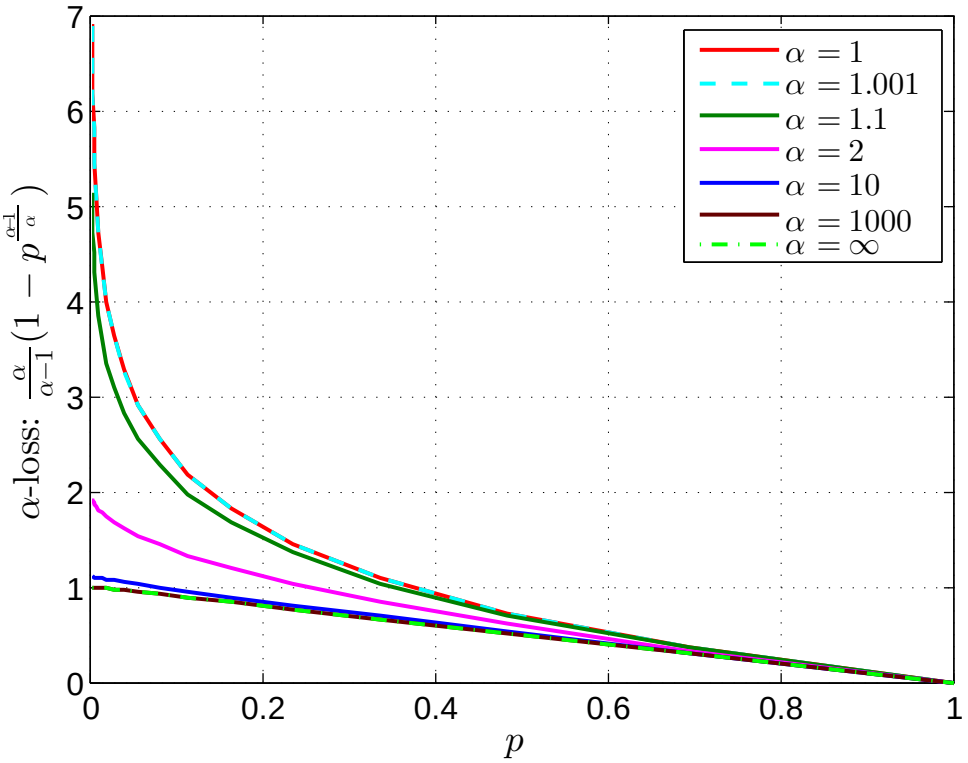


Figure 2.1: Plot of α-Loss as a Function of p

2.2 α-Loss Function

For a Markov chain $X - Y - \widehat{X}$, let $\widehat{X}$ be an estimator of X and $P_{\widehat{X}|Y}$ be a strategy for estimating X from Y. We denote the probability of correctly estimating $X = x$ given $Y = y$ as $P_{\widehat{X}|Y}(x|y)$. The estimation strategy $P_{\widehat{X}|Y}$ is selected in order to minimize an *expected loss* metric. Denoting the loss function by $\ell(x, y, P_{\widehat{X}|Y})$, the expected loss is given by $\mathbb{E}\big[\ell\big(X, Y, P_{\widehat{X}|Y}\big)\big]$.

One formulation of the loss function is the probability of *incorrectly* guessing, also called soft 0-1 loss, given by

$$\ell_{0-1}(x, y, P_{\widehat{X}|Y}) = 1 - P_{\widehat{X}|Y}(x|y), \tag{2.12}$$

such that the expected loss $\mathbb{E}\big[\ell_{0-1}\big(X, Y, P_{\widehat{X}|Y}\big)\big]$ is the expected probability of error. Here, the optimal strategy $P^{\star}_{\widehat{X}|Y}$ is the standard maximal posterior (MAP) estimator

given by

$$P^{\star}_{\widehat{X}|Y}(x|y) = \begin{cases} 1, & x = \arg\max\limits_{x\in\mathcal{X}} P_{X|Y}(x|y) \\ 0, & \text{otherwise} \end{cases}, \tag{2.13}$$

which makes the loss $\ell_{0-1}(x, y, P^{\star}_{\widehat{X}|Y})$ be either 0 or 1, and therefore, called 0-1 *loss* in the literature [94, 95]. The corresponding expected loss $\mathbb{E}\big[\ell_{0-1}\big(X, Y, P^{\star}_{\widehat{X}|Y}\big)\big]$ is the minimal expected probability of error.

To measure the uncertainty for the strategy $P_{\widehat{X}|Y}$, the log-loss (used, for example, in [94, 96, 97, 98]) is given by

$$\ell_{\log}(x, y, P_{\widehat{X}|Y}) = \log \frac{1}{P_{\widehat{X}|Y}(x|y)}. \tag{2.14}$$

The expected loss in this case is the conditional cross-entropy, given by

$$\mathbb{E}\left[\ell_{\log}(X, Y, P_{\widehat{X}|Y})\right] = \sum_{x,y} P_{X,Y}(x, y) \log \frac{1}{P_{\widehat{X}|Y}(x|y)}, \tag{2.15}$$

$$= H(X|Y) + \sum_{y} P_Y(y) D(P_{X|Y=y} \| P_{\hat{X}|Y=y}). \tag{2.16}$$

Therefore, the optimal strategy is the true posterior distribution of X given Y, i.e., $P^{\star}_{\widehat{X}|Y} = P_{X|Y}$, which makes the expected loss in (2.16) become the conditional entropy $H(X|Y)$. That is, the minimal expected log-loss is the true conditional entropy.

Note that both the soft 0-1 loss and log-loss functions are decreasing in the probability of correctly estimation $P_{\widehat{X}|Y}(x|y)$. Specifically, for $P_{\widehat{X}|Y}(x|y) = 1$, both the values of soft 0-1 loss and α-loss are 0, and for $P_{\widehat{X}|Y}(x|y) = 0$, the values of soft 0-1 loss and log-loss become 1 and ∞, respectively. To allow a continuous quantification of the loss for $P_{\widehat{X}|Y}(x|y) = 0$ from 1 to ∞, we formally define a tunable loss function, namely *α-loss*, as follows.

Definition 2.2.1 (α-loss)**.** *Let random variables X, Y and $\widehat{X}$ form a Markov chain $X - Y - \widehat{X}$, where $\widehat{X}$ is an estimator of X. The α-loss of the strategy $P_{\widehat{X}|Y}$ for*

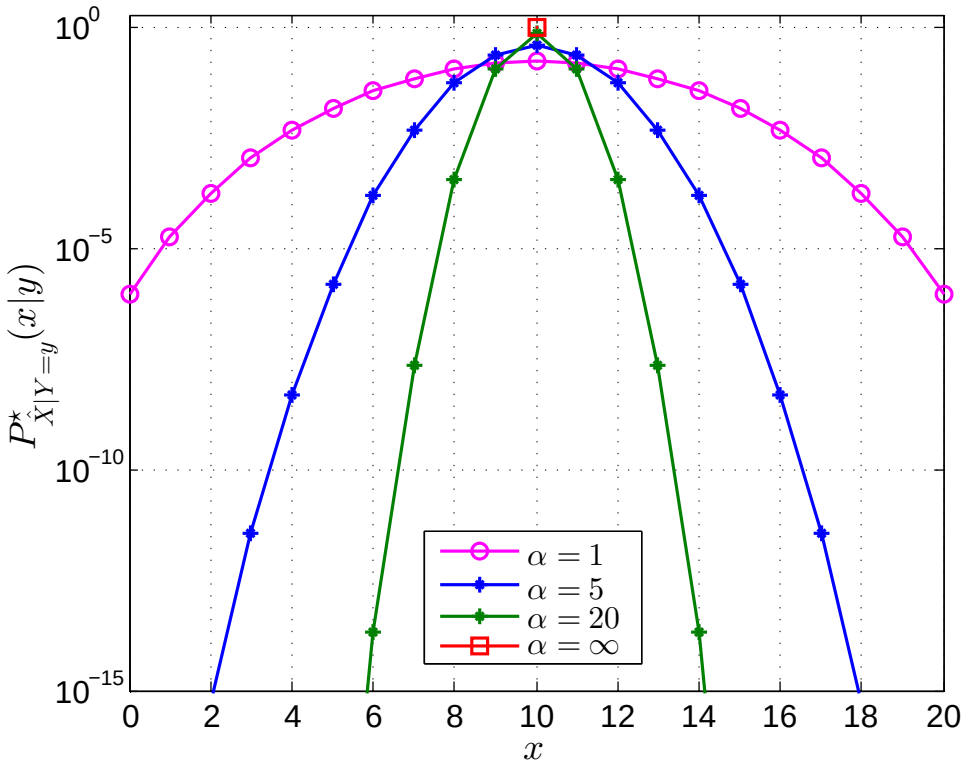


Figure 2.2: The Optimal Strategy in (2.21) for Different α

estimating $\hat{X} = x$ *from* $Y = y$ *is*

$$\ell_\alpha(x, y, P_{\widehat{X}|Y}) = \frac{\alpha}{\alpha - 1}\big(1 - P_{\widehat{X}|Y}(x|y)^{\frac{\alpha-1}{\alpha}}\big), \tag{2.17}$$

where $\alpha \in (1, \infty)$. *It is defined by its continuous extension for* $\alpha = 1$ *and* $\alpha = \infty$, *respectively, and is given by*

$$\ell_1(x, y, P_{\widehat{X}|Y}) = \lim_{\alpha \to 1} \ell_\alpha(x, y, P_{\widehat{X}|Y}) = \log \frac{1}{P_{\widehat{X}|Y}(x|y)}, \tag{2.18}$$

$$\ell_\infty(x, y, P_{\widehat{X}|Y}) = \lim_{\alpha \to \infty} \ell_\alpha(x, y, P_{\widehat{X}|Y}) = 1 - P_{\widehat{X}|Y}(x|y). \tag{2.19}$$

Note that for $\alpha = 1$, the expression in (2.18) follows directly from the L'Hôpital's rule and α-loss becomes the log-loss in (2.14); and for $\alpha = \infty$, the loss in (2.19) is exactly the soft 0-1 loss in (2.12). Fig. 2.1 plots the α-loss function in (5.20) for different values of α, where the $p \in [0.001, 1]$ represents the probability of correctly guessing, i.e., $p = P_{\widehat{X}|Y}(x|y)$ with an observation $Y = y$ and $p = P_{\widehat{X}}(x)$ without any

observation. From Fig. 2.1, we observe that α-loss function is decreasing and convex in the probability of correctly guessing.

Lemma 1. *For $1 \leq \alpha \leq \infty$, the minimal expected α-loss is given by*

$$\min_{P_{\widehat{X}|Y}} \mathbb{E}\left[\ell_\alpha(X, Y, P_{\widehat{X}|Y})\right] = \begin{cases} \frac{\alpha}{\alpha-1}\left(1 - \exp\left(\frac{1-\alpha}{\alpha} H_\alpha^A(X|Y)\right)\right), & \alpha > 1 \\ H(X|Y), & \alpha = 1 \end{cases}, \tag{2.20}$$

with the optimal estimation strategy given by

$$P^\star_{\widehat{X}|Y}(x|y) = \frac{P_{\widehat{X}|Y}(x|y)^\alpha}{\sum\limits_{x \in \mathcal{X}} P_{\widehat{X}|Y}(x|y)^\alpha}. \tag{2.21}$$

A detailed proof is in Subsection 2.6.1. Note that in (2.20), $H_\alpha^{\mathrm{A}}(X|Y)$ is Arimoto conditional entropy of X given Y in (2.9). For $\alpha = \infty$, the expression of $H_\infty^{\mathrm{A}}(X|Y)$ is

$$H_\infty^{\mathrm{A}}(X|Y) = \log \sum_y P_Y(y) \max_x P_{X|Y}(x|y), \tag{2.22}$$

such that $\exp\left(H_\infty^{\mathrm{A}}(X|Y)\right)$ is the maximal expected probability of correctly guessing X from Y. Therefore, for $\alpha = \infty$, the minimal expected α-loss is the minimal expected probability of error. In addition, the optimal estimation strategy in (2.21) becomes the true posterior distribution of X for $\alpha = 1$ and the MAP estimator for $\alpha = \infty$ [1] , respectively.

Example 1. *Let the conditional probability distribution of X given $Y = y$ be a binomial distribution with parameters $(n, p) = (20, 0.5)$, i.e., $P_{X|Y}(x|y) = \binom{20}{x} 0.5^x 0.5^{20-x}$ for $x \in [0, 20]$. Fig. 2.2 shows the optimal strategies in (2.21) for different values of α. Note that in Fig. 2.2, the magenta circles represent the true conditional probability $P_{X|Y=y}$, which is a binomial distribution with parameters $(n, p) = (20, 0.5)$. We*

[1] Note that if there are more than one realization sharing the same maximal posterior belief, for $\alpha = \infty$ the optimal strategy in (2.21) will output these most likely values with the same probability.

observe from Fig. 2.2 that as α grows from 1 to ∞, the optimal strategy gradually eliminates the less likely values of X (given y) and transforms from the true posterior distribution to the MAP estimator.

2.3 Tunable Leakage Measures: α-Leakage and Maximal α-Leakage

Let X and Y represent the original data and released data, respectively, and let U represent an arbitrary (potentially random) function of X that the observer (a curious or malicious user of the released data Y) is interested in learning. In [23], Issa *et al.* introduced MaxL to quantify the maximal gain in an adversary's ability of guessing U after observing Y. We review the definition below.

Definition 2.3.1 ([23, Def. 1])**.** *Given a joint distribution $P_{X,Y}$ on finite alphabets, the maximal leakage from X to Y is*

$$\mathcal{L}_{MaxL}(X \to Y) \triangleq \sup_{U-X-Y} \log \frac{\max\limits_{P_{\widehat{U}|Y}} \mathbb{E}\left[P_{\widehat{U}|Y}(U|Y)\right]}{\max\limits_{u} P_U(u)}, \tag{2.23}$$

where $\widehat{U}$ represents an estimator taking values from the same arbitrary finite support as U.

Note that the numerator of the logarithmic term in (2.23) is the maximal expected probability of correctly guessing U with Y given by

$$\max_{P_{\widehat{U}|Y}} \mathbb{E}\left[P_{\widehat{U}|Y}(U|Y)\right] = \max_{u} \sum_{y} P_Y(y) P_{U|Y}(u|y), \tag{2.24}$$

which is exactly the complement of the minimal expected (soft) 0-1 loss in guessing U with Y. Similarly, the denominator is the complement of the minimal expected (soft) 0-1 loss in guessing U without Y. Therefore, MaxL is a leakage measure related to (soft) 0-1 loss in (2.12).

In addition, in Def. 2.3.1, U represents any (possibly random) function of X. The numerator represents the maximal probability of correctly guessing U based on Y, while the denominator represents the maximal probability of correctly guessing U *without* knowing Y. Thus, MaxL quantifies the maximal logarithmic gain in guessing any possible function of X when an adversary has access to Y.

Analogously to the derivation of MaxL from (soft) 0-1 loss, we introduce α-leakage and maximal α-leakage based on α-loss (under the assumptions of discrete random variables and finite supports). The formal definitions are as follows.

Definition 2.3.2 (α-Leakage)**.** *Given a joint distribution $P_{X,Y}$ and an estimator $\widehat{X}$ with the same support as X, the α-leakage from X to Y is defined as*

$$\mathcal{L}_\alpha(X \to Y) \triangleq \frac{\alpha}{\alpha-1} \log \frac{\max\limits_{P_{\widehat{X}|Y}} \mathbb{E}\left[P_{\widehat{X}|Y}(X|Y)^{\frac{\alpha-1}{\alpha}}\right]}{\max\limits_{P_{\widehat{X}}} \mathbb{E}\left[P_{\widehat{X}}(X)^{\frac{\alpha-1}{\alpha}}\right]}, \tag{2.25}$$

for $\alpha \in (1,\infty)$ and by the continuous extension of (2.25) for $\alpha = 1$ and ∞.

Whereas α-leakage captures how much an adversary can learn about X from Y, we also wish to quantify the information leaked about *any function* of X through Y. To this end, we define maximal α-leakage below.

Definition 2.3.3 (Maximal α-Leakage)**.** *Given a joint distribution $P_{X,Y}$ on finite alphabets $\mathcal{X} \times \mathcal{Y}$, the maximal α-leakage from X to Y is defined as*

$$\mathcal{L}_\alpha^{max}(X \to Y) \triangleq \sup_{U-X-Y} \mathcal{L}_\alpha(U;Y), \tag{2.26}$$

$$= \sup_{U-X-Y} \lim_{\alpha' \to \alpha} \frac{\alpha'}{\alpha'-1} \log \frac{\max\limits_{P_{\widehat{U}|Y}} \mathbb{E}\left[P_{\widehat{U}|Y}(U|Y)^{\frac{\alpha'-1}{\alpha'}}\right]}{\max\limits_{P_{\widehat{U}}} \mathbb{E}\left[P_{\widehat{U}}(U)^{\frac{\alpha'-1}{\alpha'}}\right]}, \tag{2.27}$$

where $1 \leq \alpha \leq \infty$, and U represents any function of X and takes values from an arbitrary finite alphabet.

Note that for $\alpha \geq 1$,

$$\max_{P_{\widehat{U}|Y}} \mathbb{E}\left[P_{\widehat{U}|Y}(U|Y)^{\frac{\alpha-1}{\alpha}}\right] = 1 - \frac{\alpha-1}{\alpha} \min_{P_{\widehat{U}|Y}} \mathbb{E}\left[\ell_\alpha(U, Y, P_{\widehat{U}|Y})\right]. \tag{2.28}$$

Thus, there is a similar connection between maximal α-leakage and α-loss (in Def. 2.2.1) as that observed in (2.24) between MaxL and (soft) 0-1 loss, and maximal α-leakage quantifies an adversary's capability to infer *any function* of data X from the released Y.

Making use of the result in Lemma 1, the following theorem simplifies the expression of α-leakage in (2.25).

Theorem 1. *For $1 \leq \alpha \leq \infty$, α-leakage defined in* (2.25) *simplifies to*

$$\mathcal{L}_\alpha(X \to Y) = I_\alpha^A(X;Y). \tag{2.29}$$

From (2.28) and Lemma 1, we simplify the scaled logarithm of the ratio in (2.25) to Arimoto MI. A detailed proof is in Subsection 2.6.2, where we show that Arimoto conditional entropy and Rényi entropy capture the inference uncertainties of an adversary for knowing Y or not, respectively, and α-leakage measures the decrease in the inference uncertainty by knowing Y.

Making use of the conclusion in Thm. 1, the following theorem gives equivalent expressions for maximal α-leakage.

Theorem 2. *For $1 \leq \alpha \leq \infty$, the maximal α-leakage defined in* (2.26) *simplifies to*

$$\mathcal{L}_\alpha^{max}(X \to Y) = \begin{cases} \sup_{P_{\tilde{X}}} I_\alpha^S(\tilde{X};Y) = \sup_{P_{\tilde{X}}} I_\alpha^A(\tilde{X};Y), & 1 < \alpha \leq \infty \quad (2.30a) \\ I(X;Y), & \alpha = 1 \quad (2.30b) \end{cases}$$

where $P_{\tilde{X}}$ is a probability distribution over the support of P_X.

Note that maximal α-leakage is essentially the Arimoto channel capacity (with a support-set constrained input distribution) for $\alpha > 1$ [92], which is used to characterize probabilities of decoding error for scenarios in which transmission rates are higher than channel capacity. The limit of maximal α-leakage for $\alpha = 1$ gives the Shannon channel capacity. Recall that the limit of α-loss in (5.20) leads to the log-loss (for $\alpha = 1$) and soft 0-1 loss (for $\alpha = \infty$) functions, respectively. Consequently, for $\alpha = 1$ and ∞, maximal α-leakage simplifies to MI and MaxL, respectively.

A detailed proof for Thm. 2 is in Subsection 2.6.3. We summarize key steps in the proof as follows: by applying Thm. 1, we write maximal α-leakage as

$$\mathcal{L}_\alpha^{\max}(X \to Y) = \sup_{U-X-Y} I_\alpha^{\mathrm{A}}(U;Y) \quad \alpha \in [1,\infty]. \tag{2.31}$$

For $\alpha = 1$, Arimoto MI is simply the Shannon's MI, and combining with the data processing inequalities, (2.31) simplifies to $I(X;Y)$. Note that for $\alpha > 1$, Arimoto MI does not satisfy data processing inequalities. By using the facts that Arimoto MI and Sibson MI have the same supremum [90, Thm. 5] and that Sibson MI satisfies data processing inequalities [90, Thm. 3], we upper bound the supremum of (2.31) by $\sup_{P_{\tilde{X}}} I_\alpha^{\mathrm{S}}(\tilde{X};Y)$, and then, show that the upper bound can be achieved by a specific U with $H(X|U) = 0$.

Example 2. *Given a binary channel*

$$P_{Y|X} = \begin{bmatrix} 1-\rho_1 & \rho_1 \\ \rho_2 & 1-\rho_2 \end{bmatrix}, \tag{2.32}$$

where $\rho_1, \rho_2 \in [0,1]$ *are the crossover probabilities, maximal* α*-leakage in* (2.30) *is*

given by

$$\mathcal{L}_\alpha^{max}(X \to Y) = \frac{\alpha}{\alpha-1} \log \Bigg(\Big| (1-\rho_1)^\alpha (1-\rho_2)^\alpha - \rho_1^\alpha \rho_2^\alpha \Big|^{\frac{1}{\alpha}} \cdot \left(\Big| (1-\rho_2)^\alpha - \rho_1^\alpha \Big|^{\frac{1}{1-\alpha}} + \Big| (1-\rho_1)^\alpha - \rho_2^\alpha \Big|^{\frac{1}{1-\alpha}} \right)^{\frac{\alpha-1}{\alpha}} \Bigg). \tag{2.33}$$

If $\rho_1 = \rho_2$, (2.33) *simplifies to*

$$\mathcal{L}_\alpha^{max}(X \to Y) = \frac{1}{\alpha-1} \log \left((1-\rho_1)^\alpha + \rho_1^\alpha \right) + \log 2, \tag{2.34}$$

which is exactly the α*-leakage for the binary symmetric channel with the uniform input distribution. Fig. 2.3 plots the values of maximal* α*-leakage for example channels where* $\rho_1 = \rho_2$ *and* $\rho_1 \neq \rho_2$*, and shows that the ordering of leakages for the two channels varies with* α*.*

2.4 Properties of Tunable Leakage Measures

For the potential use of the two tunable information leakage measures α-leakage and maximal α-leakage introduced, for example, used as the privacy metric in privacy-utility tradeoff problems, we explore the properties of these measures. All properties are proved based on the simplified expressions of of α-leakage and maximal α-leakage in Thm. 1 and Thm. 2, respectively.

2.4.1 Properties of α*-leakage*

Thm. 1 shows that α-leakage is exactly Arimoto MI, and therefore, several basic properties of α-leakage have been shown including

(i) non-negativity [90, Section II-A],

(ii) quasi-convexity, which is proved based on the facts that for $\alpha \geq 1$ and P_X, the Arimoto MI $I_\alpha^{\mathrm{A}}(X;Y)$ is the logarithm of a linear combination of the p-norm

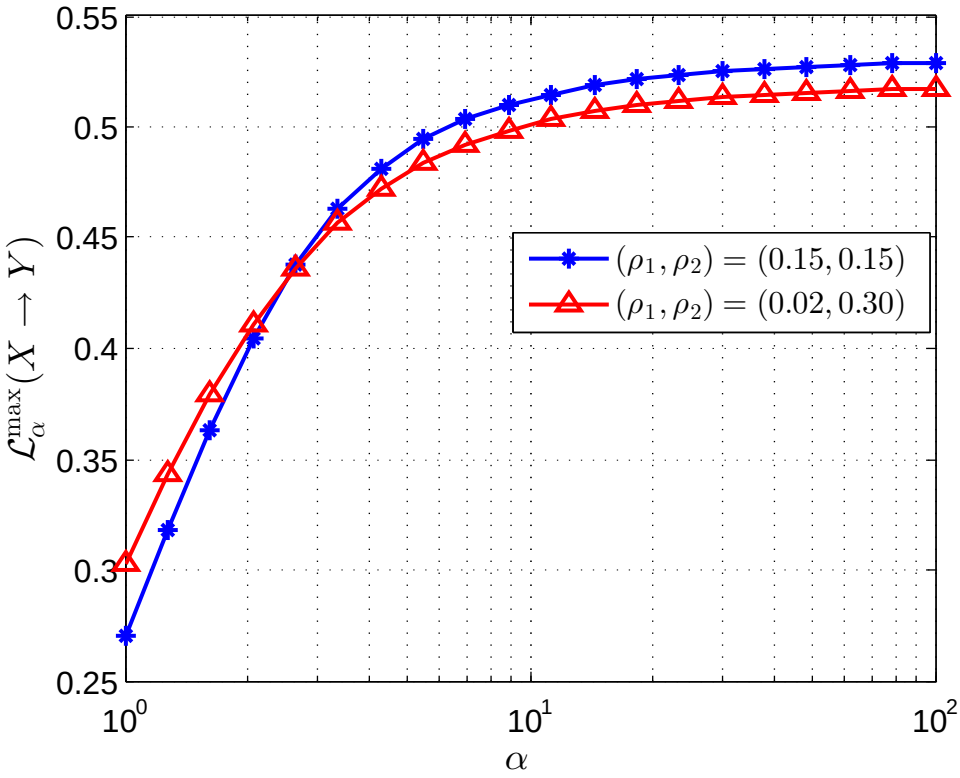


Figure 2.3: Values of Maximal α-Leakage for Binary Channels Determined by a Pair of Crossover Probabilities (ρ_1, ρ_2).

$(p = \alpha)$ $\|P_{Y|X}(\cdot|x)\|_\alpha$. From [99, Chapter 3.5], we know a log-convex function is quasi-convex such that $I_\alpha^{\mathrm{A}}(X;Y)$ is quasi-convex in $P_{Y|X}$ given P_X.

(iii) post-processing inequality, i.e., for a Markov chain $X - Y - Z$, $I_\alpha^{\mathrm{A}}(X;Z) \leq I_\alpha^{\mathrm{A}}(X;Y)$, which is directly derived from the monotonicity of conditional Arimoto entropy [100, Corollary 1].

2.4.2 Properties of Maximal α-leakage

We explore proprieties of maximal α-leakage and show that its properties include: (i) quasi-convexity in the conditional distribution $P_{Y|X}$; (ii) data processing inequalities; (iii) sub-additivity (composition property [23]) and additivity for memoryless mechanisms.

The following theorem results from the expression of maximal α-leakage in Thm. 2 as well as some known properties of Sibson MI [91, 93, 90].

Theorem 3. *For $1 \leq \alpha \leq \infty$, maximal α-leakage*

1. *is quasi-convex in $P_{Y|X}$;*

2. *is monotonically non-decreasing in α;*

3. *satisfies data processing inequalities: let random variables X, Y, Z form a Markov chain, i.e., $X - Y - Z$, then*

$$\mathcal{L}_\alpha^{max}(X \to Z) \leq \mathcal{L}_\alpha^{max}(X \to Y) \tag{2.35a}$$

$$\mathcal{L}_\alpha^{max}(X \to Z) \leq \mathcal{L}_\alpha^{max}(Y \to Z). \tag{2.35b}$$

4. *satisfies*

$$\mathcal{L}_\alpha^{max}(X \to Y) \geq 0 \tag{2.36}$$

with equality if and only if X is independent of Y, and

$$\mathcal{L}_\alpha^{max}(X \to Y) \leq \begin{cases} \log |\mathcal{X}| & \alpha > 1 \\ H(P_X) & \alpha = 1 \end{cases} \tag{2.37}$$

with equality if and only if X is a deterministic function of Y.

A detailed proof is in Appendix 2.6.4.

Remark 1. *Note that:*

- *Since both MI and MaxL are convex in $P_{Y|X}$, $\mathcal{L}_1^{max}(X \to Y)$ and $\mathcal{L}_\infty^{max}(X \to Y)$ are convex in $P_{Y|X}$.*

- *From the monotonicity in Part 2, we can upper bound maximal α-leakage as* [2]

$$\mathcal{L}_{\alpha}^{max}(X \to Y) \leq \mathcal{L}_{MaxL}(X \to Y) = I_{\infty}^{S}(X;Y). \tag{2.38}$$

- *The data processing inequalities in* (2.35a) *and* (2.35b) *are called post-processing inequality and linkage inequality, respectively [101, 102]. It is worth noting that not all information leakage metrics satisfy the linkage inequality [102, 20]. Examples include α-leakage, maximal information leakage [26], probability of correctly guessing, and DP.*

From Thm. 2, we know that for $\alpha > 1$, maximal α-leakage is the supremum of Arimoto/Sibson MI over all possible distributions on the support of original data, and therefore, is a function of a conditional probability distribution. The following theorem lower bounds the supremum by a closed-form expression of the conditional probability distribution.

Theorem 4 (Lower Bound)**.** *For $1 < \alpha \leq \infty$, maximal α-leakage is lower bounded by*

$$\mathcal{L}_{\alpha}^{max}(X \to Y) \geq \frac{\alpha}{\alpha - 1} \log \frac{\sum\limits_{y \in \mathcal{Y}} \|P_{Y|X}(y|\cdot)\|_{\alpha}}{|\mathcal{X}|^{\frac{1}{\alpha}}}, \tag{2.39}$$

with equality if and only if for all $x_1, x_2 \in \mathcal{X}$, there is

$$\sum_{y} \frac{P_{Y|X}(y|x_1)^{\alpha}}{\|P_{Y|X}(y|\cdot)\|_{\alpha}^{\alpha-1}} = \sum_{y} \frac{P_{Y|X}(y|x_2)^{\alpha}}{\|P_{Y|X}(y|\cdot)\|_{\alpha}^{\alpha-1}}. \tag{2.40}$$

A detailed proof is in Appendix 2.6.5.

When data may be revealed multiple times (e.g., entering a password multiple times), it is essential to quantify how mechanisms are designed with maximal α leakage compose in terms of total leakage. Consider two released versions Y_1 and Y_2 of X.

[2] For $\alpha = \infty$, the $I_{\infty}^{\mathrm{S}}(P_X, P_{Y|X})$ depends on the marginal distribution P_X only through the support of X.

The following theorem upper bounds maximal α-leakage to an adversary who has access to both Y_1 and Y_2 simultaneously.

Theorem 5 (Sub-additivity/Composition)**.** *Given a Markov chain $Y_1 - X - Y_2$, we have ($\alpha \in [1, \infty]$)*

$$\mathcal{L}_\alpha^{max}(X \to Y_1, Y_2) \leq \sum_{i \in \{1,2\}} \mathcal{L}_\alpha^{max}(X \to Y_i). \tag{2.41}$$

A detailed proof is in Appendix 2.6.6.

The following theorem shows the additivity of maximal α-leakage for memoryless mechanisms.

Theorem 6 (Additivity for Memoryless Mechanisms)**.** *For $\alpha \in [1, \infty]$ and a finite integer $n > 0$, let X^n and Y^n be n-length input and output, respectively, of a memoryless mechanism with no feedback, i.e.,*

$$P_{Y^n|X^n} = \prod_{i=1}^{n} P_{Y_i|X_i}, \tag{2.42}$$

where X_i and Y_i represent the i^{th} element of X^n and Y^n, respectively, such that

(1) for $\alpha > 1$

$$\mathcal{L}_\alpha^{max}(X^n \to Y^n) = \sum_{i=1}^{n} \mathcal{L}_\alpha^{max}(X_i \to Y_i) \tag{2.43}$$

(2) for $\alpha = 1$

$$\mathcal{L}_1^{max}(X^n \to Y^n) \leq \sum_{i=1}^{n} \mathcal{L}_1^{max}(X_i \to Y_i) \tag{2.44}$$

with equality if and only if entries of X^n are mutually independent.

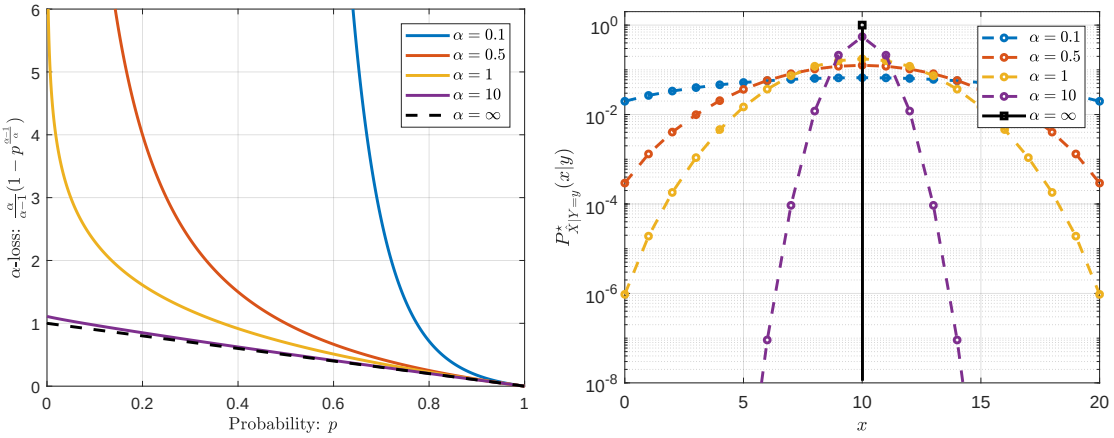
A detailed proof is in Appendix 2.6.7.

2.5 Extension to $0 < \alpha < 1$

In this section, we include the range of $0 < \alpha < 1$ into the definition of α-loss (in Def. 2.2.1) such that the α-loss of the strategy $P_{\widehat{X}|Y}$ for estimating $\widehat{X} = x$ from $Y = y$ is

$$\ell_\alpha(x, y, P_{\widehat{X}|Y}) = \begin{cases} \frac{\alpha}{\alpha-1}\left(1 - P_{\widehat{X}|Y}(x|y)^{\frac{\alpha-1}{\alpha}}\right), & 0 < \alpha < \infty \cap \alpha \neq 1 \\ -\log P_{\widehat{X}|Y}(x|y), & \alpha = 1 \\ 1 - P_{\widehat{X}|Y}(x|y) & \alpha = \infty. \end{cases} \tag{2.45}$$

As shown in Fig. 2.4a, the α-loss function is more convexity as α decreases. In addition, for $0 < \alpha < 1$, the minimal expected α-loss is given by $\frac{\alpha}{\alpha-1}\left(1 - \exp\left(\frac{1-\alpha}{\alpha} H_\alpha^A(X|Y)\right)\right)$, which is the same expression as for $\alpha > 1$ in (2.20), and the corresponding optimal estimation strategy can also be expressed as (2.21), which approaches a uniform distribution as α tends to 0 (as shown in Fig. 2.4b). Note that in Fig. 2.4b, the true conditional probability $P_{X|Y=y}$ is a binomial distribution with parameters $(n, p) = (20, 0.5)$. This extension of α-loss is motivated by the advantage of α-loss with $0 < \alpha < 1$ in



(a) Plot of α-Loss in (2.45) (b) The Optimal Strategy in (2.21).

Figure 2.4: Illustration of α-Loss and Corresponding Optimal Mechanisms for $\alpha > 0$

providing a higher accuracy of classification on imbalanced data sets where the num-

ber of samples from each category significantly differs [103]. To incorporate the case of which an adversarial inference is determined via the α-loss with $0 < \alpha < 1$, we modified the the definitions of α-leakage in Def. 2.3.2 as follows.

Definition 2.5.1. *Given a joint distribution $P_{X,Y}$ and an estimator $\hat{X}$ with the same support as X, the α-leakage from X to Y is defined as*

$$\mathcal{L}_\alpha(X \to Y) \triangleq \frac{\alpha}{\alpha-1} \log \frac{\max\limits_{P_{\hat{X}|Y}} (\alpha-1)\mathbb{E}\left[P_{\hat{X}|Y}(X|Y)^{\frac{\alpha-1}{\alpha}}\right]}{\max\limits_{P_{\hat{X}}} (\alpha-1)\mathbb{E}\left[P_{\hat{X}}(X)^{\frac{\alpha-1}{\alpha}}\right]}, \tag{2.46}$$

for $0 < \alpha < 1$ or $1 < \alpha < \infty$ and by the continuous extension of (2.46) for $\alpha = 1$ and ∞.

Note that for $\alpha > 1$, the expression of α-leakage in (2.46) is exactly the same as in (2.25). For $0 < \alpha < 1$, the minimization of the expected α-loss can be reduced to $\min_{P_{\hat{X}|Y}} \mathbb{E}\left[P_{\hat{X}|Y}(X|Y)^{\frac{\alpha-1}{\alpha}}\right]$, and therefore, the adversarial benefit in estimating X by observing Y can be captured by

$$\frac{\alpha}{1-\alpha} \log \left(\min_{P_{\hat{X}}} \mathbb{E}\left[P_{\hat{X}}(X)^{\frac{\alpha-1}{\alpha}}\right] \right) - \frac{\alpha}{1-\alpha} \log \left(\min_{P_{\hat{X}|Y}} \mathbb{E}\left[P_{\hat{X}|Y}(X|Y)^{\frac{\alpha-1}{\alpha}}\right] \right) \tag{2.47}$$

$$= \frac{\alpha}{1-\alpha} \log \frac{\min\limits_{P_{\hat{X}}} \mathbb{E}\left[P_{\hat{X}}(X)^{\frac{\alpha-1}{\alpha}}\right]}{\min\limits_{P_{\hat{X}|Y}} \mathbb{E}\left[P_{\hat{X}|Y}(X|Y)^{\frac{\alpha-1}{\alpha}}\right]}, \tag{2.48}$$

which is equivalent to the expression in (2.46) for $0 < \alpha < 1$. Therefore, although for $0 < \alpha < 1$ and $\alpha > 1$, the expression $P_{\hat{X}|Y}(X|Y)^{\frac{\alpha-1}{\alpha}}$ is positively and negatively related to α-loss, respectively, the definition of α-leakage in (2.46) correctly captures the advantage of an adversary in learning X from Y. In Def. 2.3.3, the maximal α-leakage is defined as the maximization of α-leakage over all possible function of X, and therefore, can be extended to $0 < \alpha < 1$ naturally from the extension of α-leakage. In addition, it's clear that for $0 < \alpha < 1$, the expression in (2.46) can also be simplified

to Arimoto MI between X and Y as for $\alpha > 1$ (shown in Theorem 1), and therefore, α-leakage for $0 < \alpha < 1$ inherits the properties of Arimoto MI for $0 < \alpha < 1$, including the non-negativity [90] and post-processing inequality [100]. Furthermore, by taking the same methodology used in proving the properties of maximal α-leakage for $\alpha > 1$, one can verify that the simplified expression and properties of maximal α-leakage for $\alpha > 1$ can be applied to the extended region of $0 < \alpha < 1 \cup \alpha > 1$.

2.6 Proof Details

2.6.1 Proof of Lemma 1

For $1 < \alpha < \infty$, the minimal expected value of the α-loss in Definition 2.2.1 can be expressed as

$$\begin{aligned}
&\min_{P_{\widehat{X}|Y}} \mathbb{E}\left[\ell_\alpha(X, Y, P_{\widehat{X}|Y})\right] \\
=&\min_{P_{\widehat{X}|Y}} \frac{\alpha}{\alpha-1}\left(1 - \sum_{xy} P_{X,Y}(x,y) P_{\widehat{X}|Y}(x|y)^{\frac{\alpha-1}{\alpha}}\right) && (2.49)\\
=&\frac{\alpha}{\alpha-1}\left(1 - \max_{P_{\widehat{X}|Y}} \sum_{xy} P_{X,Y}(x,y) P_{\widehat{X}|Y}(x|y)^{\frac{\alpha-1}{\alpha}}\right) && (2.50)\\
=&\frac{\alpha}{\alpha-1}\left(1 - \sum_{y} P_Y(y) \max_{P_{\widehat{X}|Y=y}} \sum_{x} P_{X|Y}(x|y) P_{\widehat{X}|Y}(x|y)^{\frac{\alpha-1}{\alpha}}\right). && (2.51)
\end{aligned}$$

For each y with $P_Y(y) > 0$, the maximization in (2.51) can be explicitly written as

$$\begin{aligned}
\max_{P_{\widehat{X}|Y=y}} \quad & \sum_{x\in\mathcal{X}} P_{X|Y}(x|y) P_{\widehat{X}|Y}(x|y)^{\frac{\alpha-1}{\alpha}} && (2.52a)\\
\text{s.t.} \quad & \sum_{x\in\mathcal{X}} P_{\widehat{X}|Y}(x|y) = 1 && (2.52b)\\
& P_{\widehat{X}|Y}(x|y) \geq 0 \quad \text{for all } x \in \mathcal{X}. && (2.52c)
\end{aligned}$$

For $1 \leq \alpha \leq \infty$, the exponent $\frac{\alpha-1}{\alpha} \geq 0$ such that the problem in (2.52) is a convex program. Therefore, by using Karush—Kuhn—Tucker (KKT) conditions [99,

Chapter 5.5.3], we obtain the optimal value of (2.52) as

$$\max_{P_{\widehat{X}|Y=y}} \sum_x P_{X|Y}(x|y) P_{\widehat{X}|Y}(x|y)^{\frac{\alpha-1}{\alpha}} = \|P_{X|Y}(\cdot|y)\|_\alpha \tag{2.52d}$$

with the optimal solution $P^\star_{\widehat{X}|Y}$ as

$$P^\star_{\widehat{X}|Y}(x|y) = \frac{P_{X|Y}(x|y)^\alpha}{\sum\limits_{x\in\mathcal{X}} P_{X|Y}(x|y)^\alpha} \quad \text{for all } x \in \mathcal{X}. \tag{2.52e}$$

For $\alpha = 1$, the optimal solution is $P^\star_{\widehat{X}|Y} = P_{X|Y}$. For $\alpha = \infty$, we have

$$\lim_{\alpha\to\infty} P^\star_{\widehat{X}|Y}(x|y) = \lim_{\alpha\to\infty} \frac{\left(\frac{P_{X|Y}(x|y)}{\max_x P_{X|Y}(x|y)}\right)^\alpha}{\sum\limits_{x\in\mathcal{X}} \left(\frac{P_{X|Y}(x|y)}{\max_x P_{X|Y}(x|y)}\right)^\alpha} \tag{2.53}$$

$$= \begin{cases} \frac{1}{k(y)}, & x = \arg\max_x P_{X|Y}(x|y) \\ 0, & \text{otherwise,} \end{cases} \tag{2.54}$$

where the integer $k(y)$ indicates the cardinality of the set $\left\{x : x = \arg\max_x P_{X|Y}(x|y)\right\}$.

Applying the optimal solution $P^\star_{\widehat{X}|Y}$ to (2.51), we have

$$\min_{P_{\widehat{X}|Y}} \mathbb{E}\left[\ell_\alpha(X, Y, P_{\widehat{X}|Y})\right] = \begin{cases} \frac{\alpha}{\alpha-1}\left(1 - \sum\limits_y \|P_{X,Y}(Xy)\|_\alpha\right), & \alpha > 1 \\ \sum\limits_{x,y} P_{X,Y}(x,y) \log \frac{1}{P_{X|Y}(x|y)}, & \alpha = 1 \end{cases}, \tag{2.55}$$

$$= \begin{cases} \frac{\alpha}{\alpha-1}\left(1 - \exp\left(\frac{1-\alpha}{\alpha} H^{\mathrm{A}}_\alpha(X|Y)\right)\right), & \alpha > 1 \\ H(X|Y), & \alpha = 1 \end{cases}. \tag{2.56}$$

□

2.6.2 Proof of Theorem 1

The expression (2.25) can be explicitly written as

$$\mathcal{L}_\alpha(X \to Y) = \lim_{\alpha'\to\alpha} \frac{\alpha'}{\alpha'-1} \log \left(\frac{\max\limits_{P_{\widehat{X}|Y}} \sum\limits_{xy} P_{X,Y}(x,y) \left(P_{\widehat{X}|Y}(x|y)\right)^{\frac{\alpha'-1}{\alpha'}}}{\max\limits_{P_{\widehat{X}}} \sum\limits_x P_X(x) P_{\widehat{X}}(x)^{\frac{\alpha'-1}{\alpha'}}} \right). \tag{2.57}$$

To simplify the expression in (2.57), we need to solve the two maximizations in the logarithm. From (2.28), we know that to solve the maximization in the numerator equals to find the minimal expected α-loss. Making use of the result in Lemma 1, we have that for $\alpha' \in (1, \infty)$,

$$\max_{P_{\widehat{X}|Y}} \sum_{x,y} P_{X,Y}(x,y) P_{\widehat{X}|Y}(x|y)^{\frac{\alpha'-1}{\alpha'}} = \exp\left(\frac{1-\alpha'}{\alpha'} H_{\alpha'}^{\mathrm{A}}(X|Y)\right). \tag{2.58}$$

Similarly, by applying KKT conditions to the maximization in the denominator, we have that for $\alpha' \in (1, \infty)$

$$\max_{P_{\widehat{X}}} \sum_{x \in \mathcal{X}} P_X(x) P_{\widehat{X}}(x)^{\frac{\alpha'-1}{\alpha'}} = \exp\left(\frac{1-\alpha'}{\alpha'} H_{\alpha'}(X)\right). \tag{2.59}$$

Therefore, we have for $\alpha' \in (1, \infty)$

$$\mathcal{L}_\alpha(X \to Y) = \frac{\alpha'}{\alpha'-1} \log \exp\left(\frac{1-\alpha'}{\alpha'}\left(H_{\alpha'}^{\mathrm{A}}(X|Y) - H_{\alpha'}(X)\right)\right) = I_{\alpha'}^{\mathrm{A}}(X;Y) \tag{2.60}$$

From the continuous extensions of Arimoto MI for $\alpha = 1$ and ∞, respectively, we have that for $1 \leq \alpha \leq \infty$, α-leakage equals to Arimoto MI.

□

2.6.3 Proof of Theorem 2

From Thm. 1, we have for $1 \leq \alpha \leq \infty$,

$$\mathcal{L}_\alpha^{\max}(X \to Y) = \sup_{U-X-Y} I_\alpha^{\mathrm{A}}(U;Y). \tag{2.61}$$

If $\alpha = 1$, we have

$$\mathcal{L}_1^{\max}(X \to Y) = \sup_{U-X-Y} I(U;Y) \leq I(X;Y) \tag{2.62}$$

where the inequality is from data processing inequalities of MI [104, Thm 2.8.1].

If $\alpha = \infty$, we have

$$\mathcal{L}_{\infty}^{\max}(X \to Y) = \sup_{U-X-Y} \log \frac{\sum_y P_Y(y) \max_u P_{U|Y}(u|y)}{\max_u P_U(u)}, \tag{2.63}$$

which is exactly the expression of MaxL, and therefore, we have that for $\alpha = \infty$, the maximal α-leakage equals to the Sibson MI of order ∞ [23, Thm. 1], i.e.,

$$\mathcal{L}_{\infty}^{\max}(X \to Y) = \log \sum_y \max_x P_{Y|X}(y|x). \tag{2.64}$$

For $\alpha \in (1, \infty)$, we provide an upper bound for $\mathcal{L}_{\alpha}^{\max}(X \to Y)$, and then, give an achievable scheme as follows.

Upper Bound: We have an upper bound of $\mathcal{L}_{\alpha}^{\max}(X \to Y)$ as

$$\begin{aligned}
\mathcal{L}_{\alpha}^{\max}(X \to Y) &= \sup_{U-X-Y} I_{\alpha}^{\mathrm{A}}(U;Y) && \text{(2.65a)}\\
&\le \sup_{P_{\tilde{X}|\tilde{U}}: P_{\tilde{X}|\tilde{U}}(\cdot|u) \ll P_X} \sup_{P_{\tilde{U}}} I_{\alpha}^{\mathrm{A}}(\tilde{U};Y) && \text{(2.65b)}\\
&= \sup_{P_{\tilde{X}|\tilde{U}}: P_{\tilde{X}|\tilde{U}}(\cdot|u) \ll P_X} \sup_{P_{\tilde{U}}} I_{\alpha}^{\mathrm{S}}(\tilde{U};Y) && \text{(2.65c)}\\
&= \sup_{P_{\tilde{X}} \ll P_X} I_{\alpha}^{\mathrm{S}}(\tilde{X};Y) && \text{(2.65d)}\\
&= \sup_{P_{\tilde{X}} \ll P_X} I_{\alpha}^{\mathrm{A}}(\tilde{X};Y) && \text{(2.65e)}
\end{aligned}$$

where $P_{\tilde{X}} \ll P_X$ means the alphabet of $P_{\tilde{X}}$ is a subset of that of P_X. The inequality in (2.65b) holds because the supremum of Arimoto MI over all $P_{\tilde{U},\tilde{X}}$ on $\mathcal{U} \times \mathcal{X}$ is no less than that (in (2.65a)) over these $P_{U,X}$ constrained by the P_X. The equations in (2.65c) and (2.65e) result from that Arimoto MI and Sibson MI of order $\alpha > 0$ have the same supremum [90, Thm. 5]; and (2.65d) obeys the data processing inequalities [90, Thm. 3].

Lower bound: We lower bound (2.61) by considering a random variable U such

that $U - X - Y$ is a Markov chain and $H(X|U) = 0$. Specifically, let the alphabet $\mathcal{U}$ consist of $\mathcal{U}_x$, a collection of U mapped to a $x \in \mathcal{X}$, i.e., $\mathcal{U} = \cup_{x\in\mathcal{X}}\mathcal{U}_x$ with $U = u \in \mathcal{U}_x$ if and only if $X = x$. Therefore, for the specific variable U, we have

$$P_{Y|U}(y|u) = \begin{cases} P_{Y|X}(y|x) & \text{for all } u \in \mathcal{U}_x \\ 0 & \text{otherwise.} \end{cases} \tag{2.66}$$

Construct a probability distribution $P_{\tilde{X}}$ over $\mathcal{X}$ from P_U as

$$P_{\tilde{X}}(x) = \frac{\sum_{u\in\mathcal{U}_x} P_U(u)^\alpha}{\sum_{x\in\mathcal{X}}\sum_{u\in\mathcal{U}_x} P_U(u)^\alpha} \quad \text{for all } x \in \mathcal{X}. \tag{2.67}$$

Thus,

$$I_\alpha^{\mathrm{A}}(U;Y) = \frac{\alpha}{\alpha-1}\log\frac{\sum\limits_{y\in\mathcal{Y}}\left(\sum\limits_{x\in\mathcal{X}}\sum\limits_{u\in\mathcal{U}_x} P_{Y|U}(y|u)^\alpha P_U(u)^\alpha\right)^{\frac{1}{\alpha}}}{\left(\sum\limits_{x\in\mathcal{X}}\sum\limits_{u\in\mathcal{U}_x} P_U(u)^\alpha\right)^{\frac{1}{\alpha}}} \tag{2.68}$$

$$= \frac{\alpha}{\alpha-1}\log\sum_{y\in\mathcal{Y}}\left(\sum_{x\in\mathcal{X}} P_{Y|X}(y|x)^\alpha \frac{\sum\limits_{u\in\mathcal{U}_x} P_U(u)^\alpha}{\sum\limits_{x\in\mathcal{X}}\sum\limits_{u\in\mathcal{U}_x} P_U(u)^\alpha}\right)^{\frac{1}{\alpha}} \tag{2.69}$$

$$= \frac{\alpha}{\alpha-1}\log\left(\sum_{y\in\mathcal{Y}}\left(\sum_{x\in\mathcal{X}} P_{Y|X}(y|x)^\alpha P_{\tilde{X}}(x)^\alpha\right)^{\frac{1}{\alpha}}\right) \tag{2.70}$$

$$= I_\alpha^{\mathrm{S}}(\tilde{X};Y) \tag{2.71}$$

Therefore,

$$\mathcal{L}_\alpha^{\max}(X\to Y) = \sup_{U-X-Y} I_\alpha^{\mathrm{A}}(U;Y) \geq \sup_{U:U-X-Y,H(X|U)=0} I_\alpha^{\mathrm{A}}(U;Y) = \sup_{P_{\tilde{X}}\ll P_X} I_\alpha^{\mathrm{S}}(\tilde{X};Y), \tag{2.72}$$

where the last inequality is because for any $P_{\tilde{X}} \ll P_X$, it can be obtained through (2.67) by appropriately choosing P_U. Therefore, combining (2.65) and (2.72), we obtain (2.30a). □

2.6.4 Proof of Theorem 3

The proof of part 1: We know that for $\alpha \geq 1$, $I_\alpha^{\mathrm{S}}(X;Y)$ is quasi-convex $P_{Y|X}$ for given P_X [104, Thm. 2.7.4], [93, Thm. 10]. In addition, the supremum of a set of quasi-convex functions is also quasi-convex, i.e., if the function $f(a,b)$ is quasi-convex in b for any given a, the supremum $\sup_a f(a,b)$ is also quasi-convex in b [99]. Therefore, maximal α-leakage in (2.30) is quasi-convex $P_{Y|X}$ for given P_X.

The proof of part 2: Let $\beta > \alpha \geq 1$, and $P_{X\alpha}^\star = \arg\sup_{P_X} I_\alpha^{\mathrm{S}}(P_X, P_{Y|X})$ for given $P_{Y|X}$, such that

$$\mathcal{L}_\alpha^{\max}(X \to Y) = I_\alpha^{\mathrm{S}}(P_{X\alpha}^\star, P_{Y|X}) \tag{2.73}$$

$$\leq I_\beta^{\mathrm{S}}(P_{X\alpha}^\star, P_{Y|X}) \tag{2.74}$$

$$\leq \sup_{P_X} I_\beta^{\mathrm{S}}(P_X, P_{Y|X}) \tag{2.75}$$

$$= \mathcal{L}_\beta^{\max}(X \to Y) \tag{2.76}$$

where (2.74) results from that I_α^{S} is non-decreasing in α for $\alpha > 0$ [93, Thm. 4], and the equality in (2.75) holds if and only if $P_{X\alpha}^\star = \arg\sup_{P_X} I_\beta(P_X, P_{Y|X})$.

The proof of part 3: Let random variables X, Y and Z form the Markov chain $X - Y - Z$. Making use of that Sibson MI of order $\alpha > 1$ satisfies data processing inequalities [90, Thm. 3], i.e.,

$$I_\alpha^{\mathrm{S}}(X;Z) \leq I_\alpha^{\mathrm{S}}(X;Y) \tag{2.77}$$

$$I_\alpha^{\mathrm{S}}(X;Z) \leq I_\alpha^{\mathrm{S}}(Y;Z), \tag{2.78}$$

we prove that maximal α-leakage satisfies data processing inequalities as follows.

We first prove (2.35a). Let $P_X^\star = \arg\sup_{P_X} I_\alpha^{\mathrm{S}}(P_X, P_{Z|X})$. For the Markov chain

$X - Y - Z$, we have

$$\mathcal{L}_\alpha^{\max}(X \to Z) = I_\alpha^{\mathrm{S}}(P_X^\star, P_{Z|X}) \tag{2.79}$$

$$\leq I_\alpha^{\mathrm{S}}(P_X^\star, P_{Y|X}) \tag{2.80}$$

$$\leq \sup_{P_X} I_\alpha^{\mathrm{S}}(P_X, P_{Y|X}) \tag{2.81}$$

$$= \mathcal{L}_\alpha^{\max}(X \to Y) \tag{2.82}$$

where the inequality in (2.80) results from (2.77). Similarly, the inequality in (2.35b) can be proved directly from (2.78).

The proof of part 4: For $\alpha = 1$, we have

$$\mathcal{L}_1^{\max}(X \to Y) = I(X;Y) \geq 0, \tag{2.83}$$

with equality if and only if X is independent of Y [104]. For $1 < \alpha \leq \infty$, referring to (2.6) and (2.30a) we have

$$\mathcal{L}_\alpha^{\max}(X \to Y) = \sup_{P_X} \frac{\alpha}{\alpha - 1} \log \sum_y \left(\sum_x P_X(x) P_{Y|X}(y|x)^\alpha \right)^{\frac{1}{\alpha}} \tag{2.84}$$

$$\geq \sup_{P_X} \frac{\alpha}{\alpha - 1} \log \sum_y \left(\sum_x P_X(x) P_{Y|X}(y|x) \right)^{\frac{\alpha}{\alpha}} \tag{2.85}$$

$$= \sup_{P_X} \frac{\alpha}{\alpha - 1} \log 1 = 0, \tag{2.86}$$

where (2.85) results from applying Jensen's inequality to the convex function $f : t \to t^\alpha$ $(t \geq 0)$, such that the equality holds if and only if given any $y \in \mathcal{Y}$, $P_{Y|X}(y|x)$ are the same for all $x \in \mathcal{X}$, such that

$$P_{Y|X}(y|x) = P_Y(y) \quad x \in \mathcal{X}, y \in \mathcal{Y} \tag{2.87}$$

which means X and Y are independent, i.e., $P_{Y|X}$ is a rank-1 row stochastic matrix. For $\alpha = 1$, from (2.30b) we know $\mathcal{L}_1^{\max}(X \to Y) = I(X;Y)$. Therefore,

$$\mathcal{L}_1^{\max}(X \to Y) - H(X) = \sum_{\mathcal{X},\mathcal{Y}} P(x,y) \log \frac{P(y|x)}{P(y)} - \sum_{\mathcal{X}} P(x) \log \frac{1}{P(x)} \tag{2.88}$$

$$= \sum_{\mathcal{X},\mathcal{Y}} P(x,y) \log \frac{P(y|x)}{P(y)} - \sum_{\mathcal{X},\mathcal{Y}} P(x,y) \log \frac{1}{P(x)} \tag{2.89}$$

$$= \sum_{\mathcal{X},\mathcal{Y}} P(x,y) \log P(x|y) \leq 0, \tag{2.90}$$

with equality if and only if for all $x, y \in \mathcal{X} \times \mathcal{Y}$, the conditional probability $P_{X|Y}(x|y)$ is either 1 or 0. That is, $\mathcal{L}_1^{\max}(X \to Y) \leq H(X)$ with equality if and only if X is a deterministic function of Y [105, Lem. 1]. For $1 < \alpha \leq \infty$, from the monotonicity of maximal α-leakage in α and (2.30a), we have

$$\mathcal{L}_\alpha^{\max}(X \to Y) \leq \mathcal{L}_\infty^{\max}(X \to Y) \tag{2.91}$$

$$= \log \sum_{y \in \mathcal{Y}} \max_x P_{Y|X}(y|x) \tag{2.92}$$

$$\leq \log \sum_{\mathcal{Y}} \sum_{\mathcal{X}} P_{Y|X}(y|x) = \log |\mathcal{X}|. \tag{2.93}$$

where the equality in (2.93) holds if and only if for every $y \in \mathcal{Y}$, $\sum_{\mathcal{X}} P(y|x) = \max_x P(y|x)$, i.e., X is a deterministic function of Y. To prove that for $\alpha \in (1, \infty)$, the upper bound in (2.93) is achievable, we construct a mapping $P_{X \Leftarrow Y}$ such that X is a deterministic function of Y. That is, for every $y \in \mathcal{Y}$, there exists a unique $x_y \in \mathcal{X}$ such that $P(x_y|y) = 1$. Therefore, we have $x_y = \arg_x P_{X \Leftarrow Y}(y|x) > 0$. For $\alpha \in (1, \infty)$, from (2.6) and (2.30b) we have

$$\mathcal{L}_\alpha^{\max}(P_{X \Leftarrow Y}) = \sup_{P_X} \frac{\alpha}{\alpha - 1} \log \sum_{y \in \mathcal{Y}} \left(P_X^{\frac{1}{\alpha}}(x_y) P_{X \Leftarrow Y}(y|x_y) \right) \tag{2.94}$$

$$= \sup_{P_X} \frac{\alpha}{\alpha - 1} \log \sum_{x \in \mathcal{X}} P_X^{\frac{1}{\alpha}}(x); \tag{2.95}$$

in addition, since the function maximized in (2.95) is symmetric and concave in P_X, it is Schur-concave in P_X, and therefore, the optimal distribution of X achieving the supreme in (2.95) is uniform. Thus,

$$\mathcal{L}_\alpha^{\max}(P_{X\hookleftarrow Y}) = \log|\mathcal{X}|, \quad 1 < \alpha \leq \infty. \tag{2.96}$$

Therefore, maximal α-leakage achieves its maximal value $\log|\mathcal{X}|$ and $H(P_X)$ for $\alpha > 1$ and $\alpha = 1$, respectively, if and only if X is a deterministic function of Y. □

2.6.5 Proof for Theorem 4

To prove Thm. 4, we define a divergence function k_α for $\alpha > 1$ and provide a lower bound for its sum in the following definition and lemma, respectively.

Definition 2.6.1. *Given two discrete distributions P_Y and Q_Y over the support $\mathcal{Y}$, a divergence function k_α for $\alpha > 1$ is defined as*

$$k_\alpha(P_Y\|Q_Y) \triangleq \sum_y Q_Y(y)\left(\frac{P_Y(y)}{Q_Y(y)}\right)^\alpha. \tag{2.97}$$

In addition, the function $k_\alpha(P_Y\|Q_Y)$ is jointly convex in (P_Y, Q_Y), such that $k_\alpha(P_Y\|Q_Y) \geq 1$ with equality if and only if $P_Y = Q_Y$.

Lemma 2. *Let K be a positive integer with $K < \infty$. Given a group of distributions $\{P_k : k \in [1, K]\}$ and an arbitrary distribution P on a discrete set $\mathcal{Y}$, there is*

$$\sum_{k=1}^K k_\alpha(P_k\|P) \geq \sum_{k=1}^K k_\alpha(P_k\|P_c) = \left(\sum_y\left(\sum_{k=1}^K P_k(y)^\alpha\right)^{\frac{1}{\alpha}}\right)^\alpha, \tag{2.98}$$

with equality if and only if $P = P_c$, where P_c is given by

$$P_c(y) = \frac{1}{Z}\left(\sum_{k=1}^K P_k(y)^\alpha\right)^{\frac{1}{\alpha}}, \ \alpha \in [1, \infty] \tag{2.99}$$

where Z is the constant as

$$Z = \sum_y \left(\sum_{k=1}^K P_k(y)^\alpha \right)^{\frac{1}{\alpha}}, \tag{2.100}$$

which guarantees that P_c is a distribution.

Proof. From the definition k_α in (2.97), we have

$$\sum_{k=1}^K k_\alpha(P_k \| P) - \sum_{k=1}^K k_\alpha(P_k \| P_c) = \sum_{k=1}^K \sum_y P_k(y)^\alpha \left(P(y)^{1-\alpha} - P_c(y)^{1-\alpha} \right) \tag{2.101}$$

$$= \sum_y \left(\sum_{k=1}^K P_k(y)^\alpha \right) \left(P(y)^{1-\alpha} - P_c(y)^{1-\alpha} \right) \tag{2.102}$$

$$= \sum_y Z^\alpha P_c(y)^\alpha \left(P(y)^{1-\alpha} - P_c(y)^{1-\alpha} \right) \tag{2.103}$$

$$= Z^\alpha \sum_y \left(P_c(y)^\alpha P(y)^{1-\alpha} - P_c(y) \right) \tag{2.104}$$

$$= Z^\alpha (k_\alpha(P_c \| P) - 1) \geq 0 \tag{2.105}$$

with equality if and only if $P = P_c$. In addition, making use of the expression of P_c and Z in (2.99) and (2.100), respectively, we have

$$\sum_{k=1}^K k_\alpha(P_k \| P_c) = \sum_{k=1}^K \sum_y P_c(y) \left(\frac{P_k(y)}{P_c(y)} \right)^\alpha \tag{2.106}$$

$$= \sum_{k=1}^K \sum_y Z^{\alpha-1} \left(\sum_{k'=1}^K P_{k'}(y)^\alpha \right)^{\frac{1}{\alpha}} \frac{P_k(y)^\alpha}{\sum_{k'=1}^K P_{k'}(y)^\alpha} \tag{2.107}$$

$$= Z^{\alpha-1} \sum_y \left(\sum_{k'=1}^K P_{k'}(y)^\alpha \right)^{\frac{1}{\alpha}} \frac{\sum_{k=1}^K P_k(y)^\alpha}{\sum_{k'=1}^K P_{k'}(y)^\alpha} \tag{2.108}$$

$$= Z^\alpha \tag{2.109}$$

$$= \left(\sum_y \left(\sum_{k=1}^K P_k(y)^\alpha \right)^{\frac{1}{\alpha}} \right)^\alpha. \tag{2.110}$$

□

Making use of the results in Lemma 2, we prove Thm. 4 as follows.

Proof. From Thm. 2, we have that for $\alpha > 1$

$$\mathcal{L}_\alpha^{\max}(X \to Y) = \sup_{P_{\tilde{X}}} I_\alpha^{\mathrm{S}}(\tilde{X}, Y) \tag{2.111}$$

$$= \sup_{P_{\tilde{X}}} \inf_{Q_Y} D_\alpha(P_{\tilde{X}} P_{Y|X} \| P_{\tilde{X}} Q_Y) \tag{2.112}$$

$$= \sup_{P_{\tilde{X}}} \inf_{Q_Y} \frac{1}{\alpha - 1} \log \sum_x P_{\tilde{X}}(x) k_\alpha(P_{Y|X=x} \| Q_Y). \tag{2.113}$$

For $\alpha > 1$, the function $f : t \to \frac{1}{\alpha-1} \log t$ is increasing in $t \geq 0$. Therefore, we simplify the optimization in (2.113) as

$$\sup_{P_{\tilde{X}}} \inf_{Q_Y} \sum_x P_{\tilde{X}}(x) k_\alpha(P_{Y|X=x} \| Q_Y) \tag{2.114}$$

and provide a lower bound of (2.114) as follows. Since the divergence function k_α is joint convex in the pair of distributions, the objective function in (2.114) is joint convex in $(P_{Y|X}, Q_Y)$ for fixed $P_{\tilde{X}}$, and linear in $P_{\tilde{X}}$ for fixed $(P_{Y|X}, Q_Y)$. Therefore, the max-min equals to the min-max as followed:

$$\sup_{P_{\tilde{X}}} \inf_{Q_Y} \sum_x P_{\tilde{X}}(x) k_\alpha(P_{Y|X=x} \| Q_Y) = \inf_{Q_Y} \sup_{P_{\tilde{X}}} \sum_x P_{\tilde{X}}(x) k_\alpha(P_{Y|X=x} \| Q_Y) \tag{2.115}$$

$$= \inf_{Q_Y} \max_x \; k_\alpha(P_{Y|X=x} \| Q_Y) \tag{2.116}$$

$$\geq \inf_{Q_Y} \frac{\sum_x k_\alpha(P_{Y|X=x} \| Q_Y)}{|\mathcal{X}|} \tag{2.117}$$

$$\geq \frac{\sum_x k_\alpha(P_{Y|X=x} \| P_c)}{|\mathcal{X}|} \tag{2.118}$$

$$= \frac{1}{|\mathcal{X}|} \left(\sum_y \| P_{Y|X}(y|\cdot) \|_\alpha \right)^\alpha, \tag{2.119}$$

where the inequality in (2.118) is directly from (2.98) in Lemma 2 with equality if and only if

$$Q_Y(y) = P_c(y) = \frac{1}{Z} \| P_{Y|X}(y|\cdot) \|_\alpha, \tag{2.120}$$

with the constant $Z = \sum_y \|P_{Y|X}(y|\cdot)\|_\alpha$. Therefore, for any $P_{Y|X}$, we have

$$\mathcal{L}_\alpha^{\max}(X \to Y) \geq \frac{\alpha}{\alpha - 1} \log \frac{\sum_y \|P_{Y|X}(y|\cdot)\|_\alpha}{|\mathcal{X}|^{\frac{1}{\alpha}}}, \tag{2.121}$$

with equality if and only if the $P_{Y|X}$ guarantees that the divergence function $k_\alpha(P_{Y|X=x}\|P_c)$ are the same for all $x \in \mathcal{X}$, i.e., the $P_{Y|X}$ satisfies (2.40). □

2.6.6 Proof of Theorem 5

Let $\mathcal{Y}_1$ and $\mathcal{Y}_2$ be the alphabets of Y_1 and Y_2, respectively. For any $(y_1, y_2) \in \mathcal{Y}_1 \times \mathcal{Y}_2$, due to the Markov chain $Y_1 - X - Y_2$, the corresponding entry of the conditional probability matrix of (Y_1, Y_2) given X is

$$P(y_1 y_2|x) = P(y_1|x)P(y_2|xy_1) = P(y_1|x)P(y_2|x).$$

Therefore, for $\alpha \in (1, \infty)$

$$\begin{aligned}
&\mathcal{L}_\alpha^{\max}(X \to Y_1, Y_2) \\
=&\sup_{P_X} \frac{\alpha}{\alpha - 1} \log \sum_{y_1, y_2} \left(\sum_x P_X(x) P_{Y_1, Y_2|X}(y_1, y_2|x)^\alpha \right)^{\frac{1}{\alpha}} &(2.122)\\
=&\sup_{P_X} \frac{\alpha}{\alpha - 1} \log \sum_{y_1, y_2} \left(\sum_x P_X(x) P_{Y_1|X}(y_1|x)^\alpha P_{Y_2|X}(y_2|x)^\alpha \right)^{\frac{1}{\alpha}}. &(2.123)
\end{aligned}$$

Let $K(y_1) = \sum_{x \in \mathcal{X}} P_X(x) P_{Y_1|X}(y_1|x)^\alpha$, for all $y_1 \in \mathcal{Y}_1$, such that we can construct a set of distributions over $\mathcal{X}$ as

$$P_{\tilde{X}}(x|y_1) = \frac{P_X(x) P_{Y_1|X}(y_1|x)^\alpha}{K(y_1)}. \tag{2.124}$$

Therefore, from (2.123), $\mathcal{L}_\alpha^{\max}(X \to Y_1, Y_2)$ can be rewritten as

$$
\begin{aligned}
&\mathcal{L}_\alpha^{\max}(X \to Y_1, Y_2) \\
=&\sup_{P_X} \frac{\alpha}{\alpha-1} \log \sum_{y_1,y_2 \in \mathcal{Y}_1 \times \mathcal{Y}_2} \left(\sum_{x \in \mathcal{X}} K(y_1) P_{\tilde{X}}(x|y_1) P_{Y_2|X}(y_2|x)^\alpha \right)^{\frac{1}{\alpha}} && (2.125) \\
=&\sup_{P_X} \frac{\alpha}{\alpha-1} \log \sum_{y_1,y_2} \left(\left(\sum_x P_X(x) P_{Y_1|X}(y_1|x)^\alpha \right)^{\frac{1}{\alpha}} \right. \\
&\left. \cdot \left(\sum_x P_{\tilde{X}}(x|y_1) P_{Y_2|X}(y_2|x)^\alpha \right)^{\frac{1}{\alpha}} \right) && (2.126) \\
=&\sup_{P_X} \frac{\alpha}{\alpha-1} \log \sum_{y_1} \left(\left(\sum_x P_X(x) P_{Y_1|X}(y_1|x)^\alpha \right)^{\frac{1}{\alpha}} \right. \\
&\left. \cdot \sum_{y_2} \left(\sum_x P_{\tilde{X}}(x|y_1) P_{Y_2|X}(y_2|x)^\alpha \right)^{\frac{1}{\alpha}} \right) && (2.127) \\
\leq&\sup_{P_X} \frac{\alpha}{\alpha-1} \log \left(\sum_{y_1} \left(\sum_x P_X(x) P_{Y_1|X}(y_1|x)^\alpha \right)^{\frac{1}{\alpha}} \right. \\
&\left. \cdot \max_{y_1} \sum_{y_2} \left(\sum_x P_{\tilde{X}}(x|y_1) P_{Y_2|X}(y_2|x)^\alpha \right)^{\frac{1}{\alpha}} \right) && (2.128) \\
=&\sup_{P_X} \frac{\alpha}{\alpha-1} \log \left(\sum_{y_1} \left(\sum_x P_X(x) P_{Y_1|X}(y_1|x)^\alpha \right)^{\frac{1}{\alpha}} \right. \\
&\left. \cdot \sum_{y_2} \left(\sum_x P_{\tilde{X}}(x|y_1^\star) P_{Y_2|X}(y_2|x)^\alpha \right)^{\frac{1}{\alpha}} \right) && (2.129) \\
\leq&\sup_{P_X} \frac{\alpha}{\alpha-1} \log \sum_{y_1} \left(\sum_x P_X(x) P_{Y_1|X}(y_1|x)^\alpha \right)^{\frac{1}{\alpha}} \\
&+ \sup_{P_{\tilde{X}}} \frac{\alpha}{\alpha-1} \log \sum_{y_2} \left(\sum_x P_{\tilde{X}}(x) P_{Y_2|X}(y_2|x)^\alpha \right)^{\frac{1}{\alpha}} && (2.130) \\
=&\mathcal{L}_\alpha^{\max}(X \to Y_1) + \mathcal{L}_\alpha^{\max}(X \to Y_2), && (2.131)
\end{aligned}
$$

where $y_1^\star$ in (2.129) is the optimal y_1 achieving the maximum in (2.128). Therefore, the equality in (2.128) holds if and only if, for all $y_1 \in \mathcal{Y}_1$,

$$\sum_{y_2} \left(\sum_x P_{\tilde{X}}(x|y_1) P_{Y_2|X}(y_2|x)^\alpha \right)^{\frac{1}{\alpha}} = \sum_{y_2} \left(\sum_x P_{\tilde{X}}(x|y_1^\star) P_{Y_2|X}(y_2|x)^\alpha \right)^{\frac{1}{\alpha}}; \quad (2.132)$$

and the equality in (2.130) holds if and only if the optimal solutions $P_X^\star$ and $P_{\tilde{X}}^\star$ of the two maximizations in (2.130) satisfy, for all $x \in \mathcal{X}$,

$$P_{\tilde{X}}^\star(x) = \frac{P_X^\star(x) P_{Y_1|X}^\alpha(y_1^\star|x)}{\sum_{x \in \mathcal{X}} P_X(x) P_{Y_1|X}^\alpha(y_1^\star|x)}. \quad (2.133)$$

Now we consider $\alpha = 1$. For $Y_1 - X - Y_2$, we have

$$I(Y_2; X|Y_1) \le I(Y_2; X). \quad (2.134)$$

From Thm. 2, there is

$$\begin{aligned} \mathcal{L}_1^{\max}(X \to Y_1, Y_2) &= I(X; Y_1) + I(X; Y_2|Y_1) && (2.135)\\ &\le I(X; Y_1) + I(X; Y_2) && (2.136)\\ &= \mathcal{L}_1^{\max}(X \to Y_1) + \mathcal{L}_1^{\max}(X \to Y_2). && (2.137) \end{aligned}$$

For $\alpha = \infty$, we also have

$$\begin{aligned} \mathcal{L}_\infty^{\max}(X \to Y_1, Y_2) &= \log \sum_{y_1, y_2 \in \mathcal{Y}_1 \times \mathcal{Y}_2} \max_{x \in \mathcal{X}} P(y_1|x) P(y_2|x) && (2.138)\\ &\le \log \sum_{y_1, y_2 \in \mathcal{Y}_1 \times \mathcal{Y}_2} \left(\max_{x \in \mathcal{X}} P(y_1|x) \right) \left(\max_{x \in \mathcal{X}} P(y_2|x) \right) && (2.139)\\ &= \log \sum_{y_1 \in \mathcal{Y}_1} \max_{x \in \mathcal{X}} P(y_1|x) + \log \sum_{y_2 \in \mathcal{Y}_2} \max_{x \in \mathcal{X}} P(y_2|x) && (2.140)\\ &= \mathcal{L}_\infty^{\max}(X \to Y_1) + \mathcal{L}_\infty^{\max}(X \to Y_2). && (2.141) \end{aligned}$$

□

2.6.7 Proof of Theorem 6

For $\alpha > 1$, a function $f(t) = \frac{\alpha}{\alpha-1} \log t$ is monotonically increasing in $t > 0$. Therefore, To solve maximal α-leakage from X^n to Y^n, i.e.,

$$\mathcal{L}_\alpha^{\max}(X^n \to Y^n) = \sup_{P_{\tilde{X}^n}} \frac{\alpha}{\alpha-1} \log \sum_{y^n} \left(\sum_{x^n} P(x^n) P(y^n|x^n)^\alpha \right)^{\frac{1}{\alpha}}, \tag{2.142}$$

it is sufficient to consider

$$\sup_{P_{\tilde{X}^n}} \sum_{y^n} \left(\sum_{x^n} P(x^n) P(y^n|x^n)^\alpha \right)^{\frac{1}{\alpha}}. \tag{2.143}$$

For a memoryless $P_{Y^n|X^n}$ with no feedback, we simplify (2.143) as

$$\begin{aligned}
&\sup_{P_{\tilde{X}^n}} \sum_{y^n} \left(\sum_{x^n} \frac{P(x^n y^n)}{P(y^n|x^n)^{1-\alpha}} \right)^{\frac{1}{\alpha}} \\
&= \sup_{\prod_{i=1}^n P_{\tilde{X}_i|\tilde{X}_{i-1},\cdots,\tilde{X}_1}} \sum_{y_1,\cdots,y_n} \left(\sum_{x_1,\cdots,x_n} \right. \\
&\qquad \left. \prod_{i=1}^n \left(\frac{P\left(y_i|x_i, x_{i-1}y_{i-1}, \cdots, x_1y_1\right)}{P\left(y_i|x^n y_{i-1}, \cdots, y_1\right)^{1-\alpha}} \frac{P\left(x_i|x_{i-1}y_{i-1}, \cdots, x_1y_1\right)}{1} \right) \right)^{\frac{1}{\alpha}} && (2.144) \\
&= \sup_{\substack{P_{\tilde{X}_i|\tilde{X}_{i-1},\cdots,\tilde{X}_1} \\ i\in[1,n]}} \sum_{y_1,\cdots,y_n} \left(\sum_{x_1,\cdots,x_n} \prod_{i=1}^n \left(\frac{P\left(y_i|x_i, \cdots, x_1\right) P\left(x_i|x_{i-1}, \cdots, x_1\right)}{P\left(y_i|x^n\right)^{1-\alpha}} \right) \right)^{\frac{1}{\alpha}} && (2.145) \\
&= \sup_{\substack{P_{\tilde{X}_i|\tilde{X}_{i-1},\cdots,\tilde{X}_1} \\ i\in[1,n]}} \sum_{y_1,\cdots,y_n} \left(\sum_{x_1,\cdots,x_n} \prod_{i=1}^n P\left(y_i|x_i\right)^\alpha P\left(x_i|x_{i-1}, \cdots, x_1\right) \right)^{\frac{1}{\alpha}} && (2.146) \\
&\leq \sup_{\substack{P_{\tilde{X}_i} \\ i\in[1,n]}} \sum_{y_1,\cdots,y_n} \left(\sum_{x_1,\cdots,x_n} \prod_{i=1}^n P\left(y_i|x_i\right)^\alpha P\left(x_i\right) \right)^{\frac{1}{\alpha}} && (2.147) \\
&= \sup_{P_{\tilde{X}_1},\cdots,P_{\tilde{X}_n}} \sum_{y_1,\cdots,y_n} \left(\prod_{i=1}^n \sum_{x_i} P\left(x_i\right) P\left(y_i|x_i\right)^\alpha \right)^{\frac{1}{\alpha}} && (2.148)
\end{aligned}$$

$$= \sup_{\substack{P_{\tilde{X}_i} \\ i\in[1,n]}} \prod_{i=1}^{n} \left(\sum_{y_i} \left(\sum_{x_i} P\left(x_i\right) P\left(y_i|x_i\right)^{\alpha} \right)^{\frac{1}{\alpha}} \right) \tag{2.149}$$

$$= \sup_{P_{\tilde{X}_i}, i\in[1,n]} \prod_{i=1}^{n} 2^{\frac{\alpha-1}{\alpha} I_{\alpha}^{\mathrm{S}}(\tilde{X}_i;Y_i)} \tag{2.150}$$

where

- (2.144) are from the chain rule of probability;
- (2.145) and (2.146) are directly from the mechanism has no feedback and is memoryless, respectively;
- the equality in (2.147) holds if and only if the source is memoryless, i.e., $P_{\tilde{X}_i|\tilde{X}_{i-1},\cdots,\tilde{X}_1} = P_{\tilde{X}_i}$ for all $i \in [1,n]$;
- both (2.148) and (2.149) are from the distributive property of multiplication.

Therefore, we have for $\alpha > 1$,

$$\sup_{P_{\tilde{X}^n}} I_{\alpha}^{\mathrm{S}}(\tilde{X}^n;Y^n) = \sum_{i=1}^{n} \sup_{P_{\tilde{X}_i}} I_{\alpha}^{\mathrm{S}}(\tilde{X}_i;Y_i). \tag{2.151}$$

That is,

$$\mathcal{L}_{\alpha}^{\max}\left(X^n \to Y^n\right) = \sum_{i=1}^{n} \mathcal{L}_{\alpha}^{\max}\left(X_i \to Y_i\right). \tag{2.152}$$

For $\alpha = 1$, we have

$$I\left(X^n;Y^n\right) = \sum_{i,j=1}^{n} I\left(X_i;Y_j\middle|X_{i-1},\cdots,X_1,Y_{j-1},\cdots,Y_1\right) \tag{2.153}$$

$$= \sum_{i,j=1}^{n} I\left(X_i;Y_j\middle|X_{i-1},\cdots,X_1\right) \tag{2.154}$$

$$= \sum_{i=1}^{n} I\left(X_i;Y_i\middle|X_{i-1},\cdots,X_1\right) \tag{2.155}$$

$$\leq \sum_{i=1}^{n} I\left(X_i;Y_i\right) \tag{2.156}$$

where

- (2.153) is from the chain rule of MI;
- (2.154) and (2.155) are from the facts that the mechanism has no feedback and is memoryless, respectively;
- from [104, (2.122)], we know that for a Markov chain $X - Y - Z$, conditioning reduces mutual information, i.e., $I(X;Y|Z) \leq I(X;Y)$ with equality if and only if $I(X;Z) = 0$. Therefore, since for any $i \in [1,n]$ $(X_{i-1}, \cdots, X_1) - X_i - Y_i$, the equality in (2.147) holds if and only if the source is memoryless, i.e., $P_{\tilde{X}_i|\tilde{X}_{i-1},\cdots,\tilde{X}_1} = P_{\tilde{X}_i}$ for all $i \in [1,n]$.

□

2.7 Concluding Remarks

We have introduced two novel tunable measures for information leakage: α-leakage and maximal α-leakage. Specifically, for $0 < \alpha \leq \infty$, α-leakage is shown to be Arimoto mutual information; maximal α-leakage is shown to be mutual information and the Arimoto channel capacity for $\alpha = 1$ and $\alpha \in (0,1) \cup (1,\infty]$, respectively. Based on the equivalent expression of the two tunable leakage measures, we have shown that α-leakage is (i) non-negative, (ii) quasi-convex in $P_{Y|X}$, and (iii) satisfying post-processing inequality; and maximal α-leakage satisfies several useful properties, including: (i) quasi-convexity, (ii) data-processing inequalities: post-processing inequality and linkage inequality, (iii) sub-additivity, and (iv) additivity for memoryless mappings. These measures can find direct applications in privacy and secrecy problems. The choice of restricting either specific variables or all possible functions of a dataset determines the choice of α-leakage and maximal α-leakage measures, respectively. The choice of $1 \leq \alpha \leq \infty$ determines the specific adversarial action ranging from refining a belief for $\alpha = 1$ to guessing the best posterior for $\alpha = \infty$.

Chapter 3

ROBUSTNESS OF MAXIMAL α-LEAKAGE

To explore the effect of side information on the leakage of information, which is measured in term of maximal α-leakage, we first introduce an extended version of maximal α-leakage, called conditional maximal α-leakage, to involve the notion of side information. Maximal α-leakage measures the information leakage for scenarios that an adversary intends to learn an arbitrary function of original data from a released data. Following the scenario, the conditional maximal α-leakage is built under a reasonable assumption in privacy protection that the function of interest is conditionally independent of the released data given the original data and the side information the adversary has. Note that in the setting the side information an adversary poses can be arbitrarily related to the function of interest, the original data or the released data.

With good properties as a privacy metric including the robustness to any arbitrary side information, differential privacy (DP) [16] has emerged as the gold standard for privacy. However, it has the fatal shortage of leading to poor utilities [106]. To balance the advantage and disadvantage of DP, several relaxed versions of the pure DP has been proposed including Rényi differential privacy (RDP) [80] which is defined via the Rényi divergence of order $\alpha > 1$. In this chapter, we also show that for an arbitrary privacy mechanism, if it satisfies a certain level of privacy in terms of maximal α-leakage, it satisfies a corresponding level of RDP, and vice versa. Due to the robustness of RDP (inheriting from DP), the result also support the robustness

of maximal α-leakage.

3.1 Conditional Tunable Information Leakage Measures

Given a pair of original and released data (X, Y), let Z be the knowledge of some particular adversary or third-party about (X, Y). Before introducing the conditional maximal α-leakage, we introduce the following simpler measure, the conditional α-leakage. Here, the adversary is interested only in guessing X, rather than a function of X.

Definition 3.1.1 (Conditional α-Leakage)**.** *Given a joint distribution P_{XYZ} and an estimator $\hat{X}$ with the same support as X, the α-leakage from X to Y given Z is defined as*

$$\mathcal{L}_\alpha(X \to Y|Z) \triangleq \frac{\alpha}{\alpha-1} \log \frac{\max\limits_{P_{\hat{X}|Y,Z}} \mathbb{E}\left[P_{\hat{X}|Y,Z}(X|Y,Z)^{\frac{\alpha-1}{\alpha}}\right]}{\max\limits_{P_{\hat{X}|Z}} \mathbb{E}\left[P_{\hat{X}|Z}(X|Z)^{\frac{\alpha-1}{\alpha}}\right]} \tag{3.1}$$

for $1 < \alpha < \infty$ and by the continuous extension of (3.1) for $\alpha = 1$ and ∞.

The conditional α-leakage quantifies the maximal logarithmic gain in inferring various information about X when an adversary *with arbitrary side information Z* has access to Y. To understand the effect of the side information Z on leakage about *any function U* of X through Y, we define conditional maximal α-leakage as follows.

Definition 3.1.2 (Maximal Conditional α-Leakage)**.** *Given a joint distribution P_{XYZ}, for $1 \leq \alpha \leq \infty$, the conditional maximal α-leakage from X to Y given Z is defined as*

$$\mathcal{L}_\alpha^{max}(X \to Y|Z) \triangleq \sup_{U:U-X-Y|Z} \mathcal{L}_\alpha(U \to Y|Z) \tag{3.2}$$

where U represents any function of X and takes values from an arbitrary alphabet. Moreover, the expression $U - X - Y|Z$ represents the conditional Markov chain con-

straint where

$$P_{UXY|Z}(uxy|z) = P(x|z)P(u|xz)P(y|xz). \tag{3.3}$$

Therefore, the conditional Markov chain $U - X - Y|Z$ *is equivalent to* $U - (X, Z) - Y$.

Note that conditional maximal α-leakage takes side information Z into consideration via the conditional Markov chain $U - X - Y|Z$, which is equivalent to $U - (X, Z) - Y$. Therefore, conditional maximal α-leakage is designed under the two assumptions: side information Z can be arbitrarily related to X and U, and the released data Y will not provide more information about U than X and Z.

The Markov chain $U - X - Y$ models inferences for a function U of X from Y. To involve side information Z in the inferences, beyond the conditional Markov chain in Def. 3.1.2, there are two other possibilities:

(i) If the side information Z that an adversary has is arbitrarily related to the function of interest U, but conditionally independent of released data Y given X, we have $(U, Z) - X - Y$. For example, if X is an individual's public records *without* voter registration indicated by Z and Y is a noisy release of X, then when U is the political preference of this person, Z can provide extra information about U and is conditionally independent of Y given X.

(ii) If the side information Z does not provide more information about the function of interest U than original data X does, but can be arbitrarily related to the released data Y, we have $U - X - (Y, Z)$. For example, if X is an individual's public records *with* voter registration, Z is a noisy release of the voter registration in X, and Y is a update of Z, then when U is the political preference of this person, Z cannot provide extra information about U than X does but it can be helpful in inferring U from Y (i.e., the Markov chain $Z - X - Y$ does not hold).

Note that in either Markov chain mentioned above, U and Y are conditionally independent given X and Z. In this sense, the proposed conditional Markov chain generally models side information in privacy-protection problems.

3.2 The Robustness to Side Information

In this section, we explore the effect of side information on inferring any function of original data from released data. First, we simplify the expression of conditional maximal α-leakage, and then, compare leakages of a privacy mechanism measured by conditional maximal α-leakage and maximal α-leakage.

The following theorem simplifies the expression of the conditional α- leakage in (3.1) as a conditional Arimoto MI based on Arimoto conditional entropy.

Definition 3.2.1. *Given a joint distribution $P_{X,Y,Z}$, the conditional Arimoto mutual information, for $1 \leq \alpha \leq \infty$, between X and Y given Z is defined as*

$$I_\alpha^A(X;Y|Z) \triangleq H_\alpha^A(X|Z) - H_\alpha^A(X|YZ) \tag{3.4}$$

where $H_\alpha^A(\cdot|\cdot)$ indicates Arimoto conditional entropy.

Note that for $\alpha = 1$, the conditional Arimoto MI in (3.4) is exactly the conditional Shannon MI $I(X;Y|Z)$.

Theorem 7. *For $\alpha \in [1, \infty]$, conditional α-leakage defined in* (3.1) *simplifies to*

$$\mathcal{L}_\alpha(X \to Y|Z) = I_\alpha^A(X;Y|Z). \tag{3.5}$$

The proof hinges on solving the two optimal problems in (3.1) by using Karush-–Kuhn-–Tucker (KKT) conditions. As this proof is nearly identical to that of Theorem 1 in Chapter 2.

Based on the result of Thm. 7, we obtain a simplified expression for conditional maximal α-leakage. Specifically, the simplified expression for $\alpha > 1$ is related to a variant of the Sibson MI defined as follows.

Definition 3.2.2. *Let $P_{X,Y|Z=z}$ indicate a conditional joint distribution of X, Y given an event $Z = z$. The event-conditional Sibson MI between X and Y given $Z = z$ is defined*

$$I_\alpha^S(X;Y|Z=z) = \frac{\alpha}{\alpha-1} \log \sum_y \left(\sum_x P(x|z) P(y|x,z)^\alpha \right)^{\frac{1}{\alpha}} \tag{3.6}$$

for $1 < \alpha < \infty$ and by the continuous extension of (3.6) for $\alpha = 1$ and ∞.

Theorem 8. *For $\alpha \in [1, \infty]$, the conditional maximal α-leakage defined in (3.2) simplifies to*

$$\mathcal{L}_\alpha^{max}(X \to Y|Z) = \begin{cases} \sup\limits_{z \in \mathrm{supp}(Z)} \sup\limits_{P_{\tilde{X}|Z=z} \ll P_{X|Z=z}} I_\alpha^S(\tilde{X};Y|Z=z), & \alpha \in (1,\infty] \\ I(X;Y|Z), & \alpha = 1. \end{cases} \tag{3.7}$$

where $\mathrm{supp}(Z)$ indicates the support of Z and $I_\alpha^S(X;Y|Z=z)$ is defined in (3.6).

A detailed proof is in Subsection 3.4.1. Note that given a channel $P_{Y|X}$, there is $\sup_X I_\alpha^{\mathrm{S}}(X;Y) = \sup_X I_\alpha^{\mathrm{A}}(X;Y)$ for $1 \le \alpha \le \infty$, and the quantity is called Arimoto channel capacity [107, 90]. Thus, for $\alpha > 1$, conditional maximal α-leakage is the maximal conditional Arimoto channel capacity of channels (from X to Y) where the channel state is controlled by Z.

The following theorem shows a relationship between conditional maximal α-leakage and maximal α-leakage.

Theorem 9. *For conditional maximal α-leakage defined in (3.2), if $Z - X - Y$ holds, then*

$$\mathcal{L}_\alpha^{max}(X \to Y|Z) \le \mathcal{L}_\alpha^{max}(X \to Y). \tag{3.8}$$

A detailed proof is in Subsection 3.4.2. Thm. 9 shows that if side information (Z) and released data (Y) is conditionally independent on the original data (X), the amount of information that Y can leak about X will not increase. That is, if side information is not involved in generating the released data from the original data, in terms of maximal α-leakage, it will not help an adversary get more information about the original data from the released data. Therefore, the (unconditional) maximal α-leakage represents a bound not only on the amount of information leaked in Y about an arbitrary function of X, but it is also a bound on the amount of information leaked in Y about X to an adversary with *arbitrary side-information*, provided Y is generated from X using only private randomness. This gives significant new meaning to the maximal α-leakage. The following example illustrates the result in Thm. 9.

Example 3. *Let the original data X uniformly take values from the binary alphabet $\{0,1\}$, and the released data Y be generated by a binary symmetric channel with a crossover probability $0 < p < 0.5$. Here, the maximal α-leakage from X to Y is*

$$\mathcal{L}_\alpha^{max}(X \to Y) = \begin{cases} \log 2 + \frac{1}{\alpha-1} \log\left(p^\alpha + (1-p)^\alpha\right), & \alpha > 1 \\ \log 2 - H(p), & \alpha = 1 \end{cases} \tag{3.9}$$

where $H(p) = -p \log p - (1-p)\log(1-p)$. Let the side information $Z \in \{0,1\}$ be generated from X via a binary symmetric channel with a crossover probability $0 \leq q \leq 0.5$, such that $Z - X - Y$ holds. From Thm. 8, we know that $\mathcal{L}_\alpha^{max}(X \to Y|Z) = 0$ for $q = 0$, and if $q \neq 0$

$$\mathcal{L}_\alpha^{max}(X \to Y|Z) = \begin{cases} \log 2 + \frac{1}{\alpha-1} \log\left(p^\alpha + (1-p)^\alpha\right), & \alpha > 1 \\ H(p+q-2pq) - H(p), & \alpha = 1. \end{cases} \tag{3.10}$$

Therefore, $\mathcal{L}_\alpha^{max}(X \to Y|Z) \leq \mathcal{L}_\alpha^{max}(X \to Y)$ with equality if and only if $\alpha > 1$ or $q = 0.5$.

As a contrast, for the same binary (X, Y) in Example 3 we show a case in which the Markov chain $Z - X - Y$ does not hold, so that side information causes the released data leak more information about the original data.

Example 4. *Let side information $Z \sim Ber(p)$ and $Z \perp X$. Let $Y = X$ for $Z = 0$ and $Y = X \oplus 1$ for $Z = 1$, such that $P_{X,Y}$ is the same as that in Example 3. From Thm. 8, we have $\mathcal{L}_\alpha^{max}(X \to Y|Z) = \log 2 > \mathcal{L}_\alpha^{max}(X \to Y)$ from* (3.9).

3.3 Maximal α-Leakage and Rényi Differential Privacy

From Chapter 2, it is known that maximal α-leakage for $\alpha \in [1, \infty]$ incorporates MI (for $\alpha = 1$), quantifying an average privacy protection over both the input and output supports of a mechanism, and MaxL (for $\alpha = \infty$), characterizing a "semi-average" privacy protection over only the output support. The set of maximal α-leakage for various $\alpha \in [1, \infty]$ depends on statistical information of data, and therefore, is regarded as context-aware metrics. On the contrary, differential privacy (DP) quantifies the worst case of information leakage and is known as a context-free metric which is independent of the statistical information of data. Several variants of DP are proposed for the sake of preserving utilities, and one of them is defined via Rényi divergence and so called Rényi differential privacy (RDP) [80]. In this section, we show the "equivalence" of maximal α-leakage and RDP in the sense of capturing the same collection of mechanisms which provide a specified level of privacy protection. This result implies that maximal α-leakage can reach out to context-free metric, and therefore, extends the scope of information leakage measures that can be linked by maximal α-leakage.

The definition of RDP is based on the notion of adjacent datasets who differ only in one element [80]. To get rid of the limit of adjacent dataset, we extend RDP to the local privacy context [24] and formally define *local Rényi differential privacy* (LRDP)

in the following definition.

Definition 3.3.1. *A mechanism $P_{Y|X} : \mathcal{X} \to \mathcal{Y}$ satisfies (α, γ)-LRDP for any non-negative γ and $\alpha > 1$, if*

$$\sup_{x,x' \in \mathcal{X}} D_\alpha(P_{Y|X=x} \| P_{Y|X=x'}) \leq \gamma. \tag{3.11}$$

where $P_{Y|X=x}$ and $P_{Y|X=x'}$ are the two conditional probabilities of Y over the support $\mathcal{Y}$ given $X = x$ and $X = x'$, respectively.

Similarly, the guaranteed privacy of a mechanism can be determined via maximal α-leakage as shown below.

Definition 3.3.2. *A mechanism $P_{Y|X} : \mathcal{X} \to \mathcal{Y}$ satisfies (α, ξ)-maximal α-leakage for any non-negative ξ and $\alpha > 1$, if*

$$\mathcal{L}_\alpha^{\max}(P_{Y|X}) = \sup_{P_{\tilde{X}}} \inf_{Q_Y} D_\alpha(P_{Y|X} \| Q_Y | P_{\tilde{X}}) \leq \xi. \tag{3.12}$$

where $P_{\tilde{X}}$ and Q_Y are two arbitrary probability distributions over $\mathcal{X}$ and $\mathcal{Y}$, respectively.

The following two inequalities between Rényi divergence and total variation distance will be used to derive the inequalities between maximal α-leakage and LRDP.

Lemma 3. *([108, (1),(6)]) Let P and Q be two arbitrary probability distributions of the random variable X. For $\alpha > 1$, there is*

$$\frac{1}{2}|P - Q|_{\mathrm{TV}}^2 \log(e) \leq D_\alpha(P\|Q) \tag{3.13}$$

$$D_\alpha(P\|Q) \leq \log\left(1 + \frac{|P - Q|_{\mathrm{TV}}}{2 \min_x Q(x)}\right). \tag{3.14}$$

We present the connection of privacy captured by maximal α-leakage and LRDP in the following theorem.

Theorem 10. *For any mechanism $P_{Y|X}$ satisfying (α, γ)-LRDP with $\alpha > 1$, it satisfies (α, γ)-maximal α-leakage, i.e.,*

$$\mathcal{L}_\alpha^{\max}(X \to Y) \leq \gamma. \tag{3.15}$$

For any mechanism $P_{Y|X}$ satisfies (α, ξ)-maximal α-leakage ($\alpha > 1$) with the minimal conditional probability no less than $0 < \tau < 1$, i.e., $\min_{x,y} P_{Y|X}(y|x) \geq \tau$, it satisfies $(\alpha, \gamma(\xi, \tau))$-LRDP with $\gamma(\xi, \tau)$ defined as

$$\gamma(\xi, \tau) \triangleq \log\left(1 + \frac{1}{\tau}\sqrt{\frac{2\xi}{\log(e)}}\right) \tag{3.16}$$

Note that the results in Theorem 10 is tight for perfect privacy, i.e., $\gamma = 0$ or $\xi = 0$.

The proof details are in 3.4.3. From the result of Theorem 10, we conclude that the LRDP and maximal α-leakage are equivalent in the sense of the collection of mechanisms satisfying a finite level of LRDP is the same collection of mechanisms satisfying a corresponding finite maximal α-leakage, and vice versa.

3.4 Proof Details

3.4.1 Proof of Theorem 8

From Thm. 7, we can simplify $\mathcal{L}_\alpha^{\max}(X \to Y|Z)$ in (3.2) as

$$\mathcal{L}_\alpha^{\max}(X \to Y|Z) = \sup_{U - X - Y|Z} I_\alpha^{\mathrm{A}}(U; Y|Z). \tag{3.17}$$

For $\alpha = 1$, we have

$$\mathcal{L}_1^{\max}(X \to Y|Z) = \sup_{U : U - (X,Z) - Y} I(U; Y|Z). \tag{3.18}$$

Under the Markov chain $U - X - Y|Z$, by the data processing inequality, we have $I(U; Y|Z) \leq I(X; Y|Z)$ with equality if and only if $I(X; Y|U, Z) = 0$. Thus,

$$\mathcal{L}_1^{\max}(X \to Y|Z) = I(X; Y|Z). \tag{3.19}$$

Now consider $\alpha > 1$. We first upper bound $L_\alpha^{\max}(X \to Y|Z)$. To show that this is upper bounded by the expression in (3.7), we show that for any variable U satisfying the Markov chain $U - X - Y|Z$, the conditional α-leakage is upper bounded by this same expression. For any such U, we have

$$I_\alpha^{\mathrm{A}}(U;Y|Z) = \frac{\alpha}{\alpha-1} \log \frac{\sum_{y,z} \left(\sum_u P_{U,Y,Z}(u,y,z)^\alpha \right)^{\frac{1}{\alpha}}}{\sum_z \left(\sum_u P_{U,Z}(u,z)^\alpha \right)^{\frac{1}{\alpha}}} \tag{3.20}$$

$$\leq \frac{\alpha}{\alpha-1} \log \sup_{z \in \mathrm{supp}(Z)} \frac{\sum_{y} \left(\sum_u P_{U,Y,Z}(u,y,z)^\alpha \right)^{\frac{1}{\alpha}}}{\left(\sum_u P_{U,Z}(u,z)^\alpha \right)^{\frac{1}{\alpha}}} \tag{3.21}$$

$$= \sup_{z \in \mathrm{supp}(Z)} I_\alpha^{\mathrm{A}}(U;Y|Z=z) \tag{3.22}$$

$$\leq \sup_{z \in \mathrm{supp}(Z)} \sup_{P_{\tilde{X}|\tilde{U}}: P_{\tilde{X}|\tilde{U}} \ll P_{X|Z=z}} \sup_{P_{\tilde{U}}} I_\alpha^{\mathrm{A}}(\tilde{U};Y|Z=z) \tag{3.23}$$

$$= \sup_{z \in \mathrm{supp}(Z)} \sup_{P_{\tilde{X}|\tilde{U}}: P_{\tilde{X}|\tilde{U}} \ll P_{X|Z=z}} \sup_{P_{\tilde{U}}} I_\alpha^{\mathrm{S}}(\tilde{U};Y|Z=z) \tag{3.24}$$

$$\leq \sup_{z \in \mathrm{supp}(Z)} \sup_{P_{\tilde{X}} \ll P_{X|Z=z}} I_\alpha^{\mathrm{S}}(\tilde{X};Y|Z=z) \tag{3.25}$$

where

- the inequality in (3.21) is from the fact that for any nonnegative a_i, b_i,

$$\frac{\sum_i a_i}{\sum_i b_i} \leq \max_i \frac{a_i}{b_i}, \tag{3.26}$$

- (3.22) follows by the definition of Arimito MI,
- in (3.23), the variables are distributed according to $P_{\tilde{U}}(u) P_{\tilde{X}|\tilde{U}}(x|u) P_{Y|X,Z}(y|x,z)$,
- (3.24) follows because Arimoto and Sibson MIs have the same supremum over the input distribution,

- (3.25) follows from the facts that Sibson MI satisfies the data processing inequality, and $\tilde{U} - \tilde{X} - Y|Z = z$ forms a Markov chain.

We now lower bound $L_\alpha^{\max}(X;Y|Z)$ by constructing a specific U satisfying $U - X - Y|Z$. For a given $P_{X,Y,Z}$, let

$$z^* = \arg \sup_{z \in \text{supp}(Z)} \sup_{\substack{P_{\tilde{X}} \\ \ll P_{X|Z=z}}} \sum_y \left(\sum_x P_{\tilde{X}}(x) P_{Y|X,Z}(y|x,z)^\alpha \right)^{\frac{1}{\alpha}}. \tag{3.27}$$

We will define a variable U with alphabet consisting of several disjoint subsets. We use $\mathcal{X}_{z^*}$ to indicate the conditional support of X given $Z = z^*$, i.e., $\mathcal{X}_{z^*} \triangleq \{x \in \mathcal{X} : P_{X,Z}(x, z^*) > 0\}$. For each $x \in \mathcal{X}_{z^\star}$, let $\mathcal{U}_{x,z^\star}$ be disjoint, finite sets. Also let $\mathcal{U}_0$ be a finite set (disjoint from those above). The cardinality of each of these sets will be determined later. Finally, let the alphabet of U be $\mathcal{U} = \mathcal{U}_0 \cup \bigcup_{x \in \mathcal{X}_{z^\star}} \mathcal{U}_{x,z^\star}$. We define the conditional distribution $P_{U|X,Z}$ as follows. Let

$$P_{U|X,Z}(u|x,z) = \begin{cases} \frac{1}{|\mathcal{U}_{x,z^\star}|}, & z = z^\star,\ u \in \mathcal{U}_{x,z^\star} \\ \frac{1}{|\mathcal{U}_0|}, & z \neq z^\star,\ u \in \mathcal{U}_0 \\ 0, & \text{otherwise.} \end{cases} \tag{3.28}$$

For the constructed U above, the conditional Arimoto MI is

$$I_\alpha^{\text{A}}(U;Y|Z) = \frac{\alpha}{\alpha - 1} \log \frac{\sum_{y,z} \left(\sum_u P_{U,Y,Z}(u,y,z)^\alpha \right)^{\frac{1}{\alpha}}}{\sum_z \left(\sum_u P_{U,Z}(u,z)^\alpha \right)^{\frac{1}{\alpha}}}. \tag{3.29}$$

The numerator in (3.29) can be written as

$$\begin{aligned}
&\sum_{y,z}\left(\sum_{u}P_{U,Y,Z}(u,y,z)^{\alpha}\right)^{\frac{1}{\alpha}}\\
=&\sum_{y,z}\left(\sum_{u}\left(\sum_{x}P_{U|X,Z}(u|x,z)P_{X,Y,Z}(x,y,z)\right)^{\alpha}\right)^{\frac{1}{\alpha}} &(3.30)\\
=&\sum_{y,z\neq z^*}\left(|\mathcal{U}_0|\left(\sum_{x}\frac{1}{|\mathcal{U}_0|}P_{X,Y,Z}(x,y,z)\right)^{\alpha}\right)^{\frac{1}{\alpha}}\\
&+\sum_{y}\left(\sum_{x}|\mathcal{U}_{x,z^*}|\left(\frac{1}{|\mathcal{U}_{x,z^*}|}P_{X,Y,Z}(x,y,z^*)\right)^{\alpha}\right)^{\frac{1}{\alpha}} &(3.31)\\
=&\frac{1-P_Z(z^*)}{|\mathcal{U}_0|^{1-\frac{1}{\alpha}}}+\sum_{y}\left(\sum_{x}|\mathcal{U}_{x,z^*}|^{1-\alpha}P_{X,Y,Z}(x,y,z^{\star})^{\alpha}\right)^{\frac{1}{\alpha}} &(3.32)
\end{aligned}$$

where the simplification in (3.32) is from (3.28). A similar derivation for the denominator in (3.29) gives

$$\sum_{z}\left(\sum_{u}P_{U,Z}(u,z)^{\alpha}\right)^{\frac{1}{\alpha}}=\frac{1-P(z^*)}{|\mathcal{U}_0|^{1-\frac{1}{\alpha}}}+\left(\sum_{x}|\mathcal{U}_{x,z^*}|^{1-\alpha}P_{X,Z}(x,z^{\star})^{\alpha}\right)^{\frac{1}{\alpha}}. \tag{3.33}$$

Note that for $\alpha>1$, as $|\mathcal{U}_0|\to\infty$, $(1-P_Z(z^*))\frac{1}{|\mathcal{U}_0|^{1-\frac{1}{\alpha}}}\to 0$. Therefore, for $\alpha>1$ we have

$$\mathcal{L}_{\alpha}^{\max}(X\to Y|Z)\geq\frac{\alpha}{\alpha-1}\log\frac{\sum_{y}\left(\sum_{x}|\mathcal{U}_{x,z^*}|^{1-\alpha}P_{X,Y,Z}(x,y,z^{\star})^{\alpha}\right)^{\frac{1}{\alpha}}}{\left(\sum_{x}|\mathcal{U}_{x,z^*}|^{1-\alpha}P_{X,Z}(x,z^{\star})^{\alpha}\right)^{\frac{1}{\alpha}}} \tag{3.34}$$

$$=\frac{\alpha}{\alpha-1}\log\sum_{y}\left(\sum_{x\in\mathcal{X}_{z^*}}P_{Y|X,Z}(y|x,z^*)^{\alpha}\frac{|\mathcal{U}_{x,z^*}|^{1-\alpha}}{\sum_{x'\in\mathcal{X}_{z^*}}|\mathcal{U}_{x',z^*}|^{1-\alpha}}\right)^{\frac{1}{\alpha}}. \tag{3.35}$$

Let $\tilde{X}\in\mathcal{X}_{z^*}$ be random variable with a distribution $P_{\tilde{X}}(x)=\frac{|\mathcal{U}_{x,z^*}|^{1-\alpha}}{\sum_{x'\in\mathcal{X}_{z^*}}|\mathcal{U}_{x',z^*}|^{1-\alpha}}$. By properly choosing cardinalities $|\mathcal{U}_{x,z^*}|$, for $x\in\mathcal{X}_{z^*}$, we can approach an arbitrary distribution $P_{\tilde{X}}$ on the support $\mathcal{X}_{z^*}$. In addition, the lower bound in (3.34) holds for

any arbitrary choice of these cardinalities. Therefore, we have

$$\begin{aligned}
&\mathcal{L}_\alpha^{\max}(X \to Y|Z) \\
&\geq \sup_{\substack{P_{\tilde{X}} \\ \ll P_{X|Z=z^*}}} \frac{\alpha}{\alpha-1} \log \sum_y \left(\sum_x P_{\tilde{X}}(x) P_{Y|X,Z}(y|x,z^*)^\alpha \right)^{\frac{1}{\alpha}} && (3.36) \\
&= \sup_{z \in \mathrm{supp}(Z)} \sup_{\substack{P_{\tilde{X}} \ll \\ P_{X|Z=z^*}}} \frac{\alpha}{\alpha-1} \log \sum_y \left(\sum_x P_{\tilde{X}}(x) P_{Y|X,Z}(y|x,z)^\alpha \right)^{\frac{1}{\alpha}} && (3.37)
\end{aligned}$$

where (3.37) is from the definition of z^* in (3.27). From (3.25) and (3.37), we have that for $\alpha > 1$

$$\mathcal{L}_\alpha^{\max}(X \to Y|Z) = \sup_{z \in \mathrm{supp}(Z)} \sup_{P_{\tilde{X}} \ll P_{X|Z=z}} I_\alpha^{\mathrm{S}}(\tilde{X}; Y|Z=z). \tag{3.38}$$

□

3.4.2 Proof of Theorem 9

From Thm. 8, we have that for $\alpha > 1$

$$\begin{aligned}
\mathcal{L}_\alpha^{\max}(X \to Y|Z) =& \sup_z \sup_{P_{\tilde{X}} \ll P_{X|Z=z}} \frac{\alpha}{\alpha-1} \log \sum_y \left(\sum_x P_{\tilde{X}}(x) P_{Y|X}(y|x)^\alpha \right)^{\frac{1}{\alpha}} && (3.39) \\
\leq& \sup_{P_{\tilde{X}} \ll P_X} \frac{\alpha}{\alpha-1} \log \sum_y \left(\sum_x P_{\tilde{X}}(x) P_{Y|X}(y|x)^\alpha \right)^{\frac{1}{\alpha}} && (3.40) \\
=& \mathcal{L}_\alpha^{\max}(X \to Y) && (3.41)
\end{aligned}$$

where (3.39) holds because the Markov chain $Z - X - Y$ allows us to replace $P_{Y|X,Z}$ with $P_{Y|X}$; and the inequality in (3.40) is from the fact that for any z conditioning on Z can only reduce the support of X; and the equality in (3.41) is from Theorem. 2 in Chapter 2. For $\alpha = 1$, from Thm. 8 we have

$$\mathcal{L}_1^{\max}(X \to Y|Z) = I(X;Y|Z), \tag{3.42}$$

such that if $Z - X - Y$ holds,

$$I(X;Y|Z) \leq I(X;Y) = \mathcal{L}_1^{\max}(X \to Y), \tag{3.43}$$

where the inequality and equality are from [104, Sec. 2.8] and [**?** , Thm. 2], respectively. Therefore, for $Z - X - Y$, $\mathcal{L}_\alpha^{\max}(X \to Y|Z) \leq \mathcal{L}_\alpha^{\max}(X \to Y)$. □

3.4.3 Proof of Theorem 10

For any mechanism $P_{Y|X}$ satisfying (α, γ)-RDP, we calculate its maximal α-leakage as following:

$$\begin{aligned}
\mathcal{L}_\alpha^{\max}(P_{Y|X}) &= \sup_{P_{\tilde{X}}} \inf_{Q_Y} D_\alpha(P_{Y|X}\|Q_Y|P_{\tilde{X}}) && (3.44)\\
&= \sup_{P_{\tilde{X}}} \inf_{Q_Y} \frac{1}{\alpha - 1} \log \left(\sum_x P_{\tilde{X}}(x) \sum_y P(y|x)^\alpha Q_Y(y)^{1-\alpha} \right) && (3.45)\\
&= \inf_{Q_Y} \sup_{P_{\tilde{X}}} \frac{1}{\alpha - 1} \log \left(\sum_x P_{\tilde{X}}(x) \sum_y P(y|x)^\alpha Q_Y(y)^{1-\alpha} \right) && (3.46)\\
&= \inf_{Q_Y} \max_x \frac{1}{\alpha - 1} \log \left(\sum_y P(y|x)^\alpha Q_Y(y)^{1-\alpha} \right) && (3.47)\\
&\leq \max_x \frac{1}{\alpha - 1} \log \left(\sum_y P(y|x)^\alpha P(y|x')^{1-\alpha} \right), \forall x' \in \mathcal{X} && (3.48)\\
&\leq \max_{x',x} \frac{1}{\alpha - 1} \log \left(\sum_y P(y|x)^\alpha P(y|x')^{1-\alpha} \right) && (3.49)\\
&= \max_{x',x} D_\alpha(P_{Y|X=x}\|P_{Y|X=x'}) \leq \gamma && (3.50)
\end{aligned}$$

where

- the equality in (3.46) is from that $D_\alpha(P_{Y|X}\|Q_Y|P_{\tilde{X}})$ is convex in Q_Y and concave in $P_{\tilde{X}}$ given $P_{Y|X}$ and $\alpha > 1$,
- the equality in (3.47) is from the facts that the function $\frac{1}{\alpha-1}\log(t)$ is monotonically increasing in $t > 0$ and a convex combination is no greater than the

maximal value involved in the convex combination.

Given any mechanism $P_{Y|X}$, its maximal α-leakage is $\mathcal{L}_\alpha^{\max}(P_{Y|X})$ and it satisfies (α,γ)-LRDP with the smallest value of γ given by

$$\begin{aligned}
\gamma &= \max_{x',x} D_\alpha(P_{Y|X=x}\|P_{Y|X=x'}) && (3.51)\\
&\le \max_{x',x} \log\left(1+\frac{|P_{Y|X=x}-P_{Y|X=x'}|_{\mathrm{TV}}}{2\min_{x,y}P(y|x)}\right) && (3.52)\\
&\le \max_{x',x}\inf_{Q_Y} \log\left(1+\frac{|P_{Y|X=x}-Q_Y|_{\mathrm{TV}}+|Q_Y-P_{Y|X=x'}|_{\mathrm{TV}}}{2\min_{x,y}P(y|x)}\right) && (3.53)\\
&\le \max_{x',x}\inf_{Q_Y} \log\left(1+\frac{\sqrt{2D_\alpha(P_{Y|X=x}\|Q_Y)}+\sqrt{2D_\alpha(P_{Y|X=x'}\|Q_Y)}}{2\sqrt{\log(e)}\min_{x,y}P(y|x)}\right) && (3.54)\\
&\le \log\left(1+\frac{\max\limits_{x',x}\inf\limits_{Q_Y}\left(\sqrt{2D_\alpha(P_{Y|X=x}\|Q_Y)}+\sqrt{2D_\alpha(P_{Y|X=x'}\|Q_Y)}\right)}{2\sqrt{\log(e)}\min_{x,y}P(y|x)}\right) && (3.55)\\
&\le \log\Bigg(1+\frac{1}{2\sqrt{\log(e)}\min_{x,y}P(y|x)}\\
&\qquad\cdot\max_{x,x'}\inf_{Q_Y} 2\max\left\{\sqrt{2D_\alpha(P_{Y|X=x}\|Q_Y)},\sqrt{2D_\alpha(P_{Y|X=x'}\|Q_Y)}\right\}\Bigg) && (3.56)\\
&= \log\left(1+\frac{2\max_x \inf_{Q_Y}\sqrt{2D_\alpha(P_{Y|X=x}\|Q_Y)}}{2\sqrt{\log(e)}\min_{x,y}P(y|x)}\right) && (3.57)\\
&= \log\left(1+\frac{\sqrt{2\max_x \inf_{Q_Y} D_\alpha(P_{Y|X=x}\|Q_Y)}}{\sqrt{\log(e)}\min_{x,y}P(y|x)}\right) && (3.58)\\
&= \log\left(1+\frac{\sqrt{2\mathcal{L}_\alpha^{\max}(P(y|x))}}{\sqrt{\log(e)}\min_{x,y}P_{Y|X}(y|x)}\right) && (3.59)
\end{aligned}$$

where

- the inequality in (3.52) is from the inequality (3.14) by replacing P and Q with $P_{Y|X=x}$ and $P_{Y|X=x'}$, respectively,
- the inequality in (3.53) is due to the fact

$$|P-P'|_{\mathrm{TV}} \le |P-Q|_{\mathrm{TV}}+|Q-P'|_{\mathrm{TV}} \tag{3.60}$$

where P, P', Q are three arbitrary probability distributions of an alphabet [1] .

- the inequality in (3.54) is from the inequality (3.13) and $|P-Q|_{\mathrm{TV}} = |Q-P|_{\mathrm{TV}}$ for any two probability distributions P, Q.

- the inequality in (3.56) is from that for any given x, x'

$$\begin{aligned}&\inf_{Q_Y}\left(\sqrt{2D_\alpha(P_{Y|X=x}\|Q_Y)}+\sqrt{2D_\alpha(P_{Y|X=x'}\|Q_Y)}\right)\\ \leq& 2\inf_{Q_Y}\max\left\{\sqrt{2D_\alpha(P_{Y|X=x}\|Q_Y)},\sqrt{2D_\alpha(P_{Y|X=x'}\|Q_Y)}\right\}\end{aligned} \tag{3.61}$$

Therefore, by plugging $\mathcal{L}_\alpha^{\max}(P_{Y|X}) \leq \xi$ and $\min_{x,y} P_{Y|X}(y|x) \geq \tau$ into (3.59), we have

$$\gamma \leq \log\left(1+\frac{1}{\tau}\sqrt{\frac{2\xi}{\log(e)}}\right) = \gamma(\xi,\tau). \tag{3.62}$$

□

3.5 Concluding Remarks

We have shown that in a data publishing setting, when the released data is generated from original data via private randomness (i.e., side information is not involved in the generation), maximal α-leakage is robust to *arbitrary side information* an adversary may have. In addition, maximal α-leakage is equivalent to the local Rényi variation of DP, i.e., LRDP, in the sense of the collection of mechanisms satisfying a finite level of LRDP can be captured by a requirement of maximal α-leakage on mechanism, and vice versa. While DP is well-known for being robust to arbitrary side information, this result also supports the robustness of maximal α-leakage.

[1] Proof: $|P-P'|_{\mathrm{TV}} = \sum_i |P(i)-P'(i)| = \sum_i |P(i)-Q(i)+Q(i)-P'(i)| \leq \sum_i (|P(i)-Q(i)| + |Q(i)-P'(i)|) = |P-Q|_{\mathrm{TV}} + |Q-P'|_{\mathrm{TV}}$

Chapter 4

PRIVACY-UTILITY TRADEOFFS WITH A HARD DISTORTION CONSTRAINT

In a privacy-guaranteed data publishing setting, a data curator/provider uses a mapping called *privacy mechanism* to generate distorted versions of original data for releases. The privacy mechanism determines the fidelity of the released data. The higher the fidelity is, the more utility of the data is maintained, and meanwhile, the less privacy of the data is preserved. Therefore, a privacy-utility tradeoff (PUT) problem arises in the design of the privacy mechanism.

We consider the two different data publishing scenarios shown in Figs. 4.1 and 4.2: the first where the entirety of the data set X is considered private, and the second where the data set consists of two parts S and X, where only S is considered private. For the first case (as shown in Fig. 4.1 where X and Y represent the original and released data. An adversary intends to infer a function U of X from Y, and $\widehat{U}$ is the adversary's estimation of U. Generally, the function U is unknown to the data curator/provider), we use maximal α-leakage as the privacy metric, thereby limiting the inference of any private information about the data set represented by the function

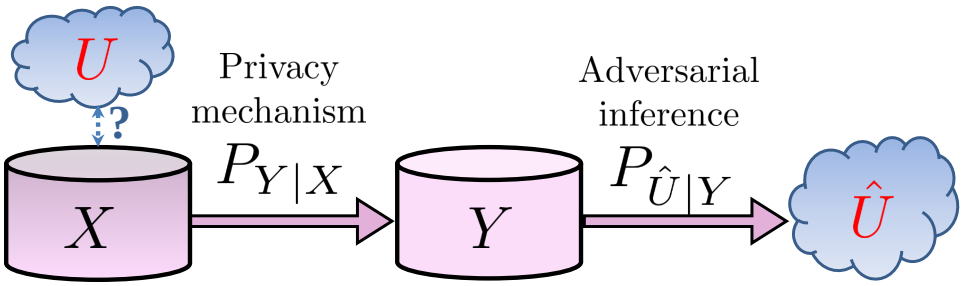


Figure 4.1: The First Privacy-Guaranteed Data Publishing Scenario: the Privacy Protection is for Entirely Sensitive Data Sets

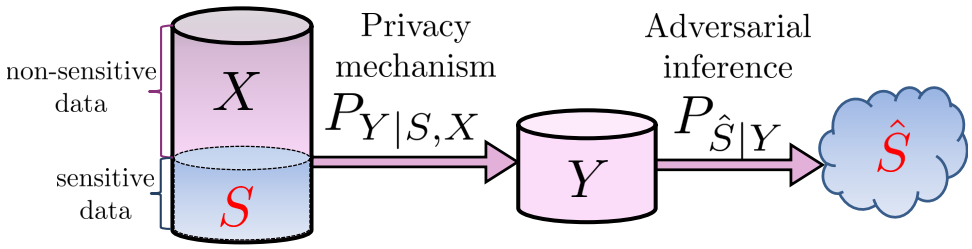


Figure 4.2: The Second Privacy-Guaranteed Data Publishing Scenario: the Privacy Protection is for Data Sets Consisting of Non-Sensitive and Sensitive Data

U. For the second case (as shown in Fig. 4.2, where X and S represent the non-sensitive and sensitive data in original data set, respectively, and Y is the released version of X. The adversary intends to infer S from Y, and $\hat{S}$ is the adversary's estimation of S), we use α-leakage as the privacy metric, thereby limited the inference only of the specific private information represented by S.

We measure utility in terms of a *hard distortion* metric, which constrains the privacy mechanism so that the distortion between each pair of original and released data is bounded with probability 1. Specifically, for the original and released data X, Y and a distortion function $d(\cdot,\cdot)$, the utility guarantee is modeled as the hard distortion constraint $d(X,Y) \leq D$ *with probability* 1, where D is the maximal permitted distortion. In other words, if a privacy mechanism $P_{Y|X}$ satisfies the hard distortion constraint, for any possible input x, all output y of the privacy mechanism must lie in a *non-empty* set $B_D(x)$ given by

$$B_D(x) \triangleq \{y : d(x,y) \leq D\}, \tag{4.1}$$

i.e., for any x with $P_X(x) > 0$, $P_{Y|X}(y|x) = 0$ if $y \notin B_D(x)$. Thus, a mathematical model of the PUT problem is given by

$$\inf_{P_{Y|X} \in \mathcal{P}_{Y|X}} \quad \mathcal{L}^{(\cdot)}_{(\cdot)}(X \to Y) \tag{4.2a}$$

$$\text{s.t.,} \quad d(X,Y) \leq D, \tag{4.2b}$$

where the set $\mathcal{P}_{Y|X}$ is the collection of stochastic matrices, and the superscript and subscript of $\mathcal{L}$ depend on the privacy measure under consideration (see Sec. 3.1 for notation).

Remark 2. *Note that given any input x, the hard distortion constraint in* (4.2b) *will force the conditional probabilities of the outputs that are not in $B_D(x)$ to be zero. Thus, this utility guarantee is* incompatible *with some privacy notions, which require each input to be mapped to all outputs with some positive probabilities; e.g., DP and any maximal f-leakage with $f(0) = \infty$.*

4.1 Leakage Measures Based on f-Divergence

We introduce two classes of information leakages derived from f-divergence, called *f-leakage* and *maximal f-leakage*. The f-leakage depends on the distribution of original data, and in contrast, maximal f-divergence only depends on the support of original data. We also show the relation between the f-divergence-based measures and maximal α-leakage for $\alpha = 1$ and $\alpha > 1$, respectively.

Recall that for a convex function $f : \mathbb{R}_+ \to \mathbb{R}$ such that $f(1) = 0$, an f-divergence D_f is a measure of the distance between two distributions given by

$$D_f(P_Y \| Q_Y) = \sum_y Q(y)\, f\left(\frac{P(y)}{Q(y)}\right). \tag{4.3}$$

Definition 4.1.1. *Given a joint distribution $P_{X,Y} = P_{Y|X} P_X$ and a f-divergence D_f, the f-leakage is defined as*

$$\mathcal{L}_f(X \to Y) = \inf_{Q_Y} D_f(P_{X,Y} \| P_X \times Q_Y), \tag{4.4}$$

and the maximal f-leakage is defined as

$$\mathcal{L}_f^{max}(X \to Y) = \sup_{P_{\tilde{X}}} \inf_{Q_Y} D_f(P_{Y|X} P_{\tilde{X}} \| P_{\tilde{X}} \times Q_Y), \tag{4.5}$$

where $P_{\tilde{X}}$ is a distribution over the support of P_X.

Note that in Definition 4.1.1, maximal f-leakage ($\mathcal{L}_f^{\max}$) depends on the distribution of X only through its support. In contrast, f-leakage ($\mathcal{L}_f$) depends fully on the distribution of X. Both measures depend on the chosen mechanism $P_{Y|X}$.

Recall that for $\alpha = 1$, maximal α-leakage is MI. Therefore, it is a special case of $\mathcal{L}_f(X \to Y)$ in (4.4) with $f(t) = t \log t$. Furthermore, for $\alpha > 1$, maximal α-leakage has a one-to-one relationship with a special case of $\mathcal{L}_f^{\max}$ in (4.5) for f given by

$$f_\alpha(t) = \frac{1}{\alpha - 1}(t^\alpha - 1), \tag{4.6}$$

such that D_f is the Hellinger divergence of order α [109]. The following lemma makes this observation precise.

Lemma 4. *For discrete random variables X and Y, the maximal α-leakage ($\alpha > 1$) from X to Y can be written as*

$$\mathcal{L}_\alpha^{max}(X \to Y) = \frac{1}{\alpha - 1} \log \big(1 + (\alpha - 1)\mathcal{L}_{f_\alpha}^{max}(X \to Y)\big), \tag{4.7}$$

where $\mathcal{L}_{f_\alpha}^{max}(X \to Y)$ is the $\mathcal{L}_f^{max}(X \to Y)$ in (4.5) for f_α given by (4.6) such that D_f is the Hellinger divergence of order α.

A detailed proof is in Section 4.5.1.

4.2 PUTs for Entirely Sensitive Data Sets

For the privacy-guaranteed publishing of an entirely sensitive data set shown in Fig. 4.1, we use maximal α-leakage as the privacy metric. From Section 4.1, we know that maximal α-leakage is a specific case of f-leakage and maximal f-leakage (in Def. 4.1.1) for $\alpha = 1$ and $\alpha > 1$, respectively. Hereby, we solve the PUT problems which minimize either f-leakage or maximal f-leakage, subject to a hard

distortion constraint. By applying the relations between maximal α-leakage and the f-divergence-based variants, we derive the optimal PUTs and optimal privacy mechanisms for the PUT problem with maximal α-leakage as the privacy measure. We denote an optimal PUT as $\mathrm{PUT}_{\mathrm{HD},\mathcal{L}_{(\cdot)}^{(\cdot)}}$, where HD and $\mathcal{L}_{(\cdot)}^{(\cdot)}$ in the subscript indicate the hard distortion and the involved privacy measure, respectively.

The following theorem characterizes the optimal tradeoff, denoted as $\mathrm{PUT}_{\mathrm{HD},\mathcal{L}_f}$, in (4.2) for the case that f-leakage is used as the privacy measure.

Theorem 11. *For any f-leakage $\mathcal{L}_f$ in (4.4) and a distortion function $d(\cdot,\cdot)$ with $B_D(x)$ in (4.1), the optimal PUT in (4.2) is given by*

$$PUT_{HD,\mathcal{L}_f}(D) = \inf_{P_{Y|X}:d(X,Y)\leq D} \mathcal{L}_f(X;Y), \tag{4.8}$$

$$= f(0) + \inf_{Q_Y} \mathbb{E}\left[Q_Y(B_D(X))\Big(f\big(\frac{1}{Q_Y(B_D(X))}\big) - f(0)\Big)\right]. \tag{4.9}$$

Moreover, letting $Q_Y^\star$ be the distribution achieving the infimum in (4.9), an optimal mechanism $P_{Y|X}^\star$ is given by

$$P_{Y|X}^\star(y|x) = \frac{\mathbf{1}\big(d(x,y)\leq D\big)Q_Y^\star(y)}{Q_Y^\star(B_D(x))}. \tag{4.10}$$

A detailed proof in Subsection 4.5.2. Note that as a result of the distribution dependence of the leakage measure $\mathcal{L}_f$ in (4.4), the optimal tradeoff in (4.9) is an *expected* function of X.

Making use of maximal f-divergence as the privacy constraint, the optimal PUT in (4.2) is given by $\mathrm{PUT}_{\mathrm{HD},\mathcal{L}_f^{\max}}$ in the following theorem.

Theorem 12. *For any maximal f-leakage $\mathcal{L}_f^{max}$ in (4.5), a distortion function $d(\cdot,\cdot)$ and $B_D(x)$ in (4.1), the optimal PUT in (4.2) is given by*

$$PUT_{HD,\mathcal{L}_f^{max}}(D) = \inf_{P_{Y|X}:d(X,Y)\leq D} \mathcal{L}_f^{max}(X\to Y), \tag{4.11}$$

$$= q^\star f((q^\star)^{-1}) + (1-q^\star)f(0), \tag{4.12}$$

with $q^\star$ defined as

$$q^\star \triangleq \sup_{Q_Y} \inf_x Q_Y(B_D(x)). \tag{4.13}$$

Moreover, letting $Q_Y^\star$ be the distribution achieving the supremum in (4.13)*, an optimal mechanism $P_{Y|X}^\star$ is given by* (4.10)*.*

A detailed proof is in Subsection 4.5.3.

Remark 3. *The PUTs in* (4.9) *and* (4.12) *simplify to finding an output distribution Q_Y that can be viewed as a "target" distribution, i.e., the optimal mechanism aims to produce this distribution as closely as possible, subject to the utility constraint. In particular, the resulting optimal mechanism (in* (4.10)*), for any input, distributes the outputs according to Q_Y while conditioning the output to be within a ball of radius D around the input. The optimization in* (4.13) *ensures that all inputs are uniformly masked while* (4.9) *provides average guarantees.*

The next corollary characterizes the optimal tradeoff $\mathrm{PUT}_{\mathrm{HD},\mathcal{L}_\alpha^{\max}}$ for maximal α-leakage. Recall that for $\alpha = 1$, $\mathcal{L}_1^{\max}$ equals $\mathcal{L}_f$ with $f(t) = t\log t$. For $\alpha > 1$, from the one-to-one relationship between $\mathcal{L}_\alpha^{\max}$ and $\mathcal{L}_{f_\alpha}^{\max}$ in (4.7), we know that finding $\mathrm{PUT}_{\mathrm{HD},\mathcal{L}_\alpha^{\max}}$ is equivalent to finding the optimal tradeoff $\mathrm{PUT}_{\mathrm{HD},\mathcal{L}_f^{\max}}$ in (4.11) for $\mathcal{L}_f^{\max} = \mathcal{L}_{f_\alpha}^{\max}$.

Corollary 1. *For maximal α-leakage, the optimal PUT in* (4.2) *is given by*

$$PUT_{HD,\mathcal{L}_\alpha^{max}}(D) = \inf_{P_{Y|X}:d(X,Y)\le D} \mathcal{L}_\alpha^{max}(X \to Y), \tag{4.14}$$

$$= \begin{cases} \inf\limits_{Q_Y} \mathbb{E}\left[\log \dfrac{1}{Q_Y(B_D(X))}\right], & \alpha = 1 \qquad (4.15a) \\ -\log q^\star, & \alpha > 1, \qquad (4.15b) \end{cases}$$

where $q^\star$ is defined in (4.13)*. Moreover, an optimal mechanism is given by* (4.10)*, where for $\alpha = 1$, $Q_Y^\star$ achieves the infimum in* (4.15a)*; and for $\alpha > 1$, $Q_Y^\star$ achieves the supremum in* (4.13)*.*

Remark 4. *Note that subject to a hard distortion constraint, the optimal privacy mechanism is always given by* (4.10)*. In particular, for maximal α-leakage, the optimal mechanism as well as the optimal PUT are identical for all $\alpha > 1$.*

4.3 PUTs for Data Sets Containing Non-Sensitive Data

For data sets containing both sensitive and non-sensitive data, indicated by S and X, respectively, as shown in Fig 4.2, the purpose of privacy protection is to limit information leakage of sensitive data while releasing non-sensitive data. We use α-leakage from S to Y as the privacy measure, where Y is the released version of X. Therefore, with $P_{Y|S,X}$ in the place of $P_{Y|X}$ in (4.2), we obtain the optimal PUT as

$$\mathrm{PUT}_{\mathrm{HD},\mathcal{L}_\alpha}(D) = \inf_{P_{Y|S,X}:d(X,Y)\le D} \mathcal{L}_\alpha(S;Y), \tag{4.16}$$

and for any (s,x) with $P_{S,X}(s,x) > 0$, the non-empty set B_D in (4.1) is given by

$$B_D(s,x) = \{y : d(x,y) \le D\}. \tag{4.17}$$

The following theorem lower bounds $\mathrm{PUT}_{\mathrm{HD},\mathcal{L}_\alpha}$.

Theorem 13. *The minimal leakage $PUT_{HD,\mathcal{L}_\alpha}$ $(1 \le \alpha \le \infty)$ in* (4.16) *is lower bounded by*

$$PUT_{HD,\mathcal{L}_\alpha}(D) \ge \begin{cases} \sum_{s,x} P(s,x) \log \Big(\max\limits_{y\in B_D(s,x)} \sum\limits_{s'\in\mathcal{S}_D(y)} P(s') \Big)^{-1}, & \alpha = 1 \\ \log \sum\limits_{s,x} \frac{P(s)P(s,x)}{\max\limits_s P_S(s)} \Big(\max\limits_{y\in B_D(s,x)} \sum\limits_{s'\in\mathcal{S}_D(y)} P(s') \Big)^{-1}, & \alpha = \infty \\ \frac{\alpha}{\alpha-1} \log \sum\limits_{s,x} \frac{P(s)^\alpha P(x|s)}{\|P_S\|_\alpha} \Big(\max\limits_{y\in B_D(s,x)} \sum\limits_{s'\in\mathcal{S}_D(y)} P(s')^\alpha \Big)^{\frac{1-\alpha}{\alpha}}, & \textit{else,} \end{cases}$$

where the nonempty set $B_D(s,x)$ of y is given by (4.17)*, and the set $\mathcal{S}_D(y)$ of s for each y is defined as*

$$\mathcal{S}_D(y) \triangleq \{s : \exists\, x,\, P_{S,X}(s,x) > 0, d(x,y) \le D\}. \tag{4.18}$$

The lower bound is tight if there exists an privacy mechanism $P_{Y|S,X} \in \mathcal{P}_{Y|S,X}(D)$ *such that*

(i) given (s,x)*, for any* y *with* $P(y|s,x) > 0$*,*

$$\sum_{s' \in \mathcal{S}_D(y)} P(s') = \max_{y' \in B_D(s,x)} \sum_{s' \in \mathcal{S}_D(y')} P(s'); \tag{4.19}$$

(ii) given any y *with* $P_Y(y) > 0$*, for any* $s \in S_D(y)$*,*

$$\sum_{x:d(x,y)\leq D} P(y|s,x)P(x|s) = \frac{P_Y(y)}{\sum\limits_{s' \in \mathcal{S}_D(y)} P(s')}, \tag{4.20}$$

where P_Y *is the marginal distribution of* Y *from the privacy mechanism* $P_{Y|S,X}$ *and* $P_{S,X}$*.*

The proof details are in Subsection 4.5.4.

Note that by using maximal α-leakage as the privacy measure, the setting for publishing data sets consisting of sensitive and non-sensitive data can be generalized to restrict leakages about *all* functions of the sensitive data. This will be addressed in future work.

4.4 Applications: PUTs for Hard Distortion Constraint

In this section, we apply the results in Sec. 4 to data sets and present the optimal PUTs for two examples: (i) using absolute distance between types (empirical distributions) of binary data sets as the distortion function; (ii) discrete data sets with Hamming distortion.

4.4.1 Example 1: Binary Data Sets with Hard Distortion on Types

When considering data set disclosure under privacy constraints, a reasonable goal is to design privacy mechanisms that preserve the statistics of the original data set

while preventing inference of each individual record (e.g., a sample or a row of the data set). Since the type (empirical distribution) of a data set captures its statistics, we quantify distortion as the distance between the type of the original and released data sets. We use maximal α-leakage to capture the gain of an adversary (with access to the released data set) in inferring any function of the original data set.

Let X^n be a random data set with n entries and Y^n be the corresponding released data set generated by a privacy mechanism $P_{Y^n|X^n}$. Entries of both X^n and Y^n are from the same alphabet $\mathcal{X}$. Let P_{x^n} and P_{y^n} indicate the types of input data set x^n and output data set y^n, respectively. We define the distortion function as the distance between types, given by

$$d_{\mathrm{T}}(x^n, y^n) = \max_{x \in \mathcal{X}} |P_{x^n}(x) - P_{y^n}(x)|, \tag{4.21}$$

and therefore, obtain $\mathrm{PUT}_{\mathrm{HD},\mathcal{L}_\alpha^{\max}}$ as in (4.14) but with data sets X^n, Y^n in place of single letters X, Y.Since types of n-length sequences take on only values that are multiples of $\frac{1}{n}$, this distortion function d_{T} takes on values of the form $\frac{m}{n}$, where $m \in [0, n]$.

We concentrate on binary data sets, i.e., $\mathcal{X} = \{0, 1\}$. Note that for binary data sets, we can simply write $d_{\mathrm{T}}(x^n, y^n) = |P_{x^n}(1) - P_{y^n}(1)|$. For a n-length binary data set, the number of types is $n + 1$. Therefore, all input and output data sets can be categorized into $n + 1$ type classes defined as

$$T(i) \triangleq \{x^n : nP_{x^n}(1) = i\}. \tag{4.22}$$

Theorem 14. *For binary data sets and the distortion function in* (4.21)*, given inte-*

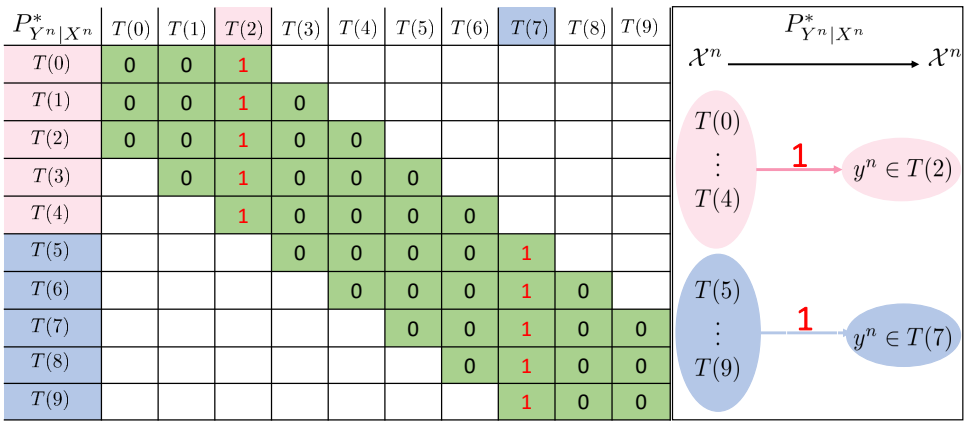


Figure 4.3: An Optimal Mechanism $\mathrm{PUT}_{\mathrm{HD},\mathcal{L}_\alpha^{\max}}\left(\frac{m}{n}\right)$ for $\alpha > 1$ with $(n, m) = (9, 2)$

gers n, m where $0 \leq m \leq n$, the optimal tradeoff for $\alpha > 1$ is

$$PUT_{HD,\mathcal{L}_\alpha^{max}}\left(\frac{m}{n}\right) = \min_{\substack{P_{Y^n|X^n}: \\ d_T(X^n,Y^n)\leq \frac{m}{n}}} \mathcal{L}_\alpha^{max}(X^n \to Y^n) \tag{4.23}$$

$$= \log\left\lceil \frac{n+1}{2m+1} \right\rceil. \tag{4.24}$$

An optimal privacy mechanism maps all input data sets in a type class to a unique output data set which is feasible and belongs to a type class in the set $\mathcal{T}^\star$ given by

$$\mathcal{T}^\star \triangleq \left\{ T(j) : j = l + (2m+1)k, k \in \left[0, \left\lceil \frac{n+1}{2m+1} \right\rceil - 1\right] \right\}, \tag{4.25}$$

where $l = m$ if $\lceil \frac{n+1}{2m+1} \rceil - \frac{n+1}{2m+1} \leq \frac{m}{2m+1}$, and otherwise, $l = n - \left(\lceil \frac{n+1}{2m+1} \rceil - 1\right)(2m+1)$.

A detailed proof is in Subsection 4.5.5. Let $(n, m) = (9, 2)$ such that from Thm. 14, we have $\mathrm{PUT}_{\mathrm{HD},\mathcal{L}_\alpha^{\max}}(\frac{2}{9}) = 1$ bit and $\mathcal{T}^\star = \{T(2), T(7)\}$. Fig. 4.3 shows the optimal mechanism, which maps all input data sets in $\{T(i) : i \in [0, 4]\}$ (resp. $\{T(i) : i \in [5, 9]\}$) to a *unique output data set* in $T(2)$ (resp. $T(7)$) with probability 1. Note that in Fig. 4.3, rows and columns are types of X^n and Y^n, respectively. The hard distortion forces conditional probabilities of outputs outside the feasible ball of given input to be zero. We highlight the conditional probabilities of feasible outputs in green, and give their values in the optimal mechanism.

4.4.2 Example 2: Hard Hamming Distortion on Data Sets

In the example, we consider hard Hamming distortion on data sets with entries from general finite alphabets. Formally, for data sets $x^n, y^n \in \mathcal{X}^n$, we define the Hamming distortion function on data sets as

$$d_{\mathrm{H}}(x^n, y^n) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(x_i \neq y_i). \tag{4.26}$$

Therefore, we obtain $\mathrm{PUT}_{\mathrm{HD},\mathcal{L}_\alpha^{\max}}$ as in (4.14) but with data sets X^n, Y^n in place of single letters X, Y.

Theorem 15. *For data sets from a finite alphabet $\mathcal{X}$ and Hamming distortion function, for any integers n, m where $0 \leq m \leq n$, the optimal tradeoff for $\alpha > 1$ is*

$$PUT_{HD,\mathcal{L}_\alpha^{max}}\left(\frac{m}{n}\right) = \min_{\substack{P_{Y^n|X^n}: \\ d_H(x^n,y^n) \leq \frac{m}{n}}} \mathcal{L}_\alpha^{max}(X^n \to Y^n) \tag{4.27}$$

$$= \log \frac{|\mathcal{X}|^n}{\sum_{i=0}^{m} \binom{n}{i} (|\mathcal{X}|-1)^i}. \tag{4.28}$$

An optimal privacy mechanism maps each input $x^n \in \mathcal{X}^n$ ***uniformly*** *to every feasible output, i.e., for all x^n, y^n where $d_H(x^n, y^n) \leq \frac{m}{n}$, $P_{Y^n|X^n}(y^n|x^n) = \frac{1}{\sum_{i=0}^{m} \binom{n}{i}(|\mathcal{X}|-1)^i}$.*

The key observation to reach the conclusion in Thm. 15 is that every output data set is in the same number of feasible balls, such that a uniform distribution over the output space leads to equal probability for the feasible ball of each input data set. The proof details are in Subsection 4.5.6. Fig. 4.4 illustrates the optimal mechanism in Thm. 15 for $\mathcal{X} = \{0, 1, 2\}$ and $(n, m) = (2, 1)$, where rows and columns are x^n and y^n, respectively. Note that we color the conditional probabilities of feasible outputs (respect to the hard Hamming distortion) and their values are the same as 0.2 in the optimal mechanism. Note that permuting items of a data set does not change the type but will lead to a non-zero Hamming distortion. The distortion on types in

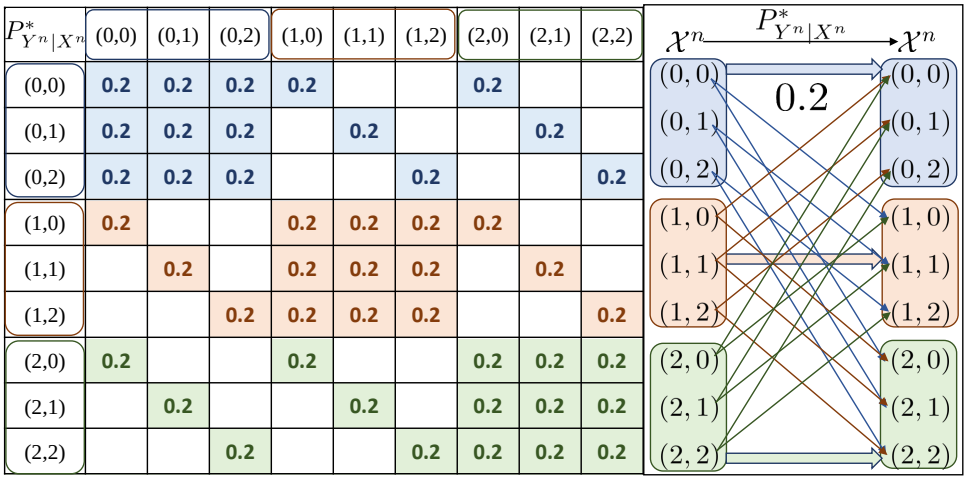


Figure 4.4: An Optimal Mechanism of (4.27) for $\alpha > 1$ with $(n, m) = (2, 1)$ and $\mathcal{X} = \{0, 1, 2\}$

(4.21) can be viewed as a relaxation of the Hamming distortion, in the sense that the set of feasible privacy mechanisms in (4.27) belongs to that in (4.23), i.e.,

$$\left\{P_{Y^n|X^n} : d_{\mathrm{H}}(x^n, y^n) \leq \frac{m}{n}\right\} \subset \left\{P_{Y^n|X^n} : d_{\mathrm{T}}(x^n, y^n) \leq \frac{m}{n}\right\}.$$

Therefore, for non-binary alphabets, the result in Thm. 15 upper bounds the minimal leakage in (4.23).

4.5 Proof Details

4.5.1 Proof for Lemma 4

Define the convex function

$$f_\alpha(t) = \frac{1}{\alpha - 1}(t^\alpha - 1), \tag{4.29}$$

then for the two distributions P and Q over the support $\mathcal{Y}$, we have a f-divergence $\mathcal{H}_\alpha(P\|Q)$, which is the Hellinger divergence of order α [109], given by

$$\mathcal{H}_\alpha(P\|Q) = \frac{1}{\alpha - 1}\left(\sum_{\mathcal{Y}} P(y)^\alpha Q(y)^{1-\alpha} - 1\right). \tag{4.30}$$

Therefore, the Rényi divergence can be written in terms of the Hellinger divergence as

$$D_\alpha(P\|Q) = \frac{1}{\alpha-1}\log(1+(\alpha-1)\mathcal{H}_\alpha(P\|Q)). \tag{4.31}$$

Thus, since $z \mapsto \frac{1}{\alpha-1}\log(1+(\alpha-1)z)$ is monotonically increasing in z for $\alpha > 1$, we can write maximal α-leakage as

$$\begin{aligned}
\mathcal{L}_\alpha^{\max}(X \to Y) &= \sup_{P_X} \inf_{Q_Y} D_\alpha(P_{X,Y}\|P_X \times Q_Y) && (4.32)\\
&= \frac{1}{\alpha-1}\log\Big(1+(\alpha-1)\sup_{P_X}\inf_{Q_Y}\mathcal{H}_\alpha(X\to Y)\Big) && (4.33)\\
&= \frac{1}{\alpha-1}\log\Big(1+(\alpha-1)\mathcal{L}_{\mathcal{H}_\alpha}(X\to Y)\Big). && (4.34)
\end{aligned}$$

That is, for $\alpha > 1$ maximal α-leakage is a monotonic function of the Hellinger divergence-based measure. □

4.5.2 Proof of Theorem 11

Given P_X, the collection of stochastic matrices is denoted as $\mathcal{P}_{Y|X}$. The feasible ball $B_D(x)$ around x is defined in (4.1). For the distribution dependent PUT in (4.8), we have

$$\begin{aligned}
&\mathrm{PUT}_{\mathrm{HD},\mathcal{L}_f}(D)\\
&= \inf_{\substack{P_{Y|X}\in\mathcal{P}_{Y|X}\\ :d(X,Y)\le D}} \inf_{Q_Y} D_f(P_{Y|X}P_X\|P_X\times Q_Y) && (4.35)\\
&= \inf_{Q_Y} \inf_{\substack{P_{Y|X}\in\mathcal{P}_{Y|X}\\ :d(X,Y)\le D}} \sum_{x\in\mathcal{X}} P_X(x) D_f(P_{Y|X=x}\|Q_Y) && (4.36)\\
&= \inf_{Q_Y} \sum_{x\in\mathcal{X}} P_X(x) \inf_{\substack{P_{Y|X=x}\\ Y\in B_D(x)}} \sum_{y\in\mathcal{Y}} Q_Y(y) f\left(\frac{P_{Y|X}(y|x)}{Q_Y(y)}\right) && (4.37)\\
&= \inf_{Q_Y} \sum_{x\in\mathcal{X}} P_X(x) \inf_{\substack{P_{Y|X=x}\\ Y\in B_D(x)}} \left(\sum_{y\in B_D(x)^c} Q_Y(y) f\left(\frac{P_{Y|X}(y|x)}{Q_Y(y)}\right)\right.
\end{aligned}$$

$$+\frac{Q_Y(B_D(x))}{Q_Y(B_D(x))}\sum_{y\in B_D(x)}Q_Y(y)f\left(\frac{P_{Y|X}(y|x)}{Q_Y(y)}\right)\Bigg) \tag{4.38}$$

$$=\inf_{Q_Y}\sum_{x\in\mathcal{X}}P_X(x)\inf_{\substack{P_{Y|X=x}\\ Y\in B_D(x)}}\Bigg(\sum_{y\in B_D(x)^c}Q_Y(y)f\left(0\right)+Q_Y(B_D(x))$$
$$\cdot\sum_{y\in B_D(x)}\frac{Q_Y(y)}{Q_Y(B_D(x))}f\left(\frac{P_{Y|X}(y|x)}{Q_Y(y)}\right)\Bigg) \tag{4.39}$$

$$\geq\inf_{Q_Y}\sum_{x\in\mathcal{X}}P_X(x)\inf_{\substack{P_{Y|X=x}\\ Y\in B_D(x)}}\left(Q_Y\left(B_D(x)^c\right)f(0)+Q_Y(B_D(x))f\left(\frac{1}{Q_Y(B_D(x))}\right)\right) \tag{4.40}$$

$$=f(0)+\inf_{Q_Y}\sum_{x\in\mathcal{X}}P_X(x)\left(Q_Y(B_D(x))\left(f\left(\frac{1}{Q_Y(B_D(x))}\right)-f(0)\right)\right) \tag{4.41}$$

where

- (4.36) follows from the fact that $D_f(P_{Y|X}P_X\|P_X\times Q_Y)$ is convex in $(P_{Y|X},Q_Y)$ for fixed P_X,
- (4.39) is directly from the hard distortion constraint $d(X;Y)\leq 0$ such that for any $y\notin B_D(x)$ $P_{Y|X}(y|x)=0$, and therefore, $\sum_{y\in B_D(x)}P_{Y|X}(y|x)=1$,
- (4.40) is from the Jensen's inequality such that

$$\sum_{y\in B_D(x)}\frac{Q_Y(y)}{Q_Y(B_D(x))}f\left(\frac{P_{Y|X}(y|x)}{Q_Y(y)}\right)$$
$$\geq f\left(\sum_{y\in B_D(x)}\frac{Q_Y(y)}{Q_Y(B_D(x))}\frac{P_{Y|X}(y|x)}{Q_Y(y)}\right) \tag{4.42}$$
$$=f\left(\frac{\sum\limits_{y\in B_D(x)}P_{Y|X}(y|x)}{Q_Y(B_D(x))}\right)=f\left(\frac{1}{Q_Y(B_D(x))}\right), \tag{4.43}$$

  with equality if and only if there is a mechanism $P_{Y|X}$ satisfying

$$\frac{P_{Y|X}(y|x)}{Q_Y(y)}=\frac{\mathbf{1}(y\in B_D(x))}{Q_Y(B_D(x))}. \tag{4.44}$$

Note that $f : \mathbb{R}_+ \to \mathbb{R}$ is a convex function, such that the function $tf(\frac{1}{t})$ is convex in $t \in \mathbb{R}_+$. Therefore, the objective function in (4.41) is convex in Q_Y. Furthermore, in (4.41) the feasible region of Q_Y is the probability distribution simplex over the set $\{B_D(x), x \in \mathcal{X}\}$. For finite supports $\mathcal{X}$ and $\mathcal{Y}$ of X and Y, respectively, the set $\{B_D(x), x \in \mathcal{X}\}$ is a compact, and therefore, the infimum in (4.41) is achievable. □

4.5.3 Proof of Theorem 12

Given P_X, the collection of stochastic matrices is denoted as $\mathcal{P}_{Y|X}$. The feasible ball $B_D(x)$ around x is defined in (4.1). For the distribution independent PUT in (4.11), we have

$$
\begin{aligned}
&\mathrm{PUT}_{\mathrm{HD},\mathcal{L}_f^{\max}}(D) \\
=&\inf_{\substack{P_{Y|X}\in\mathcal{P}_{Y|X} \\ :d(X,Y)\le D}} \sup_{P_{\tilde{X}}} \inf_{Q_Y} D_f(P_{\tilde{X}}P_{Y|X}\|P_{\tilde{X}}\times Q_Y) && (4.45)\\
=&\inf_{Q_Y} \sup_{P_{\tilde{X}}} \inf_{\substack{P_{Y|X}\in\mathcal{P}_{Y|X} \\ :d(X,Y)\le D}} D_f(P_{\tilde{X}}P_{Y|X}\|P_{\tilde{X}}\times Q_Y) && (4.46)\\
=&\inf_{Q_Y} \sup_{P_{\tilde{X}}} \inf_{\substack{P_{Y|X}\in\mathcal{P}_{Y|X} \\ :d(X,Y)\le D}} \sum_{x\in\mathcal{X}} P_{\tilde{X}}(x) D_f(P_{Y|X=x}\|Q_Y) && (4.47)\\
=&\inf_{Q_Y} \sup_{P_{\tilde{X}}} \sum_{x\in\mathcal{X}} P_{\tilde{X}}(x) \inf_{\substack{P_{Y|X=x} \\ Y\in B_D(x)}} \sum_{y\in\mathcal{Y}} Q_Y(y) f\left(\frac{P_{Y|X}(y|x)}{Q_Y(y)}\right) && (4.48)\\
=&\inf_{Q_Y} \sup_{P_{\tilde{X}}} \sum_{x\in\mathcal{X}} P_{\tilde{X}}(x) \inf_{\substack{P_{Y|X=x} \\ Y\in B_D(x)}} \left(\sum_{y\in B_D(x)} Q_Y(y) f\left(\frac{P_{Y|X}(y|x)}{Q_Y(y)}\right) \right. \\
&\left. + \sum_{y\in B_D(x)^c} Q_Y(y) f(0) \right) && (4.49)\\
=&\inf_{Q_Y} \sup_{P_{\tilde{X}}} \sum_{x\in\mathcal{X}} P_{\tilde{X}}(x) \inf_{\substack{P_{Y|X=x} \\ Y\in B_D(x)}} \left(Q_Y(B_D(x)) \sum_{y\in B_D(x)} \frac{Q_Y(y)}{Q_Y(B_D(x))} f\left(\frac{P_{Y|X}(y|x)}{Q_Y(y)}\right) \right.
\end{aligned}
$$

$$+ Q_Y(B_D(x)^c) f(0) \Bigg) \tag{4.50}$$

$$\geq \inf_{Q_Y} \sup_{P_{\tilde{X}}} \sum_{x \in \mathcal{X}} P_{\tilde{X}}(x) \inf_{\substack{P_{Y|X=x} \\ Y \in B_D(x)}} \left(Q_Y(B_D(x)) f\left(\frac{1}{Q_Y(B_D(x))} \right) \right.$$

$$\left. + Q_Y(B_D(x)^c) f(0) \right) \tag{4.51}$$

$$= \inf_{Q_Y} \sup_{P_{\tilde{X}}} \sum_{x \in \mathcal{X}} P_{\tilde{X}}(x) \left(Q_Y(B_D(x)) f\left(\frac{1}{Q_Y(B_D(x))} \right) \right.$$

$$\left. + \big(1 - Q_Y(B_D(x))\big) f(0) \right) \tag{4.52}$$

$$= \inf_{Q_Y} \sup_{P_{\tilde{X}}} \sum_{x \in \mathcal{X}} P_{\tilde{X}}(x)\, g\big(Q_Y(B_D(x))\big) \tag{4.53}$$

$$= \inf_{Q_Y} \sup_{x} g\big(Q_Y(B_D(x))\big) \tag{4.54}$$

where

- (4.46) and (4.48) follow from the fact that $D_f(P_{\tilde{X}} P_{Y|X} \| P_{\tilde{X}} \times Q_Y)$ is linear in $P_{\tilde{X}}$ for fixed $(P_{Y|X}, Q_Y)$ and convex in $(P_{Y|X}, Q_Y)$ for fixed $P_{\tilde{X}}$,
- (4.51) follows from the convexity of f and Jensen's inequality. The equality holds if and only if there exists a mechanism $P_{Y|X}$ satisfying (4.44).
- (4.53) results from $q \triangleq Q_Y(B_D(x))$ and

$$g(q) \triangleq q f(q^{-1}) + (1-q) f(0). \tag{4.55}$$

Due to the convexity of f, we have $f(q^{-1}) - f(0) \leq f'(q^{-1})\,(q^{-1} - 0)$, from which, the derivative $g'(q) = f(q^{-1}) - q^{-1} f'(q^{-1}) - f(0) \leq 0$. Therefore, the function g in (4.55) is non-increasing, such that (4.54) is simplified as $g(q^\star)$, where $q^\star$ is given by

$$q^\star \triangleq \sup_{Q_Y} \inf_{x} Q_Y(B_D(x)). \tag{4.56}$$

Note that in (4.56), the feasible region of Q_Y is the probability distribution simplex over the set $\{B_D(x), x \in \mathcal{X}\}$. For finite supports $\mathcal{X}$ and $\mathcal{Y}$ of X and Y, respectively,

the set $\{B_D(x), x \in \mathcal{X}\}$ is a compact, and therefore, the supremum in (4.56) is achievable.

□

4.5.4 Proof of Theorem 13

From Thm.1, we know that for $\alpha \geq 1$, α-leakage $\mathcal{L}_\alpha(S;Y)$ equals to Arimoto MI $I_\alpha^{\mathrm{A}}(S;Y)$. Since $I_\alpha^{\mathrm{A}}(S;Y) = H_\alpha(S) - H_\alpha^{\mathrm{A}}(S|Y)$ and $H_\alpha(S)$ is independent of $P_{Y|S,X}$, to minimize $I_\alpha^{\mathrm{A}}(S;Y)$ with respect to $P_{Y|S,X}$ can be simplified to maximize $H_\alpha^{\mathrm{A}}(S|Y)$. In addition, for $\alpha > 1$, the function $g : t \to \frac{\alpha}{1-\alpha}\log t$ is a monotonically non-increase function in $t > 0$. Therefore, the problem in (4.16) can be simplified to

$$\inf_{\substack{P_{Y|SX} \\ :d(X,Y)\leq D}} \sum_{y\in\mathcal{Y}} \Big(\sum_{s\in\mathcal{S}} P(s,y)^\alpha\Big)^{\frac{1}{\alpha}}. \tag{4.57}$$

The hard distortion on X and Y in (4.16) determines a collection of feasible x and therefore s for each y. We define the two collections for each $y \in \mathcal{Y}$ as

$$\mathcal{X}_D(y) \triangleq \{x \in \mathcal{X} : d(x,y) \leq D\}, \tag{4.58}$$

$$\mathcal{S}_D(y) \triangleq \{s \in \mathcal{S} : \exists\, x \in \mathcal{X}_D(y), P_{SX}(sx) > 0\}. \tag{4.59}$$

Note that both sets defined above are independent of the privacy mechanism $P_{Y|S,X}$.

For $\alpha \in (1,\infty)$, we have

$$\begin{aligned}
&\inf_{\substack{P_{Y|SX} \\ :d(X,Y)\leq D}} \sum_{y} \Big(\sum_{s} P(s,y)^\alpha\Big)^{\frac{1}{\alpha}} \\
&= \inf_{\substack{P_{Y|SX} \\ :d(X,Y)\leq D}} \sum_{y\in\mathcal{Y}} \Bigg(\sum_{s\in\mathcal{S}}\Big(\sum_{x\in\mathcal{X}} P(y|s,x)P(s,x)\Big)^\alpha\Bigg)^{\frac{1}{\alpha}} &(4.60)\\
&= \inf_{P_{Y|S,X}} \sum_{y} \Bigg(\sum_{\mathcal{S}_D(y)} \Big(\sum_{\mathcal{X}_D(y)} P(s,x,y)\Big)^\alpha + \sum_{\substack{x\notin\mathcal{X}_D(y) \\ s\notin\mathcal{S}_D(y)}} 0\Bigg)^{\frac{1}{\alpha}} &(4.61)
\end{aligned}$$

$$= \inf_{P_{Y|S,X}} \sum_{y} \Big(\sum_{s' \in \mathcal{S}_D(y)} P(s')^{\alpha} \Big)^{\frac{1}{\alpha}} \Bigg(\sum_{\mathcal{S}_D(y)} \frac{P(s)^{\alpha}}{\sum\limits_{s' \in \mathcal{S}_D(y)} P(s')^{\alpha}} \Big(\sum_{\mathcal{X}_D(y)} P(x,y|s) \Big)^{\alpha} \Bigg)^{\frac{1}{\alpha}} \tag{4.62}$$

$$\geq \inf_{P_{Y|S,X}} \sum_{y} \Big(\sum_{s' \in \mathcal{S}_D(y)} P(s')^{\alpha} \Big)^{\frac{1}{\alpha}} \Bigg(\sum_{\mathcal{S}_D(y)} \frac{P(s)^{\alpha}}{\sum\limits_{s' \in \mathcal{S}_D(y)} P(s')^{\alpha}} \Big(\sum_{\mathcal{X}_D(y)} P(x,y|s) \Big) \Bigg) \tag{4.63}$$

$$= \inf_{P_{Y|S,X}} \sum_{\substack{y, \mathcal{S}_D(y) \\ \mathcal{X}_D(y)}} \Big(\sum_{s' \in \mathcal{S}_D(y)} P(s')^{\alpha} \Big)^{\frac{1}{\alpha}-1} P(s)^{\alpha} P(x,y|s) \tag{4.64}$$

$$= \inf_{P_{Y|S,X}} \sum_{\substack{s,x \\ B_D(s,x)}} \Big(\sum_{s' \in \mathcal{S}_D(y)} P(s')^{\alpha} \Big)^{\frac{1}{\alpha}-1} P(s)^{\alpha} P(x,y|s) \tag{4.65}$$

$$\geq \inf_{P_{Y|S,X}} \sum_{s,x} P(s)^{\alpha} P(x|s) \min_{y \in B_D(s,x)} \Big(\sum_{s' \in \mathcal{S}_D(y)} P(s')^{\alpha} \Big)^{\frac{1}{\alpha}-1} \tag{4.66}$$

$$= \sum_{s,x} P(s)^{\alpha} P(x|s) \Big(\max_{y \in B_D(s,x)} \sum_{s' \in \mathcal{S}_D(y)} P(s')^{\alpha} \Big)^{\frac{1}{\alpha}-1}, \tag{4.67}$$

where

- (4.63) is directly from the concavity of the function $g_1 : t \to t^{\frac{1}{\alpha}}$ ($\alpha > 1$) and Jensen's inequality. The equality holds if and only if the optimal $P^{\star}_{Y|S,X}$ achieving the infimum satisfies that for all $s \in \mathcal{S}_D(y)$,

$$P^{\star}(y|s) = \sum_{x \in \mathcal{X}_D(y)} P^{\star}(y|sx) P(x|s) = \frac{P^{\star}_Y(y)}{\sum\limits_{s' \in \mathcal{S}_D(y)} P_S(s')}. \tag{4.68}$$

  where $P^{\star}_Y$ is the probability distribution of Y derived from $P^{\star}_{Y|S,X}$.

- in (4.65), $B_D(s,x)$ is the feasible ball defined in (4.17).

- the equality in (4.66) holds if and only if for any (s,x), all y with $P^{\star}(y|s,x) > 0$ lead to the same $\sum_{s' \in \mathcal{S}_D(y)} P(s')$.

- the equality in (4.67) is from the fact that the function $g : t \to t^{\frac{1}{\alpha}-1}$ is monotonically non-increasing in $t > 0$ for $\alpha \geq 1$.

Similarly, for $\alpha = \infty$, we have

$$\begin{aligned}
&\inf_{\substack{P_{Y|SX} \\ :d(X,Y)\leq D}} \sum_y P_Y(y) \max_s P_{S|Y}(s|y) \\
&= \inf_{P_{Y|S,X}} \sum_y P_Y(y) \max_{\mathcal{S}_D(y)} \left(\sum_{\mathcal{X}_D(y)} P_{S,X|Y}(s,x|y) \right) && (4.69) \\
&\geq \inf_{P_{Y|S,X}} \sum_y P(y) \left(\sum_{\mathcal{S}_D(y)} \frac{P(s)}{\sum_{s'\in\mathcal{S}_D(y)} P(s')} \sum_{\mathcal{X}_D(y)} P(s,x|y) \right) && (4.70) \\
&= \inf_{P_{Y|S,X}} \sum_{s,x} \sum_{B_D(s,x)} \frac{P(s)}{\sum_{s'\in\mathcal{S}_D(y)} P(s')} P(s,x,y) && (4.71) \\
&\geq \inf_{P_{Y|S,X}} \sum_{s,x} \sum_{B_D(s,x)} P(s,x,y) \min_{y\in B_D(s,x)} \frac{P(s)}{\sum_{s'\in\mathcal{S}_D(y)} P(s')} && (4.72) \\
&= \sum_{s,x} P(s,x)P(s) \left(\max_{y\in B_D(s,x)} \sum_{s'\in\mathcal{S}_D(y)} P(s') \right)^{-1}. && (4.73)
\end{aligned}$$

Note that the sufficient and necessary conditions for the equalities in (4.70) and (4.72) hold are the same as that for (4.63) and (4.66), respectively.

For $\alpha = 1$, $\mathcal{L}_\alpha(S \to Y) = I^{\mathrm{A}}(S;Y) = I(S;Y)$, such that

$$\begin{aligned}
&\mathrm{PUT}_{\mathrm{HD},\mathcal{L}_\alpha}(D) \\
&= \inf_{\substack{P_{Y|SX} \\ :d(X,Y)\leq D}} \sum_{s,y} P(s,y) \log \frac{P(s,y)}{P(s)P(y)} && (4.74) \\
&= \inf_{P_{Y|S,X}} \sum_y \sum_{\mathcal{S}_D(y)} \left(\Big(\sum_{\mathcal{X}_D(y)} P(s,x,y) \Big) \log \frac{\sum_{\mathcal{X}_D(y)} P(s,x,y)}{P(s)P(y)} \right) && (4.75) \\
&\geq \inf_{P_{Y|S,X}} \sum_y \left(\Big(\sum_{\mathcal{S}_D(y)} \sum_{\mathcal{X}_D(y)} P(s,x,y) \Big) \log \frac{\sum_{\mathcal{S}_D(y)} \sum_{\mathcal{X}_D(y)} P(s,x,y)}{\sum_{\mathcal{S}_D(y)} P(s)P(y)} \right) && (4.76) \\
&= \inf_{P_{Y|S,X}} \sum_{\substack{y,\mathcal{S}_D(y) \\ \mathcal{X}_D(y)}} P(s,x,y) \log \frac{1}{\sum_{s'\in\mathcal{S}_D(y)} P(s')} && (4.77) \\
&\geq \sum_{s,x} P(s,x) \min_{y\in B_D(s,x)} \log \frac{1}{\sum_{s'\in\mathcal{S}_D(y)} P(s')}. && (4.78)
\end{aligned}$$

Note that the inequality in (4.77) is from log-sum inequality in [104, Thm. 2.7.1], and the sufficient and necessary conditions for the equalities in (4.77) and (4.78) hold are the same as that for (4.63) and (4.66), respectively. □

4.5.5 Proof of Theorem 14

Define the distortion ball for the type-distance distortion in (4.21) as

$$B_m(x^n) \triangleq \left\{ y^n : |P_{x^n}(0) - P_{y^n}(0)| \leq \frac{m}{n} \right\}. \tag{4.79}$$

From Corollary 1, to find an optimal mechanism $P^{\star}_{Y^n|X^n}$, we need to find an output distribution $Q^{\star}_{Y^n}$ which optimizes (4.13) with x^n and y^n in place of x, y.

Note that for the hard distortion $|P_{x^n}(0) - P_{y^n}(0)| \leq \frac{m}{n}$, all data sets in a type class share the same group of feasible output data sets, and this feasible group can be represented by output type classes. Therefore, for any $x^n \in T(i)$ $(i \in [0, n])$, we rewrite $B_m(x^n)$ as

$$B_m(x^n) = B_m(T(i)) \triangleq \bigcup_{\substack{|i-j| \leq m \\ j \in [0,n]}} T(j). \tag{4.80}$$

We define a distribution Q_T of type classes for outputs as

$$Q_T(T(j)) \triangleq \sum_{y^n \in T(j)} Q_{Y^n}(y^n), \text{ for } j \in [0, n], \tag{4.81}$$

such that

$$q^{\star} = \sup_{Q_T} \inf_{i \in [0,n]} Q_T(B_m(T(i))). \tag{4.82}$$

The optimal distribution Q_T is determined by both upper and lower bounding $q*$ in (4.82). The upper bound is determined by restricting the optimization in (4.82) to a judicious choice of a small set of input types. The lower bound is a constructive scheme.

We define an index set $\mathcal{I}_T \subset [0, n]$ for types as

$$I_T \triangleq \left\{ l + (2m+1)k : k \in \left[0, \left\lceil \frac{n+1}{2m+1} \right\rceil - 1\right] \right\} \tag{4.83}$$

where $l = m$ if $\lceil \frac{n+1}{2m+1} \rceil \leq \frac{m+n+1}{2m+1}$, and otherwise, $l = n - \left(\lceil \frac{n+1}{2m+1} \rceil - 1\right)(2m+1)$. From the expression of $\mathcal{I}_T$ in (4.83), we observe that: (i) the difference between adjacent elements is $2m+1$; (ii) for the first and last elements,

- if $\lceil \frac{n+1}{2m+1} \rceil \leq \frac{m+n+1}{2m+1}$ holds, the first element is m and the last element is

$$m + (2m+1)\left(\left\lceil \frac{n+1}{2m+1} \right\rceil - 1\right) = (2m+1)\left\lceil \frac{n+1}{2m+1} \right\rceil - m - 1 \in [n-m, n], \tag{4.84}$$

  due to the inequalities $\frac{n+1}{2m+1} \leq \lceil \frac{n+1}{2m+1} \rceil \leq \frac{m+n+1}{2m+1}$;

- if $\lceil \frac{n+1}{2m+1} \rceil > \frac{m+n+1}{2m+1}$ holds, the last element is n and the first element is

$$n - \left(\left\lceil \frac{n+1}{2m+1} \right\rceil - 1\right)(2m+1) = n + 2m + 1 - \left\lceil \frac{n+1}{2m+1} \right\rceil (2m+1) \in [0, m), \tag{4.85}$$

  due to the inequalities $\frac{n+1}{2m+1} + 1 - \frac{1}{2m+1} \geq \lceil \frac{n+1}{2m+1} \rceil > \frac{m+n+1}{2m+1}$ for $n \in \mathbb{Z}_{++}$.

Therefore, it is not difficult to see that feasible balls of input type classes indexed by I_T are a partition of the set of all type classes, i.e.,

$$B_m(T(i_1)) \cap B_m(T(i_2)) = \emptyset \quad i_1, i_2 \in \mathcal{I}_T, \tag{4.86a}$$

$$\bigcup_{j \in [0,n]} T(j) = \bigcup_{i \in \mathcal{I}_T} B_m(T(i)). \tag{4.86b}$$

Therefore, the problem in (4.82) is upper bounded by

$$q^\star \leq \sup_{Q_T} \inf_{i \in \mathcal{I}_T} Q_T(B_m(T(i))) \tag{4.87}$$

$$\leq \sup_{Q_T} \frac{1}{|\mathcal{I}_T|} \sum_{i \in \mathcal{I}_T} Q_T(B_m(T(i))) \tag{4.88}$$

$$= \sup_{Q_T} \left(\left\lceil \frac{n+1}{2m+1} \right\rceil \right)^{-1} \sum_{j \in [0,n]} Q_T(T(j)) \tag{4.89}$$

$$= \left(\left\lceil \frac{n+1}{2m+1} \right\rceil \right)^{-1}, \tag{4.90}$$

where

- the inequality in (4.88) is from that the average probability of $B_m(T(i))$ over $i \in \mathcal{I}_T$ is no less than the minimal probability of $B_m(T(i))$ for $i \in \mathcal{I}_T$;
- the equality in (4.89) is from that the cardinality of $\mathcal{I}$ defined in (4.83) is $\lceil \frac{n+1}{2m+1} \rceil$;
- the equality in (4.90) is from that for any distribution over types $T(j)$ with $j \in [0, n]$, the sum of $Q_T(T(j))$ over $j \in [0, n]$ is 1.

To lower bound $q^\star$, we construct a distribution Q'_T as

$$Q'_T(T(j)) = \begin{cases} \left(\lceil \frac{n+1}{2m+1} \rceil \right)^{-1} & j \in I_T \\ 0 & \textit{otherwise.} \end{cases} \tag{4.91}$$

By (4.86) for each $i \in [0, n]$, there is a *unique* j satisfying $|i - j| \leq m$. Therefore, we lower bound (4.82) by

$$q^\star \geq \inf_i Q'_T(B_m(T(i))) \tag{4.92}$$

$$= \inf_i Q'_T\Big(\bigcup_{\substack{|i-j| \leq m \\ j \in \mathcal{I}_T}} T(j) \Big) \tag{4.93}$$

$$= \left(\left\lceil \frac{n+1}{2m+1} \right\rceil \right)^{-1}, \tag{4.94}$$

where the equality in (4.94) holds because for any $i \in [0, n]$, there is only one $j \in \mathcal{I}_T$ satisfying $|i - j| \leq m$ such that the union in (4.93) has exactly one element in it.

Therefore, $q^\star = \left(\lceil \frac{n+1}{2m+1} \rceil \right)^{-1}$ and the Q'_T defined in (4.91) achieve the optimum in (4.82). Thus, we can derive an optimal $Q^\star_{Y^n}$, which assigns the same non-zero probability to only one data set of each type classes indexed by I_T, i.e., $Q^\star_{Y^n}(y^n) = q^\star$

for one $y^n \in T(j)$ for each $j \in I_T$. Therefore, from (4.10) we have the corresponding optimal privacy mechanism, which maps all input data sets in one input type class to one feasible output data set with probability 1. $\square$

4.5.6 Proof of Theorem 15

For the Hamming distortion function on data sets in (4.26), the feasible ball $B_m(x^n)$ of any data set $x^n \in \mathcal{X}^n$ is given by

$$B_m(x^n) = \left\{ y^n \in \mathcal{X}^n : d_{\mathrm{H}}(x^n, y^n) \leq \frac{m}{n} \right\}. \tag{4.95}$$

For each $x^n \in \mathcal{X}^n$, the number of data sets having different values at exactly $k > 0$ different positions is $\binom{n}{k} (|\mathcal{X}| - 1)^k$, Therefore, the number of elements in its feasible ball $B_m(x^n)$ is

$$|B_m(x^n)| = \sum_{i=0}^{m} \binom{n}{i} (|\mathcal{X}| - 1)^i, \tag{4.96}$$

Note that the cardinality $|B_m(x^n)|$ in (4.96) of a feasible ball is independent of the input data set. We denote the cardinality as N_{ball}, i.e., $N_{\mathrm{ball}} \triangleq |B_m(x^n)|$. Due to the symmetric property of the Hamming distortion on data sets in (4.26), i.e., for any two data sets $x_1^n, x_2^n \in \mathcal{X}^n$, $x_1^n \in B_D(x_2)$ if and only if $x_2 \in B_D(x_1)$, we know that each output data set is in exactly N_{ball} different feasible balls (the example in Fig. 4.4 may help to figure out the above relationships). Therefore,

$$\begin{aligned}
q^\star &= \sup_{Q_{Y^n}} \inf_{x^n \in \mathcal{X}^n} Q_{Y^n}\left(B_m(x^n)\right) && (4.97)\\
&\leq \sup_{Q_{Y^n}} \frac{1}{|\mathcal{X}^n|} \sum_{x^n \in \mathcal{X}^n} Q_{Y^n}\left(B_m(x^n)\right) && (4.98)\\
&= \sup_{Q_{Y^n}} \frac{1}{|\mathcal{X}^n|} \sum_{x^n \in \mathcal{X}^n} \sum_{y^n \in B_m(x^n)} Q_{Y^n}\left(y^n\right) && (4.99)\\
&= \sup_{Q_{Y^n}} \frac{1}{|\mathcal{X}|^n} \sum_{\substack{x^n \in \mathcal{X}^n \\ y^n \in B_m(x^n)}} Q_{Y^n}\left(y^n\right) && (4.100)
\end{aligned}$$

$$= \sup_{Q_{Y^n}} \frac{1}{|\mathcal{X}|^n} \sum_{y^n \in \mathcal{X}^n} N_{\text{ball}} Q_{Y^n}(y^n) \tag{4.101}$$

$$= \frac{N_{\text{ball}}}{|\mathcal{X}|^n} \tag{4.102}$$

where

- the equality in (4.98) holds if and only if for an arbitrary pair of data sets x_1^n, x_2^n, there is

$$Q_{Y^n}(B_D(x_1^n)) = Q_{Y^n}(B_D(x_2^n)), \tag{4.103}$$

  which can be satisfied by a uniform distribution over $\mathcal{X}^n$, i.e., $Q^\star_{Y^n} = \frac{1}{|\mathcal{X}|^n}$.

- the equality in (4.101) holds because, for each y^n, the number of sequences x^n where $d_H(x^n, y^n) \leq \frac{m}{n}$ is exactly N_{ball}.

□

4.6 Concluding Remarks

We have explored PUTs in the context of hard distortion utility constraints. This utility constraint has the advantage that it allows the data curator to make specific, deterministic guarantees on the quality of the published data set. Focusing on maximal α-leakage and its f-divergence-based variants, under a hard distortion constraint, we have shown that: (i) for all $\alpha > 1$, we obtain the same optimal privacy mechanism and optimal PUT, which are independent of the distribution of the original data (or data sets); (ii) for $\alpha = 1$, the optimal mechanism differs and depends on the distribution of the original data (or data sets). In other words, for this distortion measure, the tunable privacy measure behaves as either MI or MaxL, which provides the insight that the α-loss with $\alpha < \infty$ can be sufficient to capture the soft 0-1 in applications.

The conjecture, that in a specific PUT problem with maximal α-leakaga as the privacy measure, it is sufficient to consider a limited range of α instead of to $\alpha = \infty$,

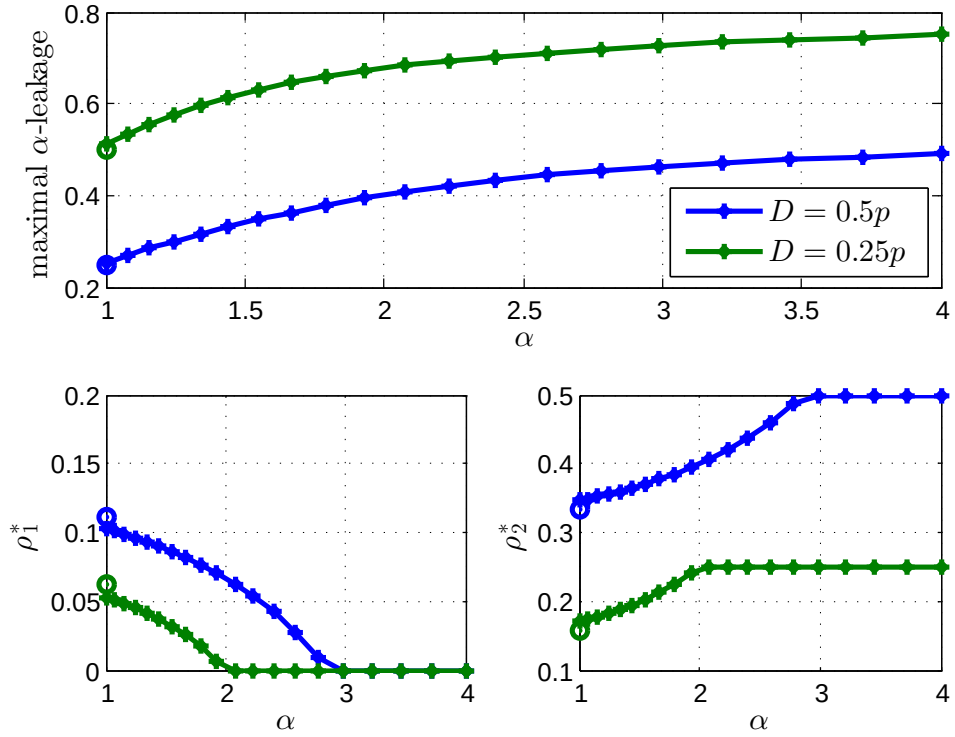


Figure 4.5: Numerical Results of the Optimization Problem in (4.105) for $p = 0.4$ and $D = 0.5p$ or $D = 0.25p$

is also supported by the following PUT problem with average distortion constraint Consider the following PUT problem that minimizes maximal α-leakage subject to the average Hamming distortion constraint:

$$\min_{P_{Y|X}} \quad \mathcal{L}_\alpha^{\max}(X \to Y) \tag{4.104a}$$

$$\text{s.t.,} \quad \sum_{x,y \in \mathcal{X}} P_{X,Y}(x,y)\mathbf{1}\left(y \neq x\right) \leq D \tag{4.104b}$$

where $0 < D < \min_x P_X(x)$ determines the upper bound of the permitted average Hamming distortion. Let $X, Y \in \{0,1\}$ and X follow the Bernoulli distribution Bern(p) $(0 < p < 1)$, i.e., $P_X(1) = p$. We represent the privacy mechanism $P_{Y|X}$ via the two crossover probabilities $P_{Y|X}(1|0) = \rho_1$ and $P_{Y|X}(0|1) = \rho_2$. By solving the supremum in the expression of maximal α-leakage, for $< 1\alpha < \infty$, the optimization

in (4.104) can be simplified as

$$\min_{\rho_1,\rho_2} \quad \frac{\alpha}{\alpha-1}\log\Bigg(\big((1-\rho_1)^\alpha(1-\rho_2)^\alpha-(\rho_1\rho_2)^\alpha\big)^{\frac{1}{\alpha}} \\ \Big(\big((1-\rho_1)^\alpha-\rho_2^\alpha\big)^{\frac{1}{1-\alpha}}+\big((1-\rho_2)^\alpha-\rho_1^\alpha\big)^{\frac{1}{1-\alpha}}\Big)^{\frac{\alpha-1}{\alpha}}\Bigg) \tag{4.105a}$$

$$\text{s.t.,} \quad (1-p)\rho_1+p\rho_2\le D. \tag{4.105b}$$

Fig. 4.5 shows the optimal values and mechanisms in (4.105) for $p=0.4$ and $D=0.5p$ or $D=0.25p$, respectively, where the above subplot is the curve of minimal values of maximal α-leakage versus the values of α (circles indicate $\alpha=1$ and stars are for $1.02\le\alpha\le 4$), and the left and right subplots below show the two crossover probabilities ρ_1^* and ρ_2^* in the optimal privacy mechanisms, respectively. From the plots, we can see that for $\alpha=1.02$, the optimal mechanism is a little different from that of mutual information [104, Fig. 10.3] due to the fact that as α tends to 1, the limit of Arimoto channel capacity is Shannon channel capacity instead of mutual information. We also observe that as α grows, the crossover probability ρ_1 gradually reduce to 0 and for enough large values of α (i.e., $\alpha\ge 2$ for $D=0.25p$ and $\alpha\ge 3$ for $D=0.5p$), the optimal mechanism is the same as $\alpha=\infty$. The example shows that for the average binary Hamming distortion, maximal α-leakage no longer behaves only as either of the two extrema, i.e., $\alpha=1$ and $\alpha=\infty$, but it is sufficient to study a limited range of α instead of to $\alpha=\infty$.

Chapter 5

APPLICATIONS OF α-LOSS IN MACHINE LEARNING: CENSORING AND FAIRNESS

In this chapter, we apply α-loss to the constrained minimax formulation proposed in [64] and develop a framework of generating censored and fair universal representations (CFUR) of data to provide fairness for multiple a priori unknown downstream tasks and censor multiple sensitive features, simultaneously. We also evaluate the performance of the CFUR framework on publicly available data sets including the UCI Adult and the UTKFace. All proof details are in Section 5.4.

5.1 Preliminaries

We consider a data set $\mathcal{D}$ with n entries where each entry is denoted as (S, X, Y) where $S \in \mathcal{S}$ indicates sensitive features, $X \in \mathcal{X}$ is a collection of non-sensitive features, and $Y \in \mathcal{Y}$ indicates a priori unknown target (non-sensitive) features to be learned. Let $\hat{Y} \in \mathcal{Y}$ be a prediction of Y. *We note that S and Y can be a collection of features or labels (e.g., S can be gender, race, or sexual orientation, while Y could be age, facial expression, etc.); for ease of writing, we use the term variable to denote both single and multiple features/labels.* Instances of X, S, and Y are denoted by x, s and y, respectively. We assume that each entry (X, S, Y) is independent and identically distributed (i.i.d.) according to $P(X, S, Y)$.

Recent results on fairness in learning applications guarantees that for a specific target variable, the prediction of a machine learning model is accurate with respect to (*w.r.t.*) the target variable but unbiased *w.r.t.* a sensitive variable. While more than two dozen measures for fairness have been proposed, the three oft-used fair-

ness measures are demographic parity (DemP), equalized odds (EO), and equality of opportunity (EoO). DemP ensures complete independence between the prediction of the target variable and sensitive variable, and thus, this notion of fairness favors utility the least, especially when the target and sensitive variables are correlated [35]. EO enforces this independence conditioned on the target variable thereby ensuring equal rates for true and false positives (wherein the target variable is binary) for all demographics. EoO ensures EO for the true positive case alone [35].

For the sake of completeness, we review these definitions briefly. We note that these definitions are often aimed at sensitive (S) and target (Y) features that are binary, and in reviewing these definitions below, we make this assumption too. However, we note that these definitions can be generalized to the non-binary setting; indeed, our own generalizations of these definitions as applied to representation setting do not make such an assumption of binary features.

Definition 5.1.1 ([35]). *A predictor $f(S, X) = \hat{Y}$ satisfies*

- *demographic parity (DemP) w.r.t. the sensitive variable S, if $\hat{Y}$ and S are independent, i.e.,*

$$\Pr(\hat{Y} = 1|S = 1) = \Pr(\hat{Y} = 1|S = 0) \tag{5.1}$$

- *equalized odds (EO) w.r.t. the sensitive variable S and target variable Y, if $\hat{Y}$ and S are independent conditional on Y, i.e.,*

$$\Pr(\hat{Y} = 1|S = 1, Y = y) = \Pr(\hat{Y} = 1|S = 0, Y = y), \quad y \in \{0, 1\} \tag{5.2}$$

- *equality of opportunity (EoO) w.r.t. the sensitive variable S and target variable Y, if $\hat{Y}$ and S are independent conditional on $Y = 1$, i.e.,*

$$\Pr(\hat{Y} = 1|S = 1, Y = 1) = \Pr(\hat{Y} = 1|S = 0, Y = 1). \tag{5.3}$$

We begin by first defining the notions of censoring and fairness for representations. In particular, in the censoring context, our goal is to introduce a definition ensuring that the censored representation limits leakage of sensitive variables learned by any adversary whose learning strategy is designed by minimizing a expected loss function. It is crucial to note that censoring will in general not give the kind of strong privacy guarantees provided by differential privacy but can be relevant for some applications where releasing a representation is crucial.

Definition 5.1.2 (Censored Representations)**.** *A representation X_r of X is censored w.r.t. the sensitive variable S against a learning adversary $h(\cdot)$, whose performance is evaluated via a loss function $\ell(h(X_r), S)$, if for an optimal adversarial strategy $h_g^* = \text{argmin}_h \, \mathbb{E}[\ell(h(g(X)), S)]$ corresponding to any (randomized) function $g(X)$*

$$\mathbb{E}[\ell(h_g^*(g(X)), S)] \leq \mathbb{E}[\ell(h_{g_r}^*(X_r), S)], \tag{5.4}$$

where $X_r = g_r(X)$ and the expectation is over h, g (or g_r), X, and S.

To motivate the generation of fair representations, we now extend the definition of DemP for representations. Indeed it is known that fair representations can be used to ensure fair classification (see, for example, [35]). We formally define fair representation and prove that such representations ensure fair classification.

Definition 5.1.3 (Demographically Fair Representations)**.** *Let $\mathcal{X}_r$ and $\mathcal{S}$ be the supports of X_r and S, respectively. A representation X_r of X satisfies DemP w.r.t. the sensitive variable S if X_r and S are independent, i.e., for any $x_r \in \mathcal{X}_r$ and $s, s' \in \mathcal{S}$,*

$$\Pr(X_r = x_r | S = s) = \Pr(X_r = x_r | S = s'). \tag{5.5}$$

Given this definition, we now prove that a fair representation in the sense of DemP will guarantee that any downstream learning algorithm making use of the fair representation is fair (in the sense of DemP) *w.r.t.* the sensitive label S.

Theorem 16 (Fair Learning via Fair Representation)**.** *Given a data set consisting of sensitive, non-sensitive, and target variables* (S, X, Y)*, respectively, if a fair representation* $X_r = g(X)$ *satisfies DemP w.r.t.* S*, then any learning algorithm* $f : \mathcal{X}_r \to \mathcal{Y}$ *satisfies DemP w.r.t.* S*.*

The proof of Theorem 16 is basically depending on the data-processing inequality of mutual information and the proof details are in Section 5.4.1.

Remark 5. *Note that the definitions of EO and EoO in Def. 5.1.1 explicitly involve a downstream learning application, and therefore, the design of a fair representation needs to include a classifier explicitly. In contrast to the universal representation setting considered here, such targeted representations and the ensuing fair classifiers provide guarantees only for those targeted* Y *features. In this limited context, however, one can still define a representation* X_r *as ensuring EO (w.r.t. to* S*) in classifying* Y *if the predicted output learned from* X_r*, i.e.,* $\hat{Y}(X_r)$*, is independent of* S *conditioned on* Y*.*

One simple approach to obtain a fair/censored representation X_r is by choosing $X_r = N$ where N is a random variable independent of X and S. However, such an X_r has no downstream utility (quantified, for example, via downstream task accuracy). More generally, the design of X_r has to ensure utility, and thus, there is a tradeoff between guaranteeing fairness/censoring and assuring a desired level of utility. The CFUR framework enables quantifying these tradeoffs formally as described in the next section.

5.2 Censored and Fair Universal Representations via Generative Adversarial Models

Formally, the CFUR model consists of two components, an generative decorrelator and an adversary as shown in Fig. 5.1. The goal of the generative decorrelator $g : \mathcal{S} \times \mathcal{X} \to \mathcal{X}_r$ is to actively decorrelate S from X_r while that of the adversary $h : \mathcal{X}_r \to \mathcal{S}$ is to infer S. Thus, in general, $g(X, S)$ is a randomized mapping that outputs a representation $X_r = g(X, S)$. Note that the design of $g(\cdot)$ depends on both X and S; however, we note that S *may not necessarily* be an input to the generative decorrelator though it will always affect the design of $g(\cdot)$ via the adversarial training process. In contrast, the role of the adversary is captured via $h(X_r)$, the adversarial decision rule (classifier) in inferring the sensitive variable S as $\hat{S} = h(g(X))$ from the representation $g(X)$. In general, the hypothesis h can be a *hard decision rule* under which $h(g(X))$ is a direct estimate of S or a *soft decision rule* under which $h(g(X)) = P_h(\cdot|g(X))$ is a distribution over $\mathcal{S}$. To quantify the adversary's performance, we use a loss function $\ell(h(g(X = x)), S = s)$ defined for every pair (x, s). Thus, the adversary's expected loss *w.r.t.* X and S is

$$L(h, g) \triangleq \mathbb{E}[\ell(h(g(X)), S)], \tag{5.6}$$

where the expectation is taken over $P(X, S)$ and the randomness in g and h.

Intuitively, the generative (since it randomizes to decorrelate) decorrelator would like to minimize the adversary's ability to learn S reliably from X_r. This can be trivially achieved by releasing an X_r independent of X. However, such an approach provides no utility for data analysts who want to learn non-sensitive variables Y from X_r. To overcome this issue, we capture the loss incurred by perturbing the original data via a distortion function $d(x_r, x)$, which measures how far the original data $X = x$ is from the processed data $X_r = x_r$. Ensuring statistical utility in turn

requires constraining the average distortion $\mathbb{E}[d(g(X), X)]$ where the expectation is taken over $P(X, S)$ and the randomness in g.

5.2.1 CFUR: Framework and Theoretical Results

To publish a censored and fair representation X_r, the data curator wishes to learn a decorrelator g that guarantees both censoring/fairness (in the sense that it is difficult for the adversary to learn S from X_r) as well as utility (in the sense that it does not distort the original data too much). In contrast, for a fixed decorrelator g, the adversary would like to find a (potentially randomized) function h that minimizes its expected loss, which is equivalent to maximizing the negative of the expected loss. This leads to a constrained minimax game between the decorrelator and the adversary given by [64]

$$\min_{g(\cdot)} \max_{h(\cdot)} \quad -\mathbb{E}[\ell(h(g(X)); S)] = -L(h, g), \tag{5.7a}$$

$$s.t. \quad \mathbb{E}[d(g(X), X)] \leq D. \tag{5.7b}$$

where $L(h, g)$ is the adversary's expected loss defined in (5.6) and the constant $D \geq 0$ determines the allowable distortion of the representation and the expectation is taken over $P(X, S)$ and the randomness in g and h. Note that if needed, the sensitive variable S can be used by the generative decorrelator g to generate the representation.

Note that the inner maximization in (5.7) is free of the distortion constraint, and therefore, for any given generative decorrelator $g(\cdot)$, the corresponding optimal adversary's strategy h^* that minimizes the adversary's expected loss $L(h, g)$ can be obtained by solving this unconstrained maximization. The minimax game in (5.7) places no restrictions on the adversary. Indeed, different loss functions and decision rules lead to different adversarial models (see Table 5.1) [110, 64]. In the following theorem, we show that the minimax game in (5.7) can produce censored representations against

	Loss function $\ell(h(g(x)), s)$	Optimal adversarial strategy h^*	Adversary type
Squared loss	$(h(g(x)) - s)^2$	$\mathbb{E}[S\|g(x)]$	MMSE adversary
0-1 loss	$\begin{cases} 0, & h(g(x)) = s \\ 1, & \text{otherwise} \end{cases}$	$\underset{s' \in \mathcal{S}}{\operatorname{argmax}} P(s'\|g(x))$	MAP adversary
soft 0-1 loss	$1 - P_h(s\|g(x))$		
Log-loss	$-\log P_h(s\|g(x))$	$P(s\|g(x))$	Belief refining adversary

Table 5.1: Adversaries Captured by the CFUR Framework via Various Loss Functions

chosen adversarial models.

Theorem 17. *For sufficiently large distortion bound D, the constrained minimax optimization in* (5.7) *generates a universal representation X_r that is censored w.r.t. to the sensitive variable S.*

The proof of Theorem 17 is based on the observation that the minimax game in (5.7) produces the generative correlator g that maximizes the expected loss of the best adversary of the adversarial model captured by the chosen loss function.

In the sequel, we use the *α-loss* (introduced in Chapter 2) as the adversary's loss function and show that under this setting, the minimax game in (5.7) is equivalent to minimizing α-leakage, which is also a proxy of DemP, and therefore, the framework provides censoring and fairness guarantees, simultaneously.

Given a decorrelator g and a adversarial strategy, for any pair of (s, x), the α-loss ($\alpha \in [1, \infty]$) of inferring $S = s$ given $X_r = g(x)$ is given by

$$\ell_\alpha(h(g(x)), s) = \lim_{\alpha' \to \alpha} \frac{\alpha'}{\alpha' - 1} \left(1 - P_h(s|g(x))^{\frac{\alpha'-1}{\alpha'}}\right). \tag{5.8}$$

As shown in Chapter 2, by tuning the parameter $\alpha \in [1, \infty]$, the α-loss captures a variety of information-theoretic adversaries ranging from a belief-refining adversary (for $\alpha = 1$) via the log-loss function $\ell(h(g(x)), s) = -\log P_h(s|g(x))$ to a *maximal a posteriori* (MAP) adversary (for $\alpha = \infty$) via the soft 0-1 loss function $\ell(h(g(x)), s) = 1 - P_h(s|g(x))$. In the following proposition, we show the optimal adversarial strategy for α-loss and the corresponding equivalent expression of the minimax optimization in (5.7).

Proposition 1. *Under α-loss, the optimal adversary strategy that minimizes the expected loss is a 'α-tilted' conditional distribution expressed as*

$$P_h^*(s|g(x)) = \frac{P(s|g(x))^\alpha}{\sum_{s' \in \mathcal{S}} P(s'|g(x))^\alpha} \tag{5.9}$$

for each pair (s, x). Specifically, for $\alpha = 1$ and $\alpha = \infty$, the optimal adversarial strategies reduce to those for log-loss and soft 0-1 *loss, i.e., the true conditional distribution and the MAP estimation (shown in Table 5.1), respectively. In addition, the optimization in* (5.7) *reduces to*

$$\begin{aligned} \min_{g(\cdot)} \quad & -H_\alpha^A(S|g(X)) \\ s.t. \quad & \mathbb{E}[d(g(X), X)] \leq D, \end{aligned} \tag{5.10}$$

where $H_\alpha^A(\cdot|\cdot)$ is the Arimoto conditional entropy. Therefore, for given $P_{S,X}$, the constrained minimax game in (5.7) *is equivalent to minimizing the α-leakage from S to $g(X)$ subject to the same distortion constraint.*

From the result in Proposition 1, we know that the objective of the constrained minimax game in (5.7): (i) simplifies to $\min_{g(\cdot)} \log \max_{s \in S} P(s, g(X))$, if the soft 0-1 loss is used as the adversary's loss function, where the maximization inside the logarithm is the probability of correctly guessing of S given $g(X)$, denoted as $\mathrm{P_c}(S|g(X))$,

defined in [111]; (ii) equals to $\min_{g(\cdot)} I(g(X), S)$ for given $P_{S,X}$, if the log-loss is used. We observe that both $\mathrm{P_c}(S|g(X))$ and $I(g(X), S)$ can be used as a proxy of DemP since their minimal values is obtained if and only if S and $g(X)$ are independent. This observation can be extended to the series of α-loss for $\alpha \in [1, \infty]$ and leads to the following theorem.

Theorem 18. *Under α-loss (including log-loss and* 0-1 *loss), the CFUR framework enforces fairness subject to the distortion constraint. As the distortion increases, the ensuing fairness guarantee approaches ideal DemP.*

Many notions of fairness rely on computing probabilities to ensure independence of sensitive and target variables that are not easy to optimize in a data-driven fashion. In Theorem 18, we propose α-loss (including log-loss modeled in practice via cross-entropy) in the CFUR framework as a proxy for enforcing DemP fairness. In the following, we show that the CFUR framework based on α-loss can also provide EO and EoO fairness.

One can also design fair classifiers directly without intermediate representations; furthermore, such classifiers can be designed with either DemP, EO or EoO guarantees. Let $\hat{Y} = \tilde{g}(S, X)$ be a predictor/classifier for the targeted variable Y. Note that the $\tilde{g}(\cdot)$ generally depends on both X and S but the dependence on S can be implicit for scenarios where the sensitive information S is not directly available. Let h be a strategy used by the adversary to infer the sensitive variable S as $\hat{S} = h(\tilde{g}(S, X), Y)$ from the soft information of the predictor $\tilde{g}(S, X) = P_{\hat{Y}|X,S}$ and the true targeted variable Y. Analogous to (5.7), the design of a fair predictor/classifier can be formulated as

$$\min_{\tilde{g}(\cdot)} \max_{h(\cdot)} \quad -\mathbb{E}\big[\ell\big(h(\tilde{g}(S, X), Y), S\big)\big], \tag{5.11a}$$

$$\text{s.t.} \quad \mathbb{E}[\ell\big(\tilde{g}(S, X), Y\big)] \leq L. \tag{5.11b}$$

Proposition 2. *Under α-loss (incorporating both log-loss and 0-1 loss), the CFUR formulation in* (5.11) *enforces fairness subject to the expected loss constraint. As the loss increases, the ensuing fairness guarantee approaches ideal equalized odd of $\tilde{g}$ respect to the sensitive variable S and the targeted variable Y.*

Note that the formulation in (5.11) can also be used to generate a fair predictor or classifier in term of DemP or EoO. For DemP, the adversary will only have $\tilde{g}(S, X)$ as the input, and for EoO, the adversary requires access to $\tilde{g}(S, X)$ and only the $Y = 1$ class.

5.2.2 Data-Driven CFUR

Thus far, we have focused on a setting where the data holder has access to $P(X, S)$. When $P(X, S)$ is known, the data holder can simply solve the constrained minimax optimization problem in (5.7) (game-theoretic version of the CFUR formulation) to obtain a decorrelation scheme that would perform best against a chosen type of adversary. In the absence of $P(X, S)$, we propose a data-driven version of the CFUR formulation that allows the data holder to learn decorrelation schemes directly from a data set $\mathcal{D} = \{(x_i, s_i)\}_{i=1}^n$.

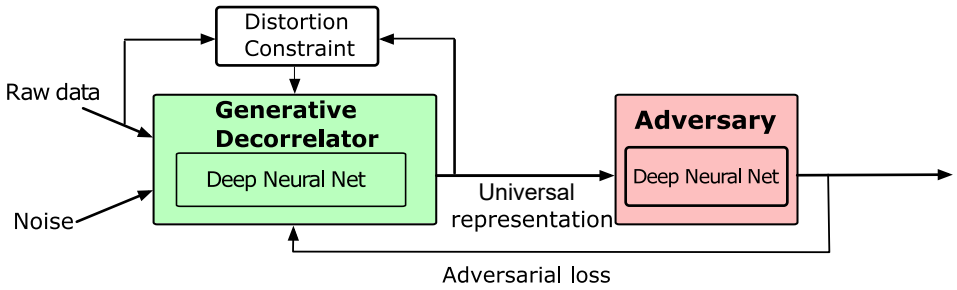


Figure 5.1: Generative Adversarial Model for Censoring and Fairness Guarantees

Under the data-driven version of the CFUR formulation, we represent the decorrelation scheme via a generative model $g(X; \theta_g)$ parameterized by θ_g. This generative model takes X as input and outputs X_r. In the training phase, the data holder learns

the optimal parameters θ_g by competing against a *computational adversary*: a classifier modeled by a neural network $h(g(X;\theta_g);\theta_h)$ parameterized by θ_h. In theory, the functions g and h can be arbitrary. However, in practice, we need to restrict them to a rich hypothesis class. Fig. 5.1 shows an example of the CFUR model in which the generative decorrelator and adversary are modeled as deep neural networks. For a fixed g and h, if S is binary and the and the log-loss is used, we can quantify the adversary's *empirical loss* using cross entropy

$$L_n(\theta_g,\theta_h) = -\frac{1}{n}\sum_{i=1}^{n} s_i \log h(g(x_i;\theta_g);\theta_h) + (1-s_i)\log(1-h(g(x_i;\theta_g);\theta_h)). \quad (5.12)$$

Note that one can generalize cross entropy to the multi-class case by using the softmax function. Therefore, the optimal model parameters are the solutions to

$$\min_{\theta_g}\max_{\theta_h} \quad -L_n(\theta_g,\theta_h), \quad (5.13a)$$

$$\text{s.t.} \quad \frac{1}{n}\sum_{i=1}^{n} d(g(x_i;\theta_g),x_i) \le D, \quad (5.13b)$$

where the expression $\frac{1}{n}\sum_{i=1}^{n} d(g(x_i;\theta_g),x_i)$ is the *empirical distortion*.

The minimax optimization in (5.13) is a two-player non-cooperative game between the generative decorrelator and the adversary with strategies θ_g and θ_h, respectively. In practice, we can learn the equilibrium of the game using an iterative algorithm (see Algorithm 1). We first minimize the adversary's empirical loss via the gradient descent algorithm in the inner loop to update θ_h for a fixed θ_g. Then, we minimize the decorrelator's empirical loss, which is modeled as a sum of the negative of the adversary's empirical loss and a *penalty*, to update θ_g for a fixed θ_h. The penalty is designed to incorporate the distortion constraint in (5.13) by using the *penalty method* [112], and specifically, is expressed as $\rho(\max\{0, \frac{1}{n}\sum_{i=1}^{n} d(g(x_i;\theta_g),x_i) - D\})^2$, the product of a penalty parameter $\rho > 0$ and a measure of violation of the constraint. Note that the measure of violation (i.e., the squared maximum) is non-zero when

Algorithm 1 Iterative algorithm

Input: Training data set $\mathcal{D} = \{(x_i, s_i)\}_{i=1}^n$, distortion parameter D, iteration numbers T and K (for outer and inner loops, respectively), sample size $m \leq n$, learning rates λ_g and λ_h

procedure ITERATIVE ALGORITHM($\mathcal{D}, D, T, K, m, \lambda_g, \lambda_h$)

Initialize θ_g^0, θ_h^0 and ρ_0

for $t = 0, \cdots, T-1$ **do** ▷ Outer loop

Randomly draw a sample $\{x_i\}_{i=1}^m$ from $\mathcal{D}$

Generate $\{x_{r,i}\}_{i=1}^m$ via $x_{r,i} = g(x_i; \theta_g^t)$ ▷ Generate representations

Calculate $D_m(\theta_p^t) = \frac{1}{m}\sum_{i=1}^m d(x_{r,i}, x_i)$ ▷ Update the empirical distortion

$\theta_h^{t'} = \theta_h^t$

for $k = 1, \cdots, K$ **do** ▷ Inner loop: update θ_h

$\theta_h^{t,k} = \theta_h^{t'} - \lambda_h \nabla_{\theta_h^{t'}} \frac{1}{m} \sum_{i=1}^m \ell(h(x_{r,i}; \theta_h^{t'}), s_i)$

$\theta_h^{t'} = \theta_h^{t,k}$

end for

$\theta_h^{t+1} = \theta_h^{t'}$

Calculate $L_m(\theta_p^t, \theta_h^{t+1}) = \frac{1}{m}\sum_{i=1}^m \ell(h(x_{r,i}; \theta_h^{t+1}), s_i)$ ▷ Update the empirical loss

if $D_m(\theta_p^t) \leq D$ **then** ▷ Update θ_g

Calculate $\theta_g^{t+1} = \theta_g^t + \lambda_g \nabla_{\theta_g^t} L_m(\theta_p^t, \theta_h^{t+1})$

else

Adapt ρ_t based on $(D_m(\theta_p^t) - D)^2$ and $-L_m(\theta_p^t, \theta_h^{t+1})$

Calculate $\theta_g^{t+1} = \theta_g^t - \lambda_g \nabla_{\theta_g^t} \Big(- L_m(\theta_p^t, \theta_h^{t+1}) + \rho_t \big(D_m(\theta_p^t) - D\big)^2 \Big)$

end if

Exit if solution converged

end for

return θ_g^{t+1}

end procedure

the constraint is violated and is zero if the constraint is satisfied. Therefore, for any fixed θ_h, the constrained optimization problem of the generative decorrelator is approximated by the following unconstrained optimization problem

$$\min_{\theta_g} \quad -L_n(\theta_g, \theta_h) + +\rho\bigg(\max\Big\{0, \frac{1}{n}\sum_{i=1}^{n} d(g(x_i;\theta_g), x_i) - D\Big\}\bigg)^2, \tag{5.14}$$

where the penalty parameter ρ is properly adapted to make any non-zero penalty term competing with the negative of the adversary's empirical loss, i.e., $-L_n(\theta_g, \theta_h)$. For convex optimization problems, the solution to a series of unconstrained problems will eventually converge to the solution of the original constrained problem [112].

The performance of the learned decorrelation scheme is tested under well-trained [1] adversarial predictors for S as well as classifiers or regressors for Y. Note that the knowledge of downstream applications are only required for evaluating utilities preserved by X_r. We follow this procedure in the next section.

5.3 CFUR for Publicly Available Data Sets

We apply our CFUR framework to publicly accessible data sets including UCI Adult and UTKFace. The UCI Adult data set [2] is for salary prediction and consists of 10 categorical features and 4 continuous features. As shown in Table 5.2, we choose gender or the tuple (gender, relationship) as sensitive variable S and other features except salary as non-sensitive variable X. The UTKFace data set [3] consists of more than 20 thousand 200×200 colorful face images labeled by age, ethnicity and gender. Individuals in the data set have ages from 0 to 116 years old and are divided into 5 ethnicities: White, Black, Asian, Indian, and others including Hispanic, Latino and Middle Eastern. We take gender as the sensitive variable S and the image pixels as

[1] In the testing phrase, all neural networks are trained on the generated representation X_r.

[2] https://archive.ics.uci.edu/ml/datasets/adult

[3] http://aicip.eecs.utk.edu/wiki/UTKFace

the non-sensitive variable X.

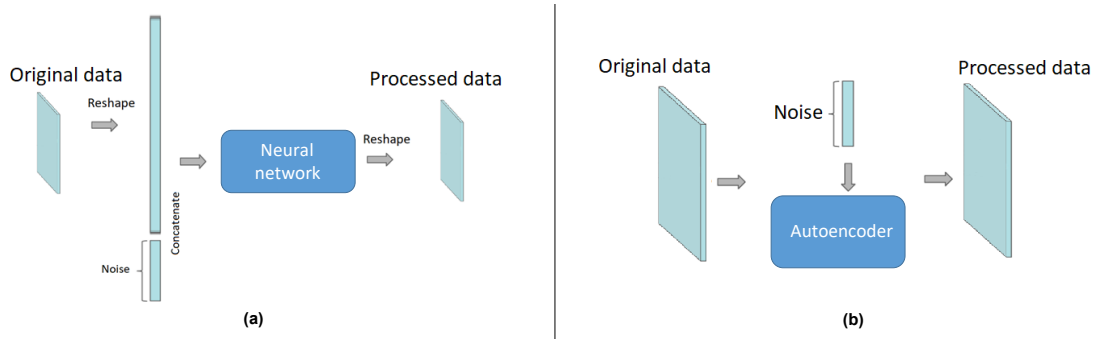


Figure 5.2: Different Architectures of the Decorrelator in the CFUR Framework

We apply two different architectures (as shown in Fig. 5.2) to the UCI Adult and UTKFace data sets, respectively. For the UCI Adult data set, we use the feedforward neural network decorrelator (FNND) (shown in Fig. 5.2 (a)), in which a feedforward multi-layer neural network is used to generate the processed data/representation X_r from the original data (i.e., X or (S, X)) and random noise. For the UTKFace data set, we use the noisy auto-encoder decorrelator (NAED) (shown in Fig. 5.2 (b)), in which the encoder of an auto-encoder generates a lower-dimensional feature vector of the original data and adds independent random noise to each element of the feature vector and the decoder of the auto-encoder reconverts the noisy feature vector to generate the representation X_r.

We use the classification accuracy of a specified sensitive feature S to measure the performance of censoring S. As a measure of fairness, DemP requires that the conditional probabilities of an arbitrary prediction $\hat{Y} = y$ for $y \in \mathcal{Y}$ given any $s \in \mathcal{S}$ equal. Therefore, for fairness, we take the following maximal difference as the measure of DemP, i.e.,

$$\Delta_{\mathbf{DemP}}(y) = \max_{s,s' \in \mathcal{S}} |P(\hat{Y} = y | S = s) - P(\hat{Y} = y | S = s')| \tag{5.15}$$

and a smaller value of $\Delta_{\mathbf{DemP}}$ implies a closer approaching of the DemP. We illustrate

Case I	Feature	Description	Case II
Y	salary	2-salary intervals: $> 50K$ and $\leq 50K$	Y
S	gender	2 classes: male and female	S
X	relationship	6 classes of family relationships	
	age	9-age intervals: $18-25$, $25-30$, $\cdots$, $60-65$	X
	workclass	8 types of employer	
	education	16 levels of the highest achieved education	
	marital-status	7 classes of marital status	
	occupation	14 types of occupation	
	race	5 classes	
	native-country	41 countries of origin	
	capital-gain	Recorded capital gain; (continuous)	
	capital-loss	Recorded capital loss; (continuous)	
	hours-per-week	Worked hours per week; (continuous)	
	education-num	Numerical version of education; (continuous)	

Table 5.2: Features in UCI Adult Data Set

the results for each data set mentioned above in the following subsections. Details of all experiments on these data sets are in Appendix 5.5.

5.3.1 Illustration of Results for UCI Adult Data Set

For the UCI Adult data set with both categorical and continuous features as shown in Table 5.2, we consider two cases: (i) in Case I, we take 'gender' as the sensitive variable S and S is either male or female in this data set, and (ii) in Case II, we take both 'gender' and 'relationship' as the sensitive variable S. In this data set, 'relationship' has 6 distinct values, and therefore, S has 12 possibilities. In both cases, we take 'salary' as the target variable Y, which is either $> 50K$ or $\leq 50K$,

and use the classification accuracy of salary as the utility measure. Note that for any binary target variable, the two values of the fair measure $\Delta_{\mathbf{DemP}}(y)$ in (5.15) are the same, and therefore, we use $\Delta_{\mathbf{DemP}}$ to denote the measure of DemP for experiments on this data set.

Case I: Binary Sensitive Feature.

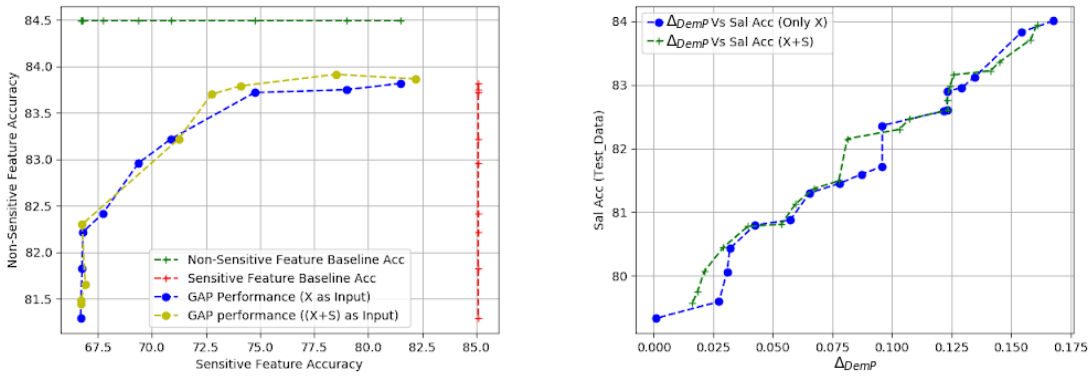
Fig. 5.3 illustrates the performances of the generated representation X_r for censoring and fairness concerns. Specifically, for censoring, the performance is evaluated via the tradeoff between the classification accuracy of salary and gender (as shown in Fig. 5.3a where we use the classification accuracy for the original testing data set (only using the non-sensitive feature X) as the baseline performance and the green and red lines denote the baseline performances for the target variable (salary) and the sensitive variable (gender), respectively); and for fairness, it is evaluated via the tradeoff between the classification accuracy of salary and the DemP measure Δ_{DemP} (as shown in Fig. 5.3b and the value of Δ_{DemP} for the original testing data is 0.2). We consider two possible inputs of the generative decorrelator $g(\cdot)$ in (5.7), i.e., only the non-sensitive features X or both X and S.

From Fig. 5.3a, we observe that: (i) while the classification accuracy of gender (the sensitive variable) is about 66% and only decreases about 20% from the baseline performance [4] , the classification accuracy of salary (the target variable) is above 82% and only decreases 2.5% from its baseline performance. Note that 66% is probability of male in the original testing data and indicates a random guess of gender. Therefore, the generated representation X_r hides the information of gender pretty well while maintaining the information of salary; (ii) only in the high utility region, where the accuracy of salary is no less than 72.5%, to take both S and X in the generative

[4]Note that the baseline performances are the accuracy or DemP measure obtained from the original testing data

decorrelator $g(\cdot)$ has a small advantage over only using X. Specifically, given the same classification accuracy of gender, the classification accuracy of salary is at most 0.3% higher.

From Fig. 5.3b, we make the following two observations: (i) the classification accuracy of salary and the Δ_{DemP} have an approximately affine relationship, and when Δ_{DemP} is almost 0, the accuracy of salary is above 79%. Therefore, our framework is effective in approaching perfect DemP; (ii) the representation X_r generated from either X or (S, X) leads to a similar fairness performance. Comparing our results in Fig. 5.3b and the results in [61, 67], for $\Delta_{\text{DemP}} = 0.06$, Edwards' and Madras's approaches have 2% and 2.5% higher classification accuracy of salary than ours, respectively. However, both their approaches are not shown to achieve the extreme point, $\Delta_{\text{DemP}} = 0$, but our approach does with the classification accuracy above 79%.



(a) Salary vs. Gender Classification Accuracy

(b) Salary Classification Accuracy vs. Δ_{DemP}

Figure 5.3: Performances for Case I of UCI Adult Data Set

We can also evaluate the fairness performance of the generated representation by using EO defined in Definition 5.1.1. EO requires that for both possible outcomes, i.e., $Y =$'$> 50K$' and $Y =$'$\leq 50K$', the conditional probabilities of the correct prediction

given the two possible gender (i.e., male or female) equal, i.e.,

$$P(\hat{Y} = Y|\text{female}, Y =' > 50K') = P(\hat{Y} = Y|\text{male}, Y =' > 50K')$$

$$P(\hat{Y} = Y|\text{female}, Y =' \leq 50K') = P(\hat{Y} = Y|\text{male}, Y =' \leq 50K').$$

Therefore, we use the differences of the above two pairs of conditional probabilities to characterize the achieved EO fairness:

$$\Delta_{\text{EO}}(> 50K) \triangleq \left| P(\hat{Y} = Y|\text{female}, Y =' > 50K') - P(\hat{Y} = Y|\text{male}, Y =' > 50K') \right|$$

$$\Delta_{\text{EO}}(\leq 50K) \triangleq \left| P(\hat{Y} = Y|\text{female}, Y =' \leq 50K') - P(\hat{Y} = Y|\text{male}, Y =' \leq 50K') \right|$$

Fig. 5.4 shows the achieved EO fairness where $\Delta_{\text{EO}}(> 50K)$ and $\Delta_{\text{EO}}(\leq 50K)$ are the two measures of EO for the two possible outcomes $Y =$'$> 50K$' and $Y =$'$\leq 50K$', respectively, and the red curve is for the sum of $\Delta_{\text{EO}}(> 50K)$ and $\Delta_{\text{EO}}(\leq 50K)$, which is exactly the Δ_{EO} in Figure 2(b) of [67]. From the results in Fig. 5.4, we observe that while the classification accuracy of salary is above 82.4%, the values of $\Delta_{\text{EO}}(> 50K)$ and $\Delta_{\text{EO}}(\leq 50K)$ decrease 99.2% and 63% (from the baseline performances) to 0.0007 and 0.0254, respectively, which means the X_r generated under the rule of DemP by our CFUR framework still provides good EO fairness. To further demonstrate the efficiency of our CFUR framework, we compare our CFUR with the methods proposed in [67]. Note that in [67], the EO fairness is quantified via the Δ_{EO} in Figure 2(b) of [67], which is exactly the EO sum $\Delta_{\text{EO}}(> 50K) + \Delta_{\text{EO}}(\leq 50K)$. Comparing the performances shown in Fig. 5.4 and Figure 2(b) of [67], we observe that our CFUR X_r is competitive with the LAFTR-DP method in [67], which uses DemP as the fairness metric for training fair classifiers. Specifically, while our classification accuracy of salary is 1.3% smaller than LAFTR-DP given the EO sum $\Delta_{\text{EO}} = 0.04$, our minimal achieved EO sum decreases 28% from LAFTR-DP and is the same as LAFTR-EO, which uses EO as the fairness metric to train models. In addition, we observe that the decrease of the EO sum is even larger than the DemP measure Δ_{DemP}, which

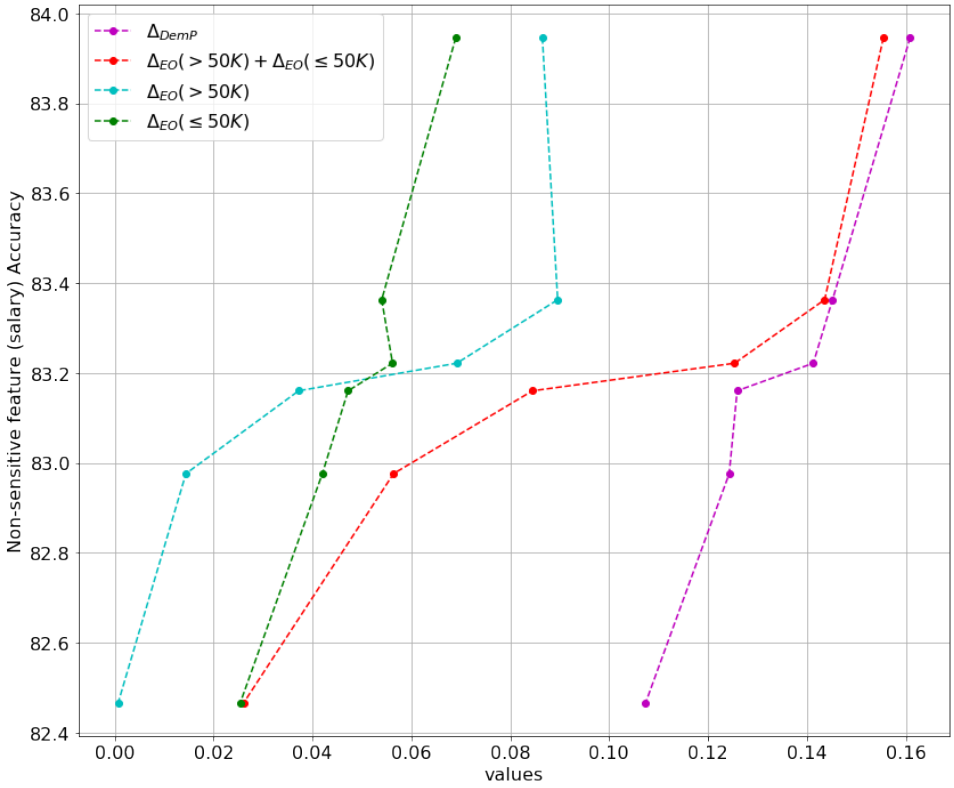


Figure 5.4: The EO Fairness Achieved in Case I of UCI Adult Data Set

shows that to generate representations under the rule of DemP can provide fairness in terms of various metrics.

Case II: Non-binary Sensitive Feature. Figs. 5.5 and 5.6 illustrate the censoring and fairness performance of the generated representation X_r in hiding 'gender' and 'relationship' jointly and separately while preserves information of 'salary'.

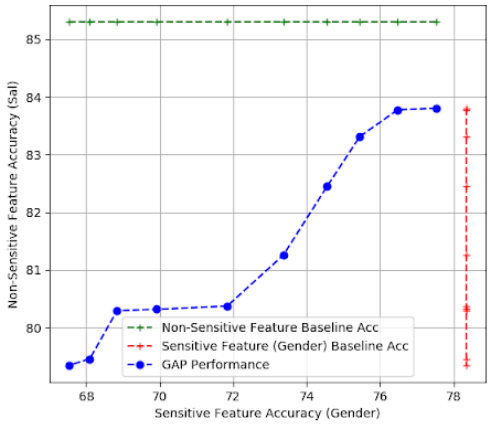
Fig. 5.5 shows the tradeoffs of the classification accuracy of salary versus gender, relationship, or (gender, relationship). Note that in Fig. 5.5, we use the classification accuracy obtained from the original testing data set as the baseline performance, which is denoted by the green and red lines for the target variable (salary) and the sensitive variable (gender or/and relationship), respectively. From Fig. 5.5, we observe that while the classification accuracy of salary is above 79%, the classification accuracy of gender and/or relationship are about 66% (as shown in Fig. 5.5a), 45%

(as shown in Fig. 5.5b) and 41% (as shown in Fig. 5.5c), respectively. Note that the probabilities of male, husband, and the combination (male, husband) is 66%, 40% and 40%, respectively, in the original testing data. Therefore, while the classification accuracy of salary is preserved as 79%, the inferences of gender, relationship, and combination (gender, relationship) are randomly guessing with known prior. Thus, the X_r is pretty well in hiding multiple sensitive information both separately and jointly. On the other hand, to have the flexibility of hiding the other sensitive information – relationship, the cost is a reduce in utility. Specifically, to compare the results in Figs. 5.3a and 5.5a, we can see that the drop of the classification accuracy of salary is at most 3% for any given accuracy of gender [5] .

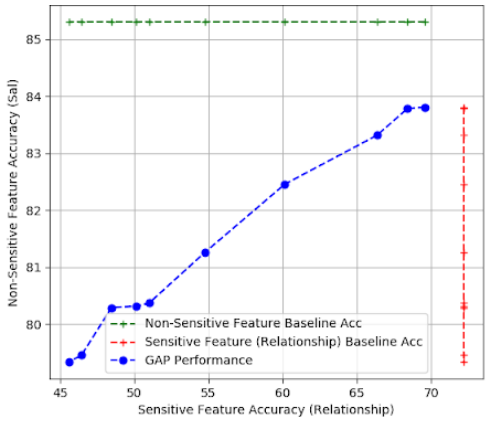
Fig. 5.6 shows the tradeoffs between the classification accuracy of salary versus the DemP measure Δ_{DemP} (defined in (5.15)) *w.r.t.* gender, relationship or their combination [6] , respectively. Note that in Fig. 5.6b, the red and blue lines are the DemP measure Δ_{DemP} for relationship and (gender, relationship), respectively, and for the original testing data, the value of Δ_{DemP} for gender, relationship and (gender, relationship) is 0.2, 0.438 and 0.443, respectively. In Fig. 5.6, we observe that while the classification accuracy of salary is above 94% of the baseline performance the value of Δ_{DemP} is dropped to 25% for gender and to about 34% for both relationship and the combination. Therefore, the X_r works well in decorrelating the information of salary and the information of gender and relationship jointly and separately. From 5.6b, we observe that the value of Δ_{DemP} for the combination is almost the same as that for relationship. In addition, comparing the results in Figs. 5.3b and 5.6b, we can see that for any given classification accuracy of salary, the Δ_{DemP} for gender in

[5]Note that in Figs., 5.3a and 5.5a, the baseline performances are different and it is because in Case II, the feature variable X does not contain 'relationship'.

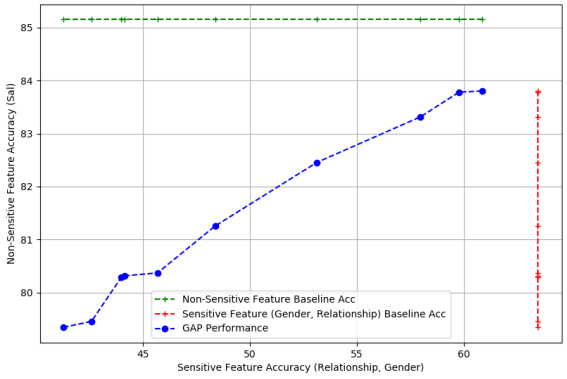
[6]Note that there are no samples for (Female, Husband) and only 2 samples for (Male, Wife). Therefore, in the calculation of Δ_{DemP}, we ignore the two groups.



(a) Gender



(b) Relationship



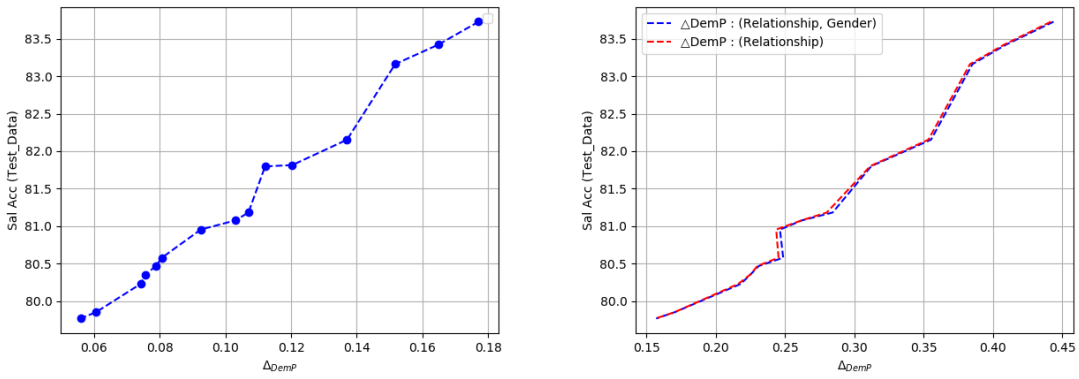
(c) (Gender, Relationship)

Figure 5.5: Tradeoffs between Classification Accuracy of Non-Sensitive Feature (Salary) and Sensitive Features (Gender and/or Relationship) in Case II of UCI Adult Data Set

Case II can be about 25% higher of than that in Case I, which is the cost of providing fairness for relationship.

5.3.2 Illustration of Results for UTKFace Data Set

In the UTKFace data set, the face images are the non-sensitive variable X. We take 'gender' as the sensitive variable S and either 'ethnicity' or 'age' as the target variable Y for ethnicity classification and age regression, respectively. For the two



(a) Gender (b) Relationship (or) (Relationship, Gender)

Figure 5.6: Tradeoffs between Salary Classification Accuracy and the Δ_{DemP} of Gender and/or Relationship in Case II of UCI Adult Data Set

downstream applications, the corresponding supports of Y are $\mathcal{Y} = \{\text{White, Black, Asian, Indian}\}$ and $\mathcal{Y} = \{i \in \mathbb{Z} : 10 \leq i \leq 65\} = [10, 65]$, respectively. We use the maximum of the DemP measure (defined in (5.15)) over the support $\mathcal{Y}$, i.e., the value $\Delta_{\text{DemP}} = \max_{y \in \mathcal{Y}} \Delta_{\text{DemP}}(y)$, to indicate the achieved fairness.

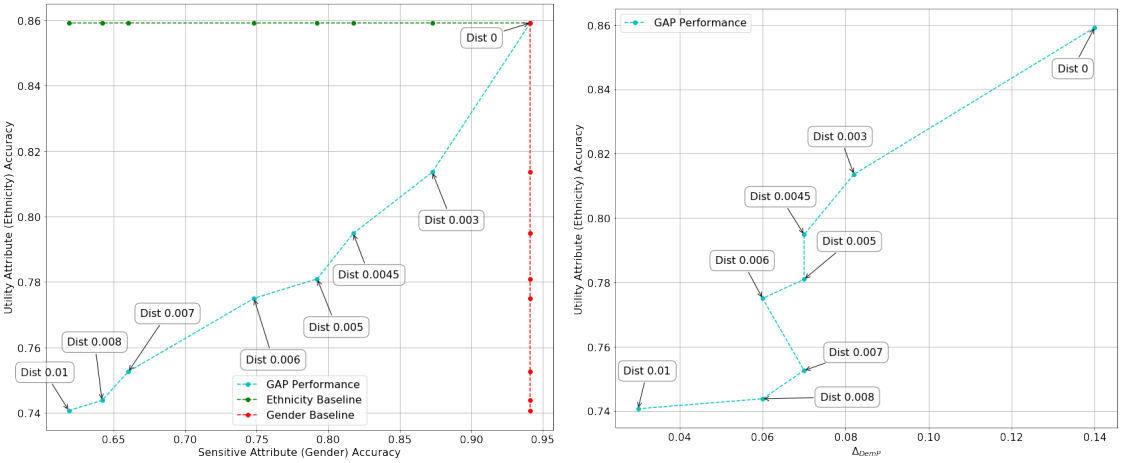
Figure 5.7 illustrates the output representations X_r for 16 typical [7] faces in the UTKFace data set for increasing per-pixel distortion. From Fig. 5.7, we can observe (i) for a small per-pixel distortion (e.g., 0.003), the distinguished features of gender like lip color are smoothed out; and (ii) as a higher per-pixel distortion is allowed (e.g., 0.006), our framework generates a face with an opposite gender; (iii) however, when the average per-pixel distortion is too large (e.g., 0.01), the representation generated by the CFUR framework tends to be too blurred to show the face contour and clear five sense organs. Note that the set of vertical faces highlighted in boxes make explicitly how the sensitive feature (gender) is changed with increasing distortion.

Figs. 5.8a and 5.9 show the tradeoffs, achieved by the generated representation X_r,

[7]The 16 typical faces covers the 8 possible combination of 2 gender (male and female) and 4 ethnicity (White, Black, Asian and Indian) and includes young, adult and old faces.

Figure 5.7: The Generated CFUR of Face Images for Different Values of the Average Per-Pixel Distortions of UTKFace Data Set

of the classification accuracy of gender and the utility for the ethnicity classification and age regression, respectively. Note that 'dist' indicate the per pixel distortion on images. In Fig. 5.8a, while the classification accuracy of gender is about 62% and decreases about 35% from the baseline performance, the classification accuracy of ethnicity is above 74% and only decreases 14% from its baseline performance. Note that in the original testing data, the highest marginal probability for gender and ethnicity are 54.6% (the probability of male) and 43.2% (the probability of White), respectively. Therefore, the 62% classification accuracy of gender is only better than a random guess by 7.4% while the 74% classification accuracy of ethnicity is better than a random guess by 30.8%. Therefore, the X_r hides the information about gender well while maintaining the information about salary. For age regression, we use the mean absolute error (MAE), i.e., the average absolute difference between the predicted



(a) Ethnicity vs. Gender Classification Accuracy

(b) Ethnicity Classification Accuracy vs. Δ_{DemP}

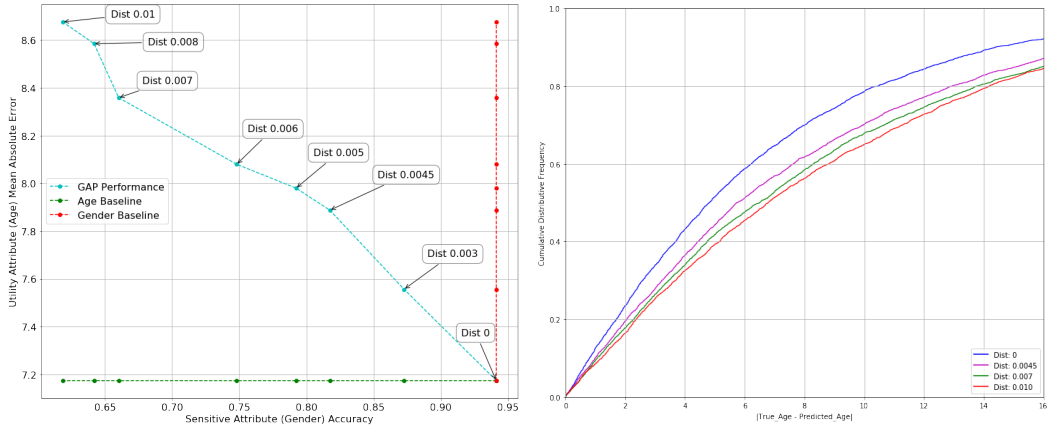
Figure 5.8: The Achieved Tradeoffs between Classification Accuracy of Ethnicity and Gender as well as and Δ_{DemP} for UTKFace Data Set

age and the true age, as the utility measure. In Fig. 5.9a, we observe that while the classification accuracy for gender is about 62%, which is a 35% decrease from the baseline performance 94%, the increase in the MAE is 1.5 which is about a 20% increase from the baseline performance 7.2. Fig. 5.9a shows the cumulative distribution function (CDF) of the difference between the true and predicted age for various distortions, from which we can see that the drop of the cumulative probability is at most 1%. Thus, it is shown that the CFUR framework also does a good job in maintaining the information about age, and therefore, constraining the distortion of generated representations is an efficient way for guaranteeing the utility of various applications.

Figs. 5.8b and 5.10 show the tradeoff between the utility measure and DemP measure Δ_{DemP} of the generated representation X_r in ethnicity classification and age regression, respectively. In Fig. 5.8b where the x-axis is the maximal value of DemP measure in (5.15) over the four ethnicity, we observe that while achieving about 86%

Distortion	0	0.003	0.0045	0.005	0.006	0.007	0.008	0.01
Δ_{DemP}(White)	0.061	0.055	0.04	0.03	0.03	0.02	0.02	0.01
Δ_{DemP}(Black)	0.109	0.021	0.02	0.05	0.03	0.05	0.03	0.03
Δ_{DemP}(Asian)	0.14	0.082	0.07	0.07	0.06	0.07	0.06	0.03
Δ_{DemP}(Indian)	0.031	0.006	0.01	0	0.01	0	0.01	0.01

Table 5.3: DemP Fairness (Indicated by $\Delta_{\text{DemP}}(\cdot)$) of Ethnicity Classification on UTKFace Data Set

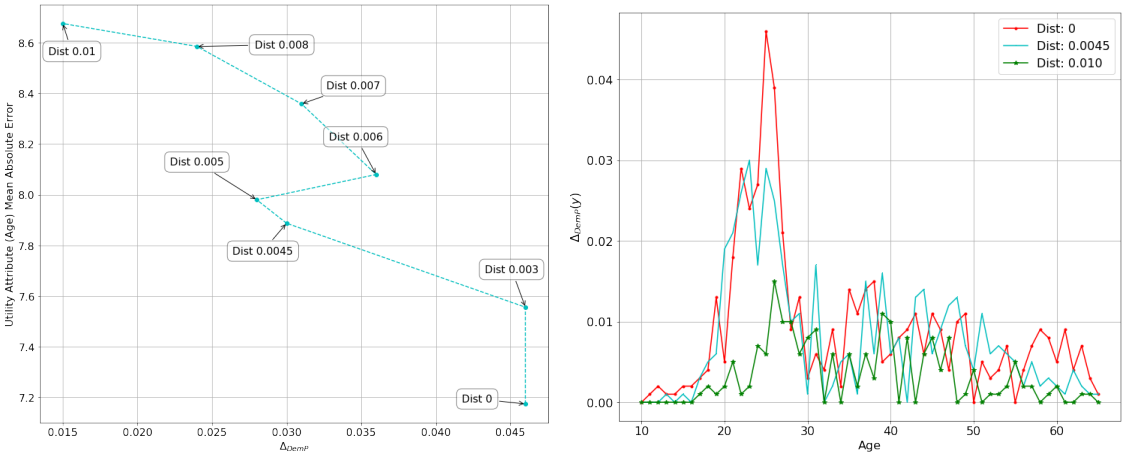


(a) Mean Absolute Error of Age Prediction vs. Gender Classification Accuracy

(b) The CDF of the Difference between the True and Predicted Age

Figure 5.9: Performances of Age Regression on UTKFace Data Set

of the baseline classification accuracy, the Δ_{DemP} is reduced to 0.03 which is only 20% of the $\Delta_{\text{DemP}} = 0.14$ in the original testing data. Therefore, the X_r is good at approaching DemP fairness while maintaining the utility for ethnicity classification. Table 5.3 shows the decrease of the DemP measure for each the four ethnicity as the distortion increases. In Fig. 5.10a where the x-axis is the maximal value of the DemP measure in (5.15) over the chosen age range (10-65), while preserving 86% of the utility baseline performance, the Δ_{DemP}, i.e., the maximal value of DemP measure over the 56 age values, decreases to 0.015 which is less than 33% of the $\Delta_{\text{DemP}} = 0.046$ on the original testing data. Fig. 5.10b shows the demographic



(a) Mean Absolute Error of Age Prediction vs. Δ_{DemP}

(b) Values of the DemP Measure for Various Distortions

Figure 5.10: Achieved DemP Fairness for Age Regression on UTKFace Data Set

measure $\Delta_{\text{DemP}}(y)$, $y \in [10, 65]$, for various distortions, from which we observe that when the pixel distortion is 0.01, even though the $\Delta_{\text{DemP}} = 0.015$, for 17 different age values, $\Delta_{\text{DemP}}(y) = 0$. That is, the predictions of these 17 different ages are completely independent of gender (the sensitive information) and DemP is achieved for those predictions.

5.4 Proof Details

5.4.1 Proof of Theorem 16

Given a pair of sensitive and non-sensitive variables (S, X), let X_r be the representation of X generated by a random mapping g, i.e., $X_r = g(X)$. In a learning task of inferring an arbitrary target variable Y related to (S, X), an algorithm estimates Y as $\hat{Y}$ from X_r. Therefore, these random variables form the Markov chain $(S, X) - X_r - \hat{Y}$. From Definition 5.1.3, if the representation X_r satisfies DemP *w.r.t* S, X_r is independent of S and $I(S; X_r) = 0$. From the data processing inequality and

non-negativity of MI, we have that,

$$0 \leq I(S; \hat{Y}) \leq I(S; X_r) = 0. \tag{5.16}$$

Therefore, S is independent of $\hat{Y}$, and $\hat{Y}$ satisfies DemP w.r.t. S. Note that S can be a collection of sensitive features. From $I(S; X_r) = 0$, there is $I(X_r; S_t) = 0$ for any subset of features $S_t \subset S$ and therefore, we have that $\hat{Y}$ satisfies DemP w.r.t. any subset of features in S. □

5.4.2 Proof of Theorem 17

In (5.7), the output of the optimal generative decorrelator $g^*(X)$ is the universal representation of the original features X, i.e.,

$$X_r = g^*(X) = \operatorname*{argmin}_{g(\cdot)} \max_{h(\cdot)} -\mathbb{E}[\ell(h(g(X)); S)].$$

Let h^* be the corresponding optimal adversarial strategy, i.e.,

$$h^*_{g^*} = \operatorname*{argmin}_{h} \mathbb{E}[\ell(h(X_r), S)].$$

For sufficient large distortion bound D, X_r can be arbitrarily distorted from X and $g(\cdot)$ can be any mapping from X to X_r. Therefore, we can get ride of the distortion constraint in (5.7) and have

$$-\mathbb{E}[\ell(h^*_{g^*}(X_r); S)] = -\max_{g(\cdot)} \min_{h(\cdot)} \mathbb{E}[\ell(h(g(X)); S)] \tag{5.17}$$

$$= -\max_{g(\cdot)} \mathbb{E}[\ell(h^*_g(g(X)); S)], \tag{5.18}$$

$$\leq -\mathbb{E}[\ell(h^*_g(g(X)); S)], \tag{5.19}$$

where $g(X)$ is any (randomized) mapping of X. That is, the generated representation X_r satisfies the inequality in (5.4). Thus, for sufficiently large distortion bound D, the generated representation X_r generated from the formulation in (5.7) is censored *w.r.t.* the sensitive variable S against a learning adversary $h(\cdot)$ captured by a loss function $\ell(h(X_r), S)$. □

5.4.3 Proof of Proposition 1

Consider the α-loss function [110]

$$\ell(h(g(X)), s) = \frac{\alpha}{\alpha - 1}\left(1 - P_h(s|g(X))^{1-\frac{1}{\alpha}}\right), \tag{5.20}$$

for any $\alpha > 1$. Denote $H_\alpha^{\mathrm{A}}(S|g(X))$ as the Arimoto conditional entropy of order α. Due to that α-loss is convex in $P_h(s|g(X))$, by using Karush–Kuhn–Tucker (KKT) conditions, we show that

$$\max_{h(\cdot)} -\mathbb{E}\left[\frac{\alpha}{\alpha - 1}\left(1 - P_h(s|g(X))^{1-\frac{1}{\alpha}}\right)\right] = \frac{\alpha}{1-\alpha}\left(1 - \exp\left(\frac{1-\alpha}{\alpha} H_\alpha^{\mathrm{A}}(S|g(X))\right)\right)$$

which is achieved by a 'α-tilted' conditional distribution

$$P_h^*(s|g(X)) = \frac{P(s|g(X))^\alpha}{\sum\limits_{s\in\mathcal{S}} P(s|g(X))^\alpha}.$$

Under this choice of a decision rule, the objective of the minimax optimization in (5.7) reduces to

$$\min_{g(\cdot)} -H_\alpha^{\mathrm{A}}(S|g(X)), \tag{5.21}$$

which is simplified as $\min\limits_{g(\cdot)} \log \sum_{x\in\mathcal{X}} \max_{s'\in\mathcal{S}} P(g(x), s)$ for $\alpha = \infty$. Note that for given $P_{S,X}$, $H_\alpha^{\mathrm{A}}(S)$ is a constant, and therefore, (5.21) is equivalent to

$$\min_{g(\cdot)} -H_\alpha^{\mathrm{A}}(S|g(X)) + H_\alpha^{\mathrm{A}}(S) = \min_{g(\cdot)} I_\alpha^{\mathrm{A}}(g(X); S), \tag{5.22}$$

where the Arimoto MI $I_\alpha^{\mathrm{A}}(g(X); S)$ equals to the α-leakage form S to $g(X)$. □

5.4.4 Proof of Theorem 18

For any fixed classifier g, the optimal adversarial strategy in (5.7) is

$$h^\star(g(X)) = \arg\max_{h(\cdot)} -\mathbb{E}\big[\ell\big(h(g(X)), S\big)\big]. \tag{5.23}$$

When α-loss is used, from Proposition. 1 we have that for $\alpha \geq 1$, the optimal adversarial strategy is given by

$$h^\star(g(X), S) = \frac{P(s|g(X))^\alpha}{\sum\limits_{s' \in \mathcal{S}} P(s'|g(X))^\alpha}, \tag{5.24}$$

for any $s \in \mathcal{S}$, and the corresponding expected α-loss is given by

$$\mathbb{E}\big[\ell\big(h(g(X)), S\big)\big] = \begin{cases} \frac{\alpha}{\alpha-1}\left(1 - \exp\left(\frac{1-\alpha}{\alpha} H_\alpha^{\mathrm{A}}(S|g(X))\right)\right), & \alpha > 1 \\ H(S|g(X)), & \alpha = 1. \end{cases} \tag{5.25}$$

Therefore, the optimization in (5.7) can be simplifies to

$$\min_{g(\cdot)} \quad - H_\alpha^{\mathrm{A}}(S|g(X)), \tag{5.26a}$$

$$\text{s.t.} \quad \mathbb{E}[d\big(g(X), X\big)] \leq D \tag{5.26b}$$

where $H_\alpha^{\mathrm{A}}(S|g(X))$ is the Arimoto conditional entropy of order α. Note that as α tends to 1, it simplifies to Shannon entropy [90]. From the non-negativity of Arimoto mutual information, we know that

$$H_\alpha^{\mathrm{A}}(S|g(X)) \leq H_\alpha^{\mathrm{A}}(S) \tag{5.27}$$

with equality if and only if $g(X)$ is independent of S, which is exactly the requirement of DemP. Thus, as the distortion bound D in (5.26) increases, the CFUR formulation in (5.26) will approach ideal DemP by enforcing $H_\alpha^{\mathrm{A}}(S|g(X)) = H_\alpha^{\mathrm{A}}(S)$. □

5.4.5 Proof of Proposition 2

For any fixed classifier $\tilde{g}$, the optimal adversarial strategy in (5.11) is

$$h^\star(\tilde{g}(S, X), Y) = \arg\max_{h(\cdot)} \quad -\mathbb{E}\big[\ell\big(h(\tilde{g}(S, X), Y), S\big)\big]. \tag{5.28}$$

When α-loss is used, from Proposition. 1 we have that for $\alpha \geq 1$, the optimal adversarial strategy is given by

$$h^{\star}(\tilde{g}(S,X),Y) = \frac{P(s|\tilde{g}(S,X),Y)^{\alpha}}{\sum\limits_{s' \in \mathcal{S}} P(s'|\tilde{g}(S,X),Y)^{\alpha}}, \tag{5.29}$$

for any $s \in \mathcal{S}$, and the corresponding expected α-loss is given by

$$\mathbb{E}\big[\ell\big(h(\tilde{g}(S,X),Y),S\big)\big] = \begin{cases} \frac{\alpha}{\alpha-1}\left(1-\exp\left(\frac{1-\alpha}{\alpha}H_{\alpha}^{\mathrm{A}}(S|\tilde{g}(S,X),Y)\right)\right), & \alpha > 1 \\ H(S|\tilde{g}(S,X),Y), & \alpha = 1. \end{cases} \tag{5.30}$$

Therefore, the optimization in (5.11) can be simplifies to

$$\min_{\tilde{g}(\cdot)} \quad -H_{\alpha}^{\mathrm{A}}(S|\tilde{g}(S,X),Y), \tag{5.31a}$$

$$\text{s.t.} \quad \mathbb{E}[\ell\big(\tilde{g}(S,X),Y\big)] \leq L \tag{5.31b}$$

where $H_{\alpha}^{\mathrm{A}}(S|\tilde{g}(S,X),Y)$ is the Arimoto conditional entropy of order α. Note that as α tends to 1, it simplifies to Shannon entropy [90].

From the non-negativity of Arimoto mutual information, we know that

$$H_{\alpha}^{\mathrm{A}}(S|\tilde{g}(S,X),Y) \leq H_{\alpha}^{\mathrm{A}}(S|Y) \tag{5.32}$$

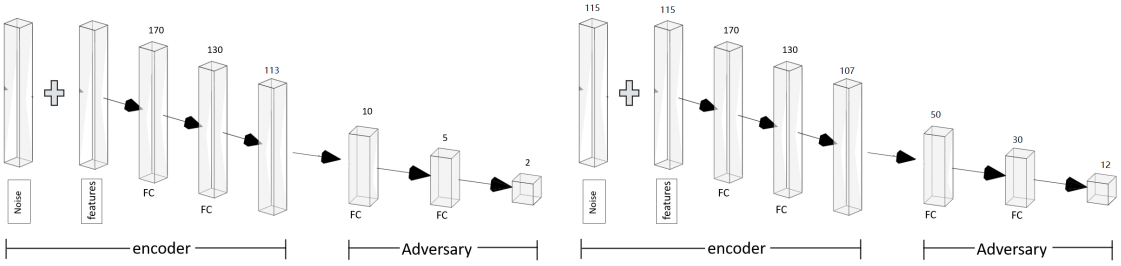
with equality if and only if $\tilde{g}(S,X)$ is independent of S conditioning on Y, which is exactly the requirement of EO. Thus, as the loss upper-bound L in (5.31) increases, the CFUR formulation in (5.31) will approach ideal EO of $\tilde{g}(S,X)$ respect to Y and S by enforcing $H_{\alpha}^{\mathrm{A}}(S|\tilde{g}(S,X),Y) = H_{\alpha}^{\mathrm{A}}(S|Y)$. □

5.5 Details of Experiments

We train our models based on the data-driven version of the CFUR formulation presented in Section 5.2 using TensorFlow [113].

5.5.1 Experiments on the UCI Adult Data Set

Each sample in the UCI Adult data set has both continuous and categorical features. Table 5.2 lists all the considered features. We perform a one hot encoding on each categorical feature in (S, X) and store the mapping function from the onehot encoding to the categorical data. For the continuous features in X, we restrict them into the interval $(0, 1)$ by normalization.



(a) The Architecture for Case I.

(b) The Architecture for Case II.

Figure 5.11: The Architectures of the Generative Decorrelator and Adversary for UCI Adult Data Set

ARCHITECTURE. For the UCI Adult data set, the two architectures used are shown in Fig. 5.11, where the appended noise has the same size of the input features and the output of the generative decorrelator has the same dimension as the feature variable X. We concatenate the pre-processed data with a same size standard Gaussian random vector, and feed the entire vector to the generative decorrelator. The generative decorrelator consists of two full-connected (FC) hidden layers with the number of neurons as 170 and 130, respectively. Since the output representation X_r has the same dimension as the feature variable X, the output layer of the generative decorrelator has 113 (as shown in Fig. 5.11a) and 107 (as shown in Fig. 5.11b) neurons for Case I and Case II, respectively. We use a leaky Rectified linear unit (ReLU) activation function in the generative decorrelator. Finally, recall that we consider two cases for this data set. Case I with binary (gender) sensitive feature and Case II with

non-binary (gender and relationship) sensitive features. Futhermore, for Case I, the inputs can be either X only or both X and S. Therefore, in Fig. 5.11a, with only X as input to the generative decorrelator, the length of the input vector is 226 and when both X and S (binary) are inputs, the input length is 230. For both scenarios, the length of the generative decorrelator's output is 113. In Fig. 5.11b, since the input is X and S, its input is 230; on the other hand, the length of the generative decorrelator's output is 107 since S is non-binary in this case.

For Case I, the adversarial classifier in Fig. 5.11a consists of three FC layers with the number of neurons as 10, 5 as well as 2, respectively, and it takes X_r as the input and outputs a belief distribution for the binary sensitive variable S (i.e., gender). Here a ReLU is used as the activation function in the two hidden layers and the soft-max is used in the output layer to generate a belief distribution for gender. The same architecture is used for the downstream application of salary classification. For Case II, the adversarial classifier in Fig. 5.11b consists of three FC layers with the number of neurons as 50, 30 as well as 12, respectively. Here a Leaky ReLU is used as the activation function in the two hidden layers and the soft-max is used in the output layer. All of the above models use α-loss with $\alpha = 1$, i.e., log-loss, as the adversary's loss function and are optimized by an Adam optimizer.

5.5.2 Experiments on the UTKFace Data Set

We reshape 200×200 aligned-colorful faces in the UTKFace data set consists into 64×64 colorful images.

ARCHITECTURE. For the UTKFace data set, the used architectures are shown in Figs. 5.12, 5.13 and 5.14. Fig. 5.12 gives the architecture of the CFUR model which consists of an generative decorrelator and an adversarial classifier. The generative decorrelator is implemented by a noisy auto-encoder whose encoder transforms

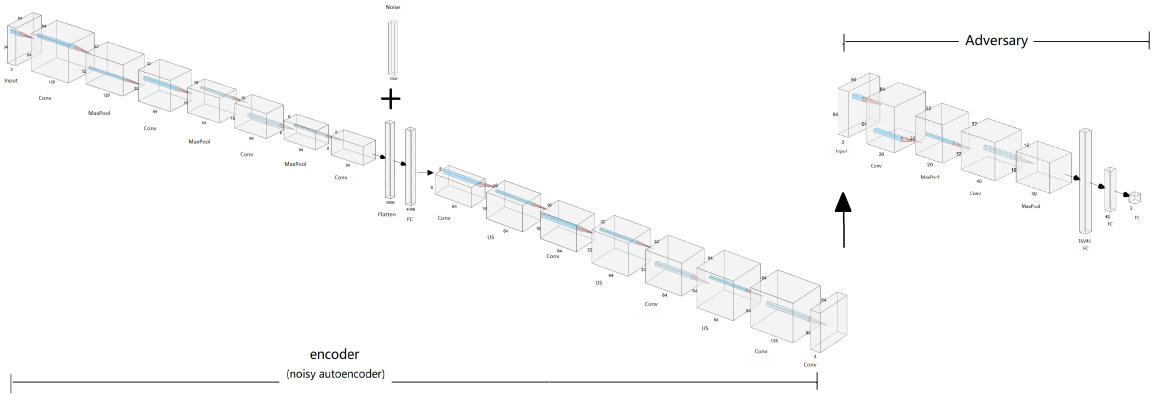


Figure 5.12: The Architecture of the Generative Decorrelator and Adversary for UTKFace Data Set

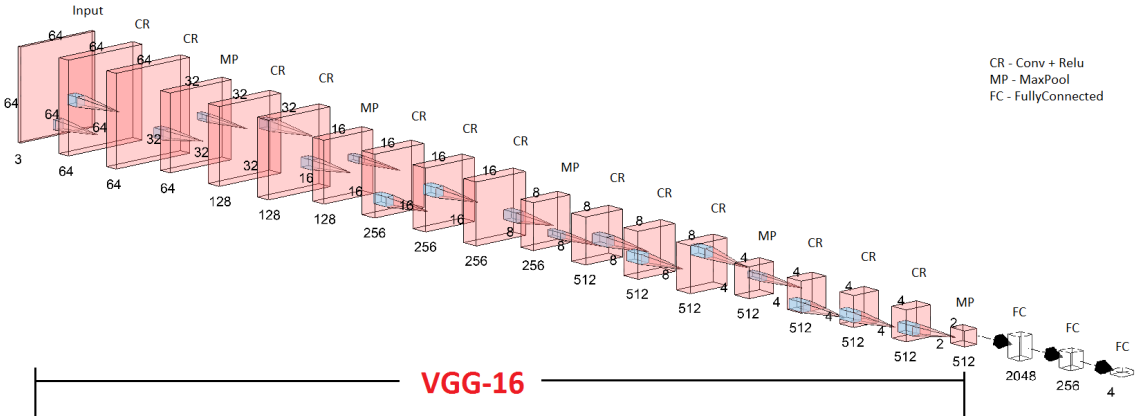


Figure 5.13: The Architecture of the Ethnicity Classifier of UTKFace Data Set

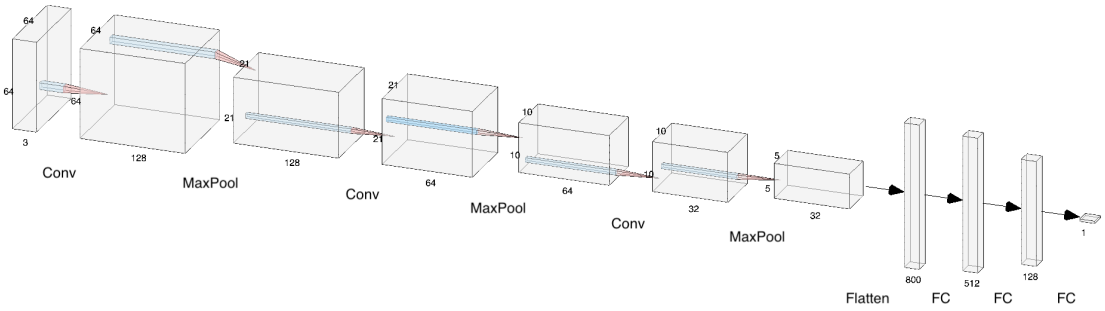


Figure 5.14: The Architecture of the Neural Network for Age Regression of UTK-Face Data Set

the original 64×64 RGB-images to a 4096-dimensional feature vector. Different from standard auto-encoder, which directly feeds the feature vector into its encoder. Here, the feature vector is mixed with a 4096-dimensional standard normal random vector [8] , and then, fed into a decoder to reconstruct a 64×64 colorful image, which is the universal representation X_r. Specifically, the encoder of the auto-encoder consists of 4 convolution layers with 128, 64, 64 and 64 output channels, respectively, and 3 2×2-max pooling layers following the first 3 convolution layers. The encoder is followed by 2 FC layers with 4096 neurons which mixes the noise and the output feature vector. The following decoder part consists of 5 convolution layers with 64, 64, 64, 128 and 3 output channels, respectively, and 3 2×2-up-sampling layers following the first 3 convolution layers. The adversarial classifier takes into the representation X_r and outputs the prediction of the sensitive information gender. It consists of 2 convolution layers with 20 and 40 output channels, respectively, 2 2×2-max pooling layers following each of the convolution layers and 2 full-connected layers with 40 and 2 neurons, respectively. The size of kernels in convolution layers is 3×3. All convolution and full-connected layers use ReLU as the activation function except the last layers of the decoder and the adversarial which use sigmoid and softmax, respectively. The generative decorrelator and adversarial classifier use the square-loss and log-loss as the loss functions, respectively, and both of them are optimized by an Adam Optimizer.

Fig. 5.13 gives the architecture of the downstream non-binary classification for ethnicity. The classifier is built by changing the top (last) 3 FC layers of the VGG 16 model [9] pre-trained on ImageNet. The first one layer has 256 neurons with ReLU as the activation function and is followed by a Dropout layer with the rate 0.5, and

[8] A random vector is a standard normal random vector if all of its components are independent and identically following the standard normal distribution.

[9] https://keras.io/applications/#vgg16

the second one has 4 neurons with softmax as the activation function. The classifier use log-loss and is optimized by a Stochastic Gradient Descent Optimizer. Fig. 5.14 shows the architecture for the the downstream application of age regression. The regressor consists of three 3×3 convolution layers with 128, 64 and 32 output channels, three 2×2-max pooling layers following each of the convolution layers, and three FC layers with 512, 128 and 1 neurons, respectively. All layers use ReLU as the activation function except the last layer which uses a linear activation. The model uses the squared loss as the loss function and is optimized by a Adam Optimizer.

5.6 Concluding Remarks

In this chapter, we introduced an adversarial learning framework with verifiable guarantees for learning ML models that can create censored and fair universal representations for data sets with known sensitive features. The novelty of our approach is in producing representations that hide several sensitive features *jointly and separately*, and simultaneously, provide fairness with respect to the sensitive features for any downstream learning task not known *a priori*. The CFUR framework allows the data holder to learn the censoring and fair scheme (a randomized mapping that decorrelates the sensitive and non-sensitive features) directly from the data set without requiring access to data set statistics. Under the CFUR framework, finding the optimal generative decorrelator, subject to a fidelity constraint on the representation, is formulated as a constrained two-player game, namely a game between the generative decorrelator and an adversary. With *α-loss* as the adversary's loss, the CFUR framework can provide guarantees against strong information-theoretic adversaries, such as MAP (soft 0-1 loss) and MI (log-loss) adversaries. It also allows enforcing fairness, quantified via demographic parity. Furthermore, for *a priori* known classification task, we also proved that our CFUR framework can be modified to approximate

fairness via demographic parity, equalized odds or equality of opportunity.

Yet another highlight of our work is the performance validation of the CFUR framework on publicly available real data sets including images and data sets involving a mix of categorical and continuous features. Our results allow us to visually highlight two key results: (a) the tradeoff between the representation fidelity (via utilities of downstream applications, e.g., the accuracy of classifications and mean absolute error for a regression) and censoring guarantees (via the adversarial accuracy in learning sensitive features jointly or separately); (b) the tradeoff between the representation fidelity and fairness guarantee (via the maximal difference among conditional probabilities of a prediction/estimation on various sensitive features). Our results also reveal to some extent the effect on performance of the choice of decorrelator and adversary architectures (chosen as deep neural networks) for different data sets.

Chapter 6

OPTIMAL CONVERSION FROM RÉNYI DIFFERENTIAL PRIVACY TO (ϵ, δ)-DIFFERENTIAL PRIVACY: AN INFORMATION-THEORETICAL APPROACH

In this chapter, we characterize the privacy performance, in terms of (ϵ, δ)-differential privacy (DP), of a mechanism that satisfies a given level of Rényi differential privacy (RDP). Based on the existing result in literature that (ϵ, δ)-differential privacy can be expressed via Hockey-stick divergence [83, 84], we formulate the optimal conversion from RDP to (ϵ, δ)-DP as an optimization problem with both objective and constraint functions as f-divergences. We simplify the optimization problem to a univariate-convex project by making use of the known property of the joint range of two f-divergences [87], and obtain the optimal conversion from RDP to (ϵ, δ)-DP. This result allows us to improve the adaptive composition theorem of Gaussian mechanisms, i.e., privacy mechanisms of adding Gaussian noise to functions of a dataset, as well as the privacy parameter of the differentially private stochastic gradient descent (DP-SGD) algorithm [70]. Note that the proof details for all theorems and lemmas are in Section 6.4.

6.1 Preliminaries

To incorporation the notion of neighboring data sets, we use some new notations in this chapter, which are clarified as follows. We use $\mathcal{D}$ to indicate the collection of data sets with elements from a given support, and Y to be the range of an arbitrary (possibly random) mapping of any data set in $\mathcal{D}$. Let D and Y indicate two random variables taking values from $\mathcal{D}$ and $\mathcal{Y}$, respectively, and $P_{Y|D} : \mathcal{D} \to \mathcal{Y}$ be the map-

ping/mechanism from D to Y. Two data sets $d, d' \in \mathcal{D}$ are said to be neighboring (denoted by $d \sim d'$) if they differ in only one element. We use $P_{Y|D=d}$ and $P_{Y|D=d'}$ to indicate the the corresponding output distributions of the mechanism $P_{Y|D}$ for taking in the data sets d and d'. For an arbitrary subset $\mathcal{A}$ of $\mathcal{Y}$, $P_{Y|D=d}(\mathcal{A})$ and $P_{Y|D=d'}(\mathcal{A})$ are the corresponding probabilities for the mechanism $P_{Y|D}$ mapping the datastes d and d' to $\mathcal{A}$, respectively.

Definition 6.1.1. *A mechanism $P_{Y|D} : \mathcal{D} \to \mathcal{Y}$ is said to be*

- *(ϵ, δ)-DP for any given non-negative ϵ and $\delta \in [0, 1]$, if*

$$\sup_{\mathcal{A}\in\mathcal{Y}, d\sim d'\in\mathcal{D}} P_{Y|D=d}(\mathcal{A}) - e^{\epsilon} P_{Y|D=d'}(\mathcal{A}) \leq \delta. \tag{6.1}$$

- *(α, γ)-RDP for any given $\alpha > 1$ and non-negative γ , if*

$$\sup_{d\sim d'\in\mathcal{D}} D_{\alpha}(P_{Y|D=d} \| P_{Y|D=d'}) \leq \gamma, \tag{6.2}$$

  where $D_{\alpha}(P\|Q) \triangleq \frac{1}{\alpha-1} \log \mathbb{E}_Q \left[\left(\frac{dP}{dQ} \right)^{\alpha} \right]$ denotes the Rényi divergence of order α between two probability distributions P and Q.

The above two relaxed versions of DP can be respectively expressed via two specific instances of f-divergence. The definition of f-divergence and the two instances: $\mathsf{E}_{e^{\epsilon}}$-divergence and χ^{α}-divergence [1] are presented below.

Definition 6.1.2. *([81, 82]) Given two probability distributions P and Q and a real-valued convex function $f(\cdot)$ satisfying $f(1) = 0$, the f-divergence between P and Q is defined as*

$$D_f(P\|Q) \triangleq \int f\left(\frac{dP}{dQ}\right) dQ. \tag{6.3}$$

[1] The χ^{α}-divergence is also known as Hellinger divergence in [**?**].

In addition, for $f(t) = (t - e^{\epsilon})_+ = \max\{0, t - e^{\epsilon}\}$ *with* $\epsilon \geq 0$*, the corresponding* f*-divergence is called* $\mathsf{E}_{e^{\epsilon}}$*-divergence given by*

$$\mathsf{E}_{e^{\epsilon}}(P\|Q) = \int (dP - e^{\epsilon} dQ)_+ = \int \max\left\{0,\ dP - e^{\epsilon} dQ\right\}, \tag{6.4}$$

and for $f(t) = \frac{1}{\alpha-1}(t^{\alpha} - 1)$ *with* $\alpha > 1$*, the corresponding* f*-divergence is called* χ^{α}*-divergence given by*

$$\chi^{\alpha}(P\|Q) = \frac{1}{\alpha - 1}\left(\int \left(\frac{dP}{dQ}\right)^{\alpha} dQ - 1\right). \tag{6.5}$$

The connection between (ϵ, δ)-DP and $\mathsf{E}_{e^{\epsilon}}$-divergence as well as that between RDP and χ^{α}-divergence are shown in the following proposition.

Proposition 3. *A mechanism* $P_{Y|D} : \mathcal{D} \to \mathcal{Y}$ *satisfies*

- (ϵ, δ)*-DP if and only if [83, 84]*

$$\sup_{d \sim d' \in \mathcal{D}} \mathsf{E}_{e^{\epsilon}}\left(P_{Y|D=d} \| P_{Y|D=d'}\right) \leq \delta. \tag{6.6}$$

- (α, γ)*-RDP if and only if*

$$\sup_{d \sim d' \in \mathcal{D}} \chi^{\alpha}\left(P_{Y|D=d} \| P_{Y|D=d'}\right) \leq \chi_{\alpha}(\gamma), \tag{6.7}$$

  where the function $\chi_{\alpha}(\gamma)$ *is defined as*

$$\chi_{\alpha}(\gamma) \triangleq \frac{e^{(\alpha-1)\gamma} - 1}{\alpha - 1}, \tag{6.8}$$

  which indicates the one-to-one mapping from Rényi divergence to χ^{α}*-divergence for an arbitrary pair of probability distributions.*

Note that $d \sim d'$ *are the neighboring data sets and* $P_{Y|D=d}$ *and* $P_{Y|D=d'}$ *are the two output distributions over* $\mathcal{Y}$ *of the mechanism* $P_{Y|D}$ *while taking in* d *and* d'*, respectively.*

The results in Proposition 3 will be used to derive the optimal conversion from (α, γ)-RDP to (ϵ, δ)-DP, which is tighter than the existing one provided by the following theorem.

Theorem 19 ([70, 80])**.** *If the mechanism $P_{Y|D}$ is (α, γ)-RDP for any $\alpha > 1, \gamma \geq 0$, then it satisfies (ϵ, δ)-DP for any $\delta \in (0, 1)$ and*

$$\epsilon = \gamma - \frac{\ln \delta}{\alpha - 1}. \tag{6.9}$$

6.2 Optimal Conversion from RDP to (ϵ, δ)-DP

To characterize the optimal conversion from RDP to (ϵ, δ)-DP [2] , it is sufficient to ask that for an arbitrary mechanism $P_{Y|D}$ satisfying (α, γ)-RDP, what are the smallest values of ϵ and δ such that $P_{Y|D}$ satisfies (ϵ, δ)-DP? Let $\mathcal{P}(\mathcal{Y}|\mathcal{D})$ indicate the collection of stochastic mappings from $\mathcal{D}$ to $\mathcal{Y}$. By using the results in Proposition 3, we can formulate this question as the following optimization problem:

$$\delta(\gamma|\alpha, \epsilon) = \sup_{P_{Y|D} \in \mathcal{P}(\mathcal{Y}|\mathcal{D})} \sup_{d \sim d' \in \mathcal{D}} \mathsf{E}_{e^\epsilon}(P_{Y|D=d} \| P_{Y|D=d'}) \tag{6.10a}$$

$$\text{s.t.} \quad \sup_{d \sim d' \in \mathcal{D}} \chi^\alpha(P_{Y|D=d} \| P_{Y|D=d'}) \leq \chi_\alpha(\gamma), \tag{6.10b}$$

where $\chi_\alpha(\gamma)$, defined in (6.8), is the corresponding value of χ^α-divergence when Rényi divergence is γ. Therefore, for any mechanism satisfying (α, γ)-RDP, it satisfies (ϵ, δ')-DP for any $\delta' \geq \delta(\gamma|\alpha, \epsilon)$ and may not satisfy (ϵ, δ')-DP for any $\delta' < \delta(\gamma|\alpha, \epsilon)$. That is, for any given values of α, γ and ϵ, this quantity $\delta(\gamma|\alpha, \epsilon)$ is the smallest value of δ such that any (α, γ)-RDP mechanism satisfies (ϵ, δ)-DP, and the mapping $\gamma \to \delta(\gamma|\alpha, \epsilon)$ determines the optimal conversion from RDP to (ϵ, δ)-DP.

[2] Note that in the collection of mechanisms that satisfying (ϵ, δ)-DP with $\delta > 0$, there exist mechanisms such that the maximal Rényi divergence between its output distributions can be arbitrary large, i.e., the mechanism can only satisfy (α, γ)-RDP for $\gamma \to \infty$. Therefore, there is no (α, γ)-RDP privacy for any bounded γ guaranteed by (ϵ, δ)-DP with $\delta > 0$.

In the sequel, we present an computational efficient method of obtaining the optimal conversion by transforming and simplifying the optimization in (6.10a). For the simplification of expressions, we use two arbitrary probability distributions P and Q over the same support $\mathcal{Y}$, i.e., $P, Q \in \mathcal{P}(\mathcal{Y})$, to indicate a pair of output distributions $P_{Y|D=d}, P_{Y|D=d'}$ of an arbitrary mechanism $P_{Y|D} \in \mathcal{P}(\mathcal{Y}|\mathcal{D})$ by taking in any pair of neighboring types $d \sim d' \in \mathcal{D}$ [3] . The optimization problem in (6.10a) is rewritten as

$$\delta(\gamma|\alpha,\epsilon) = \sup_{P,Q\in\mathcal{P}(\mathcal{Y})} \quad \mathsf{E}_{e^\epsilon}(P\|Q) \tag{6.11a}$$

$$\text{s.t.} \quad \chi^\alpha(P\|Q) \le \chi_\alpha(\gamma). \tag{6.11b}$$

Fig. 6.1 shows the plot of $\delta(\gamma|\alpha,\epsilon)$ versus γ for given α and ϵ, and from which, it can be observed that for given α and ϵ, the mapping $\gamma \to \delta(\gamma|\alpha,\epsilon)$ constitutes the upper boundary of the joint region of E_{e^ϵ}-divergence and χ^α-divergence given by

$$\mathcal{R}_\alpha \triangleq \Big\{ \big(\chi^\alpha(P\|Q), \mathsf{E}_{e^\epsilon}(P\|Q)\big) \big| P, Q \in \mathcal{P}(\mathcal{Y}) \Big\}. \tag{6.12}$$

That is, the optimal conversion from RDP to (ϵ,δ)-DP is determined by the upper boundary of the joint region $\mathcal{R}_\alpha$ defined in (6.12), and therefore, is equivalently characterized by the inverse mapping $\delta \to \gamma(\delta|\alpha,\epsilon)$ given by

$$\gamma(\delta|\alpha,\epsilon) = \inf_{P,Q\in\mathcal{P}(\mathcal{Y})} \quad \chi_\alpha^{-1}\Big(\chi^\alpha(P\|Q)\Big) \tag{6.13a}$$

$$\text{s.t.} \quad \mathsf{E}_{e^\epsilon}(P\|Q) \ge \delta \tag{6.13b}$$

where $\chi_\alpha^{-1}(t) = \frac{1}{\alpha-1}\log(1+(\alpha-1)t)$ is the inverse of the function $\chi_\alpha(\cdot)$ defined in (6.8) and gives the corresponding value of Rényi divergence when χ^α-divergence is t.

The above observation allows us to cast the problem of converting from (α,γ)-RDP to (ϵ,δ)-DP as characterizing the joint range of E_{e^ϵ} and χ^α divergences. Therefore,

[3] Note that the divergences between $P_{Y|D=d}, P_{Y|D=d'}$ are not necessarily smaller than the divergences between P_d and $P_{d'}$. For example, for a querying answer $Y \in \{0,1\}$ indicating a specific item belonging to a dataset or not, the output distributions $P_{Y|D=d}$, $P_{Y|D=d'}$ can have infinity divergence if the specific item is the one that the two datasets d, d' differ in.

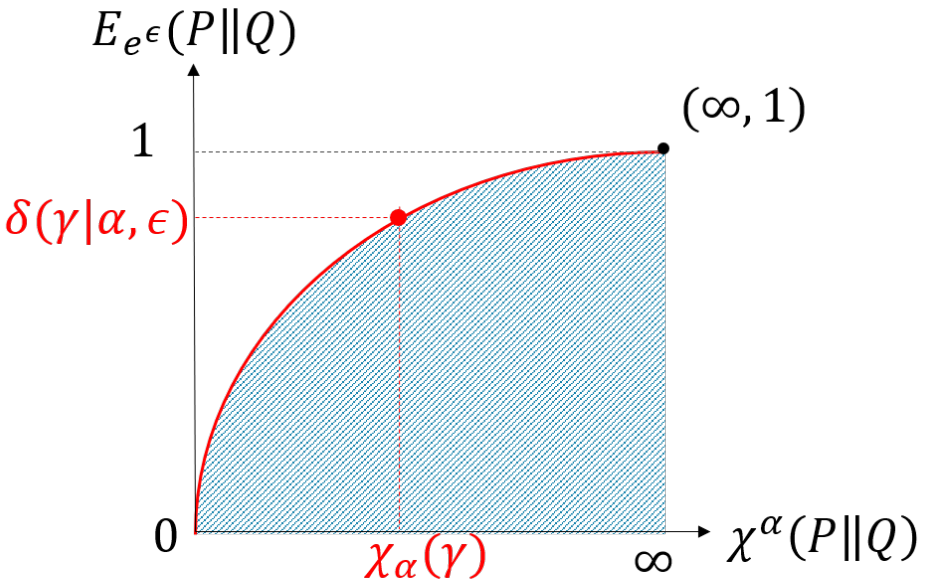


Figure 6.1: Joint Region of $\chi^\alpha(P\|Q)$ and $\mathsf{E}_{e^\epsilon}(P\|Q)$ for All $P, Q \in \mathcal{P}(\mathcal{Y})$

we can simplify the optimization problem in (6.13) to a univariate convex program by using a property of the joint region of any two f-divergences presented in the proposition below.

Theorem 20. *([87, Theorem 8]) For any two f-divergences D_{f_1} and D_{f_2}, there is*

$$\left\{\left(D_{f_1}(P\|Q), D_{f_2}(P\|Q)\right)\middle| P, Q \in \mathcal{P}(\mathcal{Y})\right\} = \mathsf{conv}(\mathcal{B}) \tag{6.14}$$

where $\mathsf{conv}(\cdot)$ denotes the convex hull operator and

$$\mathcal{B} \triangleq \left\{\left(D_{f_1}(P_{\mathsf{b}}\|Q_{\mathsf{b}}), D_{f_2}(P_{\mathsf{b}}\|Q_{\mathsf{b}})\right)\middle| P_{\mathsf{b}}, Q_{\mathsf{b}} \in \mathcal{P}(\{0,1\})\right\}. \tag{6.15}$$

Making use of the result in Theorem 20, we simplify the optimization (6.13), which can potentially be of significant complexity, to a univariate convex program, which is a simple tractable problem.

Theorem 21. *For any $\alpha > 1$, $\epsilon \geq 0$ and $\delta \in [0, 1)$,*

$$\gamma(\delta|\alpha, \epsilon) = \epsilon + \frac{1}{\alpha - 1} \log \min_{p \in (\delta, 1)} \left(p^\alpha (p - \delta)^{1-\alpha} + (1 - p)^\alpha (e^\epsilon - p + \delta)^{1-\alpha} \right). \tag{6.16}$$

It can be shown the term inside the logarithm is convex in p and hence this optimization problem can be numerically solved with an arbitrary accuracy. It seems, however, not simple to analytically derive $\gamma(\delta|\alpha, \epsilon)$. Nevertheless, we obtain a tight lower bound in the following theorem.

Theorem 22. *For any $\epsilon \geq 0$ and $\alpha > 1$, we have*

$$\gamma(0|\alpha, \epsilon) = 0,$$

$$\gamma(\delta|\alpha, \epsilon) = \epsilon - \log(1 - \delta), \qquad \text{if } \alpha\delta \geq 1, \tag{6.17}$$

$$\gamma(\delta|\alpha, \epsilon) \geq \max\{g(\alpha, \epsilon, \delta), f(\alpha, \epsilon, \delta)\}, \quad \text{if } 0 < \alpha\delta < 1, \tag{6.18}$$

where

$$g(\alpha, \epsilon, \delta) \triangleq \epsilon - \frac{1}{\alpha - 1} \log \frac{\zeta_\alpha}{\delta},$$

with $\zeta_\alpha \triangleq \frac{1}{\alpha} \left(1 - \frac{1}{\alpha}\right)^{\alpha - 1}$ and

$$f(\alpha, \epsilon, \delta) \triangleq \epsilon + \frac{1}{\alpha - 1} \log \left((e^\epsilon - \alpha\delta) \left(\frac{\delta - 1}{\delta - e^\epsilon} \right)^\alpha + \alpha\delta \right).$$

In Fig. 6.2, we numerically solve (6.16) for three pairs of (α, ϵ) and compare them with their corresponding bounds obtained from Theorem 22, highlighting the tightness of the above lower bound. In practice, it is often appealing to design differentially private mechanisms with a hard-coded value of δ. To address this practical need, we convert the lower bound in Theorem 22 to an upper bound on $\epsilon(\gamma|\alpha, \delta)$.

Lemma 5. *Given a mechanism satisfying (α, γ)-RDP for any non-negative γ and $\alpha > 1$, it satisfies (ϵ, δ)-DP for any $\delta \in (0, 1)$ and $\epsilon \geq \epsilon(\gamma|\alpha, \delta)$, where the quantity $\epsilon(\gamma|\alpha, \delta)$ is give by*

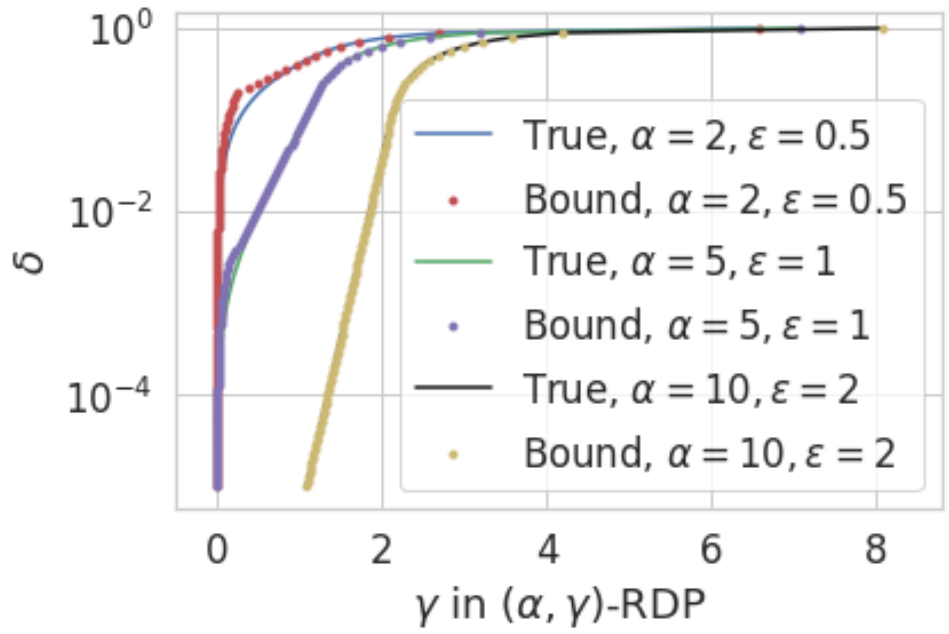


Figure 6.2: True Values (Obtained via Numerically Solving the Convex Project in (6.16)) Versus the Bounds (Obtained from Theorem 22) for Three Pairs of (α, ϵ)

- *if* $\alpha\delta \geq 1$,

$$\epsilon(\gamma|\alpha, \delta) = \max\left\{0,\ \gamma + \log(1-\delta)\right\} \tag{6.19}$$

- *if* $0 < \alpha\delta < 1$,

$$\epsilon(\gamma|\alpha, \delta) \leq \frac{1}{\alpha - 1} \min\left\{ \max\left\{0,\ (\alpha-1)\gamma - \log\frac{\delta}{\zeta_\alpha}\right\}, \log\left(\frac{(\alpha-1)\chi_\alpha(\gamma)}{\alpha\delta} + 1\right)\right\}, \tag{6.20}$$

where $\chi_\alpha(\gamma)$ *is defined in* (6.8). *Moreover,* $\epsilon(0|\alpha, \delta) = 0$.

The the key idea in the proof of Lemma 5 is to get the inverse functions (in term of ϵ) of the functions $g(\alpha, \epsilon, \delta)$ and $f(\alpha, \epsilon, \delta)$ in Theorem 22. Note that the function $g(\alpha, \epsilon, \delta)$ is linear in ϵ and invertible. For the function $f(\alpha, \epsilon, \delta)$, we bound its inverse function from above via an approximation of $f(\alpha, \epsilon, \delta)$.

Note that the conversion from (α, γ)-RDP to (ϵ, δ)-DP in Lemma 5 is much tighter than the one obtained from Theorem 19, which is characterized by $\epsilon(\gamma|\alpha, \delta) \leq \gamma -$

$\frac{1}{\alpha-1}\log\delta$. Specifically, for $\alpha\delta \geq 1$, the expression of $\epsilon(\gamma|\alpha,\delta)$ in (6.19) is less than γ, and therefore, much smaller than the upper bound in Theorem 19; and for $\alpha\delta < 1$, the upper bound in (6.20) is also smaller than that in Theorem 19 since the expression ζ_α (defined in Theorem 22) is less than 1 for $\alpha > 1$. It must be mentioned that Balle et al. [114, Theorem 21] has recently proved the bound $\epsilon(\gamma|\alpha,\delta) \leq \gamma - \frac{1}{\alpha-1}\log\frac{\delta}{\zeta_\alpha}$, via a fundamentally different approach which is weaker than Lemma 5.

Remark 6. *As an important special case, this lemma demonstrates that an (α,γ)-RDP mechanism provides $(0,\delta)$-DP guarantee if (i) $1-e^{-\gamma} < \frac{1}{\alpha}$ and $\delta \in \left[\zeta_\alpha e^{(\alpha-1)\gamma}, \frac{1}{\alpha}\right]$, or (ii) $\delta > \max\{1-e^{-\gamma}, 1/\alpha\}$. Notice that this is significantly stronger than the upper bound obtained from Theorem 19, i.e., $\epsilon(\gamma|\alpha,\delta) \leq \gamma - \frac{1}{\alpha-1}\log\delta$, from which $(0,\delta)$-DP cannot be achieved.*

6.3 Applications of the Optimal Conversion from RDP to (ϵ,δ)-DP

In this section, we improve the composition theorem of (ϵ,δ)-DP by fusing the tighter conversion in Lemma 5 and the method of moments accountant (MA), which was recently proposed by Abadi et al. [70]. The cornerstone of MA is the linear composability of RDP, which is expressed in the following theorem.

Theorem 23. *[70, Theorem 2] Suppose that a mechanism $P_{Y^T|D}$ consists of a sequence of adaptive mechanisms $P_{Y_1|D}, P_{Y_2|D,Y_1}, \cdots, P_{Y_T|D,Y_1,\cdots,Y_{T-1}}$, where $P_{Y_i|D,Y_1,\cdots,Y_{i-1}} : \prod_{j=1}^{i-1} : \mathcal{Y}_j \times \mathcal{D} \to \mathcal{Y}_i$. Then, for any $\alpha > 1$,*

$$D_\alpha(P_{Y^T|D=d}\|P_{Y^T|D=d'}) \leq \sum_{i=1}^{T} D_\alpha(P_{Y_i|D=d,Y^{i-1}(d)}\|P_{Y_i|D=d',Y^{i-1}(d')}) \tag{6.21}$$

where $Y^{i-1}(d)$ and $Y^{i-1}(d')$ indicate the corresponding output sequences, generated by mechanisms $P_{Y_1|X}, \cdots, P_{Y_{i-1}|X,Y_1,\cdots,Y_{i-2}}$, from the neighboring data sets d and d', respectively, and $Y^T = (Y_1, Y_2, \cdots, Y_T)$ takes value from the support $\mathcal{Y}_1 \times \cdots \mathcal{Y}_T$.

Theorem 23 indicates the linear composability of RDP. In addition, making use of the results in Theorem 23 and Lemma 5, one can tailor the composition theorem of (ϵ, δ)-DP for any specific mechanism (e.g., Gaussian mechanisms adding Gaussian noise to data functions) by calculating the sequence of Rényi divergences of the composed mechanisms. Therefore, in practice this MA-based approach can be more accurate than the well-known advanced composition theorems [78, 79], which is applicable for general mechanisms and fails to capture the uniqueness of a specific mechanism, e.g., the properties of Gaussian noise in Gaussian mechanisms. For the rest of this section, we assume that $P_{Y_i|X,Y_1,\cdots,Y_{i-1}}$, for $i \in \{1, 2, \cdots, T\}$, are Gaussian mechanisms, and apply the tighter conversion in Lemma 5 to improve the privacy parameters of the adaptive composition $P_{Y^T|D}$ presented in [70], which is obtained via the result in Theorem 19.

6.3.1 Bounds on Privacy Parameters of Gaussian Composition

Given a Gaussian mechanism, the maximal value of the Rényi divergence between the corresponding output distributions of any pair of neighboring data sets is a linear function of α as shown in the following lemma.

Lemma 6. *Given a data function f with unit L_2-sensitivity, i.e., $\sup_{d\sim d'\in\mathcal{D}} \|f(d) - f(d')\|_2 = 1$, and a Gaussian mechanism $P_{Y|D}$ with the noise variance σ^2, i.e., $Y = f(D) + N$ and $N \sim \mathcal{N}(0, \sigma^2)$, the Rényi divergence between the output distributions for any pair of neighboring data sets d and d' is bounded from above by*

$$\sup_{d\sim d'\in\mathcal{D}} D_\alpha(P_{Y|D=d} \| P_{Y|D=d'}) = \frac{\alpha}{2\sigma^2}, \tag{6.22}$$

where $\alpha > 1$. That is, for a data function with unit L_2-sensitivity, the Gaussian mechanism with variance σ^2 satisfies $(\alpha, \rho\alpha)$-RDP with $\rho \triangleq \frac{1}{2\sigma^2}$.

The proof of Lemma 6 is mainly based on the probability distribution of Gaussian noise. Without loss of generality, we assume that the L_2-sensitivity of any function of interest is unit. Note that if the L_2-sensitivity of a function of interest is c, we can scale the variance of a Gaussian mechanism as $c^2\sigma^2$ to get the same result as shown in Lemma 6.

In light of the linear composability of RDP in Theorem 23 and the result in Lemma 6, we know that the T-fold adaptive composition [4] $P_{Y^T|D}$ of a Gaussian mechanism $P_{Y|D}$ with variance σ^2 satisfies $(\alpha, \rho\alpha T)$-RDP. Therefore, from the result in Theorem 19, one can obtain that the composition $P_{Y^T|D}$ satisfies (ϵ, δ)-DP with $\delta \in (0, 1)$ and

$$\epsilon = \inf_{\alpha>1} \rho\alpha T - \frac{\log \delta}{\alpha - 1} = \rho T + \sqrt{4\rho T \log \frac{1}{\delta}}. \tag{6.23}$$

In (6.23), from the convexity of the objective function in α, the optimal solution $\alpha^* = 1 + \sqrt{-\rho T \log \delta}$ is obtained by calculating the root of the first derivative of the objective function in α.

We next use the result in Lemma 5 to improve the privacy parameter in (6.23) of the T-folder composition $P_{Y^T|D}$. To do so, define

$$\epsilon(\rho, T|\delta) \triangleq \inf_{\alpha>1} \epsilon(\rho\alpha T|\alpha, \delta). \tag{6.24}$$

Thus, $P_{Y^T|D}$ is $(\epsilon(\rho, T|\delta), \delta)$-DP for any $\delta \in (0, 1)$. Invoking Lemma 5, we can obtain a bound $\epsilon(\rho, T|\delta)$ as shown in the following lemma.

Lemma 7. *The T-fold adaptive homogeneous composition of a Gaussian mechanism with variance σ^2 is $(\epsilon(\rho, T|\delta), \delta)$-DP with $\delta \in (0, 1)$ and*

$$\epsilon(\rho, T|\delta) \leq \min\left\{\epsilon_0(\rho, T),\ \epsilon_1(\rho, T),\ \left(\frac{\rho T}{\delta} + \log(1 - \delta)\right)_+\right\}, \tag{6.25}$$

[4] The adaptive composition requires that the output of the previous mechanism is used as a part of the input of the subsequent mechanism.

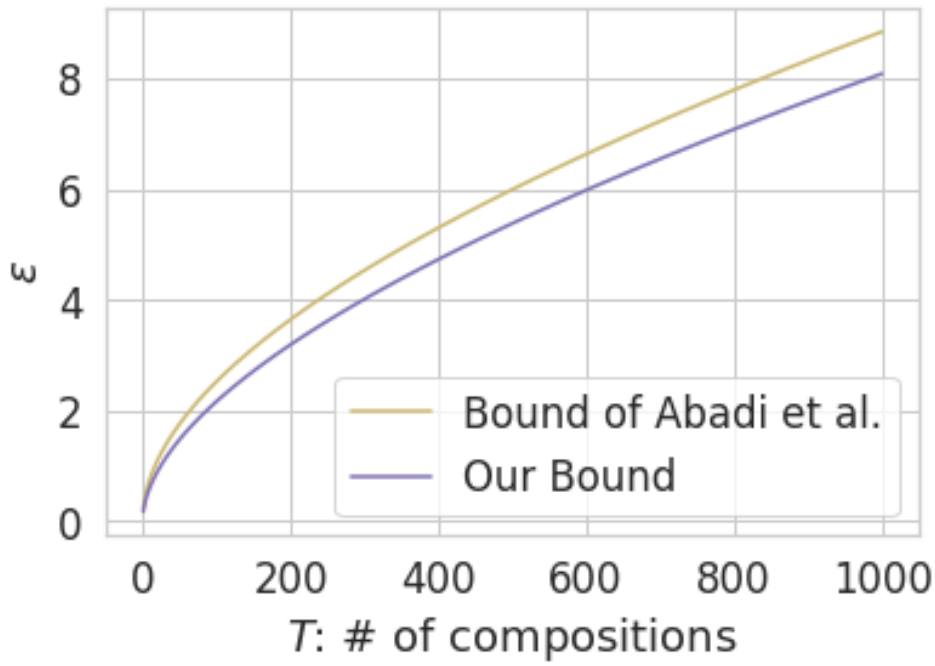


Figure 6.3: Comparison of Our Bound in Lemma 7 on $\epsilon(\rho, T|\delta)$ with (6.23) Obtained by Abadi et al. for $\sigma = 20$ and $\delta = 10^{-5}$

where $\rho = \frac{1}{2\sigma^2}$ *and*

$$\epsilon_0(\rho, T) \triangleq \inf_{\alpha \in (1, \frac{1}{\delta}]} \left(\rho\alpha T - \frac{1}{\alpha - 1} \log \frac{\delta}{\zeta_\alpha} \right)_+, \tag{6.26}$$

$$\epsilon_1(\rho, T) \triangleq \inf_{\alpha \in (1, \frac{1}{\delta}]} \frac{1}{\alpha - 1} \log \left(1 + \frac{e^{\rho\alpha(\alpha-1)T} - 1}{\alpha\delta} \right), \tag{6.27}$$

and ζ_α *is as defined in Theorem 22.*

The bound given in Lemma 7 can shed light on the optimal variance of the Gaussian mechanism $P_{Y|D}$ required to ensure that $P_{Y^T|D}$ is (ϵ, δ)-DP. To put our result about the variance in perspective, we first mention two previously-known bounds on σ^2. Advanced composition theorems (see, e.g., [78, Theorem III.3]) require $\sigma^2 = \Omega(\frac{T \log(1/\delta) \log(T/\delta)}{\epsilon^2})$. Abadi et al. [70, Theorem 1] improved this result by showing that σ^2 suffices to be linear in T; more precisely, $\sigma^2 = \Omega(\frac{T \log(1/\delta)}{\epsilon^2})$. To have a better comparison with our final result, we write this result more explicitly. It follows

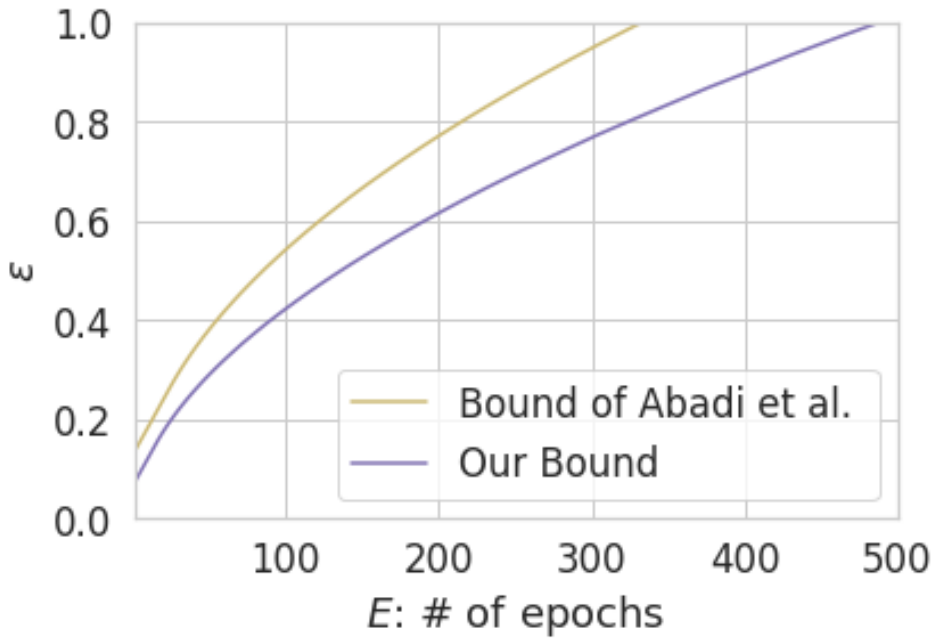


Figure 6.4: Privacy Parameter ϵ of DP-SGD with $\sigma = 4$, $q = 0.001$ and $\delta = 10^{-5}$

from (6.23) that

$$\sigma^2 \geq \frac{T}{2\left(\epsilon - 2\log\delta - 2\sqrt{-(\epsilon - \log\delta)\log\delta}\right)}, \tag{6.28}$$

and hence assuming δ is sufficiently small, we obtain

$$\sigma^2 \geq \frac{2T}{\epsilon^2}\log\frac{1}{\delta} + \frac{T}{\epsilon} + O\left(\frac{1}{\log\delta^{-1}}\right). \tag{6.29}$$

We are now in order to state our result.

Theorem 24. *The T-fold adaptive homogeneous composition of a Gaussian mechanism with variance σ^2 is (ϵ, δ)-DP, for $\epsilon > 2\delta\log\frac{1}{\delta}$, if*

$$\sigma^2 \geq \frac{2T}{\epsilon^2}\log\frac{1}{\delta} + \frac{T}{\epsilon} - \frac{2T}{\epsilon^2}\left(\log(2\log\delta^{-1}) + 1 - \log\epsilon\right) + O\left(\frac{\log^2(\log\delta^{-1})}{\log\delta^{-1}}\right). \tag{6.30}$$

The proof of this theorem is based on a relaxation of Theorem 22 obtained by ignoring $f(\alpha, \epsilon, \delta)$. Considering both f and g will result in a stronger result at the

expense of more involved analysis. Comparing with (6.29), Theorem 24 indicates that, providing δ is sufficiently small, the variance of each constituent Gaussian mechanism can be reduced by $\frac{2T}{\epsilon^2}\left(\log(2\log\delta^{-1}) + 1 - \log\epsilon\right)$.

6.3.2 Illustrations

In this section, we empirically compare our bound on $\epsilon(\rho, T|\delta)$ given in Lemma 7 with the privacy parameter (6.23) obtained by Abadi et al. via Theorem 19, which has been extensively used in the state-of-the-art differentially private machine learning algorithms, e.g., [115, 74, 116, 117, 118, 119, 120, 76]. We do so in two different settings: (1) vanilla T-fold composition of the Gaussian mechanism with fixed variance, and (2) DP-SGD algorithm.

Vanilla Gaussian Composition: Here, we wish to obtain bounds on the privacy parameter ϵ of $P_{Y^T|D}$ where $P_{Y|D}$ is a Gaussian mechanism with $\sigma = 20$. In Fig. 6.3, we compare the value of ϵ from Lemma 7 with that from (6.23) for $\delta = 10^{-5}$. According to this plot, our result enables us to achieve a smaller privacy parameter by up to 0.75, i.e., $\max_{T\in[1000]} \epsilon_{\mathsf{Abadi}}(\rho, T|\delta) - \epsilon(\rho, T|\delta) = 0.75$ where $\epsilon_{\mathsf{Abadi}}(\rho, T|\delta)$ is the ϵ given in (6.23). This privacy amplification may have important impacts on recent private deep leaning algorithms. Alternatively, one can observe that our result allows for more iteration for the same ϵ, for instance 100 more iterations for any ϵ larger than 6.

DP-SGD: SGD is the standard algorithm for training many machine learning models. In order to fit a model without compromising privacy, a standard practice is to add Gaussian noise to the gradient of each mini-batch, see e.g., [70, 71, 118, 115, 72, 73]. The prime use of MA was to exploit the RDP's simple composition property in deriving the privacy parameters of the DP-SGD algorithm [70, Algorithm 1]. To have a fair comparison, we implement this algorithm with the sub-sampling rate

$q = 0.001$ and noise parameter $\sigma = 4$, and then, compute its privacy parameters via (6.23) with $\rho = q^2/((1-q)\sigma^2)$ (see [70, Lemma 3]) and $\delta = 10^{-5}$. We then compare it in Fig. 6.4 with our privacy parameter calculated from Lemma 7 with the same ρ and σ. As demonstrated in this figure, our result allows remarkably more epochs (often over a hundred) within the same privacy budget and thus providing higher utility.

6.4 Proof Details

6.4.1 Proof of Theorem 21

First notice that, in light of Theorem 20, the convex set $\mathcal{R}_\alpha$ defined in (6.12) is equal to the convex hull of the set $\mathcal{B}_{\alpha,\epsilon}$ given by

$$\mathcal{B}_{\alpha,\epsilon} = \{(\chi^\alpha(P_\mathsf{b}\|Q_\mathsf{b}), \mathsf{E}_{e^\epsilon}(P_\mathsf{b}\|Q_\mathsf{b}))\big|P_\mathsf{b}, Q_\mathsf{b} \in (\{0,1\})\} \tag{6.31}$$

where $P_\mathsf{b} = \mathsf{Bernoulli}(p)$ and $Q_\mathsf{b} = \mathsf{Bernoulli}(q)$ with parameters $p, q \in (0,1)$. For any pair of such distributions, define $\tilde{\gamma} \triangleq \chi^\alpha(P_\mathsf{b}\|Q_\mathsf{b})$ and $\delta \triangleq \mathsf{E}_{e^\epsilon}(P_\mathsf{b}\|Q_\mathsf{b})$. We first show that the convex hull of the set $\mathcal{B}_{\alpha,\epsilon}$ is the $\bar{\mathcal{B}}_{\alpha,\epsilon}$ given by

$$\bar{\mathcal{B}}_{\alpha,\epsilon} = \{(\tilde{\gamma}, \delta)\big|\delta \in [0,1), \tilde{\gamma} \geq \tilde{\gamma}(\delta)\} \tag{6.32}$$

with $\tilde{\gamma}(\delta)$ given by

$$\tilde{\gamma}(\delta) = \inf_{0<p,q<1} \chi^\alpha(P_\mathsf{b}\|Q_\mathsf{b}) \tag{6.33}$$
$$\text{s.t. } \mathsf{E}_{e^\epsilon}(P_\mathsf{b}\|Q_\mathsf{b}) \geq \delta.$$

To this goal, we need to demonstrate that for any $\lambda \in [0,1]$ and pairs of points $(\tilde{\gamma}_1, \delta_1), (\tilde{\gamma}_2, \delta_2) \in \mathcal{B}_{\alpha,\epsilon}$, we have $(\lambda\tilde{\gamma}_1 + \bar{\lambda}\tilde{\gamma}_2, \lambda\delta_1 + \bar{\lambda}\delta_2) \in \bar{\mathcal{B}}_{\alpha,\epsilon}$, where $\bar{\lambda} = 1 - \lambda$, or equivalently, $\lambda\delta_1 + \bar{\lambda}\delta_2 \in [0,1)$ and $\lambda\tilde{\gamma}_1 + \bar{\lambda}\tilde{\gamma}_2 \geq \tilde{\gamma}(\lambda\delta_1 + \bar{\lambda}\delta_2)$. Hence, it suffices to show that $\delta \mapsto \tilde{\gamma}(\delta)$ is convex.

Let $p_i, q_i \in (0,1)$ with $p_i \geq q_i$ be the optimal solution of (6.33) for δ_i, $i = 1, 2$, and $P_{\mathsf{b},i}, Q_{\mathsf{b},i}$ be the corresponding Bernoulli distributions. For any $\lambda \in [0,1]$, we

construct two Bernoulli distribution $P_{\mathsf{b},\lambda}$ and $Q_{\mathsf{b},\lambda}$ with parameters $p_\lambda = \lambda p_1 + \bar{\lambda} p_2$ and $q_\lambda = \lambda q_1 + \bar{\lambda} q_2$, respectively. It can be verified that

$$\begin{aligned}
\mathsf{E}_{e^\epsilon}(P_{\mathsf{b},\lambda} \| Q_{\mathsf{b},\lambda}) =& p_\lambda - e^\epsilon q_\lambda && (6.34)\\
=& \lambda p_1 + \bar{\lambda} p_2 - e^\epsilon(\lambda q_1 + \bar{\lambda} q_2) && (6.35)\\
\geq& \lambda \delta_1 + \bar{\lambda} \delta_2, && (6.36)
\end{aligned}$$

i.e., (p_λ, q_λ) is feasible for $\lambda\delta_1 + \bar{\lambda}\delta_2$. In addition, from the convexity of χ^α, we have that

$$\begin{aligned}
\lambda\tilde{\gamma}(\delta_1) + \bar{\lambda}\tilde{\gamma}(\delta_2) =& \lambda\chi^\alpha(P_{\mathsf{b},1} \| Q_{\mathsf{b},1}) + \bar{\lambda}\chi^\alpha(P_{\mathsf{b},2} \| Q_{\mathsf{b},2}) && (6.37)\\
\geq& \chi^\alpha(P_{\mathsf{b},\lambda} \| Q_{\mathsf{b},\lambda}) && (6.38)\\
\geq& \tilde{\gamma}(\lambda\delta_1 + \bar{\lambda}\delta_2). && (6.39)
\end{aligned}$$

Therefore, the function $\tilde{\gamma}(\delta)$ is convex in δ and hence $\bar{\mathcal{B}}_{\alpha,\epsilon}$ is the convex hull of $\mathcal{B}_{\alpha,\epsilon}$. In light of Theorem 20, this in turn implies that $\mathcal{R}_\alpha = \bar{\mathcal{B}}_{\alpha,\epsilon}$.

The above analysis shows that the mapping $\delta \mapsto \tilde{\gamma}(\delta)$ in fact constitutes the upper boundary of $\mathcal{B}_{\alpha,\epsilon}$ and thus $\mathcal{R}_\alpha$. Since $\chi_\alpha(\cdot)$ is a bijection, this allows us to deduce

$$\begin{aligned}
\gamma(\delta|\alpha,\epsilon) = \inf_{0<p,q<1} & \chi_\alpha^{-1}\left(\chi^\alpha(P_{\mathsf{b}} \| Q_{\mathsf{b}})\right) && (6.40)\\
& \text{s.t. } \mathsf{E}_{e^\epsilon}(P_{\mathsf{b}} \| Q_{\mathsf{b}}) \geq \delta,
\end{aligned}$$

and hence the optimization problem (6.13) can be converted to the above optimization problem with only two parameters.

Expanding both χ^α and E_{e^ϵ}, we can explicitly write (6.40) as

$$\begin{aligned}
\gamma(\delta|\alpha,\epsilon) = \inf_{0<q<p<1} & \frac{1}{\alpha-1}\log\left(p^\alpha q^{1-\alpha} + (1-p)^\alpha(1-q)^{1-\alpha}\right) && (6.41)\\
& \text{s.t. } p - qe^\epsilon \geq \delta,
\end{aligned}$$

where $0 \leq \delta < 1$ and $0 \leq \gamma < \infty$. Let $h(p, q; \alpha)$ indicate the objective function of the optimization problem in (6.41). For any given $\alpha > 1$ and $p \in (0, 1)$, the partial

derivative of $h(p, q; \alpha)$ with respect to q is given by

$$\frac{\partial\, h(p, q; \alpha)}{\partial q} = \frac{p^{\alpha} q^{-\alpha} - (1-p)^{\alpha}(1-q)^{-\alpha}}{p^{\alpha} q^{1-\alpha} + (1-p)^{\alpha}(1-q)^{1-\alpha}}, \tag{6.42}$$

which is negative for all $0 < q < p < 1$, and therefore, $h(p, q; \alpha)$ is decreasing in q. In addition, for $\epsilon \geq 0$ and $\delta \in [0, 1)$, the two constraints $0 < q < p < 1$ and $p - qe^{\epsilon} \geq \delta$ in (6.41) can be equivalently rewritten as

$$\begin{cases} \delta < p < 1 \\ 0 < q < \frac{p-\delta}{e^{\epsilon}}. \end{cases} \tag{6.43}$$

Thus, the infimum in (6.41) is attained at $q = \frac{p-\delta}{e^{\epsilon}}$, and therefore, for $\alpha > 1$, $\delta \in [0, 1)$ and $\epsilon \geq 0$, the optimization problem in (6.41) is simplified as

$$\begin{aligned} e^{(\alpha-1)(\gamma(\delta|\alpha,\epsilon)-\epsilon)} = & \inf_{p} \; p^{\alpha}(p-\delta)^{1-\alpha} + (1-p)^{\alpha}(e^{\epsilon} - p + \delta)^{1-\alpha} \\ & \text{s.t. } \delta < p < 1, \end{aligned} \tag{6.44}$$

which is the desired result. □

6.4.2 Proof of Theorem 22

Recall that the optimization problem in Theorem 21 is equivalent to (6.44). Let $h_1(p; \alpha, \delta, \epsilon)$ indicate the objective function in (6.44). One can verify that for $\alpha > 1, \delta \in [0, 1)$ and $\epsilon > 0$, the mapping $p \mapsto h_1(p; \alpha, \delta, \epsilon)$ is convex. Therefore, the numerical result of $\gamma(\delta|\alpha, \epsilon)$ can be easily obtained for any given α, δ and ϵ. To get closed-form expressions, we explore lower bounds of the objective value in (6.44) as follows.

Lower bound 1: Ignoring the second term in $h_1(p; \alpha, \delta, \epsilon)$, we obtain

$$e^{(\alpha-1)(\gamma(\delta|\alpha,\epsilon)-\epsilon)} \geq \inf_{\delta < p < 1} \; p^{\alpha}(p-\delta)^{1-\alpha} \tag{6.45}$$

We note that the objective function in (6.45) is convex in p. It can be observed via

$$\frac{\partial^2}{\partial p^2} p^{\alpha}(p-\delta)^{1-\alpha} = (\alpha-1)\alpha\left(p^{\frac{\alpha}{2}}(p-\delta)^{\frac{-1-\alpha}{2}} - p^{\frac{\alpha-2}{2}}(p-\delta)^{\frac{1-\alpha}{2}}\right)^2 \geq 0, \tag{6.46}$$

and therefore, by setting the first derivative to be 0, we obtain the optimal solution for the the corresponding unconstrained problem as $p^* = \alpha\delta$. Since $\alpha > 1$, it follows that the optimal solution of (6.45) is given by $p^* = \min\{\alpha\delta, 1\}$, and therefore

$$e^{(\alpha-1)(\gamma(\delta|\alpha,\epsilon)-\epsilon)} \geq \left(\delta\alpha^{\alpha}(\alpha-1)^{1-\alpha}\right)\mathbf{1}\{\alpha\delta < 1\} + \left((1-\delta)^{1-\alpha}\right)\mathbf{1}\{\alpha\delta \geq 1\} \tag{6.47}$$

with equality holds if and only if $\alpha\delta \geq 1$, where $\mathbf{1}\{\cdot\}$ denotes the indicator function. Thus, if $\alpha\delta \geq 1$, we have

$$\gamma(\delta|\alpha,\epsilon) = \epsilon - \log(1-\delta),$$

and if $\alpha\delta < 1$, we have the lower bound

$$\gamma(\delta|\alpha,\epsilon) \geq \epsilon - \frac{1}{\alpha-1}\log\left(\frac{1}{\delta\alpha}\left(1-\frac{1}{\alpha}\right)^{\alpha-1}\right) = \epsilon - \frac{1}{\alpha-1}\log\frac{\zeta_\alpha}{\delta}. \tag{6.48}$$

Lower bound 2: To obtain the second lower bound, we note that the function $h_1(p;\alpha,\delta,\epsilon)$ is convex in δ. This enables us to bound $h_1(p;\alpha,\delta,\epsilon)$ from below by using its linear approximation at $\delta = 0$. Hence we can write

$$h_1(p;\alpha,\delta,\epsilon) \geq h_1(p;\alpha,\delta=0,\epsilon) + \frac{\partial h_1(p;\alpha,\delta=0,\epsilon)}{\partial\delta}\delta \tag{6.49}$$

$$= p + (\alpha-1)\delta + \left(\frac{1-p}{e^{\epsilon}-p}\right)^{\alpha}(e^{\epsilon} - p - (\alpha-1)\delta) \tag{6.50}$$

with equality if and only if $\delta = 0$. Therefore, we have

$$e^{(\alpha-1)(\gamma(\delta|\alpha,\epsilon)-\epsilon)} \geq \inf_{\delta<p<1}\ p + \left(\frac{1-p}{e^{\epsilon}-p}\right)^{\alpha}(e^{\epsilon} - (\alpha-1)\delta - p) + (\alpha-1)\delta. \tag{6.51}$$

Let $h_2(p;\alpha,\delta,\epsilon)$ indicate the objective function of (6.51). In the following, we prove the monotonicity of $h_2(p;\alpha,\delta,\epsilon)$ in p for $\alpha > 1$, $1 > \delta \geq 0$ and $\epsilon \geq 0$. Taking the first

derivative of $h_2(p;\alpha,\delta,\epsilon)$ with respect to p, we have

$$\frac{\partial\, h_2(p;\alpha,\delta,\epsilon)}{\partial\, p} = 1 + \left(\frac{1-p}{e^\epsilon - p}\right)^\alpha \left(\frac{\alpha(e^\epsilon-1)(p+(\alpha-1)\delta - e^\epsilon)}{(e^\epsilon-p)(1-p)} - 1\right) \tag{6.52}$$

$$=: h_3(p;\alpha,\delta,\epsilon)$$

$$\geq 1 + \left(\frac{1-p}{e^\epsilon - p}\right)^\alpha \left(-\frac{\alpha(e^\epsilon-1)}{1-p} - 1\right) =: h_4(p;\alpha,\epsilon) \tag{6.53}$$

$$> h_4(p=\delta;\alpha,\epsilon) \tag{6.54}$$

$$= \frac{(e^\epsilon-\delta)^\alpha - (1-\delta)^\alpha - \alpha(e^\epsilon-1)(1-\delta)^{\alpha-1}}{(e^\epsilon-\delta)^\alpha} \tag{6.55}$$

$$=: \frac{h_5(\delta,\alpha,\epsilon)}{(e^\epsilon-\delta)^\alpha}$$

$$\geq \frac{h_5(\delta,\alpha,\epsilon=0)}{(e^\epsilon-\delta)^\alpha} = 0 \tag{6.56}$$

where

- the inequality in (6.53) is from the fact that the function $h_3(p;\alpha,\delta,\epsilon)$ is increasing in δ, and therefore, for $1 > \delta \geq 0$, $h_3(p;\alpha,\delta,\epsilon) \geq h_3(p;\alpha,\delta=0,\epsilon) = h_4(p;\alpha,\epsilon)$

- the inequality in (6.54) is due to the fact that the function $h_4(p;\alpha,\epsilon)$ is increasing in p as shown below

$$\frac{\partial\, h_4(p;\alpha,\epsilon)}{\partial p} = \alpha(\alpha-1)(e^\epsilon-1)^2(1-p)^{\alpha-2}(e^\epsilon-p)^{-\alpha-1} > 0, \tag{6.57}$$

  and therefore, for $1 > p > \delta$, $h_4(p;\alpha,\epsilon) > h_4(p=\delta;\alpha,\epsilon)$.

- the inequality in (6.56) is from the monotonicity of the function $h_5(\delta,\alpha,\epsilon)$ in ϵ. Specifically,

$$\frac{\partial\, h_5(\delta,\alpha,\epsilon)}{\partial\, \epsilon} = \alpha e^\epsilon \left((e^\epsilon-\delta)^{\alpha-1} - (1-\delta)^{\alpha-1}\right) \geq 0, \tag{6.58}$$

  and therefore, for $\epsilon \geq 0$, $h_5(\delta,\alpha,\epsilon) \geq h_5(\delta,\alpha,\epsilon=0) = 0$.

Therefore, the objective function $h_2(p;\alpha,\delta,\epsilon)$ in (6.51) is increasing in p, and therefore, we have

$$e^{(\alpha-1)(\gamma(\delta|\alpha,\epsilon)-\epsilon)} \geq h_2(p=\delta;\alpha,\delta,\epsilon) = \alpha\delta + \left(\frac{1-\delta}{e^\epsilon-\delta}\right)^\alpha (e^\epsilon - \alpha\delta)\,, \tag{6.59}$$

with equality if and only if $\delta = 0$. Thus, we have

$$\gamma(\delta|\alpha,\epsilon) \geq \epsilon + \frac{1}{\alpha-1}\log\left(\alpha\delta + \left(\frac{1-\delta}{e^\epsilon-\delta}\right)^\alpha (e^\epsilon - \alpha\delta)\right), \tag{6.60}$$

where the equality holds if and only if $\delta = 0$ which leads to $\gamma_\alpha^\epsilon(\delta=0) = 0$. The lower bounds (6.48) and (6.60) give the desired result. □

6.4.3 Proof of Lemma 5

From the function $g(\alpha,\epsilon,\delta)$ in Theorem 22, we have

$$\epsilon(\gamma|\alpha,\delta)\begin{cases} \leq \max\left\{0,\ \gamma - \frac{1}{\alpha-1}\log\frac{\delta}{\zeta_\alpha}\right\}, & \text{if } \alpha\delta \leq 1 \\ = \max\left\{0,\ \gamma + \log(1-\delta)\right\}, & \text{otherwise.}\end{cases} \tag{6.61}$$

Next, we obtain a closed-form upper bound on $\epsilon(\gamma|\alpha,\delta)$ from the function $f(\alpha,\epsilon,\delta)$ in Theorem 22. To do so, let $f_1(\alpha,\epsilon,\delta)$ be the expression inside the logarithm in $f(\alpha,\epsilon,\delta)$, i.e., $f_1(\alpha,\epsilon,\delta) \triangleq (e^\epsilon - \alpha\delta)\left(\frac{\delta-1}{\delta-e^\epsilon}\right)^\alpha + \alpha\delta$. The second partial derivative of $f_1(\delta,\alpha,\epsilon)$ with respect to δ is given by

$$\frac{\partial^2 f_1(\delta,\alpha,\epsilon)}{\partial\,\delta^2} = (\alpha-1)\alpha\,(e^\epsilon-1)\left(\frac{\delta-1}{\delta-e^\epsilon}\right)^\alpha \frac{\left(e^\epsilon\left(-2\delta+e^\epsilon+1\right)-\alpha\delta\left(e^\epsilon-1\right)\right)}{(\delta-1)^2\left(\delta-e^\epsilon\right)^2}. \tag{6.62}$$

Therefore, for $\alpha > 1$, $\epsilon \geq 0$ and $1 \geq \delta \geq 0$, the convexity of $f_1(\delta,\alpha,\epsilon)$ in δ is guaranteed by

$$\delta - \frac{e^\epsilon(e^\epsilon+1)}{2e^\epsilon+\alpha(e^\epsilon-1)} \leq 0. \tag{6.63}$$

Let $f_2(\alpha,\epsilon) \triangleq \frac{e^\epsilon(e^\epsilon+1)}{2e^\epsilon+\alpha(e^\epsilon-1)}$, and therefore, if $\delta - f_2(\alpha,\epsilon) \leq 0$, we have

$$\gamma(\delta|\alpha,\epsilon) \geq f(\alpha,\epsilon,\delta) = \epsilon + \frac{1}{\alpha-1}\log\left(f_1(\alpha,\epsilon,\delta)\right) \tag{6.64}$$

$$\geq \epsilon + \frac{1}{\alpha-1}\log\left(f_1(\alpha,\epsilon,\delta=0) + \frac{\partial\, f_1(\delta=0,\alpha,\epsilon)}{\partial\,\delta}\delta\right) \tag{6.65}$$

$$= \epsilon + \frac{1}{\alpha-1}\log\left(e^{-\epsilon(\alpha-1)} + \alpha\delta - \alpha\delta e^{-\epsilon(\alpha-1)}\right), \tag{6.66}$$

with equality if and only if $\delta = 0$. In the following, we prove that $\delta \leq \frac{1}{\alpha}$ is a sufficient condition for $\delta - f_2(\alpha,\epsilon) \leq 0$ by showing that $f_2(\alpha,\epsilon) > 1/\alpha$ for any $\alpha > 1$. Taking the first partial derivative of $f_2(\alpha,\epsilon)$ with respect to ϵ, we have

$$\frac{\partial\, f_2(\alpha,\epsilon)}{\partial\,\epsilon} = \frac{e^\epsilon((2+\alpha)e^{2\epsilon} - 2\alpha e^{2\epsilon} - \alpha)}{(2e^\epsilon + \alpha(e^\epsilon-1))^2} \tag{6.67}$$

$$\begin{cases} \leq 0, & 1 \leq e^\epsilon \leq \frac{\alpha+\sqrt{2\alpha(\alpha+1)}}{2+\alpha} \\ > 0, & \text{otherwise,} \end{cases} \tag{6.68}$$

and therefore,

$$f_2(\alpha,\epsilon) - \frac{1}{\alpha} \geq f_2\left(\alpha, \epsilon = \log\frac{\alpha+\sqrt{2\alpha(\alpha+1)}}{2+\alpha}\right) - \frac{1}{\alpha} \tag{6.69}$$

$$= \frac{2(\alpha^2 + \alpha(\sqrt{2\alpha(\alpha+1)}-1)-2)}{\alpha(2+\alpha)^2} \triangleq \frac{f_3(\alpha)}{\alpha(2+\alpha)^2} \tag{6.70}$$

$$> \frac{f_3(\alpha=1)}{\alpha(2+\alpha)^2} = 0 \tag{6.71}$$

where the inequality in (6.71) follows from the fact that $f_3(\alpha)$ is monotonically increasing in $\alpha > 1$ as shown below:

$$\frac{\mathrm{d}f_3(\alpha)}{\mathrm{d}\alpha} = \frac{\sqrt{2}\alpha(1+2\alpha)}{\sqrt{\alpha(1+\alpha)}} + 2\sqrt{2\alpha(1+\alpha)} + 4\alpha - 2 > 0. \tag{6.72}$$

Therefore, from the inequality in (6.66), we have that for $\delta \leq 1/\alpha$,

$$\epsilon(\gamma|\alpha,\delta) \leq \frac{1}{\alpha-1}\log\left(\frac{e^{(\alpha-1)\gamma}-1}{\alpha\delta}+1\right)$$

$$= \frac{1}{\alpha-1}\log\left(\frac{(\alpha-1)\chi_\alpha(\gamma)}{\alpha\delta}+1\right)$$

and equality holds if and only if $\gamma = 0$, i.e., $\epsilon_\alpha^\delta(\gamma=0) = 0$. □

6.4.4 Derivation of Remark 6

Note that it can be verified that $\gamma - \frac{1}{\alpha-1}\log\frac{\delta}{\zeta_\alpha} < 0$ for $\delta > \zeta_\alpha e^{(\alpha-1)\gamma}$. Combined with $\alpha\delta \leq 1$, we therefore have $\epsilon_\alpha^\delta(\gamma) = 0$ for $\delta \in [\zeta_\alpha e^{(\alpha-1)\gamma}, \frac{1}{\alpha}]$. To have a valid non-empty interval, we must have the condition $\zeta_\alpha e^{(\alpha-1)\gamma} < \frac{1}{\alpha}$ that is simplified to $1 - e^{-\gamma} \leq \frac{1}{\alpha}$. A similar holds for the case $\alpha\delta > 1$: we have $\gamma + \log(1-\delta) < 0$ if $\delta > 1 - e^{-\gamma}$. Hence, $\epsilon(\gamma|\alpha,\delta) = 0$ if $\delta > \max\{1 - e^{-\gamma}, 1/\alpha\}$. □

6.4.5 Proof of Lemma 6

From the expression of the output $Y = f(D) + N$ with $N \sim \mathcal{N}(0,\sigma^2)$, we have that for any pair of neighboring data sets d and d', the corresponding outputs $Y(d)$ and $Y(d')$ follow the Gaussian distributions $\mathcal{N}(f(d),\sigma^2)$ and $\mathcal{N}(f(d'),\sigma^2)$, respectively. Therefore, the Rényi divergence between the two output distributions is

$$
\begin{aligned}
&\sup_{d\sim d'\in\mathcal{D}} D_\alpha(P_{Y|D=d}\|P_{Y|D=d'}) \\
&= \sup_{d\sim d'\in\mathcal{D}} D_\alpha(\mathcal{N}(f(d),\sigma^2)\|\mathcal{N}(f(d'),\sigma^2)) && (6.73)\\
&= \sup_{d\sim d'\in\mathcal{D}} \frac{1}{\alpha-1}\log\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(y-f(d'))^2}{2\sigma^2}}e^{\alpha\frac{(y-f(d'))^2-(y-f(d))^2}{2\sigma^2}}dy && (6.74)\\
&= \sup_{d\sim d'\in\mathcal{D}} \frac{1}{\alpha-1}\log\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi\sigma^2}}e^{\frac{\alpha(\alpha-1)(f(d)-f(d'))^2}{2\sigma^2}}e^{-\frac{(y-\alpha f(d)+(\alpha-1)f(d'))^2}{2\sigma^2}}dy && (6.75)\\
&= \sup_{d\sim d'\in\mathcal{D}} \frac{1}{\alpha-1}\log\left(e^{\frac{\alpha(\alpha-1)(f(d)-f(d'))^2}{2\sigma^2}}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(y-\alpha f(d)+(\alpha-1)f(d'))^2}{2\sigma^2}}dy\right) && (6.76)\\
&= \sup_{d\sim d'\in\mathcal{D}} \frac{1}{\alpha-1}\log\left(e^{\frac{\alpha(\alpha-1)(f(d)-f(d'))^2}{2\sigma^2}}\cdot 1\right) && (6.77)\\
&= \sup_{d\sim d'\in\mathcal{D}} \frac{\alpha(f(d)-f(d'))^2}{2\sigma^2} = \frac{\alpha}{2\sigma^2} && (6.78)
\end{aligned}
$$

where the last equation is due to the assumption that the L_2-sensitivity of the function f is 1. □

6.4.6 Proof of Lemma 7

Recall that for the T-fold composition of Gaussian mechanism with variance σ^2, we have $\gamma = \alpha\rho T$ where $\rho = 1/\sigma^2$. From Lemma 5, we have that for $\alpha\delta \geq 1$ and $0 < \delta < 1$,

$$\epsilon(\rho\alpha T|\alpha, \delta) = \max\{0,\, \rho\alpha T + \log(1-\delta)\} \tag{6.79}$$

and therefore,

$$\begin{aligned}
\epsilon(\rho, T|\delta) &= \inf_{\alpha>1} \epsilon(\rho\alpha T|\alpha, \delta) && (6.80)\\
&\leq \inf_{\alpha\geq\frac{1}{\delta}} \max\{0,\, \rho\alpha T + \log(1-\delta)\} && (6.81)\\
&= \max\left\{0,\, \frac{\rho T}{\delta} + \log(1-\delta)\right\}; && (6.82)
\end{aligned}$$

In addition, from Lemma 5, we have that for $0 < \alpha\delta < 1$,

$$\epsilon(\rho\alpha T|\alpha, \delta) \leq \min\left\{\left(\alpha\rho T - \frac{1}{\alpha-1}\log\frac{\delta}{\zeta_\alpha}\right)_+, \frac{1}{\alpha-1}\log\left(\frac{(\alpha-1)\chi_\alpha(\alpha\rho T)}{\alpha\delta} + 1\right)\right\}, \tag{6.83}$$

where $\chi_\alpha(\alpha\rho T) = \frac{e^{(\alpha-1)\rho\alpha T}-1}{\alpha-1}$ and $(f(\cdot))_+ = \max\{0,\, f(\cdot)\}$. Therefore,

$$\begin{aligned}
\epsilon(\rho, T|\delta) &= \inf_{\alpha>1} \epsilon(\rho\alpha T|\alpha, \delta) && (6.84)\\
&\leq \inf_{1<\alpha<\frac{1}{\delta}} \min\left\{\left(\alpha\rho T - \frac{1}{\alpha-1}\log\frac{\delta}{\zeta_\alpha}\right)_+, \frac{1}{\alpha-1}\log\left(\frac{e^{\rho\alpha(\alpha-1)T}}{\alpha\delta} + 1\right)\right\}. && (6.85)
\end{aligned}$$

Combining the two inequalities in (6.82) and (6.85), we obtain the upper bound of $\epsilon(\rho, T|\delta)$ in Lemma 7. □

6.4.7 Proof of Theorem 24

Lemma 7 illustrates that the T-fold adaptive homogeneous composition of the Gaussian mechanism with variance σ^2 is (ϵ, δ)-DP where

$$\epsilon = \inf_{1<\alpha\leq\frac{1}{\delta}} \frac{\alpha T}{2\sigma^2} - \frac{1}{\alpha-1}\log\frac{\delta}{\zeta_\alpha}. \tag{6.86}$$

Rearrenging the above, we obtain

$$\sigma^2 = \inf_{1<\alpha\leq\frac{1}{\delta}} \frac{\alpha T}{2\epsilon + \frac{2}{\alpha-1}\log\frac{\delta}{\zeta_\alpha}} \tag{6.87}$$

Assuming that $\frac{2\log\delta^{-1}}{\epsilon} \leq \frac{1}{\delta}$, or equivalently $\epsilon \geq 2\delta\log\delta^{-1}$, then we can plug $\alpha = \frac{2\log\delta^{-1}}{\epsilon}$ in (6.87) to obtain

$$\frac{\alpha T}{2\epsilon + \frac{2}{\alpha-1}\log\alpha\delta - 2\log\left(1-\frac{1}{\alpha}\right)}\bigg|_{\alpha=\frac{2\log\delta^{-1}}{\epsilon}} \tag{6.88}$$

$$= \frac{(\epsilon - 2\log\frac{1}{\delta})T\log\frac{1}{\delta}}{\epsilon^2\left(\epsilon - \log\frac{1}{\delta} + \frac{-\epsilon+2\log\frac{1}{\delta}}{\epsilon}\log\left(\frac{-\epsilon+2\log\frac{1}{\delta}}{2\log\frac{1}{\delta}}\right) - \log\left(\frac{2\log\frac{1}{\delta}}{\epsilon}\right)\right)} \tag{6.89}$$

$$= \frac{2T\log\frac{1}{\delta}}{\epsilon^2} + \frac{T}{\epsilon} - \frac{2T\left(\log\left(2\log\frac{1}{\delta}\right) + 1 - \log\epsilon\right)}{\epsilon^2} + \frac{T}{2\epsilon^2\log\frac{1}{\delta}}\left(4\log^2\left(\frac{\log\frac{1}{\delta^2}}{\epsilon}\right)\right.$$
$$\left. - 6\epsilon\log\left(\frac{\log\frac{1}{\delta^2}}{\epsilon}\right) + 8\log\left(\frac{\log\frac{1}{\delta^2}}{\epsilon}\right) + 2\epsilon^2 - 5\epsilon + 4\right) + O\left(\frac{1}{\log^2\frac{1}{\delta}}\right) \tag{6.90}$$

$$= \frac{2T}{\epsilon^2}\log\frac{1}{\delta} + \frac{T}{\epsilon} - \frac{2T}{\epsilon^2}\left(\log(2\log\delta^{-1}) + 1 - \log\epsilon\right) + O\left(\frac{\log^2(\log\delta^{-1})}{\log\delta^{-1}}\right). \tag{6.91}$$

where

- the expression in (6.88) is from the expression of $\zeta_\alpha = \frac{1}{\alpha}\left(1-\frac{1}{\alpha}\right)^{\alpha-1}$ (defined in Theorem 22) and the condition $\epsilon > 2\delta\log\delta^{-1}$,
- the expression in (6.90) is the Taylor expansion of (6.89) at $\delta = 0$,
- in (6.90) as $\delta \to 0$, we have $\log\delta^{-1} \to \infty$, therefore, for any fixed finite ϵ and T, the fourth term is of order $O\left(\frac{\log^2(\log\delta^{-1})}{\log\delta^{-1}}\right)$ and dominates $O\left(\frac{1}{\log^2\delta^{-1}}\right)$.

It is worth mentioning that the choice of α has already appeared in literature, see e.g., [118, Discussion following Theorem 35]. □

6.5 Concluding Remarks

In this chapter, We propose a framework based on the information-theoretic study of joint range of f-divergences to settle the *optimal* conversion from RDP to (ϵ,δ)-DP,

and thus further enhances the privacy guarantee obtained by the moment accountant approach, i.e., the linear composition theorem of RDP. An approximation of this optimal conversion allows us to derive bounds on the number of DP-SGD iterations for a given DP constraint. Our result improves upon the state-of-the-art [70] by allowing more training iterations (often hundreds more) for the same privacy budget, and thus providing higher utility for free.

Chapter 7

CONCLUSION AND FUTURE WORK

We have introduced a novel family of loss functions, namely α-loss ($\alpha > 0$), to characterize adversarial actions. Making use of the α-loss, we have defined a tunable measure of information leakage called maximal α-leakage, which incorporates mutual information and maximal leakage for $\alpha = 1$ and $\alpha = \infty$, respectively, and is proved to be Arimoto channel capacity for $1 < \alpha < \infty$. Therefore, we have given an operational meaning to these well-known information-theoretical measures via the lens of adversarial losses. By applying α-loss to a generative adversarial machine-learning model, we have demonstrated that fairness and censoring guarantees can be simultaneously provided via a universal representation of data. Thanks to the ability of information-theoretical methods in capturing characteristics of privacy/random mechanisms, we have improved the composition property of (ϵ, δ)-DP, a measure for the worst case of information leakage, for specific privacy mechanisms (e.g., Gaussian mechanism).

There are many problems yet to be addressed. For example, in practice, more specifically, in data-driven machine learning, how could one design feasible privacy/random mechanisms that satisfy constraints on maximal α-leakage? In addition, for the application of α-loss in the centralized setting of machine learning, it is critical to develop methods for efficiently obtaining good values of α to get fairness and censoring guarantees as well as utility guarantees beyond that of log-loss, i.e., setting $\alpha = 1$. In decentralized/federated computation settings, it is valuable to understand how much α-loss can help to improve the learning performances. Finally, for the optimal conversion from RDP to (ϵ, δ)-DP, can we use it to improve the privacy tracking of specific

mechanisms in other scenarios, e.g., in federated learning? It is also interesting to explore the connection between our privacy-tracking method, based on the optimal conversion and the moment accountant, and the approach based on Gaussian-DP recently provided in [121] by Dong at el.?

REFERENCES

[1] H. F. Ladd, "Evidence on discrimination in mortgage lending," *Journal of Economic Perspectives*, vol. 12, no. 2, pp. 41–62, 1998. 1.1

[2] D. Pedreshi, S. Ruggieri, and F. Turini, "Discrimination-aware data mining," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 560–568, ACM, 2008. 1.1

[3] C. Song and V. Shmatikov, "Overlearning reveals sensitive attributes," *arXiv:1905.11742*, 2019. 1.1

[4] W. Diffie and M. Hellman, "New directions in cryptography," *IEEE Transactions on Information Theory*, vol. 22, no. 6, pp. 644–654, 1976. 1.1

[5] T. Elgamal, "A public key cryptosystem and a signature scheme based on discrete logarithms," *IEEE Transactions on Information Theory*, vol. 31, no. 4, pp. 469–472, 1985. 1.1

[6] R. D. Prisco and A. D. Santis, "On the relation of random grid and deterministic visual cryptography," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 4, pp. 653–665, 2014. 1.1

[7] S. Leung-Yan-Cheong and M. Hellman, "The Gaussian wire-tap channel," *IEEE Transactions on Information Theory*, vol. 24, no. 4, pp. 451–456, 1978. 1.1

[8] N. Bhargav, S. L. Cotton, and D. E. Simmons, "Secrecy capacity analysis over κ-μ fading channels: Theory and applications," *IEEE Transactions on Communications*, vol. 64, no. 7, pp. 3011–3024, 2016. 1.1

[9] B. Dai, L. Yu, and Z. Ma, "Relay broadcast channel with confidential messages," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 2, pp. 410–425, 2016. 1.1

[10] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *IEEE Symposium on Security and Privacy*, 2008. 1.1

[11] T. Ristenpart, E. Tromer, H. Shacham, and S. Savage, "Hey, you, get off of my cloud: Exploring information leakage in third-party compute clouds," in *16th ACM Conference on Computer and Communications Security*, pp. 199–212, 2009. 1.1

[12] D. Shah and T. Zaman, "Rumors in a network: Who's the culprit?," *IEEE Transactions on Information Theory*, vol. 57, no. 8, pp. 5163–5181, 2011. 1.1

[13] G. Liang, W. He, C. Xu, L. Chen, and J. Zeng, "Rumor identification in microblogging systems based on users' behavior," *IEEE Transactions on Computational Social Systems*, vol. 2, no. 3, pp. 99–108, 2015. 1.1

[14] A. Ghassami, X. Gong, and N. Kiyavash, "Capacity limit of queueing timing channel in shared FCFS schedulers," in *IEEE International Symposium on Information Theory*, pp. 789–793, 2015. 1.1

[15] A. K. Biswas, "Efficient timing channel protection for hybrid (packet/circuit-switched) network-on-chip," *IEEE Transactions on Parallel and Distributed Systems*, vol. 29, no. 5, pp. 1044–1057, 2018. 1.1

[16] C. Dwork, "Differential privacy," in *Proc. 33rd Intl. Colloq. Automata, Lang., Prog.*, (Venice, Italy), July 2006. 1.1, 3

[17] C. Dwork, "Differential privacy: A survey of results," in *Theory and Applications of Models of Computation: Lecture Notes in Computer Science*, New York:Springer, Apr. 2008. 1.1

[18] S. Asoodeh, M. Diaz, F. Alajaji, and T. Linder, "Privacy-aware guessing efficiency," in *IEEE International Symposium on Information Theory*, pp. 754–758, 2017. 1.1

[19] I. Issa, S. Kamath, and A. B. Wagner, "An operational measure of information leakage," in *Annual Conference on Information Science and Systems*, 2016. 1.1

[20] B. Rassouli and D. Gündüz, "Optimal utility-privacy trade-off with the total variation distance as the privacy measure," in *arXiv:1801.02505v1 [cs.IT]*, 2018. 1.1, 1

[21] I. Mironov, "Rényi differential privacy," in *IEEE 30th Computer Security Foundations Symposium*, 2017. 1.1

[22] S. P. Kasiviswanathan and A. D. Smith, "A note on differential privacy: Defining resistance to arbitrary side information," *arXiv:0803.3946v3*, 2015. 1.1

[23] I. Issa, A. B. Wagner, and S. Kamath, "An operational approach to information leakage," *arXiv:1807.07878*, 2018. 1.1, 2.3, 2.3.1, 2.4.2, 2.6.3

[24] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy and statistical minimax rates," in *IEEE 54th Annual Symposium on Foundations of Computer Science*, 2013. 1.1, 3.3

[25] Q. Geng, P. Kairouz, S. Oh, and P. Viswanath, "The staircase mechanism in differential privacy," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 7, pp. 1176–1184, 2015. 1.1

[26] F. du Pin Calmon and N. Fawaz, "Privacy against statistical inference," in *50th Annual Allerton Conference onCommunication, Control, and Computing*, 2012. 1.1, 1

[27] A. Nageswaran and P. Narayan, "Data privacy for a ρ-recoverable function," *arXiv:1802.07851 [cs.IT]*, 2018. 1.1

[28] S. Asoodeh, F. Alajaji, and T. Linder, "Privacy-aware MMSE estimation," *IEEE International Symposium on Information Theory*, pp. 1989–1993, 2016. 1.1

[29] S. Barocas and A. D. Selbst, "Big data's disparate impact," *Calif. L. Rev.*, vol. 104, p. 671, 2016. 1.1

[30] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ArXiv*, vol. abs/1908.09635, 2019. 1.1

[31] C. Jung, M. Kearns, S. Neel, A. Roth, L. Stapleton, and Z. S. Wu, "Eliciting and enforcing subjective individual fairness," *arXiv:1905.10660*, 2019. 1.1

[32] P. Lahoti, K. P. Gummadi, and G. Weikum, "ifair: Learning individually fair data representations for algorithmic decision making," in *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pp. 1334–1345, IEEE, 2019. 1.1

[33] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 259–268, ACM, 2015. 1.1

[34] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226, ACM, 2012. 1.1

[35] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in neural information processing systems*, pp. 3315–3323, 2016. 1.1, 5.1, 5.1.1, 5.1

[36] B. Fish, J. Kun, and Á. D. Lelkes, "A confidence-based approach for balancing fairness and accuracy," in *Proceedings of the 2016 SIAM International Conference on Data Mining*, pp. 144–152, SIAM, 2016. 1.1

[37] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," *arXiv preprint arXiv:1801.07593*, 2018. 1.1

[38] A. Beutel, J. Chen, T. Doshi, H. Qian, A. Woodruff, C. Luu, P. Kreitmann, J. Bischof, and E. H. Chi, "Putting fairness principles into practice: Challenges, metrics, and improvements," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 453–459, 2019. 1.1

[39] S. Garg, V. Perot, N. Limtiaco, A. Taly, E. H. Chi, and A. Beutel, "Counterfactual fairness in text classification through robustness," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 219–226, 2019. 1.1

[40] A. Agarwal, M. Dudik, and Z. S. Wu, "Fair regression: Quantitative definitions and reduction-based algorithms," in *International Conference on Machine Learning*, pp. 120–129, 2019. 1.1

[41] H. Narasimhan, A. Cotter, M. Gupta, and S. Wang, "Pairwise fairness for ranking and regression," *arXiv preprint arXiv:1906.05330*, 2019. 1.1

[42] D. Lewandowski and U. Spree, "Ranking of wikipedia articles in search engines revisited: Fair ranking for reasonable quality?," *Journal of the American Society for Information Science and technology*, vol. 62, no. 1, pp. 117–132, 2011. 1.1

[43] K. Yang and J. Stoyanovich, "Measuring fairness in ranked outputs," in *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, pp. 1–6, 2017. 1.1

[44] M. Zehlike, F. Bonchi, C. Castillo, S. Hajian, M. Megahed, and R. Baeza-Yates, "Fa*ir: A fair top-k ranking algorithm," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pp. 1569–1578, 2017. 1.1

[45] A. J. Biega, K. P. Gummadi, and G. Weikum, "Equity of attention: Amortizing individual fairness in rankings," in *The 41st international acm sigir conference on research & development in information retrieval*, pp. 405–414, 2018. 1.1

[46] A. Singh and T. Joachims, "Fairness of exposure in rankings," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2219–2228, 2018. 1.1

[47] A. Beutel, J. Chen, T. Doshi, H. Qian, L. Wei, Y. Wu, L. Heldt, Z. Zhao, L. Hong, E. H. Chi, *et al.*, "Fairness in recommendation ranking through pairwise comparisons," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2212–2220, 2019. 1.1

[48] F. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney, "Optimized pre-processing for discrimination prevention," in *Advances in Neural Information Processing Systems*, pp. 3992–4001, 2017. 1.1

[49] S. Hajian, J. Domingo-Ferrer, A. Monreale, D. Pedreschi, and F. Giannotti, "Discrimination-and privacy-aware patterns," *Data Mining and Knowledge Discovery*, vol. 29, no. 6, pp. 1733–1782, 2015. 1.1

[50] D. McNamara, C. S. Ong, and R. C. Williamson, "Provably fair representations," *arXiv:1710.04394*, 2017. 1.1

[51] C. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel, "The variational fair autoencoder," 2016. 1.1

[52] F. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney, "Data pre-processing for discrimination prevention: Information-theoretic optimization and analysis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 5, pp. 1106–1119, 2018. 1.1

[53] A. Ghassami, S. Khodadadian, and N. Kiyavash, "Fairness in supervised learning: An information theoretic approach," *arXiv preprint arXiv:1801.04378*, 2018. 1.1

[54] M. Gutmann and A. Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 297–304, 2010. 1.1

[55] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, pp. 2672–2680, 2014. 1.1

[56] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014. 1.1

[57] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Advances in neural information processing systems*, pp. 2234–2242, 2016. 1.1

[58] A. Odena, "Semi-supervised learning with generative adversarial networks," *arXiv:1606.01583*, 2016. 1.1

[59] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016. 1.1

[60] P. Luc, C. Couprie, S. Chintala, and J. Verbeek, "Semantic segmentation using adversarial networks," *arXiv:1611.08408*, 2016. 1.1

[61] H. Edwards and A. J. Storkey, "Censoring representations with an adversary," in *The 4th International Conference on Learning Representations, ICLR*, 2016. 1.1, 5.3.1

[62] M. Abadi and D. G. Andersen, "Learning to protect communications with adversarial neural cryptography," *arXiv preprint arXiv:1610.06918*, 2016. 1.1

[63] N. Raval, A. Machanavajjhala, and L. P. Cox, "Protecting visual secrets using adversarial nets," in *CVPR Workshop Proceedings*, 2017. 1.1

[64] C. Huang, P. Kairouz, X. Chen, L. Sankar, and R. Rajagopal, "Context-aware generative adversarial privacy," *Entropy*, vol. 19, no. 12, p. 656, 2017. 1.1, 5, 5.2.1, 5.2.1

[65] A. Tripathy, Y. Wang, and P. Ishwar, "Privacy-preserving adversarial networks," *arXiv preprint arXiv:1712.07008*, 2017. 1.1

[66] A. Beutel, J. Chen, Z. Zhao, and E. H. Chi, "Data decisions and theoretical implications when adversarially learning fair representations," *arXiv preprint arXiv:1707.00075*, 2017. 1.1

[67] D. Madras, E. Creager, T. Pitassi, and R. Zemel, "Learning adversarially fair and transferable representations," *arXiv:1802.06309*, 2018. 1.1, 5.3.1, 5.3.1

[68] R. Kohavi, "Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid.," in *Kdd*, vol. 96, pp. 202–207, 1996. 1.1

[69] S. Y. Zhang, Zhifei and H. Qi, "Age progression/regression by conditional adversarial autoencoder," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2017. 1.1

[70] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proc. of CCS*, pp. 308–318, 2016. 1.1, 6, 19, 6.3, 23, 6.3, 6.3.1, 6.3.2, 6.5

[71] R. Shokri and V. Shmatikov, "Privacy-preserving deep learning," in *Proc. of CCS*, pp. 1310–1321, 2015. 1.1, 6.3.2

[72] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate, "Differentially private empirical risk minimization," *Journal of Machine Learning Research*, vol. 12, no. Mar, pp. 1069–1109, 2011. 1.1, 6.3.2

[73] R. Bassily, A. Smith, and A. Thakurta, "Private empirical risk minimization: Efficient algorithms and tight error bounds," in *Proceedings of FOCS*, (Washington, DC, USA), pp. 464–473, IEEE Computer Society, 2014. 1.1, 6.3.2

[74] B. Balle, G. Barthe, and M. Gaboardi, "Privacy amplification by subsampling: Tight analyses via couplings and divergences," in *Proc. 32Nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 6280–6290, 2018. 1.1, 6.3.2

[75] X. Wu, F. Li, A. Kumar, K. Chaudhuri, S. Jha, and J. Naughton, "Bolt-on differential privacy for scalable stochastic gradient descent-based analytics," in *SIGMOD*, pp. 1307–1322, 2017. 1.1

[76] H. B. McMahan, G. Andrew, U. Erlingsson, S. Chien, I. Mironov, and P. Kairouz, "A general approach to adding differential privacy to iterative training procedures," 2018. 1.1, 6.3.2

[77] Google, "Tensorflow privacy," 2018. 1.1

[78] C. Dwork, G. N. Rothblum, and S. Vadhan, "Boosting and differential privacy," in *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pp. 51–60, IEEE, 2010. 1.1, 6.3, 6.3.1

[79] P. Kairouz, S. Oh, and P. Viswanath, "The composition theorem for differential privacy," *IEEE Transactions on Information Theory*, vol. 63, no. 6, pp. 4037–4049, 2017. 1.1, 6.3

[80] I. Mironov, "Rényi differential privacy," in *Proceedings of 30th IEEE Computer Security Foundations Symposium (CSF)*, pp. 263–275, 2017. 1.1, 3, 3.3, 19

[81] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observations," *Studia Sci. Math. Hungar.*, vol. 2, pp. 299–318, 1967. 1.1, 6.1.2

[82] S. M. Ali and S. D. Silvey, "A general class of coefficients of divergence of one distribution from another," *Journal of Royal Statistics*, vol. 28, pp. 131–142, 1966. 1.1, 6.1.2

[83] G. Barthe and F. Olmedo, "Beyond differential privacy: Composition theorems and relational logic for f-divergences between probabilistic programs," in *ICALP*, pp. 49–60, 2013. 1.1, 6, 3

[84] B. Balle and Y.-X. Wang, "Improving the Gaussian mechanism for differential privacy: Analytical calibration and optimal denoising," in *Proceedings of the 35th International Conference on Machine Learning*, vol. 80, pp. 394–403, 10–15 July 2018. 1.1, 6, 3

[85] N. Sharma and N. A. Warsi, "Fundamental bound on the reliability of quantum information transmission," *CoRR*, vol. abs/1302.5281, 2013. 1.1

[86] I. Sason and S. Verdu, "f-divergence inequalities," *IEEE Transactions on Information Theory*, vol. 62, no. 11, pp. 5973–6006, 2016. 1.1

[87] P. Harremoës and I. Vajda, "On pairs of f-divergences and their joint range," *IEEE Transactions on Information Theory*, vol. 57, pp. 3230–3235, June 2011. 1.1, 6, 20

[88] A. Rényi, "On measures of entropy and information," in *4th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 547–561, The Regents of the University of California, 1961. 2.1

[89] T. Van Erven and P. Harremos, "Rényi divergence and kullback-leibler divergence," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 3797–3820, 2014. 2.1

[90] S. Verdú, "α-mutual information," in *IEEE Information Theory and Applications Workshop*, 2015. 2.1.1, 2.1, 2.3, 2.4.1, 2.4.2, 2.5, 2.6.3, 2.6.4, 3.2, 5.4.4, 5.4.5

[91] R. Sibson, "Information radius," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, vol. 14, no. 2, pp. 149–160, 1969. 2.1, 2.1, 2.4.2

[92] S. Arimoto, "Information measures and capacity of order α for discrete memoryless channels," in *Colloquia mathematica Societatis János Bolyai*, (Kestheley, Hungary), pp. 41–52, 1975. 2.1, 2.1, 2.3

[93] S.-W. Ho and S. Verdú, "Convexity/concavity of Rényi entropy and α-mutual information," in *IEEE International Symposium on Information Theory*, 2015. 2.1, 2.4.2, 2.6.4, 2.6.4

[94] X. Nguyen, M. J. Wainwright, and M. I. Jordan, "On surrogate loss functions and f-divergences," *The Annals of Statistics*, vol. 37, no. 2, pp. 876–904, 2009. 2.2

[95] P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe, "Convexity, classification, and risk bounds," *Journal of the American Statistical Association*, vol. 101, no. 473, pp. 138–156, 2006. 2.2

[96] N. Merhav and M. Feder, "Universal prediction," *IEEE Transactions on Information Theory*, vol. 44, pp. 2124–2147, Oct 1998. 2.2

[97] T. A. Courtade and R. D. Wesel, "Multiterminal source coding with an entropy-based distortion measure," in *IEEE International Symposium on Information Theory*, pp. 2040–2044, July 2011. 2.2

[98] T. A. Courtade and T. Weissman, "Multiterminal source coding under logarithmic loss," *IEEE Transactions on Information Theory*, vol. 60, pp. 740–761, Jan 2014. 2.2

[99] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2014. 2.4.1, 2.52, 2.6.4

[100] S. Fehr and S. Berens, "On the conditional Rényi entropy," *IEEE Transactions on Information Theory*, vol. 60, no. 11, pp. 6801–6810, 2014. 2.4.1, 2.5

[101] D. Kifer and B.-R. Lin, "An axiomatic view of statistical privacy and utility," *Journal of Privacy and Confidentiality*, vol. 4, no. 1, pp. 5–49, 2012. 1

[102] Y. Wang, Y. O. Basciftci, and P. Ishwar, "Privacy-utility tradeoffs under constrained data release mechanisms," *arXiv:1710.09295v1 [cs.IT]*, 2017. 1

[103] T. Sypherd, M. Diaz, L. Sankar, and P. Kairouz, "A tunable loss function for binary classification," in *2019 IEEE International Symposium on Information Theory (ISIT)*, pp. 2479–2483, IEEE, 2019. 2.5

[104] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley-Interscience, 2nd ed., 2006. 2.6.3, 2.6.4, 2.6.4, 2.6.7, 3.4.2, 4.5.4, 4.6

[105] J. Liao, L. Sankar, F. P. Calmon, and V. Y. F. Tan, "Hypothesis testing under maximal leakage privacy constraints," in *IEEE International Symposium on Information Theory*, pp. 779–783, 2017. 2.6.4

[106] X. Yang, S. E. Fienberg, and A. Rinaldo, "Differential privacy for protecting multi-dimensional contingency table data: Extensions and applications," *Journal of Privacy and Confidentiality*, vol. 4, no. 1, 2012. 3

[107] I. Csiszar, "Generalized cutoff rates and Rényi's information measures," *IEEE Transactions on Information Theory*, vol. 41, no. 1, pp. 26–34, 1995. 3.2

[108] I. Sason and S. Verdú, "Upper bounds on the relative entropy and rényi divergence as a function of total variation distance for finite alphabets," in *2015 IEEE Information Theory Workshop-Fall (ITW)*, pp. 214–218, IEEE, 2015. 3

[109] F. Liese and I. Vajda, "On divergences and informations in statistics and information theory," *IEEE Transactions on Information Theory*, vol. 52, pp. 4394–4412, Oct 2006. 4.1, 4.5.1

[110] J. Liao, O. Kosut, L. Sankar, and F. P. Calmon, "Privacy under hard distortion constraints," *arXiv preprint arXiv:1806.00063*, 2018. 5.2.1, 5.4.3

[111] S. Asoodeh, M. Diaz, F. Alajaji, and T. Linder, "Estimation efficiency under privacy constraints," *IEEE Transactions on Information Theory*, vol. 65, no. 3, pp. 1512–1534, 2018. 5.2.1

[112] W. E. Lillo, M. H. Loh, S. Hui, and S. H. Zak, "On solving constrained optimization problems with neural networks: A penalty method approach," *IEEE Transactions on neural networks*, vol. 4, no. 6, pp. 931–940, 1993. 5.2.2, 5.2.2

[113] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016. 5.5

[114] B. Balle, G. Barthe, M. Gaboardi, J. Hsu, and T. Sato, "Hypothesis testing interpretations and Rényi differential privacy," *ArXiv*, vol. abs/1905.09982, 2019. 6.2

[115] B. Balle, G. Barthe, M. Gaboardi, and J. Geumlek, "Privacy amplification by mixing and diffusion mechanisms," *ArXiv*, vol. abs/1905.12264, 2019. 6.3.2

[116] N. Papernot, M. Abadi, Úlfar Erlingsson, I. Goodfellow, and K. Talwar, "Semi-supervised knowledge transfer for deep learning from private training data," in *Proceedings of the International Conference on Learning Representations*, 2017. 6.3.2

[117] J. Geumlek, S. Song, and K. Chaudhuri, "Rényi differential privacy mechanisms for posterior sampling," in *Advances in Neural Information Processing Systems 30*, pp. 5289–5298, 2017. 6.3.2

[118] V. Feldman, I. Mironov, K. Talwar, and A. Thakurta, "Privacy amplification by iteration," *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 521–532, 2018. 6.3.2, 6.4.7

[119] Y.-X. Wang, B. Balle, and S. P. Kasiviswanathan, "Subsampled Rényi differential privacy and analytical moments accountant," in *AISTATS*, 2018. 6.3.2

[120] A. Bhowmick, J. Duchi, J. Freudiger, G. Kapoor, and R. Rogers, "Protection against reconstruction and its applications in private federated learning," 2018. 6.3.2

[121] J. Dong, A. Roth, and W. J. Su, "Gaussian differential privacy," *CoRR*, vol. abs/1905.02383, 2019. 7