Information Theory for Trustworthy Machine Learning

A dissertation presented

by

Hao Wang

to

The John A. Paulson School of Engineering and Applied Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Applied Mathematics

Harvard University

Cambridge, Massachusetts

April 2022

Dissertation Advisor: Prof. Flavio P. Calmon Hao Wang

Information Theory for Trustworthy Machine Learning

Abstract

In this thesis, we provide an information-theoretic foundation for trustworthy machine learning (ML). ML algorithms are increasingly used in applications of significant social consequences. It is crucial to ensure that these algorithms are not just accurate but also generalizable, fair, and privacy-preserving. We develop new information-theoretic tools for proving rigorous performance guarantees for practical ML models and establishing protocols to ensure the responsible use of data.

Towards rigorous performance guarantees, we derive generalization bounds for understanding the behaviors of complex ML models (e.g., neural networks). In particular, we study how the interaction between data distributions and optimization methods influences generalization. We consider a family of optimization methods—noisy iterative algorithms—and investigate their generalization capability. We derive distribution-dependent generalization bounds by connecting noisy iterative algorithms with additive noise channels found in information theory. Numerical experiments demonstrate that our results can help understand many empirical observations of neural networks.

Towards responsible use, we characterize a fundamental limit of algorithmic fairness and privacy. Specifically, we provide conditions that ensure fair use of group attributes and analyze the "best" privacy-utility trade-offs among all privacy-preserving mechanisms. These theoretical results, in turn, inspire new algorithms that can repair unfair models or preserve privacy in data sharing. Finally, we evaluate the robustness of information-theoretic privacy measures and establish statistical consistency of "optimal" privacy mechanisms.

Contents

List of Tables

List of Figures

Acknowledgments

Studying abroad and speaking a non-native language is already challenging. Half of my PhD time has been spent during the global pandemic. This thesis would not be possible without the enormous support, love, and kindness of the people around me.

First and foremost, I would like to thank my advisor Prof. Flavio P. Calmon for his unwavering support, guidance, and patience during my PhD study. Flavio introduced me to information theory, teaching me not only beautiful and fascinating theories but also the practical engineering problems behind them. He was always willing to share his in-depth knowledge, insight, and experiences with me, from which I benefited a lot, especially when I started doing research. He was extremely supportive and helped me explore my research interests. I remember the day I decided to pursue a PhD at Harvard and Flavio wrote me that "Grad school is a journey—a time of growth not only academically, but also personally and professionally." It was an absolute honor to be Flavio's student in this incredible journey!

I am very grateful to all the members of my thesis committee: Prof. Salil Vadhan, Prof. Demba Ba, and Prof. Na (Lina) Li. Salil has always been an inspiration for me during my graduate career, and I am very thankful for his invaluable feedback on my research. I would like to thank Demba for his continuous support, encouragement, and constructive feedback on my research since my first day at Harvard. In addition, TFing his Decision Theory course was one of the happiest things I had during the global pandemic. I am grateful to Lina for the inclusive and friendly environments she strives to create. The time flew by, but I still remember the last Thanksgiving before the pandemic we celebrated together—it was one of the warmest moments during my PhD study. In addition, she kindly invited me to present my research in her group meeting, through which I received much valuable feedback.

I would like to thank all my collaborators, especially Mario Diaz, Berk Ustun, and Rui Gao. Working with Mario is a pleasure: I enjoy our discussions on complex math problems, our "adversarial refinement" of writing papers, the foosball games we played, and the dinners/ice creams we had. ¡Muchas gracias! Berk plays a crucial role not just as a research collaborator but also as a mentor in the early stage of my PhD study. I have learned many essential skills from him: modeling a real problem in math, disentangling hard problems into simple solvable pieces, writing clean codes and papers, etc. Berk, I am so grateful for your support and advice throughout these years. It was an absolute pleasure working with Rui over the past few years. Many of the results presented here

were inspired by the discussions with Rui. His brilliant insight and consistent encouragement are so crucial to this thesis. Thank you so much, Rui.

I would like to thank many other faculties I have met during my PhD study. In particular, I would like to thank Prof. Yue Lu, Prof. Adam Smith, and Prof. Joseph K. Blitzstein for offering great classes on Information Processing and Statistical Physics, Adaptive Data Analysis, and Probability Theory, respectively. I would also like to thank Prof. Morgane Austern for sharing her insights and knowledge from a statistical perspective.

I would like to thank the IBM Research Computational Creativity Group, at which I did an internship in the summer of 2019. I am very grateful to my manager Richard Goodwin, mentors Robin Lougee and Richard Segal, and all other group members for their help. They have shared their plentiful experience and immense knowledge with me, which profoundly influences my career. Finally, I would like to thank my friends, Junyu Cao and Xiufan Yu, who I met at IBM and have become my close friends since then.

I would like to thank my labmates: Hsiang Hsu, Haewon Jeong, Carol Long, Lucas W. Monteiro, Wael Alghamdi, Felipe Gomez, Madeleine Barowsky, Javier Zazo, Shahab Asoodeh, Juwendo Denis, Winston Michalak, Lisa Vo, José Cândido S. Santos Filho, Sungmin Cha, etc. Discussing research problems, sharing thoughts on complex problems, and turning coffees into ideas with them were the happiest thing for me each day. Special thanks to Hsiang Hsu and Haewon Jeong. I have met Hsiang in my first year at Harvard and since then we fought through so many classes, deadlines, and finals together. I am very thankful for his support and company. Likewise, I am grateful for Haewon's help and encouragement, especially during my job search. Additionally, thanks for introducing me to so many places in Boston that I have never explored in the past six years.

I thank my friends at Harvard, without whom my PhD journey would be pale and gray: Jingmei Hu, Shang Liu, Diana Zhang, Mengdie Zhao, Muqing Xu, Bolei Deng, Hayoun Oh, Zheng Yang, etc. I would also like to acknowledge Xin Chen, Zhaolin Ren, Yingying Li, Hong Hu, Weiyu Li, Saiqian Zhang, Andrew Song, Bahareh Tolooshams, and many others at SEAS. Although we come from different backgrounds, we share many common research interests and had so many enjoyable discussions. I thank my undergraduate friends from USTC who have walked with me through the PhD journey: Yifeng Qi, Chongyang Bai, and Zhehui Chen. I thank my old friends from middle school and high school who have kept in touch with me for so many years: Jinling Liu, Xiaoming Si, Weiyu Cui, Zhehui Gong, Taoran Huo, and Ruilin Zhu.

Chapter 1

Introduction

For decades, information theory has provided a mathematical theory for the systems that underlie the digital world. The growth of information theory can be traced back to Claude Shannon's "A Mathematical Theory of Communication" [246] in 1948. In this paper, Shannon delineated a methodological blueprint that established the theoretical foundation of communication:

- **Model.** Shannon introduced a simple yet profound communication model: a transmitter encodes a discrete message set into a signal, which is sent through a noisy channel and then decoded by a receiver. Furthermore, he incorporated randomness into both information source and channel—communication was mainly considered as a deterministic signal reconstruction problem before him.

- **Analysis.** Shannon assumed that the probability distributions of the information source and the noisy channel are known and the computational power of the transmitter and receiver is unrestricted. These assumptions enabled rigorous mathematical tools (e.g., random coding) to analyze the model, leading to a fundamental limit of communication.

- **Application.** Despite its simplicity, the analysis not only revealed insights about reliable communication, but also provided a design guideline for new coding schemes.

Since Shannon, this blueprint has been successfully applied to many other areas, including hypothesis testing [62], cryptography [247], statistical mechanics [134], and quantum information theory [243].

Information theory has provided a foundation for systems that can reliably transmit, store, and process massive amounts of data. The total amount of data generated in 2018 is estimated to be 33

zettabytes (1 zettabyte = 10^{21} bytes). This number grows to 59 zettabytes in 2020 and is predicted to reach 175 zettabytes in 2025 [278]. The massive amounts of data, along with the increasing computing power, have fueled the fast development of machine learning (ML) in recent years. ML models and algorithms are now increasingly used in applications of significant social consequence, including hiring [39], loan approval [251], and college admission [50].

The widespread deployment of ML has exposed several new challenges. For example, it has been documented in many applications, ranging from facial recognition [49], to criminal recidivism prediction [10], to translation [40], that ML algorithms can be biased against legally protected groups. Data publishing is increasingly pervasive, yet the disclosure of data may incur a privacy risk through unwanted inferences [258]. Finally, modern ML models (e.g., neural networks) are increasingly complex[1] and, consequently, increasingly opaque. It is crucial to ensure that these complex models not only work well in training but also generalize to unseen data in the future. These challenges show that the standard approach of optimizing for high accuracy is not sufficient when developing ML in applications of individual-level consequence. We must make ML systems worthy of trust.

In this thesis, we apply Shannon's blueprint to build an information-theoretic foundation of trustworthy ML[2]. We focus on three important aspects: (i) generalization of complex ML models; (ii) algorithmic fairness; and (iii) privacy in data publishing. Our ultimate goal is to develop theory that provides rigorous performance guarantees for ML systems and guides the responsible use of data.

Generalization. Classical statistical learning theory attributes the generalization of ML models to the use of a hypothesis class with constrained complexity [269, 273]. However, the traditional wisdom fails to explain many empirical observations in the overparameterized regime (e.g., deep neural networks can often fit the entire training set while generalizing extremely well [307]). Motivated by these issues, many new theories have been developed in order to understand the generalization of complex models. For example, there is a line of recent works that investigate how optimization algorithms drive the learning algorithm to a solution with "good" generalization properties [13, 28, 114]. In contrast, we focus on how data distribution influences the generalization of ML models. By fitting different training data, ML models can exhibit distinct generalization behaviors even if they have the same architecture and are trained by the same optimization method (see Figure 3.1).

[1] State-of-the-art deep neural networks have reached more than 100 billion parameters [46].

[2] We adopt the definition from Varshney [275]: trustworthy ML system is one that has sufficient: basic performance, reliability, human interaction, and aligned purpose (i.e., alignment of the ML system's purpose with a society's wants).

Since providing exact analytical expressions for the generalization gap is challenging in general, we take a step back and derive generalization bounds. Our bounds are distribution-dependent and are highly correlated with the true generalization gap as validated through numerical experiments. We hope that this effort can help us understand the generalization of complex models and inspire new regularization techniques for preventing overfitting.

Fairness. In domains where it is acceptable to fit a model that yields different decisions for individuals based on group attributes (e.g., age, sex), ensuring fair use of the group attributes is crucial. The fairness principles we adopt are non-maleficence (i.e., "do no harm") and beneficence (i.e., "do good") [31]: ML models should avoid the causation of harm and be as accurate as possible on each protected group. Existing works [86, 267, 305] mainly focus on developing new algorithms (or models) to improve model performance by using group attributes. For example, Ustun *et al.* [267] introduced a tree structure to recursively choose group attributes for decoupling classifiers. Dwork *et al.* [86] proposed an algorithm to learn separate models for different groups using transfer learning. In contrast, we ask the question: when is there a performance improvement from incorporating the group attributes into the model? We compare split classifiers (i.e., a set of classifiers trained and deployed separately on each group) with group-blind classifiers (classifiers that do not use group attributes as input). We characterize conditions under which splitting classifiers brings the most performance benefit. These conditions are cast in terms of the difference in probability distributions across different protected groups. Our analysis also reveals a limitation of using group-blind classifiers: they may incur an inherent accuracy trade-off between groups with different probability distributions. Moreover, this trade-off cannot be reconciled by collecting more data or applying different learning algorithms.

Privacy. We consider a one-shot-release model where a user shares a single data point with an analyst to receive some utility. The data point is perturbed by applying a privacy mechanism so that the analyst can reconstruct some useful information while keeping other private information hidden. This is a different setting compared with differential privacy (DP) [87], which aims at answering queries while simultaneously ensuring privacy of individual records in the database. In other words, the one-shot-release model ensures that the user can reliably share their data with a curator while DP concerns with the privacy risk when the curator releases some statistics of the database. Moreover, DP (and local-DP [92]) is robust w.r.t. the data distribution and aims to protect

individuals' entire records. In contrast, we assume that the user's data point contains some private information, represented by a random variable S, and some useful information, represented by X. We consider the strongest adversary in terms of statistical knowledge and computing power: the adversary knows the true distribution of the data, the privacy mechanism being used, the disclosed data, and has unrestricted computing resources. The adversary's objective is to illegitimately infer the private features associated with the disclosed data. Given the probability distribution of (S, X), we characterize a fundamental limit of privacy-utility trade-off and design privacy mechanisms to approach this limit.

Besides the methodological blueprint, information theory has provided powerful tools for mathematical analysis. We apply these tools to address emerging challenges in ML. In Chapter 3, we use strong data processing inequalities [66, 78] and properties of Gaussian channels [115] to analyze ML generalization. In Chapter 4, we leverage two-point methods [45, 171] to provide conditions for fair use of group attributes. In Chapter 5, we delineate a privacy-utility trade-off by using tools from rate distortion theory [69]. In Chapter 6, we investigate the Lipschitz continuity of many information-leakage measures. We hope that this effort opens a new application frontier for information-theoretic tools in ML.

Overview and Main Contributions

We provide more details and elaborate on our contributions to each topic.

Generalization of Noisy Iterative Algorithms

Modern deep neural networks (DNNs) are highly expressive: they can memorize an entire training dataset and still generalize well to unseen data [307]. This empirical observation is not captured by traditional generalization bounds found in statistical learning theory, which attribute the generalization ability to the use of a hypothesis class with constrained complexity [269, 273]. Recent studies demonstrate that different algorithmic choices and data distributions may yield DNNs with contrasting generalization behaviors [27, 120, 205]. In this thesis, we study the generalization properties of a class of optimization methods used for training DNNs, namely noisy iterative algorithms.

Noisy iterative algorithms are used in different practical settings due to their many attractive properties [see e.g., 173, 224, 301, 308]. For example, differentially private SGD (DP-SGD) algorithm

[1, 252], one kind of noisy iterative algorithms, is often used to train ML models while protecting user privacy [87]. Recently, it has been implemented in open-source libraries, including Opacus [93] and TensorFlow Privacy [222]. The additive noise in iterative algorithms may also mitigate overfitting for deep neural networks (DNNs) [200]. From a theoretical perspective, noisy iterative algorithms can escape local minima [159] or saddle points [108] and generalize well [214].

We derive generalization bounds for the noisy iterative algorithms, which can help understand recent empirical observations that are not explained by uniform notions of hypothesis class complexity [269, 273]. For example, a neural network trained using true labels exhibits better generalization ability than a network trained using corrupted labels even when the network architecture is fixed and perfect training accuracy is achieved [307]. Distribution-independent bounds may not be able to capture this phenomenon because they are invariant to both true data and corrupted data. In contrast, our bounds capture this empirical observation, exhibiting a lower value on networks trained on true labels compared to ones trained on corrupted labels (Figure 3.1). Another example is that a wider network often has a more favourable generalization capability [206]. This may seem counter-intuitive at first glance since one may expect that wider networks have a higher VC-dimension and, consequently, would have a higher generalize gap. Our bounds capture this behaviour and are decreasing with respect to the neural network width (Figure 3.2).

This chapter is based on our papers [283, 284, 287]:

- Hao Wang, Yizhe Huang, Rui Gao, and Flavio Calmon. "Analyzing the generalization capability of SGLD using properties of Gaussian channels." In Advances in Neural Information Processing Systems (NeurIPS), 2021.
- Hao Wang, Rui Gao, and Flavio P. Calmon. "Generalization bounds for noisy iterative algorithms using properties of additive noise channels." under review, 2021.
- Hao Wang, Mario Diaz, José Cândido S. Santos Filho, and Flavio P. Calmon. "An information-theoretic view of generalization via Wasserstein distance." In IEEE International Symposium on Information Theory (ISIT), 2019.

Ensuring Fair Use of Group Attributes

A ML model exhibits *disparate treatment* [24] if it treats similar data points from distinct individuals differently based on a group attribute (e.g., age, sex). In applications such as hiring, the existence of

disparate treatment can be illegal [88]. However, in settings such as healthcare, it can be legal and ethical to fit a model which presents disparate treatment[3]. In this case, it is essential to ensure *fair use* of the group attribute (i.e., the model uses the group attribute to produce a tailored performance benefit for each group).

We consider two questions that are central to understanding the above-mentioned principles through the use of group attributes by a ML model:

(i) When does using group attributes bring the most performance benefit?

(ii) How to learn a model while ensuring fair use of group attributes?

To answer the first question, we introduce precise conditions for fair use of group attributes—i.e. conditions under which training a separate model for each group produces the most performance improvement compared with using a group-blind classifier. These conditions are cast in terms of the difference in probability distributions between groups of individuals and are revealed by powerful two-point methods for studying min-max problems. Furthermore, I provide an efficient algorithm that can quickly verify the conditions from data.

To answer the second question, we consider the setting where a black-box classifier exhibits a performance disparity across groups. We aim to repair the model by improving its performance on the disadvantaged group without training a new model. Since the model performance relies on the distribution of input features, is there a hypothetical distribution of input features that minimizes the performance disparity? We refer to this distribution as a counterfactual distribution and design a (functional) gradient descent algorithm for learning this distribution from data. The counterfactual distribution can then be used to repair the model so that it no longer exhibits disparate impact in deployment.

This chapter is based on our publications [285, 286, 288, 289]:

- Hao Wang, Hsiang Hsu, Mario Diaz, and Flavio P. Calmon. "To split or not to split: The impact of disparate treatment in classification." IEEE Transactions on Information Theory (T-IT), 2021.
- Hao Wang, Berk Ustun, and Flavio P. Calmon. "Repairing without retraining: Avoiding disparate impact with counterfactual distributions." In International Conference on Machine Learning (ICML), 2019.

[3]For example, the Equal Credit Opportunity Act (ECOA) permits a creditor to use an applicant's age and income for analyzing credit, as long as such information is used in a fair manner (see 12 CFR §1002.6(b)(2) in [96]).

- Hao Wang, Berk Ustun, and Flavio P. Calmon. "On the direction of discrimination: An information-theoretic analysis of disparate impact in machine learning." In IEEE International Symposium on Information Theory (ISIT), 2018.
- Hao Wang, Hsiang Hsu, Mario Diaz, and Flavio P. Calmon. "The impact of split classifiers on group fairness." In IEEE International Symposium on Information Theory (ISIT), 2021.

An Estimation-Theoretic View of Privacy

Data sharing and publishing is increasingly common within scientific communities [262], businesses [253], government operations [210], medical fields [256], and beyond. Data is usually shared with an application in mind, from which the data provider receives some utility. For example, when a user shares her movie ratings with a streaming service, she receives utility in the form of suggestions of new, interesting movie recommendations that fit her taste. As a second example, when a medical research group shares patient data, their aim is to enable a wider community of researchers and statisticians to learn patterns from that data. Utility is then gained through new scientific discoveries.

The disclosure of non-encrypted data incurs a privacy risk through unwanted inferences. In our previous examples, the streaming service may infer the user's political preference (potentially deemed private by the user) from her movie ratings [199], or an insurance company may determine the identity of a patient within a medical dataset [239, 256]. The dichotomy between privacy and utility has been widely studied by computer scientists, statisticians, and information theorists alike. While specific metrics and models vary among these communities, their desideratum is the same: to design mechanisms that perturb the data (or functions thereof) while achieving an acceptable privacy-utility trade-off (PUT).

Our aim is to characterize the fundamental limits of PUT from an *estimation-theoretic perspective*, and to design privacy mechanisms that provide estimation-theoretic guarantees. Here, an analyst is allowed to reconstruct (in a mean-squared error sense) certain functions of the data (utility), while other private functions should not be reconstructed with distortion below a certain threshold (privacy). We demonstrate how chi-square information captures the fundamental PUT in this case and provide bounds for the best PUT. Then we propose a convex program to compute privacy mechanisms when the functions to be disclosed and hidden are known a priori and the data distribution is known. Finally, we derive lower bounds on the minimum mean-squared error of estimating a target private function from the disclosed data. Here the underlying data distribution is

unknown, but the correlation between the target function and a set of functions which are known to be hard to infer from the disclosed variable is given.

This chapter is based on our publications [281, 290]:

- Hao Wang, Lisa Vo, Flavio P. Calmon, Muriel Médard, Ken R. Duffy, and Mayank Varia. "Privacy with estimation guarantees." IEEE Transactions on Information Theory (T-IT), 2019.
- Hao Wang, and Flavio P. Calmon. "An estimation-theoretic view of privacy." In Annual Allerton Conference on Communication, Control, and Computing (Allerton), 2017.

Robustness of Privacy Measures and Mechanisms

A common approach to ensure privacy consists of perturbing the data using a privacy mechanism: a randomized mapping designed to control private information leakage. Different notions of privacy leakage have been proposed to capture the ability of different adversaries to learn private information, e.g., Shannon's mutual information [229, 240], k-anonymity [257], differential privacy [87], maximal leakage [131], total variation [227], among others. In addition, privacy mechanisms should also guarantee the statistical utility of the disclosed data, usually quantified by some measure of similarity between the original and the perturbed data, e.g., distortion [303], f-divergence [147], minimum mean-squared error [18]. In general, privacy and utility objectives compete with each other, making the design of privacy mechanisms a non-trivial task. When data privacy and statistical utility are measured using information-theoretic quantities (e.g., mutual information), most methods for the analysis and design of privacy mechanisms rely on the implicit assumption that the data distribution is, for the most part, known [e.g., 17, 29, 51, 52, 198, 209, 227, 228, 240, 281, 290]. In practice, the designer has access only to a sample from the true distribution.

In this chapter, we revisit this assumption and study the robustness of information-theoretic privacy measures and mechanisms to partial knowledge of the input distribution. We first derive upper bounds for the difference between the privacy-utility guarantees for the empirical and the true distributions. Our bounds can be applied to a large class of information leakage measures, including probability of correctly guessing, f-information with f locally Lipschitz, Arimoto's mutual information (α-leakage) of order $\alpha > 1$, Sibson's mutual information of order $\alpha > 1$, and maximal α-leakage of order $\alpha > 1$. Then we establish the statistical consistency of optimal privacy mechanisms. Finally, we introduce the notion of uniform privacy mechanisms, which assure privacy for every

distribution within a specific neighborhood of the empirical distribution. Based on the large deviation results, the uniform privacy mechanisms deliver privacy guarantees for the true distributions with a certain probability (depending on the selection of neighborhood).

This chapter is based on our publications [76, 282]:

- Mario Diaz*, Hao Wang*, Flavio P. Calmon, and Lalitha Sankar. "On the robustness of information-theoretic privacy measures and mechanisms." IEEE Transactions on Information Theory (T-IT), 2019. *Equal contribution.
- Hao Wang, Mario Diaz, Flavio P. Calmon, and Lalitha Sankar. "The utility cost of robust privacy guarantees." In IEEE International Symposium on Information Theory (ISIT), 2018.

Chapter 2

Background

We first introduce the notation that is common to all the chapters of this thesis. Then we recall some fundamental properties of f-divergence that underlie many proofs in this thesis.

2.1 Notation

We denote independence of random variables U and V by $U \perp\!\!\!\perp V$, and write $U \sim V$ to indicate that U and V have the same distribution. When U, V, and W form a Markov chain, we write $U \to V \to W$. The minimum mean-squared error (MMSE) of estimating U given V is defined as

$$\mathsf{mmse}(U|V) \triangleq \min_{U \to V \to \hat{U}} \mathbb{E}\left[(U-\hat{U})^2\right] = \mathbb{E}\left[(U-\mathbb{E}\left[U|V\right])^2\right].$$

For any real-valued random variable U, we denote the $\mathcal{L}_p$-norm of U as

$$||U||_p \triangleq (\mathbb{E}\left[|U|^p\right])^{1/p}.$$

The set of all functions that applied to a random variable U with distribution P_U result in an $\mathcal{L}_2$-norm less than or equal to 1 is given by

$$\mathcal{L}_2(P_U) \triangleq \left\{f : \mathcal{U} \to \mathbb{R} \mid \|f(U)\|_2 \leq 1\right\}. \tag{2.1}$$

The conditional expectation operators $T_{V|U} : \mathcal{L}_2(P_V) \to \mathcal{L}_2(P_U)$ and $T_{U|V} : \mathcal{L}_2(P_U) \to \mathcal{L}_2(P_V)$ are given by $(T_{V|U}g)(u) \triangleq \mathbb{E}\left[g(V)|U=u\right]$ and $(T_{U|V}f)(v) \triangleq \mathbb{E}\left[f(U)|V=v\right]$, respectively.

2.2 f-divergence

Csiszár's f-divergence [70] measures the difference between two probability distributions. Here we recall the definition and some fundamental properties of f-divergence.

Definition 1. Let $f : (0, \infty) \to \mathbb{R}$ be a convex function with $f(1) = 0$ and P, Q be two probability distributions over a set $\mathcal{X} \subseteq \mathbb{R}^d$. The f-divergence between P and Q is defined as

$$\mathrm{D}_f(P\|Q) \triangleq \int_{\mathcal{X}} f\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \mathrm{d}Q. \tag{2.2}$$

Examples of f-divergence include KL-divergence ($f(t) = t \log t$), total variation distance ($f(t) = |t-1|/2$), and χ^2-divergence ($f(t) = t^2 - 1$). We provide more examples of f-divergence in Appendix B.1.

The f-divergence motivates a way of measuring the dependence between a pair of random variables (X, Y). Specifically, the f-information between (X, Y) is defined as

$$I_f(X;Y) \triangleq \mathrm{D}_f(P_{X,Y}\|P_X \otimes P_Y) = \mathbb{E}\left[\mathrm{D}_f(P_{Y|X}\|P_Y)\right], \tag{2.3}$$

where $P_{X,Y}$ is the joint distribution, P_X, P_Y are the marginal distributions, $P_{Y|X}$ is the conditional distribution, and the expectation is taken over $X \sim P_X$. In particular, if the KL-divergence is used in (2.3), the corresponding f-information is the well-known mutual information [246].

2.3 Properties of f-divergence

We review some fundamental properties of f-divergence that will be used in this thesis.

Strong data processing inequalities. The data processing inequality states that if a Markov chain $U \to X \to Y$ holds, then

$$I_f(U;Y) \le I_f(U;X). \tag{2.4}$$

In other words, no post-processing of X can increase the information about U. Under certain conditions, the data processing inequality can be sharpened, which leads to a strong data processing inequality [66, 78], often cast in terms of a contraction coefficient. Next, we recall the contraction coefficients of f-divergences and show their connection with strong data processing inequalities.

For a given transition probability kernel $P_{Y|X} : \mathcal{X} \to \mathcal{Y}$, let $P_{Y|X} \circ P$ be the distribution on $\mathcal{Y}$ induced by the push-forward of the distribution P (i.e., the distribution of Y when the distribution of X is P). The contraction coefficient of $P_{Y|X}$ for D_f is defined as

$$\eta_f(P_{Y|X}) \triangleq \sup_{P,Q:P\neq Q} \frac{\mathrm{D}_f(P_{Y|X} \circ P \| P_{Y|X} \circ Q)}{\mathrm{D}_f(P\|Q)} \in [0,1].$$

In particular, when the total variation distance is used, the corresponding contraction coefficient $\eta_{\mathrm{TV}}(P_{Y|X})$ is known as the Dobrushin's coefficient [78], which owns an equivalent expression:

$$\eta_{\mathrm{TV}}(P_{Y|X}) = \sup_{\boldsymbol{x},\boldsymbol{x}'\in\mathcal{X}} \mathrm{D}_{\mathrm{TV}}(P_{Y|X=\boldsymbol{x}} \| P_{Y|X=\boldsymbol{x}'}). \tag{2.5}$$

The Dobrushin's coefficient upper bounds all other contraction coefficients [66]: $\eta_f(P_{Y|X}) \leq \eta_{\mathrm{TV}}(P_{Y|X})$. Furthermore, for any Markov chain $U \to X \to Y$, the contraction coefficients satisfy [see Theorem 5.2 in 223, for a proof]

$$I_f(U;Y) \leq \eta_f(P_{Y|X}) \cdot I_f(U;X). \tag{2.6}$$

When $\eta_f(P_{Y|X}) < 1$, the strict inequality $I_f(U;Y) < I_f(U;X)$ improves the data processing inequality and, hence, is referred to as a strong data processing inequality. We refer the reader to Polyanskiy and Wu [219] and Raginsky [223] for a more comprehensive review on strong data processing inequalities and Calmon et al. [54] for non-linear strong data processing inequalities in Gaussian channels.

Gaussian channels. Consider a pair of random variables (X,Y) related by $Y = X + mN$ where X is lying on $\mathcal{X}$; $m > 0$ is a constant; and $N \sim N(0, \mathbf{I}_d)$ follows a standard Gaussian distribution. This model can be regarded as a single use of a Gaussian channel, which has a long history in information theory and possesses many interesting properties. For example, if $\mathcal{X}$ is a compact set, the contraction coefficients have a non-trivial upper bound

$$\eta_{\mathrm{KL}}(P_{Y|X}) \leq \eta_{\mathrm{TV}}(P_{Y|X}) = 1 - 2\bar{\Phi}\left(\frac{\mathsf{diam}(\mathcal{X})}{2m}\right), \tag{2.7}$$

where $\mathsf{diam}(\mathcal{X}) \triangleq \sup_{\boldsymbol{x},\boldsymbol{x}'\in\mathcal{X}} \|\boldsymbol{x}-\boldsymbol{x}'\|_2$ is the diameter of $\mathcal{X}$ and $\bar{\Phi}(t) \triangleq \int_t^\infty \frac{1}{\sqrt{2\pi}} \exp(-v^2/2)\mathrm{d}v$ is the Gaussian complementary cumulative distribution function (CCDF). Another useful property is the following inequality [see Lemma 3.4.2 in 226, for a proof] which upper bounds the KL-divergence of the output distributions from the Gaussian channel by the Wasserstein distance of their input

distributions. It also serves as a fundamental lemma for proving Otto-Villani's HWI inequality [211] in the Gaussian case.

Lemma 1. *Let X and X' be a pair of random variables which are independent of $N \sim N(0, \mathbf{I}_d)$. Then for any $m > 0$*

$$\mathrm{D}_{\mathrm{KL}}(P_{X+mN} \| P_{X'+mN}) \leq \frac{1}{2m^2} \mathbb{W}_2^2(P_X, P_{X'}). \tag{2.8}$$

Here $\mathbb{W}_2(P_X, P_{X'})$ is the 2-Wasserstein distance equipped with the L_2 cost function:

$$\mathbb{W}_2^2(P_X, P_{X'}) \triangleq \inf \mathbb{E}\left[\|X - X'\|_2^2\right],$$

where the infimum is taken over all couplings (i.e., joint distributions) of the random variables X and X' with marginals P_X and $P_{X'}$, respectively.

We recall an analogous result [115] which is also used in our proof. It gives an upper bound for the input-output mutual information of a Gaussian channel.

Lemma 2. *Let X be a random variable which is independent of $N \sim N(0, \mathbf{I}_d)$. Then for any $m > 0$*

$$I(X + mN; X) \leq \frac{1}{2m^2} \mathsf{Var}(X). \tag{2.9}$$

Converse lower bounds. f-divergence plays an important role in constructing min-max lower bound in non-parametric estimation theory [304]. Below we recall two examples, Le Cam's method [171] and Brown-Low's lower bound [45], which will be used in this thesis.

Consider estimating a functional of interest $\theta(P)$ based on a data point X drawn from the probability distribution P. For an estimator $\hat{\theta} = \hat{\theta}(X)$, we use a loss function $\ell(\hat{\theta}, \theta(P))$ to measure the accuracy. Assume that we know, a prior, that the probability distribution $P \in \mathcal{P}$. Then the following lower bound holds by Le Cam's two-point method.

Lemma 3. *For any two probability distributions $P_1, P_2 \in \mathcal{P}$,*

$$\inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}\left[\ell(\hat{\theta}, \theta(P))\right] \geq \frac{1}{2} \ell(\theta(P_1), \theta(P_2)) \left(1 - \mathrm{D}_{\mathrm{TV}}(P_1 \| P_2)\right).$$

Here a single point X is available for constructing the estimator $\hat{\theta}$. On the other hand, one can extend the above result to product measures if n i.i.d. points are available. In this case, the following

f-divergence inequality is often useful:

$$1 - \mathrm{D}_{\mathrm{TV}}(P_1^{\otimes n} \| P_2^{\otimes n}) \geq \frac{1}{2}(1 - \mathrm{D}_{\mathrm{TV}}(P_1^{\otimes n} \| P_2^{\otimes n})^2) \geq \frac{1}{2}\exp(-\mathrm{D}_{\mathrm{KL}}(P_1^{\otimes n} \| P_2^{\otimes n})) = \frac{1}{2}\exp(-n\mathrm{D}_{\mathrm{KL}}(P_1 \| P_2))$$

where the second inequality is from [44].

Le Cam's two-point method hold for any (pseudo-metric) loss function. Now let us consider a particular loss function–ℓ_2 loss. In this case, Brown and Low established the following lower bound in [45].

Lemma 4. *For any two probability distributions $P_1, P_2 \in \mathcal{P}$,*

$$\inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}\left[(\hat{\theta} - \theta(P))^2\right] \geq \frac{(\theta(P_2) - \theta(P_1))^2}{\left(1 + \sqrt{1 + \mathrm{D}_{\chi^2}(P_2 \| P_1)}\right)^2}. \tag{2.10}$$

The following property of χ^2-divergence is often useful when applying the above lemma to product measures.

$$\mathrm{D}_{\chi^2}(P_2^{\otimes n} \| P_1^{\otimes n}) = (1 + \mathrm{D}_{\chi^2}(P_2 \| P_1))^n - 1 \tag{2.11}$$

Principal inertia components. We present some properties of the principal inertia components (PICs) and show a connection with the χ^2-information (i.e., f-information with $f(t) = t^2 - 1$): the χ^2-information is the sum of all PICs. This connection implies an estimation-theoretic interpretation of the χ^2-information. Note that we adopt the definition of PICs presented in [53], but the PICs predate [53] by many decades [see e.g., 48, 109, 112, 124, 230, 241, 295].

Definition 2 ([53, Definition 1])**.** Let U and V be random variables with support sets $\mathcal{U}$ and $\mathcal{V}$, respectively, and joint distribution $P_{U,V}$. In addition, let $f_0 : \mathcal{U} \to \mathbb{R}$ and $g_0 : \mathcal{V} \to \mathbb{R}$ be the constant functions $f_0(u) = 1$ and $g_0(v) = 1$. For $k \in \mathbb{Z}_+$, we (recursively) define

$$\lambda_k(U;V) \triangleq \mathbb{E}\left[f_k(U)g_k(V)\right]^2, \tag{2.12}$$

where

$$\begin{aligned}(f_k, g_k) \triangleq \operatorname{argmax} \Big\{ \mathbb{E}\left[f(U)g(V)\right]^2 \;\Big|\; & f \in \mathcal{L}_2(P_U), g \in \mathcal{L}_2(P_V), \mathbb{E}\left[f(U)f_j(U)\right] = 0, \\ & \mathbb{E}\left[g(V)g_j(V)\right] = 0, j \in \{0, \ldots, k-1\} \Big\}.\end{aligned} \tag{2.13}$$

The values $\lambda_k(U;V)$ are called the *principal inertia components* (PICs) of $P_{U,V}$. The functions f_k and g_k are called the *principal functions* of $P_{U,V}$.

Observe that the PICs satisfy $\lambda_k(U;V) \leq 1$, since $f_k \in \mathcal{L}_2(P_U)$, $g_k \in \mathcal{L}_2(P_V)$, and

$$|\mathbb{E}\left[f(U)g(V)\right]| \leq \|f(U)\|_2 \|g(V)\|_2 \leq 1.$$

Thus, from Definition 2, $0 \leq \lambda_{k+1}(U;V) \leq \lambda_k(U;V) \leq 1$.

The largest PIC satisfies $\lambda_1(U;V) = \rho_m(U;V)^2$ where $\rho_m(U;V)$ is the maximal correlation [230], defined as

$$\rho_m(U;V) \triangleq \max_{\substack{\mathbb{E}[f(U)]=\mathbb{E}[g(V)]=0 \\ \mathbb{E}\left[f(U)^2\right]=\mathbb{E}\left[g(V)^2\right]=1}} \mathbb{E}\left[f(U)g(V)\right]. \tag{2.14}$$

Definition 3 ([53, Definition 2]). For $\mathcal{U} = [m]$ and $\mathcal{V} = [n]$, let $\mathbf{P}_{U,V} \in \mathbb{R}^{m\times n}$ be a matrix with entries $[\mathbf{P}_{U,V}]_{i,j} = P_{U,V}(i,j)$, and $\mathbf{D}_U \in \mathbb{R}^{m\times m}$ and $\mathbf{D}_V \in \mathbb{R}^{n\times n}$ be diagonal matrices with diagonal entries $[\mathbf{D}_U]_{i,i} = P_U(i)$ and $[\mathbf{D}_V]_{j,j} = P_V(j)$, respectively, where $i \in [m]$ and $j \in [n]$. We define

$$\mathbf{Q}_{U,V} \triangleq \mathbf{D}_U^{-1/2}\mathbf{P}_{U,V}\mathbf{D}_V^{-1/2}. \tag{2.15}$$

We denote the singular value decomposition of $\mathbf{Q}_{U,V}$ by $\mathbf{Q}_{U,V} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$.

Definition 4 ([53, Definition 14]). Let $d \triangleq \min\{|\mathcal{U}|, |\mathcal{V}|\} - 1$, and $\lambda_d(U;V)$ the d-th PIC of $P_{U,V}$. We define

$$\delta(P_{U,V}) \triangleq \begin{cases} \lambda_d(U;V) & \text{if } |\mathcal{V}| \leq |\mathcal{U}|, \\ 0 & \text{otherwise.} \end{cases} \tag{2.16}$$

We also denote $\lambda_d(U;V)$ and the corresponding principal functions f_d, g_d as $\lambda_{\min}(U;V)$ and $f_{\min}$, $g_{\min}$, respectively, when the alphabet size is clear from the context.

The next theorem illustrates some different characterizations of the PICs.

Theorem 1 ([53, Theorem 1]). *The following characterizations of the PICs are equivalent:*

1. *The characterization given in Definition 2, where, for f_k and g_k given in (2.13), $g_k(V) = \frac{\mathbb{E}[f_k(U)|V]}{\|\mathbb{E}[f_k(U)|V]\|_2}$ and $f_k(U) = \frac{\mathbb{E}[g_k(V)|U]}{\|\mathbb{E}[g_k(V)|U]\|_2}$.*
2. *For any $k \in \mathbb{Z}_+$,*

$$1 - \lambda_k(U;V) = \mathsf{mmse}(h_k(U)|V), \tag{2.17}$$

where

$$h_k \triangleq \operatorname{argmin} \left\{ \mathsf{mmse}(h(U)|V) \;\middle|\; \|h(U)\|_2 = 1, \mathbb{E}\left[h(U)h_j(U)\right] = 0, j \in \{0,\ldots,k-1\} \right\}. \quad (2.18)$$

If $\lambda_k(U;V)$ is unique, then $h_k = f_k$, given in (2.13).

3. *$\sqrt{\lambda_k(U;V)}$ is the $(k+1)$-st largest singular value of $\mathbf{Q}_{U,V}$. The principal functions f_k and g_k in (2.13) correspond to the columns of the matrices $\mathbf{D}_U^{-1/2}\mathbf{U}$ and $\mathbf{D}_V^{-1/2}\mathbf{V}$, respectively, where $\mathbf{Q}_{U,V} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$.*

The equivalent characterizations of the PICs in the above theorem have the following intuitive interpretation: the principal functions can be viewed as a basis that decompose the mean-squared error of estimating functions of a hidden variable U given an observation V. In particular, for any zero-mean finite-variance function $f : \mathcal{U} \to \mathbb{R}$,

$$\mathsf{mmse}(f(U)|V) = \sum_{i=1}^{|\mathcal{U}|-1} \mathbb{E}\left[f(U)f_i(U)\right]^2 (1 - \lambda_i(U;V)).$$

Finally, the χ^2-information between U and V is the sum of all PICs [53, 295]: $\chi^2(U;V) = \sum_{i=1}^{d} \lambda_i(U;V)$, where $d = \min\{|\mathcal{U}|, |\mathcal{V}|\} - 1$.

Chapter 3

Generalization of Noisy Iterative Algorithms

Many learning algorithms aim to solve the following (possibly non-convex) optimization problem:

$$\min_{\boldsymbol{w}\in\mathcal{W}} L_\mu(\boldsymbol{w}) \triangleq \mathbb{E}\left[\ell(\boldsymbol{w}, Z)\right] = \int_{\mathcal{Z}} \ell(\boldsymbol{w}, \boldsymbol{z}) \mathrm{d}\mu(\boldsymbol{z}), \tag{3.1}$$

where $\boldsymbol{w} \in \mathcal{W} \subseteq \mathbb{R}^d$ is the model parameter (e.g., weights of a neural network) to optimize; μ is the underlying data distribution that generates Z; and $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}^+$ is the loss function (e.g., 0-1 loss). In the context of supervised learning, Z is often composed by a feature vector X and its corresponding label Y. Since the data distribution μ is unknown, $L_\mu(\boldsymbol{w})$ cannot be computed directly. In practice, a dataset $S \triangleq (Z_1, \cdots, Z_n)$ containing n i.i.d. points $Z_i \sim \mu$ is used to minimize an empirical risk:

$$\min_{\boldsymbol{w}\in\mathcal{W}} L_S(\boldsymbol{w}) \triangleq \frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{w}, Z_i). \tag{3.2}$$

We consider the following (projected) noisy iterative algorithm for solving the empirical risk optimization in (3.2). The parameter $\boldsymbol{w}$ is initialized with a random point $W_0 \in \mathcal{W}$ and updated using the following rule:

$$W_t = \mathsf{Proj}_{\mathcal{W}}\left(W_{t-1} - \eta_t \cdot g(W_{t-1}, \{Z_i\}_{i\in\mathcal{B}_t}) + m_t \cdot N\right), \tag{3.3}$$

where η_t is the learning rate; N is an additive noise drawn independently from a distribution P_N; m_t

is the magnitude of the noise; $\mathcal{B}_t \subseteq [n]$ contains the indices of the data points used at the current iteration and $b_t \triangleq |\mathcal{B}_t|$; g is the direction for updating the parameter (e.g., gradient of the loss function); and

$$g(W_{t-1}, \{Z_i\}_{i\in\mathcal{B}_t}) \triangleq \frac{1}{b_t}\sum_{i\in\mathcal{B}_t} g(W_{t-1}, Z_i). \tag{3.4}$$

At the end of each iteration, the parameter is projected onto the domain $\mathcal{W}$, i.e., $\mathsf{Proj}_{\mathcal{W}}(\boldsymbol{w}) \triangleq \operatorname{argmin}_{\boldsymbol{w}'\in\mathcal{W}} \|\boldsymbol{w}' - \boldsymbol{w}\|$. The recursion in (3.3) is run T iterations and the final output is a random variable W_T.

The goal of this chapter is to provide an upper bound for the *expected generalization gap*:

$$\mathbb{E}\left[L_\mu(W_T) - L_S(W_T)\right], \tag{3.5}$$

where the expectation is taken over the randomness of the training data set S and of the algorithm.

3.1 Overview and Main Contributions

In this chapter, we present three generalization bounds for the noisy iterative algorithms (Section 3.5). These bounds rely on different kinds of f-divergence but are proved in a uniform manner by exploring properties of additive noise channels (Section 3.4). Among them, the KL-divergence bound can deal with sampling with replacement; the total variation bound is often the tightest one; and the χ^2-divergence bound requires the mildest assumption. We apply our results to applications, including DP-SGD, federated learning, and SGLD (Section 3.6). Under these applications, our generalization bounds own a simple form and can be estimated from the training data. Finally, we demonstrate our bounds through numerical experiments (Section 3.7), showing that they can predict the behavior of the true generalization gap.

Our generalization bounds incorporate a time-decaying factor. This decay factor tightens the bounds by enabling the impact of early iterations to reduce with time. Our analysis is motivated by a line of recent works [19, 23, 99] which observed that data points used in the early iterations enjoy stronger differential privacy guarantees than those occurring late. Accordingly, we prove that if a data point is used at an early iteration, its contribution to the generalization gap is decreasing with time due to the cumulative effect of the noise added in the iteration afterward.

The proof techniques of this chapter are based on fundamental tools from information theory. We

first use an information-theoretic framework, proposed by Russo and Zou [238] and Xu and Raginsky [300] and further tightened by Bu et al. [47], for deriving algorithmic generalization bounds. This framework relates the generalization gap in (3.5) with the f-information[1] $I_f(W_T; Z_i)$ between the algorithmic output W_T and each individual data point Z_i. However, estimating this f-information from data is intractable since the underlying distribution is unknown. Given this major challenge, we connect the noisy iterative algorithms with additive noise channels, a fundamental model used in data transmission. As a result, we further upper bound the f-information by a quantity that can be estimated from data by developing new properties of additive noise channels. Moreover, we incorporate a time-decaying factor into our bounds, which enables the contribution of early iterations on our bounds to decrease with time. This factor is established by strong data processing inequalities [66, 78] and has an intuitive interpretation: the dependence between algorithmic output W_T and the data points used in the early iterations is decreasing with time due to external additive noise (i.e., $I_f(W_T; Z_i)$ is decreasing with T for a fixed Z_i).

3.2 Related Works

There are significant recent works which adopt the information-theoretic framework [300] for analyzing the generalization capability of noisy iterative algorithms. Among them, Pensia et al. [214] initially derived a generalization bound in Corollary 1 and their bound was extended in Proposition 3 of Bu et al. [47] for the SGLD algorithm. Although the framework in Pensia et al. [214] can be applied to a broad class of noisy iterative algorithms, their bound in Corollary 1 and Proposition 3 in Bu et al. [47] rely on the Lipschitz constant of the loss function, which makes them independent of the data distribution. Distribution-independent bounds can be potentially loose since the Lipschitz constant may be large and may not capture some empirical observations (e.g., label corruption [307]). Specifically, this Lipschitz constant only relies on the architecture of the network instead of the weight matrices or the data distribution so it is the same for a network trained from corrupted data and a network trained from true data.

To obtain a distribution-dependent bound, Negrea et al. [201] improved the analysis in Pensia et al. [214] by replacing the Lipschitz constant with a gradient prediction residual when analyzing

[1]The f-information (see (2.3) for its definition) includes a family of measures, such as mutual information, which quantify the dependence between two random variables.

the SGLD algorithm. Their follow-up work [118] investigated the Langevin dynamics algorithm (i.e., full batch SGLD), which was later extended by Gálvez et al. [105] to SGLD, and observed a time-decaying phenomenon in their experiments. Specifically, [118] incorporated a quantity, namely the squared error probability of the hypothesis test, into their bound in Theorem 4.2 and this quantity decays with the number of iterations. This seems to suggest that earlier iterations have a larger impact on their generalization bound. In contrast, our decay factor indicates that the impact of earlier iterations is reducing with the total number of iterations. Furthermore, the bound in their Theorem 4.2 requires a bounded loss function while our χ^2-based generalization bound only needs the variance of the loss function to be bounded. More broadly, Neu et al. [204] investigated the generalization properties of SGD. However, the generalization bound in their Proposition 3 suffers from a weaker order $O(1/\sqrt{n})$ when the analysis is applied to the SGLD algorithm.

In addition to the works discussed above, there is a line of papers on deriving SGLD generalization bounds [175, 196]. Among them, Mou et al. [196] introduced two generalization bounds. The first one [Theorem 1 of 196], a stability-based bound, achieves $O(1/n)$ rate in terms of the sample size n but relies on the Lipschitz constant of the loss function which makes it distribution-independent. The second one [Theorem 2 of 196], a PAC-Bayes bound, replaces the Lipschitz constant by an expected-squared gradient norm but suffers from a slower rate $O(1/\sqrt{n})$. In contrast, our SGLD bound in Proposition 4 has order $O(1/n)$ and tightens the expected-squared gradient norm by the variance of gradients. The PAC-Bayes bound in Mou et al. [196] also incorporates an explicit time-decaying factor. However, their analysis seems to heavily rely on the Gaussian noise. In contrast, our generalization bounds include a decay factor for a broad class of noisy iterative algorithms. A follow-up work by Li et al. [175] combined the algorithmic stability approach with PAC-Bayesian theory and presented a bound which achieves order $O(1/n)$. However, their bound requires the scale of the learning rate to be upper bounded by the inverse Lipschitz constant of the loss function. In contrast, we do not need any assumptions on the learning rate.

A standard approach [see e.g., 122] of deriving a generalization bound for the DP-SGD algorithm follows two steps: (i) prove that DP-SGD satisfies the (ϵ, δ)-DP guarantees [19, 23, 99, 252, 296]; (ii) derive/apply a generalization bound that holds for *any* (ϵ, δ)-DP algorithm [30, 84, 145]. However, generalization bounds obtained from this procedure are distribution-independent since DP is robust with respect to the data distribution. In contrast, our bounds in Section 3.6.1 are distribution-dependent. We extend our analysis and derive a generalization bound in the setting of federated

learning in Section 3.6.2. A previous work by Yagli et al. [302] also proved a generalization bound for federated learning in their Theorem 3 but their bound involves a mutual information which could be hard to estimate from data.

3.3 Preliminaries

Notation. A random variable X is σ-sub-Gaussian if $\log \mathbb{E}\left[\exp \lambda(X - \mathbb{E}[X])\right] \leq \sigma^2\lambda^2/2$ for any $\lambda \in \mathbb{R}$. For a random vector $X = (X_1, \cdots, X_d)$, we define its variance and minimum mean absolute error (MMAE) as

$$\mathsf{Var}(X) \triangleq \inf_{\boldsymbol{a}\in\mathbb{R}^d} \mathbb{E}\left[\|X - \boldsymbol{a}\|_2^2\right], \tag{3.6}$$

$$\mathsf{mmae}(X) \triangleq \inf_{\boldsymbol{a}\in\mathbb{R}^d} \mathbb{E}\left[\|X - \boldsymbol{a}\|_1\right]. \tag{3.7}$$

The vector $\boldsymbol{a}$ which minimizes (3.6) and (3.7) are

$$\operatorname*{argmin}_{\boldsymbol{a}\in\mathbb{R}^d} \mathbb{E}\left[\|X - \boldsymbol{a}\|_2^2\right] = (\mathbb{E}[X_1], \cdots, \mathbb{E}[X_d]), \tag{3.8}$$

$$\operatorname*{argmin}_{\boldsymbol{a}\in\mathbb{R}^d} \mathbb{E}\left[\|X - \boldsymbol{a}\|_1\right] = (\mathsf{median}(X_1), \cdots, \mathsf{median}(X_d)), \tag{3.9}$$

where $\mathsf{median}(X_i)$ is the median of the random variable X_i.

Information-theoretic generalization bounds. A recent work by Xu and Raginsky [300] provided a framework for analyzing algorithmic generalization. Specifically, they considered a learning algorithm as a channel (i.e., conditional probability distribution) that takes a training set S as input and outputs a parameter W. Furthermore, they derived an upper bound for the expected generalization gap using the mutual information $I(W;S)$. This bound was later tightened by Bu et al. [47] using an individual sample mutual information. By adapting their proof and leveraging variational representations of f-divergence [207], we present another two generalization bounds based on different kinds of f-information.

Lemma 5. *Consider a learning algorithm which takes a dataset $S = (Z_1, \cdots, Z_n)$ as input and outputs W.*

- *[Proposition 1 in 47] If the loss $\ell(\boldsymbol{w}, Z)$ is σ-sub-Gaussian under $Z \sim \mu$ for all $\boldsymbol{w} \in \mathcal{W}$, then*

$$\left|\mathbb{E}\left[L_\mu(W) - L_S(W)\right]\right| \leq \frac{\sqrt{2}\sigma}{n}\sum_{i=1}^{n}\sqrt{I(W;Z_i)}, \tag{3.10}$$

where $I(W;Z_i)$ is the mutual information (i.e., f-information with $f(t) = t\log t$).

- *If the loss $\ell(\boldsymbol{w}, Z)$ is upper bounded by a constant $A > 0$, then*

$$\left|\mathbb{E}\left[L_\mu(W) - L_S(W)\right]\right| \leq \frac{A}{n}\sum_{i=1}^{n} \mathrm{T}(W;Z_i), \tag{3.11}$$

where $\mathrm{T}(W;Z_i)$ is the T-information (i.e., f-information with $f(t) = |t-1|/2$).

- *If the variance of the loss function is finite (i.e., $\mathsf{Var}\left(\ell(W;Z)\right) < \infty$), then*

$$\left|\mathbb{E}\left[L_\mu(W) - L_S(W)\right]\right| \leq \frac{\sqrt{\mathsf{Var}\left(\ell(W;Z)\right)}}{n}\sum_{i=1}^{n}\sqrt{\chi^2(W;Z_i)}, \tag{3.12}$$

where $\chi^2(W;Z_i)$ is the χ^2-information (i.e., f-information with $f(t) = t^2 - 1$) and Z is a fresh data point which is independent of W (i.e., $(W,Z) \sim P_W \otimes \mu$).

We apply Lemma 5 to analyze the generalization capability of noisy iterative algorithms. Estimating the f-information in Lemma 5 from data is intractable. Hence, we further upper bound these f-information by exploring properties of additive noise channels in the next section. Furthermore, we also incorporate a time-decaying factor into our bound, which is established by strong data processing inequalities, recalled in the upcoming subsection.

Although our analysis is applicable for *any* f-information, we focus on the three f-information in Lemma 5 since:

- Mutual information is often easier to work with due to its many useful properties. For example, the chain rule of mutual information plays an important role for handling sampling with replacement (see Section 3.6.3).

- T-information often yields a tighter bound than (3.10) and (3.12). This can be seen by the following f-divergence inequalities [see Eq. 1 and 94 in 242]:

$$\sqrt{2}\mathrm{T}(W;Z_i) \leq \sqrt{I(W;Z_i)} \leq \sqrt{\log(1+\chi^2(W;Z_i))} \leq \sqrt{\chi^2(W;Z_i)}.$$

  Furthermore, the T-information can be used to analyze a broader class of noisy iterative algorithms. For example, when the additive noise is drawn from a distribution with bounded support, the other two f-information may lead to an infinite generalization bound while the T-information can still give a non-trivial bound (see the last row in Table 3.1).

- χ^2-information requires the mildest assumptions. Apart from bounded loss functions, it is often

hard to verify the sub-Gaussianity of $\ell(\boldsymbol{w}, Z)$ for all $\boldsymbol{w}$. The advantage of (3.12) is that it replaces the sub-Gaussian constant with the variance of the loss function.

Remark 1. Using f-information for bounding generalization gap has appeared in prior literature [see e.g., 4, 7, 91, 106, 138, 143, 283]. More broadly, there are significant recent works [see e.g., 14, 117, 123, 144, 185, 225, 254, 302, 311] on deriving new information-theoretic generalization bounds and applying them to different applications. The reason we adopt Lemma 5 for analyzing noisy iterative algorithms is that it enables us to incorporate a time-decaying factor into our bounds.

3.4 Properties of Additive Noise Channels

Additive noise channels have a long history in information theory. Here we show two important properties of additive noise channels which will be used for deriving the generalization bounds in the next section. The first property (Lemma 6) leads to a decay factor into our bounds. The second property (Lemma 7) produces computable generalization bounds.

Consider a single use of an additive noise channel. Let (X, Y) be a pair of random variables related by $Y = X + mN$ where $X \in \mathcal{X}$; $m > 0$ is a constant; and N represents an independent noise. In other words, the conditional distribution of Y given X can be characterized by $P_{Y|X=\boldsymbol{x}} = P_{\boldsymbol{x}+mN}$. If $\mathcal{X}$ is a compact set, the contraction coefficients often have a non-trivial upper bound, leading to a strong data processing inequality. This is formalized in the following lemma whose proof follows directly from the definition of the Dobrushin's coefficient in (2.5) and the fact that the Dobrushin's coefficient is a universal upper bound of all the contraction coefficients.

Lemma 6. *Let N be a random variable which is independent of (U, X). For a given norm $\|\cdot\|$ on a compact set $\mathcal{X} \subseteq \mathbb{R}^d$ and $m, A > 0$, we define*

$$\delta(A, m) \triangleq \sup_{\|\boldsymbol{x}-\boldsymbol{x}'\| \le A} \mathrm{D}_{\mathrm{TV}}\left(P_{\boldsymbol{x}+mN} \| P_{\boldsymbol{x}'+mN}\right). \tag{3.13}$$

Then the Markov chain $U \to X \to X + mN$ holds and

$$I_f(U; X + mN) \le \delta(\mathsf{diam}(\mathcal{X}), m) \cdot I_f(U; X), \tag{3.14}$$

where $\mathsf{diam}(\mathcal{X}) \triangleq \sup_{\boldsymbol{x},\boldsymbol{x}' \in \mathcal{X}} \|\boldsymbol{x} - \boldsymbol{x}'\|$ is the diameter of $\mathcal{X}$.

Computing f-information in general is intractable when the underlying distribution is unknown. Hence, we further upper bound the f-information in Lemma 5 by a quantity which is easier to

Noise Type	$\mathsf{C}_{\mathrm{KL}}(\boldsymbol{x},\boldsymbol{x}';m)$	$\mathsf{C}_{\chi^2}(\boldsymbol{x},\boldsymbol{x}';m)$	$\mathsf{C}_{\mathrm{TV}}(\boldsymbol{x},\boldsymbol{x}';m)$	$\delta(A,m)$
Gaussian	$\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|_2^2}{2m^2}$	$\exp\left(\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|_2^2}{m^2}\right)-1$	$\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|_2}{2m}$	$1-2\bar{\Phi}\left(\frac{A}{2m}\right)$
Laplace	$\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|_1}{m}$	$\exp\left(\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|_1}{m}\right)-1$	$\sqrt{\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|_1}{2m}}$	$1-\exp\left(-\frac{A}{m}\right)$
Uniform on $[-1,1]$	$\infty\mathbb{I}_{[x\neq x']}$	$\infty\mathbb{I}_{[x\neq x']}$	$\min\left\{1,\left\|\frac{x-x'}{2m}\right\|\right\}$	$\min\left\{1,\frac{A}{2m}\right\}$

Table 3.1: *Closed-form expressions (or upper bounds if in blue color) of $\mathsf{C}_f(\boldsymbol{x},\boldsymbol{x}';m)$ (see (3.15) for its definition) and $\delta(A,m)$ (see (3.13) for its definition). The function $\delta(A,m)$ is equipped with the 2-norm for Gaussian distribution and 1-norm for Laplace distribution. We denote the the Gaussian complementary cumulative distribution function (CCDF) by $\bar{\Phi}(x)\triangleq\int_x^\infty\frac{1}{\sqrt{2\pi}}\exp(-v^2/2)dv$ and define $\infty\cdot 0=0$ as convention. The proof is deferred to Appendix A.1.2.*

compute. To achieve this goal, we introduce another property of additive noise channels. Specifically, let $Y=X+mN$ and $Y'=X'+mN$ be the output variables from the same additive noise channel with input variables X and X', respectively. Then the f-divergence in the output space can be upper bounded by the optimal transport cost in the input space.

Lemma 7. *Let N be a random variable which is independent of (X,X'). For $\boldsymbol{x},\boldsymbol{x}'\in\mathbb{R}^d$ and $m>0$, we define a cost function*

$$\mathsf{C}_f(\boldsymbol{x},\boldsymbol{x}';m)\triangleq\mathrm{D}_f\left(P_{\boldsymbol{x}+mN}\|P_{\boldsymbol{x}'+mN}\right). \tag{3.15}$$

Then for any $m>0$, we have

$$\mathrm{D}_f(P_{X+mN}\|P_{X'+mN})\leq\mathbb{W}(P_X,P_{X'};m). \tag{3.16}$$

Here $\mathbb{W}(P_X,P_{X'};m)$ is the optimal transport cost:

$$\mathbb{W}(P_X,P_{X'};m)\triangleq\inf\mathbb{E}\left[\mathsf{C}_f(X,X';m)\right], \tag{3.17}$$

where the infimum is taken over all couplings (i.e., joint distributions) of the random variables X and X' with marginals P_X and $P_{X'}$, respectively.

Proof. See Appendix A.1.1. □

Lemmas 6 and 7 show that the functions $\delta(A,m)$ and $\mathsf{C}_f(\boldsymbol{x},\boldsymbol{x}';m)$ can be useful for sharpening the data processing inequality and upper bounding the f-information in Lemma 5. We demonstrate in Table 3.1 that these functions can be expressed in closed-form for specific additive noise distributions.

Remark 2. Let N be drawn from a Gaussian distribution. Substituting the closed-form expression of

$\mathsf{C}_{\mathrm{KL}}(\boldsymbol{x}, \boldsymbol{x}'; m)$ from Table 3.1 into Lemma 7 leads to

$$\mathsf{D}_{\mathrm{KL}}(P_{X+mN} \| P_{X'+mN}) \leq \frac{1}{2m^2} \mathbb{W}_2^2(P_X, P_{X'}) \tag{3.18}$$

where $\mathbb{W}_2(P_X, P_{X'})$ is the 2-Wasserstein distance equipped with the L_2 cost function:

$$\mathbb{W}_2^2(P_X, P_{X'}) \triangleq \inf \mathbb{E}\left[\|X - X'\|_2^2\right].$$

This inequality serves as a fundamental building block for proving Otto-Villani's HWI inequality [211] in the Gaussian case [42, 226].

3.5 Generalization Bounds for Noisy Iterative Algorithms

In this section, we present our main result—generalization bounds for noisy iterative algorithms. First, by leveraging strong data processing inequalities, we prove that the amount of information about the data points used in early iterations decays with time. Accordingly, our generalization bounds incorporate a time-decaying factor which enables the impact of early iterations on our bounds to reduce with time. Second, by using properties of additive noise channels developed in the last section, we further upper bound the f-information by a quantity which is often easier to estimate. The above two aspects correspond to Lemma 8 and 9 which are the basis of our main result in Theorem 2.

Before diving into the analysis, we first discuss assumptions made in this chapter.

Assumption 1. *The mini-batch indices $(\mathcal{B}_1, \cdots, \mathcal{B}_T)$ in (3.3) are specified before the algorithm is run and data are drawn without replacement.*

If the mini-batches are selected when the algorithm is run, one can analyze the expected generalization gap by first conditioning on $\mathcal{B} \triangleq (\mathcal{B}_1, \cdots, \mathcal{B}_T)$ and then taking an expectation over the randomness of $\mathcal{B}$:

$$\mathbb{E}\left[L_\mu(W_T) - L_S(W_T)\right] = \mathbb{E}\left[\mathbb{E}\left[L_\mu(W_T) - L_S(W_T) \mid \mathcal{B}\right]\right].$$

Our analysis can be extended to the case where data are drawn with replacement (see Proposition 4) by using the chain rule for mutual information.

Assumption 2. *The parameter domain $\mathcal{W}$ is compact and $\|g(\boldsymbol{w}, \boldsymbol{z})\| \leq K$ for all $\boldsymbol{w}, \boldsymbol{z}$. We denote the diameter of $\mathcal{W}$ by $D \triangleq \sup_{\boldsymbol{w}, \boldsymbol{w}' \in \mathcal{W}} \|\boldsymbol{w} - \boldsymbol{w}'\|$.*

Our generalization bounds rely on the second assumption mildly. In fact, this assumption only affects the time-decaying factor in our bounds which is always upper bounded by 1. If we remove this assumption, our bounds still hold though the decay factor disappears.

Now we are in a position to derive generalization bounds under the above assumptions. As a consequence of strong data processing inequalities, the following lemma indicates that the information of a data point Z_i contained in the algorithmic output W_T will reduce with time T.

Lemma 8. *Under Assumption 1, 2, if a data point Z_i is used in the t-th iteration, then*

$$I_f(W_T; Z_i) \leq I_f(W_t; Z_i) \cdot \prod_{t'=t+1}^{T} \delta(D + 2\eta_{t'} K, m_{t'}), \tag{3.19}$$

where the function $\delta(\cdot, \cdot)$ is defined in (3.13).

Proof. For the t-th iteration, we rewrite the recursion in (3.3) as

$$U_t = W_{t-1} - \eta_t \cdot g(W_{t-1}, \{Z_i\}_{i \in \mathcal{B}_t}) \tag{3.20a}$$

$$V_t = U_t + m_t \cdot N \tag{3.20b}$$

$$W_t = \mathsf{Proj}_{\mathcal{W}}(V_t). \tag{3.20c}$$

Let Z_i be a data point used at the t-th iteration. Under Assumption 1, the following Markov chain holds:

$$Z_i \to U_t \to V_t \to W_t \to \cdots \to W_{T-1} \to U_T \to V_T \to W_T. \tag{3.21}$$

Let $\mathcal{U}_T$ be the range of U_T. By Assumption 2 and the triangle inequality,

$$\mathsf{diam}(\mathcal{U}_T) \leq \mathsf{diam}(\mathcal{W}) + 2\eta_T K = D + 2\eta_T K.$$

Now we leverage the strong data processing inequality in Lemma 6 and obtain

$$\begin{aligned} I_f(W_T; Z_i) &\leq I_f(V_T; Z_i) \\ &\leq \delta(D + 2\eta_T K, m_T) \cdot I_f(U_T; Z_i) \\ &\leq \delta(D + 2\eta_T K, m_T) \cdot I_f(W_{T-1}; Z_i), \end{aligned}$$

where the first and last steps are due to the data processing inequality. Applying this procedure recursively leads to the desired conclusion. □

For many types of noise (e.g., Gaussian or Laplace noise), the function $\delta(\cdot, \cdot)$ is *strictly* smaller

than 1 (see Table 3.1). In this case, the information about the data points used in early iterations is reducing via the multiplicative factor in (3.19). Furthermore, one can even prove that $I_f(W_T; Z_i) \to 0$ as $T \to \infty$ if the magnitude of the additive noise in (3.3) has a lower bound.

Lemma 8 explains how our generalization bounds in Theorem 2 incorporate a time-decaying factor. However, it still involves a f-information $I_f(W_t; Z_i)$, which can be hard to compute from data. Next, we further upper bound this f-information by using properties of additive noise channels developed in the last section (see Lemma 7).

Lemma 9. *Under Assumption 1, if a data point Z_i is used at the t-th iteration, then*

$$I_f(W_t; Z_i) \leq \mathbb{E}\left[\mathsf{C}_f\left(g(W_{t-1}, Z), g(W_{t-1}, \bar{Z}); \frac{m_t b_t}{\eta_t}\right)\right], \tag{3.22}$$

where the function $\mathsf{C}_f(\cdot,\cdot;\cdot)$ is defined in (3.15) and the expectation is taken over $(W_{t-1}, Z, \bar{Z}) \sim P_{W_{t-1}} \otimes \mu \otimes \mu$.

Proof. Recall the definition of U_t, V_t in (3.20). The data processing inequality yields

$$I_f(W_t; Z_i) \leq I_f(V_t; Z_i). \tag{3.23}$$

By the definition of f-information, we can write

$$I_f(V_t; Z_i) = \mathbb{E}\left[\mathrm{D}_f(P_{V_t|Z_i} \| P_{V_t})\right] = \int_{\mathcal{Z}} \mathrm{D}_f(P_{V_t|Z_i=z} \| P_{V_t}) \mathrm{d}\mu(z). \tag{3.24}$$

Since $V_t = U_t + m_t \cdot N$ by its definition, Lemma 7 leads to

$$\mathrm{D}_f\left(P_{V_t|Z_i=z} \| P_{V_t}\right) \leq \mathbb{W}(P_{U_t|Z_i=z}, P_{U_t}; m_t). \tag{3.25}$$

To further upper bound the above optimal transport cost, we construct a special coupling. Let W_{t-1} be the output of the noisy iterative algorithm at the $(t-1)$-st iteration. Then we introduce two random variables:

$$U_z^* \triangleq W_{t-1} - \frac{\eta_t}{b_t}\left(\sum_{j \in \mathcal{B}_t, j \neq i} g(W_{t-1}, Z_j) + g(W_{t-1}, z)\right),$$
$$U^* \triangleq W_{t-1} - \frac{\eta_t}{b_t} \sum_{j \in \mathcal{B}_t} g(W_{t-1}, Z_j).$$

Here U_z^* and U^* have marginals $P_{U_t|Z_i=z}$ and P_{U_t}, respectively. By the definition of optimal transport

cost in (3.17), we have

$$\mathbb{W}\left(P_{U_t|Z_i=z}, P_{U_t}; m_t\right) \le \mathbb{E}\left[\mathsf{C}_f(U_z^*, U^*; m_t)\right]. \tag{3.26}$$

The property of $\mathsf{C}_f(\boldsymbol{x}, \boldsymbol{y}; m)$ in Lemma 23 yields

$$\begin{aligned}\mathbb{E}\left[\mathsf{C}_f(U_z^*, U^*; m_t)\right] &= \mathbb{E}\left[\mathsf{C}_f\left(-\frac{\eta_t}{b_t}g(W_{t-1}, z), -\frac{\eta_t}{b_t}g(W_{t-1}, Z_i); m_t\right)\right] \\ &= \mathbb{E}\left[\mathsf{C}_f\left(\frac{\eta_t}{b_t}g(W_{t-1}, Z_i), \frac{\eta_t}{b_t}g(W_{t-1}, z); m_t\right)\right] \\ &= \mathbb{E}\left[\mathsf{C}_f\left(g(W_{t-1}, Z_i), g(W_{t-1}, z); \frac{m_t b_t}{\eta_t}\right)\right]. \end{aligned} \tag{3.27}$$

Since the data point Z_i is only used at the t-th iteration, it is independent of W_{t-1}. We introduce two independent copies $Z, \bar{Z}$ of Z_i such that $(W_{t-1}, Z, \bar{Z}) \sim P_{W_{t-1}} \otimes \mu \otimes \mu$. Combining (3.24–3.27) and using Tonelli's theorem lead to

$$I_f(V_t; Z_i) \le \mathbb{E}\left[\mathsf{C}_f\left(g(W_{t-1}, Z), g(W_{t-1}, \bar{Z}); \frac{m_t b_t}{\eta_t}\right)\right]. \tag{3.28}$$

Substituting (3.28) into (3.23) gives the desired conclusion. □

With Lemma 5, 8, and 9 in hand, we now present the main result in this section: three generalization bounds for noisy iterative algorithms under different assumptions.

Theorem 2. *Suppose that Assumption 1, 2 hold.*

- *If the loss $\ell(\boldsymbol{w}, Z)$ is σ-sub-Gaussian under $Z \sim \mu$ for all $\boldsymbol{w} \in \mathcal{W}$, the expected generalization gap $\mathbb{E}\left[L_\mu(W_T) - L_S(W_T)\right]$ can be upper bounded by*

$$\frac{\sqrt{2}\sigma}{n}\sum_{t=1}^{T} b_t \sqrt{\mathbb{E}\left[\mathsf{C}_{\mathrm{KL}}\left(g(W_{t-1}, Z), g(W_{t-1}, \bar{Z}); \frac{m_t b_t}{\eta_t}\right)\right] \prod_{t'=t+1}^{T} \delta(D + 2\eta_{t'}K, m_{t'})}. \tag{3.29}$$

- *If the loss function is upper bounded by a constant $A > 0$, the expected generalization gap $\mathbb{E}\left[L_\mu(W_T) - L_S(W_T)\right]$ can be upper bounded by*

$$\frac{A}{n}\sum_{t=1}^{T} b_t \mathbb{E}\left[\mathsf{C}_{\mathrm{TV}}\left(g(W_{t-1}, Z), g(W_{t-1}, \bar{Z}); \frac{m_t b_t}{\eta_t}\right)\right] \prod_{t'=t+1}^{T} \delta(D + 2\eta_{t'}K, m_{t'}). \tag{3.30}$$

- *If the variance of the loss function is finite (i.e., $\mathsf{Var}\left(\ell(W_T; Z)\right) < \infty$), the expected generalization gap*

$\mathbb{E}\left[L_\mu(W_T) - L_S(W_T)\right]$ *can be upper bounded by*

$$\frac{\sigma}{n}\sum_{t=1}^{T} b_t \sqrt{\mathbb{E}\left[\mathsf{C}_{\chi^2}\left(g(W_{t-1}, Z), g(W_{t-1}, \bar{Z}); \frac{m_t b_t}{\eta_t}\right)\right] \prod_{t'=t+1}^{T} \delta(D + 2\eta_{t'}K, m_{t'})}, \tag{3.31}$$

where $\sigma \triangleq \sqrt{\mathsf{Var}\left(\ell(W_T; Z)\right)}$ *with* $(W_T, Z) \sim P_{W_T} \otimes \mu$.

Proof. See Appendix A.2.1. □

Our generalization bounds involve the information (e.g., step size η_t, magnitude of noise m_t, and batch size b_t) at all iterations. Moreover, our bounds are distribution-dependent through the expectation term. Finally, since the function δ is often strictly smaller than 1 (see Table 3.1 for some examples), the multiplicative factor $\prod_{t'=t+1}^{T} \delta(D + 2\eta_{t'}K, m_{t'})$ enables the impact of early iterations on our bounds to reduce with time.

The generalization bounds in Theorem 2 may seem contrived at first glance as they rely on the functions δ and C_f defined in (3.13) and (3.15). However, in the next section, we will show that these bounds can be significantly simplified when we apply them to real applications. Furthermore, we will also compare the advantage of each bound under these applications.

3.6 Applications

We demonstrate the generalization bounds in Theorem 2 through several applications in this section.

3.6.1 Differentially Private Stochastic Gradient Descent (DP-SGD)

Differentially private stochastic gradient descent (DP-SGD) is a variant of SGD where noise is added to a stochastic gradient estimator in order to ensure privacy of each individual record. We recall an implementation of (projected) DP-SGD [see e.g., Algorithm 1 in 99]. At each iteration, the parameter of the empirical risk is updated using the following rule:

$$W_t = \mathsf{Proj}_{\mathcal{W}}\left(W_{t-1} - \eta\left(g(W_{t-1}, \{Z_i\}_{i\in\mathcal{B}_t}) + N\right)\right), \tag{3.32}$$

where N is an additive noise; $\mathcal{B}_t$ contains the indices of the data points used at the current iteration and $b_t \triangleq |\mathcal{B}_t|$; the function g indicates a direction for updating the parameter. The recursion in (3.32) is run for T iterations and we assume that data are drawn without replacement. At the end of each

iteration, the parameter is projected onto a compact domain $\mathcal{W}$. We denote the diameter of $\mathcal{W}$ by D. The output from the DP-SGD algorithm is the last iterate W_T. Finally, we assume that

$$\sup_{\boldsymbol{w}\in\mathcal{W},\boldsymbol{z}\in\mathcal{Z}} \|g(\boldsymbol{w},\boldsymbol{z})\| \leq K. \tag{3.33}$$

This assumption can be satisfied by gradient clipping and is crucial for guaranteeing differential privacy as it controls the sensitivity of each update.

The differential privacy guarantees of the DP-SGD algorithm have been extensively studied in the literature [see e.g., 19, 23, 99, 252, 296]. Here we consider a different angle: the generalization of DP-SGD. We derive generalization bounds for the DP-SGD algorithm under Laplace and Gaussian mechanisms by using our Theorem 2.

Proposition 1 (Laplace mechanism)**.** *Suppose that the additive noise N in (3.32) follows a standard multivariate Laplace distribution. Let $\mathcal{W}$ be equipped with the 1-norm and $q \triangleq 1-\exp\left(-(D+2\eta K)/\eta\right) \in (0,1)$.*

- *If the loss $\ell(\boldsymbol{w},Z)$ is σ-sub-Gaussian under $Z\sim\mu$ for all $\boldsymbol{w}\in\mathcal{W}$, then*

$$\mathbb{E}\left[L_\mu(W_T)-L_S(W_T)\right] \leq \frac{2\sigma}{n}\sum_{t=1}^{T}\sqrt{b_t\cdot \mathsf{mmae}\left(g(W_{t-1},Z)\right)\cdot q^{T-t}}. \tag{3.34}$$

- *If the loss function is upper bounded by $A>0$, then*

$$\mathbb{E}\left[L_\mu(W_T)-L_S(W_T)\right] \leq \frac{\sqrt{2}A}{n}\sum_{t=1}^{T}\sqrt{b_t}\cdot\mathbb{E}\left[\sqrt{\|g(W_{t-1},Z)-\boldsymbol{e}\|_1}\right]\cdot q^{T-t}, \tag{3.35}$$

  where $\boldsymbol{e}\triangleq \mathsf{median}\left(g(W_{t-1},Z)\right)$.

- *If the variance of the loss function is bounded (i.e., $\mathsf{Var}\left(\ell(W_T;Z)\right)<\infty$), then*

$$\mathbb{E}\left[L_\mu(W_T)-L_S(W_T)\right] \leq \frac{\sigma}{n}\sum_{t=1}^{T}\sqrt{b_t\cdot\mathbb{E}\left[\exp\left(2\left\|g(W_{t-1},Z)-\boldsymbol{e}\right\|_1\right)-1\right]\cdot q^{T-t}}, \tag{3.36}$$

  where $\sigma=\sqrt{\mathsf{Var}\left(\ell(W_T;Z)\right)}$ and $\boldsymbol{e}\triangleq\mathsf{median}\left(g(W_{t-1},Z)\right)$.

Proof. See Appendix A.3.1. □

Proposition 2 (Gaussian mechanism)**.** *Suppose that the additive noise N in (3.32) follows a standard multivariate Gaussian distribution. Let $\mathcal{W}$ be equipped with the 2-norm and $q\triangleq 1-2\bar{\Phi}\left((D+2\eta K)/2\eta\right)\in(0,1)$ with $\bar{\Phi}(\cdot)$ being the Gaussian CCDF.*

- *If the loss $\ell(\boldsymbol{w}, Z)$ is σ-sub-Gaussian under $Z \sim \mu$ for all $\boldsymbol{w} \in \mathcal{W}$, then*

$$\mathbb{E}\left[L_\mu(W_T) - L_S(W_T)\right] \leq \frac{2\sigma}{n} \sum_{t=1}^{T} \sqrt{\mathsf{Var}\left(g(W_{t-1}, Z)\right) \cdot q^{T-t}}. \tag{3.37}$$

- *If the loss function is upper bounded by $A > 0$, then*

$$\mathbb{E}\left[L_\mu(W_T) - L_S(W_T)\right] \leq \frac{A}{n} \sum_{t=1}^{T} \mathbb{E}\left[\|g(W_{t-1}, Z) - \boldsymbol{e}\|_2\right] \cdot q^{T-t}, \tag{3.38}$$

  where $\boldsymbol{e} \triangleq \mathbb{E}\left[g(W_{t-1}, Z)\right]$.

- *If the variance of the loss function is bounded (i.e., $\mathsf{Var}\left(\ell(W_T; Z)\right) < \infty$), then*

$$\mathbb{E}\left[L_\mu(W_T) - L_S(W_T)\right] \leq \frac{\sigma}{n} \sum_{t=1}^{T} \sqrt{\mathbb{E}\left[\exp\left(4\left\|g(W_{t-1}, Z) - \boldsymbol{e}\right\|_2^2\right) - 1\right] \cdot q^{T-t}}, \tag{3.39}$$

  where $\sigma = \sqrt{\mathsf{Var}\left(\ell(W_T; Z)\right)}$ and $\boldsymbol{e} \triangleq \mathbb{E}\left[g(W_{t-1}, Z)\right]$.

Proof. See Appendix A.3.1. □

Our Theorem 2 leads to three generalization bounds for each DP-SGD mechanism. We discuss the advantage of each bound in the following remark by focusing on the Gaussian mechanism.

Remark 3. We first assume that the loss function is upper bounded by A, leading to an $A/2$-sub-Gaussian loss $\ell(\boldsymbol{w}, Z)$ and $\sqrt{\mathsf{Var}\left(\ell(W_T; Z)\right)} \leq A/2$. Since

$$\mathbb{E}\left[X\right] \leq \sqrt{\mathbb{E}\left[X^2\right]} \leq \frac{1}{2}\sqrt{\mathbb{E}\left[\exp(4X^2) - 1\right]},$$

then for $\boldsymbol{e} \triangleq \mathbb{E}\left[g(W_{t-1}, Z)\right]$

$$\mathbb{E}\left[\|g(W_{t-1}, Z) - \boldsymbol{e}\|_2\right] \leq \sqrt{\mathsf{Var}\left(g(W_{t-1}, Z)\right)} \leq \frac{1}{2}\sqrt{\mathbb{E}\left[\exp\left(4\left\|g(W_{t-1}, Z) - \boldsymbol{e}\right\|_2^2\right) - 1\right]}.$$

Therefore, we have

$$(3.38) \leq (3.37) \leq (3.39).$$

In other words, the total-variation bound in (3.30) yields the tightest generalization bound (3.38) for the DP-SGD algorithm. On the other hand, the χ^2-divergence bound in (3.31) leads to a bound (3.39) that requires the mildest assumption. At this moment, it seems unclear what the advantage of the KL-divergence bound is. Nonetheless, we will show in Section 3.6.3 that the nice properties of mutual information (e.g., chain rule) help extend our analysis to the general setting where data are drawn with replacement.

A standard approach [see e.g., 122] for analyzing the generalization of the DP-SGD algorithm often follows two steps: establish (ϵ, δ)-differential privacy guarantees for the DP-SGD algorithm and prove/apply a generalization bound that holds for *any* (ϵ, δ)-differentially private algorithms. However, generalization bounds obtained in this manner are distribution-independent since differential privacy is robust with respect to the data distribution. As observed in existing literature [see e.g., 307] and our Figure 3.1, machine learning models trained under different data distributions can exhibit completely different generalization behaviors. Our bounds take into account the data distribution through the expectation terms (or mmae, variance).

Our generalization bounds can be estimated from data. Take the bound in (3.37) as an example. If sufficient data are available at each iteration, we can estimate the variance term by the population variance of $\{g(W_{t-1}, Z_i) \mid i \in \mathcal{B}_t\}$ since W_{t-1} is independent of Z_i for $i \in \mathcal{B}_t$. Alternatively, we can draw a hold-out set for estimating the variance term at each iteration.

3.6.2 Federated Learning (FL)

Federated learning (FL) [193] is a setting where a model is trained across multiple clients (e.g., mobile devices) under the management of a central server while the training data are kept decentralized. We recall the federated averaging algorithm with local-update DP-SGD in Algorithm 1 and refer the readers to Kairouz et al. [146] for a more comprehensive review.

Algorithm 1 Federated averaging (local DP-SGD).

Input:
Total number of clients N and clients per round C
Total global updates T and local updates M
DP-SGD learning rate η
Initialize: W_0 randomly selected from $\mathcal{W}$
for $t = 1, \cdots, T$ global steps **do**
  Server chooses a subset $\mathcal{S}_t$ of C clients
  Server sends W_{t-1} to all selected clients
  for each client $k \in \mathcal{S}_t$ in parallel **do**
    Initialize $W_{t,0}^k \leftarrow W_{t-1}$
    for $j = 1, \cdots, M$ local steps **do**
      Draw b fresh data points $\{Z_i^k\}_{i \in [b]}$ and noise $N \sim N(0, \mathbf{I}_d)$
      Update the parameter $W_{t,j}^k \leftarrow \mathsf{Proj}_{\mathcal{W}} \left(W_{t,j-1}^k - \eta \left(g \left(W_{t,j-1}^k, \{Z_i^k\}_{i \in [b]} \right) + N \right) \right)$
    end for
    Send $W_{t,M}^k$ back to the server
  end for
  Server aggregates the parameter $W_t = \frac{1}{C} \sum_{k \in \mathcal{S}_t} W_{t,M}^k$
end for
Output: W_T

It is crucial to be able to *monitor* the performance of the global model on each client. Although the global model could achieve a desirable performance on average, it may fail to achieve high accuracy for each local client. This is because in the federated learning setting, data are typically unbalanced (different clients own different number of samples) and not identically distributed (data distribution varies across different clients). Since in practice clients may not have an extra hold-out dataset to evaluate the performance of the global model, they can instead compute the loss of the model on their training set and compensate the mismatch by the generalization gap (or its upper bound). It is worth noting that this approach of monitoring model performance is completely decentralized as the clients do not need to share their data with the server and all the computation can be done locally. As discussed in Remark 3, the total variation bound in (3.30) often leads to the tightest generalization bound so we recast it under the setting of FL.

Proposition 3. *Let $\mathcal{T}_k \subset [T]$ contain the indices of global iterations in which the k-th client interacts with the server. If the loss function is upper bounded by $A > 0$, the expected generalization gap of the k-th client has an upper bound:*

$$\mathbb{E}\left[L_{\mu_k}(W_T) - L_{S_k}(W_T)\right] \leq \frac{A}{n_k} \sum_{t \in \mathcal{T}_k} \sum_{j=1}^{M} \mathbb{E}\left[\|g(W_{t,j-1}^k, Z^k) - \boldsymbol{e}\|_2\right] \cdot q^{M(T+1-t)-j},$$

where n_k is the number of training data from the k-th client, $\boldsymbol{e} \triangleq \mathbb{E}\left[g(W_{t,j-1}^k, Z^k)\right]$, and

$$q \triangleq 1 - 2\bar{\Phi}\left(\frac{\sqrt{C}(D + 2\eta K)}{2\eta}\right) \in (0, 1)$$

with D being the diameter of $\mathcal{W}$, $K \triangleq \sup_{\boldsymbol{w},\boldsymbol{z}} \|g(\boldsymbol{w}, \boldsymbol{z})\|_2$, and $\bar{\Phi}(\cdot)$ being the Gaussian CCDF.

Proof. See Appendix A.3.2. □

Yagli et al. [302] introduced a generalization bound in the context of federated learning. However, their bound in Theorem 3 involves a mutual information. Here we replace the mutual information with an expectation term. This improvement allows local clients to compute our bound from their training data more reliably.

3.6.3 Stochastic Gradient Langevin Dynamics (SGLD)

We analyze the generalization gap of the stochastic gradient Langevin dynamics (SGLD) algorithm [110, 293]. We start by recalling a standard framework of SGLD. The dataset S is first divided into m

disjoint mini-batches:

$$S = \bigcup_{j=1}^{m} S_j, \quad \text{where } |S_j| = b \text{ and } S_j \cap S_k = \emptyset \text{ for } j \neq k.$$

We initialize the parameter of the empirical risk with a random point $W_0 \in \mathcal{W}$ and update using the following rule:

$$W_t = W_{t-1} - \eta_t \nabla_{\boldsymbol{w}} \hat{\ell}(W_{t-1}, S_{B_t}) + \sqrt{\frac{2\eta_t}{\beta_t}} N, \tag{3.40}$$

where η_t is the learning rate; β_t is the inverse temperature; N is drawn independently from a standard Gaussian distribution; $B_t \in [m]$ is the mini-batch index; $\hat{\ell}$ is a surrogate loss (e.g., hinge loss); and

$$\nabla_{\boldsymbol{w}} \hat{\ell}(W_{t-1}, S_{B_t}) \triangleq \frac{1}{b} \sum_{Z \in S_{B_t}} \nabla_{\boldsymbol{w}} \hat{\ell}(W_{t-1}, Z). \tag{3.41}$$

We study a general setting where the output from SGLD can be any function of the parameters across all iterations (i.e., $W = f(W_1, \cdots, W_T)$), including the setting considered before where $W = W_T$. For example, the output can be an average of all iterates (i.e., Polyak averaging) $W = \frac{1}{T}\sum_t W_t$ or the parameter which achieves the smallest value of the loss function $W = \text{argmin}_{W_t} L_\mu(W_t)$.

Alas, Theorem 2 cannot be applied directly to the SGLD algorithm because the Markov chain in (3.21) does not hold any more when data are drawn with replacement. In order to circumvent this issue, we develop a different proof technique by using the chain rule for mutual information.

Proposition 4. *If the loss function $\ell(\boldsymbol{w}, Z)$ is σ-sub-Gaussian under $Z \sim \mu$ for all $\boldsymbol{w} \in \mathcal{W}$, then*

$$\mathbb{E}\left[L_\mu(W) - L_S(W)\right] \leq \frac{\sqrt{2b}\sigma}{2n} \sum_{j=1}^{m} \sqrt{\sum_{t \in \mathcal{T}_j} \beta_t \eta_t \cdot \mathsf{Var}\left(\nabla_{\boldsymbol{w}} \hat{\ell}(W_{t-1}, S_j)\right)},$$

where the set $\mathcal{T}_j$ contains the indices of iterations in which the mini-batch S_j is used.

Proof. See Appendix A.3.3. □

Our bound incorporates the gradient variance which measures a particular kind of "flatness" of the loss landscape. We note that a recent work [137] has observed empirically that the variance of gradients is predictive of and highly correlated with the generalization gap of neural networks. Here we evidence this connection from a theoretical viewpoint by incorporating the gradient variance into the generalization bound.

Unfortunately, our generalization bound does not incorporate a decay factor[2] anymore. To understand why it happens, let us imagine an extreme scenario in which the SGLD algorithm outputs all the iterates (i.e., $W = (W_1, \cdots, W_T)$). For a data point Z_i used at the t-th iteration, the data processing inequality implies that

$$I(W_1, \cdots, W_T; Z_i) \geq I(W_t; Z_i).$$

Hence, it is impossible to have $I(W_1, \cdots, W_T; Z_i) \to 0$ as $T \to \infty$ unless $I(W_t; Z_i) = 0$.

Many existing SGLD generalization bounds [e.g., 175, 196, 201, 214] are expressed as a sum of errors associated with each training iteration. In order to compare with these results, we present an analogous bound in the following corollary. This bound is obtained by combining a key lemma for proving Proposition 4 with Minkowski inequality and Jensen's inequality so it is often much weaker than Proposition 4.

Corollary 1. *If the loss function $\ell(\boldsymbol{w}, Z)$ is σ-sub-Gaussian under $Z \sim \mu$ for all $\boldsymbol{w} \in \mathcal{W}$, the expected generalization gap of the SGLD algorithm can be upper bounded by*

$$\frac{\sqrt{2}\sigma}{2} \min \left\{ \frac{1}{n} \sum_{t=1}^{T} \sqrt{\beta_t \eta_t \cdot \mathsf{Var}\left(\nabla_{\boldsymbol{w}} \hat{\ell}(W_{t-1}, Z_t^\dagger) \right)}, \sqrt{\frac{1}{bn} \sum_{t=1}^{T} \beta_t \eta_t \cdot \mathsf{Var}\left(\nabla_{\boldsymbol{w}} \hat{\ell}(W_{t-1}, Z_t^\dagger) \right)} \right\},$$

where $Z_t^\dagger$ is any data point used in the t-th iteration.

Proof. See Appendix A.3.4. □

Our bound is distribution-dependent through the variance of gradients in contrast with Corollary 1 of Pensia et al. [214], Proposition 3 of Bu et al. [47], and Theorem 1 of Mou et al. [196], which rely on the Lipschitz constant: $\sup_{\boldsymbol{w}, \boldsymbol{z}} \|\nabla_{\boldsymbol{w}} \hat{\ell}(\boldsymbol{w}, \boldsymbol{z})\|_2$. These bounds fail to explain some generalization phenomena of DNNs, such as label corruption [307], because the Lipschitz constant takes a supremum over all possible weight matrices $\boldsymbol{w}$ and data points $\boldsymbol{z}$. In other words, this Lipschitz constant only relies on the architecture of the network instead of the weight matrices or data distribution. Hence, it is the same for a network trained from corrupted data and a network trained from true data. We remark that the Lipschitz constant used by Bu et al. [47], Mou et al. [196], Pensia et al. [214] is different from the Lipschitz constant of the function corresponding to a network w.r.t. the input variable. The latter one has been used in the literature [see e.g., 27] for

[2] We note that the analysis in Mou et al. [196] requires $W = W_T$. Hence, in the setting we consider (i.e., W is a function of $W_1, \cdots, W_T$), it is unclear if it is possible to include a decay factor in the bound.

deriving generalization bounds and, to some degree, can capture generalization phenomena, such as label corruption.

The order of our generalization bound in Corollary 1 is $\min\left(\frac{1}{n}\sum_{t=1}^{T}\sqrt{\beta\eta_t}, \sqrt{\frac{\beta}{bn}\sum_{t=1}^{T}\eta_t}\right)$. It is tighter than Theorem 2 of Mou et al. [196] whose order is $\sqrt{\frac{\beta}{n}\sum_{t=1}^{T}\eta_t}$. Our bound is applicable regardless of the choice of learning rate while the bound in Li et al. [175] requires the scale of the learning rate to be upper bounded by the reciprocal of the Lipschitz constant. Our Corollary 1 has the same order with Negrea et al. [201] but we incorporate an additional decay factor when applying our bounds to the DP-SGD algorithm (see Proposition 2) and numerical experiments suggest that our bound is more favourably correlated with the true generalization gap [see Table 1 in 287].

3.7 Numerical Experiments

In this section, we demonstrate our generalization bound (Proposition 4) through numerical experiments on the MNIST dataset [172], CIFAR-10 dataset [164], and SVHN dataset [203], showing that it can predict the behavior of the true generalization gap.

3.7.1 Corrupted Labels

As observed in Zhang et al. [307], DNNs have the potential to memorize the entire training dataset even when a large portion of the labels are corrupted. For networks with identical architecture, those trained using true labels have better generalization capability than those ones trained using corrupted labels, although both of them achieve perfect training accuracy. Unfortunately, distribution-independent bounds, such as the ones using VC-dimension, may not be able to capture this phenomenon because they are invariant for both true data and corrupted data. In contrast, our bound quantifies this empirical observation, exhibiting a lower value on networks trained on true labels compared to ones trained on corrupted labels (Figure 3.1).

In our experiment, we randomly select 5000 samples as our training dataset and change the label of $\alpha \in \{0\%, 25\%, 50\%, 75\%\}$ of the training samples. Then we use the SGLD algorithm to train a neural network under different corruption level. The training process continues until the training accuracy is 1.0 (see Figure 3.1 Left). We compare our generalization bound with the generalization gap in Figure 3.1 Middle and Right. As shown, both our bound and the generalization gap are increasing w.r.t. the corruption level in the last epoch. Furthermore, the curve of our bound has very

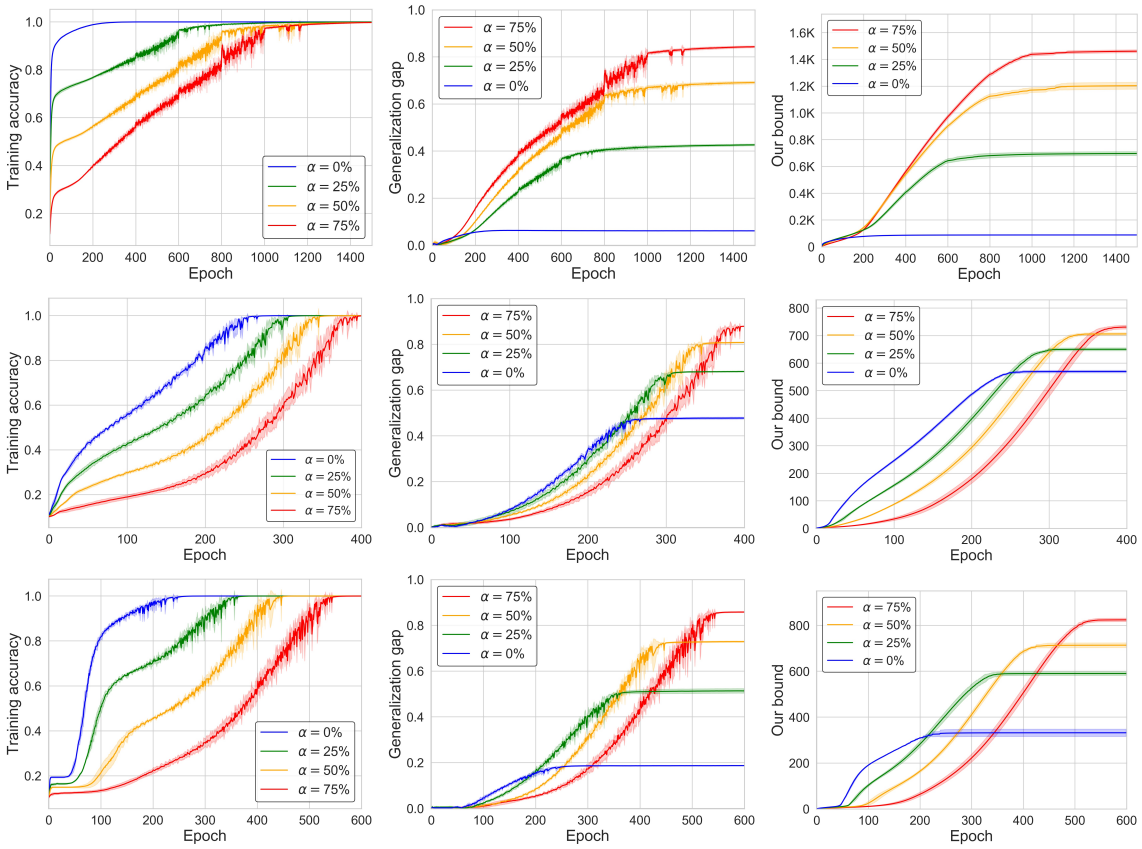

Figure 3.1: *Illustration of our generalization bound in Proposition 4. We use the SGLD algorithm to train 3-layer neural networks on MNIST (top row) and convolutional neural networks on CIFAR-10 (middle row) and SVHN (bottom row) when the training data have different label corruption level* $\alpha \in \{0\%, 25\%, 50\%, 75\%\}$. *Left column: training accuracy. Middle column: empirical generalization gap. Right column: empirical generalization bound.*

similar shape with the generalization gap. Finally, we observe that the generalization gap tends to be stable since the algorithm converges (Figure 3.1 Middle). Our generalization bound captures this phenomenon (Figure 3.1 Right) as the variance of gradients becomes negligible when the algorithm starts converging. The intuition is that the variance of gradients reflects the flatness of the loss landscape and as the algorithm converges, the loss landscape becomes flatter.

3.7.2 Network Width

As observed by several recent studies [see e.g., 137, 206], wider networks can lead to a smaller generalization gap. This may seem contradictory to the traditional wisdom as one may expect that a class of wider networks has a higher VC-dimension and, hence, would have a higher generalize gap. In our experiment, we use the SGLD algorithm to train neural networks with different widths. The

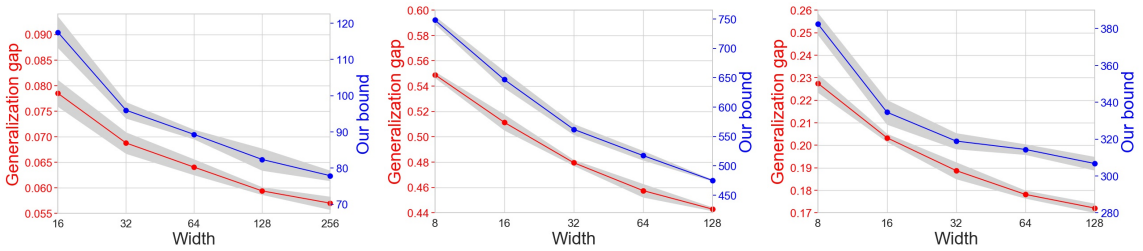

Figure 3.2: *Comparison between our generalization bound (Proposition 4) and the generalization gap. We use the SGLD algorithm to train neural networks with varying widths on the MNIST (left), CIFAR-10 (middle), and SVHN (right) datasets.*

training process runs for 400 epochs until the training accuracy is 1.0. We compare our generalization bound with the generalization gap in Figure 3.2. As shown, both the generalization gap and our bound are decreasing with respect to the network width.

3.8 Conclusion

In this chapter, we investigated the generalization of models trained by noisy iterative algorithms. We derived distribution-dependent generalization bounds in order to understand how the interaction between data distributions and optimization methods influences the generalization of complex ML models. We established a unified framework and leveraged fundamental tools from information theory (e.g., strong data processing inequalities and properties of additive noise channels) for proving these bounds. We demonstrated our generalization bounds through applications, including DP-SGD, FL, and SGLD, in which our bounds own a simple form and can be estimated from data. Numerical experiments suggest that our bounds can help explain some empirical observations of neural networks.

Chapter 4

Ensuring Fair Use of Group Attributes

Disparate treatment occurs when a machine learning model yields different decisions for individuals based on a group attribute (e.g., age, sex). In domains where prediction accuracy is paramount, it could be acceptable to fit a model which exhibits disparate treatment. In this case, it is crucial to ensure fair use of group attributes—the designed model should tailor treatment to people's unique characteristics while preventing harm to any group of individuals.

The role of a group attribute in fair classification can be understood through several metrics and principles. When a ML model is deployed in practice, fairness can be quantified in terms of the performance disparity *conditioned* on a group attribute, such as statistical parity [98] and equalized odds [119]. In domains where the goal is to predict accurately (e.g., medical diagnostics), *non-maleficence* (i.e., "do no harm") and *beneficence* (i.e., "do good") [31] become more appropriate moral principles for fairness [191, 267, 286, 305]. Accordingly, a ML model should avoid the causation of harm and be as accurate as possible on each protected group.

The relationship between achieving the above-mentioned principles and allowing a classifier to exhibit disparate treatment is complex. On the one hand, using a *group-blind classifier* (i.e., a classifier that does not use the group attribute as an input feature) may cause harm unintentionally since model performance relies on the distribution of the input data [86, 267, 286, 288, 289]. This probability distribution can vary significantly conditioned on a group attribute due to, for example, inherent differences between groups [86], differences in labeling [38], and differences in sampling [255]. On the other hand, training a separate classifier for each protected group—a setting we refer to as *splitting classifiers*—does not necessarily guarantee non-maleficence when sample size

is limited [310]: groups with insufficient samples may incur a high generalization error and suffer from overfitting. Motivated by this practical challenge, the first questions we study is: when is it beneficial to split classifiers in terms of model performance? In this chapter, we provide precise conditions under which splitting classifiers brings the most performance improvement compared with group-blind classifiers.

The next question we study is: how to learn a model while ensuring fair use of group attributes. We consider the setting where a given black-box classifier exhibits disparate impact[1]. We aim to eliminate the performance gap by perturbing the distribution of input variables for the disadvantaged group. We refer to the perturbed distribution as a counterfactual distribution and propose a descent algorithm to learn a counterfactual distribution from data. The estimated distribution can be used to build a data preprocessor that reduces disparate impact. Our approach can be applied to many practical settings. Imagine that a rural clinic purchases a classification model to detect bone fractures in X-rays and discovers that patients with a certain physical trait have high false positive rate; or a bank enters a new market and discovers its credit score underperforms on customers over 60 years of age. In all these cases, our tools can be used to repair the predictive models without training a new one.

4.1 Overview and Main Contributions

First, we show that in the information-theoretic regime where the underlying distribution is known—or, equivalently, an arbitrarily large number of samples are available—splitting *never harms* any group in terms of average performance metrics. Thus, splitting will naturally follow the non-maleficence principle in the large-sample regime. Second, we introduce a notion called the *benefit-of-splitting* which measures the performance improvement by splitting classifiers compared to using a group-blind classifier across all groups. The benefit-of-splitting is also an information-theoretic quantity as it only relies on the underlying data distribution rather than number of samples or hypothesis class.

The definition of the benefit-of-splitting involves a model performance measure and, hence, we divide our analyses into two parts based on different choices of this measure. In Section 4.5, we quantify model performance in terms of standard loss functions (e.g., ℓ_1 and cross entropy loss). For the benefit-of-splitting under these loss functions, we provide sharp upper and lower

[1] A machine learning model has disparate impact when its performance changes across groups [24].

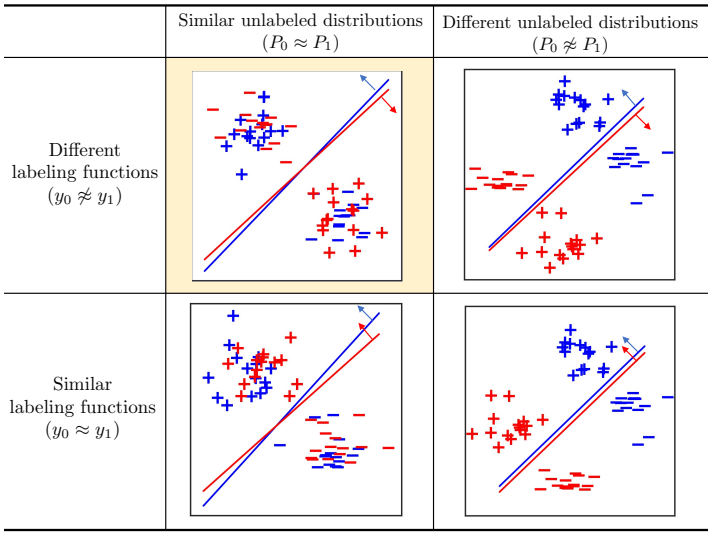

Figure 4.1: *The taxonomy of splitting based on two different factors. Samples from two groups are depicted in red and blue, respectively, and their labels are represented by $+$, $-$. Each group's labeling function is shown with the corresponding color and the arrows indicate the regions where the points are labeled as $+$. Splitting classifiers benefits model performance the most if the labeling functions are different and the unlabeled distributions are similar (yellow region).*

bounds (Theorem 3) that capture when splitting classifiers benefits model performance the most. These bounds indicate two factors (see Figure 4.1 for an illustration) which are central to the benefit-of-splitting: (i) disagreement between labeling functions[2], (ii) similarity between unlabeled distributions[2]. Based on these two factors, our upper bounds in Theorem 3 indicate that splitting does not produce much benefit if the labeling functions are similar or the unlabeled distributions are different; our lower bounds in Theorem 3 indicate that splitting benefits the most if two groups' labeling functions are different and unlabeled distributions are similar. Furthermore, our lower bounds in Theorem 3 lead to an impossibility (i.e., converse) result for group-blind classifiers: under certain precise conditions, using a group-blind classifier will always suffer from an inherent accuracy trade-off between different groups and splitting classifiers can reconcile this issue. This converse result is information-theoretic: a data scientist cannot overcome this limit by using more samples or altering the hypothesis class.

In Section 4.6, we consider false error rate as a performance measure since in applications such as medical diagnostics, high false error rate could result in unintentional harm [170]. Under this metric,

[2] We borrow the terms "labeling function" and "unlabeled distribution" from the domain adaptation literature [33, 190]. The labeling function takes a data point as an input and produces a probability of its binary label being 1 and the unlabeled distribution is a (marginal) probability distribution of the unlabeled data. Furthermore, the labeling function can be viewed as a "channel" (i.e., conditional distribution) in the information theory parlance. The formal definitions are given in Section 4.3.

computing the benefit-of-splitting directly from its definition may at first seem intractable since it involves an optimization over an infinite-dimensional functional space. Nonetheless, we prove that the benefit-of-splitting under false error rate has an equivalent, dual expression (Theorem 4) which only requires solving two small-scale convex programs. Furthermore, the objective functions of these convex programs have closed-form supergradients (Proposition 6). Combining these two results leads to an efficient procedure (Algorithm 2) for computing the benefit-of-splitting. We validate our procedure through numerical experiments on synthetic datasets in Section 4.9.1. When the underlying data distribution is known, our procedure has a provable convergence guarantee and returns the precise values of the benefit-of-splitting. When the underlying data distribution is unknown, our procedure may suffer from approximation errors but still outperforms more naive empirical approaches.

The aforementioned results capture the benefit-of-splitting from an information-theoretic perspective where the underlying data distributions are assumed to be known and the space of potential classifiers is unrestricted. In Section 4.7, we consider the effect of splitting classifiers in a more practical setting where group-blind and split classifiers are restricted over the same hypothesis class (e.g., logistic regressions) and the underlying distribution is accessed only through finitely many i.i.d. samples. In this case, splitting classifiers is not necessarily beneficial since the group with less samples may suffer from overfitting. To quantify the effect of splitting classifiers, we analyze the *sample-limited benefit-of-splitting*. We derive upper and lower bounds for the benefit-of-splitting in this regime in Theorem 6. These bounds disentangle three factors which determine the effect of splitting classifiers in practice: (i) disagreement between optimal (split) classifiers and training error associated with these optimal classifiers; (ii) similarity between (empirical) unlabeled distributions; and (iii) model complexity and number of samples. The first two factors are analogous to the ones that affect the benefit-of-splitting in the information-theoretic regime: when the hypothesis class is complex enough and the sample size tends to infinity, the optimal classifiers approximate the labeling functions and the empirical unlabeled distributions converge to the true unlabeled distributions. We illustrate how these factors determine the performance impact of splitting classifiers through experiments on 40 datasets downloaded from OpenML [272].

Finally, we consider how to use the group attribute while ensuring fair use. In particular, we consider the setting where a (black-box) classifier exhibits a performance disparity across groups of individuals. We aim to eliminate the performance gap by perturbing the distribution of input

variables for the disadvantaged group. We refer to the perturbed distribution as a counterfactual distribution, and characterize its properties for common (group) fairness criteria. Our tools recover a counterfactual distribution using a descent procedure in the simplex of probability distributions. We prove that influence functions can be used to compute a gradient in this setting, and derive closed-form estimators that enable efficient computation of influence functions for several fairness criteria. We design pre-processing methods that use counterfactual distributions to repair the black-box classifier without the need to train a new model. We validate our procedure by repairing classifiers trained with real-world datasets. Our results demonstrate how counterfactual distributions can help mitigate disparate impact in real-world applications.

The proof techniques of this chapter are based on fundamental tools found in statistics, such as Brown-Low's two-points lower bound [45], and methods in convex analysis, such as Ky Fan's min-max theorem [94]. These tools are widely used in applications such as non-parametric estimation [265], and are useful for analyzing the min-max risk in statistical settings [80, 139, 220, 298, 299]. Furthermore, the factors that we provide for understanding the effect of splitting classifiers are inspired by the necessary and sufficient conditions of domain adaptation learnability in Ben-David *et al.* [72].

4.2 Related Works

ML models have been increasingly used in applications of individual-level consequences, ranging from recidivism prediction [10] and lending [130] to healthcare [169]. A number of works in fair ML aim at understanding why discrimination happens [2, 60, 65, 68, 71, 83, 135, 136, 149, 161]; how it can be quantified [36, 61, 217]; and how it can be reduced [3, 5, 56, 57, 111, 121, 151, 153, 158, 194]. There are also an increasing number of studies that take causality into account for understanding and mitigating discrimination [64, 156, 167, 197]. We build on a line of recent results on decoupling predictive models for improving accuracy-fairness trade-offs [see e.g., 86, 160, 183, 267, 305]. For example, Ustun *et al.* [267] introduce a tree structure to recursively choose group attributes for decoupling. Lipton *et al.* [183] show that using group-blind classifiers could be suboptimal in terms of trading off accuracy and fairness. The work closest to ours is Dwork *et al.* [86] which present a decoupling technique to learn separate models for different groups. A detailed comparison with [86] is given in Section 4.7.2.

A standard assumption in ML is that the training and testing data are drawn from the same underlying probability distribution. Domain adaptation [33, 107, 190] and transfer learning [163, 168] consider a more general setting where models are trained on a source domain and deployed on a (different) target domain. A common assumption therein is known as *covariate shift*, which requires the source and target domain share the same labeling function. In this chapter, we prove (see Theorem 3) that if the covariate shift assumption is violated and two groups' unlabeled distributions are similar, then no classifier can perform well on both groups. In this regard, our work is connected to Ben-David *et al.* [72] which present impossibility results on domain adaptation learnability. Compared to [72], Theorem 3 characterizes an information-theoretic fundamental limit which cannot be circumvented by using a large number of samples or a carefully designed hypothesis class. Furthermore, the lower bound in Theorem 3 serves as a complementary statement to the upper bounds in domain adaptation [cf. 33, 190]. These bounds jointly describe the range of the loss a data scientist may incur by training a model on the source domain and deploying on the target domain.

We develop a theoretical framework that is used to design methods to determine counterfactual distributions in practice. We then use counterfactual distributions to design optimal transport-based pre-processing methods for ensuring fairness. In this regard, the closest work to ours are those of Del Barrio et al. [74], Feldman et al. [98], Johndrow and Lum [142], which propose methods to control specific disparate impact metrics via optimal transport. These methods differ from ours in that they (i) focus on reducing measures of disparity related to predicted outcomes; (ii) map the input variable distributions across *all* groups to a common distribution. More broadly, our approach differs from other model-agnostic approaches to mitigate disparate impact [e.g., pre-processing methods such as 56, 151] in that it does not require access to the training data, and does not require training a new model.

The term "counterfactual distribution" is often used to describe different kinds of hypothetical effects. In the statistics and economics literature [see e.g., 22, 63, 77, 100, 102, 141, 215, 234], a counterfactual distribution refers to a hypothetical distribution of an *outcome variable* given a specific distribution of input variables (e.g., the distribution of wages (outcome variable) for young workers if young workers had the same qualifications as older workers). The counterfactual distribution in this chapter describes a different kind of effect — i.e., a distribution of input variables to minimize disparate impact — and, consequently, must be derived using a different set of tools.

4.3 Preliminaries

Notation Consider a binary classification task (e.g., detecting pneumonia from X-rays) where the goal is to learn a probabilistic classifier $h : \mathcal{X} \to [0,1]$ that predicts a label (e.g., presence of pneumonia) $Y \in \{0,1\}$ using input features (e.g., chest X-rays) $X \in \mathcal{X}$. We assume there is an additional binary[3] group attribute (e.g., sex) $S \in \{0,1\}$ that does not belong to the input features X. We denote the unlabeled probability distributions of input features conditioned on the group attribute by

$$P_0 \triangleq P_{X|S=0}, \quad P_1 \triangleq P_{X|S=1}.$$

The labeling functions of the two groups are denoted by

$$y_0(x) \triangleq P_{Y|X,S}(1|x,0), \quad y_1(x) \triangleq P_{Y|X,S}(1|x,1).$$

In order to measure the difference between two unlabeled distributions (i.e., P_0 and P_1), we recall Csiszár's f-divergence [70]. Let $f : (0,\infty) \to \mathbb{R}$ be a convex function with $f(1) = 0$ and P, Q be two probability distributions over $\mathcal{X}$. The f-divergence between P and Q is defined by

$$\mathrm{D}_f(P\|Q) \triangleq \int_{\mathcal{X}} f\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \mathrm{d}Q. \tag{4.1}$$

Some examples of f-divergence are included in Appendix B.1.

4.4 The Benefit-of-Splitting

We study the impact of disparate treatment by comparing the performance between optimal group-blind and split classifiers. Recall that a ML model exhibits disparate treatment if it explicitly uses a group attribute to produce an output. We illustrate the difference between group-blind and split classifiers through the example of logistic regressions:

- a group-blind classifier does not use a group attribute as an input: $h(x) = \mathsf{logistic}(w^T x)$ where $\mathsf{logistic}(t) \triangleq 1/(1+\exp(-t))$ for $t \in \mathbb{R}$;
- split classifiers are a set of classifiers trained and deployed separately on each group: $h_s(x) =$

[3]For the sake of illustration, we assume that the group attribute S is binary but our results can be extended to a setting of multi-groups. Furthermore, split classifiers can be applied to a scenario where multiple subgroups overlap [153] since individuals belonging to both groups can opt for either one of the split classifiers.

$\mathsf{logistic}(w_s^T x)$ for $s \in \{0,1\}$.

We measure the performance of both group-blind and split classifiers in terms of the *disadvantaged group* (i.e., the group with worst performance). For a given performance measure $L_s(\cdot)$ (higher values indicate a worse performance), the performance of a group-blind classifier h and a set of split classifiers $\{h_s\}_{s\in\{0,1\}}$, respectively, is measured by

$$\max_{s\in\{0,1\}} L_s(h) \quad \text{and} \quad \max_{s\in\{0,1\}} L_s(h_s).$$

Consequently, the optimal group-blind and split classifiers (across all measurable functions from $\mathcal{X}$ to $[0,1]$) achieve the performance

$$\inf_{h:\mathcal{X}\to[0,1]} \max_{s\in\{0,1\}} L_s(h) \quad \text{and} \quad \max_{s\in\{0,1\}} \inf_{h:\mathcal{X}\to[0,1]} L_s(h).$$

Next, we introduce the benefit-of-splitting to quantify the effect of splitting classifiers compared to using a group-blind classifier.

Definition 5. For each $s \in \{0,1\}$, let $P_{X,Y|S=s}$ be a fixed probability distribution and $L_s(\cdot)$ be a performance measure, we define the benefit-of-splitting as

$$\epsilon_{\text{split}} \triangleq \inf_{h:\mathcal{X}\to[0,1]} \max_{s\in\{0,1\}} L_s(h) - \max_{s\in\{0,1\}} \inf_{h:\mathcal{X}\to[0,1]} L_s(h), \tag{4.2}$$

where the infimum is taken over all (measurable) functions.

The benefit-of-splitting is the difference between the performance of the optimal group-blind and split classifiers. In other words, if h^* and $\{h_s^*\}_{s\in\{0,1\}}$ are optimal group-blind and split classifiers respectively, i.e.,

$$h^* \in \operatorname*{argmin}_{h:\mathcal{X}\to[0,1]} \max_{s\in\{0,1\}} L_s(h), \quad h_s^* \in \operatorname*{argmin}_{h:\mathcal{X}\to[0,1]} L_s(h) \;\; s \in \{0,1\},$$

the benefit-of-splitting can be equivalently expressed as

$$\epsilon_{\text{split}} = \max_{s\in\{0,1\}} L_s(h^*) - \max_{s\in\{0,1\}} L_s(h_s^*). \tag{4.3}$$

In practice, a data scientist may restrict the type of classifiers by fixing a hypothesis class (e.g., logistic regressions). The benefit-of-splitting can be adapted for capturing the effect of splitting classifiers in this case (see Definition 9).

By the optimality of h_s^* and the max-min inequality, we have $L_s(h^*) \geq L_s(h_s^*)$ for $s \in \{0,1\}$ and

$\epsilon_{\text{split}} \geq 0$ which implies that, information-theoretically, using a separate classifier on each group will never diminish model performance compared to using a group-blind classifier. A natural question is: how much performance improvement does splitting classifiers bring? Before answering this question, we specify performance measures of interest and present the benefit-of-splitting under these performance measures.

4.4.1 Loss Reduction by Splitting

The first type of performance measures contains standard loss functions which have been widely used in fair ML [see e.g., 86] and domain adaptation [see e.g., 33]. These loss functions quantify the disagreement between the labeling function y_s and the probabilistic classifier h. We recast the benefit-of-splitting under these loss functions below.

Definition 6. The ℓ_1-benefit-of-splitting $\epsilon_{\text{split},1}$ is the benefit-of-splitting in Definition 5 with the performance measure:

$$L_s(h) = \mathbb{E}\left[|h(X) - y_s(X)| \mid S = s\right].$$

The ℓ_2-benefit-of-splitting $\epsilon_{\text{split},2}$ is the benefit-of-splitting in Definition 5 with the performance measure:

$$L_s(h) = \mathbb{E}\left[(h(X) - y_s(X))^2 \mid S = s\right].$$

The KL-benefit-of-splitting $\epsilon_{\text{split},\mathsf{KL}}$ is the benefit-of-splitting in Definition 5 with the performance measure:

$$L_s(h) = \mathbb{E}\left[\mathrm{D}_{\mathrm{KL}}(y_s(X)\|h(X)) \mid S = s\right],$$

where for $p, q \in [0,1]$, $\mathrm{D}_{\mathrm{KL}}(p\|q) \triangleq p\log(p/q) + (1-p)\log((1-p)/(1-q))$.

4.4.2 False Error Rate Reduction by Splitting

Now we use the false error rate (FER) as a performance measure. The false error rate of a classifier is the maximum[4] between (generalized) false positive rate and (generalized) false negative rate [217]. In healthcare, assuring low false error rate is as important as guaranteeing high accuracy since patients

[4]Our analysis can be extended to any convex combination of false positive rate and false negative rate.

could suffer from harm due to a classifier's false error rate [170]. For example, the false negative diagnosis may delay treatment in patients who are critically ill; the false positive diagnosis could lead to an unnecessary treatment. Furthermore, a classifier with high accuracy does not necessarily mean it has low false error rate. Hence, we consider how split classifiers reduce the false error rate by recasting the benefit-of-splitting under this performance measure.

Definition 7. The FER-benefit-of-splitting $\epsilon_{\text{split,FER}}$ is the benefit-of-splitting in Definition 5 with the performance measure:

$$L_s(h) = \max\left\{\mathbb{E}\left[h(X) \mid Y = 0, S = s\right], \mathbb{E}\left[1 - h(X) \mid Y = 1, S = s\right]\right\}. \tag{4.4}$$

Connection with equalized odds. Equalized odds, discussed by Hardt *et al.* [119], is a commonly used group fairness measure that requires different groups to have (approximately) the same false positive and false negative rates. Specifically, a probabilistic classifier $h : \mathcal{X} \to [0, 1]$ satisfies equalized odds [119, 217] if

$$\begin{aligned} \mathbb{E}\left[h(X) \mid Y = 0, S = 0\right] &= \mathbb{E}\left[h(X) \mid Y = 0, S = 1\right] && \text{(equal false positive rate)}, \\ \mathbb{E}\left[1 - h(X) \mid Y = 1, S = 0\right] &= \mathbb{E}\left[1 - h(X) \mid Y = 1, S = 1\right] && \text{(equal false negative rate)}. \end{aligned}$$

Under this definition, classifiers are considered "unfair" if their false positive rate or false negative rate vary across different groups. However, imposing equalized odds constraints may lead to a significant performance reduction in classification [65, 67, 103, 309]. In contrast, the benefit-of-splitting definition studied in this chapter aims to capture the principles of non-maleficence and beneficence [31]: classifiers should avoid the causation of harm and achieve the best performance on each group. By taking the optimal group-blind classifier as a baseline approach, this may allow split classifiers to potentially exhibit performance disparities between groups—as long as the split classifiers do not perform worse than the baseline approach and are as accurate as possible.

4.5 The Taxonomy of Splitting

In this section, we analyze the loss reduction by splitting classifiers compared to using a group-blind classifier. We achieve this goal by upper and lower bounding the benefit-of-splitting under different loss functions (see Definition 6). These bounds reveal factors which could impact the effect of

splitting classifiers and lead to a taxonomy of splitting, i.e., a characterization of when splitting benefits model performance the most or splitting does not bring much benefit.

Before stating the main result (i.e., bounds for the benefit-of-splitting), we prove a lemma first which converts the definition of the benefit-of-splitting into a single variable optimization problem. This lemma will be used in the proof of our lower bounds.

Lemma 10. *The benefit-of-splitting under different loss functions in Definition 6 have equivalent expressions*

$$\epsilon_{split,1} = \sup_{\omega\in[0,1]} (1-\omega)\int_{\mathcal{A}_\omega} |y_1(x)-y_0(x)|dP_1(x) + \omega\int_{\mathcal{A}_\omega^c} |y_1(x)-y_0(x)|dP_0(x),$$

$$\epsilon_{split,2} = \sup_{\omega\in[0,1]} \omega(1-\omega)\int \frac{(y_1(x)-y_0(x))^2 dP_0(x)dP_1(x)}{\omega dP_0(x)+(1-\omega)dP_1(x)},$$

$$\epsilon_{split,\mathsf{KL}} = \sup_{\omega\in[0,1]} \mathsf{JS}_\omega(P_{X,Y|S=0}\|P_{X,Y|S=1}) - \mathsf{JS}_\omega(P_0\|P_1),$$

where $\mathcal{A}_\omega \triangleq \left\{x\in\mathcal{X} \mid \frac{dP_0(x)}{dP_1(x)} \geq \frac{1-\omega}{\omega}\right\}$ *and* $\mathsf{JS}_\omega(\cdot\|\cdot)$ *is the Jensen-Shannon divergence.*

Proof. See Appendix B.2.1. □

Next, we provide upper and lower bounds for $\epsilon_{\text{split},1}$, $\epsilon_{\text{split},2}$, and $\epsilon_{\text{split},\mathsf{KL}}$, respectively. These bounds rely on two main factors: (i) disagreement between different groups' labeling functions and (ii) similarity between their unlabeled distributions. In particular, the second factor is captured by a certain f-divergence [70, 242] (see Appendix B.1 for some examples of f-divergence).

Theorem 3. *The ℓ_1-benefit-of-splitting can be upper and lower bounded*

$$\epsilon_{split,1} \leq \min\left\{\min_{s\in\{0,1\}} \sqrt{\mathbb{E}\left[(y_1(X)-y_0(X))^2 \mid S=s\right]}\cdot\sqrt{1-\mathrm{D}_{TV}(P_0\|P_1)},\right.$$
$$\left.\frac{1}{2}\max_{s\in\{0,1\}} \mathbb{E}\left[|y_1(X)-y_0(X)| \mid S=s\right]\right\},$$

$$\epsilon_{split,1} \geq \frac{1}{2}\max_{s\in\{0,1\}}\left\{\mathbb{E}\left[|y_1(X)-y_0(X)| \mid S=s\right] - \sqrt{\mathbb{E}\left[(y_1(X)-y_0(X))^2 \mid S=s\right]}\cdot d_2(P_{1-s}\|P_s)\right\},$$

where $\mathrm{D}_{TV}(P_0\|P_1)$ *is the total variation distance and* $d_2(P_{1-s}\|P_s)$ *is Marton's divergence. Suppose that*

$$\mathbb{E}\left[|y_1(X)-y_0(X)| \mid S=1\right] \geq \mathbb{E}\left[|y_1(X)-y_0(X)| \mid S=0\right].$$

Then the ℓ_2-benefit-of-splitting can be upper and lower bounded

$$\epsilon_{split,2} \leq \min\left\{\min_{s\in\{0,1\}} \sqrt{\mathbb{E}\left[(y_1(X)-y_0(X))^4 \mid S=s\right]} \cdot \sqrt{1-\mathrm{D}_{TV}(P_0\|P_1)}, \right.$$
$$\left. \frac{1}{4}\max_{s\in\{0,1\}} \mathbb{E}\left[(y_1(X)-y_0(X))^2 \mid S=s\right]\right\},$$
$$\epsilon_{split,2} \geq \left(\frac{\mathbb{E}\left[|y_1(X)-y_0(X)| \mid S=0\right]}{\sqrt{\mathrm{D}_{\chi^2}(P_1\|P_0)+1}+1}\right)^2,$$

where $\mathrm{D}_{\chi^2}(P_1\|P_0)$ is the chi-square divergence. The KL-benefit-of-splitting can be upper and lower bounded

$$\epsilon_{split,\mathsf{KL}} \leq \min\left\{2\mathsf{JS}(P_{X,Y|S=0}\|P_{X,Y|S=1}) - 2\mathsf{JS}(P_0\|P_1), \max_{s\in\{0,1\}} \mathbb{E}\left[\mathrm{D}_{\mathrm{KL}}\left(y_s(X)\|\tfrac{y_0(X)+y_1(X)}{2}\right) \mid S=s\right]\right\},$$
$$\epsilon_{split,\mathsf{KL}} \geq \mathsf{JS}(P_{X,Y|S=0}\|P_{X,Y|S=1}) - \mathsf{JS}(P_0\|P_1),$$

where $\mathsf{JS}(\cdot\|\cdot)$ is the Jensen–Shannon divergence.

Proof. See Appendix B.2.2. □

Now we consider extreme scenarios to verify the sharpness of the bounds and to understand when splitting classifiers benefits model performance the most (see Figure 4.1 for an illustration).

- Consider the setting where two groups share the same labeling function (i.e., $y_0 = y_1$). All the upper and lower bounds in Theorem 3 for the benefit-of-splitting under different loss functions become zero and, hence, the bounds are sharp. This is quite intuitive as one can use the labeling function y_0 as a group-blind classifier and it achieves perfect performance on both groups. Hence, there is no benefit of splitting classifiers.

- Consider the setting where two groups share the same unlabeled distribution (i.e., $P_0 = P_1$). The upper and lower bounds of $\epsilon_{\text{split},1}$ are both $\mathbb{E}\left[|y_1(X)-y_0(X)|\right]/2$, which is equal to $\epsilon_{\text{split},1}$. The bounds of $\epsilon_{\text{split},2}$ become

  $$\frac{1}{4}\mathbb{E}\left[|y_1(X)-y_0(X)|\right]^2 \leq \epsilon_{\text{split},2} \leq \frac{1}{4}\mathbb{E}\left[(y_1(X)-y_0(X))^2\right].$$

  If, in addition, $|y_0(x)-y_1(x)|$ is the same across all x, the upper and lower bounds become the same and, hence, are sharp. Finally, the bounds of $\epsilon_{\text{split},\mathsf{KL}}$ become

  $$\mathbb{E}\left[\mathsf{JS}(y_0(X)\|y_1(X))\right] \leq \epsilon_{\text{split},\mathsf{KL}} \leq \max_{s\in\{0,1\}} \mathbb{E}\left[\mathrm{D}_{\mathrm{KL}}\left(y_s(X)\|\tfrac{y_0(X)+y_1(X)}{2}\right)\right].$$

If, in addition, $\mathbb{E}\left[\mathrm{D}_{\mathrm{KL}}\left(y_0(X)\|(y_0(X)+y_1(X))/2\right)\right] = \mathbb{E}\left[\mathrm{D}_{\mathrm{KL}}\left(y_1(X)\|(y_0(X)+y_1(X))/2\right)\right]$, then the upper and lower bounds are equal. This extreme case indicates that when different groups have the same unlabeled distribution (i.e., $P_0 = P_1$), the benefit-of-splitting is determined by the disagreement between their labeling functions (i.e., large disagreement leads to high benefit).

- Consider the setting where two groups have unlabeled distributions lying on disjoint support sets. In this case, $\mathrm{D}_{\mathrm{TV}}(P_0\|P_1) = 1$ and $\mathsf{JS}(P_0\|P_1) = \log 2$. Hence, the upper bounds of $\epsilon_{\text{split},1}$ and $\epsilon_{\text{split},2}$ become zero. Furthermore,

$$0 \le \mathsf{JS}(P_{X,Y|S=0}\|P_{X,Y|S=1}) - \mathsf{JS}(P_0\|P_1) = \mathsf{JS}(P_{X,Y|S=0}\|P_{X,Y|S=1}) - \log 2 \le 0.$$

  where the last step is because the Jensen–Shannon divergence is always upper bounded by $\log 2$. Therefore, the upper bound of $\epsilon_{\text{split},\mathsf{KL}}$ is zero as well. In other words, there is no benefit of splitting classifiers when the unlabeled distributions are mutually singular. One can interpret this fact by considering a special group-blind classifier which mimics the labeling function of each group in the region where its unlabeled distribution lies. This classifier achieves perfect performance for each group. Note that such a group-blind classifier exists since we do not restrict the space of potential classifiers and, hence, any (measurable) function could become a classifier.

To summarize, from an information-theoretic perspective, splitting classifiers benefits the most if two groups have similar unlabeled distributions and different labeling functions. This taxonomy of splitting appears for all the commonly used loss functions (i.e., ℓ_1, ℓ_2, and KL loss).

Recall that the benefit-of-splitting (see Definition 5) measures the performance improvement by using the optimal split classifiers compared to deploying the optimal group-blind classifier across all groups. Here model performance is quantified in terms of the disadvantaged group (i.e., the group with the worst performance). We end this section by considering the Bayes risk as an alternative way of measuring model performance[5]. Specifically, the performance of a group-blind classifier h and a set of split classifiers $\{h_s\}_{s\in\{0,1\}}$, respectively, is measured by

$$\begin{aligned}
&\text{group-blind}: && \Pr(S=0)\cdot\mathbb{E}\left[|h(X)-y_0(X)| \mid S=0\right] + \Pr(S=1)\cdot\mathbb{E}\left[|h(X)-y_1(X)| \mid S=1\right],\\
&\text{split}: && \Pr(S=0)\cdot\mathbb{E}\left[|h_0(X)-y_0(X)| \mid S=0\right] + \Pr(S=1)\cdot\mathbb{E}\left[|h_1(X)-y_1(X)| \mid S=1\right].
\end{aligned}$$

[5]For the sake of illustration, in what follows we only consider the ℓ_1 loss.

They can be equivalently written as

$$\mathbb{E}\left[|h(X) - y_S(X)|\right] \quad \text{and} \quad \mathbb{E}\left[|h_S(X) - y_S(X)|\right].$$

The performance difference between the optimal group-blind and split classifiers leads to the following definition.

Definition 8. We define the population-benefit-of-splitting as

$$\epsilon_{\text{split,pop}} \triangleq \inf_{h:\mathcal{X}\to[0,1]} \mathbb{E}\left[|h(X) - y_S(X)|\right] - \inf_{\substack{h_s:\mathcal{X}\to[0,1] \\ \text{for } s\in\{0,1\}}} \mathbb{E}\left[|h_S(X) - y_S(X)|\right].$$

The population-benefit-of-splitting is upper bounded by the benefit-of-splitting (i.e., $\epsilon_{\text{split,pop}} \leq \epsilon_{\text{split,1}}$) since the Bayes risk is upper bounded by the worst-case risk and the split classifiers $\{y_s\}_{s\in\{0,1\}}$ composed by the labeling functions can achieve zero risk. Hence, the upper bound of $\epsilon_{\text{split,1}}$ in Theorem 3 naturally translates into an upper bound of $\epsilon_{\text{split,pop}}$. Next, we provide alternative bounds for $\epsilon_{\text{split,pop}}$ which reveal an additional factor influencing $\epsilon_{\text{split,pop}}$.

Proposition 5. *Assume* $\Pr(S=0) \leq 0.5$. *The population-benefit-of-splitting can be upper and lower bounded*

$$\epsilon_{split,pop} \leq \Pr(S=0) \cdot \mathbb{E}\left[|y_1(X) - y_0(X)| \mid S=0\right],$$
$$\epsilon_{split,pop} \geq \Pr(S=0)\left(\mathbb{E}\left[|y_1(X) - y_0(X)| \mid S=0\right] - E_{\frac{\Pr(S=1)}{\Pr(S=0)}}(P_0\|P_1)\right),$$

where $E_{\frac{\Pr(S=1)}{\Pr(S=0)}}(P_0\|P_1)$ *is the* E_γ*-divergence with* $\gamma = \Pr(S=1)/\Pr(S=0)$.

Proof. See Appendix B.2.3. □

Remark 4. The E_γ-divergence plays an important role in Bayesian statistical hypothesis testing [184, 242]. Since $\gamma \to E_\gamma(P\|Q)$ is non-increasing and $E_1(P\|Q) = \mathrm{D}_{\mathrm{TV}}(P\|Q)$ [184], we can further lower bound $\epsilon_{\text{split,pop}}$ by using the total variation distance

$$\epsilon_{\text{split,pop}} \geq \Pr(S=0)\left(\mathbb{E}\left[|y_1(X) - y_0(X)| \mid S=0\right] - \mathrm{D}_{\mathrm{TV}}(P_0\|P_1)\right).$$

The E_γ-divergence relates with the DeGroot statistical information [73] through (see Equation (421) in [242])

$$\mathcal{I}_p(P\|Q) = \begin{cases} pE_{\frac{1-p}{p}}(P\|Q) & p \in (0, \frac{1}{2}] \\ (1-p)E_{\frac{p}{1-p}}(Q\|P) & p \in [\frac{1}{2}, 1). \end{cases}$$

Hence, we can write our lower bound of $\epsilon_{\text{split,pop}}$ equivalently as

$$\epsilon_{\text{split,pop}} \geq \Pr(S=0) \cdot \mathbb{E}\left[|y_1(X) - y_0(X)| \mid S=0\right] - \mathcal{I}_{\Pr(S=0)}(P_0 \| P_1).$$

As shown in Proposition 5, the population-benefit-of-splitting is affected not only by the above-mentioned two factors (i.e., disagreement between labeling functions and similarity between unlabeled distributions) but also by the percentage of the minority group over the whole population. This reveals a caveat of the population-benefit-of-splitting: the minority group can be underrepresented when one designs a group-blind classifier by minimizing the loss over the whole population. In contrast, the benefit-of-splitting (see Definition 6) does not rely on the probability of the group attribute and, hence, represents each group equally.

4.6 An Efficient Procedure for Computing the Effect of Splitting

In the last section, we provide upper and lower bounds for the benefit-of-splitting under different kinds of loss functions. Here, we consider a different performance measure: false error rate. It turns out that the benefit-of-splitting under false error rate, denoted by $\epsilon_{\text{split,FER}}$ (see Definition 7), has an equivalent expression which leads to an efficient procedure of computing $\epsilon_{\text{split,FER}}$.

Even with the knowledge of the underlying data distribution, computing the benefit-of-splitting directly from its definition is challenging. This is because the space of potential classifiers is unrestricted (i.e., any measurable function could be used as group-blind or split classifiers) and solving optimization problems over this infinite-dimensional functional space could be intractable. One may attempt to circumvent this issue by restricting the classifiers over a hypothesis class. However, this naive approach has two limitations. First, it is unclear how to choose a hypothesis class in order to compute the benefit-of-splitting reliably. We will show in Example 1 that different hypothesis classes could result in completely different values of the benefit-of-splitting. Second, as evidenced in [306], training the optimal group-blind or split classifiers may suffer from a non-convexity issue.

We leverage the special form of the false error rate in (4.4) and prove an equivalent expression of $\epsilon_{\text{split,FER}}$ below which can be computed by solving two small-scale convex programs. The objective functions of these convex programs have closed-form supergradients. Hence, they can be solved efficiently via standard solvers, such as (stochastic) mirror descent [32, 202]. When the

data distribution is known, our procedure returns the precise values of $\epsilon_{\text{split,FER}}$ without the need of training optimal group-blind and split classifiers. The equivalent expression of $\epsilon_{\text{split,FER}}$ is given in the following theorem.

Theorem 4. *Assume* $\Pr(Y=i,S=s)>0$ *for any* $i,s\in\{0,1\}$. *The FER-benefit-of-splitting* $\epsilon_{\text{split,FER}}$ *can be equivalently written as*

$$\max_{\boldsymbol{\mu}\in\Delta_4}\left\{\sum_{s\in\{0,1\}}\mu_{s,1}+\mathbb{E}\left[\left(\sum_{s,i\in\{0,1\}}\mu_{s,i}\phi_{s,i}(X)\right)_-\right]\right\}-\max_{\substack{\boldsymbol{\nu}^{(s)}\in\Delta_2\\ \text{for } s\in\{0,1\}}}\left\{\nu_1^{(s)}+\mathbb{E}\left[\left(\sum_{i\in\{0,1\}}\nu_i^{(s)}\phi_{s,i}(X)\right)_-\right]\right\}.$$

Here for a positive integer d, $\Delta_d\triangleq\{z\in\mathbb{R}^d\mid\sum_{i=1}^d z_i=1,\ z_i\geq 0\}$, *for any* $a\in\mathbb{R}$, $(a)_-\triangleq\min\{a,0\}$, $\boldsymbol{\mu}\triangleq(\mu_{0,0},\mu_{0,1},\mu_{1,0},\mu_{1,1})$, $\boldsymbol{\nu}^{(s)}\triangleq(\nu_0^{(s)},\nu_1^{(s)})$, *and*

$$\phi_{s,i}(x)\triangleq\frac{(1-i-y_s(x))\Pr(S=s\mid X=x)}{\Pr(Y=i,S=s)},\quad s,i\in\{0,1\}. \tag{4.5}$$

Proof. See Appendix B.3.1. □

Remark 5. We demonstrate a proof sketch of Theorem 4. The FER-benefit-of-splitting $\epsilon_{\text{split,FER}}$ is composed by $\inf_{h:\mathcal{X}\to[0,1]}\max_{s\in\{0,1\}}L_s(h)$ and $\max_{s\in\{0,1\}}\inf_{h:\mathcal{X}\to[0,1]}L_s(h)$. The first term can be equivalent written as

$$\inf_{h:\mathcal{X}\to[0,1]}\max_{\boldsymbol{\mu}\in\Delta_4}\left\{\sum_{s\in\{0,1\}}\mu_{s,0}\mathbb{E}\left[h(X)\mid Y=0,S=s\right]+\mu_{s,1}\mathbb{E}\left[1-h(X)\mid Y=1,S=s\right]\right\}. \tag{4.6}$$

The key step in our proof is to swap maximum and infimum in (4.6) by using Ky Fan's min-max theorem [94] (see Lemma 26). Then for a fixed $\boldsymbol{\mu}$, the optimal classifier owns a closed-form expression. After some algebraic manipulations, (4.6) becomes the first convex program in the equivalent expression of $\epsilon_{\text{split,FER}}$. In the same vein, the another term $\max_{s\in\{0,1\}}\inf_{h:\mathcal{X}\to[0,1]}L_s(h)$ becomes the second convex program.

Next, we show that the objective functions of the convex programs in Theorem 4 have closed-form supergradients.

Proposition 6. *Under the same notations and assumptions in Theorem 4, functions* $g:\Delta_4\to\mathbb{R}$ *and* $g_s:\Delta_2\to\mathbb{R}$ *with* $s\in\{0,1\}$ *defined as*

$$g(\boldsymbol{\mu})\triangleq\sum_{s\in\{0,1\}}\mu_{s,1}+\mathbb{E}\left[\left(\sum_{s,i\in\{0,1\}}\mu_{s,i}\phi_{s,i}(X)\right)_-\right]\quad g_s(\boldsymbol{\nu})\triangleq\nu_1+\mathbb{E}\left[\left(\sum_{i\in\{0,1\}}\nu_i\phi_{s,i}(X)\right)_-\right]$$

have a closed-form supergradient, respectively:

$$\partial g(\boldsymbol{\mu}) \ni \left(i + \mathbb{E}\left[\psi_{s,i}(X) \cdot \mathbb{I}\Big[\sum_{s',i' \in \{0,1\}} \mu_{s',i'} \phi_{s',i'}(X) < 0 \Big] \;\middle|\; S = s \right] \right)_{s,i \in \{0,1\}}, \tag{4.7}$$

$$\partial g_s(\boldsymbol{\nu}) \ni \left(i + \mathbb{E}\left[\psi_{s,i}(X) \cdot \mathbb{I}\Big[\sum_{i' \in \{0,1\}} \nu_{i'} \phi_{s,i'}(X) < 0 \Big] \;\middle|\; S = s \right] \right)_{i \in \{0,1\}}, \tag{4.8}$$

where $\mathbb{I}[\cdot]$ *is the indicator function and*

$$\psi_{s,i}(x) \triangleq \frac{1 - i - y_s(x)}{\Pr(Y = i \mid S = s)}, \quad s, i \in \{0,1\}. \tag{4.9}$$

Proof. See Appendix B.3.2. □

When the underlying data distribution is known, one can compute $\epsilon_{\text{split,FER}}$ by solving the convex programs in Theorem 4 via standard tools, such as mirror descent, with convergence guarantees [32]. This is non-trivial because, as stated before, computing $\epsilon_{\text{split,FER}}$ directly from its definition could be intractable.

In practice, when the underlying data distribution is unknown, one can first approximate the conditional distribution $\Pr(S = 1 \mid X = x)$ and the labeling functions $y_0(x)$, $y_1(x)$ by training three well-calibrated binary classifiers. These classifiers will be called when computing the supergradient of the objective functions (see Proposition 6). We summarize our procedure of computing $\epsilon_{\text{split,FER}}$ in Algorithm 2 where stochastic mirror descent is used for solving the convex programs in Theorem 4. The numerical results are deferred to Section 4.9.1.

Algorithm 2 Computing $\epsilon_{\text{split,FER}}$ using stochastic mirror descent.

Input:
dataset: $\mathcal{D} = \{(x_i, y_i, s_i)\}_{i=1}^n$, maximum number of iterations: T, step size: $\{\eta_t\}_{t=1}^T$
Initialize
$\mathcal{I}_0 \leftarrow \{i = 1, \cdots, n \mid s_i = 0\}$ ▷ indices of points in $\mathcal{D}$ with $s_i = 0$
$\mathcal{D}_0 \leftarrow (x_i, y_i)$ for $i \in \mathcal{I}_0$ ▷ points with $s_i = 0$
$\mathcal{D}_1 \leftarrow (x_i, y_i)$ for $i \notin \mathcal{I}_0$ ▷ points with $s_i = 1$
approximate $\Pr(S = 1 \mid X = x)$ ▷ train a classifier using $\{(x_i, s_i)\}_{i=1}^n$
approximate $y_0(x)$ and $y_1(x)$ ▷ train two classifiers using $\mathcal{D}_0$ and $\mathcal{D}_1$, respectively
$\boldsymbol{\mu} \leftarrow (0.25, 0.25, 0.25, 0.25)$ and $\boldsymbol{\nu}^{(s)} \leftarrow (0.5, 0.5)$ ▷ initialize values
for t = 1,2, $\cdots$,T **do**
draw unlabeled sample $x_{0,t}$, $x_{1,t}$ ▷ randomly draw sample from $\mathcal{D}_0$, $\mathcal{D}_1$
pick $\boldsymbol{w} \in \partial g(\boldsymbol{\mu})$ and $\boldsymbol{w}^{(s)} \in \partial g_s(\boldsymbol{\nu}^{(s)})$ ▷ approximate supergradient using $x_{0,t}, x_{1,t}$
$\mu_j \leftarrow \mu_j \exp(\eta_t w_j) / \sum_{j'} \mu_{j'} \exp(\eta_t w_{j'})$ ▷ update variable via entropic descent
$\nu_j^{(s)} \leftarrow \nu_j^{(s)} \exp(\eta_t w_j^{(s)}) / \sum_{j'} \nu_{j'}^{(s)} \exp(\eta_t w_{j'}^{(s)})$ ▷ update variable via entropic descent
end for
return: $g(\boldsymbol{\mu}) - \max_{s \in \{0,1\}} g_s(\boldsymbol{\nu}^{(s)})$ ▷ the FER-benefit-of-splitting: $\epsilon_{\text{split,FER}}$

Our procedure can be understood through the following two steps:

- training a classifier to identify the group attribute using input features and a classifier for each group to predict label using input features;
- solving (convex) programs with these classifiers in hand.

We remark that this two-step approach has also appeared in [e.g., 194, 288] for designing "fair" classifiers.

4.7 Splitting in Practice

So far we have studied the benefit-of-splitting from an information-theoretic view as we assume the underlying data distribution is known and do not restrict the space of potential classifiers. In this section, we study the effect of splitting classifiers from a more practical perspective. First, we restrict the classifiers over a hypothesis class (e.g., logistic regressions) and analyze the hypothesis class dependent splitting. Second, we consider splitting classifiers in a finite sample regime and study the sample limited splitting.

4.7.1 Hypothesis Class Dependent Splitting

We restrict both group-blind and split classifiers over the same hypothesis class and introduce a hypothesis class dependent benefit-of-splitting for quantifying the loss reduction by splitting classifiers.

Definition 9. For a fixed probability distribution $P_{X,Y|S=s}$ with $s \in \{0,1\}$ and a given hypothesis class $\mathcal{H}$, the $\mathcal{H}$-benefit-of-splitting is defined as

$$\epsilon^{\mathcal{H}}_{\text{split}} \triangleq \inf_{h\in\mathcal{H}} \max_{s\in\{0,1\}} \mathbb{E}\left[|h(X) - y_s(X)| \mid S = s\right] - \max_{s\in\{0,1\}} \inf_{h\in\mathcal{H}} \mathbb{E}\left[|h(X) - y_s(X)| \mid S = s\right]. \tag{4.10}$$

Clearly, the $\mathcal{H}$-benefit-of-splitting maintains the non-maleficence principle $\epsilon^{\mathcal{H}}_{\text{split}} \geq 0$, i.e., given sufficient samples, splitting classifiers will never diminish model accuracy compared to using a group-blind classifier. Next, we provide upper and lower bounds for $\epsilon^{\mathcal{H}}_{\text{split}}$ in order to understand when splitting classifiers brings the most benefit. As before, these bounds rely on three major factors: (i) disagreement between optimal (split) classifiers; (ii) similarity between unlabeled distributions; and (iii) approximation error defined as the smallest loss achieved by split classifiers. In particular,

we assume that the last factor is small. This is a common assumption in, e.g., the domain adaptation literature [33] since when the hypothesis class is complex enough, this term will be negligible. Furthermore, one central notion of fairness we follow is non-maleficence (i.e., classifiers should avoid the causation of harm on any group). When the approximation error is large, neither group-blind classifiers nor splitting classifiers are accurate and "harm" is inevitable. Hence, one should change the hypothesis class first instead of splitting.

Theorem 5. *Let h_s^* be an optimal classifier for group $s \in \{0,1\}$:*

$$h_s^* \in \operatorname*{argmin}_{h \in \mathcal{H}} \mathbb{E}\left[|h(X) - y_s(X)| \mid S = s\right].$$

Then we have the following upper and lower bounds for the $\mathcal{H}$-benefit-of-splitting

$$\epsilon_{split}^{\mathcal{H}} \leq \min_{s \in \{0,1\}} \mathbb{E}\left[|h_1^*(X) - h_0^*(X)| \mid S = s\right]$$
$$\epsilon_{split}^{\mathcal{H}} \geq \frac{1}{2} \max_{s \in \{0,1\}} \mathbb{E}\left[|h_1^*(X) - h_0^*(X)| \mid S = s\right] - \mathrm{D}_{\mathrm{TV}}(P_0 \| P_1) - \frac{3\sum_{s \in \{0,1\}} \mathbb{E}\left[|h_s^*(X) - y_s(X)| \mid S = s\right]}{2}.$$

Proof. See Appendix B.4.1. □

Analogous to our discussions in Section 4.5, the bounds in Theorem 5 delineate a taxonomy of splitting when both group-blind and split classifiers are restricted over the same hypothesis class: splitting classifiers does not bring much benefit when two groups have similar optimal classifiers; splitting classifiers benefits the most when two groups have similar unlabeled distributions and different optimal classifiers. We further demonstrate this taxonomy of splitting and show how these factors influence the effect of splitting through numerical experiments in Section 4.9.2.

In contrast to the upper bound for ϵ_{split} (see Theorem 3), the upper bound for $\epsilon_{\text{split}}^{\mathcal{H}}$ does not involve the similarity between the unlabeled distributions. Consequently, when the optimal classifiers are different and the unlabeled distributions are different as well, it is unclear how much benefit splitting classifiers brings. We provide the following example which shows that different hypothesis classes may result in largely different values of the $\mathcal{H}$-benefit-of-splitting. Hence, one must study the effect of splitting on a case-by-case basis for different hypothesis classes.

Example 1. Let two groups' unlabeled distributions and labeling functions be $P_0 \sim \mathcal{N}(-\mu, 1)$, $y_0(x) = \mathbb{I}_{[x > -\mu]}$ and $P_1 \sim \mathcal{N}(\mu, 1)$, $y_1(x) = \mathbb{I}_{[x < \mu]}$, respectively. As μ grows larger, the distance between the unlabeled distributions P_0 and P_1 increases (i.e., $\mathrm{D}_{\mathrm{TV}}(P_0 \| P_1) \to 1$ as $\mu \to \infty$). Now we consider the following two hypothesis classes:

- $\mathcal{H}_{\text{threshold}}$ is the class of threshold functions over $\mathbb{R}$: $\mathbb{I}_{[x>a]}$ or $\mathbb{I}_{[x<b]}$.
- $\mathcal{H}_{\text{interval}}$ is the class of intervals over $\mathbb{R}$: $\mathbb{I}_{[x\in(a,b)]}$.

Here, a, b are allowed to be $-\infty$ and $+\infty$, respectively. In both cases, the labeling functions are included in the hypothesis classes and, hence, are optimal classifiers. The disagreement between these optimal classifiers is at least $1/2$:

$$\mathbb{E}\left[|y_1(X) - y_0(X)| \mid S = s\right] \geq 1/2, \quad s \in \{0, 1\}.$$

The benefit-of-splitting under $\mathcal{H}_{\text{threshold}}$ is $1/2$ as any group-blind classifier incurs at least $1/2$ loss on the disadvantaged group. On the other hand, as μ becomes larger, the benefit-of-splitting under $\mathcal{H}_{\text{interval}}$ is nearly 0 since a group-blind classifier with the form $h^*(x) = \mathbb{I}_{[x\in(-\mu,\mu)]}$ can achieve almost perfect accuracy.

The previous example shows that using a threshold function as a group-blind classifier will always incur an inevitable accuracy trade-off between two groups. On the other hand, if we enrich the hypothesis class to include interval functions, this trade-off can be reconciled. Motivated by this observation, when two groups have different unlabeled distributions and different labeling functions, we conjecture that the $\mathcal{H}$-benefit-of-splitting is determined by the "richness" of the hypothesis class: a more complex hypothesis class can produce a group-blind classifier which mimics the labeling function of each group in the region they lie in, and, hence, this classifier guarantees high accuracy for both groups. We formalize this intuition through the example of feedforward neural networks. Recall that a sigmoidal function [25] (e.g., logistic function) $S : \mathbb{R} \to \mathbb{R}$ is a bounded measurable function which satisfies $S(z) \to 1$ as $z \to +\infty$ and $S(z) \to 0$ as $z \to -\infty$. The hypothesis class associated to feedforward neural networks with one layer of sigmoidal functions has the form

$$\mathcal{H} = \left\{ \sum_{i=1}^{k} c_i S(a_i \cdot x + b_i) + c_0 \mid a_i \in \mathbb{R}^d, b_i, c_i \in \mathbb{R} \right\}. \tag{4.11}$$

In this case, Barron's approximation bounds [25] guarantee that these neural networks can approximate a large class of functions reliably.

Proposition 7. *Consider the hypothesis class $\mathcal{H}$ in (4.11). If $\mathcal{X} \subset \mathbb{R}^d$ is compact, we have*

$$\epsilon^{\mathcal{H}}_{split} \leq \min_{s\in\{0,1\}} \sqrt{\mathbb{E}\left[(y_1(X) - y_0(X))^2 \mid S = s\right]} \cdot \sqrt{1 - \mathrm{D}_{TV}(P_0 \| P_1)} + \frac{2\mathsf{diam}(\mathcal{X})C}{\sqrt{k}},$$

where $\mathsf{diam}(\mathcal{X}) = \sup_{x,x' \in \mathcal{X}} \|x - x'\|_2$,

$$h^*(x) \triangleq \frac{y_0(x)dP_0(x) + y_1(x)dP_1(x)}{dP_0(x) + dP_1(x)} = \int_{\mathbb{R}^d} \exp(iwx)\widetilde{h^*}(w)dw \tag{4.12}$$

for some complex-valued function $\widetilde{h^*}$*, and* $C \triangleq \int_{\mathbb{R}^d} \|w\|_2 |\widetilde{h^*}(w)|dw$.

Proof. See Appendix B.4.2. □

Remark 6. The condition in (4.12) goes back to the seminal work of Barron [25]. By the Fourier inversion theorem, if both h^* and its Fourier transform are integrable, this condition is satisfied. Further situations where (4.12) holds are discussed in [25, Section IX].

In contrast to Theorem 5, the upper bound for the $\mathcal{H}$-benefit-of-splitting above involves the similarity between the unlabeled distributions (i.e., $\mathrm{D}_{\mathrm{TV}}(P_0 \| P_1)$) at the cost of having an additional term which is inversely proportional to the hypothesis class complexity. The intuition behind our proof is that if a data scientist is able to train a neural network with enough neurons, a group-blind classifier is capable of guaranteeing high accuracy for both groups when their unlabeled distributions are different. Consequently, there is no much room for accuracy improvement by splitting classifiers.

4.7.2 Comparison with the Cost-of-Coupling

We compare our notion of the $\mathcal{H}$-benefit-of-splitting with the cost-of-coupling introduced by Dwork *et al.* [86]. We first illustrate the difference between group blind, coupled, split classifiers through the example of logistic regressions:

- a group blind classifier never uses a group attribute as an input: $h(x) = \mathsf{logistic}(w^T x)$;
- a coupled classifier uses a group attribute while sharing other parameters: $h(s,x) = \mathsf{logistic}(w^T x + w_0 s)$;
- split classifiers are a set of classifiers applied to each separate group: $h_s(x) = \mathsf{logistic}(w_s^T x)$.

Now we recast the definition of the cost-of-coupling [86] using our notation.

Definition 10 ([86])**.** Let $\mathcal{H}_{\mathrm{C}}$ be a hypothesis class which contains coupled classifiers from a finite set $\mathcal{S} \times \mathcal{X}$ to $[0,1]$. For a given loss function $\ell(\cdot,\cdot)$, the cost-of-coupling is defined as

$$\max_{P_{S,X,Y}} \left\{ \min_{h \in \mathcal{H}_{\mathrm{C}}} L(h) - \min_{\substack{h_s \in \mathcal{H}_{\mathrm{C}} \\ \text{for } s \in \mathcal{S}}} L(\{h_s\}_{s \in \mathcal{S}}) \right\},$$

where the maximum is over all distributions on $\mathcal{S} \times \mathcal{X} \times \{0,1\}$ and $L(h) \triangleq \mathbb{E}[\ell(Y, h(S,X))]$, $L(\{h_s\}_{s\in\mathcal{S}}) \triangleq \mathbb{E}[\ell(Y, h_S(S,X))]$.

There are two important differences between the $\mathcal{H}$-benefit-of-splitting (see Definition 9) and the cost-of-coupling [86]. First, our notion quantifies the gain in accuracy by using split classifiers rather than a group-blind classifier. In contrast, the cost-of-coupling compares coupled classifiers with split classifiers which both take a group attribute as an input. Second, the cost-of-coupling is a worst-case quantity as it maximizes over all distributions. By allowing our notion to rely on the data distribution, Definition 9 captures more intricate scenarios for characterizing the benefit of splitting classifiers. Furthermore, by taking the maximum over all distributions, we recover an analogous result of Theorem 2 in [86].

Corollary 2. *There exists a probability distribution $Q_{S,X,Y}$ whose $\mathcal{H}$-benefit-of-splitting is at least $1/2$ under*

1. *Linear predictors: $\mathcal{H} = \{\mathbb{I}_{[w^T x \geq 0]} \mid w \in \mathbb{R}^d\}$;*
2. *Decision trees: $\mathcal{H}$ is the set of binary decision trees.*

Furthermore, under this hypothetical distribution $Q_{S,X,Y}$, no matter which group-blind classifier $h \in \mathcal{H}$ is used, there is always a group $s \in \{0,1\}$ such that $\mathbb{E}[|h(X) - y_s(X)| \mid S = s] \geq 1/2$.

The proof technique used for this corollary can be extended to many other models (e.g., kernel methods or neural networks) and we defer its proof to Appendix B.4.3.

4.7.3 Sample Limited Splitting

Consider the following scenario. A data scientist has access to finitely many samples and she/he solves an empirical risk optimization in order to obtain an optimal group-blind classifier or a set of optimal split classifiers. When these classifiers are deployed on new fresh samples, a natural question is whether the optimal split classifiers still outperform the group-blind classifier. We introduce the sample-limited-splitting which quantifies the effect of splitting classifiers within this finite sample regime.

Definition 11. For a given hypothesis class $\mathcal{H}$ and n_s i.i.d. samples $\{(x_{s,i}, y_{s,i})\}_{i=1}^{n_s}$ from group $s \in \{0,1\}$, let $\hat{h}^*$ and $\{\hat{h}_s^*\}_{s\in\{0,1\}}$ be optimal group-blind and split classifiers for the empirical ℓ_1 loss,

respectively:

$$\hat{h}^* \in \operatorname*{argmin}_{h\in\mathcal{H}} \max_{s\in\{0,1\}} \frac{\sum_{i=1}^{n_s} |h(x_{s,i}) - y_{s,i}|}{n_s}, \tag{4.13}$$

$$\hat{h}_s^* \in \operatorname*{argmin}_{h\in\mathcal{H}} \frac{\sum_{i=1}^{n_s} |h(x_{s,i}) - y_{s,i}|}{n_s}, \quad s \in \{0,1\}. \tag{4.14}$$

The sample-limited-splitting is defined as

$$\hat{\epsilon}_{\text{split}} \triangleq \max_{s\in\{0,1\}} \mathbb{E}\left[|\hat{h}^*(X) - y_s(X)| \mid S = s\right] - \max_{s\in\{0,1\}} \mathbb{E}\left[|\hat{h}_s^*(X) - y_s(X)| \mid S = s\right]. \tag{4.15}$$

Unlike the benefit-of-splitting or the $\mathcal{H}$-benefit-of-splitting, the sample-limited-splitting is not necessarily non-negative. In other words, with limited amount of samples available, splitting classifiers may not improve accuracy for both groups. In what follows, we provide data-dependent upper and lower bounds for the sample-limited-splitting in order to understand the effect of splitting classifiers in the finite sample regime.

Theorem 6. *Let $\mathcal{H}$ be a hypothesis class from $\mathcal{X}$ to $\{0,1\}$ with VC dimension D. If $\hat{h}_s^*$ is a minimizer of the empirical ℓ_1 loss $\sum_{i=1}^{n_s} |h(x_{s,i}) - y_{s,i}|/n_s$ computed via n_s i.i.d. samples $\{(x_{s,i}, y_{s,i})\}_{i=1}^{n_s}$, then, with probability at least $1-\delta$,*

$$\hat{\epsilon}_{split} \leq \min_{s\in\{0,1\}} \frac{\sum_{i=1}^{n_s} |\hat{h}_1^*(x_{s,i}) - \hat{h}_0^*(x_{s,i})|}{n_s} + \Omega, \tag{4.16}$$

$$\hat{\epsilon}_{split} \geq \frac{1}{2} \max_{s\in\{0,1\}} \frac{\sum_{i=1}^{n_s} |\hat{h}_1^*(x_{s,i}) - \hat{h}_0^*(x_{s,i})|}{n_s} - \mathrm{D}_{TV}(\hat{P}_0 \| \hat{P}_1) - 3\lambda - \Omega, \tag{4.17}$$

where $\hat{P}_s$ is the empirical unlabeled distribution and

$$\lambda \triangleq \frac{1}{2}\left(\frac{\sum_{i=1}^{n_0} |\hat{h}_0^*(x_{0,i}) - y_{0,i}|}{n_0} + \frac{\sum_{i=1}^{n_1} |\hat{h}_1^*(x_{1,i}) - y_{1,i}|}{n_1}\right), \quad \Omega \triangleq 4 \max_{s\in\{0,1\}} \sqrt{\frac{2D\log(6n_s) + 2\log(8/\delta)}{n_s}}.$$

Proof. See Appendix B.4.4. □

Here, the term λ is the (average) training loss and Ω is the complexity term, which is approximately $\sqrt{D/\min\{n_0, n_1\}}$. As shown, the upper and lower bounds for $\hat{\epsilon}_{\text{split}}$ rely on four factors. The first three factors, which also appear in our bounds of the $\mathcal{H}$-benefit-of-splitting (see Theorem 5), are the disagreement between the (empirically) optimal classifiers, the similarity of the (empirically) unlabeled distributions, and the (empirically) training error. In addition to these factors, our bounds for $\hat{\epsilon}_{\text{split}}$ also depend on the number of samples from each group, especially minority group with less samples, and model complexity (measured by the VC dimension [11]).

4.8 Repairing without Retraining

In the previous sections, we study *when* does using a group attribute bring the most performance benefit. Here we consider *how* to use a group attribute to reduce disparate impact when it happens. Specifically, we consider the setting where a black-box classifier h exhibits a performance disparity across two groups. We assume that $S = 0$ attains the less favorable value of performance and refer to $S = 0$ and $S = 1$ as the *target* and *baseline* groups, respectively. Our goal is to repair the model by improving its performance on the target group.

We achieve this goal by perturbing the distribution of input variables for the target group (see Figure 4.2 for an illustration). We refer to the perturbed distribution as a counterfactual distribution and characterize its properties for common disparity metrics. We introduce a descent algorithm to learn a counterfactual distribution from data. Then we demonstrate how influence functions provide a natural "descent direction" in this setting and derive closed-form expressions for the influence functions. Lastly, we discuss how the counterfactual distribution can be used to build a data preprocessor that can help improve the model performance for the target group.

The tools we develop will be able to scrutinize and repair performance disparities in a way that: (i) only affects one group; (ii) benefits the group it affects (on average); and (iii) incentivizes individuals to reveal their group attributes at prediction time. The latter two points (i.e., do-no-harm and opt-in) are important elements of ethical treatment disparity [see e.g., 183, 267]. In what follows, we provide more details on this procedure.

4.8.1 Disparity Metrics

We measure the performance disparity between groups in terms of a *disparity metric*. Formally, a disparity metric is a mapping $\mathsf{M} : \mathcal{P} \to \mathbb{R}$ where $\mathcal{P}$ is the set of probability distributions over $\mathcal{X}$. We provide examples of $\mathsf{M}(P_0)$ for common fairness criteria in Table 4.1. Note that we write disparity metrics as $\mathsf{M}(P_0)$ since they can be expressed as a function of P_0 once the classifier and the distributions $P_{Y|X,S}$, P_1, and P_S are fixed.

4.8.2 Counterfactual Distributions

A *counterfactual distribution* is a hypothetical probability distribution of input variables for the target group that minimizes a specific disparity metric.

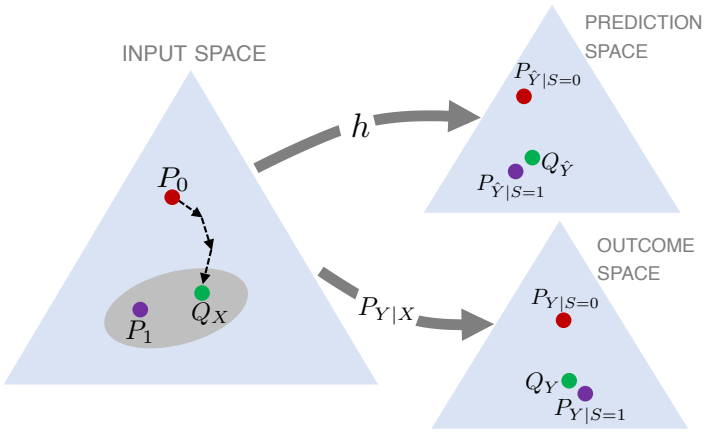

Figure 4.2: *Illustration of probability distributions affecting the disparate impact of a fixed classification model h. Here, P_0 and P_1 denote the distributions of input variables for groups where $S = 0$ and $S = 1$, respectively. A* counterfactual distribution *Q_X is a perturbation of P_0 that minimizes a specific measure of disparity. The counterfactual distribution may not be unique, as illustrated by the shaded ellipse.*

Performance Metric	Acronym	Disparity Metric
Statistical Parity	SP	$\Pr(\hat{Y} = 0 \mid S = 0) - \Pr(\hat{Y} = 0 \mid S = 1)$
False Discovery Rate	FDR	$\Pr(Y = 0 \mid \hat{Y} = 1, S = 0) - \Pr(Y = 0 \mid \hat{Y} = 1, S = 1)$
False Negative Rate	FNR	$\Pr(\hat{Y} = 0 \mid Y = 1, S = 0) - \Pr(\hat{Y} = 0 \mid Y = 1, S = 1)$
False Positive Rate	FPR	$\Pr(\hat{Y} = 1 \mid Y = 0, S = 0) - \Pr(\hat{Y} = 1 \mid Y = 0, S = 1)$

Table 4.1: *Disparity metrics $\mathsf{M}(P_0)$ for common fairness criteria. We assume that $S = 0$ attains the less favorable value of performance so that $\mathsf{M}(P_0) \geq 0$.*

Definition 12. A counterfactual distribution Q_X is a distribution of input variables for the target group such that:

$$Q_X \in \operatorname*{argmin}_{Q'_X \in \mathcal{P}} \left| \mathsf{M}(Q'_X) \right|, \tag{4.18}$$

where $\mathsf{M}(\cdot)$ is a given disparity metric and $\mathcal{P}$ is the set of probability distributions over $\mathcal{X}$.

There exist several ways to resolve the performance disparity of a fixed classifier by perturbing the distributions of input variables. For example, one could simultaneously perturb the input distributions for all groups to a "midpoint" distribution [see e.g., the distributions considered by 74, 98, 142, to achieve statistical parity].

While our tools could recover such distributions, we will purposely consider a counterfactual distribution that alters the input variables for a group that attains the less favorable performance

(i.e., the target group $S = 0$). This choice reflects our desire to resolve the performance disparity by having the target group perform better, rather than having the baseline group perform worse. As we discuss later, this choice reduces the data requirements to estimate the counterfactual distribution and the individuals who are affected by the repair (i.e., this approach only produces a preprocessor that affects individuals where $S = 0$).

At this point, an observant reader may wonder why a counterfactual distribution for the target group is not simply the distribution of input variables over the baseline group (i.e., $Q_X \equiv P_1$). In fact, the distribution of input variables for the baseline group P_1 *is not necessarily* a counterfactual distribution when $P_{Y|X,S=0} \neq P_{Y|X,S=1}$. We illustrate this point with the following example.

Example 2. Consider a classification task where the input variables $X = (X_1, X_2) \in \{0,1\}^2$ are drawn from distributions such that $P_{X|S=s} = P_{X_1|S=s} \cdot P_{X_2|S=s}$ for $s \in \{0,1\}$ where:

$$\Pr(X_1 = 1|S = 0) = 0.9,\ \Pr(X_2 = 1|S = 0) = 0.2,$$
$$\Pr(X_1 = 1|S = 1) = 0.1,\ \Pr(X_2 = 1|S = 1) = 0.5.$$

Assume that the true outcome variables Y are drawn from the conditional distributions:

$$\begin{aligned} P_{Y|X,S=0}(1|\boldsymbol{x}) &= \mathsf{logistic}(2x_1 - 2x_2), \\ P_{Y|X,S=1}(1|\boldsymbol{x}) &= \mathsf{logistic}(2x_1 + 4x_2 - 3). \end{aligned} \tag{4.19}$$

In this case, the Bayes optimal classifier for $S = 1$ is $h(\boldsymbol{x}) = \mathbb{I}_{[x_2=1]}$. Using the difference in FPR as the disparity metric, h achieves $\mathsf{M}(P_0) = 25.1\%$. In this case, setting $P_0 \leftarrow P_1$ would achieve a disparity of $\mathsf{M}(P_1) = 43.6\%$. In contrast, we can achieve a disparity metric of $\mathsf{M}(Q_X) = 0.0\%$ for a counterfactual distribution such that

$$Q_X(0,0) = 0.50, \qquad Q_X(0,1) = 0.09,$$
$$Q_X(1,0) = 0.41, \qquad Q_X(1,1) = 0.00.$$

Example 2 shows that counterfactual distributions may be non-trivial when the conditional distributions of Y given X differ across groups (i.e., $P_{Y|X,S=0} \neq P_{Y|X,S=1}$). In particular, the condition $P_{Y|X,S=0} \neq P_{Y|X,S=1}$ will always hold whenever counterfactual distributions do not completely eliminate the disparity between groups. We formalize this statement in the next proposition.

Proposition 8. *If* $\mathsf{M}(Q_X) > 0$ *where* Q_X *is a counterfactual distribution for a disparity metric in Table 4.1, then* $P_{Y|X,S=0} \neq P_{Y|X,S=1}$.

Proposition 8 illustrates how a counterfactual distribution can be used to detect cases where a classifier exhibits an irreconcilable performance disparity between groups – i.e., a disparity that cannot be resolved by perturbing the distributions of input variables for the target group. The result complements various impossibility results on inevitable trade-offs between groups [see e.g., 65, 161, 217]. It also provides a sufficient condition that can inform the need for treatment disparity [see e.g., of 86, 160, 183, 267].

4.8.3 Measuring the Descent Direction

In what follows, we describe how to reduce the value of a disparity metric by perturbing the distribution of input variables over the target group P_0.

We start by formally defining the local perturbation of an input distribution.

Definition 13. The perturbed distribution $\widetilde{P}_0$ over the target group ($S = 0$) is given by

$$\widetilde{P}_0(\boldsymbol{x}) \triangleq P_0(\boldsymbol{x})(1 + \epsilon f(\boldsymbol{x})), \quad \forall \boldsymbol{x} \in \mathcal{X} \tag{4.20}$$

where $f : \mathcal{X} \to \mathbb{R}$ is a perturbation function from the class of all functions with zero mean and unit variance w.r.t. P_0, and $\epsilon > 0$ is a positive scaling constant chosen so that $\widetilde{P}_0$ is a valid probability distribution.

Here, $f(\boldsymbol{x})$ represents a direction in the probability simplex while ϵ represents the magnitude of perturbation [see e.g., 8, 41, 127, for other applications of local perturbations of measures in information theory].

As we will see shortly, the direction of steepest descent for disparate impact can be measured using an *influence function* [see e.g., 129, 162, for other uses in machine learning].

Definition 14. For a disparity metric $\mathsf{M}(\cdot)$, the influence function $\psi : \mathcal{X} \to \mathbb{R}$ is given by

$$\psi(\boldsymbol{x}) \triangleq \lim_{\epsilon \to 0} \frac{\mathsf{M}\left((1-\epsilon)P_0 + \epsilon\delta_{\boldsymbol{x}}\right) - \mathsf{M}(P_0)}{\epsilon} \tag{4.21}$$

where $\delta_{\boldsymbol{x}}(\boldsymbol{z}) = \mathbb{I}_{[\boldsymbol{z}=\boldsymbol{x}]}$ is the delta function at $\boldsymbol{x}$.

Intuitively, given a sufficiently large dataset from the deployment population, the influence function approximates the change in a disparity metric when a sample $\boldsymbol{x} \in \mathcal{X}$ from the target group is removed (or added) to the dataset.

In Proposition 9, we show that perturbing the distribution P_0 along the direction defined by $-\psi(\boldsymbol{x})$ produces the largest local decrease of the disparity metric. That is, $-\psi(\boldsymbol{x})$ reflects the direction of steepest descent in disparate impact.

Proposition 9. *Given a disparity metric* $\mathsf{M}(\cdot)$*, we have that*

$$\underset{f(\boldsymbol{x})}{\operatorname{argmin}} \lim_{\epsilon \to 0} \frac{\mathsf{M}(\widetilde{P}_0) - \mathsf{M}(P_0)}{\epsilon} = \frac{-\psi(\boldsymbol{x})}{\sqrt{\mathbb{E}\left[\psi(X)^2 | S = 0\right]}}, \tag{4.22}$$

for any influence function $\psi : \mathcal{X} \to \mathbb{R}$ *such that* $\mathbb{E}\left[\psi(X)^2 | S = 0\right] \neq 0$.

Proposition 10 shows that when disparity is measured using a linear combination of metrics, the influence function for the compound metric can be expressed as a linear combination of the influence functions for its components.

Proposition 10. *Given any convex combination of K disparity metrics* $\mathsf{M}(P_0) = \sum_{i=1}^{K} \lambda_i \mathsf{M}_i(P_0)$*, the influence function of the compound disparity metric* $\mathsf{M}(P_0)$ *has the form:*

$$\psi(\boldsymbol{x}) = \sum_{i=1}^{K} \lambda_i \psi_i(\boldsymbol{x}). \tag{4.23}$$

Proposition 10 allows us to consider a larger class of disparity measures than those shown in Table 4.1. For instance, one can recover a counterfactual distribution to achieve equalized odds [119] by using a convex combination of influence functions for FPR and FNR.

4.8.4 Computing Influence Functions

We now present closed-form expressions for the influence functions of disparity metrics shown in Table 4.1. The expressions will be cast in terms of two classifiers:

- $h(\boldsymbol{x})$: the black-box classifier that we aim to repair;
- $\hat{y}_0(\boldsymbol{x})$: a classifier that uses the same input variables as h, but aims to predict the true outcome *for individuals in the target group*, $P_{Y|X,S=0}(1|\boldsymbol{x})$.

Given $h(\boldsymbol{x})$, we train $\hat{y}_0(\boldsymbol{x})$ using an *auditing dataset* $\mathcal{D}^{\text{audit}} = \{(\boldsymbol{x}_i, y_i, s_i)\}_{i=1}^{n}$ drawn from the deployment population. With these models in hand, we can then compute influence functions using closed-form expressions shown in the following proposition.

Proposition 11. *The influence functions for the disparity metrics in Table 4.1 can be expressed as*

$$\begin{aligned}
\psi^{SP}(\boldsymbol{x}) &= -h(\boldsymbol{x}) + \hat{\mu}_0, \\
\psi^{FDR}(\boldsymbol{x}) &= \frac{h(\boldsymbol{x})(1-\hat{y}_0(\boldsymbol{x})) - \nu_{0,1}h(\boldsymbol{x})}{\hat{\mu}_0}, \\
\psi^{FNR}(\boldsymbol{x}) &= \frac{(1-h(\boldsymbol{x}))\hat{y}_0(\boldsymbol{x}) - \gamma_{0,1}\hat{y}_0(\boldsymbol{x})}{\mu_0}, \\
\psi^{FPR}(\boldsymbol{x}) &= \frac{h(\boldsymbol{x})(1-\hat{y}_0(\boldsymbol{x})) - \gamma_{1,0}(1-\hat{y}_0(\boldsymbol{x}))}{(1-\mu_0)},
\end{aligned}$$

where μ_s, $\hat{\mu}_s$, $\gamma_{a,b}$, and $\nu_{a,b}$ are constants such that

$$\begin{aligned}
\mu_s &\triangleq \Pr(Y=1|S=s), \\
\hat{\mu}_s &\triangleq \Pr(\hat{Y}=1|S=s), \\
\gamma_{a,b} &\triangleq \Pr(\hat{Y}=a|Y=b,S=0), \\
\nu_{a,b} &\triangleq \Pr(Y=a|\hat{Y}=b,S=0).
\end{aligned}$$

4.8.5 Learning Counterfactual Distributions from Data

So far we have shown that influence functions can be used to evaluate the direction of steepest descent of a disparity metric (Proposition 9), and that the value of an influence function can be estimated using data from the deployment population (Proposition 11).

Considering these results, one would expect that disparity could be minimized by repeatedly (i) perturbing the distribution in the direction of steepest descent (4.22), and (ii) estimating the influence function at the new, perturbed distribution. Repeating these steps, we would recover an approximate solution to (4.18) – i.e., an approximate counterfactual distribution.

In Algorithm 3, we formalize this intuition by presenting a descent procedure to recover a counterfactual distribution for a given disparity metric $\mathsf{M}(\cdot)$. The procedure is analogous to stochastic gradient descent in the space of distributions over $\mathcal{X}$, where the resampling at each iteration corresponds to a gradient step. Given a classifier h and a dataset $\{(\boldsymbol{x}_i, y_i, s_i)\}_{i=1}^n$ from the distribution $P_{X,Y,S}$, the procedure outputs a dataset drawn from a counterfactual distribution.

The procedure pairs each point with a *sampling weight* w_i, which is initialized as $w_i = 1.0$. At each iteration, it first computes the value of the influence function $\psi(\boldsymbol{x})$. Next, it updates the values of each sampling weight for each point in the target group as $(1 - \epsilon\psi(\boldsymbol{x}_i)) \cdot w_i$, where ϵ is a user-specified step size parameter. The updated sampling weights represent the direction in which the distribution for the target group should be perturbed to reduce $\mathsf{M}(\cdot)$. The data points from the target group are

Algorithm 3 Distributional Descent.

Input:
- $h : \mathcal{X} \to [0, 1]$ ▷ classification model
- $\mathcal{D} = \{(\boldsymbol{x}_i, y_i, s_i)\}_{i=1}^n$ ▷ data from deployment population
- $\mathsf{M}(\cdot)$ ▷ disparity metric
- $\epsilon > 0$ ▷ step size

Initialize

$I_0 \leftarrow \{i = 1, \ldots, n \mid s_i = 0\}$
$\mathcal{D}_0 \leftarrow (\boldsymbol{x}_i, y_i)$ for $i \in I_0$ ▷ samples where $s_i = 0$
$\mathcal{D}_1 \leftarrow (\boldsymbol{x}_i, y_i)$ for $i \notin I_0$ ▷ samples where $s_i \neq 0$
$\boldsymbol{w}_0 \leftarrow [w_i]_{i \in I_0}$ where $w_i = 1.0$ ▷ initialize weights
$M \leftarrow \mathsf{M}(\mathcal{D}_0 \cup \mathcal{D}_1)$ ▷ evaluate disparity metric
repeat
  $M^{\text{old}} \leftarrow M$
  $\psi_i \leftarrow \psi(\boldsymbol{x}_i)$ for $i \in I_0$ ▷ compute $\psi(\boldsymbol{x}_i)$ for points in $\mathcal{D}_0$
  $w_i \leftarrow (1 - \epsilon\psi_i) \cdot w_i$ for $i \in I_0$
  $\widetilde{\mathcal{D}}_0 \leftarrow \textsc{Resample}(\mathcal{D}_0, \boldsymbol{w}_0)$
  $M \leftarrow \mathsf{M}(\widetilde{\mathcal{D}}_0 \cup \mathcal{D}_1)$ ▷ evaluate disparity metric
until $M \geq M^{\text{old}}$
return: $\boldsymbol{w}_0$, $\widetilde{\mathcal{D}}_0$ ▷ samples from counterfactual distribution

procedure $\textsc{Resample}(\mathcal{D}, \boldsymbol{w})$
  return: $|\mathcal{D}|$ points sampled from $\mathcal{D}$ using the weights $\boldsymbol{w}$
end procedure

then resampled with their sampling weights. The set of resampled points mimics one drawn from the perturbed distribution.

The procedure determines if the classifier still has disparate impact at the end of each iteration by computing the value of $\mathsf{M}(\cdot)$ on the set of resampled points. These steps are repeated until $\mathsf{M}(\cdot)$ ceases to decrease. Once the procedure stops, it outputs: (i) dataset drawn from a counterfactual distribution; (ii) a set of sampling weights for each point from the target group, which can be used to draw samples from the counterfactual distribution.

4.8.6 Model Repair

Next, we describe how to use counterfactual distributions to repair classifiers that exhibit disparate impact.

Preprocessor Given a classifier $h(\boldsymbol{x})$, we aim to mitigate disparate impact by constructing a *preprocessor* $T : \mathcal{X} \to \mathcal{X}$ that alters the features of the target group. Thus, the *repaired classifier* $\widetilde{h}(\boldsymbol{x})$ will

operate as:

$$\widetilde{h}(\boldsymbol{x}) = \begin{cases} h(T(\boldsymbol{x})) & \text{if } s = 0, \\ h(\boldsymbol{x}) & \text{otherwise.} \end{cases} \tag{4.24}$$

The preprocessor is a (potentially randomized) mapping that transforms the distribution of samples over the target population into the counterfactual distribution, i.e., given a random variable X drawn from the target population distribution, the distribution of $T(X)$ will approximate counterfactual distribution.

Optimal Transport We produce the preprocessor by solving an optimal transport problem. To this end, we require the following inputs:

- $\mathcal{D}_0$, which represents the original samples for the target group. We assume $\mathcal{D}_0$ contains n_0 samples, of which m are distinct: $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_m\}$.
- $\widetilde{\mathcal{D}}_0$, which represents the samples drawn from the counterfactual distribution (i.e., the data produced via resampling in Algorithm 3). We assume that $\widetilde{\mathcal{D}}_0$ contains $\widetilde{n}_0$ samples, of which $\widetilde{m}$ are distinct: $\{\widetilde{\boldsymbol{x}}_1, \cdots, \widetilde{\boldsymbol{x}}_{\widetilde{m}}\}$.

With these samples at hand, we formulate an optimal transport problem of the form:

$$\min_{\gamma_{ij} \in \mathbb{R}^+} \quad \sum_{i=1}^{m} \sum_{j=1}^{\widetilde{m}} C_{ij} \gamma_{ij} \tag{4.25a}$$

$$\text{s.t.} \quad \sum_{j=1}^{\widetilde{m}} \gamma_{ij} = p_i \quad i = 1, \cdots, m \tag{4.25b}$$

$$\sum_{i=1}^{m} \gamma_{ij} = q_j \quad j = 1, \cdots, \widetilde{m}. \tag{4.25c}$$

Here, C_{ij} represents the cost of altering the input variables from $\boldsymbol{x}_i$ to $\widetilde{\boldsymbol{x}}_j$ given a user-specified *cost function* that we will discuss shortly; p, q are the empirical estimates of P_0 and Q_X, respectively,

$$p_i = \frac{1}{n_0} \sum_{\boldsymbol{x} \in \mathcal{D}_0} \delta_{\boldsymbol{x}_i}(\boldsymbol{x}), \quad q_j = \frac{1}{\widetilde{n}_0} \sum_{\boldsymbol{x} \in \widetilde{\mathcal{D}}_0} \delta_{\widetilde{\boldsymbol{x}}_j}(\boldsymbol{x});$$

The optimal transport problem in (4.25) is a standard linear program that aims to find a *coupling* of p and q, γ [see e.g., 216, 277]. Formally, a coupling is a joint probability distribution with marginal distributions specified by p and q. Given the minimal-cost coupling γ^*, one can construct

a (randomized) preprocessor $T(\cdot)$ which takes a sample $\boldsymbol{x}_i$ and returns an altered sample $\widetilde{\boldsymbol{x}}_j$ with probability γ^*_{ij}/p_i.

We note that the linear programming formulation in (4.25) is designed for settings with discrete input distributions. In settings when the distributions P_0 and Q_X are continuous, an analogous optimal transport problem can be formulated and solved with other approaches [see e.g., 9, 35].

Choice of Cost Function The cost function C_{ij} controls how samples of the target group are perturbed. By default, one could use a standard distance metric such as the L_2-norm [e.g., 74, 98, 142]. However, one could also consider additional criteria to fine-tune the mapping specified by $T(\cdot)$. For example, one can specify a cost function that avoids specific kinds of change by setting the value of C_{ij} to a large constant so as to penalize undesirable mappings [e.g., a mapping that would alter immutable attributes such as marital status 268].

Customization via Constraints Users can also fine-tune the behavior of the preprocessor by adding custom constraints to the feasible region of optimal transport problems as in (4.25). For example, one can impose constraints on *individual fairness* to ensure that the repaired classifier will "treat similar individuals similarly" [see e.g., 85]. This behavior could be induced by including constraints of the form:

$$\frac{1}{2}\sum_{j=1}^{\widetilde{m}}\left|\frac{\gamma_{ij}}{p_i}-\frac{\gamma_{lj}}{p_l}\right| \leq \mathsf{d}(\boldsymbol{x}_i,\boldsymbol{x}_l) \quad \text{for all } i,l\in[m].$$

Here, the LHS is the total-variation distance [69] between the distributions of $T(\boldsymbol{x}_i)$ and $T(\boldsymbol{x}_l)$, and $\mathsf{d}(\cdot,\cdot):\mathcal{X}\times\mathcal{X}\to\mathbb{R}^+$ is a distance metric that reflects the similarity between samples.

4.9 Numerical Experiments

We illustrate the theoretical results presented in this chapter through experiments. In Section 4.6, we presented an algorithm (Algorithm 2) for computing the benefit-of-splitting. In particular, when the data distribution is known, this algorithm provably converges to the exact value of the benefit-of-splitting. To evaluate Algorithm 2, we conduct experiments on a synthetic example where both the data distribution and the values of the benefit-of-splitting are known. In Section 4.7, we characterized a taxonomy of splitting when classifiers are restricted over a hypothesis class. We demonstrate this taxonomy of splitting through experiments on 40 real-world datasets. In Section 4.9.3, we

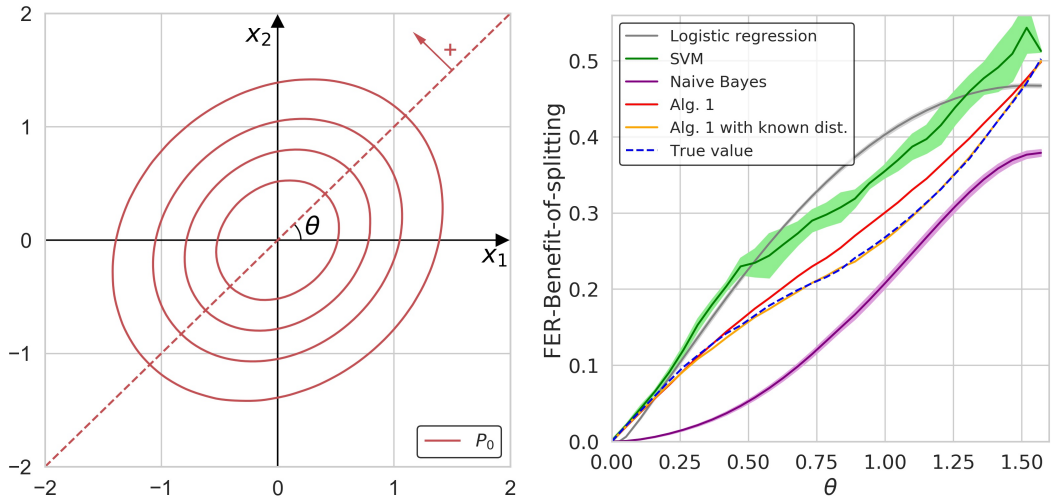

Figure 4.3: *We demonstrate the performance of Algorithm 2 for computing the FER-benefit-of-splitting $\epsilon_{split,\mathsf{FER}}$ on synthetic datasets. Left: the ellipses are the level sets of the unlabeled distribution P_0 and the dash line is the labeling function y_0 with a arrow indicating the region where points are labeled as $+$. Right: $\epsilon_{split,\mathsf{FER}}$ computed by different approaches along with its true values.*

demonstrate how counterfactual distributions can be used to avoid disparate impact for classifiers on real-world datasets.

4.9.1 Synthetic Datasets

We introduced the FER-benefit-of-splitting $\epsilon_{\text{split,FER}}$ in Section 4.4.2 and proposed an efficient procedure for computing this quantity (Algorithm 2). Here, we validate Algorithm 2 through experiments on synthetic datasets. For a fixed parameter $\theta \in [0, \pi/2]$, let two groups' unlabeled distributions be zero-mean Gaussian distributions with different covariance matrices: $P_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_0)$ and $P_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_1)$ where

$$\mathbf{\Sigma}_0 = \begin{pmatrix} 0.5\cos(\theta)^2 + 1 & 0.5\sin(\theta)\cos(\theta) \\ 0.5\sin(\theta)\cos(\theta) & 0.5\sin(\theta)^2 + 1 \end{pmatrix}, \quad \mathbf{\Sigma}_1 = \begin{pmatrix} 0.5\cos(\theta)^2 + 1 & -0.5\sin(\theta)\cos(\theta) \\ -0.5\sin(\theta)\cos(\theta) & 0.5\sin(\theta)^2 + 1 \end{pmatrix}.$$

The distributions P_0 and P_1 correspond to θ counterclockwise and clockwise rotation of the Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathsf{diag}(1.5, 1))$. Furthermore, let the labeling functions be

$$y_0(x) = \begin{cases} 1 & \text{if } (-\sin(\theta), \cos(\theta)) \cdot x > 0 \\ 0 & \text{otherwise,} \end{cases} \quad y_1(x) = \begin{cases} 1 & \text{if } (\sin(\theta), \cos(\theta)) \cdot x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

The left-hand side of Figure 4.3 displays the level sets of P_0 as well as its labeling function.

In this synthetic example, $\epsilon_{\text{split,FER}}$ has a closed-form expression: $\epsilon_{\text{split,FER}} = 2\Pr(X \in \mathcal{A} \mid S = 0)$ where $\mathcal{A} \triangleq \{x = (x_1, x_2) \in \mathbb{R}^2 \mid y_1(x) = 1, x_2 < 0\}$. When $\theta = 0$, two groups share the same unlabeled distribution (i.e., $P_0 = P_1$) and the same labeling function (i.e., $y_0 = y_1$). Hence, there is no benefit of splitting classifiers: $\epsilon_{\text{split,FER}} = 0$. On the other hand, when $\theta = \pi/2$, two groups have the same unlabeled distribution but completely different labeling functions. Splitting classifiers achieves the most benefit: $\epsilon_{\text{split,FER}} = 0.5$.

By varying the values of θ and drawing 10k samples from each group, we compare the true values of $\epsilon_{\text{split,FER}}$ with the outputs from Algorithm 2 as well as other empirical approximations. Recall that Algorithm 2 requires a conditional distribution $\Pr(S = s \mid X = x)$ and the labeling functions y_0 and y_1. Since the conditional distribution and labeling functions are known in this synthetic example, we feed their explicit forms into Algorithm 2 for computing $\epsilon_{\text{split,FER}}$ (orange curve in Figure 4.3 Right). In practice, the conditional distribution and labeling functions are unknown, so we also train a Naive Bayes classifier [104] to approximate $\Pr(S = s \mid X = x)$ and two linear support-vector machine (SVM) classifiers [104] to approximate the labeling functions. By feeding these binary classifiers into Algorithm 2, another approximation of $\epsilon_{\text{split,FER}}$ is output (red curve in Figure 4.3 Right). Furthermore, we compute $\epsilon_{\text{split,FER}}$ empirically by training optimal group-blind and split classifiers via logistic regression, linear SVM, or Naive Bayes classifier. Computing the false error rate reduction leads to three empirical approximations of $\epsilon_{\text{split,FER}}$.

As shown in Figure 4.3, when Algorithm 2 has access to the explicit forms of $\Pr(S = s \mid X = x)$, y_0, and y_1, it accurately recovers $\epsilon_{\text{split,FER}}$. This is remarkable since even with the knowledge of the underlying distributions, it is unclear how to compute $\epsilon_{\text{split,FER}}$ directly from its definition. We also observe that Algorithm 2 applied to binary classifiers outputs more accurate approximation of $\epsilon_{\text{split,FER}}$ than the approximations produced by using logistic regression, linear SVM, or Naive Bayes classifier.

To summarize, we conclude that (i) when the underlying distribution is known, Algorithm 2 can produce the precise values of $\epsilon_{\text{split,FER}}$ and has convergence guarantees; (ii) when Algorithm 2 is fed with binary classifiers, it produces reliable approximation of $\epsilon_{\text{split,FER}}$; (iii) computing the FER-benefit-of-splitting empirically by training optimal classifiers could incur high approximation errors.

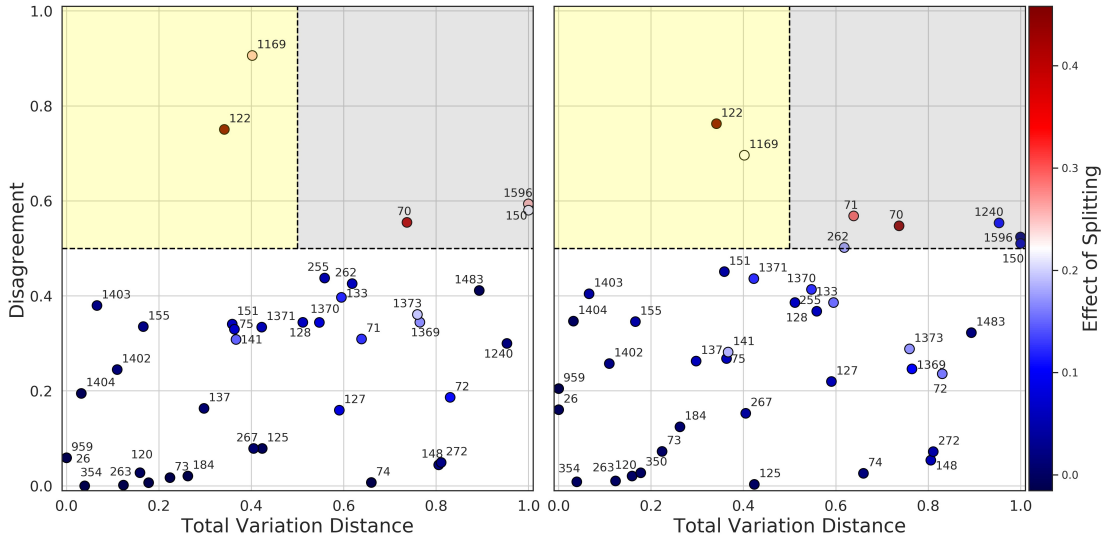

Figure 4.4: *We demonstrate how the effect of splitting classifiers is determined by the two factors: disagreement between optimal classifiers (y-axis) and total variation distance between unlabeled distributions (x-axis). We restrict both group-blind and split classifiers to logistic regression classifiers (left) or decision tree classifiers (right). Each dot represents a dataset in OpenML with color indicating the effect of splitting classifiers and texts indicating dataset ID. Theorem 5 reveal a taxonomy of splitting where splitting does not bring much benefit (white region); splitting brings the most benefit (yellow region); or splitting has undetermined effect (grey region).*

4.9.2 Datasets from OpenML

In Section 4.7, we analyzed the effect of splitting classifiers when both group-blind and split classifiers are restricted over the same hypothesis class. The bounds in Theorem 5 reveal two main factors that could determine this effect: disagreement between optimal classifiers and similarity between unlabeled distributions. Here we demonstrate how these two factors influence the effect of splitting through experiments on 40 real-world datasets, collected from OpenML [272].

Setup. We preprocess all 40 datasets by adopting the procedure described in [86]. All categorical features are transformed into binary by assigning the most frequent object to 1 and the rest of the objects to 0. The first binary feature is selected as the group attribute and, hence, these datasets are "semi-synthetic". We truncate the datasets so that each group contains at most 10k data points. In each dataset, there are at least 8k data points per group, minimizing the effect of potential lack of samples per group.

Implementation. We obtain optimal split classifiers via training a logistic regression model with the `LIBLINEAR` solver [95], fitting the model by drawing samples from each group. Since an optimal

group-blind classifier is a minimizer of $\min_{h\in\mathcal{H}} \max_{w\in[0,1]} wL_0(h) + (1-w)L_1(h)$ where $L_s(h)$ is the loss of a classifier h on group $s \in \{0,1\}$, we solve this optimization approximately by considering its dual formula $\max_{w\in[0,1]} \min_{h\in\mathcal{H}} wL_0(h) + (1-w)L_1(h)$ and use 5-fold cross validation to tune the parameter w therein. Although this procedure of training group-blind classifier needs access to data points' group attribute, it does not violate group-blindness [183] because the output classifier does not use the group attribute as an input when deploying on new data. In addition to logistic regressions, we repeat this experiment by training decision tree classifiers with depth 7. The disagreement between optimal classifiers is calculated by applying the optimal split classifiers on each data point and computing the discrepancy. We estimate the total variation distance between unlabeled distributions by applying the procedures introduced in [152].

Result. In Figure 4.4, we illustrate the taxonomy of splitting delineated by our bounds in Theorem 5. We restrict the hypothesis class to be logistic regression (Figure 4.4 Left) or to be decision trees with depth 7 (Figure 4.4 Right). Each dot in the figures represents a dataset with its corresponding ID number in the OpenML dataset. The color captures the loss reduction by using the optimal split classifiers compared to deploying the optimal group-blind classifier (red means splitting has more benefit and blue means splitting does not bring much benefit). The location of each dot is determined by the two factors: disagreement between optimal classifiers (y-axis) and total variation distance between unlabeled distributions (x-axis).

- The upper bound in Theorem 5 indicates that splitting does not bring much benefit when the optimal classifiers are similar. As shown in Figure 4.4, all datasets which are below the horizontal dash line have small benefit by splitting classifiers (i.e., dots are blue).
- The lower bound in Theorem 5 indicates that splitting benefits model performance when the optimal classifiers are different and the unlabeled distributions are similar. As shown in Figure 4.4, there are two datasets (ID 122 and 1169) which are in the yellow region and they all achieve large benefit from splitting classifiers.
- When both the optimal classifiers and the unlabeled distributions are different, the effect of splitting classifiers can not be determined by the bounds in Theorem 5. As shown in Figure 4.4, the datasets in the grey region could have either large benefit by splitting classifiers or limited benefit. Furthermore, we have conjectured (see Section 4.7.1) that in this case a more complex

hypothesis class leads to less benefit from splitting classifiers. This is further evidenced in the experiments: when both group-blind and split classifiers are logistic regressions (Figure 4.4 Left), the datasets which are in the grey region all achieve non-trivial benefit by splitting classifiers. In contrast, when decision trees are used (Figure 4.4 Right), there are datasets (e.g., ID 1240) in the grey region which achieve a limited amount of benefit by splitting.

4.9.3 Real-world Datasets

Setup We aim to recover counterfactual distributions for different disparity metrics in Table 4.1. To this end, we consider processed versions of the `adult` dataset [21] and the ProPublica `compas` dataset [10].

For each dataset, we use:

- 30% of samples to train a classifier $h(\boldsymbol{x})$ to repair;
- 50% of samples to recover a counterfactual distribution via Algorithm 3;
- 20% of samples as a hold-out set to evaluate the performance of the repaired model.

We use ℓ_2-logistic regression to train a classifier $h(\boldsymbol{x})$ as well as the classifier $\hat{y}_0(\boldsymbol{x})$ that we use to estimate the influence functions in Algorithm 3. We tune the parameters and estimate the performance of each classifiers using a standard 10-fold CV setup.

Our setup assumes that the data used to train and repair the model are drawn from the same distribution, which may not be the case in settings with dataset shift. Our setup also differs from real-world settings in that we use 70% of the samples in each dataset to estimate the counterfactual distribution. In practice, however, we would use all available samples since we would be given the classifier to repair.

Discussion In Table 4.2, we show the effectiveness of preprocessors built to mitigate different kinds of disparity for the classifiers trained on the `adult` and `compas` datasets.

We build each preprocessor as follows. We first resample data from the target population according to Algorithm 3. This outputs a dataset of samples drawn from the counterfactual distribution. Next, we use the resampled dataset to produce an empirical estimate of the counterfactual distribution Q_X. This distribution is then used to obtain the preprocessor by solving a version of (4.25) with the cost function $C_{ij} = \|\boldsymbol{x}_i - \widetilde{\boldsymbol{x}}_j\|_2^2$.

Dataset	Metric	Target Group	Original Model			Repaired Model		Target Group AUC	
			Baseline Group	Target Group	**Disc. Gap**	Target Group	**Disc. Gap**	Before Repair	After Repair
`adult`	SP	Female	0.696	0.874	**0.178**	0.688	**-0.007**	0.895	0.758
`adult`	FNR	Female	0.478	0.639	**0.161**	0.483	**0.004**	0.895	0.880
`adult`	FPR	Male	0.021	0.119	**0.098**	0.023	**0.002**	0.829	0.714
`compas`	SP	White	0.514	0.594	**0.079**	0.533	**0.018**	0.704	0.667
`compas`	FNR	White	0.350	0.487	**0.137**	0.439	**0.088**	0.704	0.699
`compas`	FPR	Non-white	0.190	0.278	**0.087**	0.160	**-0.029**	0.732	0.680

Table 4.2: *Change in disparate impact for classification models for* `adult` *and* `compas` *when paired with a randomized preprocessor built to mitigate different kinds of disparity. Each row shows the value of a specific performance metric for the classifier over the target and baseline groups (e.g., SP, FNR, and FPR). The target group is defined as the group that attains the less favorable value of the performance metric. The preprocessor aims to reduce to difference in performance metric by randomly perturbing the input variables for individuals in the target group. We also include AUC to show the change in performance due to the randomized preprocessor. All values are computed using a hold-out sample that is not used to train the model or build the preprocessor.*

As shown, the approach reduces disparate impact in the target group, while having a minor effect on test accuracy at various decision points across the full ROC curve. Counterfactual distributions provide a way to scrutinize this mapping in greater detail. As shown in Table 4.3, one can visualize the differences between the observed distribution and counterfactual distribution to understand how the input variable distributions are altered to reduce disparity.

This kind of constrastive analysis may be helpful in understanding the factors that produce performance disparities in the first place. For example, the differences between the observed distribution P_0 and the counterfactual distribution Q_X could be used to identify prototypical samples [see e.g, 37, 157], or to score features in terms of their ability to produce disparities in the deployment population [2, 71].

4.10 Conclusion

Split classifiers should only be considered when it is ethical and legal to do so, and when it does not result in harm to any underlying group. Eliminating disparate treatment does not necessarily lead to a group-fair classifier. On the one hand, a group attribute could correlate with other proxy variables which are used for decision making [130, 288]. On the other hand, the group attribute can be an important feature for the prediction task [67, 160]. In the latter case, using a group-blind classifier for achieving treatment parity may lead to an unfavorable accuracy trade-off.

	Observed		Counterfactual		
	Female	Male	SP Female	FNR Female	FPR Male
Married	18%	63%	39%	23%	54%
Immigrant	10%	11%	11%	11%	12%
HighestDegree_is_HS	32%	32%	24%	28%	37%
HighestDegree_is_AS	7%	8%	9%	9%	6%
HighestDegree_is_BS	15%	18%	21%	17%	13%
HighestDegree_is_MSorPhD	6%	7%	13%	8%	5%
AnyCapitalLoss	3%	5%	8%	5%	4%
Age $\leq$ 30	39%	29%	29%	38%	35%
WorkHrsPerWeek<40	38%	17%	33%	37%	19%
JobType_is_WhiteCollar	34%	19%	36%	35%	15%
JobType_is_BlueCollar	5%	34%	4%	5%	39%
JobType_is_Specialized	23%	21%	29%	23%	20%
JobType_is_ArmedOrProtective	1%	2%	1%	1%	3%
Industry_is_Private	73%	69%	64%	69%	70%
Industry_is_Government	15%	12%	22%	17%	12%
Industry_is_SelfEmployed	5%	15%	8%	6%	13%

Table 4.3: *Counterfactual distributions produced using Algorithm 3 for a classifier on* `adult`*. We observe that different metrics produce different counterfactual distributions. By comparing the distribution of the target group with the counterfactual distribution, we can evaluate how the repaired classifier will perturb their features to reduce disparity.*

Motivated by the above discussion, we investigated the following fundamental question: when disparate treatment is allowed, is it beneficial to incorporate the group attribute as an input feature in order to improve a classifier's performance? Due to the bias-variance trade-off, in practice, the answer will depend on the number of training data and the complexity of the hypothesis class. In this chapter, we focused on an information-theoretic regime where the underlying data distribution is known—or infinitely many data points are available—and the hypothesis class is unrestricted. To evaluate the potential gain in average performance from allowing a classifier to exhibit disparate treatment, we compared split classifiers with group-blind classifiers and characterized precise conditions where splitting classifiers achieves the most benefit. Our results show that—in this narrow information-theoretic regime—splitting classifiers follows the non-maleficence principle and allows a data scientist to deploy more accurate and suitable models for each group.

Besides investigating the conditions for fair use of group attributes, we also introduced a new distributional paradigm to mitigate disparate impact by using group attributes. Our framework is based on counterfactual distributions, which can be efficiently computed given a fixed model and data from a population of interest. Specifically, we proposed a descent procedure to estimate a counterfactual distribution from data. We proved that the best (first-order) direction is the influence functions, which own closed-form expressions for common group fairness criteria. The estimated counterfactual distribution yields a preprocessor that can improve the model performance for the disadvantaged group.

Chapter 5

An Estimation-Theoretic View of Privacy

Let S denote a private variable to be hidden (e.g., political preference) and X be a useful variable that depends on S (e.g., movie ratings). Our goal is to disclose a realization of a random variable Y, produced from X through a randomized mapping $P_{Y|X}$ called the *privacy mechanism*. Here, S, X, and Y satisfy the Markov condition $S \to X \to Y$. We assume that an analyst will provide some utility based on an observation of Y (e.g., movie recommendations), while potentially trying to estimate S from Y.

We derive privacy-utility trade-offs (PUTs) when both privacy and utility are measured in terms of the mean-squared error of reconstructing functions of S and X from an observation of Y. We analyze three related scenarios: (i) an *aggregate* setting, where certain functions of X can be, on average, reconstructed from the disclosed variable while controlling the MMSE of estimating functions of S and $P_{S,X}$ is known to the privacy mechanism designer, (ii) a *composite* setting, where specific functions of S and X have different privacy/utility reconstruction requirements and $P_{S,X}$ is known to the privacy mechanism designer, and (iii) a *restricted-knowledge* setting, where $P_{S,X}$ is unknown, but the correlation between a target function to be hidden and a set of functions which are known to be hard to infer from the disclosed variable is given.

5.1 Overview and Main Contributions

We present the outline of this chapter and a summary of our main contributions.

Aggregate PUTs. We start by studying the problem of limiting an untrusted party's ability to estimate functions of S given an observation of Y, while controlling for the MMSE of reconstructing functions of X given Y. Here, privacy and utility are measured in terms of the χ^2-information between S and Y and the χ^2-information between X and Y, denoted by $\chi^2(S;Y)$ and $\chi^2(X;Y)$ (cf. (5.2)), respectively. We introduce the χ^2-privacy-utility function in Section 5.4. Bounds of this function are presented in Theorem 7. In particular, the upper bound is cast in terms of the PICs of $P_{S,X}$ and provides an interpretation of the trade-off between privacy and utility that goes beyond simply using maximal correlation. We also prove that the upper bound is achievable in the high-privacy regime in Theorem 8.

Composite PUTs. χ^2-based metrics guarantee privacy and utility in a uniform sense, capturing the aggregate mean-squared error of estimating *any* functions of the private and the useful variables. However, in many applications, specific functions of S and X that should be hidden/revealed are known *a priori*. This knowledge enables a more refined design of privacy mechanisms that specifically target these functions. We explore this finer-grained approach in Section 5.5, and propose a PIC-based convex program for computing privacy mechanisms within this setting. We demonstrate the practical feasibility of the convex programs through two numerical experiments in Section 5.7, deriving privacy mechanisms for a synthetic dataset and a real-world dataset. In the latter case, we approximate $P_{S,X}$ using its empirical distribution.

Restricted knowledge of the distribution. The aforementioned aggregate and composite PUTs require knowledge of the joint distribution $P_{S,X}$. In Section 5.6, we forgo this assumption, and study a simpler setting where $S = \phi(X)$ (i.e., the private variable is a function of the data) and the correlation between $\phi(X)$ and a set of functions (composed with the data) $\{\phi_j(X)\}_{j=1}^m$ is given. In practice, $\phi(X)$ may be a sensitive feature of the data X, and $\{\phi_j(X)\}_{j=1}^m$ is a collection of other features from which $\mathbb{E}\left[\phi(X)\phi_j(X)\right]$ can be accurately estimated.

Our goal here is to derive lower bounds on the MMSE of estimating a real-valued function of X, namely $\phi(X)$, from Y for any privacy mechanism $P_{Y|X}$. These bounds are cast in terms of the

MMSE of estimating $\phi_j(X)$ from Y and the correlation between $\phi(X)$ and $\{\phi_j(X)\}_{j=1}^m$. This leads to a converse result in Theorem 10: if the MMSE of estimating $\phi_j(X)$ from Y is large and $\phi(X)$ is strongly correlated with $\phi_j(X)$, then the MMSE of estimating $\phi(X)$ from Y will also be large and privacy is assured in an estimation-theoretic sense. The inverse result is straightforward: if $\phi(X)$ and $\phi_j(X)$ are strongly correlated and $\phi_j(X)$ can be reliably reconstructed from Y, then $\phi(X)$ can also be reliably estimated from Y. This intuitive trade-off is at the heart of the estimation-theoretic view of privacy, and demonstrates that no function of X can remain private whilst other strongly correlated functions are revealed through Y. The results in Section 5.6 make this intuition mathematically precise.

5.2 Related Works

Currently, the most adopted definition of privacy is differential privacy [87], which enables queries to be computed over a database while simultaneously ensuring privacy of individual entries of the database. Information-theoretic quantities, such as Rényi divergence, can be used to relax the definition of differential privacy [195]. Fundamental bounds on composition of differentially private mechanisms were given by Kairouz *et al.* [148]. Recently, a new privacy framework called Pufferfish [155] was developed for creating customized privacy definitions.

Several papers, such as Sankar *et al.* [240], Calmon and Fawaz [51], Asoodeh *et al.* [15], and Makhdoumi *et al.* [187], have studied information disclosure with privacy guarantees through an information-theoretic lens. For example, Sankar *et al.* [240] characterized PUTs in large databases using tools from rate-distortion theory. Calmon and Fawaz [51] used expected distortion and mutual information to measure utility and privacy, respectively, and characterized the PUT as an optimization problem. Makhdoumi *et al.* [187] introduced the privacy funnel, where both privacy and utility are measured in terms of mutual information, and showed its connection with the information bottleneck [263]. The PUT was also explored in [186] and [229] using mutual information as a privacy metric.

Other quantities from the information-theoretic literature have been used to quantify privacy and utility. For example, Asoodeh *et al.* [17] and Calmon *et al.* [53] used estimation-theoretic tools to characterize fundamental limits of privacy. Liao *et al.* [179, 180] explored the PUT within a hypothesis testing framework. Issa *et al.* [132, 133] introduced maximal leakage as an information leakage metric. There is also significant recent work in information-theoretic privacy in the context of network

secrecy. For example, Li and Oechtering [176] proposed a new privacy metric based on distributed Bayesian detection which can inform privacy-aware system design. Recently, Tripathy *et al.* [264] and Huang *et al.* [126] used adversarial networks for designing privacy mechanisms that navigate the PUT. Takbiri *et al.* [259] considered obfuscation and anonymization techniques and characterized the conditions required to obtain perfect privacy.

We use the definition of PICs presented in [53], but note that the PICs predate [53] by many decades (e.g., [48, 109, 112, 124, 230, 241, 295]). Recently, Huang *et al.* [127] considered the PICs by analyzing the "divergence transition matrix" [127, Eq. 2]. Specifically, there are different directions of local perturbation [41] of input distribution and the direction which leads to the greatest influence of the output distribution of a noisy channel can be identified [127] by specifying the singular vector decomposition of the divergence transition matrix. In the follow-on work, Huang *et al.* [128] used the divergence transition matrix in the context of feature selection. The singular values of the divergence transition matrix are exactly the square root of the PICs considered here, and are also related to the singular values of the conditional expectation operator, as also noted by Makur and Zheng [188] and originally by Witsenhausen [295] and others [48]. We build on these prior works by using the PICs for quantifying privacy-utility trade-offs.

5.3 Preliminaries

Notation. For a positive integer n, we define the set $[n] \triangleq \{1, \cdots, n\}$. We denote matrices in bold capital letters (e.g., $\mathbf{P}$) and vectors in bold lower-case letters (e.g., $\mathbf{p}$). For a vector $\mathbf{p}$, $\mathsf{diag}(\mathbf{p})$ is defined as the matrix with diagonal entries equal to $\mathbf{p}$ and all other entries equal to 0. The span of a set $\mathcal{V}$ of vectors is

$$\mathsf{span}(\mathcal{V}) \triangleq \left\{ \sum_{i=1}^{k} \lambda_i \mathbf{v}_i \;\middle|\; k \in \mathbb{N}, \mathbf{v}_i \in \mathcal{V}, \lambda_i \in \mathbb{R} \right\}.$$

The dimension of a linear span is denoted by $\mathsf{dim}(\mathsf{span}(\mathcal{V}))$. Let $(\mathcal{A}, d_{\mathcal{A}})$ and $(\mathcal{B}, d_{\mathcal{B}})$ be two metric spaces. We say that $f : \mathcal{A} \to \mathcal{B}$ is Lipschitz over $\mathcal{A}' \subseteq \mathcal{A}$ if there exists $L \geq 0$ such that, for all $a_1, a_2 \in \mathcal{A}'$,

$$d_{\mathcal{B}}(f(a_1), f(a_2)) \leq L d_{\mathcal{A}}(a_1, a_2). \tag{5.1}$$

For a random variable U with probability distribution P_U, we denote

$$P_{U\,\min} \triangleq \inf\{P_U(u) \mid u \in \mathcal{U}\},$$

where $\mathcal{U}$ is the support set of U. The MMSE of estimating U given V is

$$\mathsf{mmse}(U|V) \triangleq \min_{U \to V \to \hat{U}} \mathbb{E}\left[(U-\hat{U})^2\right] = \mathbb{E}\left[(U-\mathbb{E}\left[U|V\right])^2\right].$$

The χ^2-*information* between two random variables U and V is defined as

$$\chi^2(U;V) \triangleq \mathbb{E}\left[\left(\frac{P_{U,V}(U,V)}{P_U(U)P_V(V)}\right)\right] - 1. \tag{5.2}$$

Let P_U and Q_U be two probability distributions taking values in the same discrete and finite set $\mathcal{U}$. We denote $||P_U - Q_U||_1 \triangleq \sum_{u\in\mathcal{U}} |P_U(u) - Q_U(u)|$.

Throughout this chapter, we assume all random variables are discrete with finite support sets.

5.4 Aggregate PUTs: The Chi-Square-Privacy-Utility Function

We start our analysis by adopting χ^2-information as a measure of both privacy and utility. As seen in the previous section, $\chi^2(S;Y) = \sum_{i=1}^{d} \lambda_i(S;Y)$, where $d = \min\{|\mathcal{S}|, |\mathcal{Y}|\} - 1$. If $\chi^2(S;Y) < 1$, then, from characterization 2 in Theorem 1, the MMSE of reconstructing *any* zero-mean, unit-variance function of S given Y is lower bounded by $1 - \chi^2(S;Y)$, i.e., all functions of S cannot be reconstructed with small MMSE given an observation of Y. Note that this argument also holds true when we replace χ^2-information with the maximal correlation. In fact, in the high privacy regime, the PUT under χ^2-information is essentially equivalent to the PUT when both privacy and utility are measured using maximal correlation. We make this intuition precise at the end of this section. When $1 \leq \chi^2(S;Y)$, certain private functions, on average, may be estimated from Y but, in general, most private functions are still kept in secret. Analogously, when $\chi^2(X;Y)$ is large, certain functions of X can be, on average, reconstructed (i.e., estimated) with small MMSE from Y. We demonstrate next that the PICs play a central role in bounding the PUT in this regime.

We first introduce the χ^2-privacy-utility function. This function captures how well an analyst can reconstruct functions of the useful variable X while restricting the analyst's ability to estimate functions of the private variable S.

Definition 15. For a given joint distribution $P_{S,X}$ and $0 \leq \epsilon \leq \chi^2(S;X)$, we define the χ^2-privacy-

utility (trade-off) function as

$$F_{\chi^2}(\epsilon; P_{S,X}) \triangleq \sup_{P_{Y|X} \in \mathcal{D}(\epsilon; P_{S,X})} \chi^2(X;Y),$$

where $\mathcal{D}(\epsilon; P_{S,X}) \triangleq \{P_{Y|X} \mid S \to X \to Y, \chi^2(S;Y) \leq \epsilon\}$.

It has been proved in [125, 294] that there is always a privacy mechanism $P_{Y|X}$ which achieves the supremum in $F_{\chi^2}(\epsilon; P_{S,X})$ using at most $|\mathcal{X}|+1$ symbols (i.e., $|\mathcal{Y}| \leq |\mathcal{X}|+1$). The following lemma gives an alternative way to compute the χ^2-information, in the discrete, finite setting.

Lemma 11. *Suppose $S \to X \to Y$. Then*

$$\chi^2(X;Y) = \mathsf{tr}(\mathbf{A}) - 1, \tag{5.3}$$

$$\chi^2(S;Y) = \mathsf{tr}(\mathbf{BA}) - 1, \tag{5.4}$$

where, using (2.15),

$$\mathbf{A} \triangleq \mathbf{Q}_{X,Y}\mathbf{Q}_{X,Y}^T, \ \mathbf{B} \triangleq \mathbf{Q}_{S,X}^T\mathbf{Q}_{S,X}.$$

Proof. See Appendix C.1.1. □

The following lemma characterizes some properties of the χ^2-privacy-utility function.

Lemma 12. *For a given joint distribution $P_{S,X}$, the χ^2-privacy-utility function $F_{\chi^2}(\epsilon; P_{S,X})$ is a concave function in ϵ. Furthermore, $\epsilon \to \frac{1}{\epsilon}F_{\chi^2}(\epsilon; P_{S,X})$ is a non-increasing mapping.*

Proof. See Appendix C.1.2. □

The χ^2-privacy-utility function has a simple upper bound,

$$F_{\chi^2}(\epsilon; P_{S,X}) \leq \epsilon + |\mathcal{X}| - 1 - \chi^2(S;X), \tag{5.5}$$

which follows immediately from the data-processing inequality:

$$\chi^2(S;X) + \chi^2(X;Y) \leq \chi^2(S;Y) + \chi^2(X;X). \tag{5.6}$$

We derive an upper bound for the χ^2-privacy-utility function that significantly improves (5.5) by using properties of the PICs. The bound is piecewise linear, where each piece has a slope given in terms of a PIC of $P_{S,X}$. Intuitively, this bound corresponds to the privacy mechanism $P_{Y|X}$ that

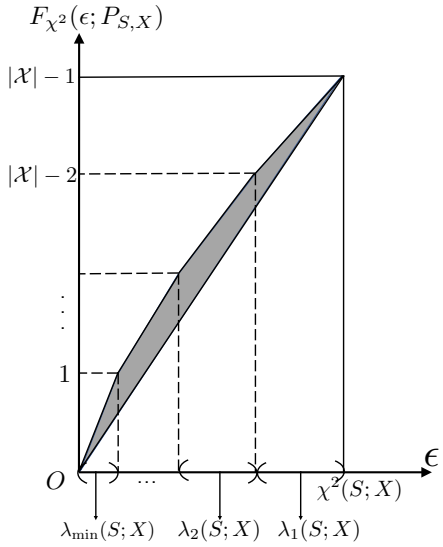

Figure 5.1: *Piecewise linear upper bound and lower bound for the χ^2-privacy-utility function when $\delta(P_{S,X})$, defined in (2.16), is positive.*

achieves the best PUT if $P_{Y|X}$ was not constrained to be non-negative. We also provide a lower bound that follows directly from the concavity of the χ^2-privacy-utility function. These bounds are illustrated in Fig. 5.1.

Definition 16. For $t_i \in [0,1]$ $(i \in [n])$, $0 \leq \epsilon \leq \sum_{i\in[n]} t_i$, and $n \leq m$, $G_\epsilon^m(t_1,...,t_n)$ is defined as

$$G_\epsilon^m(t_1,...,t_n) \triangleq \max\left\{ \sum_{i=1}^m x_i \;\middle|\; (x_1,...,x_m) \in \mathcal{D}_\epsilon^m(t_1,...,t_n) \right\},$$

where

$$\mathcal{D}_\epsilon^m(t_1,...,t_n) \triangleq \left\{ (x_1,...,x_m) \;\middle|\; \sum_{i=1}^n t_i x_i \leq \epsilon, x_i \in [0,1], i \in [m] \right\}.$$

For fixed m and t_i $(i \in [n])$, $G_\epsilon^m(t_1,...,t_n)$ is a piecewise linear function with respect to ϵ and can be expressed in closed-form (cf. Appendix C.1.3).

Theorem 7. *For the χ^2-privacy-utility function $F_{\chi^2}(\epsilon; P_{S,X})$ introduced in Definition 15 and $\epsilon \in [0, \chi^2(S;X)]$,*

$$\frac{|\mathcal{X}|-1}{\chi^2(S;X)}\epsilon \leq F_{\chi^2}(\epsilon; P_{S,X}) \leq G_\epsilon^{|\mathcal{X}|-1}(\lambda_1(S;X),...,\lambda_d(S;X)),$$

where $d \triangleq \min\{|\mathcal{S}|, |\mathcal{X}|\} - 1$ and $\lambda_1(S;X),...,\lambda_d(S;X)$ are the PICs of $P_{S,X}$.

Proof. See Appendix C.1.3. □

Remark 7. The upper bound for the χ^2-privacy-utility function given in Theorem 7 can also be

proved by, for example, combining Theorem 4 in [280] with properties of the PICs.

We now illustrate the piecewise linear upper bound. Recall that the PIC decomposition of $P_{S,X}$ results in a set of basis functions $\mathcal{P} \triangleq \{f_1(S), \cdots, f_d(S)\}$, with corresponding MMSE estimators $\mathcal{U} \triangleq \{g_1(X), \cdots, g_d(X)\}$. Consider the following intuition for designing a sequence of privacy mechanisms. The first mechanism enables the function $g_d(X)$ to be reliably estimated from Y while keeping all other functions in $\mathcal{U}$ secret. In this case, the utility is one, since exactly one zero-mean, unit-variance function of X can be recovered from Y. The privacy leakage is $\lambda_d(S;X)$, since using $g_d(X)$ to estimate the private function $f_d(S)$ has mean-squared error $1 - \lambda_d(S;X)$. Following the same procedure, the second privacy mechanism allows only $g_d(X)$ and $g_{d-1}(X)$ to be recovered from the disclosed variable and so on. This sequence of privacy mechanisms corresponds to the breakpoints of the upper bound. Note that such privacy mechanisms may not be feasible — hence the upper bound.

Note that $F_{\chi^2}(0; P_{S,X})$ characterizes the maximal aggregate MMSE of estimating useful functions while guaranteeing perfect privacy. Here perfect privacy means that no zero-mean, unit-variance function of S can be reconstructed from Y. If the value of $F_{\chi^2}(0; P_{S,X})$ is known, a better lower bound can be obtained from the concavity of $F_{\chi^2}(\epsilon; P_{S,X})$ as

$$\frac{|\mathcal{X}| - 1 - F_{\chi^2}(0; P_{S,X})}{\chi^2(S;X)}\epsilon + F_{\chi^2}(0; P_{S,X}) \leq F_{\chi^2}(\epsilon; P_{S,X}). \tag{5.7}$$

When $S = X$, then $\chi^2(S;X) = |\mathcal{X}| - 1$ and $F_{\chi^2}(\epsilon; P_{S,X}) = \epsilon$. Following from Definition 16 and noticing that all PICs of $P_{S,X}$ are 1, the upper bound and the lower bound for the χ^2-privacy-utility function in Theorem 7 are both ϵ, which is equal to $F_{\chi^2}(\epsilon; P_{S,X})$. In this sense, the upper bound and lower bound given in Theorem 7 are sharp. We investigate the tightness of the upper bound through numerical example in Section 5.7.1.

The following corollary of Lemma 12 and Theorem 7 shows that the χ^2-privacy-utility function is strictly increasing with respect to ϵ.

Corollary 3. *For a given joint distribution $P_{S,X}$, the mapping $\epsilon \to F_{\chi^2}(\epsilon; P_{S,X})$ is strictly increasing for $\epsilon \in [0, \chi^2(S;X)]$.*

Proof. See Appendix C.1.4. □

We denote

$$\partial\mathcal{D}(\epsilon; P_{S,X}) \triangleq \{P_{Y|X} \mid S \to X \to Y, \chi^2(S;Y) = \epsilon\}.$$

By Corollary 3, $F_{\chi^2}(\epsilon; P_{S,X})$ is strictly increasing. Therefore,

$$F_{\chi^2}(\epsilon; P_{S,X}) = \max_{P_{Y|X} \in \partial\mathcal{D}(\epsilon; P_{S,X})} \chi^2(X;Y). \tag{5.8}$$

By Corollary 7 in [53], when $\delta(P_{S,X}) = 0$, defined in (2.16), then $F_{\chi^2}(0; P_{S,X}) > 0$ (i.e., there exists a privacy mechanism that allows the disclosure of a non-trivial amount of useful functions while guaranteeing perfect privacy). On the other hand, when $\delta(P_{S,X}) > 0$, then $F_{\chi^2}(0; P_{S,X}) = 0$. The following theorem shows that when $\delta(P_{S,X}) > 0$, the upper bound of $F_{\chi^2}(\epsilon; P_{S,X})$ in Theorem 7 is achievable around zero, implying that the upper bound is tight around zero. The proof of this theorem also provides a specific way to construct an optimal privacy mechanism (i.e., achieves the upper bound in Theorem 7).

Theorem 8. *Suppose $\delta(P_{S,X}) > 0$ and $P_{X\,\min} > 0$. Then there exists Y such that $S \to X \to Y$, $\chi^2(X;Y) = P_{X\,\min}$ and $\chi^2(S;Y) = P_{X\,\min}\lambda_{\min}(S;X)$.*

Proof. See Appendix C.1.5. □

When $\delta(P_{S,X}) > 0$ and $P_{X\,\min} > 0$, then $F_{\chi^2}(\hat{\epsilon}; P_{S,X}) = P_{X\,\min}$ where $\hat{\epsilon} = P_{X\,\min}\lambda_{\min}(S;X)$. Since $(\hat{\epsilon}, P_{X\,\min})$ is a point on the upper bound of the χ^2-privacy-utility function given in Theorem 7, Theorem 8 shows that, in this case, the upper bound is achievable in the high-privacy region. We remark that the local behavior of privacy-utility functions in high-privacy region and high-utility region has been studied in the context of strong data processing inequalities (e.g., [54, 189] and the references therein).

Connections with Maximal Correlation

Maximal correlation has previously been considered as a privacy measure in [16, 17, 55, 174]. In particular, it has been proved [55] that when $\rho_m(S;Y)$ is small, then $\Pr(S \neq \hat{S})$ can be lower bounded for any $\hat{S} = h(Y)$. We show in Corollary 4 that, in the high privacy regime, the privacy-utility function under maximal correlation possesses similar properties to $F_{\chi^2}(\epsilon; P_{S,X})$. However, when $\rho_m(S;Y)$ is large, say $\rho_m(S;Y) = 1$, it is unclear whether one private function or several private functions can be recovered from the disclosed variable. In contrast, χ^2-information can distinguish

between these two cases and quantifies how many private functions, on average, can be reconstructed from the disclosed variable. For example, a user might be comfortable revealing that his/her age is above a certain threshold, but not the age itself. In this case, the privacy leakage measured by maximal correlation is one since there is a function of age which can be recovered from the disclosed variable. Thus, maximal correlation cannot distinguish between the cases where only one function of S and S itself can be estimated from the disclosed data. We will revisit this example in the next section and show how to design privacy mechanisms using PICs which target specific private functions and useful functions. Finally, we provide an example showing the limitation of maximal correlation as a utility measure.

Definition 17. For a given joint distribution $P_{S,X}$ and $0 \leq \epsilon \leq \rho_m(S;X)$, we define the maximal-correlation-privacy-utility (trade-off) function as

$$F_{\rho_m}(\epsilon; P_{S,X}) \triangleq \sup_{P_{Y|X} \in \mathcal{D}_{\rho_m}(\epsilon; P_{S,X})} \rho_m(X;Y),$$

where $\mathcal{D}_{\rho_m}(\epsilon; P_{S,X}) \triangleq \{P_{Y|X} \mid S \to X \to Y, \rho_m(S;Y) \leq \epsilon\}$.

The next corollary follows from the same proof techniques used in Theorem 7 and Theorem 8.

Corollary 4. *For a given joint distribution $P_{S,X}$ and $\epsilon \in [0, \rho_m(S;X)]$, if $\delta(P_{S,X}) > 0$, then $F_{\rho_m}(\epsilon; P_{S,X}) \leq \epsilon / \sqrt{\lambda_{\min}(S;X)}$. Furthermore, if $P_{X\min} > 0$, then there exists Y such that $S \to X \to Y$, $\rho_m(X;Y) = \sqrt{P_{X\min}}$ and $\rho_m(S;Y) = \sqrt{P_{X\min}\lambda_{\min}(S;X)}$.*

We illustrate the limitation of the maximal correlation as a utility measure through the following example.

Example 3. Let $\mathcal{S} = \{-1,1\}^n$ and $\mathcal{X} = \{-1,1\}^n$, and X^n be the result of passing S^n through a memoryless binary symmetric channel with crossover probability $\epsilon < 1/2$. We assume that S^n is composed of n uniform and i.i.d. bits. For $\mathcal{A} \subseteq [n]$, let $Y = \prod_{i \in \mathcal{A}} X_i$. In this case, one can show that $\rho_m(S^n;Y) = (1-2\epsilon)^{|\mathcal{A}|}$ and $\rho_m(X^n;Y) = 1$. If $|\mathcal{A}|$ is an increasing function of n, then $\rho_m(S^n;Y) \to 0$ as $n \to \infty$. In other words, we can disclose a function of X^n achieving nearly perfect privacy and utility as measured by $\rho_m(S^n;Y)$ and $\rho_m(X^n;Y)$, respectively, with large $|\mathcal{A}|$ and n. However, as n increases, the basis of functions in $\mathcal{L}_2(P_{X^n})$ will increase exponentially, and revealing only one function may not be enough for achieving utility. The crux of the limitation is that maximal correlation only takes into account the most reliably estimated function. The χ^2-information

overcomes this limitation by capturing all possible real-valued functions of X^n that can be recovered from Y. In particular, if $\chi^2(X^n;Y) = |\mathcal{X}| - 1$, then all zero-mean finite-variance functions of X^n can be reconstructed from Y. We will revisit this example again in Section 5.6 and Section 5.7.

5.5 Composite PUTs: A Convex Program for Computing Privacy Mechanisms

In the previous section, we studied χ^2-based metrics for both privacy and utility. The optimization problem in the definition of χ^2-privacy-utility function (Definition 15) is non-convex. Next, we provide a convex program for designing privacy mechanisms by adding more stringent constraints on privacy and utility.

More specifically, we explore an alternative, finer-grained approach for measuring both privacy and utility based on PICs (recall that χ^2-information is the sum of all PICs). This approach has a practical motivation, since oftentimes there are specific well-defined features (functions) of the data (realizations of a random variable) that should be hidden or disclosed. For example, a user may be willing to disclose that they prefer documentaries over action movies, but not exactly which documentary they like. More abstractly, we consider the case where certain known functions should be disclosed (utility), whereas others should be hidden (privacy). This is a finer-grained setting than the one used in the last section, since χ^2-information captures the aggregate reconstruction error across all zero-mean, unit-variance functions.

We denote the set of functions to be disclosed as

$$\mathcal{U}(X) \triangleq \{u_i : \mathcal{X} \to \mathbb{R} \mid \mathbb{E}\left[u_i(X)\right] = 0, ||u_i(X)||_2 = 1, i \in [n]\},$$

and the set of functions to be hidden as

$$\mathcal{P}(S) \triangleq \{s_i : \mathcal{S} \to \mathbb{R} \mid \mathbb{E}\left[s_i(S)\right] = 0, ||s_i(S)||_2 = 1, i \in [m]\}.$$

Our goal is to find the privacy mechanism $P_{Y|X}$ such that $S \to X \to Y$ and Y satisfies the following privacy-utility constraints:

1. **Utility constraints:** $\max\{\mathsf{mmse}(u_i(X)|Y)\}_{i\in[n]} \leq \Delta$ and $X \sim Y$.

2. **Privacy constraints:** $\mathsf{mmse}(s_i(S)|Y) \geq \theta_i, i \in [m]$.

Note that the utility constraint $X \sim Y$ implies that the disclosed variable follows the same distribution as the useful variable. The practical motivation for adding this constraint is to enable Y to preserve overall population statistics about X, while hiding information about individual samples. This assumption also enables the problem of finding the optimal privacy mechanism to be formulated as a convex program, described next.

We follow two steps – projection[1] and optimization – to find the privacy mechanism. Private functions are projected to a new set of functions based on the useful variable in the first step. Then a PIC-based convex program is proposed in order to find the privacy mechanism.

5.5.1 Projection

As a first step, we project (i.e., compute the conditional expectation) all private functions to the useful variable and obtain a new set of functions:

$$\mathcal{P}(X) \triangleq \left\{ \hat{s}_i(x) \triangleq \frac{\mathbb{E}\left[s_i(S)|X=x\right]}{||\mathbb{E}\left[s_i(S)|X\right]||_2} \;\middle|\; i \in [m] \right\}.$$

It is worth noting that, after the projection, the obtained privacy mechanism may not be an optimal solution to the original problem since the privacy constraints become stricter (see Lemma 13). Nonetheless, the advantage of this projection is twofold. First, it can significantly simplify the optimization program, since after the projection all functions are cast in terms of the useful variable alone. Second, the private variable is not needed as an input to the optimization after the projection. Therefore, the party that solves the optimization does not need access to the private data directly, further guaranteeing the safety of the sensitive information. The following lemma proves that privacy guarantees cast in terms of the projected functions still hold for the original functions.

Lemma 13. *Assume $S \to X \to Y$. For any function $f : \mathcal{S} \to \mathbb{R}$, if $\mathbb{E}\left[f(S)\right] = 0$ and $||\mathbb{E}\left[f(S)|X\right]||_2 \neq 0$, we have $\mathbb{E}\left[\mathbb{E}\left[f(S)|X\right]\right] = 0$ and*

$$\mathsf{mmse}\left(\frac{f(S)}{||f(S)||_2}\middle|Y\right) \geq \mathsf{mmse}\left(\frac{\mathbb{E}\left[f(S)|X\right]}{||\mathbb{E}\left[f(S)|X\right]||_2}\middle|Y\right).$$

Proof. See Appendix C.2.1. □

By Lemma 13, $\mathsf{mmse}(s_i(S)|Y) \geq \mathsf{mmse}(\hat{s}_i(X)|Y)$. Therefore, if the new set of functions satisfies the privacy constraints (i.e., $\mathsf{mmse}(\hat{s}_i(X)|Y) \geq \theta_i$), the original set of functions also satisfies the

[1] We call this step as projection because of the geometric interpretation of conditional expectation (see, e.g., [82]).

privacy constraints (i.e., $\mathsf{mmse}(s_i(S)|Y) \geq \theta_i$).

5.5.2 Optimization

We introduce next a PIC-based convex program to find the privacy mechanism $P_{Y|X}$. First, we construct a matrix $\mathbf{F}$ given by $(\mathbf{f}_0, \mathbf{f}_1, ..., \mathbf{f}_{|\mathcal{X}|-1})$ such that

$$\mathbf{F}^T \mathbf{D}_X \mathbf{F} = \mathbf{I}, \tag{5.9}$$

$$\mathsf{span}(\{\mathbf{f}_0, ..., \mathbf{f}_{n'}\}) = \mathsf{span}(\{\mathbf{f}_0, \mathbf{u}_1, ..., \mathbf{u}_n\}), \tag{5.10}$$

where $\mathbf{f}_0 \triangleq (1, ..., 1)^T$, $\mathbf{f}_i \triangleq (f_i(1), ..., f_i(|\mathcal{X}|))^T$, $\mathbf{u}_i \triangleq (u_i(1), ..., u_i(|\mathcal{X}|))^T$, and

$$n' \triangleq \mathsf{dim}(\mathsf{span}(\{\mathbf{f}_0, \mathbf{u}_1, ..., \mathbf{u}_n\})) - 1.$$

Following from (5.9), $\{f_k(x) \mid k = 0, ..., |\mathcal{X}| - 1\}$ is a basis of $\mathcal{L}_2(P_X)$ and, consequently, the functions $\hat{s}_i(x)$ can be decomposed as

$$\hat{s}_i(x) = \sum_{k=0}^{|\mathcal{X}|-1} \alpha_{i,k} f_k(x). \tag{5.11}$$

Since $\mathbb{E}[\hat{s}_i(X)] = 0$, then $\alpha_{i,0} = 0$. Similarly, since $\mathbf{u}_i \in \mathsf{span}(\{\mathbf{f}_0, ..., \mathbf{f}_{n'}\})$ and $\mathbb{E}[u_i(X)] = 0$, we have

$$u_i(x) = \sum_{k=1}^{n'} \beta_{i,k} f_k(x). \tag{5.12}$$

If $\mathbf{P}_{X,Y} = \mathbf{D}_X \mathbf{F} \mathbf{\Sigma} \mathbf{F}^T \mathbf{D}_X$ with $\mathbf{\Sigma} = \mathsf{diag}(1, \sigma_1, ..., \sigma_{|\mathcal{X}|-1})$ is a feasible joint distribution matrix (i.e., non-negative entries and all entries add to 1), then, following from Theorem 1,

$$\mathsf{mmse}(\hat{s}_i(X)|Y) = 1 - \sum_{k=1}^{|\mathcal{X}|-1} \alpha_{i,k}^2 \sigma_k^2,$$

$$\mathsf{mmse}(u_i(X)|Y) = \sum_{k=1}^{n'} \beta_{i,k}^2 (1 - \lambda_k(X;Y)) \leq 1 - \min_{k \in [n']} \lambda_k(X;Y) = 1 - \left(\min_{k \in [n']} \sigma_k \right)^2.$$

Therefore, the design of the privacy mechanism $P_{Y|X}$ with privacy-utility constraints is equivalent to solving the PIC-based convex program in Formulation 5.5.2. In this case, the objective function is chosen as $\mathsf{obj}(\sigma_1, ..., \sigma_{n'}) = \min\{\sigma_1, ..., \sigma_{n'}\}$[2].

[2]This is a convex program since one can add a constraint $\sigma_i \geq \sigma$ ($i \in [n']$) and maximize σ.

$$\max \ \mathsf{obj}(\sigma_1, ..., \sigma_{n'}) \tag{5.13}$$

$$\text{s.t.} \sum_{k=1}^{|\mathcal{X}|-1} \alpha_{i,k}^2 \sigma_k^2 \leq 1 - \theta_i \ (i = 1, ..., m), \tag{5.14}$$

$$0 \leq \sigma_i \leq 1 \ (i = 1, ..., |\mathcal{X}| - 1), \tag{5.15}$$

$$\mathbf{\Sigma} = \mathsf{diag}(1, \sigma_1, ..., \sigma_{|\mathcal{X}|-1}), \tag{5.16}$$

$$\mathbf{P}_{X,Y} = \mathbf{D}_X \mathbf{F} \mathbf{\Sigma} \mathbf{F}^T \mathbf{D}_X, \tag{5.17}$$

$$\mathbf{P}_{X,Y} \text{ has non-negative entries.} \tag{5.18}$$

Formulation 5.5.2: PIC-based convex program. Here σ_i $(i = 1, ..., |\mathcal{X}| - 1)$ and θ_i $(i = 1, ..., m)$ are variables and privacy parameters, respectively.

The objective function $\min\{\sigma_1, ..., \sigma_{n'}\}$ maximizes the worst-case utility over all useful functions. On the other hand, we can choose the objective function to be a weighted sum $\sum_{i=1}^{n'} a_i \sigma_i$. Although maximizing the weighted sum is not equivalent to the desired utility constraints, this new formulation allows more flexibility in the optimization. In particular, this enables useful functions which do not highly correlate with private functions to achieve better utility, in terms of mean-squared error, under the same privacy constraints. Furthermore, the weights can be used to prioritize the reconstruction of certain useful functions.

The previous convex programs can be numerically solved by standard methods (e.g., CVXPY [75]). Note that when all useful functions and private functions are based on the same random variable, we can use optimization without projection. We defer the numerical results to Section 5.7, where we derive privacy mechanisms for a synthetic dataset and a real-world dataset using tools introduced in this section.

5.6 Lower Bounds for MMSE with Restricted Knowledge of the Data Distribution

So far we have assumed the information-theoretic setting where the probability distribution $P_{S,X}$ is known to the privacy mechanism designer beforehand. In this section, we forgo this assumption and consider a setting where $S = \phi(X)$ and the correlation between $\phi(X)$ and a set of functions

(composed with the data) $\{\phi_j(X)\}_{j=1}^m$ is given. We derive lower bounds for the MMSE of estimating $\phi(X)$ given Y in terms of the MMSE of estimating $\phi_j(X)$ given Y. In privacy systems, X may be a user's data and Y a distorted version of X generated by a privacy mechanism $P_{Y|X}$. The set $\{\phi_j(X)\}_{j=1}^m$ could then represent a set of functions that are known to be hard to infer from Y due to inherent privacy constraints of the setup. For example, when the privacy mechanism $P_{Y|X}$ is designed by the PIC-based convex programs in Formulations 5.5.2 and $\{\phi_j(X)\}_{j=1}^m$ is the set of private functions, $\mathsf{mmse}\left(\phi_j(X)|Y\right)$ is lower bounded due to the privacy constraints.

The following lemma will be used to derive the lower bounds for the MMSE of $\phi(X)$ given Y.

Lemma 14. *Let $L_n : (0,\infty)^n \times [0,1]^n \to \mathbb{R}$ be given by*

$$L_n(\boldsymbol{a},\boldsymbol{b}) \triangleq \max\left\{\boldsymbol{a}^T\boldsymbol{y} \mid \boldsymbol{y}\in\mathbb{R}^n, \|\boldsymbol{y}\|_2 \le 1, \boldsymbol{y}\le\boldsymbol{b}\right\}. \tag{5.19}$$

Let π be a permutation of $[n]$ such that $b_{\pi(1)}/a_{\pi(1)} \le \cdots \le b_{\pi(n)}/a_{\pi(n)}$. If $b_{\pi(1)}/a_{\pi(1)} \ge 1$, $L_n(\boldsymbol{a},\boldsymbol{b}) = \|\boldsymbol{a}\|_2$. Otherwise,

$$L_n(\boldsymbol{a},\boldsymbol{b}) = \sum_{i=1}^{k^*} a_{\pi(i)}b_{\pi(i)} + \sqrt{\left(\|\boldsymbol{a}\|_2^2 - \sum_{i=1}^{k^*} a_{\pi(i)}^2\right)\left(1-\sum_{i=1}^{k^*} b_{\pi(i)}^2\right)}$$

where

$$k^* \triangleq \max\left\{k\in[n] \;\middle|\; \frac{b_{\pi(k)}}{a_{\pi(k)}} \le \sqrt{\frac{\left(1-\sum_{i=1}^{k-1} b_{\pi(i)}^2\right)^+}{\|\boldsymbol{a}\|_2^2 - \sum_{i=1}^{k-1} a_{\pi(i)}^2}}\right\}. \tag{5.20}$$

Proof. See Appendix C.3.1. □

Throughout this section we assume $\|\phi_i(X)\|_2 = 1$ ($i\in[m]$) and $\mathbb{E}\left[\phi_i(X)\phi_j(X)\right] = 0$ ($i\neq j$). For a given ϕ_i, the inequality

$$\max_{\psi\in\mathcal{L}_2(P_Y)} \mathbb{E}\left[\phi_i(X)\psi(Y)\right] = \|\mathbb{E}\left[\phi_i(X)|Y\right]\|_2 \le \nu_i \tag{5.21}$$

is satisfied, where $0\le\nu_i\le 1$. This is equivalent to $\mathsf{mmse}(\phi_i(X)|Y) \ge 1-\nu_i^2$.

Theorem 9. *Let $\|\phi(X)\|_2 = 1$ and $\mathbb{E}\left[\phi(X)\phi_i(X)\right] = \rho_i > 0$. Denoting $\boldsymbol{\rho} \triangleq (|\rho_1|,\ldots,|\rho_m|)$, $\boldsymbol{\nu} \triangleq (\nu_1,\ldots,\nu_m)$, $\rho_0 \triangleq \sqrt{1-\sum_{i=1}^m \rho_i^2}$, $\boldsymbol{\rho}_0 \triangleq (\rho_0,\boldsymbol{\rho})$ and $\boldsymbol{\nu}_0 \triangleq (1,\boldsymbol{\nu})$, then*

$$\|\mathbb{E}\left[\phi(X)|Y\right]\|_2 \le B_m(\boldsymbol{\rho}_0,\boldsymbol{\nu}_0), \tag{5.22}$$

where

$$B_m(\boldsymbol{\rho}_0, \boldsymbol{\nu}_0) \triangleq \begin{cases} L_{m+1}(\boldsymbol{\rho}_0, \boldsymbol{\nu}_0) & \text{if } \rho_0 > 0, \\ L_m(\boldsymbol{\rho}, \boldsymbol{\nu}) & \text{otherwise,} \end{cases} \tag{5.23}$$

and L_n *is given in* (5.19)*. Consequently,*

$$\mathsf{mmse}(\phi(X)|Y) \geq 1 - B_m(\boldsymbol{\rho}_0, \boldsymbol{\nu}_0)^2. \tag{5.24}$$

Proof. See Appendix C.3.2. □

Denote $\psi_i(Y) \triangleq (T_{X|Y}\phi_i)(Y)/\|(T_{X|Y}\phi_i)(Y)\|_2$ ($i \in [m]$) and $\phi_0(X) \triangleq \rho_0^{-1}(\phi(X) - \sum_{i=1}^m \rho_i \phi_i(X))$ if $\rho_0 > 0$, otherwise $\phi_0(X) \triangleq 0$. The previous bounds, (5.22) and (5.24), can be further improved when $\mathbb{E}\left[\psi_i(Y)\phi_j(X)\right] = 0$ for $i \neq j$, $j \in \{0, \ldots, m\}$.

Theorem 10. *Let* $\|\phi(X)\|_2 = 1$ *and* $|\mathbb{E}\left[\phi(X)\phi_i(X)\right]| = \rho_i > 0$ *for* $i \in [m]$*. In addition, assume* $\mathbb{E}\left[\psi_i(Y)\phi_j(X)\right] = 0$ *for* $i \neq j$, $i \in [t]$ *and* $j \in \{0, \ldots, m\}$*, where* $0 \leq t \leq m$*. Then*

$$\|\mathbb{E}\left[\phi(X)|Y\right]\|_2 \leq \sqrt{\sum_{k=1}^{t} \nu_i^2 \rho_i^2 + B_{m-t}(\widetilde{\boldsymbol{\rho}}, \widetilde{\boldsymbol{\nu}})^2}, \tag{5.25}$$

where $\widetilde{\boldsymbol{\rho}} = (\rho_0, \rho_{t+1}, \ldots, \rho_m)$, $\widetilde{\boldsymbol{\nu}} = (1, \nu_{t+1}, \ldots, \nu_m)$*, and* B_m *is defined in* (5.23) *(considering* $B_0 = 0$*). In particular, if* $t = m$*,*

$$\|\mathbb{E}\left[\phi(X)|Y\right]\|_2 \leq \sqrt{\rho_0^2 + \sum_{k=1}^{m} \nu_i^2 \rho_i^2}, \tag{5.26}$$

and (5.26) *is an equality when* $\rho_0 = 0$*. Furthermore,*

$$\mathsf{mmse}(\phi(X)|Y) \geq 1 - \sum_{k=1}^{t} \nu_i^2 \rho_i^2 - B_{m-t}(\widetilde{\boldsymbol{\rho}}, \widetilde{\boldsymbol{\nu}})^2. \tag{5.27}$$

Proof. See Appendix C.3.3. □

In what follows, we use three examples to illustrate different use cases of Theorem 9 and 10. Example 4 illustrates how Theorem 10 can be applied to the q-ary symmetric channel which could be perceived as a model of randomized response [147, 291], and demonstrates that bound (5.26) is sharp. Example 5 illustrates Theorem 10 for the binary symmetric channel. Here the useful variable is composed by n uniform and independent bits. In this case, the basis can be expressed as the parity bits of the input to the channel. Finally, Example 6 illustrates Theorem 9 for one-bit functions. The same method used in the proof of Theorem 9 is applied to bound the probability of correctly guessing a one-bit function from an observation of the disclosed data.

Example 4 (q-ary symmetric channel)**.** Let $\mathcal{X} = \mathcal{Y} = [q]$, and Y be the result of passing X through an (ϵ, q)-ary symmetric channel, which is defined by the transition probability

$$P_{Y|X}(y|x) = (1-\epsilon)\mathbb{I}_{y=x} + \epsilon/q \text{ for all } x \in \mathcal{X}, y \in \mathcal{Y}. \tag{5.28}$$

We assume that X has a uniform distribution, which implies Y also has a uniform distribution. Any function $\phi \in \mathcal{L}_2(P_X)$ such that $\mathbb{E}[\phi(X)] = 0$ and $\|\phi(X)\|_2 = 1$ satisfies

$$\psi(Y) = (T_{X|Y}\phi)(Y) = (1-\epsilon)\phi(Y),$$

and, consequently, $\|(T_{X|Y}\phi)(Y)\|_2 = (1-\epsilon)$. We will use this fact to show that the bound (5.26) is sharp in this case.

Observe that for $\phi_i, \phi_j \in \mathcal{L}_2(P_X)$, if $\mathbb{E}[\phi_i(X)\phi_j(X)] = 0$ then $\mathbb{E}[\psi_i(Y)\psi_j(Y)] = 0$. Now let $\phi \in \mathcal{L}_2(P_X)$ satisfy $\mathbb{E}[\phi(X)] = 0$ and $\|\phi(X)\|_2 = 1$, and let $\mathbb{E}[\phi(X)\phi_i(X)] = \rho_i$ for $i \in [m]$, where $\{\phi_i\}$ satisfies the conditions in Theorem 10 and $\sum_{i=1}^m \rho_i^2 = 1$. In addition, $\|\psi_i(Y)\|_2 = (1-\epsilon) = \nu_i$. Then, from (5.26) and noting that $\rho_0 = 0$, $t = m$, we have

$$\|(T_{X|Y}\phi)(Y)\|_2 \leq \sqrt{\sum_{i=1}^m \nu_i^2\rho_i^2} = (1-\epsilon)\sqrt{\sum_{i=1}^m \rho_i^2} = 1-\epsilon,$$

which matches $\|(T_{X|Y}\phi)(Y)\|_2$, and the bound is tight in this case.

Example 5 (Binary channels with additive noise)**.** Let $\mathcal{X} = \{-1,1\}^n$ and $\mathcal{Y} = \{-1,1\}^n$, and Y^n be the result of passing X^n through a memoryless binary symmetric channel with crossover probability $\epsilon < 1/2$. We assume that X^n is composed by n uniform and i.i.d. bits. For $\mathcal{S} \subseteq [n]$, let

$$\chi_{\mathcal{S}}(X^n) \triangleq \prod_{i \in \mathcal{S}} X_i.$$

Any function $\phi : \mathcal{X} \to \mathbb{R}$ can then be decomposed in terms of the basis $\chi_{\mathcal{S}}(X^n)$ as [208]

$$\phi(X^n) = \sum_{\mathcal{S} \subseteq [n]} c_{\mathcal{S}}\chi_{\mathcal{S}}(X^n),$$

where $c_{\mathcal{S}} = \mathbb{E}[\phi(X^n)\chi_{\mathcal{S}}(X^n)]$. Furthermore, since $\mathbb{E}[\chi_{\mathcal{S}}(X^n)|Y^n] = (1-2\epsilon)^{|\mathcal{S}|}\chi_{\mathcal{S}}(Y^n)$, it follows from Theorem 10 that

$$\mathsf{mmse}(\phi(X^n)|Y^n) = 1 - \sum_{\mathcal{S} \subseteq [n]} c_{\mathcal{S}}^2(1-2\epsilon)^{2|\mathcal{S}|}. \tag{5.29}$$

This result can be generalized for the case $X^n = Y^n \otimes Z^n$, where the operation $\otimes$ denotes bit-wise

multiplication, Z^n is drawn from $\{-1,1\}^n$ and X^n is uniformly distributed. In this case

$$\mathsf{mmse}(\phi(X^n)|Y^n) = 1 - \sum_{\mathcal{S}\subseteq[n]} c_{\mathcal{S}}^2 \mathbb{E}\left[\chi_{\mathcal{S}}(Z^n)\right]^2. \tag{5.30}$$

Example 6 (One-Bit Functions)**.** Let X be a hidden random variable with support $\mathcal{X}$, and let Y be a noisy observation of X. We denote by $B_1, \ldots, B_m$ a collection of m predicates of X, where $B_i = \phi_i(X)$, $\phi_i : \mathcal{X} \to \{-1,1\}$ for $i \in [m]$ and, without loss of generality, $\mathbb{E}[B_i] = b_i \geq 0$.

We denote by $\hat{B}_i$ an estimate of B_i given an observation of Y, where $B_i \to X \to Y \to \hat{B}_i$. We assume that for any $\hat{B}_i$

$$\left|\mathbb{E}[B_i\hat{B}_i]\right| \leq 1 - 2\alpha_i$$

for some $0 \leq \alpha_i \leq (1-b_i)/2 \leq 1/2$. This condition is equivalent to imposing that $\Pr(B_i \neq \hat{B}_i) \geq \alpha_i$, since

$$\begin{aligned}\mathbb{E}\left[B_i\hat{B}_i\right] &= \Pr(B_i = \hat{B}_i) - \Pr(B_i \neq \hat{B}_i)\\ &= 1 - 2\Pr(B_i \neq \hat{B}_i).\end{aligned}$$

In particular, this captures the "hardness" of guessing B_i based solely on an observation of Y.

Now assume there is a bit B such that $\mathbb{E}[BB_i] = \rho_i$ for $i \in [m]$ and $\mathbb{E}\left[B_iB_j\right] = 0$ for $i \neq j$. We can apply the same method used in the proof of Theorem 9 to bound the probability of B being guessed correctly from an observation of Y:

$$\Pr(B \neq \hat{B}) \geq \frac{1}{2}\left(1 - B_m(\boldsymbol{\rho}, \boldsymbol{\nu})\right), \tag{5.31}$$

where $\nu_i = 1 - 2\alpha_i$.

5.7 Numerical Experiments

We illustrate some of the results derived in this chapter through two numerical experiments. The first experiment, conducted on a synthetic dataset, verifies the tightness of the upper bound for the χ^2-privacy-utility function. The second experiment, run on a real-world dataset, demonstrates the performance of the optimization methods proposed in Section 5.5.

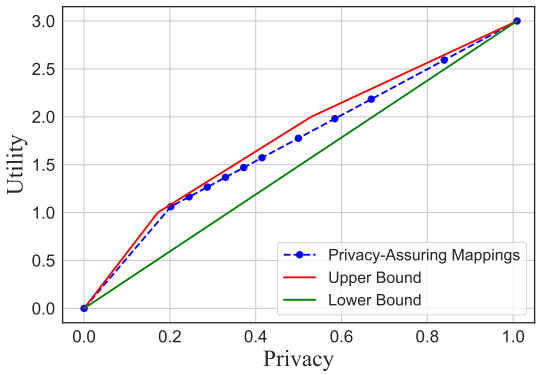

Figure 5.2: *We depict the bounds of the χ^2-privacy-utility function (see Theorem 7) and the privacy-utility values of the privacy mechanisms designed by the optimization methods in Section 5.5.*

5.7.1 Parity Bits

We choose private variable $S = (S_1, S_2) \in \{-1, 1\}^2$, where S is composed by two independent bits with $\Pr(S_1 = 1) = 0.45$ and $\Pr(S_2 = 1) = 0.4$. The useful variable $X = (X_1, X_2) \in \{-1, 1\}^2$ is generated by passing S_1 and S_2 through $\mathsf{BSC}(0.2)$ and $\mathsf{BSC}(0.15)$, respectively.

We use the optimization methods proposed in Section 5.5 to design privacy mechanisms. The private and useful functions are selected as $s_1(S) = S_1$ and $u_1(X) = X_1X_2$, $u_2(X) = X_2$, respectively. We first project the private function to the useful variable. Then we apply Formulation 5.5.2 with $\mathsf{obj}(\sigma_1, ..., \sigma_{n'}) = \sum_{i=1}^{n'} \sigma_i$ to find the privacy mechanisms.

In Fig. 5.2, we depict the privacy and utility, measured by χ^2-information, of the privacy mechanisms. We also draw the upper bound and lower bound of the χ^2-privacy-utility function. As shown, the privacy-utility values of the designed mechanisms are very close to the upper bound. In particular, since the χ^2-privacy-utility function is a concave function (see Lemma 12), the curve of this function is between its upper bound (red line) and the linear interpolation of the achievable privacy-utility values (dashed line).

5.7.2 UCI Adult Dataset

We apply our formulations to the UCI Adult Dataset [181]. A natural selection for the private and useful variables are $S = (\mathit{Gender}, \mathit{Race})$ and $X = (\mathit{Education\ Years}, \mathit{Income})$, respectively. This allows us to interpret the results of our formulations in an intuitive way, as one would expect there to exist correlations between the chosen private and useful variables. Private functions and useful functions

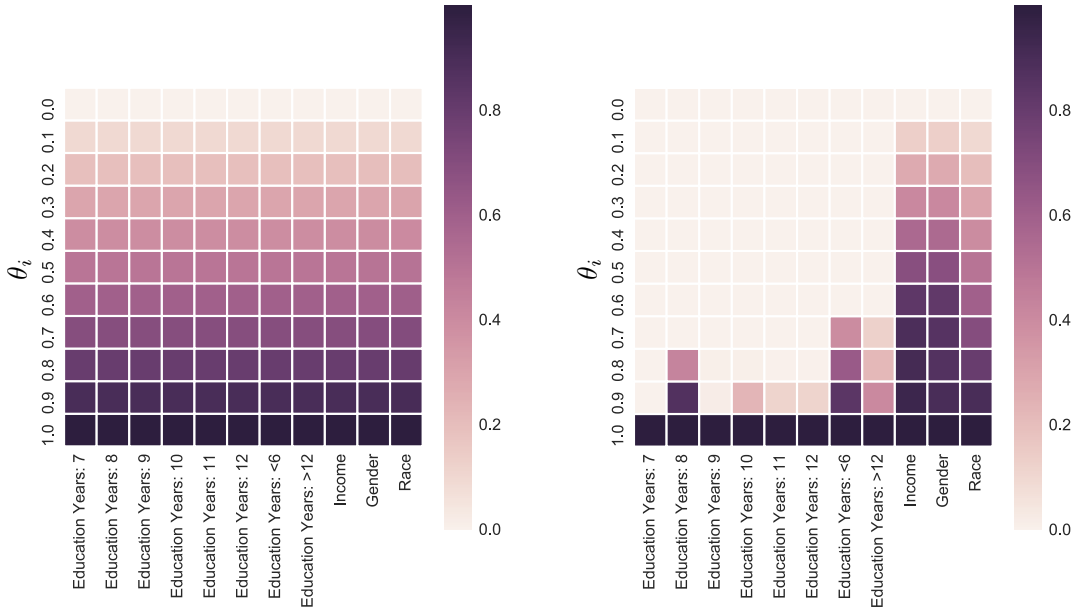

Figure 5.3: *MMSE of estimating each function given the disclosed variable, where darker means harder to estimate. Here* $(Education\ Years, Income)$ *and* $(Gender, Race)$ *are useful variable and private variable, respectively. The privacy parameters* θ_i *are selected as the same for all* i *and increase from* 0 *to* 1 *(i.e., the privacy constraints are increasing from the top down). The privacy mechanisms are designed by Formulation 5.5.2 with* $\mathsf{obj}(\sigma_1, ..., \sigma_{n'}) = \min\{\sigma_1, ..., \sigma_{n'}\}$ *(left) and with* $\mathsf{obj}(\sigma_1, ..., \sigma_{n'}) = \sum_{i=1}^{n'} \sigma_i$ *(right), respectively.*

are represented by indicator functions. Furthermore, functions which are linear combinations of others are removed. Following the same procedure proposed in Section 5.5, we first project all private functions to the useful variable. We use QR decomposition [271] to construct the basis $\{f_k(x)\}$. Note that other decomposition methods can also be used for constructing basis and, in fact, different bases affect the behavior of the PIC-based convex program (e.g., the joint distribution matrix $\mathbf{P}_{X,Y}$ returned by the optimization may be different). Consequently, the solution produced from the optimization program may not be optimal. Finally, Formulation 5.5.2 is used to compute the privacy mechanisms.

In Fig. 5.3, we show the MMSE of estimating useful functions and private functions given the disclosed variable. As shown, when we use Formulation 5.5.2 with $\mathsf{obj}(\sigma_1, ..., \sigma_{n'}) = \min\{\sigma_1, ..., \sigma_{n'}\}$ to compute privacy mechanisms, the estimation errors behave uniformly among all functions. This is because we aim at maximizing the worst-case utility over all useful functions. On the other hand, the privacy mechanisms designed by Formulation 5.5.2 with $\mathsf{obj}(\sigma_1, ..., \sigma_{n'}) = \sum_{i=1}^{n'} \sigma_i$ reveal more interpretable relationships between the private functions and useful functions. We see that *Income*, *Gender*, and *Race* are highly correlated, and it is not possible to reveal *Income* while maintaining privacy

for *Gender* and *Race*. Of particular interest are the subtle correlations between the three aforementioned functions and *Education Years*. There is a marked correlation between *Education Years* < 6 and, to a lesser degree, *Education Years* > 12, with *Gender* and *Race*. This may be due to the fact that most members of the dataset do not end their education midway. That is, most individuals will either never have begun schooling in the first place or will not continue their education after the 12-year benchmark, which marks graduation from high school. Therefore, we observe that the relationship between *Education Years* and *Race* is manifested the most in the two extremities of *Education Years* (> 12 and < 6). Also of note is the correlation between the private functions and *Education Years* : 8. Though not as obvious, this relationship can, too, be explained by the fact that 8 years of education marks another benchmark: the beginning of high school, also a time when people are prone to terminating their education.

5.8 Conclusion

In this chapter, we studied a fundamental PUT in data disclosure, where an analyst is allowed to reconstruct certain functions of the data, while other private functions should not be estimated with distortion below a certain threshold. First, χ^2-information was used to measure both privacy and utility. Bounds on the best PUT were provided and the upper bound, in particular, was shown to be achievable in the high-privacy region. Moreover, a PIC-based convex program was proposed to design privacy-assuring mechanisms when the useful functions and private functions were known beforehand. We also derived lower bounds on the MMSE of estimating a target function from the disclosed data. Our hope is that the methods presented here can inspire new, information-theoretically grounded and interpretable privacy mechanisms.

Chapter 6

Robustness of Privacy Measures and Mechanisms

A common approach to ensure privacy in data disclosure is to process the dataset through a privacy mechanism that seeks to fulfill certain privacy and utility guarantees. Information theoretic methods for designing privacy mechanisms often rely on the implicit assumption that the data distribution is, for the most part, known [e.g., 17, 29, 51, 52, 198, 209, 227, 228, 240, 281, 290]. However, in practice, the data distribution can only be accessed through a limited number of observed samples.

In this chapter, we consider the following setup. We assume that data has both private and non-private features and, based on a sample of such pairs, the designer creates a randomized mapping called a privacy mechanism. As new samples arrive, the designed mechanism is applied to the non-private features in order to produce a sanitized version of them which is later disclosed. In this context, we assume that the adversary (data analyst) knows the true distribution of the data, the privacy mechanism being used, and the disclosed data set. The adversary's objective is then to illegitimately infer the private features associated to the disclosed data. We illustrate this procedure in Figure 6.1. Observe that this adversary is the strongest in terms of statistical knowledge and hence serves as a worst case benchmark. Apart from its statistical knowledge, we assume that the adversary has no side information regarding the disclosed data.

Since privacy mechanisms are designed using a sample, their privacy-utility guarantees might not generalize to the true distribution, thereby creating a privacy threat as the *de facto* guarantees

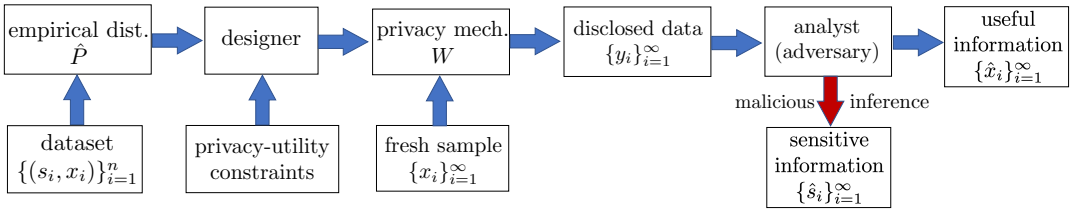

Figure 6.1: *Typical setting for the design and deployment of privacy mechanisms.*

might be considerably different from those in the design phase. In this chapter, we study the effect of the discrepancy between the empirical and true distributions on the analysis and design of privacy mechanisms.

First, we consider the setting where the privacy mechanism is designed using an estimate of the data distribution to evaluate the privacy-utility guarantees. We derive bounds for the discrepancy between the privacy-utility guarantees for the empirical distribution (type) and the true (unknown) data distribution. In this context, these bounds can be used to asses the *de facto* guarantees of privacy mechanisms when deployed in practice.

Next, we investigate the *statistical consistency* of the optimal privacy mechanisms. Assume that the privacy mechanism designer constructs an optimal privacy mechanism for the empirical distribution. As the number of samples increases, the optimal privacy mechanism designed for the empirical distribution naturally changes. We show that if such a sequence of privacy mechanisms converges, its limit is an optimal privacy mechanism for the true data distribution.

Finally, we introduce the notion of *uniform* privacy mechanism. When privacy is a priority, the privacy mechanism designer may be required to guarantee a specific level of privacy for the true distribution despite having access only to an estimate of it. Motivated by this setting, we consider privacy mechanisms that, by design, assure privacy for *every* distribution within a specific neighborhood of the empirical distribution. In this case, large deviations results imply that privacy is guaranteed for the true distribution with a certain probability (depending on the neighborhood). Since privacy is guaranteed *uniformly* on a neighborhood of the empirical distribution, we name these privacy mechanisms as *uniform privacy mechanisms*.

6.1 Overview and Main Contributions

In Section 6.4, we provide probabilistic upper bounds for the difference between the privacy-utility guarantees for the empirical and the true distributions under five different information metrics: probability of correctly guessing, f-information with f locally Lipschitz, Arimoto's mutual information (α-leakage) of order $\alpha > 1$, Sibson's mutual information of order $\alpha > 1$, and maximal α-leakage of order $\alpha > 1$. These bounds, which scale as $O(1/\sqrt{n})$ with n being the sample size, hold uniformly across all privacy mechanisms and, hence, can be used even for privacy mechanisms that do not have explicit descriptions, e.g., the privacy mechanisms implemented using neural networks in [126]. Explicit constants depending on the information metrics under consideration are provided. The proofs of these bounds rely on known large deviations results [292] and Lipschitz continuity properties of information leakage measures established in this chapter. These continuity properties can be combined with results for other estimation frameworks different from the large deviations one, e.g., the ℓ_1-minimax setting in [150] and references therein. In those cases, the role of the empirical estimator is naturally replaced by other estimators, e.g., the add-constant estimator in [150].

We study the convergence properties of optimal privacy mechanisms in Section 6.5. Specifically, we consider the case when a privacy mechanism designer constructs a sequence optimal privacy mechanisms for a sequence of joint distributions converging to the true distribution. In this context, we show that while the sequence of optimal privacy mechanisms do not necessarily form a Cauchy sequence, the distance between the privacy mechanisms in the sequence and the set of optimal privacy mechanisms with respect to the true distribution does go to zero. This convergence is analyzed for information measures which satisfy three technical conditions outlined in Section 6.5. These conditions are satisfied by probability of correctly guessing, f-information with f locally Lipschitz, and Arimoto's mutual information (α-leakage) of order $\alpha > 1$. As a by-product, we prove that, under these conditions, the privacy-utility function as a function of the joint distribution is continuous except possibly at the boundary of its domain.

In Section 6.6, we introduce the notion of uniform privacy mechanism and prove the existence of *optimal uniform privacy mechanisms*, i.e., uniform privacy mechanisms that attain the best utility in a max-min sense. Optimal uniform privacy mechanisms are considerably harder to design than their non-uniform counterparts as privacy has to be guaranteed for any distribution within a neighborhood of the empirical distribution. Nonetheless, we show that they can be approximated by certain mechanisms designed to deliver privacy *only* for the empirical distribution. This approximation

result circumvents the highly non-trivial task of designing optimal uniform privacy mechanisms.

6.2 Related Works

Differential privacy is a popular measure of privacy which aims at answering queries while simultaneously ensuring privacy of individual records in the database [87]. It does not take into account the distribution of the entire dataset which results in its robustness with respect to the dataset distribution, a property that is not satisfied *a priori* by information-theoretic privacy mechanisms. In this chapter, we analyze the degree of robustness to variations of the dataset distribution of certain information-theoretic measures of privacy.

The study of privacy from an information-theoretic point of view heavily depends on the chosen information metric. One commonly-used information leakage metric is Shannon's mutual information [247] and its generalizations [276]. In particular, both α-leakage and f-information — a special case of Csiszár's f-divergence [70] — have been used to formulate the privacy-utility trade-offs problem, see e.g., Liao *et al.* [177]. These generalizations seek to study privacy using information metrics that have appeared in the information theory literature and carry some operational (statistical) meaning. There are also information measures which have been introduced specifically in the context of privacy and information leakage. For example, Issa *et al.* [131, 133] introduced a metric called maximal leakage to measure privacy leakage, later extended by Liao *et al.* [178] in a tunable measure called α-leakage. Duchi *et al.* [81] presented locally differential privacy. We refer the reader to Wagner and Eckhoff [279] for a survey of privacy metrics and remark that information metrics also shed light on many secrecy problems, e.g., cryptosystems with short keys [174], side-channel attack [133], and entropic security [79, 237].

Many methods to design information-theoretic privacy mechanisms have been proposed in the past, see, e.g., [18, 51, 133, 179, 229, 240]. However, all these works assume that the designer has full knowledge of the data distribution. In practice, the designer may only have access to n i.i.d. samples, creating a mismatch between the distribution used to design the privacy mechanism and the true one. We show that for every information leakage measure considered (see Theorem 11), the discrepancy between the privacy-utility guarantees for the empirical and true distributions scales as $O(1/\sqrt{n})$. It is important to remark that, in the context of the information bottleneck, Shamir *et al.* [245] established similar upper bounds for mutual information that scale as $O(\log(n)/\sqrt{n})$. Hence,

mutual information appears to fall outside the scope of the techniques presented in this chapter. A related problem is that of estimating information-theoretic measures from a limited number of samples, as studied by, for example, Jiao *et al.* [140], Wu and Yang [298] and Issa *et al.* [133] in the context of privacy. Our work differs from those as we do not focus on the construction of estimators suited for a specific information leakage measure, instead we analyze the performance of the plug-in estimator under different leakage measures. Given the ease of implementation of this estimator, we believe our contributions might be of interest to practitioner.

The fundamental trade-offs between privacy guarantees and statistical utility in data disclosure have been studied in several papers from an information-theoretic view. For example, Yamamoto [303] developed the trade-off between rate, distortion, and equivocation for a specific source coding model. Rebollo-Monedero *et al.* [229], Calmon and Fawaz [51], and Sankar *et al.* [240] characterized privacy-utility trade-offs using tools from rate-distortion theory. Varodayan and Khisti [274] and Tan *et al.* [260] considered the privacy-energy efficiency trade-off in smart meter systems. Makhdoumi *et al.* [187] showed the connection between the privacy funnel and the information bottleneck proposed by Tishby *et al.* [263]. The privacy-utility function has been analyzed to find the fundamental limits of privacy-utility trade-offs under different metrics, see, e.g., Calmon *et al.* [55] and Asoodeh *et al.* [17]. It has been shown that this function exhibits some common properties across different metrics. For example, it is continuous, concave, and strictly increasing with respect to the privacy parameter, see also Wang and Calmon [281] and Asoodeh *et al.* [18].

The study of robustness has appeared in many different areas, such as optimization [34], statistics [129], artificial intelligence [270], and machine learning [58]. The notion of robustness considered in this chapter is closely related to generalization in machine learning [244], which aim at analyzing different performances of the learner's output between training and testing phases. More specifically, a learner may output a classifier h which minimizes a given empirical risk over a sample $S = (z_1, \cdots, z_m)$, namely,

$$L_S(h) \triangleq \frac{1}{m} \sum_{i=1}^{m} \ell(h, z_i), \tag{6.1}$$

where ℓ is a given loss function. Nonetheless, in practice, the same classifier h might exhibit a very different true risk, defined as

$$L(h) \triangleq \mathbb{E}\left[\ell(h, Z)\right], \tag{6.2}$$

where Z is a random variable distributed according to the dataset distribution. The goal of generalization is to analyze the difference between the empirical risk and the true risk and derive bounds for this discrepancy. This problem has been studied in computer science [154] and, recently, has been analyzed using tools from information theory by Xu and Raginsky [300] and Russo and Zou [238]. Since most information-theoretic metrics do not seem to be expressible in terms of a loss function as in (6.2), classical generalization results cannot be directly applied for analyzing the robustness of information-theoretic privacy mechanisms. Recently, the robustness of privacy mechanisms has been investigated using tools from information theory in the works of Wang and Calmon [281] and Issa *et al.* [133]. It is important to mention that some techniques developed for privacy preservation, such as differential privacy, can be useful to analyze generalization guarantees, as shown by Dwork *et al.* [84].

6.3 Preliminaries and Problem Setup

In this section, we review preliminary material regarding information leakage measures and large deviations bounds for distribution estimation. In addition, we formally introduce the problems addressed in this chapter.

Notation. Given random variables U and U' supported over a finite alphabet $\mathcal{U}$ and another random variable V supported over a finite alphabet $\mathcal{V}$, we let $P_{U'} \cdot P_{V|U}$ be the joint distribution over $\mathcal{U} \times \mathcal{V}$ determined by

$$(P_{U'} \cdot P_{V|U})(u, v) = P_{U'}(u) P_{V|U}(v|u). \tag{6.3}$$

Observe that with this notation, $P_U \cdot P_{V|U} = P_{U,V}$. Given another random variable $\widetilde{V}$ supported over a finite alphabet $\widetilde{\mathcal{V}}$, we let $P_{\widetilde{V}|U} P_{V|\widetilde{V}}$ be the transition probability matrix determined by

$$(P_{\widetilde{V}|U} P_{V|\widetilde{V}})(v|u) = \sum_{\widetilde{v} \in \widetilde{\mathcal{V}}} P_{\widetilde{V}|U}(\widetilde{v}|u) P_{V|\widetilde{V}}(v|\widetilde{v}). \tag{6.4}$$

We let S and X be two random variables with discrete supports $\mathcal{S}$ and $\mathcal{X}$, respectively. We let P denote the joint distribution $P_{S,X}$ and $\hat{P}$ be a generic estimate of P. The empirical distribution obtained from n i.i.d. samples drawn from P is denoted by $\hat{P}_n$. We let $(P_n)_{n=1}^{\infty}$ be any sequence of joint distributions converging to the joint distribution P. Any generic distribution over $\mathcal{S} \times \mathcal{X}$ is denoted by Q. Also, any sequence of distributions over $\mathcal{S} \times \mathcal{X}$ is denoted by $(Q_n)_{n=1}^{\infty}$.

Category	Notation	Meaning
Norms	$\|a\|_\alpha$	α-norm with $\alpha \in [1, \infty)$: $(\sum_i \|a_i\|^\alpha)^{1/\alpha}$
	$\|a\|_\infty$	∞-norm: $\max_i \|a_i\|$
Joint Distributions	P	True joint distribution: $P_{S,X}$
	$\hat{P}$	Estimated distribution: $\hat{P}_{S,X}$
	$\hat{P}_n$	Empirical distribution obtained from n i.i.d. samples
	$(P_n)_{n=1}^\infty$	Any sequence of joint distributions converging to P
	Q	Any joint distribution over $\mathcal{S} \times \mathcal{X}$
	$(Q_n)_{n=1}^\infty$	Any sequence of joint distributions over $\mathcal{S} \times \mathcal{X}$
Channels	W	Any privacy mechanism (channel) from $\mathcal{X}$ to $\mathcal{Y}$
Sets	$\mathcal{W}$	Set of all privacy mechanisms
	$\mathcal{W}_N$	Set of all privacy mechanisms from $\mathcal{X}$ to $\{1 \ldots, N\}$
	$\mathcal{P}$	Set of all joint distributions over $\mathcal{S} \times \mathcal{X}$
	$\mathcal{Q}$	Any closed subset of $\mathcal{P}$

Table 6.1: *List of notation used in this chapter.*

For convenience, we summarize some common notation in Table 6.1.

6.3.1 Information Leakage Measures.

Here we review the definition and some basic properties of information leakage measures which are commonly used in the literature. An information leakage measure $\mathcal{L}(U \to V)$ quantifies how much information V leaks about U. Specifically, U and V can represent raw and disclosed datasets, respectively, and, in this case, $\mathcal{L}(U \to V)$ measures the utility of the disclosed dataset. On the other hand, when U and V represent private information and information available to an adversary, then $\mathcal{L}(U \to V)$ quantifies the adversary's ability of inferring U from the observation of V. In the privacy literature it is valuable to relate information leakage measures with a specific loss/gain function (implicitly) adopted by an adversary. In what follows, we review a range of information leakage measures along with their operational meanings in terms of loss/gain functions.

α-Leakage

In an effort to unify several information leakage measures within a single framework, Liao *et al.* recently introduced an information leakage measure called α-leakage [178].

Definition 18 (Definition 4, [178])**.** Let U and V be two random variables supported over finite sets $\mathcal{U}$ and $\mathcal{V}$, respectively. For $\alpha \in (1, \infty)$, the α-leakage from U to V is defined as

$$\mathcal{L}_\alpha(U \to V) \triangleq \frac{\alpha}{\alpha - 1} \log \frac{\max\limits_{\hat{U}: U \to V \to \hat{U}} \mathbb{E}\left[\Pr(\hat{U} = U|U,V)^{\frac{\alpha-1}{\alpha}}\right]}{\max\limits_{\hat{U}: U \perp\!\!\!\perp \hat{U}} \mathbb{E}\left[\Pr(\hat{U} = U|U)^{\frac{\alpha-1}{\alpha}}\right]}. \tag{6.5}$$

The value of $\mathcal{L}_\alpha(U \to V)$ is extended by continuity to $\alpha = 1$ and $\alpha = \infty$.

In terms of the adversarial setting described at the beginning of this section, α-leakage with $\alpha \in (1, \infty]$ can be interpreted as the multiplicative increase, upon observing V, of the expected gain of an adversary. In order to make this observation precise, note that the expected value in the denominator of the logarithmic term in (6.5) equals

$$\mathbb{E}\left[\Pr(\hat{U} = U|U)^{\frac{\alpha-1}{\alpha}}\right] = \sum_{u \in \mathcal{U}} P_U(u) P_{\hat{U}}(u)^{\frac{\alpha-1}{\alpha}}. \tag{6.6}$$

In particular, the RHS of (6.6) equals expected gain of $P_{\hat{U}}$ w.r.t. the gain function

$$\mathsf{gain}(u, P_{\hat{U}}) = P_{\hat{U}}(u)^{\frac{\alpha-1}{\alpha}}. \tag{6.7}$$

Thus, the denominator of the logarithmic term in (6.5) is the largest expected gain of an adversary with no additional information apart from the distribution of U. Also, the expected value in the numerator of the logarithmic term in (6.5) equals

$$\mathbb{E}\left[\Pr(\hat{U} = U|U,V)^{\frac{\alpha-1}{\alpha}}\right] = \sum_{u \in \mathcal{U}} \sum_{v \in \mathcal{V}} P_{U,V}(u,v) P_{\hat{U}|V}(u|v)^{\frac{\alpha-1}{\alpha}}. \tag{6.8}$$

Thus, the numerator of the logarithmic term in (6.5) is the largest expected gain of an adversary that has access to V and the joint distribution of U and V. From these observations we conclude that $\mathcal{L}_\alpha(U \to V)$ measures the multiplicative increase of the best expected gain upon observing V. In a similar way, $\mathcal{L}_1(U \to V)$ can be interpreted as additive increase, upon observing V, of the expectation of the gain function

$$\mathsf{gain}(u, P_{\hat{U}}) = \log P_{\hat{U}}(u). \tag{6.9}$$

In addition to its operational definition, α-leakage can be related to Arimoto's mutual information

[12, 97] whose definition is provided below.

Definition 19. Let U and V be two random variables supported over finite sets $\mathcal{U}$ and $\mathcal{V}$, respectively. Their Arimoto's mutual information of order $\alpha \in (1, \infty)$ is defined as

$$I_\alpha^{\mathrm{A}}(P_{U,V}) \triangleq \frac{\alpha}{\alpha - 1} \log \frac{\sum_v \|P_{U,V}(\cdot, v)\|_\alpha}{\|P_U(\cdot)\|_\alpha}. \tag{6.10}$$

Also, by continuous extension,

$$I_1^{\mathrm{A}}(P_{U,V}) \triangleq I(P_{U,V}) \quad \text{and} \quad I_\infty^{\mathrm{A}}(P_{U,V}) \triangleq \log \frac{\sum_v \max_u P_{U,V}(u, v)}{\max_u P_U(u)}, \tag{6.11}$$

where $I(P_{U,V})$ denotes Shannon's mutual, i.e., $I(P_{U,V}) \triangleq \sum_{u \in \mathcal{U}} \sum_{v \in \mathcal{V}} P_{U,V}(u, v) \log \frac{P_{U,V}(u, v)}{P_U(u) P_V(v)}$.

In [178], Liao *et al.* proved that α-leakage is indeed equal to Arimoto's mutual information of order α.

Proposition 12 (Theorem 1, [178])**.** *Let U and V be two random variables supported over finite sets $\mathcal{U}$ and $\mathcal{V}$, respectively. For $\alpha \in [1, \infty]$, α-leakage satisfies*

$$\mathcal{L}_\alpha(U \to V) = I_\alpha^{\mathrm{A}}(P_{U,V}). \tag{6.12}$$

The previous proposition shows that α-leakage recovers Shannon's mutual information in the extremal case $\alpha = 1$. It is important to point out that Shannon's mutual information has been extensively used in the context of privacy (see e.g., the works of Rebollo-Monedero *et al.* [229], Sankar *et al.* [240], Calmon and Fawaz [51], and Asoodeh *et al.* [17]). On the other extreme, when $\alpha = \infty$, α-leakage is closely related to another information leakage measure: probability of correctly guessing. This information leakage measure, whose definition is recalled next, has been used recently in the context of privacy-utility trade-offs by Asoodeh *et al.* [18] and in the broader privacy literature by Smith [250], Braun *et al.* [43], Barthe and Kopf [26], and references therein.

Definition 20. Let U and V be two random variables supported over finite sets $\mathcal{U}$ and $\mathcal{V}$, respectively. The probability of correctly guessing U and the probability of correctly guessing U given V are given by

$$P_c(U) \triangleq \max_{u \in \mathcal{U}} P_U(u), \tag{6.13}$$

$$P_c(U|V) \triangleq \sum_{v \in \mathcal{V}} \max_{u \in \mathcal{U}} P_{U,V}(u, v). \tag{6.14}$$

As its name suggests, $P_c(U)$ is equal to the (largest) probability of correctly guessing U without any side information but the distribution of U. Similarly, $P_c(U|V)$ is equal to the (largest) probability of correctly guessing U given V and the joint distribution $P_{U,V}$. This interpretation can be made precise using the adversarial setting and the gain function defined in (6.7) with $\alpha = \infty$. Indeed, a simple manipulation shows that

$$P_c(U) = \max_{\hat{U}:U \perp\!\!\!\perp \hat{U}} \mathbb{E}\left[\mathbb{I}_{U=\hat{U}}\right] \quad \text{and} \quad P_c(U|V) = \max_{\hat{U}:U\to V\to \hat{U}} \mathbb{E}\left[\mathbb{I}_{U=\hat{U}}\right], \tag{6.15}$$

which corresponds to the adversarial setting under the 0-1 loss. Observe that under this loss, the optimal adversary strategy is the same as the maximum a posteriori (MAP) decoder in the communication literature. Finally, observe that

$$\mathcal{L}_\infty(U \to V) = \log \frac{P_c(U|V)}{P_c(U)}, \tag{6.16}$$

evidencing the intrinsic relation between ∞-leakage and probability of correctly guessing.

Remark 8. Despite the close relation between ∞-leakage and probability of correctly guessing, we always treat them separately. On the one hand, probability of correctly guessing is of interest in its own right. On the other hand, the results for probability of correctly guessing are easier to derive than their ∞-leakage counterparts. As a result, working explicitly with probability of correctly guessing allows us to present the key techniques used in this chapter while keeping the technical difficulties at a minimum.

In [133], Issa *et al.* introduced the notion of maximal leakage in order to quantify the adversary's capability to infer any (randomized) function of U from V. Motivated by this notion, Liao *et al.* introduced *maximal α-leakage*, a tunable information leakage measure that is equal to maximal leakage when $\alpha = \infty$.

Definition 21 (Definition 5, [178])**.** Let U and V be two random variables supported over finite sets $\mathcal{U}$ and $\mathcal{V}$, respectively. For $\alpha \in [1, \infty]$, the maximal α-leakage from U to V is defined as

$$\mathcal{L}_\alpha^{\max}(U \to V) \triangleq \sup_{T:T\to U\to V} \mathcal{L}_\alpha(T \to V), \tag{6.17}$$

where T represents any (randomized) function of U that takes values on a finite but arbitrary alphabet.

Arimoto's mutual information can be regarded as a generalization of Shannon's mutual infor-

mation, see, e.g., [276]. Another such generalization is Sibson's mutual information [249] which, as shown below, is also related to maximal α-leakage.

Definition 22 (Definition 4, [276])**.** Let U and V be two random variables supported over finite sets $\mathcal{U}$ and $\mathcal{V}$, respectively. Their Sibson's mutual information of order $\alpha \in (1, \infty)$ is defined as

$$I_\alpha^{\mathrm{S}}(P_{U,V}) \triangleq \frac{\alpha}{\alpha - 1} \log \sum_{v \in \mathcal{V}} \|P_U(\cdot)^{1/\alpha} P_{V|U}(v|\cdot)\|_\alpha. \tag{6.18}$$

Also, $I_1^{\mathrm{S}}(P_{U,V}) = I(P_{U,V})$ and $I_\infty^{\mathrm{S}}(P_{U,V}) = \log \left(\sum_{v \in \mathcal{V}} \max_{u \in \mathcal{U}} P_{V|U}(v|u) \right)$.

The following proposition shows that maximal α-leakage is the Arimoto's capacity under an input support constraint which, in turn, is equal to Sibson's capacity under the same constraint.

Proposition 13 (Theorem 2, [178])**.** *Let U and V be two random variables supported over finite sets $\mathcal{U}$ and $\mathcal{V}$, respectively. For $\alpha \in (1, \infty]$, maximal α-leakage satisfies*

$$\mathcal{L}_\alpha^{\max}(U \to V) = \sup_{P_{\widetilde{U}}} I_\alpha^{\mathrm{A}}(P_{\widetilde{U}} \cdot P_{V|U}) = \sup_{P_{\widetilde{U}}} I_\alpha^{\mathrm{S}}(P_{\widetilde{U}} \cdot P_{V|U}), \tag{6.19}$$

where the support of $P_{\widetilde{U}}$ is a subset of the support of P_U[1]*. Also, $\mathcal{L}_1^{\max}(U \to V) = I(P_{U,V})$.*

In particular, maximal α-leakage of order infinity (i.e., maximal leakage) is the Sibson's mutual information of order infinity.

Proposition 14 (Corollary 1, [133])**.** *Let U and V be two random variables supported over finite sets $\mathcal{U}$ and $\mathcal{V}$, respectively. Then,*

$$\mathcal{L}_\infty^{\max}(U \to V) = I_\infty^{S}(P_{U,V}). \tag{6.20}$$

f-Information

We finish our review on information leakage measures by recalling the definition and some basic properties of f-information. This information leakage measure is defined in terms of Csiszár's f-divergence whose definition is recalled next for the reader's convenience.

Definition 23 (Definition 1.1, [70])**.** Let $f : (0, \infty) \to \mathbb{R}$ be a convex function with $f(1) = 0$. Assume that Q_1 and Q_2 are two probability distributions over a finite set $\mathcal{Z}$ and that $Q_1 \ll Q_2$. The

[1] In [178, Theorem 2], $P_{\widetilde{U}}$ ranges over the distributions with the same support as P_U. However, their proof readily implies Proposition 13 as stated in this chapter.

f-divergence between Q_1 and Q_2 is given by

$$D_f(Q_1\|Q_2) \triangleq \sum_{z\in\mathcal{Z}} Q_2(z) f\left(\frac{Q_1(z)}{Q_2(z)}\right). \tag{6.21}$$

Given a function f as in the previous definition, its corresponding f-information is defined as follows.

Definition 24. Let $f : (0,\infty) \to \mathbb{R}$ be a convex function with $f(1) = 0$. Furthermore, let U and V be two random variables supported over finite sets $\mathcal{U}$ and $\mathcal{V}$, respectively. Their f-information is defined by

$$\begin{aligned} I_f(P_{U,V}) &\triangleq D_f(P_{U,V}\|P_U P_V) \\ &= \sum_{u\in\mathcal{U}}\sum_{v\in\mathcal{V}} P_U(u)P_V(v) f\left(\frac{P_{U,V}(u,v)}{P_U(u)P_V(v)}\right). \end{aligned} \tag{6.22}$$

While an f-information may not have a straightforward interpretation in operational terms for a specific function f, many functions f do, e.g., mutual information and χ^2-information discussed below. Hence, a general treatment allows us to simultaneously handle those f-information for which there is a concrete operational meaning and those whose operational interpretation is yet to be discovered. More recent developments about the properties of f-divergence can be found in Raginsky [223], Calmon *et al.* [54], and the references therein.

In the context of privacy, χ^2-information is a fine example of an f-information with a tangible operational interpretation. Following the standard convention, the χ^2-information between two random variables U and V is defined as $\chi^2(P_{U,V}) \triangleq I_f(P_{U,V})$ with $f(t) = (t-1)^2$. Relying on the so-called *principal inertia components* (PICs) [113], Calmon *et al.* [53] showed that if $\chi^2(U;V) < \epsilon$ for some $0 < \epsilon < 1$, then the *minimum mean-squared error* (MMSE) of reconstructing *any* zero-mean unit-variance function of U given V is lower bounded by $1-\epsilon$, i.e., no function of U can be reconstructed with small MMSE given an observation of V. Thus, χ^2-information measures an adversary's ability to reconstruct functions of U from V under an MMSE loss. Recently, χ^2-information has been used in the context of privacy-utility trade-offs by Wang and Calmon in [281].

6.3.2 Large Deviations Bounds for Distribution Estimation

Now we review some preliminary results regarding the distance between the empirical and true distributions. Throughout this chapter, we measure the distance between two probability distributions

over $\mathcal{Z}$, say Q_1 and Q_2, by their ℓ_1-distance which is given by

$$\|Q_1 - Q_2\|_1 \triangleq \sum_{z \in \mathcal{Z}} |Q_1(z) - Q_2(z)|. \tag{6.23}$$

A result by Weissman *et al.* [292, Theorem 2.1] establishes that, for all $\epsilon > 0$,

$$\Pr\left(\|\hat{P}_n - P\|_1 \geq \epsilon\right) \leq (2^{|\mathcal{Z}|} - 2)\exp(-n\phi(\pi_P)\epsilon^2/4), \tag{6.24}$$

where P is a probability distribution over the finite set $\mathcal{Z}$, $\hat{P}_n$ is the empirical distribution obtained from n i.i.d. samples from P, $\pi_P \triangleq \max_{\mathcal{A} \subseteq \mathcal{Z}} \min\{P(\mathcal{A}), 1 - P(\mathcal{A})\}$, and

$$\phi(p) \triangleq \begin{cases} \frac{1}{1-2p}\log\frac{1-p}{p} & p \in [0, 1/2), \\ 2 & p = 1/2. \end{cases} \tag{6.25}$$

Note that $\phi(p) \geq 2$ for all $p \in [0, 1/2]$. Hence,

$$\Pr\left(\|\hat{P}_n - P\|_1 \geq \epsilon\right) \leq \exp(|\mathcal{Z}|)\exp(-n\epsilon^2/2). \tag{6.26}$$

By taking $\mathcal{Z} = \mathcal{S} \times \mathcal{X}$ and $\epsilon = \sqrt{\frac{2}{n}(|\mathcal{S}| \cdot |\mathcal{X}| - \log\beta)}$, inequality (6.26) implies that, with probability at least $1 - \beta$,

$$\|\hat{P}_n - P\|_1 \leq \sqrt{\frac{2}{n}(|\mathcal{S}| \cdot |\mathcal{X}| - \log\beta)}. \tag{6.27}$$

Even though in this chapter we focus on large deviations results, it is worth pointing out that the order $O(\sqrt{|\mathcal{S}| \cdot |\mathcal{X}|/n})$ is present in other fundamental settings, e.g., the minimax expected loss framework in [150, Corollary 9].

6.3.3 Problem Setup

We follow the same framework considered in the last chapter. Again, S is a variable to be hidden (e.g., political preference) and X is an observed variable (e.g., movie ratings) that is correlated with S. In order to receive some utility (e.g., personalized recommendations), we would like to disclose as much information about X without compromising S. An approach with rigorous privacy guarantees is to release a new random variable Y produced by applying a randomized mapping to X, hence forming a Markov chain $S \to X \to Y$. This mapping, called the privacy mechanism, is designed to satisfy a specific privacy constraint.

In the sequel, we assume that S and X are discrete random variables with support sets $\mathcal{S}$ and

$\mathcal{X}$, respectively. Let $\mathcal{W}$ be the set of all privacy mechanisms which take X as input and display a discrete random variable Y as output. More specifically, let

$$\mathcal{W} \triangleq \bigcup_{N\geq 1} \left\{ W \in [0,1]^{|\mathcal{X}|\times N} : \forall x \in \mathcal{X}, \sum_{y=1}^{N} W(x,y) = 1 \right\}. \tag{6.28}$$

Through this chapter, we denote the joint distribution of S and X by P and the privacy mechanism producing Y from X by W, i.e., $W = P_{Y|X}$. The privacy leakage and the utility generated by a mapping $W \in \mathcal{W}$ for the true distribution P are denoted by $\mathcal{L}(P,W)$ and $\mathcal{U}(P,W)$, respectively. Throughout this chapter, $\mathcal{L}$ and $\mathcal{U}$ are either probability of correctly guessing, f-information with f locally Lipschitz, Arimoto's mutual information (α-leakage) of order $\alpha \in (1,\infty]$, Sibson's mutual information of order $\alpha \in (1,\infty]$, or maximal α-leakage of order $\alpha \in (1,\infty]$.

In this framework, a natural problem is to characterize the fundamental trade-off between privacy and utility as captured by the following definition. We denote by $\mathcal{P}$ the set of all probability distributions over $\mathcal{S} \times \mathcal{X}$.

Definition 25. For a given $P \in \mathcal{P}$ and $\epsilon \geq \inf_{W\in\mathcal{W}} \mathcal{L}(P,W)$, the *privacy-utility function* is defined as

$$\mathsf{H}(P;\epsilon) \triangleq \sup_{W\in\mathcal{D}(P;\epsilon)} \mathcal{U}(P,W), \tag{6.29}$$

where $\mathcal{D}(P;\epsilon) \triangleq \{W \in \mathcal{W} : \mathcal{L}(P,W) \leq \epsilon\}$. Furthermore, the collection of all optimal privacy mechanisms for P at ϵ is defined as

$$\mathcal{W}^*(P;\epsilon) \triangleq \{W \in \mathcal{W} : \mathcal{L}(P,W) \leq \epsilon,\ \mathcal{U}(P,W) = \mathsf{H}(P;\epsilon)\}. \tag{6.30}$$

Observe that, by definition, $\mathcal{D}(P;\epsilon)$ is the set of all privacy mechanisms providing an ϵ-privacy guarantee for P. Hence, the privacy-utility function in Definition 25 quantifies the best utility achieved by any privacy mechanism providing an ϵ-privacy guarantee for P.

This specific type of privacy-utility trade-off (PUT), and the optimal privacy mechanisms associated to it, has been investigated for several measures of privacy and utility, see, for example, [17, 18, 227, 281]. These investigations often rely on the implicit assumption that the data distribution is, for the most part, known. However, in practice, the data distribution may only be accessed through a limited number of samples. In this chapter, we revisit this assumption and study its implications in the design and performance of privacy mechanisms. Next we present the problems addressed and the main contributions of this chapter.

Discrepancy of Privacy-Utility Guarantees

In practice, the designer may not have access to the true distribution P, but only to samples drawn from this distribution. In this case, the privacy-utility guarantees for a distribution estimated from the samples, say $\hat{P}$, and the true distribution P might be different. For any given privacy mechanism W, these discrepancies are effectively quantified by

$$|\mathcal{L}(\hat{P}, W) - \mathcal{L}(P, W)| \quad \text{and} \quad |\mathcal{U}(\hat{P}, W) - \mathcal{U}(P, W)|. \tag{6.31}$$

In Section 6.4 we provide probabilistic upper bounds for the discrepancies in (6.31) when the estimated distribution $\hat{P}$ is the empirical distribution $\hat{P}_n$ of n i.i.d. samples drawn from P. These upper bounds depend on the sample size n, the alphabet sizes $|\mathcal{S}|$ and $|\mathcal{X}|$, and, in some cases, the probability of the least likely symbol of the marginals of $\hat{P}_n$. Their derivations rely on the large deviations results [292] and continuity properties of information leakage measures established in this chapter. We summarize these results in the following meta theorem (see Theorem 11). For ease of notation, we let

$$\mathcal{L}_c(Q, W) \triangleq P_c(S|Y) \quad \text{and} \quad \mathcal{U}_c(Q, W) \triangleq P_c(X|Y), \tag{6.32}$$

where $S \to X \to Y$ is such that $P_{S,X} = Q$ and $P_{Y|X} = W$. With this notation, we also let $\mathcal{L}_f(Q, W) \triangleq I_f(P_{S,Y})$, $\mathcal{L}_\alpha^{\mathrm{A}}(Q, W) \triangleq I_\alpha^{\mathrm{A}}(P_{S,Y})$, $\mathcal{L}_\alpha^{\mathrm{S}}(Q, W) \triangleq I_\alpha^{\mathrm{S}}(P_{S,Y})$, and, by abuse of notation, $\mathcal{L}_\alpha^{\max}(Q, W) \triangleq \mathcal{L}_\alpha^{\max}(S \to Y)$. The analogues for utility are defined in a similar way.

Theorem 11. *Let $\hat{P}_n$ be the empirical distribution obtained from n i.i.d. samples drawn from P. Then, with probability at least $1 - \beta$, for any $W \in \mathcal{W}$, we have*

$$|\mathcal{L}(\hat{P}_n, W) - \mathcal{L}(P, W)|$$
$$\leq \sqrt{\frac{2}{n}\left(|\mathcal{S}| \cdot |\mathcal{X}| - \log \beta\right)} \cdot \begin{cases} 1 & \text{when } \mathcal{L} = \mathcal{L}_c, \\ C_{f,\overline{m}_S} & \text{when } \mathcal{L} = \mathcal{L}_f \text{ with } f \text{ locally Lipschitz}, \\ \frac{2\alpha}{\alpha - 1}|\mathcal{S}|^{1-1/\alpha} & \text{when } \mathcal{L} = \mathcal{L}_\alpha^{\mathrm{A}} \text{ with } \alpha \in (1, \infty], \\ \frac{2\alpha + 1}{(\alpha - 1)\overline{m}_S^{1-1/\alpha}} & \text{when } \mathcal{L} = \mathcal{L}_\alpha^{\mathrm{S}} \text{ with } \alpha \in (1, \infty), \\ \frac{2}{\min_s \sum_x \hat{P}_n(s,x)} & \text{when } \mathcal{L} = \mathcal{L}_\infty^{\mathrm{S}} \text{ or } \mathcal{L} = \mathcal{L}_\infty^{\max}, \\ \frac{4\alpha|\mathcal{S}|^{1-1/\alpha}}{(\alpha - 1)\min_s \sum_x \hat{P}_n(s,x)} & \text{when } \mathcal{L} = \mathcal{L}_\alpha^{\max} \text{ with } \alpha \in (1, \infty), \end{cases} \tag{6.33}$$

$$
\begin{aligned}
&|\mathcal{U}(\hat{P}_n, W) - \mathcal{U}(P, W)| \\
&\leq \sqrt{\frac{2}{n}\left(|\mathcal{S}| \cdot |\mathcal{X}| - \log \beta\right)} \cdot \begin{cases} 1 & \text{when } \mathcal{U} = \mathcal{U}_c, \\ C_{f,\overline{m}_X} & \text{when } \mathcal{U} = \mathcal{U}_f \text{ with } f \text{ locally Lipschitz}, \\ \frac{2\alpha}{\alpha-1}|\mathcal{X}|^{1-1/\alpha} & \text{when } \mathcal{U} = \mathcal{U}_\alpha^{\mathrm{A}} \text{ with } \alpha \in (1,\infty], \\ \frac{1}{(\alpha-1)\overline{m}_X^{1-1/\alpha}} & \text{when } \mathcal{U} = \mathcal{U}_\alpha^{\mathrm{S}} \text{ with } \alpha \in (1,\infty), \\ 0 & \text{when } \mathcal{U} = \mathcal{U}_\infty^{\mathrm{S}} \text{ or } \mathcal{U} = \mathcal{U}_\alpha^{\max} \text{ with } \alpha \in (1,\infty], \end{cases}
\end{aligned} \tag{6.34}
$$

where $C_{f,u} = 2K_{f,u^{-1}} + (2u^{-1}+1)L_{f,u^{-1}}$ *with* $K_{g,u}$, $L_{g,u}$, $\overline{m}_S$ *and* $\overline{m}_X$ *as defined in* (6.40), (6.41), (6.55), *and* (6.56), *respectively.*

We illustrate the discrepancies of privacy-utility guarantees in (6.33) and (6.34) along with the corresponding upper bounds through a synthetic and a real-world datasets in Section 6.7.

Convergence of Optimal Privacy Mechanisms

Assume there is a sequence of joint distributions $(P_n)_{n=1}^{\infty}$ converging to the true distribution P. The results provided in Section 6.4 establish that the privacy-utility guarantees for P_n converge to those of P as n goes to infinity. Nonetheless, they cannot be used to study the convergence properties of the optimal privacy mechanisms of P_n as defined in (6.30). In Section 6.5 we address the convergence of optimal privacy mechanisms by further exploiting some of the Lipschitz continuity properties established for the information leakage measures under consideration. To be more specific, assume that the privacy mechanism designer constructs an optimal privacy mechanism W_n^* for each P_n, i.e., $W_n^* \in \mathcal{W}^*(P_n;\epsilon)$. We establish the convergence of the sequence of optimal privacy mechanisms $(W_n^*)_{n=1}^{\infty}$ to the set of optimal privacy mechanisms for P, i.e., $\mathcal{W}^*(P;\epsilon)$. Due to technical considerations, this result covers only probability of correctly guessing, f-information with f locally Lipschitz, and Arimoto's mutual information of order α with $\alpha \in (1,\infty]$.

Uniform Privacy Mechanisms

In applications where privacy is a priority, a specific privacy guarantee for the true distribution P may be required, even though the designer has only access to a distribution $\hat{P}$ estimated from samples. We propose the following procedure to overcome this difficulty: (a) use large deviations results to find a probabilistic upper bound, say r, for the distance between $\hat{P}$ and P; (b) design privacy

mechanisms that deliver the required privacy guarantee for *all* distributions at distance less or equal than r from $\hat{P}$. Based on this procedure, we introduce the definition of uniform privacy mechanisms in Section 6.6. We prove that *optimal* uniform privacy mechanisms exist and while their design might be challenging, they can be efficiently approximated by optimal privacy mechanisms for $\hat{P}$ as defined in (6.30). We finish Section 6.6 establishing convergence properties of optimal uniform privacy mechanisms similar to those in Section 6.5.

6.4 Discrepancy of Privacy-Utility Guarantees

We provide probabilistic upper bounds for the difference between the privacy-utility guarantees of the empirical and the true distributions. These upper bounds do not depend on the specific privacy mechanism. In order to simplify the exposition, we assume that privacy leakage and utility are measured using the same information metric. Nonetheless, the case when different information metrics are used can be handled in a straightforward manner. In this section, we study five different information metrics: probability of correctly guessing, f-information with f locally Lipschitz, Arimoto's mutual information (α-leakage) of order $\alpha > 1$, Sibson's mutual information of order $\alpha > 1$, and maximal α-leakage of order $\alpha > 1$.

6.4.1 Probability of Correctly Guessing

The main result of this subsection relies on the following lemma which establishes the Lipschitz continuity of the mappings $\mathcal{L}_c(\cdot, W)$ and $\mathcal{U}_c(\cdot, W)$. Recall that for $Q \in \mathcal{P}$ and $W \in \mathcal{W}$,

$$\mathcal{L}_c(Q, W) \triangleq P_c(S|Y) \quad \text{and} \quad \mathcal{U}_c(Q, W) \triangleq P_c(X|Y), \tag{6.35}$$

where $S \to X \to Y$ is such that $P_{S,X} = Q$ and $P_{Y|X} = W$.

Lemma 15. *For any $Q_1, Q_2 \in \mathcal{P}$ and $W \in \mathcal{W}$, we have*

$$|\mathcal{L}_c(Q_1, W) - \mathcal{L}_c(Q_2, W)| \leq \|Q_1 - Q_2\|_1, \tag{6.36}$$

$$|\mathcal{U}_c(Q_1, W) - \mathcal{U}_c(Q_2, W)| \leq \|Q_1 - Q_2\|_1. \tag{6.37}$$

Proof. See Appendix D.1.1. □

In particular, when $Q_1 = \hat{P}$ is an estimate of $Q_2 = P$, the previous lemma implies that any upper

bound on $\|\hat{P} - P\|_1$ translates into an upper bound for difference of the corresponding privacy-utility guarantees. The following theorem specializes this observation to the empirical distribution by means of the large deviations inequality in (6.27).

Theorem 12. *Let $\hat{P}_n$ be the empirical distribution obtained from n i.i.d. samples drawn from P. Then, with probability at least $1 - \beta$, for any $W \in \mathcal{W}$, we have*

$$|\mathcal{L}_c(\hat{P}_n, W) - \mathcal{L}_c(P, W)| \leq \sqrt{\frac{2}{n}\left(|\mathcal{S}| \cdot |\mathcal{X}| - \log \beta\right)}, \tag{6.38}$$

$$|\mathcal{U}_c(\hat{P}_n, W) - \mathcal{U}_c(P, W)| \leq \sqrt{\frac{2}{n}\left(|\mathcal{S}| \cdot |\mathcal{X}| - \log \beta\right)}. \tag{6.39}$$

In Section 6.7.2 we illustrate the fitness of the bounds in Theorem 12 when applied to ProPublica's COMPAS dataset [10].

Remark 9. Lemma 15 and Theorem 12 illustrate the general technique we use to derive (probabilistic) upper bounds for the difference between the privacy-utility guarantees of the empirical and the true distributions. Specifically, we relate the difference between the privacy-utility guarantees of two joint distributions with their ℓ_1 distance, and then we use large deviations results to provide upper bound for the ℓ_1 distance between the empirical and the true distributions.

6.4.2 f-Information with f Locally Lipschitz

We start establishing the following notation. For a given function $g : [0, \infty) \to \mathbb{R}$ and $u > 0$, we let

$$K_{g,u} \triangleq \sup\{|g(t)| : t \in [0, u]\}. \tag{6.40}$$

The constant $K_{g,u}$ is the so-called supremum norm of g on $[0, u]$. In addition, if g is Liptschitz on $[0, u]$, we let $L_{g,u}$ be its Lipschitz constant on $[0, u]$, i.e.,

$$L_{g,u} \triangleq \min\{L \geq 0 : |g(t_1) - g(t_2)| \leq L|t_1 - t_2|, \forall t_1, t_2 \in [0, u]\}. \tag{6.41}$$

A function $g : [0, \infty) \to \mathbb{R}$ is called locally Lipschitz if g is Lipschitz on $[0, u]$ for every $u > 0$. Note that a locally Lipschitz function is not necessarily Lipschitz on $[0, \infty)$. For example, the function $g(t) = t^2$ is locally Lipschitz with $L_{g,u} = 2u$ for all $u > 0$, but it is not Lipschitz on $[0, \infty)$. For any

two distributions $Q_1, Q_2 \in \mathcal{P}$, we denote

$$m_S \triangleq \min\left\{\sum_{x\in\mathcal{X}} Q_i(s,x) : s \in \mathcal{S}, i \in \{1,2\}\right\}, \tag{6.42}$$

$$m_X \triangleq \min\left\{\sum_{s\in\mathcal{S}} Q_i(s,x) : x \in \mathcal{X}, i \in \{1,2\}\right\}. \tag{6.43}$$

Observe that m_S (resp. m_X) equals the probability of the least likely symbol among the marginal distributions over $\mathcal{S}$ (resp. $\mathcal{X}$) of Q_1 and Q_2. In other words, if (S_i, X_i) has joint distribution Q_i for each $i \in \{1,2\}$, then

$$m_S = \min\{P_{S_i}(s) : s \in \mathcal{S}, i \in \{1,2\}\}, \tag{6.44}$$

$$m_X = \min\{P_{X_i}(x) : x \in \mathcal{X}, i \in \{1,2\}\}. \tag{6.45}$$

The following lemma serves as an analogue of Lemma 15 for an f-information with f locally Lipschitz. Recall that for $Q \in \mathcal{P}$ and $W \in \mathcal{W}$,

$$\mathcal{L}_f(Q,W) \triangleq I_f(P_{S,Y}) \quad \text{and} \quad \mathcal{U}_f(Q,W) \triangleq I_f(P_{X,Y}), \tag{6.46}$$

where $S \to X \to Y$ is such that $P_{S,X} = Q$ and $P_{Y|X} = W$.

Lemma 16. *If $f : [0,\infty) \to \mathbb{R}$ is locally Lipschitz, then, for any $Q_1, Q_2 \in \mathcal{P}$ and $W \in \mathcal{W}$, we have*

$$|\mathcal{L}_f(Q_1,W) - \mathcal{L}_f(Q_2,W)| \le C_{f,m_S}\|Q_1 - Q_2\|_1, \tag{6.47}$$

$$|\mathcal{U}_f(Q_1,W) - \mathcal{U}_f(Q_2,W)| \le C_{f,m_X}\|Q_1 - Q_2\|_1, \tag{6.48}$$

where, for $u \in \{m_S, m_X\}$, $C_{f,u} \triangleq 2K_{f,u^{-1}} + (2u^{-1}+1)L_{f,u^{-1}}$.

Proof. See Appendix D.1.2. □

Remark 10. Despite the similarity between Lemmas 15 and 16, the latter does not imply that the mapping $\mathcal{L}_f(\cdot, W)$ is Lipschitz continuous. Indeed, the factor C_{f,m_S} depends on Q_1 and Q_2 through m_S. Nonetheless, Lemma 16 does show that the local Lipschitzianity of f is bequeathed to $\mathcal{L}_f(\cdot, W)$. Specifically, for every $\delta > 0$, the mapping $\mathcal{L}_f(\cdot, W)$ is Lipschitz continuous over

$$\left\{Q \in \mathcal{P} : \delta \le \min_{s\in\mathcal{S}} \sum_{x\in\mathcal{X}} Q(s,x)\right\}, \tag{6.49}$$

and its Lipschitz constant is less than or equal to $C_{f,\delta}$. Indeed, this assertion follows from the fact that $u \mapsto C_{f,u}$ is a non-increasing function, as exhibited in Appendix D.1.3. Hence, any lower bound

for m_S translates into an upper bound for C_{f,m_S} in (6.47). As shown below, the *local Lipschitzianity* of $\mathcal{L}_f(\cdot, W)$ is enough to derive a bound similar to (6.38). A similar discussion can be held for $\mathcal{U}_f(\cdot, W)$.

We illustrate the value of $C_{f,u}$, defined in Lemma 16, through the following examples:

- *Total variation distance.* If $f(t) = |t-1|/2$, then f is a (globally) Lipschitz function with $L_{f,u} = 1/2$ and $K_{f,u} = \max\{1, u-1\}/2$ for all $u > 0$. Consequently,

$$C_{f,u} = u^{-1} \max\{1 + 3u/2, 2 - u/2\}; \tag{6.50}$$

- *χ^2-divergence.* If $f(t) = (t-1)^2$, then f is a locally Lipschitz function with $L_{f,u} = 2\max\{1, u-1\}$ and $K_{f,u} = \max\{1, (u-1)^2\}$ for all $u > 0$. Consequently,

$$C_{f,u} = 2u^{-2} \max\{2u + 2u^2, 3 - 3u\}; \tag{6.51}$$

- *Hellinger divergence.* If $f_\alpha(t) = \dfrac{t^\alpha - 1}{\alpha - 1}$ with $\alpha > 1$, then f_α is locally Lipschitz with $L_{f_\alpha,u} = \dfrac{\alpha u^{\alpha-1}}{\alpha - 1}$ and $K_{f,u} = \frac{1}{\alpha-1} \max\{1, u^\alpha - 1\}$ for all $u > 0$. Consequently,

$$C_{f,u} = \frac{u^{-\alpha}}{\alpha - 1} \max\{2\alpha + \alpha u + 2u^\alpha, (2\alpha + 2) + \alpha u - 2u^\alpha\}. \tag{6.52}$$

Note, however, that mutual information cannot be handled by Lemma 16 since the function $f(t) = t\log(t)$ is not locally Lipschitz. In fact, mutual information seems to have a different asymptotic behavior w.r.t. the sample size n when compared to the theorems in this chapter, cf. [245, Theorem 3].

Relying on Lemma 16 and the large deviations inequality in (6.27), the following theorem provides a data dependent bound for the discrepancy between the guarantees provided for the empirical and the true distributions. For $x \in \mathbb{R}$, we define $(x)_+ \triangleq \max\{0, x\}$.

Theorem 13. *Let $\hat{P}_n$ be the empirical distribution obtained from n i.i.d. samples drawn from P. If $f : [0, \infty) \to \mathbb{R}$ is locally Lipschitz, then, with probability at least $1 - \beta$, for any $W \in \mathcal{W}$, we have*

$$|\mathcal{L}_f(\hat{P}_n, W) - \mathcal{L}_f(P, W)| \le C_{f,\overline{m}_S} \sqrt{\frac{2}{n}(|\mathcal{S}| \cdot |\mathcal{X}| - \log\beta)}, \tag{6.53}$$

$$|\mathcal{U}_f(\hat{P}_n, W) - \mathcal{U}_f(P, W)| \le C_{f,\overline{m}_X} \sqrt{\frac{2}{n}(|\mathcal{S}| \cdot |\mathcal{X}| - \log\beta)}, \tag{6.54}$$

where, for $u \in \{\overline{m}_S, \overline{m}_X\}$, $C_{f,u} \triangleq 2K_{f,u^{-1}} + (2u^{-1}+1)L_{f,u^{-1}}$,

$$\overline{m}_S \triangleq \left(\min\left\{ \sum_{x\in\mathcal{X}} \hat{P}_n(s,x) : s \in \mathcal{S} \right\} - \sqrt{\frac{2}{n}\left(|\mathcal{S}|\cdot|\mathcal{X}| - \log\beta\right)} \right)_+, \tag{6.55}$$

$$\overline{m}_X \triangleq \left(\min\left\{ \sum_{s\in\mathcal{S}} \hat{P}_n(s,x) : x \in \mathcal{X} \right\} - \sqrt{\frac{2}{n}\left(|\mathcal{S}|\cdot|\mathcal{X}| - \log\beta\right)} \right)_+. \tag{6.56}$$

Proof. See Appendix D.1.3. □

Remark 11. Observe that up to an additive factor of $(2(|\mathcal{S}|\cdot|\mathcal{X}| - \log\beta)/n)^{1/2}$, which is negligible in the large n regime, $\overline{m}_S$ equals the probability of the least likely symbol of the marginal distribution over $\mathcal{S}$ of $\hat{P}_n$. As a consequence, the bound in (6.53) is data-dependent, while the bound in (6.38) is data-independent. This discrepancy comes mainly from the fact that $\mathcal{L}_c(\cdot, W)$ is Lipschitz continuous, as established in Lemma 15, while $\mathcal{L}_f(\cdot, W)$ is *locally Lipschitz* in the sense of Remark 10. A similar discussion holds for utility.

In most practical scenarios, the alphabet of X is significantly larger than the alphabet of S, e.g., S might be political preference while X is movie ratings, S might be private household information while X is smart meter data, S might be gender while X is a profile picture, etc. In particular, when the sample size is limited, the bounds in Theorem 13 might become ineffective due to the potentially small value of $\overline{m}_X$. In order to alleviate this issue, we propose the following pre-processing technique which combines the symbols of $\mathcal{X}$ with less observations in the dataset.

Given $\gamma \geq 0$ and a symbol x_0 not belonging to $\mathcal{X}$, we let Π_γ be the mapping with input alphabet $\mathcal{X}$, output alphabet

$$\mathcal{X}_\gamma \triangleq \{x_0\} \cup \left\{ x \in \mathcal{X} : \sum_{s\in\mathcal{S}} \hat{P}(s,x) \geq \gamma \right\}, \tag{6.57}$$

and determined by

$$\Pi_\gamma(x) = \begin{cases} x & \text{if } \sum_s \hat{P}(s,x) \geq \gamma, \\ x_0 & \text{otherwise.} \end{cases} \tag{6.58}$$

The proposed pre-processing technique consists in applying this mapping to the samples $\{(s_i, x_i)\}_{i=1}^n$ to obtain the modified samples $\{(s_i, \Pi_\gamma(x_i))\}_{i=1}^n$. Clearly, this pre-processing technique improves the bounds in Theorem 13 by increasing $\overline{m}_X$. However, this improvement might come at a price in utility, as shown in the following proposition.

Proposition 15. *Let* $\gamma \geq 0$ *be given and let* $\hat{P}_n$ *be the empirical distribution of* n *samples* $\{(s_i, x_i)\}_{i=1}^n$. *If*

$\hat{P}_\gamma$ is the empirical distribution of the modified samples $\{(s_i, \Pi_\gamma(x_i))\}_{i=1}^n$, then, for any $\epsilon \in \mathbb{R}$ such that $\mathcal{W}^*(\hat{P}_\gamma; \epsilon) \neq \emptyset$,

$$\mathsf{H}_f(\hat{P}_\gamma; \epsilon) \leq \mathsf{H}_f(\hat{P}_n; \epsilon), \tag{6.59}$$

where H_f denotes the privacy-utility function when both privacy leakage and utility are measured using f-information.

Proof. See Appendix D.1.4. □

The proof of Proposition 15 relies on the following lemma.

Lemma 17. *Let $\gamma \geq 0$ be given. If $X \to X_0 \to Y_0$ is a Markov chain with $X_0 = \Pi_\gamma(X)$, then, for every f-information,*

$$I_f(P_{X,Y_0}) = I_f(P_{X_0,Y_0}). \tag{6.60}$$

Observe that Lemma 17 is an immediate consequence of the data processing inequality (DPI) for f-information [221, Remark 2.3] and the fact that X_0 is a deterministic function of X. Indeed, if $\Pi : \mathcal{X} \to \mathcal{X}_0$ is any deterministic function and $X_0 \triangleq \Pi(X)$, then any random variable Y_0 satisfies $X_0 \to X \to Y_0$. In particular, the DPI implies that $I_f(P_{X,Y_0}) \geq I_f(P_{X_0,Y_0})$. If, in addition, Y_0 is a randomized function of X_0, i.e., $X \to X_0 \to Y_0$, then the DPI leads to $I_f(P_{X,Y_0}) \leq I_f(P_{X_0,Y_0})$.

6.4.3 Arimoto's Mutual Information, Sibson's Mutual Information, and Maximal α-Leakage

We now establish continuity properties of Arimoto's mutual information, Sibson's mutual information, and maximal α-leakage similar to those proved in Lemmas 15 and 16 for probability of correctly guessing and f-information, respectively. Upon these properties, we state two theorems regarding the discrepancy of privacy-utility guarantees for the information measures at hand.

Recall that for $Q \in \mathcal{P}$ and $W \in \mathcal{W}$,

$$\mathcal{L}_\alpha^{\mathrm{A}}(Q, W) \triangleq I_\alpha^{\mathrm{A}}(P_{S,Y}) \quad \text{and} \quad \mathcal{U}_\alpha^{\mathrm{A}}(Q, W) \triangleq I_\alpha^{\mathrm{A}}(P_{X,Y}), \tag{6.61}$$

where $S \to X \to Y$ is such that $P_{S,X} = Q$ and $P_{Y|X} = W$. The next lemma shows the Lipschitz continuity of $\mathcal{L}_\alpha^{\mathrm{A}}(\cdot, W)$ and $\mathcal{U}_\alpha^{\mathrm{A}}(\cdot, W)$.

Lemma 18. *Let $\alpha \in (1,\infty]$. For any $Q_1, Q_2 \in \mathcal{P}$ and $W \in \mathcal{W}$,*

$$|\mathcal{L}_\alpha^{\mathrm{A}}(Q_1,W) - \mathcal{L}_\alpha^{\mathrm{A}}(Q_2,W)| \le \frac{2\alpha}{\alpha-1}|\mathcal{S}|^{1-1/\alpha}\|Q_1 - Q_2\|_1, \tag{6.62}$$

$$|\mathcal{U}_\alpha^{\mathrm{A}}(Q_1,W) - \mathcal{U}_\alpha^{\mathrm{A}}(Q_2,W)| \le \frac{2\alpha}{\alpha-1}|\mathcal{X}|^{1-1/\alpha}\|Q_1 - Q_2\|_1. \tag{6.63}$$

Proof. See Appendix D.1.5. □

Remark 12. Recall that probability of correctly guessing and Arimoto's mutual information of order ∞ are related through the formula

$$I_\infty^{\mathrm{A}}(U;V) = \log \frac{P_c(U|V)}{P_c(U)}. \tag{6.64}$$

Observe that, despite this relation, the bound for probability of correctly guessing in (6.36) and the corresponding bound for Arimoto's mutual information of order ∞ in (6.62) differ by a factor of $2|\mathcal{S}|$. As shown in the proof of Lemma 18, the $|\mathcal{S}|$ factor comes from the log in (6.64) via the minimum in the following inequality

$$\left|\log \frac{a}{b}\right| \le \frac{|a-b|}{\min\{a,b\}}, \qquad a,b>0. \tag{6.65}$$

A similar argument explains the extra factor of $2|\mathcal{X}|$ appearing in (6.63) w.r.t. its counterpart in (6.37).

Now we consider Sibson's mutual information. Recall that for $Q \in \mathcal{P}$ and $W \in \mathcal{W}$,

$$\mathcal{L}_\alpha^{\mathrm{S}}(Q,W) \triangleq I_\alpha^{\mathrm{S}}(P_{S,Y}) \quad \text{and} \quad \mathcal{U}_\alpha^{\mathrm{S}}(Q,W) \triangleq I_\alpha^{\mathrm{S}}(P_{X,Y}), \tag{6.66}$$

where $S \to X \to Y$ is such that $P_{S,X} = Q$ and $P_{Y|X} = W$. The following lemma establishes that the mappings $\mathcal{L}_\alpha^{\mathrm{S}}(\cdot,W)$ and $\mathcal{U}_\alpha^{\mathrm{S}}(\cdot,W)$ have similar continuity properties to those of $\mathcal{L}_f(\cdot,W)$ and $\mathcal{U}_f(\cdot,W)$ with f locally Lipschitz.

Lemma 19. *For $Q_1, Q_2 \in \mathcal{P}$, we define m_S and m_X as in (6.42) and (6.43), respectively. If $\alpha \in (1,\infty)$, then, for any $W \in \mathcal{W}$,*

$$|\mathcal{L}_\alpha^{\mathrm{S}}(Q_1,W) - \mathcal{L}_\alpha^{\mathrm{S}}(Q_2,W)| \le \frac{2\alpha+1}{\alpha-1}\frac{\|Q_1 - Q_2\|_1}{m_S^{1-1/\alpha}}, \tag{6.67}$$

$$|\mathcal{U}_\alpha^{\mathrm{S}}(Q_1,W) - \mathcal{U}_\alpha^{\mathrm{S}}(Q_2,W)| \le \frac{1}{\alpha-1}\frac{\|Q_1 - Q_2\|_1}{m_X^{1-1/\alpha}}. \tag{6.68}$$

If $\alpha = \infty$, then, for any $W \in \mathcal{W}$,

$$|\mathcal{L}_\infty^{\mathrm{S}}(Q_1, W) - \mathcal{L}_\infty^{\mathrm{S}}(Q_2, W)| \leq \frac{2\|Q_1 - Q_2\|_1}{\min_s \sum_x Q_1(s,x)}, \tag{6.69}$$

$$|\mathcal{U}_\infty^{\mathrm{S}}(Q_1, W) - \mathcal{U}_\infty^{\mathrm{S}}(Q_2, W)| = 0. \tag{6.70}$$

Proof. See Appendix D.1.6. □

Remark 13. A straightforward manipulation shows that

$$\exp\left(I_\infty^{\mathrm{A}}(P_{U,V})\right) = \sum_{v\in\mathcal{V}} \max_{u\in\mathcal{U}} \frac{P_U(u)}{\max_{u'} P_U(u')} P_{V|U}(v|u), \tag{6.71}$$

$$\exp\left(I_\infty^{\mathrm{S}}(P_{U,V})\right) = \sum_{v\in\mathcal{V}} \max_{u\in\mathcal{U}} P_{V|U}(v|u). \tag{6.72}$$

From (6.71) and (6.72), we observe that Sibson's mutual information of order infinity is independent of the input distribution P_U, while Arimoto's mutual information of the same order is not. By comparing (6.62) and (6.69), we also observe that the privacy leakage bound for Sibson's mutual information of order infinity *does* depend on the probability of the least likely symbol of the input[2], while its Arimoto's counterpart does not. These seemingly contradicting facts have a rather intuitive cause: it is hard to estimate $P_{V|U}(\cdot|u)$ reliably if $P_U(u)$ is small. Observe that while $P_{V|U}(v|u)$ appears in both (6.71) and (6.72), the factor $P_U(u)/\max_{u'} P_U(u')$ in (6.71) makes Arimoto's mutual information of order infinity less dependent on symbols u with small probabilities $P_U(u)$. The difficulty in estimating Sibson's mutual information in comparison to its Arimoto counterpart is also natural from a privacy perspective. As proved by Issa *et al.* in [133], Sibson's mutual information guarantees privacy for *any* (possibly randomized) function of S, while Arimoto's mutual information guarantees privacy only for S itself.

We end up this section showing that maximal α-leakage ($\alpha > 1$) behaves in a similar way to Sibson's mutual information of order ∞. Recall that for $Q \in \mathcal{P}$ and $W \in \mathcal{W}$,

$$\mathcal{L}_\alpha^{\max}(Q, W) \triangleq \sup_{P_{\tilde{S}}} I_\alpha^{\mathrm{A}}\left(P_{\tilde{S}} \cdot (P_{X|S} W)\right) \quad \text{and} \quad \mathcal{U}_\alpha^{\max}(Q, W) \triangleq \sup_{P_{\tilde{X}}} I_\alpha^{\mathrm{A}}\left(P_{\tilde{X}} \cdot W\right), \tag{6.73}$$

where $S \to X \to Y$ is such that $P_{S,X} = Q$ and $P_{Y|X} = W$.

[2]In recent work [133], Issa *et al.* established a bound similar to (6.78) which also depends on the probability of the least likely symbol of the input. Thus, this dependency seems unavoidable at the moment.

Lemma 20. *Let $\alpha \in (1,\infty)$. For any $Q_1, Q_2 \in \mathcal{P}$ and $W \in \mathcal{W}$,*

$$|\mathcal{L}_\alpha^{\max}(Q_1, W) - \mathcal{L}_\alpha^{\max}(Q_2, W)| \le \frac{4\alpha}{\alpha - 1} \frac{|\mathcal{S}|^{1-1/\alpha}}{\min_s \sum_x Q_1(s,x)} \|Q_1 - Q_2\|_1, \tag{6.74}$$

$$|\mathcal{U}_\alpha^{\max}(Q_1, W) - \mathcal{U}_\alpha^{\max}(Q_2, W)| = 0. \tag{6.75}$$

Proof. See Appendix D.1.7. □

We summarize the probabilistic upper bounds derived in this subsection in the following two theorems. They follow immediately from Lemma 18, 19, 20 and the large deviations inequality in (6.27).

Theorem 14. *Let $\hat{P}_n$ be the empirical distribution obtained from n i.i.d. samples drawn from P. Then, with probability at least $1-\beta$, for any $W \in \mathcal{W}$, we have*

$$|\mathcal{L}(\hat{P}_n, W) - \mathcal{L}(P, W)| \le \sqrt{\frac{2}{n}\left(|\mathcal{S}| \cdot |\mathcal{X}| - \log\beta\right)} \cdot \begin{cases} \frac{2\alpha}{\alpha-1}|\mathcal{S}|^{1-1/\alpha} & \text{when } \mathcal{L} = \mathcal{L}_\alpha^{\mathrm{A}} \text{ with } \alpha \in (1,\infty], \\ \frac{2\alpha+1}{(\alpha-1)\overline{m}_S^{1-1/\alpha}} & \text{when } \mathcal{L} = \mathcal{L}_\alpha^{\mathrm{S}} \text{ with } \alpha \in (1,\infty), \end{cases} \tag{6.76}$$

$$|\mathcal{U}(\hat{P}_n, W) - \mathcal{U}(P, W)| \le \sqrt{\frac{2}{n}\left(|\mathcal{S}| \cdot |\mathcal{X}| - \log\beta\right)} \cdot \begin{cases} \frac{2\alpha}{\alpha-1}|\mathcal{X}|^{1-1/\alpha} & \text{when } \mathcal{U} = \mathcal{U}_\alpha^{\mathrm{A}} \text{ with } \alpha \in (1,\infty], \\ \frac{1}{(\alpha-1)\overline{m}_X^{1-1/\alpha}} & \text{when } \mathcal{U} = \mathcal{U}_\alpha^{\mathrm{S}} \text{ with } \alpha \in (1,\infty), \end{cases} \tag{6.77}$$

where $\overline{m}_S$ and $\overline{m}_X$ are defined in (6.55) and (6.56), respectively.

Theorem 15. *Let $\hat{P}_n$ be the empirical distribution obtained from n i.i.d. samples drawn from P. Then, with probability at least $1-\beta$, for any $W \in \mathcal{W}$, we have*

$$|\mathcal{L}(\hat{P}_n, W) - \mathcal{L}(P, W)| \le \frac{2\sqrt{\frac{2}{n}\left(|\mathcal{S}| \cdot |\mathcal{X}| - \log\beta\right)}}{\min_s \sum_x \hat{P}_n(s,x)} \cdot \begin{cases} 1 & \text{when } \mathcal{L} = \mathcal{L}_\infty^{\mathrm{S}} \text{ or } \mathcal{L} = \mathcal{L}_\infty^{\max}, \\ \frac{2\alpha|\mathcal{S}|^{1-1/\alpha}}{\alpha-1} & \text{when } \mathcal{L} = \mathcal{L}_\alpha^{\max} \text{ with } \alpha \in (1,\infty), \end{cases} \tag{6.78}$$

and $|\mathcal{U}(\hat{P}_n, W) - \mathcal{U}(P, W)| = 0$ when $\mathcal{U} = \mathcal{U}_\infty^{\mathrm{S}}$ or $\mathcal{U} = \mathcal{U}_\alpha^{\max}$ with $\alpha \in (1,\infty]$.

6.5 Convergence of Optimal Privacy Mechanisms

Now we study the consistency of the optimal privacy mechanisms using, among other results, the (local) Lipschitz continuity of the mappings $\mathcal{L}(\cdot, W)$ and $\mathcal{U}(\cdot, W)$ established in the previous section.

Assume that the (privacy) mechanism designer is given an $\epsilon \in \mathbb{R}$ and a sequence of joint distributions $(P_n)_{n=1}^{\infty}$ converging to the true distribution P. Furthermore, the designer constructs an optimal ϵ-private mechanism W_n^* for each P_n. In this section, we analyze some convergence properties of the sequence of optimal privacy mechanisms $(W_n^*)_{n=1}^{\infty}$.

In the sequel, we assume that prior knowledge about the true distribution P might be available. Accordingly, we let $\mathcal{Q} \subseteq \mathcal{P}$ be the set of all joint distributions compatible with such prior knowledge. Note that when no prior knowledge is available, we simply let $\mathcal{Q} = \mathcal{P}$. The main results of this and the following section are established under the following conditions.

There exist a closed set $\mathcal{Q} \subseteq \mathcal{P}$ and $N \in \mathbb{N}$ such that

(**C**.1) *Continuity.* For every $Q \in \mathcal{Q}$, the mappings $\mathcal{L}(Q, \cdot)$ and $\mathcal{U}(Q, \cdot)$ are continuous over

$$\mathcal{W}_N \triangleq \mathcal{W} \cap \mathbb{R}^{|\mathcal{X}| \times N}, \tag{6.79}$$

where $\mathcal{W}$, as defined in (6.28), denotes the set of all privacy mechanisms. Furthermore, the mapping $\mathsf{H}(Q; \cdot)$ is continuous over $[\epsilon_{\min}(Q), \infty)$, where

$$\epsilon_{\min}(Q) \triangleq \inf \{\mathcal{L}(Q, W) : \ W \in \mathcal{W}\}. \tag{6.80}$$

(**C**.2) *Lipschitz Continuity.* There exist positive constants C_L and C_U such that, for all $Q_1, Q_2 \in \mathcal{Q}$ and $W \in \mathcal{W}_N$,

$$|\mathcal{L}(Q_1, W) - \mathcal{L}(Q_2, W)| \leq C_L \|Q_1 - Q_2\|_1, \tag{6.81}$$

$$|\mathcal{U}(Q_1, W) - \mathcal{U}(Q_2, W)| \leq C_U \|Q_1 - Q_2\|_1. \tag{6.82}$$

(**C**.3) *Support.* For each $Q \in \mathcal{Q}$ and $\epsilon \geq \epsilon_{\min}(Q)$, the intersection $\mathcal{W}^*(Q; \epsilon) \cap \mathcal{W}_N$ is not empty.

These three conditions might seem restrictive at a first glance. Nonetheless, as shown in Section 6.5.1 below, they are satisfied by probability of correctly guessing, f-information with f locally Lipschitz, and Arimoto's mutual information of order α with $\alpha \in (1, \infty]$.

Note that condition (**C**.1) is a continuity requirement. Specifically, it requires the privacy leakage and utility functions to be continuous w.r.t. the privacy mechanism W. Also, it requires the privacy-

utility function to be continuous w.r.t. the privacy parameter ϵ. Similarly, condition (**C.**2) requires the privacy leakage and utility functions to be Lipschitz continuous w.r.t. the joint distribution Q. Observe that the constants C_L and C_U in (6.81) and (6.82), respectively, do not depend on the joint distributions Q_1 and Q_2 neither on the privacy mechanism W. Nonetheless, these constants may depend on $\mathcal{Q}$ and N.

Remark 14. Observe that the set $\mathcal{W}_N$ is in correspondence with the set of all privacy mechanisms from $\mathcal{X}$ to $\{1, \ldots, N\}$. In particular, condition (**C.**3) requires that, for each $Q \in \mathcal{Q}$ and $\epsilon \geq \epsilon_{\min}(Q)$, there exists an optimal ϵ-private mechanism for Q supported over $\{1, \ldots, N\}$. As a consequence,

$$\mathsf{H}(Q;\epsilon) = \max_{W \in \mathcal{D}_N(Q;\epsilon)} \mathcal{U}(Q,W), \tag{6.83}$$

where $\mathcal{D}_N(Q;\epsilon) \triangleq \{W \in \mathcal{W}_N : \mathcal{L}(Q,W) \leq \epsilon\}$. Furthermore, the intersection of $\mathcal{W}_N$ and all optimal privacy mechanisms is defined as

$$\mathcal{W}_N^*(P;\epsilon) \triangleq \{W \in \mathcal{W}_N : \mathcal{L}(P,W) \leq \epsilon,\ \mathcal{U}(P,W) = \mathsf{H}(P;\epsilon)\}. \tag{6.84}$$

By comparing (6.29) and (6.83), we can observe that condition (**C.**3) allows us to replace the unbounded set $\mathcal{W}$ with the compact space $\mathcal{W}_N$ in the optimization defining the privacy-utility function H. This technical difference is crucial in the proofs of the subsequent results. In a similar spirit, observe that condition (**C.**3) also implies that

$$\epsilon_{\min}(Q) = \min\left\{\mathcal{L}(Q,W) :\ W \in \mathcal{W}_N\right\}. \tag{6.85}$$

A two-variable function might be continuous in each argument without being jointly continuous, see, e.g., [235, Ex. 4.7]. The following lemma is an immediate, yet important, consequence of conditions (**C.**1–2), as it establishes the joint continuity of the privacy leakage and utility functions.

Lemma 21. *Assume that conditions (**C.**1–2) hold true for a given closed set $\mathcal{Q} \subseteq \mathcal{P}$ and a given $N \in \mathbb{N}$. The functions $\mathcal{L}(\cdot,\cdot)$ and $\mathcal{U}(\cdot,\cdot)$ are continuous over $\mathcal{Q} \times \mathcal{W}_N$.*

Proof. See Appendix D.2.1. □

Building upon Lemma 21, we establish the following proposition which plays a key role in the proofs of the main results of this section.

Proposition 16. *Assume that conditions (**C.**1–3) hold true for a given closed set $\mathcal{Q} \subseteq \mathcal{P}$ and a given $N \in \mathbb{N}$. For any given $\epsilon \in \mathbb{R}$, the mapping $\mathsf{H}(\cdot;\epsilon)$ is continuous over the set $\{Q \in \mathcal{Q} : \epsilon_{\min}(Q) < \epsilon\}$.*

Proof. See Appendix D.2.2. □

Remark 15. The mapping $\mathsf{H}(\cdot;\epsilon)$ might *not* be continuous over $\{Q \in \mathcal{Q} : \epsilon_{\min}(Q) \leq \epsilon\}$. In order to prove this claim, consider the following example.

Example 7. For each $\zeta \in [0, 1/2]$, let Q_ζ be the joint distribution of (S, X_ζ) where $S \sim \mathrm{Ber}(1/2)$ and $P_{X_\zeta|S} = \mathrm{BSC}(\zeta)$. By definition, H_c denotes the privacy-utility function when both privacy leakage and utility are measured using probability of correctly guessing. By Theorem 2 in [18], we have that, for every $\zeta \in [0, 1/2)$,

$$\epsilon_{\min}(Q_\zeta) = 1/2 \quad \text{and} \quad \mathsf{H}_c(Q_\zeta; 1/2) = 1/2. \tag{6.86}$$

Since S and $X_{1/2}$ are independent, we also have that

$$\epsilon_{\min}(Q_{1/2}) = 1/2 \quad \text{and} \quad \mathsf{H}_c(Q_{1/2}; 1/2) = 1. \tag{6.87}$$

Therefore, we have that $\lim_{\zeta \uparrow 1/2} \mathsf{H}_c(Q_\zeta; 1/2) \neq \mathsf{H}_c(Q_{1/2}; 1/2)$ although $\lim_{\zeta \uparrow 1/2} Q_\zeta = Q_{1/2}$.

Despite that $\mathsf{H}(\cdot;\epsilon)$ might be discontinuous at the boundary $\{Q \in \mathcal{Q} : \epsilon_{\min}(Q) = \epsilon\}$, as presented in Example 7, in some situations such pathological behavior is not exhibited. In such cases, the following corollary provides conditions for which the pointwise convergence established in Proposition 16 upgrades into uniform convergence. This technical result will be useful in explaining the numerical experiments in Section 6.7.

Corollary 5. *Assume that conditions (**C**.1–3) hold true for a given closed set $\mathcal{Q} \subseteq \mathcal{P}$ and a given $N \in \mathbb{N}$. In addition, assume that there exists $\epsilon_0 \in \mathbb{R}$ such that $\epsilon_{\min}(Q) = \epsilon_0$ for all $Q \in \mathcal{Q}$. If $P_n \in \mathcal{Q}$ for each $n \in \mathbb{N}$, $\lim_n P_n = P$, and $\lim_n \mathsf{H}(P_n; \epsilon_0) = \mathsf{H}(P; \epsilon_0)$, then $\mathsf{H}(P_n; \cdot)$ converges uniformly to $\mathsf{H}(P; \cdot)$ over $[\epsilon_0, \infty)$, i.e.,*

$$\lim_{n \to \infty} \sup_{\epsilon \in [\epsilon_0, \infty)} |\mathsf{H}(P_n; \epsilon) - \mathsf{H}(P, \epsilon)| = 0. \tag{6.88}$$

Proof. See Appendix D.2.3. □

The next theorem shows that if W_n^* is an optimal ϵ-private mechanism for P_n and $(P_n)_{n=1}^\infty$ converges to P, then the sequence of optimal ϵ-private mechanisms $(W_n^*)_{n=1}^\infty$ converges to the set of optimal ϵ-private mechanisms for P. For $W \in \mathcal{W}_N$ and a subset $\mathcal{W}' \subseteq \mathcal{W}_N$, the distance between W and $\mathcal{W}'$ is defined as

$$\mathsf{dist}(W, \mathcal{W}') \triangleq \inf\{\|W - W'\|_1 : W' \in \mathcal{W}'\}, \tag{6.89}$$

where $\|W - W'\|_1 = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} |W(x,y) - W'(x,y)|$.

Theorem 16. *Assume that conditions (**C**.1–3) hold true for a given closed set $\mathcal{Q} \subseteq \mathcal{P}$ and a given $N \in \mathbb{N}$. Let $\epsilon \in \mathbb{R}$ be given. If $P_n \in \mathcal{Q}$ for each $n \in \mathbb{N}$, $\lim_n P_n = P$, and $\epsilon_{\min}(P) < \epsilon$, then, for any sequence $(W_n^*)_{n=1}^{\infty} \subset \mathcal{W}_N$ such that $W_n^* \in \mathcal{W}_N^*(P_n; \epsilon)$,*

$$\lim_{n \to \infty} \mathsf{dist}(W_n^*, \mathcal{W}_N^*(P; \epsilon)) = 0. \tag{6.90}$$

Furthermore, if $P_n = \hat{P}_n$ is the empirical distribution obtained from n i.i.d. samples drawn from P, then

$$\Pr\left(\lim_{n \to \infty} \mathsf{dist}(W_n^*, \mathcal{W}_N^*(P; \epsilon)) = 0\right) = 1. \tag{6.91}$$

Proof. See Appendix D.2.4. □

Remark 16. Under the assumptions in Theorem 16, it may be possible that $\mathcal{W}_N^*(P_n; \epsilon)$ is empty for some values of n. Nonetheless, under the same assumptions, $\mathcal{W}_N^*(P_n; \epsilon)$ is non-empty for n sufficiently large. This fact can be easily derived from the continuity of the mapping $\epsilon_{\min}(\cdot)$ which is the content of Lemma 39 in Appendix D.2.4.

The following corollary follows directly from Theorem 16 and the compactness of $\mathcal{W}_N^*(P; \epsilon)$ established in Lemma 40 in Appendix D.2.4. It shows that the limit of optimal ϵ-private mechanisms is also an optimal ϵ-private mechanism.

Corollary 6. *In addition to the assumptions in Theorem 16, assume that $\lim_n W_n^* = W_0$ for some $W_0 \in \mathcal{W}_N$. Then, $W_0 \in \mathcal{W}_N^*(P; \epsilon)$.*

It is important to note that Theorem 16 does not imply that the sequence of optimal privacy mechanisms $(W_n^*)_{n=1}^{\infty}$ converges to a privacy mechanism. Indeed, it has been noticed in simulation that a small perturbation to the joint distribution might lead to a big change to the optimal privacy mechanism returned by some optimization algorithms. The following theorem, whose proof relies on standard results from *set-valued analysis* [20], shows that given a sequence $(P_n)_{n=1}^{\infty}$ with $\lim_n P_n = P$, it is possible to choose $W_n^* \in \mathcal{W}_N^*(P_n; \epsilon)$ such that $(W_n^*)_{n=1}^{\infty}$ is convergent. However, we prove this property only for a residual set which, by definition, is a countable intersection of dense open subsets of $\mathcal{Q} \subset \mathbb{R}^{|\mathcal{S}| \times |\mathcal{X}|}$.

Theorem 17. *Assume that conditions (**C**.1–3) hold true for a given closed set $\mathcal{Q} \subseteq \mathcal{P}$ and a given $N \in \mathbb{N}$. For any given $\epsilon \in \mathbb{R}$ and any $\delta > 0$, there exists a residual set $\mathcal{Q}' \subseteq \{Q \in \mathcal{Q} : \epsilon_{\min}(Q) + \delta \leq \epsilon\}$ which satisfies that:*

For any joint distribution $Q_0 \in \mathcal{Q}'$, any sequence of joint distributions $(Q_n)_{n=1}^{\infty} \subset \{Q \in \mathcal{Q} : \epsilon_{\min}(Q) + \delta \leq \epsilon\}$ with $\lim_n Q_n = Q_0$, and any optimal privacy mechanism $W_0^* \in \mathcal{W}_N^*(Q_0; \epsilon)$, there exists a sequence of privacy mechanisms $(W_n^*)_{n=1}^{\infty}$ such that $W_n^* \in \mathcal{W}_N^*(Q_n; \epsilon)$ for each $n \in \mathbb{N}$ and $\lim_n W_n^* = W_0^*$.

Proof. See Appendix D.2.5. □

Remark 17. Baire's theorem states that any residual subset of a complete metric space is dense, see, e.g., [236, Thm. 5.6]. In this sense, residual sets are considered to be *big* in topological terms. However, residual sets might be negligible in measure theoretic terms, see, e.g., [101, Exercise 5.3.31].

6.5.1 Continuity and Compactness Conditions

In this section we show that probability of correctly guessing, f-information with f locally Lipschitz, Arimoto's mutual information of order α with $\alpha \in (1, \infty]$ satisfy conditions (**C.**1–3).

Probability of Correctly Guessing

We choose $\mathcal{Q} = \mathcal{P}$ and $\mathcal{Y} = \{1, \ldots, N\}$ with $N \geq |\mathcal{X}| + 1$. Note that $\mathcal{Q} = \mathcal{P}$ corresponds to the case where no prior information about the true distribution P is available. Recall that, by definition,

$$\mathcal{L}_c(Q, W) = \sum_{y \in \mathcal{Y}} \max_{s \in \mathcal{S}} \sum_{x \in \mathcal{X}} Q(s, x) W(x, y), \tag{6.92}$$

$$\mathcal{U}_c(Q, W) = \sum_{y \in \mathcal{Y}} \max_{x \in \mathcal{X}} \sum_{s \in \mathcal{S}} Q(s, x) W(x, y). \tag{6.93}$$

Since the maximum and the sum of continuous functions are continuous functions as well, we have that, for every $Q \in \mathcal{P}$, the mappings $\mathcal{L}_c(Q, \cdot)$ and $\mathcal{U}_c(Q, \cdot)$ are continuous over $\mathcal{W}_N$. In [18], Asoodeh *et al.* showed that $\mathsf{H}_c(Q; \cdot)$, the privacy-utility function associated to $\mathcal{L}_c$ and $\mathcal{U}_c$, is continuous and piecewise linear over[3] $[\epsilon_{\min}(Q), \infty)$. Therefore condition (**C.**1) is satisfied. Furthermore, Lemma 15 shows that, for any $Q_1, Q_2 \in \mathcal{P}$ and $W \in \mathcal{W}$,

$$|\mathcal{L}_c(Q_1, W) - \mathcal{L}_c(Q_2, W)| \leq \|Q_1 - Q_2\|_1, \tag{6.94}$$

$$|\mathcal{U}_c(Q_1, W) - \mathcal{U}_c(Q_2, W)| \leq \|Q_1 - Q_2\|_1, \tag{6.95}$$

[3] Indeed, in this setting $\epsilon_{\min}(Q) = \max_s \sum_x Q(s, x)$.

which implies that condition (**C.**2) is satisfied with $C_L = 1$ and $C_U = 1$. It was established in [18] that, for every $\epsilon \geq \epsilon_{\min}(Q)$, there is always an optimal ϵ-private mechanism using at most $|\mathcal{X}| + 1$ symbols. As a consequence, condition (**C.**3) holds true as $N \geq |\mathcal{X}| + 1$ by assumption.

f-Information with f Locally Lipschitz

Assume that the function $f : [0, \infty) \to \mathbb{R}$ is locally Lipschitz and convex with $f(1) = 0$. Given $\gamma > 0$, we choose $N \geq |\mathcal{X}| + 1$ and

$$\mathcal{Q} = \left\{ Q \in \mathcal{P} : \gamma \leq \min_{s \in \mathcal{S}} \sum_{x \in \mathcal{X}} Q(s,x), \gamma \leq \min_{x \in \mathcal{X}} \sum_{s \in \mathcal{S}} Q(s,x) \right\}. \tag{6.96}$$

In this case, the prior information about the true distribution $P = P_{S,X}$ comes in the form of the assumption that the probability mass functions of S and X are bounded away from 0. Recall that, by definition,

$$\mathcal{L}_f(Q,W) = \sum_{s \in \mathcal{S}} \sum_{y \in \mathcal{Y}} P_S(s) P_Y(y) f\left(\frac{P_{S,Y}(s,y)}{P_S(s)P_Y(y)} \right), \tag{6.97}$$

$$\mathcal{U}_f(Q,W) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P_X(x) P_Y(y) f\left(\frac{P_{X,Y}(x,y)}{P_X(x)P_Y(y)} \right), \tag{6.98}$$

where $S \to X \to Y$ is such that $P_{S,X} = Q$ and $P_{Y|X} = W$. Note that if $Q \in \mathcal{Q}$, then, for all $(s,x,y) \in \mathcal{S} \times \mathcal{X} \times \mathcal{Y}$,

$$\max\left\{ \frac{P_{S,Y}(s,y)}{P_S(s)P_Y(y)}, \frac{P_{X,Y}(x,y)}{P_X(x)P_Y(y)} \right\} \leq \gamma^{-1}. \tag{6.99}$$

Upon this fact, it is straightforward to verify that the mappings $\mathcal{L}_f(Q,\cdot)$ and $\mathcal{U}_f(Q,\cdot)$ are continuous over $\mathcal{W}_N$. In [125], Hsu *et al.* established the continuity of $\mathsf{H}_f(Q;\cdot)$ over $[0,\infty)$. Hence, condition (**C.**1) is satisfied. Recall the definitions of $K_{g,u}$ and $L_{g,u}$ in (6.40) and (6.41), respectively. Lemma 16 shows that, for any $Q_1, Q_2 \in \mathcal{Q}$ and $W \in \mathcal{W}$,

$$|\mathcal{L}_f(Q_1,W) - \mathcal{L}_f(Q_2,W)| \leq (2K_{f,\gamma^{-1}} + (2\gamma^{-1}+1)L_{f,\gamma^{-1}}) \|Q_1 - Q_2\|_1, \tag{6.100}$$

$$|\mathcal{U}_f(Q_1,W) - \mathcal{U}_f(Q_2,W)| \leq (2K_{f,\gamma^{-1}} + (2\gamma^{-1}+1)L_{f,\gamma^{-1}}) \|Q_1 - Q_2\|_1, \tag{6.101}$$

which implies that condition (**C.**2) is satisfied with

$$C_L = C_U = 2K_{f,\gamma^{-1}} + (2\gamma^{-1}+1)L_{f,\gamma^{-1}}. \tag{6.102}$$

For example, if $f(x) = |x-1|$, then $C_L = C_U \leq 4\gamma^{-1} + 1$; and if $f(x) = x^2 - 1$, then $C_L = C_U \leq 8\gamma^{-2}$. As with probability of correctly guessing, there is always an optimal ϵ-private mechanism using at

most $|\mathcal{X}|+1$ symbols [125]. From this fact, condition (**C.3**) follows immediately.

Arimoto's Mutual Information

Let $\alpha \in (1,\infty]$. We choose $N \geq |\mathcal{X}|+1$ and $\mathcal{Q} = \mathcal{P}$, i.e., no prior information is assumed. As with the previous information measures, the continuity of the mappings $\mathcal{L}_\alpha^{\mathrm{A}}(Q,\cdot)$ and $\mathcal{U}_\alpha^{\mathrm{A}}(Q,\cdot)$ is evident. We denote the privacy-utility function by $\mathsf{H}_\alpha^{\mathrm{A}}$ when both privacy leakage and utility are measured using Arimoto's mutual information of order α. Note that the graph of $\mathsf{H}_\alpha^{\mathrm{A}}(Q;\cdot)$ corresponds to the upper boundary of the set

$$\mathcal{A} \triangleq \left\{ (\mathcal{L}_\alpha^{\mathrm{A}}(Q,W), \mathcal{U}_\alpha^{\mathrm{A}}(Q,W)) : W \in \mathcal{W} \right\}. \tag{6.103}$$

More specifically, $\mathsf{H}_\alpha^{\mathrm{A}}(Q;\epsilon) = \sup\{u : (p,u) \in \mathcal{A}, p \leq \epsilon\}$. Consider the transformation

$$(p,u) \mapsto \left(\Big\| \sum_{x\in\mathcal{X}} Q(\cdot,x) \Big\|_\alpha e^{\alpha p/(\alpha-1)}, \Big\| \sum_{s\in\mathcal{S}} Q(s,\cdot) \Big\|_\alpha e^{\alpha u/(\alpha-1)} \right). \tag{6.104}$$

It is straightforward to verify that this transformation is one-to-one, continuous, and monotone coordinatewise. Using this transformation, it can be shown that (the upper boundary of) $\mathcal{A}$ is homeomorphic to (the upper boundary of)

$$\mathcal{B} \triangleq \{(\phi(Q,W), \psi(Q,W)) : W \in \mathcal{W}\}, \tag{6.105}$$

where $\phi(Q,W) \triangleq \sum_{y\in\mathcal{Y}} \|\sum_{x\in\mathcal{X}} Q(\cdot,x)W(x,y)\|_\alpha$ and $\psi(Q,W) \triangleq \sum_{y\in\mathcal{Y}} \|\sum_{s\in\mathcal{S}} Q(s,\cdot)W(\cdot,y)\|_\alpha$. By the homogeneity of the α-norm, it can be verified that

$$\phi(Q,[\lambda_1 W_1,\ldots,\lambda_K W_K] = \sum_{k=1}^{K} \lambda_k \phi(Q,W_k), \tag{6.106}$$

whenever $K \in \mathbb{N}$, $W_1,\ldots,W_K \in \mathcal{W}$, and $\lambda_1,\ldots,\lambda_K \geq 0$ with $\sum_k \lambda_k = 1$. A similar equality holds for ψ. Therefore, $\mathcal{B}$ is a convex set and, as a consequence, its upper boundary is the graph of a continuous function. Since the upper boundaries of $\mathcal{A}$ and $\mathcal{B}$ are homeomorphic, we conclude that the mapping $\mathsf{H}_\alpha^{\mathrm{A}}(Q;\cdot)$ is continuous. Therefore, condition (**C.1**) is satisfied. Also, Lemma 18 implies that condition (**C.2**) is satisfied with $C_L = \dfrac{2\alpha}{\alpha-1}|\mathcal{S}|^{1-1/\alpha}$ and $C_U = \dfrac{2\alpha}{\alpha-1}|\mathcal{X}|^{1-1/\alpha}$. Observe that proving condition (**C.3**) with $N \geq |\mathcal{X}|+1$ is equivalent to show that any point in the upper boundary of $\mathcal{A}$ can be written as $(\mathcal{L}_\alpha^{\mathrm{A}}(Q,W), \mathcal{U}_\alpha^{\mathrm{A}}(Q,W))$ for some $W \in \mathcal{W}_N$. Since $\mathcal{A}$ and $\mathcal{B}$ are homeomorphic, the latter property can be established from an analogous property for $\mathcal{B}$ which in turn follows from a minor adaptation of the argument in [294].

Remark 18. Regarding condition (**C.**3), both [18] and [125] build upon the convex analysis argument employed by Witsenhausen and Wyner in [294]. More specifically, they rely on an extension of Carathéodory's theorem, the so-called Fenchel-Eggleston theorem [89], in order to find a bound for the size of the output alphabet of an optimal privacy mechanism.

6.6 Uniform Privacy Mechanisms

In this section we consider the scenario in which delivering privacy is the top priority for the privacy mechanism designer. We introduce privacy mechanisms that, with a certain probability, guarantee privacy for the true distribution despite having access only to an estimate of it. These mechanisms are constructed in the following conceptual way. Recall that large deviations results show that, with a certain probability, the true distribution is within some ℓ_1 ball of the empirical distribution. Consequently, if a mechanism guarantees privacy *uniformly* for every distribution within such an ℓ_1 ball, it necessarily guarantees privacy for the true distribution with at least the same probability. In this section, we prove that there is a well-defined notion of optimality for *uniform* privacy mechanisms and that these optimal mechanisms can be approximated by appropriately chosen (non-uniform) optimal privacy mechanisms as previously defined in (6.30).

In order to introduce uniform privacy mechanisms precisely, recall that inequality (6.27) shows that, with probability at least $1 - \exp(|\mathcal{S}| \cdot |\mathcal{X}| - nr^2/2)$, the true distribution P is within the ℓ_1 ball of radius r centered at the empirical distribution $\hat{P}_n$,

$$\mathcal{Q}_r(\hat{P}_n) \triangleq \{Q \in \mathcal{Q} : \|Q - \hat{P}_n\|_1 \leq r\}. \tag{6.107}$$

Based on this observation, we consider uniform privacy mechanisms which guarantee privacy for every joint distribution within $\mathcal{Q}_r(\hat{P})$, where $\hat{P}$ is any estimate of P. Despite the fact that $P \in \mathcal{Q}_r(\hat{P})$ with high probability, we do not know the true value of P. For this reason, we define optimal uniform privacy mechanisms as those that achieve the best worst-case utility within $\mathcal{Q}_r(\hat{P})$.

Definition 26. Let $\mathcal{Q} \subset \mathcal{P}$ and $N \in \mathbb{N}$. For $\hat{P} \in \mathcal{Q}$, $\epsilon \geq 0$, and $r \geq 0$, we define the set of uniform privacy mechanisms for $\mathcal{Q}_r(\hat{P})$ at ϵ as

$$\mathcal{D}_{\mathcal{Q},N}(\hat{P}; \epsilon, r) \triangleq \bigcap_{Q \in \mathcal{Q}_r(\hat{P})} \{W \in \mathcal{W}_N : \mathcal{L}(Q, W) \leq \epsilon\}. \tag{6.108}$$

Furthermore, we define the set of optimal uniform privacy mechanisms for $\mathcal{Q}_r(\hat{P})$ at ϵ as

$$\mathcal{W}^{\dagger}_{\mathcal{Q},N}(\hat{P};\epsilon,r) \triangleq \underset{W\in\mathcal{D}_{\mathcal{Q},N}(\hat{P};\epsilon,r)}{\arg\max} \mathcal{U}_r(\hat{P},W), \tag{6.109}$$

where $\mathcal{U}_r(\hat{P},W) \triangleq \min_{Q\in\mathcal{Q}_r(\hat{P})} \mathcal{U}(Q,W)$.

Recall that, as defined in Remark 14, $\mathcal{D}_N(Q;\epsilon) = \{W \in \mathcal{W}_N : \mathcal{L}(Q,W) \leq \epsilon\}$ is the set of all privacy mechanisms in $\mathcal{W}_N$ delivering an ϵ-privacy guarantee for Q. Thus,

$$\mathcal{D}_{\mathcal{Q},N}(\hat{P};\epsilon,r) = \bigcap_{Q\in\mathcal{Q}_r(\hat{P})} \mathcal{D}_N(Q;\epsilon) \tag{6.110}$$

is the set of all privacy mechanisms in $\mathcal{W}_N$ that deliver an ϵ-privacy guarantee uniformly for all the distributions in $\mathcal{Q}_r(\hat{P})$, i.e., all the distributions at a distance less than or equal to r from $\hat{P}$. For a given privacy mechanism W, $\mathcal{U}_r(\hat{P},W)$ quantifies the least utility $\mathcal{U}(Q,W)$ attained by W over all the distributions in $\mathcal{Q}_r(\hat{P})$. Thus, by definition, $\mathcal{W}^{\dagger}_{\mathcal{Q},N}(\hat{P};\epsilon,r)$ is the set of all uniform privacy mechanisms for $\mathcal{Q}_r(\hat{P})$ at ϵ with the best worst-case utility.

The following lemma shows that Definition 26 is well-defined under conditions (**C.**1–2). Specifically, it shows that the infimum defining $\mathcal{U}_r(\hat{P},W)$ and the supremum defining $\mathcal{W}^{\dagger}_{\mathcal{Q},N}(\hat{P};\epsilon,r)$ are attainable.

Lemma 22. *Assume that conditions (**C.**1–2) hold true for a given closed set $\mathcal{Q} \subseteq \mathcal{P}$ and a given $N \in \mathbb{N}$. Then*

(i) *the infimum* $\inf\{\mathcal{U}(Q,W) : Q \in \mathcal{Q}_r(\hat{P})\}$ *is attainable for every $W \in \mathcal{W}_N$;*

(ii) *the supremum* $\sup\{\mathcal{U}_r(\hat{P},W) : W \in \mathcal{D}_{\mathcal{Q},N}(\hat{P};\epsilon,r)\}$ *is attainable whenever $\mathcal{D}_{\mathcal{Q},N}(\hat{P};\epsilon,r)$ is not empty.*

Proof. See Appendix D.3.1. □

Observe that (ii) shows that optimal uniform privacy mechanisms do exists as long as $\mathcal{D}_{\mathcal{Q},N}(\hat{P};\epsilon,r)$ is not empty. Nonetheless, their construction might be challenging as the construction of optimal privacy mechanisms in the non-uniform sense, as defined in (6.30), is already non-trivial in most cases. The following theorem shows that optimal privacy mechanism can be modified to deliver a uniform privacy guarantee without incurring in a big cost in terms of utility. Throughout this section, we consistently use W^* to denote optimal privacy mechanisms, i.e., elements in $\mathcal{W}^*_N$ as defined in (6.84), and $W^{\dagger}$ to denote their uniform counterparts as defined in (6.109).

Theorem 18. *Assume that conditions (**C**.1–3) hold true for a given closed set $\mathcal{Q} \subseteq \mathcal{P}$ and a given $N \in \mathbb{N}$. If $P \in \mathcal{Q}_r(\hat{P})$ and $\epsilon - C_L r \geq \epsilon_{\min}(\hat{P})$, then*

$$\emptyset \neq \mathcal{D}_N(\hat{P}; \epsilon - C_L r) \subseteq \mathcal{D}_{\mathcal{Q},N}(\hat{P}; \epsilon, r). \tag{6.111}$$

Furthermore, for every $W^* \in \mathcal{W}_N^*(\hat{P}; \epsilon - C_L r)$ and every $W^\dagger \in \mathcal{W}_{\mathcal{Q},N}^\dagger(\hat{P}; \epsilon, r)$,

$$\mathcal{U}(P, W^*) \geq \mathcal{U}(P, W^\dagger) - \left(\mathsf{H}(\hat{P}; \epsilon + C_L r) - \mathsf{H}(\hat{P}; \epsilon - C_L r) + 2C_U r\right). \tag{6.112}$$

Proof. See Appendix D.3.2. □

Observe that (6.111) establishes that privacy mechanisms for $\hat{P}$ at $\epsilon - C_L r$ are, in fact, uniform privacy mechanisms at ϵ. In other words, (6.111) shows that by imposing a slightly stronger privacy requirement for $\hat{P}$, we obtain *uniform* privacy mechanisms for $\mathcal{Q}_r(\hat{P})$. Furthermore, under condition (**C**.1),

$$\lim_{r \downarrow 0} \left(\mathsf{H}(\hat{P}; \epsilon + C_L r) - \mathsf{H}(\hat{P}; \epsilon - C_L r) + 2C_U r\right) = 0. \tag{6.113}$$

Therefore, (6.112) shows that, when r is small, any optimal privacy mechanism for $\hat{P}$ at $\epsilon - C_L r$ performs almost as well as any optimal uniform privacy mechanism for $\mathcal{Q}_r(\hat{P})$ at ϵ. Of course, all these conclusions hold as long as $P \in \mathcal{Q}_r(\hat{P})$ but, as pointed out before, this is the case (with high probability) in the large sample size regime.

Remark 19. Note that if $\mathsf{H}(\hat{P}; \cdot)$ is Lipschitz continuous with Lipschitz constant L, then (6.112) becomes

$$\mathcal{U}(P, W^*) \geq \mathcal{U}(P, W^\dagger) - 2(C_U + LC_L)r. \tag{6.114}$$

Observe that if $\mathsf{H}(\hat{P}; \cdot)$ is a convex function differentiable at $\epsilon_0 \triangleq \epsilon_{\min}(\hat{P})$, then $L = \mathsf{H}'(\hat{P}; \epsilon_0)$. The value of $\mathsf{H}'(\hat{P}; \epsilon_0)$ is closely related with the notion of *reverse* strong data processing inequality [52], see also [17].

Now we consider a specific example to illustrate the above results.

Example 8. Recall that H_c is the privacy-utility function associated to $\mathcal{L}_c$ and $\mathcal{U}_c$ as given in (6.35). For ease of notation, we define

$$p \# q \triangleq \begin{pmatrix} (1-p)(1-q) & (1-p)q \\ pq & p(1-q) \end{pmatrix}. \tag{6.115}$$

Let $\mathcal{Q} = \{p\#q : p \in [1/2, 1], q \in [0, 1-p]\}$ and $N \geq 2$. This selection of $\mathcal{Q}$ captures the case where $S \sim \text{Ber}(p)$ with $p \in [1/2, 1]$ and $P_{X|S} = \text{BSC}(q)$ with $q \in [0, 1-p]$. By Theorem 2 in [18], for all $Q = p\#q \in \mathcal{Q}$,

$$\mathsf{H}_c(Q;\epsilon) = 1 - \frac{1-q}{p-q}(p+q-2pq) + \epsilon\frac{p+q-2pq}{p-q}, \tag{6.116}$$

whenever $\epsilon \in [p, 1-q]$. In particular, $\mathsf{H}_c(Q;\cdot)$ is Lipschitz continuous with Lipschitz constant $\frac{p+q-2pq}{p-q}$. Recall that, for probability of correctly guessing, $C_L = C_U = 1$ as established in Section 6.5.1. Hence, under the assumptions of Theorem 18, (6.114) becomes

$$\mathcal{U}(P, W^*) \geq \mathcal{U}(P, W^\dagger) - \frac{2\hat{p}(1-\hat{q})}{\hat{p}-\hat{q}}r, \tag{6.117}$$

where $W^* \in \mathcal{W}_N^*(\hat{P}; \epsilon - r)$ and $W^\dagger \in \mathcal{W}_{\mathcal{Q},N}^\dagger(\hat{P}; \epsilon, r)$ with $\hat{P} \triangleq \hat{p}\#\hat{q} \in \mathcal{Q}$. Furthermore, by taking $\hat{P} = \hat{P}_n$ and $r = (2(4 - \log\beta)/n)^{1/2}$, inequality (6.27) implies that, with probability at least $1-\beta$,

$$\mathcal{U}(P, W^*) \geq \mathcal{U}(P, W^\dagger) - \frac{2\hat{p}(1-\hat{q})}{\hat{p}-\hat{q}}\sqrt{\frac{2}{n}(4 - \log\beta)}. \tag{6.118}$$

We finish this section by studying some convergence properties of uniform privacy mechanism, similar to those studied in Theorem 16. More specifically, the next theorem shows that although a sequence $(W_n^\dagger)_{n=1}^\infty$ with $W_n^\dagger \in \mathcal{W}_{\mathcal{Q},N}^\dagger(P_n; \epsilon, r_n)$ may not be convergent, the distance between each $W_n^\dagger$ and $\mathcal{W}_N^*(P;\epsilon)$ converges to zero as long as $\lim_n P_n = P$ and $\lim_n r_n = 0$. Furthermore, it also shows that, under certain conditions, this convergence can be guaranteed almost surely for the empirical distribution estimator.

Theorem 19. *Assume that conditions (**C**.1–3) hold true for a given closed set $\mathcal{Q} \subseteq \mathcal{P}$ and a given $N \in \mathbb{N}$. Let $\epsilon \in \mathbb{R}$ and $P \in \mathcal{Q}$ be given. If $P_n \in \mathcal{Q}$ with $\|P_n - P\|_1 \leq r_n$ for each $n \in \mathbb{N}$, $\lim_n r_n = 0$, and $\epsilon > \epsilon_{\min}(P)$, then, for any sequence $(W_n^\dagger)_{n=1}^\infty \subset \mathcal{W}_N$ such that $W_n^\dagger \in \mathcal{W}_{\mathcal{Q},N}^\dagger(P_n; \epsilon, r_n)$ for all $n \geq 1$,*

$$\lim_{n\to\infty} \mathsf{dist}(W_n^\dagger, \mathcal{W}_N^*(P;\epsilon)) = 0. \tag{6.119}$$

Furthermore, if $P_n = \hat{P}_n$ is the empirical estimator obtained from n i.i.d. samples drawn from P and, for some $p > 1$, $r_n \geq \sqrt{\frac{2p\log(n)}{n}}$ for all $n \geq 1$, then,

$$\Pr\left(\lim_{n\to\infty} \mathsf{dist}(W_n^\dagger, \mathcal{W}_N^*(P;\epsilon)) = 0\right) = 1. \tag{6.120}$$

Proof. See Appendix D.3.3. □

The following corollary follows immediately from Thereom 19 by taking $P = P_n = \hat{P}$.

Corollary 7. *Assume that conditions (C.1–3) hold true for a given closed set $\mathcal{Q} \subseteq \mathcal{P}$ and a given $N \in \mathbb{N}$. If $\hat{P} \in \mathcal{Q}$ and $\epsilon > \epsilon_{\min}(\hat{P})$, then*

$$\limsup_{r \downarrow 0} \mathcal{W}^{\dagger}_{\mathcal{Q},N}(\hat{P}; \epsilon, r) \subseteq \mathcal{W}^{*}_{N}(\hat{P}; \epsilon), \tag{6.121}$$

where

$$\limsup_{r \downarrow 0} \mathcal{W}^{\dagger}_{\mathcal{Q},N}(\hat{P}; \epsilon, r) \triangleq \left\{ W \in \mathcal{W}_N : \liminf_{r \downarrow 0} \operatorname{dist}\left(W, \mathcal{W}^{\dagger}_{\mathcal{Q},N}(\hat{P}; \epsilon, r)\right) = 0 \right\}. \tag{6.122}$$

Theorem 18 shows that in terms of performance, optimal privacy mechanisms approximate optimal uniform privacy mechanisms in the small r regime. The previous corollary shows that, in the same regime, optimal uniform privacy mechanisms are *geometrically close* to the set of optimal privacy mechanisms. These two results evidence an intrinsic relation between optimal privacy mechanisms and their uniform counterparts.

6.7 Numerical Experiments

We illustrate some of the results derived in this chapter through two numerical experiments. The first experiment, conducted on a synthetic dataset, illustrates the convergence of the empirical privacy-utility function and optimal privacy mechanisms to their corresponding limits as the sample size increases. The second experiment, performed on a real-world dataset, displays the discrepancy between the privacy-utility guarantees during the design and testing of a privacy mechanism.

6.7.1 Synthetic Datasets

Here we illustrate the convergence of the empirical privacy-utility function and optimal privacy mechanisms as the sample size increases. The joint distribution matrix $\mathbf{P}$, determined by $\mathbf{P}_{s,x} \triangleq P(s,x)$, is chosen to be

$$\mathbf{P} = \begin{pmatrix} 0.42 & 0.18 \\ 0.16 & 0.24 \end{pmatrix}, \tag{6.123}$$

and both privacy leakage and utility are measured using χ^2-information, i.e., the f-information associated to the function $f(t) = (t-1)^2$. For any given value of n, we randomly select n i.i.d. samples $\{(s_i, x_i)\}_{i=1}^{n}$ drawn from the joint distribution P and compute the empirical distribution $\hat{P}_n$.

In Figure 6.2 (top left), we depict the privacy-utility function $\mathsf{H}_{\chi^2}(P; \cdot)$ and the empirical privacy-

utility function $\mathsf{H}_{\chi^2}(\hat{P}_n;\cdot)$ for three different values of n. Note that the privacy-utility function of the empirical distribution converges (pointwise) to the corresponding function for the true distribution, i.e., for any given $\epsilon \geq 0$ we have that $\mathsf{H}_{\chi^2}(\hat{P}_n;\epsilon)$ approaches $\mathsf{H}_{\chi^2}(P;\epsilon)$ as n grows. This corroborates the conclusion of Proposition 16 which establishes that, for any given $\epsilon \in \mathbb{R}$, $\mathsf{H}(\cdot;\epsilon)$ is continuous over $\{Q \in \mathcal{Q} : \epsilon_{\min}(Q) < \epsilon\}$. In Figure 6.2 (top right), we depict the corresponding optimal privacy mechanisms. In order to visualize a privacy mechanism $P_{Y|X}$ in the xy-plane, we let the x-axis and y-axis to be the values of $P_{Y|X}(0|0)$ and $P_{Y|X}(1|1)$, respectively. Observe that optimal privacy mechanisms are not unique. Hence, depending on the algorithm used to obtain such mechanisms, a sequence of empirical optimal privacy mechanisms may not converge to a fixed mechanism. However, as observed from Figure 6.2 (top right), the distance between any such sequence and the set of optimal privacy mechanisms for the true distribution will necessarily converge to zero which echoes Theorem 16.

Finally we scatter plot the largest (signed) gap, denoted by Δ_n, between $\mathsf{H}_{\chi^2}(P_n;\cdot)$ and $\mathsf{H}_{\chi^2}(P;\cdot)$. In particular, the absolute value of Δ_n is equal to the uniform norm of the function $\mathsf{H}_{\chi^2}(P_n;\cdot) - \mathsf{H}_{\chi^2}(P;\cdot)$, i.e.,

$$|\Delta_n| = \sup_{\epsilon \in [0,\infty)} |\mathsf{H}_{\chi^2}(P_n;\epsilon) - \mathsf{H}_{\chi^2}(P;\epsilon)|. \tag{6.124}$$

We depict the value of Δ_n in Figure 6.2 (bottom). As shown, when the number of samples increases, $|\Delta_n|$ tends to decrease which illustrates the uniform convergence established in Corollary 5. Due to the mismatch between the empirical and true distributions, privacy and utility might be over or under-estimated, leading to the variable sign of Δ_n.

6.7.2 ProPublica's COMPAS Recidivism Dataset

We now illustrate Theorems 12 and 14 in Section 6.4 through ProPublica's COMPAS dataset [10], which contains the criminal history, jail and prison time, demographics and COMPAS risk scores for defendants in Broward County from 2013 and 2014. We process the original dataset by dropping records with missing information and quantizing some of the interest variables. Our final dataset contains 5278 records. We choose the private variable (S): *Race* $\in \{$*Caucasian*, *African-American*$\}$; the useful variable (X): *PriorCounts* $\in \{0, 1-3, >3\}$ and *AgeCategory* $\in \{<25, 25-45, >45\}$. Hence, the size of support sets are $|\mathcal{S}| = 2$ and $|\mathcal{X}| = 9$.

Given $n \in \mathbb{N}$, we choose n records from the dataset as our training set and another n different

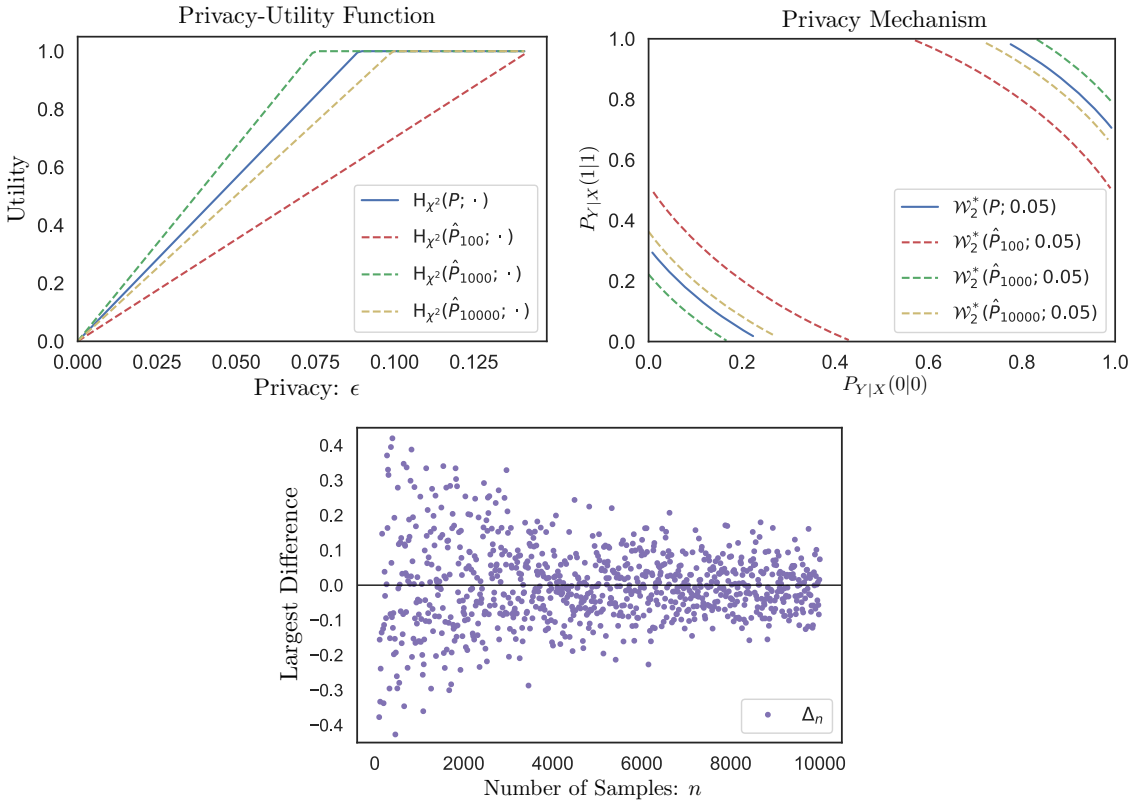

Figure 6.2: *Both privacy leakage and utility are measured using the f-information with $f(t) = (t-1)^2$. Top left: privacy-utility function $\mathsf{H}_{\chi^2}(P;\cdot)$ and empirical privacy-utility function $\mathsf{H}_{\chi^2}(\hat{P}_n;\cdot)$. Top right: corresponding optimal privacy mechanisms for $\epsilon = 0.05$. Bottom: largest (signed) difference between $\mathsf{H}_{\chi^2}(P_n;\cdot)$ and $\mathsf{H}_{\chi^2}(P;\cdot)$.*

records as our testing set. We use probability of correctly guessing to measure both privacy and utility. We compute the empirical distribution $\hat{P}_{n,train}$ based on the training set and let W_n be the privacy mechanism that solves the following optimization problem:

$$\max_{W \in \mathcal{W}_R} \mathcal{U}_c(\hat{P}_{n,train}, W) \tag{6.125}$$

$$\text{s.t. } \mathcal{L}_c(\hat{P}_{n,train}, W) \leq 0.65, \tag{6.126}$$

where $\mathcal{W}_R$ denotes the set of *randomized response mechanisms* [147, 291], i.e., the set of all privacy mechanisms W with output alphabet $\mathcal{X}$ such that, for some $\rho \geq 0$,

$$W(i,j) = \begin{cases} \dfrac{e^\rho}{e^\rho + |\mathcal{X}| - 1} & \text{if } i = j, \\ \dfrac{1}{e^\rho + |\mathcal{X}| - 1} & \text{if } i \neq j. \end{cases} \tag{6.127}$$

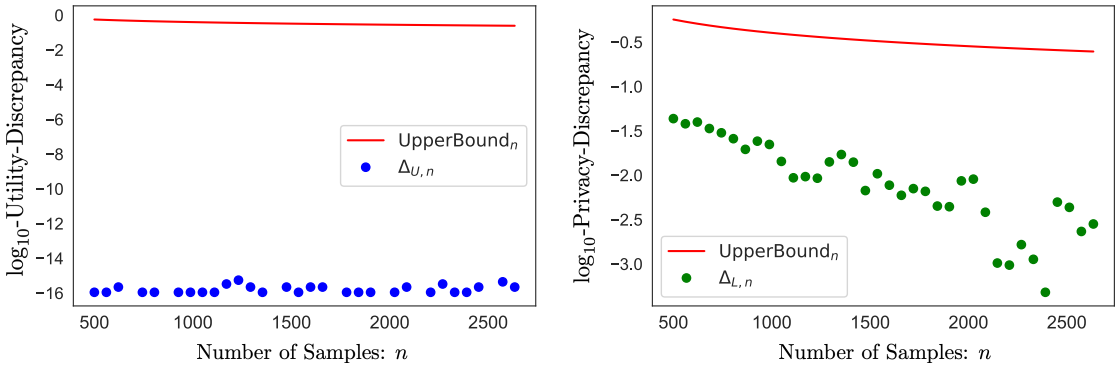

Figure 6.3: *Both privacy leakage and utility are measured using probability of correctly guessing. Left: discrepancy between utility guarantees for the training and testing sets. Right: discrepancy between privacy guarantees for the training and testing sets. In both pictures the theoretical upper bound in (6.130) is shown in red.*

Then, we let $\hat{P}_{n,test}$ be the empirical distribution of the testing set and compute the privacy-utility guarantees attained by W_n for this distribution. Finally, we evaluate the discrepancy between the privacy-utility guarantees provided for $\hat{P}_{n,train}$ and $\hat{P}_{n,test}$:

$$\Delta_{L,n} \triangleq |\mathcal{L}_c(\hat{P}_{n,test}, W_n) - \mathcal{L}_c(\hat{P}_{n,train}, W_n)|, \tag{6.128}$$

$$\Delta_{U,n} \triangleq |\mathcal{U}_c(\hat{P}_{n,test}, W_n) - \mathcal{U}_c(\hat{P}_{n,train}, W_n)|. \tag{6.129}$$

By the triangle inequality and Theorem 12, with probability at least $1-\beta$, these two discrepancies are upper bounded by

$$\mathsf{UpperBound}_n \triangleq 2\sqrt{\frac{2}{n}\left(|\mathcal{S}| \cdot |\mathcal{X}| - \log\beta\right)}. \tag{6.130}$$

Figure 6.3 depicts the discrepancies $\Delta_{L,n}$ and $\Delta_{U,n}$, as functions of n, along with the upper bound $\mathsf{UpperBound}_n$ for $\beta = 10^{-1}$. Observe that when number of samples increases, the discrepancy between the privacy-utility guarantees tends to decrease.

To further illustrate our results, we perform a similar experiment using Arimoto's mutual information of order 2. Specifically, we compute the privacy mechanism W_n by solving the following optimization problem:

$$\max_{W \in \mathcal{W}_Z} \mathcal{U}_2^A(\hat{P}_{n,train}, W) \tag{6.131}$$

$$\text{s.t. } \mathcal{L}_2^A(\hat{P}_{n,train}, W) \leq 0.05, \tag{6.132}$$

where $\mathcal{W}_Z$ denotes the set of privacy mechanisms W with output alphabet $\mathcal{X}$ such that, for some

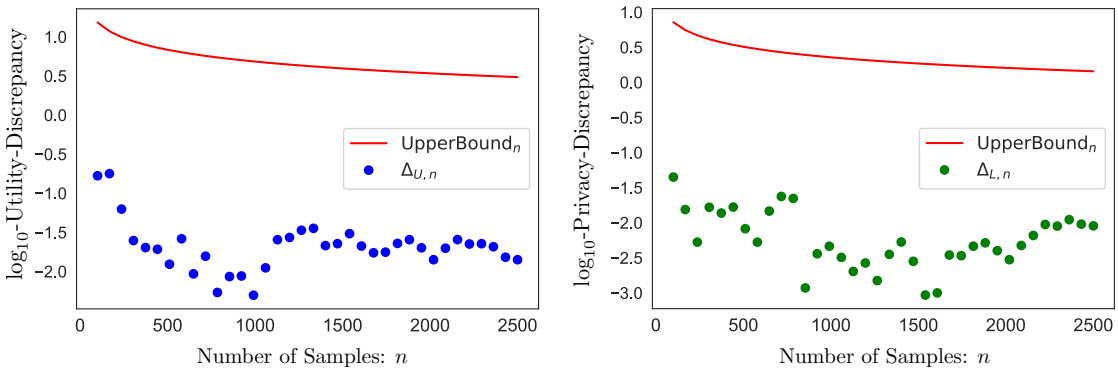

Figure 6.4: *Both privacy leakage and utility are measured using Arimoto's mutual information of order* 2*. Left: discrepancy between utility guarantees for the training and testing sets. Right: discrepancy between privacy guarantees for the training and testing sets. The theoretical upper bounds, in* (6.134) *and* (6.135) *respectively, are shown in red.*

$\bar{x} \in \mathcal{X}$ and $\zeta \geq 0$,

$$W(i,j) = \begin{cases} 1 & \text{if } i = j = \bar{x}, \\ 1-\zeta & \text{if } i = j \neq \bar{x}, \\ \zeta & \text{if } i \neq \bar{x}, j = \bar{x}. \end{cases} \tag{6.133}$$

In the binary case, $\mathcal{W}_Z$ is nothing but the collection of Z-channels. It has been proved [18] that these channels are capable to achieve optimal privacy-utility trade-offs in some cases. Let $\Delta_{L,n}$ (resp. $\Delta_{U,n}$) be the discrepancy between the privacy (resp. utility) guarantees for $\hat{P}_{n,train}$ and $\hat{P}_{n,test}$. By Theorem 14, with probability at least $1-\beta$, $\Delta_{L,n}$ and $\Delta_{U,n}$ are upper bounded by

$$\text{Privacy:UpperBound}_n \triangleq 8\sqrt{\frac{2}{n}\left(|\mathcal{S}| \cdot |\mathcal{X}| - \log\beta\right) \cdot |\mathcal{S}|}, \tag{6.134}$$

$$\text{Utility:UpperBound}_n \triangleq 8\sqrt{\frac{2}{n}\left(|\mathcal{S}| \cdot |\mathcal{X}| - \log\beta\right) \cdot |\mathcal{X}|}, \tag{6.135}$$

respectively. Figure 6.4 depicts the discrepancies $\Delta_{L,n}$ and $\Delta_{U,n}$, as functions of n, along with their corresponding upper bounds for $\beta = 10^{-1}$.

6.8 Conclusion

In this chapter, we analyzed the effect of a limited sample size on the disclosure of data under privacy constraints. We considered a setting where data is released upon observing non-private features correlated with a set of private features. We evaluated privacy and utility using one out of

five of the most commonly used information leakage and utility measures in information-theoretic privacy: probability of correctly guessing, f-information, Arimoto's mutual information, Sibson's mutual information, and maximal α-leakage.

The effect of a limited sample size was assessed via probabilistic upper bounds for the difference between the privacy-utility guarantees for the empirical and true distributions. An important feature of these bounds is that they are completely independent of the privacy mechanism at hand. On a technical level, the proofs of these bounds depend on large deviations results already available in the literature and continuity properties of information leakage measures established in this work. Furthermore, we have established new continuity properties of privacy-utility functions. Using these properties, we have shown that the limit of a convergent sequence of optimal privacy mechanisms is an optimal privacy mechanism itself.

In order to mitigate the effect of a limited sample size on the privacy guarantees delivered to the true distribution, we introduced the notion of *uniform* privacy mechanisms. By definition, these mechanisms provide a specific privacy guarantee for *every* distribution in a given subset of the probability simplex. In particular, when this subset is a neighborhood of the empirical distribution, large deviations results imply that privacy is guaranteed for the true distribution with high probability. While the construction of optimal uniform privacy mechanisms might be challenging, we proved that these mechanisms can be approximated in a natural way by optimal privacy mechanisms in the non-uniform sense. More specifically, we have proved that an optimal privacy mechanism for the empirical distribution delivers a slightly weaker privacy guarantee for a whole neighborhood of the empirical distribution and performs almost as well as any optimal uniform privacy mechanism. By establishing convergence results regarding optimal uniform privacy mechanisms, we have further exhibited the intrinsic relation between optimal privacy mechanisms and their uniform counterparts.

While this work predominantly focused on large deviations bounds, the continuity properties derived in this paper can be applied together with contemporary results to derive similar upper bounds in other estimation frameworks, e.g., ℓ_1 min-max estimation.

Chapter 7

Conclusion and Future Work

In this thesis, we established an information-theoretic foundation of trustworthy ML. We proved rigorous performance guarantees for ML models and developed protocols for the responsible use of data and ML algorithms. We derived generalization bounds to understand why complex ML models often generalize well in practice. We developed theory that delineates a fundamental limit of algorithmic fairness and privacy. The theory also provided design guidelines for practitioners who deploy ML technology in applications of individual-level consequences.

Understanding the Generalization of Complex ML Models

In Chapter 3, we investigated how the data distribution and optimization method influence the generalization of ML models. Specifically, we derived distribution-dependent generalization bounds for noisy iterative algorithms. The key step in our proof was to build a connection between noisy iterative algorithms and additive noise channels found in information theory. This connection enabled us to leverage the properties of additive noise channels for studying the generalization of noisy iterative algorithms.

There are several open questions that deserve further investigation. For example, we proved that our generalization bounds could be tightened if the output of the algorithm is the last iterate. Our analysis was inspired by a line of works on privacy amplification by iteration [19, 23, 99]. On the other hand, there are other ways to amplify privacy, such as subsampling [59] and shuffling [90]. It would be interesting to understand if the these methods can improve the algorithmic generalization.

More broadly, it would be interesting to investigate the behavior of neural networks in the *overparameterized regime*. The uniform convergence results from statistical learning theory suggest that models that perfectly fit the training data will exhibit a poor generalization performance. Although these results are useful for understanding classical ML, they fail to explain the behavior of deep learning methods in the overparameterized regime where models can interpolate all training data and achieve perfect training accuracy. In this regime, prediction models often exhibit "benign" overfitting [28]—they generalize well to unseen data while overfitting the training data.

In the overparameterized regime, there are many solutions to the empirical risk minimization, each having distinct generalization properties. In this case, gradient-based optimization algorithms introduce a bias in selecting a minimizer with certain properties. For example, gradient descent for training a linear model with squared error loss converges to a solution with minimal L_2 norm. It would be interesting to see if the information-theoretic framework developed in this thesis sheds light on understanding implicit regularization. On the other hand, to explain how the optimization methods influence the generalization of neural networks, it would be helpful to develop a unified framework that yields generalization bounds for all kinds of optimization methods.

The classical learning theory suggests that the test error exhibits a U-shaped curve with respect to model complexity due to the bias-variance trade-off. However, this phenomenon only occurs in the underparameterized regime. When it comes to the overparameterized regime, the test error decreases again with the model complexity and highly overparameterized models often achieve a better test accuracy than the best underparameterized model. It would be interesting to investigate this research direction and propose new model complexity measures for explaining the double descent phenomenon.

Ensuring Algorithmic Fairness and Privacy

We investigated algorithmic fairness and privacy in Chapter 4–6. We provided conditions to ensure the fair use of group attributes and characterized the fundamental limit of privacy-utility trade-offs. We studied the robustness of information leakage measures and established the statistical consistency of "optimal" privacy mechanisms. Our theory was not only technically relevant, but it also inspired new algorithms for correcting biased models and for designing privacy mechanisms.

There are several open questions that deserve further exploration. First, we proved that the differ-

ence in underlying data distributions between groups, the number of samples, and the hypothesis class could all influence the effect of splitting classifiers. Nonetheless, we believe that there are more factors that play an important role in determining this effect. For example, a group-blind classifier may perform worse on minority groups due to unbalanced samples in the training process and using split classifiers could potentially reconcile this issue. In a similar vein, the lack of sample diversity (i.e., training datasets do not contain enough samples from minority groups) could affect the performance and generalization of ML models for minority groups. Hence, it is crucial to characterize the impact of sample size and diversity on detecting and reducing discrimination.

Ensuring fairness requires being able to detect discrimination in the first place. However, a major challenge for testing group fairness is that the available data for conducting the test are limited but the number of groups can grow exponentially with the number of group attributes. If we conduct an independent hypothesis test for each group, then the probability that at least one null hypothesis is wrongly rejected can increase rapidly. It would be interesting to design statistical tests to verify discrimination that account for the multiplicity issue. These tests will be extremely useful to audit discrimination in ML models and are currently missing in the literature.

References

[1] Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. (2016). Deep learning with differential privacy. In *ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318.

[2] Adler, P., Falk, C., Friedler, S. A., Nix, T., Rybeck, G., Scheidegger, C., Smith, B., and Venkatasubramanian, S. (2018). Auditing black-box models for indirect influence. *Knowledge and Information Systems*, 54(1):95–122.

[3] Agarwal, A., Beygelzimer, A., Dudik, M., Langford, J., and Wallach, H. (2018). A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69. PMLR.

[4] Alabdulmohsin, I. M. (2015). Algorithmic stability and uniform generalization. In *Advances in Neural Information Processing Systems*, volume 28, pages 19–27.

[5] Alghamdi, W., Asoodeh, S., Wang, H., Calmon, F. P., Wei, D., and Ramamurthy, K. N. (2020). Model projection: Theory and applications to fair machine learning. In *IEEE International Symposium on Information Theory*, pages 2711–2716.

[6] Aliprantis, C. D. and Border, K. C. (2006). *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer.

[7] Aminian, G., Toni, L., and Rodrigues, M. R. (2021). Jensen-Shannon information based characterization of the generalization error of learning algorithms. In *IEEE Information Theory Workshop*, pages 1–5.

[8] Anantharam, V., Gohari, A., Kamath, S., and Nair, C. (2013). On maximal correlation, hypercontractivity, and the data processing inequality studied by Erkip and Cover. *arXiv preprint arXiv:1304.6133*.

[9] Angenent, S., Haker, S., and Tannenbaum, A. (2003). Minimizing flows for the Monge–Kantorovich problem. *SIAM Journal on Mathematical Analysis*, 35(1):61–97.

[10] Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias. *ProPublica, May*, page 2016.

[11] Anthony, M. and Bartlett, P. L. (2009). *Neural network learning: Theoretical foundations*. Cambridge University Press.

[12] Arimoto, S. (1977). Information measures and capacity of order α for discrete memoryless channels. *Topics in information theory*.

[13] Arora, S., Cohen, N., Hu, W., and Luo, Y. (2019). Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems*, volume 32.

[14] Asadi, A., Abbe, E., and Verdu, S. (2018). Chaining mutual information and tightening generalization bounds. In *Advances in Neural Information Processing Systems*, volume 31.

[15] Asoodeh, S., Alajaji, F., and Linder, T. (2014). Notes on information-theoretic privacy. In *Annual Allerton Conference on Communication, Control, and Computing*, pages 1272–1278.

[16] Asoodeh, S., Alajaji, F., and Linder, T. (2016a). Privacy-aware mmse estimation. In *IEEE International Symposium on Information Theory*, pages 1989–1993.

[17] Asoodeh, S., Diaz, M., Alajaji, F., and Linder, T. (2016b). Information extraction under privacy constraints. *Information*, 7(1):15.

[18] Asoodeh, S., Diaz, M., Alajaji, F., and Linder, T. (2018). Estimation efficiency under privacy constraints. *IEEE Transactions on Information Theory*, 65(3):1512–1534.

[19] Asoodeh, S., Diaz, M., and Calmon, F. P. (2020). Privacy amplification of iterative algorithms via contraction coefficients. In *IEEE International Symposium on Information Theory*, pages 896–901.

[20] Aubin, J.-P. and Frankowska, H. (2009). *Set-valued Analysis*. Springer Science & Business Media.

[21] Bache, K. and Lichman, M. (2013). UCI Machine Learning Repository.

[22] Balke, A. and Pearl, J. (1994). Counterfactual probabilities: Computational methods, bounds and applications. In *Conference on Uncertainty in Artificial Intelligence*, pages 46–54. Morgan Kaufmann Publishers Inc.

[23] Balle, B., Barthe, G., Gaboardi, M., and Geumlek, J. (2019). Privacy amplification by mixing and diffusion mechanisms. In *Advances in Neural Information Processing Systems*, pages 13298–13308.

[24] Barocas, S. and Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, 104:671.

[25] Barron, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945.

[26] Barthe, G. and Kopf, B. (2011). Information-theoretic bounds for differentially private mechanisms. In *IEEE Computer Security Foundations Symposium*, pages 191–204.

[27] Bartlett, P. L., Foster, D. J., and Telgarsky, M. J. (2017). Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, pages 6240–6249.

[28] Bartlett, P. L., Montanari, A., and Rakhlin, A. (2021). Deep learning: a statistical viewpoint. *Acta Numerica*, 30:87–201.

[29] Basciftci, Y. O., Wang, Y., and Ishwar, P. (2016). On privacy-utility tradeoffs for constrained data release mechanisms. In *Information Theory and Applications Workshop*, pages 1–6. IEEE.

[30] Bassily, R., Nissim, K., Smith, A., Steinke, T., Stemmer, U., and Ullman, J. (2021). Algorithmic stability for adaptive data analysis. *SIAM Journal on Computing*, pages 377–405.

[31] Beauchamp, T. L. and Childress, J. F. (2001). *Principles of biomedical ethics*. Oxford University Press, USA.

[32] Beck, A. and Teboulle, M. (2003). Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175.

[33] Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. (2010). A theory of learning from different domains. *Machine learning*, 79(1-2):151–175.

[34] Ben-Tal, A., El Ghaoui, L., and Nemirovski, A. (2009). *Robust Optimization*. Princeton University Press.

[35] Benamou, J.-D. and Brenier, Y. (2000). A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393.

[36] Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. (2021). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44.

[37] Bien, J. and Tibshirani, R. (2011). Prototype selection for interpretable classification. *The Annals of Applied Statistics*, pages 2403–2424.

[38] Blum, A. and Stangl, K. (2020). Recovering from biased data: Can fairness constraints improve accuracy? In *Symposium on Foundations of Responsible Computing*, volume 156, pages 3:1–3:20.

[39] Bogen, M. and Rieke, A. (2018). Help wanted: An examination of hiring algorithms, equity, and bias.

[40] Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*.

[41] Borade, S. and Zheng, L. (2008). Euclidean information theory. In *IEEE International Zurich Seminar on Communications*, pages 14–17.

[42] Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press.

[43] Braun, C., Chatzikokolakis, K., and Palamidessi, C. (2009). Quantitative notions of leakage for one-try attacks. *Electronic Notes in Theoretical Computer Science*, 249:75–91.

[44] Bretagnolle, J. and Huber, C. (1979). Estimation des densités: risque minimax. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 47(2):119–137.

[45] Brown, L. D. and Low, M. G. (1996). A constrained risk inequality with applications to nonparametric functional estimation. *The Annals of Statistics*, 24(6):2524–2535.

[46] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

[47] Bu, Y., Zou, S., and Veeravalli, V. V. (2020). Tightening mutual information-based bounds on generalization error. *IEEE Journal on Selected Areas in Information Theory*, 1(1):121–130.

[48] Buja, A. (1990). Remarks on functional canonical variates, alternating least squares methods and ace. *The Annals of Statistics*, 18(3):1032–1069.

[49] Buolamwini, J. and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91.

[50] Burgess, M. (2020). The lessons we all must learn from the A-levels algorithm debacle. *Wired UK*.

[51] Calmon, F. P. and Fawaz, N. (2012). Privacy against statistical inference. In *Annual Allerton Conference on Communication, Control, and Computing*, pages 1401–1408.

[52] Calmon, F. P., Makhdoumi, A., and Médard, M. (2015). Fundamental limits of perfect privacy. In *IEEE International Symposium on Information Theory*, pages 1796–1800.

[53] Calmon, F. P., Makhdoumi, A., Médard, M., Varia, M., Christiansen, M., and Duffy, K. R. (2017a). Principal inertia components and applications. *IEEE Transactions on Information Theory*, 63(8):5011–5038.

[54] Calmon, F. P., Polyanskiy, Y., and Wu, Y. (2018). Strong data processing inequalities for input constrained additive noise channels. *IEEE Transactions on Information Theory*, 64(3):1879–1892.

[55] Calmon, F. P., Varia, M., Médard, M., Christiansen, M. M., Duffy, K. R., and Tessaro, S. (2013). Bounds on inference. In *Annual Allerton Conference on Communication, Control, and Computing*, pages 567–574.

[56] Calmon, F. P., Wei, D., Vinzamuri, B., Ramamurthy, K. N., and Varshney, K. R. (2017b). Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, pages 3992–4001.

[57] Celis, L. E., Huang, L., Keswani, V., and Vishnoi, N. K. (2019). Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Conference on Fairness, Accountability, and Transparency*, pages 319–328. ACM.

[58] Charikar, M., Steinhardt, J., and Valiant, G. (2017). Learning from untrusted data. In *Annual ACM SIGACT Symposium on Theory of Computing*, pages 47–60.

[59] Chaudhuri, K. and Mishra, N. (2006). When random sampling preserves privacy. In *Annual International Cryptology Conference*, pages 198–213. Springer.

[60] Chen, I., Johansson, F. D., and Sontag, D. (2018). Why is my classifier discriminatory? In *Advances in Neural Information Processing Systems*, pages 3539–3550.

[61] Chen, J., Kallus, N., Mao, X., Svacha, G., and Udell, M. (2019). Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Conference on Fairness, Accountability, and Transparency*, page 339–348.

[62] Chernoff, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, pages 493–507.

[63] Chernozhukov, V., Fernández-Val, I., and Melly, B. (2013). Inference on counterfactual distributions. *Econometrica*, 81(6):2205–2268.

[64] Chiappa, S. (2019). Path-specific counterfactual fairness. In *AAAI Conference on Artificial Intelligence*, volume 33, pages 7801–7808.

[65] Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163.

[66] Cohen, J., Kempermann, J. H., and Zbaganu, G. (1998). *Comparisons of stochastic matrices with applications in information theory, statistics, economics and population*. Springer Science & Business Media.

[67] Corbett-Davies, S. and Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*.

[68] Cotter, A., Gupta, M., and Narasimhan, H. (2019). On making stochastic classifiers deterministic. In *Advances in Neural Information Processing Systems*, pages 10910–10920.

[69] Cover, T. M. and Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

[70] Csiszár, I. (1967). Information-type measures of difference of probability distributions and indirect observation. *Studia Scientiarum Mathematicarum Hungarica*, 2:229–318.

[71] Datta, A., Sen, S., and Zick, Y. (2016). Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *IEEE Symposium on Security and Privacy*, pages 598–617.

[72] David, S. B., Lu, T., Luu, T., and Pal, D. (2010). Impossibility theorems for domain adaptation. In *International Conference on Artificial Intelligence and Statistics*, pages 129–136.

[73] DeGroot, M. H. (1962). Uncertainty, information, and sequential experiments. *The Annals of Mathematical Statistics*, 33(2):404–419.

[74] Del Barrio, E., Gamboa, F., Gordaliza, P., and Loubes, J.-M. (2019). Obtaining fairness using optimal transport theory. In *International Conference on Machine Learning*. PMLR.

[75] Diamond, S. and Boyd, S. (2016). CVXPY: A Python-embedded modeling language for convex optimization. *The Journal of Machine Learning Research*, 17(1):2909–2913.

[76] Diaz, M., Wang, H., Calmon, F. P., and Sankar, L. (2019). On the robustness of information-theoretic privacy measures and mechanisms. *IEEE Transactions on Information Theory*, 66(4):1949–1978.

[77] DiNardo, J., Fortin, N. M., and Lemieux, T. (1995). Labor market institutions and the distribution of wages, 1973-1992: A semiparametric approach. Technical report, National bureau of economic research.

[78] Dobrushin, R. L. (1956). Central limit theorem for nonstationary Markov chains. i. *Theory of Probability & Its Applications*, 1(1):65–80.

[79] Dodis, Y. and Smith, A. (2005). Entropic security and the encryption of high entropy messages. In *Theory of Cryptography Conference*, pages 556–577. Springer.

[80] Duchi, J., Wainwright, M. J., and Jordan, M. I. (2013). Local privacy and minimax bounds: Sharp rates for probability estimation. In *Advances in Neural Information Processing Systems*, pages 1529–1537.

[81] Duchi, J. C., Jordan, M. I., and Wainwright, M. J. (2014). Privacy aware learning. *Journal of the ACM*, 61(6):1–57.

[82] Durrett, R. (2010). *Probability: theory and examples*. Cambridge University Press.

[83] Dutta, S., Venkatesh, P., Mardziel, P., Datta, A., and Grover, P. (2020). An information-theoretic quantification of discrimination with exempt features. In *AAAI Conference on Artificial Intelligence*.

[84] Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., and Roth, A. (2015). Generalization in adaptive data analysis and holdout reuse. In *Advances in Neural Information Processing Systems*, pages 2350–2358.

[85] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Innovations in Theoretical Computer Science Conference*, pages 214–226.

[86] Dwork, C., Immorlica, N., Kalai, A. T., and Leiserson, M. (2018). Decoupled classifiers for group-fair and efficient machine learning. In *Conference on Fairness, Accountability and Transparency*, pages 119–133.

[87] Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, pages 265–284. Springer.

[88] EEOC (1979). Uniform guidelines on employee selection procedures.

[89] Eggleston, H. G. (1958). *Convexity*. Cambridge University Press.

[90] Erlingsson, Ú., Feldman, V., Mironov, I., Raghunathan, A., Talwar, K., and Thakurta, A. (2019). Amplification by shuffling: From local to central differential privacy via anonymity. In *Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2468–2479. SIAM.

[91] Esposito, A. R., Gastpar, M., and Issa, I. (2021). Generalization error bounds via Rényi-, f-divergences and maximal leakage. *IEEE Transactions on Information Theory*.

[92] Evfimievski, A., Gehrke, J., and Srikant, R. (2003). Limiting privacy breaches in privacy preserving data mining. In *ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pages 211–222.

[93] Facebook AI (2020). Introducing Opacus: A high-speed library for training pytorch models with differential privacy.

[94] Fan, K. (1953). Minimax theorems. *National Academy of Sciences of the United States of America*, 39(1):42.

[95] Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9(Aug):1871–1874.

[96] Federal Trade Commission (FTC) (2020). Equal Credit Opportunity Act (ECOA).

[97] Fehr, S. and Berens, S. (2014). On the conditional Rényi entropy. *IEEE Transactions on Information Theory*, 60(11):6801–6810.

[98] Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268.

[99] Feldman, V., Mironov, I., Talwar, K., and Thakurta, A. (2018). Privacy amplification by iteration. In *IEEE Annual Symposium on Foundations of Computer Science*, pages 521–532.

[100] Fisher, A. and Kennedy, E. H. (2021). Visually communicating and teaching intuition for influence functions. *The American Statistician*, 75(2):162–172.

[101] Folland, G. B. (1984). *Real Analysis: Modern Techniques and Their Applications*. Wiley-Interscience, New York.

[102] Fortin, N., Lemieux, T., and Firpo, S. (2011). Decomposition methods in economics. In *Handbook of Labor Economics*, volume 4, pages 1–102. Elsevier.

[103] Friedler, S. A., Scheidegger, C., and Venkatasubramanian, S. (2016). On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236*.

[104] Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*. Springer series in statistics New York.

[105] Gálvez, B. R., Bassi, G., Thobaben, R., and Skoglund, M. (2021a). On random subset generalization error bounds and the stochastic gradient Langevin dynamics algorithm. In *IEEE Information Theory Workshop*, pages 1–5.

[106] Gálvez, B. R., Bassi, G., Thobaben, R., and Skoglund, M. (2021b). Tighter expected generalization error bounds via Wasserstein distance. In *Advances in Neural Information Processing Systems*.

[107] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. (2016). Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030.

[108] Ge, R., Huang, F., Jin, C., and Yuan, Y. (2015). Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pages 797–842.

[109] Gebelein, H. (1941). Das statistische problem der korrelation als variations-und eigenwertproblem und sein zusammenhang mit der ausgleichsrechnung. *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik*, 21(6):364–379.

[110] Gelfand, S. B. and Mitter, S. K. (1991). Recursive stochastic algorithms for global optimization in R^d. *SIAM Journal on Control and Optimization*, 29(5):999–1018.

[111] Ghassami, A., Khodadadian, S., and Kiyavash, N. (2018). Fairness in supervised learning: An information theoretic approach. In *IEEE International Symposium on Information Theory*, pages 176–180.

[112] Greenacre, M. and Hastie, T. (1987). The geometric interpretation of correspondence analysis. *Journal of the American statistical association*, 82(398):437–447.

[113] Greenacre, M. J. (1984). *Theory and Applications of Correspondence Analysis*. Academic, San Diego, CA, USA.

[114] Gunasekar, S., Woodworth, B. E., Bhojanapalli, S., Neyshabur, B., and Srebro, N. (2017). Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, volume 30.

[115] Guo, D., Shamai, S., and Verdú, S. (2005). Mutual information and minimum mean-square error in Gaussian channels. *IEEE Transactions on Information Theory*, 51(4):1261–1282.

[116] Gut, A. (2013). *Probability: A Graduate Course*, volume 75. Springer Science & Business Media.

[117] Hafez-Kolahi, H., Golgooni, Z., Kasaei, S., and Soleymani, M. (2020). Conditioning and processing: Techniques to improve information-theoretic generalization bounds. *Advances in Neural Information Processing Systems*, 33.

[118] Haghifam, M., Negrea, J., Khisti, A., Roy, D. M., and Dziugaite, G. K. (2020). Sharpened generalization bounds based on conditional mutual information and an application to noisy, iterative algorithms. *Advances in Neural Information Processing Systems*, 33:9925–9935.

[119] Hardt, M., Price, E., and Srebro, N. (2016a). Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323.

[120] Hardt, M., Recht, B., and Singer, Y. (2016b). Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 1225–1234. PMLR.

[121] Hashimoto, T., Srivastava, M., Namkoong, H., and Liang, P. (2018). Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pages 1929–1938. PMLR.

[122] He, F., Wang, B., and Tao, D. (2021). Tighter generalization bounds for iterative differentially private learning algorithms. In *Conference on Uncertainty in Artificial Intelligence*.

[123] Hellström, F. and Durisi, G. (2020). Generalization bounds via information density and conditional information density. *IEEE Journal on Selected Areas in Information Theory*, 1(3):824–839.

[124] Hirschfeld, H. O. (1935). A connection between correlation and contingency. In *Mathematical Proceedings of the Cambridge Philosophical Society*, pages 520–524.

[125] Hsu, H., Asoodeh, S., Salamatian, S., and Calmon, F. P. (2018). Generalizing bottleneck problems. In *IEEE International Symposium on Information Theory*, pages 531–535.

[126] Huang, C., Kairouz, P., Chen, X., Sankar, L., and Rajagopal, R. (2017a). Context-aware generative adversarial privacy. *Entropy*, 19(12):656.

[127] Huang, S.-L., Makur, A., Kozynski, F., and Zheng, L. (2014). Efficient statistics: Extracting information from iid observations. In *Annual Allerton Conference on Communication, Control, and Computing*, pages 699–706.

[128] Huang, S.-L., Zhang, L., and Zheng, L. (2017b). An information-theoretic approach to unsupervised feature selection for high-dimensional data. In *IEEE Information Theory Workshop*, pages 434–438.

[129] Huber, P. J. (2011). Robust statistics. In *International Encyclopedia of Statistical Science*, pages 1248–1251. Springer.

[130] Hunt, D. B. (2005). Redlining. *Encyclopedia of Chicago*.

[131] Issa, I., Kamath, S., and Wagner, A. B. (2016). Maximal leakage minimization for the Shannon cipher system. In *IEEE International Symposium on Information Theory*, pages 520–524.

[132] Issa, I. and Wagner, A. B. (2017). Operational definitions for some common information leakage metrics. In *IEEE International Symposium on Information Theory*, pages 769–773.

[133] Issa, I., Wagner, A. B., and Kamath, S. (2019). An operational approach to information leakage. *IEEE Transactions on Information Theory*, 66(3):1625–1657.

[134] Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical review*, 106(4):620.

[135] Jeong, H., Wang, H., and Calmon, F. P. (2022). Fairness without imputation: A decision tree approach for fair prediction with missing values. In *AAAI Conference on Artificial Intelligence*.

[136] Jiang, H., Kim, B., Guan, M., and Gupta, M. (2018). To trust or not to trust a classifier. In *Advances in Neural Information Processing Systems*, pages 5541–5552.

[137] Jiang, Y., Neyshabur, B., Mobahi, H., Krishnan, D., and Bengio, S. (2020). Fantastic generalization measures and where to find them. In *International Conference on Learning Representations*.

[138] Jiao, J., Han, Y., and Weissman, T. (2017). Dependence measures bounding the exploration bias for general measurements. In *IEEE International Symposium on Information Theory*, pages 1475–1479.

[139] Jiao, J., Han, Y., and Weissman, T. (2018). Minimax estimation of the l_1 distance. *IEEE Transactions on Information Theory*, 64(10):6672–6706.

[140] Jiao, J., Venkat, K., Han, Y., and Weissman, T. (2015). Minimax estimation of functionals of discrete distributions. *IEEE Transactions on Information Theory*, 61(5):2835–2885.

[141] Johansson, F., Shalit, U., and Sontag, D. (2016). Learning representations for counterfactual inference. In *International Conference on Machine Learning*, pages 3020–3029. PMLR.

[142] Johndrow, J. E. and Lum, K. (2019). An algorithm for removing sensitive information: application to race-independent recidivism prediction. *The Annals of Applied Statistics*, 13(1):189–220.

[143] Jose, S. T. and Simeone, O. (2021a). Information-theoretic bounds on transfer generalization gap based on Jensen-Shannon divergence. In *European Signal Processing Conference*, pages 1461–1465. IEEE.

[144] Jose, S. T. and Simeone, O. (2021b). Information-theoretic generalization bounds for meta-learning and applications. *Entropy*, 23(1):126.

[145] Jung, C., Ligett, K., Neel, S., Roth, A., Sharifi-Malvajerdi, S., and Shenfeld, M. (2021). A new analysis of differential privacy's generalization guarantees. In *Annual ACM SIGACT Symposium on Theory of Computing*.

[146] Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. (2019). Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*.

[147] Kairouz, P., Oh, S., and Viswanath, P. (2014). Extremal mechanisms for local differential privacy. In *Advances in Neural Information Processing Systems*, pages 2879–2887.

[148] Kairouz, P., Oh, S., and Viswanath, P. (2017). The composition theorem for differential privacy. *IEEE Transactions on Information Theory*, 63(6):4037–4049.

[149] Kallus, N. and Zhou, A. (2018). Residual unfairness in fair machine learning from prejudiced data. In *International Conference on Machine Learning*, pages 2439–2448. PMLR.

[150] Kamath, S., Orlitsky, A., Pichapati, D., and Suresh, A. T. (2015). On learning distributions from their samples. In *Conference on Learning Theory*, pages 1066–1100. PMLR.

[151] Kamiran, F. and Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33.

[152] Kanamori, T., Suzuki, T., and Sugiyama, M. (2011). f-divergence estimation and two-sample homogeneity test under semiparametric density-ratio models. *IEEE Transactions on Information Theory*, 58(2):708–720.

[153] Kearns, M., Neel, S., Roth, A., and Wu, Z. S. (2018). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pages 2564–2572. PMLR.

[154] Kearns, M. J., Vazirani, U. V., and Vazirani, U. (1994). *An Introduction to Computational Learning Theory*. MIT press.

[155] Kifer, D. and Machanavajjhala, A. (2014). Pufferfish: A framework for mathematical privacy definitions. *ACM Transactions on Database Systems*, 39(1):1–36.

[156] Kilbertus, N., Carulla, M. R., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, pages 656–666.

[157] Kim, B., Khanna, R., and Koyejo, O. O. (2016). Examples are not enough, learn to criticize! criticism for interpretability. In *Advances in Neural Information Processing Systems*, pages 2280–2288.

[158] Kim, M. P., Ghorbani, A., and Zou, J. (2019). Multiaccuracy: Black-box post-processing for fairness in classification. In *AAAI/ACM Conference on AI, Ethics, and Society*, pages 247–254.

[159] Kleinberg, B., Li, Y., and Yuan, Y. (2018a). An alternative view: When does SGD escape local minima? In *International Conference on Machine Learning*, pages 2698–2707. PMLR.

[160] Kleinberg, J., Ludwig, J., Mullainathan, S., and Rambachan, A. (2018b). Algorithmic fairness. In *Aea papers and proceedings*, volume 108, pages 22–27.

[161] Kleinberg, J., Mullainathan, S., and Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In *Innovations in Theoretical Computer Science Conference*, volume 67, pages 43:1–43:23.

[162] Koh, P. W. and Liang, P. (2017). Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, pages 1885–1894. PMLR.

[163] Kpotufe, S. and Martinet, G. (2018). Marginal singularity, and the benefits of labels in covariate-shift. In *Conference on Learning Theory*, pages 1882–1886.

[164] Krizhevsky, A. (2009). Learning multiple layers of features from tiny images. *technical report*.

[165] Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

[166] Kurdila, A. J. and Zabarankin, M. (2006). *Convex functional analysis*. Springer Science & Business Media.

[167] Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4066–4076.

[168] Kuzborskij, I. and Orabona, F. (2013). Stability and hypothesis transfer learning. In *International Conference on Machine Learning*, pages 942–950. PMLR.

[169] Lal, A., Pinevich, Y., Gajic, O., Herasevich, V., and Pickering, B. (2020). Artificial intelligence and computer simulation models in critical illness. *World Journal of Critical Care Medicine*, 9(2):13.

[170] Lalkhen, A. G. and McCluskey, A. (2008). Clinical tests: sensitivity and specificity. *Continuing Education in Anaesthesia Critical Care & Pain*, 8(6):221–223.

[171] LeCam, L. (1973). Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, pages 38–53.

[172] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

[173] Li, C., Chen, C., Carlson, D., and Carin, L. (2016). Preconditioned stochastic gradient Langevin dynamics for deep neural networks. In *AAAI Conference on Artificial Intelligence*, pages 1788–1794.

[174] Li, C. T. and El Gamal, A. (2018). Maximal correlation secrecy. *IEEE Transactions on Information Theory*, 64(5):3916–3926.

[175] Li, J., Luo, X., and Qiao, M. (2020). On generalization error bounds of noisy gradient methods for non-convex learning. In *International Conference on Learning Representations*.

[176] Li, Z. and Oechtering, T. J. (2015). Privacy-aware distributed Bayesian detection. *IEEE Journal of Selected Topics in Signal Processing*, 9(7):1345–1357.

[177] Liao, J., Kosut, O., Sankar, L., and Calmon, F. P. (2018a). Privacy under hard distortion constraints. In *IEEE Information Theory Workshop*, pages 1–5.

[178] Liao, J., Kosut, O., Sankar, L., and Calmon, F. P. (2018b). A tunable measure for information leakage. In *IEEE International Symposium on Information Theory*, pages 701–705.

[179] Liao, J., Sankar, L., Calmon, F. P., and Tan, V. Y. (2017). Hypothesis testing under maximal leakage privacy constraints. In *IEEE International Symposium on Information Theory*, pages 779–783.

[180] Liao, J., Sankar, L., Tan, V. Y., and Calmon, F. P. (2016). Hypothesis testing in the high privacy limit. In *Annual Allerton Conference on Communication, Control, and Computing*, pages 649–656.

[181] Lichman, M. (2013). UCI machine learning repository.

[182] Lin, J. (1991). Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151.

[183] Lipton, Z., McAuley, J., and Chouldechova, A. (2018). Does mitigating ML's impact disparity require treatment disparity? In *Advances in Neural Information Processing Systems*, pages 8125–8135.

[184] Liu, J., Cuff, P., and Verdú, S. (2016). E_γ-resolvability. *IEEE Transactions on Information Theory*, 63(5):2629–2658.

[185] Lopez, A. T. and Jog, V. (2018). Generalization error bounds using Wasserstein distances. In *IEEE Information Theory Workshop*, pages 1–5.

[186] Makhdoumi, A. and Fawaz, N. (2013). Privacy-utility tradeoff under statistical uncertainty. In *Annual Allerton Conference on Communication, Control, and Computing*, pages 1627–1634.

[187] Makhdoumi, A., Salamatian, S., Fawaz, N., and Médard, M. (2014). From the information bottleneck to the privacy funnel. In *IEEE Information Theory Workshop*, pages 501–505.

[188] Makur, A. and Zheng, L. (2017). Polynomial singular value decompositions of a family of source-channel models. *IEEE Transactions on Information Theory*, 63(12):7716–7728.

[189] Makur, A. and Zheng, L. (2020). Comparison of contraction coefficients for f-divergences. *Problems of Information Transmission*, 56(2):103–156.

[190] Mansour, Y., Mohri, M., and Rostamizadeh, A. (2009). Domain adaptation: Learning bounds and algorithms. In *Conference on Learning Theory*.

[191] Martinez, N., Bertran, M., and Sapiro, G. (2020). Minimax Pareto fairness: A multi objective perspective. In *International Conference on Machine Learning*, pages 6755–6764. PMLR.

[192] Marton, K. (1996). A measure concentration inequality for contracting markov chains. *Geometric & Functional Analysis GAFA*, 6(3):556–571.

[193] McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282. PMLR.

[194] Menon, A. K. and Williamson, R. C. (2018). The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency*, pages 107–118.

[195] Mironov, I. (2017). Rényi differential privacy. In *IEEE Computer Security Foundations Symposium*, pages 263–275.

[196] Mou, W., Wang, L., Zhai, X., and Zheng, K. (2018). Generalization bounds of SGLD for non-convex learning: Two theoretical viewpoints. In *Conference on Learning Theory*, pages 605–638.

[197] Nabi, R. and Shpitser, I. (2018). Fair inference on outcomes. In *AAAI Conference on Artificial Intelligence*, pages 1931–1940.

[198] Nageswaran, A. and Narayan, P. (2019). Data privacy for a ρ-recoverable function. *IEEE Transactions on Information Theory*, 65(6):3470–3488.

[199] Narayanan, A. and Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. In *IEEE Symposium on Security and Privacy*, pages 111–125.

[200] Neelakantan, A., Vilnis, L., Le, Q. V., Sutskever, I., Kaiser, L., Kurach, K., and Martens, J. (2016). Adding gradient noise improves learning for very deep networks. *International Conference on Learning Representations Workshop*.

[201] Negrea, J., Haghifam, M., Dziugaite, G. K., Khisti, A., and Roy, D. M. (2019). Information-theoretic generalization bounds for SGLD via data-dependent estimates. In *Advances in Neural Information Processing Systems*, pages 11015–11025.

[202] Nemirovsky, A. and Yudin, D. (1983). *Problem complexity and method efficiency in optimization.* Wiley.

[203] Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. (2011). Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*.

[204] Neu, G., Dziugaite, G. K., Haghifam, M., and Roy, D. M. (2021). Information-theoretic generalization bounds for stochastic gradient descent. In *Conference on Learning Theory*.

[205] Neyshabur, B., Bhojanapalli, S., McAllester, D., and Srebro, N. (2017). Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, pages 5947–5956.

[206] Neyshabur, B., Tomioka, R., and Srebro, N. (2015). In search of the real inductive bias: On the role of implicit regularization in deep learning. *International Conference on Learning Representations Workshop*.

[207] Nguyen, X., Wainwright, M. J., and Jordan, M. I. (2010). Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861.

[208] O'Donnell, R. (2008). Some topics in analysis of Boolean functions. In *ACM Symposium on Theory of Computing*, pages 569–578.

[209] Osia, S. A., Rassouli, B., Haddadi, H., Rabiee, H. R., and Gündüz, D. (2019). Privacy against brute-force inference attacks. In *IEEE International Symposium on Information Theory*, pages 637–641.

[210] Otjacques, B., Hitzelberger, P., and Feltz, F. (2007). Interoperability of e-government information systems: Issues of identification and data sharing. *Journal of Management Information Systems*, 23(4):29–51.

[211] Otto, F. and Villani, C. (2000). Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *Journal of Functional Analysis*, 173(2):361–400.

[212] Oxtoby, J. C. (2013). *Measure and Category: A Survey of the Analogies between Topological and Measure Spaces*, volume 2. Springer Science & Business Media.

[213] Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175.

[214] Pensia, A., Jog, V., and Loh, P.-L. (2018). Generalization error bounds for noisy, iterative algorithms. In *IEEE International Symposium on Information Theory*, pages 546–550.

[215] Peters, J., Janzing, D., and Schölkopf, B. (2017). *Elements of causal inference: foundations and learning algorithms*. MIT press.

[216] Peyré, G. and Cuturi, M. (2017). Computational optimal transport. Technical report, Center for Research in Economics and Statistics.

[217] Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. (2017). On fairness and calibration. In *Advances in Neural Information Processing Systems*, pages 5680–5689.

[218] Polyanskiy, Y., Poor, H. V., and Verdú, S. (2010). Channel coding rate in the finite blocklength regime. *IEEE Transactions on Information Theory*, 56(5):2307–2359.

[219] Polyanskiy, Y. and Wu, Y. (2016). Dissipation of information in channels with input constraints. *IEEE Transactions on Information Theory*, 62(1):35–55.

[220] Polyanskiy, Y. and Wu, Y. (2019a). Dualizing Le Cam's method, with applications to estimating the unseens. *arXiv preprint arXiv:1902.05616*.

[221] Polyanskiy, Y. and Wu, Y. (2019b). Lecture notes on information theory. *Lecture Notes for 6.441 (MIT), ECE 563 (UIUC), STAT 364 (Yale)*.

[222] Radebaugh, C. and Erlingsson, U. (2019). Introducing TensorFlow privacy: Learning with differential privacy for training data.

[223] Raginsky, M. (2016). Strong data processing inequalities and Φ-Sobolev inequalities for discrete channels. *IEEE Transactions on Information Theory*, 62(6):3355–3389.

[224] Raginsky, M., Rakhlin, A., and Telgarsky, M. (2017). Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. In *Conference on Learning Theory*, pages 1674–1703.

[225] Raginsky, M., Rakhlin, A., Tsao, M., Wu, Y., and Xu, A. (2016). Information-theoretic analysis of stability and bias of learning algorithms. In *IEEE Information Theory Workshop*, pages 26–30.

[226] Raginsky, M. and Sason, I. (2013). Concentration of measure inequalities in information theory, communications, and coding. *Foundations and Trends in Communications and Information Theory*, 10(1-2):1–250.

[227] Rassouli, B. and Gündüz, D. (2019). Optimal utility-privacy trade-off with total variation distance as a privacy measure. *IEEE Transactions on Information Forensics and Security*, 15:594–603.

[228] Rassouli, B., Rosas, F. E., and Gündüz, D. (2019). Data disclosure under perfect sample privacy. *IEEE Transactions on Information Forensics and Security*, 15:2012–2025.

[229] Rebollo-Monedero, D., Forne, J., and Domingo-Ferrer, J. (2010). From t-closeness-like privacy to postrandomization via information theory. *IEEE Transactions on Knowledge and Data Engineering*, 22(11):1623–1636.

[230] Rényi, A. (1959). On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungarica*, 10(3-4):441–451.

[231] Resnick, S. I. (2013). *Extreme Values, Regular Variation and Point Processes*. Springer.

[232] Rockafellar, R. T. and Wets, R. J. (1982). On the interchange of subdifferentiation and conditional expectation for convex functionals. *Stochastics: An International Journal of Probability and Stochastic Processes*, 7(3):173–182.

[233] Royden, H. L. (1968). *Real Analysis*. Macmillan, New York, 2 edition.

[234] Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331.

[235] Rudin, W. (1976). *Principles of Mathematical Analysis*. McGraw-hill, New York, 3 edition.

[236] Rudin, W. (1987). *Real and Complex Analysis*. McGraw-Hill, 2 edition.

[237] Russell, A. and Wang, H. (2002). How to fool an unbounded adversary with a short key. In *International Conference on the Theory and Applications of Cryptographic Techniques*, pages 133–148. Springer.

[238] Russo, D. and Zou, J. (2019). How much does your data exploration overfit? controlling bias via information usage. *IEEE Transactions on Information Theory*, 66(1):302–323.

[239] Safran, C., Bloomrosen, M., Hammond, W. E., Labkoff, S., Markel-Fox, S., Tang, P. C., and Detmer, D. E. (2007). Toward a national framework for the secondary use of health data: an American medical informatics association white paper. *Journal of the American Medical Informatics Association*, 14(1):1–9.

[240] Sankar, L., Rajagopalan, S. R., and Poor, H. V. (2013). Utility-privacy tradeoffs in databases: An information-theoretic approach. *IEEE Transactions on Information Forensics and Security*, 8(6):838–852.

[241] Sarmanov, O. (1962). Maximum correlation coefficient (nonsymmetric case). *Selected Translations in Mathematical Statistics and Probability*, 2:207–210.

[242] Sason, I. and Verdú, S. (2016). f-divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006.

[243] Schumacher, B. (1995). Quantum coding. *Physical Review A*, 51(4):2738.

[244] Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge University Press.

[245] Shamir, O., Sabato, S., and Tishby, N. (2010). Learning and generalization with the information bottleneck. *Theoretical Computer Science*, 411(29-30):2696–2711.

[246] Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

[247] Shannon, C. E. (1949). Communication theory of secrecy systems. *The Bell System Technical Journal*, 28(4):656–715.

[248] Sharma, N. and Warsi, N. A. (2013). Fundamental bound on the reliability of quantum information transmission. *Physical review letters*, 110(8):080501.

[249] Sibson, R. (1969). Information radius. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 14(2):149–160.

[250] Smith, G. (2009). On the foundations of quantitative information flow. In *International Conference on Foundations of Software Science and Computational Structures*, pages 288–302. Springer.

[251] Sofinskyi, V. (2021). Credit scoring using machine learning.

[252] Song, S., Chaudhuri, K., and Sarwate, A. D. (2013). Stochastic gradient descent with differentially private updates. In *IEEE Global Conference on Signal and Information Processing*, pages 245–248.

[253] Stefansson, G. (2002). Business-to-business data sharing: A source for integration of supply chains. *International Journal of Production Economics*, 75(1):135–146.

[254] Steinke, T. and Zakynthinou, L. (2020). Reasoning about generalization via conditional mutual information. In *Conference on Learning Theory*, pages 3437–3452. PMLR.

[255] Suresh, H. and Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In *ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–9.

[256] Sweeney, L. (1997). Guaranteeing anonymity when sharing medical data, the datafly system. In *AMIA Annual Fall Symposium*, page 51.

[257] Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570.

[258] Sweeney, L. (2015). Only you, your doctor, and many others may know. *Technology Science*, 2015092903(9):29.

[259] Takbiri, N., Houmansadr, A., Goeckel, D. L., and Pishro-Nik, H. (2019). Matching anonymized and obfuscated time series to users' profiles. *IEEE Transactions on Information Theory*, 65(2):724–741.

[260] Tan, O., Gunduz, D., and Poor, H. V. (2013). Increasing smart meter privacy through energy harvesting and storage devices. *IEEE Journal on Selected Areas in Communications*, 31(7):1331–1341.

[261] Tan, V. Y., Anandkumar, A., Tong, L., and Willsky, A. S. (2011). A large-deviation analysis of the maximum-likelihood learning of Markov tree structures. *IEEE Transactions on Information Theory*, 57(3):1714–1735.

[262] Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., Manoff, M., and Frame, M. (2011). Data sharing by scientists: practices and perceptions. *PLoS ONE*, 6(6):e21101.

[263] Tishby, N., Pereira, F. C., and Bialek, W. (1999). The information bottleneck method. In *Annual Allerton Conference on Communication, Control, and Computing*, pages 368–377.

[264] Tripathy, A., Wang, Y., and Ishwar, P. (2019). Privacy-preserving adversarial networks. In *Annual Allerton Conference on Communication, Control, and Computing*, pages 495–505.

[265] Tsybakov, A. B. (2008). *Introduction to nonparametric estimation*. Springer Science & Business Media.

[266] Tsybakov, A. B. (2009). *Introduction to nonparametric estimation*. Springer-Verlag New York.

[267] Ustun, B., Liu, Y., and Parkes, D. (2019a). Fairness without harm: Decoupled classifiers with preference guarantees. In *International Conference on Machine Learning*, pages 6373–6382. PMLR.

[268] Ustun, B., Spangher, A., and Liu, Y. (2019b). Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 10–19. ACM.

[269] Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142.

[270] Valiant, L. G. (2000). Robust logics. *Artificial Intelligence*, 117(2):231–253.

[271] Van Loan, C. F. (1996). Matrix computations (Johns Hopkins studies in mathematical sciences).

[272] Vanschoren, J., van Rijn, J. N., Bischl, B., and Torgo, L. (2013). Openml: Networked science in machine learning. *SIGKDD Explorations*, 15(2):49–60.

[273] Vapnik, V. and Chervonenkis, A. Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264.

[274] Varodayan, D. and Khisti, A. (2011). Smart meter privacy using a rechargeable battery: Minimizing the rate of information leakage. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1932–1935.

[275] Varshney, K. R. (2021). *Trustworthy Machine Learning*. Chappaqua, NY, USA. http://trustworthymachinelearning.com.

[276] Verdú, S. (2015). α-mutual information. In *IEEE Information Theory and Applications Workshop*, pages 1–6.

[277] Villani, C. (2009). *Optimal transport: old and new*, volume 338. Springer.

[278] Vopson, M. M. (2021). The world's data explained: how much we're producing and where it's all stored.

[279] Wagner, I. and Eckhoff, D. (2018). Technical privacy metrics: a systematic survey. *ACM Computing Surveys*, 51(3):1–38.

[280] Wang, B. and Zhang, F. (1992). Some inequalities for the eigenvalues of the product of positive semidefinite hermitian matrices. *Linear Algebra and its Applications*, 160:113–118.

[281] Wang, H. and Calmon, F. P. (2017). An estimation-theoretic view of privacy. In *Annual Allerton Conference on Communication, Control, and Computing*, pages 886–893.

[282] Wang, H., Diaz, M., Calmon, F. P., and Sankar, L. (2018a). The utility cost of robust privacy guarantees. In *IEEE International Symposium on Information Theory*, pages 706–710.

[283] Wang, H., Diaz, M., Santos Filho, J. C. S., and Calmon, F. P. (2019a). An information-theoretic view of generalization via Wasserstein distance. In *IEEE International Symposium on Information Theory*, pages 577–581.

[284] Wang, H., Gao, R., and Calmon, F. P. (2021a). Generalization bounds for noisy iterative algorithms using properties of additive noise channels. *arXiv preprint arXiv:2102.02976*.

[285] Wang, H., Hsu, H., Diaz, M., and Calmon, F. P. (2021b). The impact of split classifiers on group fairness. In *IEEE International Symposium on Information Theory*, pages 3179–3184.

[286] Wang, H., Hsu, H., Diaz, M., and Calmon, F. P. (2021c). To split or not to split: The impact of disparate treatment in classification. *IEEE Transactions on Information Theory*.

[287] Wang, H., Huang, Y., Gao, R., and Calmon, F. P. (2021d). Analyzing the generalization capability of SGLD using properties of Gaussian channels. In *Advances in Neural Information Processing Systems*.

[288] Wang, H., Ustun, B., and Calmon, F. (2019b). Repairing without retraining: Avoiding disparate impact with counterfactual distributions. In *International Conference on Machine Learning*, pages 6618–6627. PMLR.

[289] Wang, H., Ustun, B., and Calmon, F. P. (2018b). On the direction of discrimination: An information-theoretic analysis of disparate impact in machine learning. In *IEEE International Symposium on Information Theory*, pages 126–130.

[290] Wang, H., Vo, L., Calmon, F. P., Médard, M., Duffy, K. R., and Varia, M. (2019c). Privacy with estimation guarantees. *IEEE Transactions on Information Theory*, 65(12):8025–8042.

[291] Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69.

[292] Weissman, T., Ordentlich, E., Seroussi, G., Verdu, S., and Weinberger, M. J. (2003). Inequalities for the L_1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*.

[293] Welling, M. and Teh, Y. W. (2011). Bayesian learning via stochastic gradient Langevin dynamics. In *International Conference on Machine Learning*, pages 681–688.

[294] Witsenhausen, H. and Wyner, A. (1975). A conditional entropy bound for a pair of discrete random variables. *IEEE Transactions on Information Theory*, 21(5):493–501.

[295] Witsenhausen, H. S. (1975). On sequences of pairs of dependent random variables. *SIAM Journal on Applied Mathematics*, 28(1):100–113.

[296] Wu, X., Li, F., Kumar, A., Chaudhuri, K., Jha, S., and Naughton, J. (2017). Bolt-on differential privacy for scalable stochastic gradient descent-based analytics. In *ACM International Conference on Management of Data*, pages 1307–1322.

[297] Wu, Y. (2020). Lecture notes on: Information-theoretic methods for high-dimensional statistics.

[298] Wu, Y. and Yang, P. (2016). Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Transactions on Information Theory*, 62(6):3702–3720.

[299] Xu, A. and Raginsky, M. (2016). Information-theoretic lower bounds on bayes risk in decentralized estimation. *IEEE Transactions on Information Theory*, 63(3):1580–1600.

[300] Xu, A. and Raginsky, M. (2017). Information-theoretic analysis of generalization capability of learning algorithms. In *Advances in Neural Information Processing Systems*, pages 2524–2533.

[301] Xu, P., Chen, J., Zou, D., and Gu, Q. (2018). Global convergence of Langevin dynamics based algorithms for nonconvex optimization. In *Advances in Neural Information Processing Systems*, pages 3122–3133.

[302] Yagli, S., Dytso, A., and Poor, H. V. (2020). Information-theoretic bounds on the generalization error and privacy leakage in federated learning. In *International Workshop on Signal Processing Advances in Wireless Communications*, pages 1–5.

[303] Yamamoto, H. (1983). A source coding problem for sources with additional outputs to keep secret from the receiver or wiretappers. *IEEE Transactions on Information Theory*, 29(6):918–923.

[304] Yu, B. (1997). Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer.

[305] Zafar, M. B., Valera, I., Rodriguez, M., Gummadi, K., and Weller, A. (2017a). From parity to preference-based notions of fairness in classification. In *Advances in Neural Information Processing Systems*, pages 229–239.

[306] Zafar, M. B., Valera, I., Rogriguez, M. G., and Gummadi, K. P. (2017b). Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pages 962–970.

[307] Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2017a). Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*.

[308] Zhang, Y., Liang, P., and Charikar, M. (2017b). A hitting time analysis of stochastic gradient Langevin dynamics. In *Conference on Learning Theory*, pages 1980–2022.

[309] Zhao, H. and Gordon, G. (2019). Inherent tradeoffs in learning fair representations. *Advances in neural information processing systems*, 32:15675–15685.

[310] Zhou, H. H., Zhang, Y., Ithapu, V. K., Johnson, S. C., Wahba, G., and Singh, V. (2017). When can multi-site datasets be pooled for regression? hypothesis tests, l_2-consistency and neuroscience applications. In *International Conference on Machine Learning*, pages 4170–4179. PMLR.

[311] Zhou, R., Tian, C., and Liu, T. (2021). Individually conditional individual mutual information bound on generalization error. In *IEEE International Symposium on Information Theory*, pages 670–675.

Appendix A

Appendix to Chapter 3

A.1 Proofs for Section 3.4

A.1.1 Proof of Lemma 7

Lemma 7 follows from a slight tweak of the proof of Theorem 4 in Polyanskiy and Wu [219].

Proof. First, we choose a coupling $P_{X,X'}$, which has marginals P_X and $P_{X'}$. The probability distribution of $X + mN$ can be written as the convolution of P_X and P_{mN}. Specifically,

$$\mathrm{d}P_{X+mN}(\boldsymbol{y}) = \int_{\boldsymbol{x}\in\mathcal{X}} \mathrm{d}P_{mN}(\boldsymbol{y}-\boldsymbol{x})\mathrm{d}P_X(\boldsymbol{x}) = \int_{\boldsymbol{x}\in\mathcal{X}}\int_{\boldsymbol{x}'\in\mathcal{X}} \mathrm{d}P_{\boldsymbol{x}+mN}(\boldsymbol{y})\mathrm{d}P_{X,X'}(\boldsymbol{x},\boldsymbol{x}').$$

Similarly, we have

$$\mathrm{d}P_{X'+mN}(\boldsymbol{y}') = \int_{\boldsymbol{x}\in\mathcal{X}}\int_{\boldsymbol{x}'\in\mathcal{X}} \mathrm{d}P_{\boldsymbol{x}'+mN}(\boldsymbol{y}')\mathrm{d}P_{X,X'}(\boldsymbol{x},\boldsymbol{x}').$$

Since the mapping $(P,Q) \to \mathrm{D}_f(P\|Q)$ is convex [see Theorem 6.1 in 221, for a proof], Jensen's inequality yields

$$\begin{aligned}\mathrm{D}_f\left(P_{X+mN}\|P_{X'+mN}\right) &\le \int_{\boldsymbol{x}\in\mathcal{X}}\int_{\boldsymbol{x}'\in\mathcal{X}} \mathrm{D}_f(P_{\boldsymbol{x}+mN}\|P_{\boldsymbol{x}'+mN})\mathrm{d}P_{X,X'}(\boldsymbol{x},\boldsymbol{x}') \\ &= \mathbb{E}\left[\mathsf{C}_f(X,X';m)\right], \end{aligned} \tag{A.1}$$

where the last step follows from the definition. The left-hand side of (A.1) only relies on the marginal distributions of X and X', so taking the infimum on both sides of (A.1) over all couplings $P_{X,X'}$ leads to the desired conclusion. □

A.1.2 Proof of Table 3.1

First, we derive a useful property of $\mathsf{C}_f(\boldsymbol{x}, \boldsymbol{y}; m)$ in the following lemma, which will be used for computing the closed-form expressions in Table 3.1.

Lemma 23. *For any $\boldsymbol{z} \in \mathbb{R}^d$ and $a > 0$, the function $\mathsf{C}_f(\boldsymbol{x}, \boldsymbol{x}'; m)$ in (3.15) satisfies*

$$\mathsf{C}_f(a\boldsymbol{x} + \boldsymbol{z}, a\boldsymbol{x}' + \boldsymbol{z}; m) = \mathsf{C}_f\left(\boldsymbol{x}, \boldsymbol{x}'; \frac{m}{a}\right), \quad \mathsf{C}_f(\boldsymbol{x}, \boldsymbol{x}'; m) = \mathsf{C}_f(-\boldsymbol{x}', -\boldsymbol{x}; m).$$

Proof. For simplicity, we assume that N is a continuous random variable in $\mathbb{R}^d$ with probability density function (PDF) $p(\boldsymbol{w})$. Then the PDFs of $a\boldsymbol{x} + \boldsymbol{z} + mN$ and $a\boldsymbol{x}' + \boldsymbol{z} + mN$ are

$$\frac{1}{m^d} \cdot p\left(\frac{\boldsymbol{w} - a\boldsymbol{x} - \boldsymbol{z}}{m}\right) \quad \text{and} \quad \frac{1}{m^d} \cdot p\left(\frac{\boldsymbol{w} - a\boldsymbol{x}' - \boldsymbol{z}}{m}\right).$$

By definition,

$$\begin{aligned} \mathsf{C}_f(a\boldsymbol{x} + \boldsymbol{z}, a\boldsymbol{x}' + \boldsymbol{z}; m) &= \mathrm{D}_f(P_{a\boldsymbol{x}+\boldsymbol{z}+mN} \| P_{a\boldsymbol{x}'+\boldsymbol{z}+mN}) \\ &= \frac{1}{m^d} \int_{\mathbb{R}^d} p\left(\frac{\boldsymbol{w} - a\boldsymbol{x}' - \boldsymbol{z}}{m}\right) f\left(\frac{p\left(\frac{\boldsymbol{w}-a\boldsymbol{x}-\boldsymbol{z}}{m}\right)}{p\left(\frac{\boldsymbol{w}-a\boldsymbol{x}'-\boldsymbol{z}}{m}\right)}\right) \mathrm{d}\boldsymbol{w}. \end{aligned} \tag{A.2}$$

Let $\boldsymbol{v} = (\boldsymbol{w} - \boldsymbol{z})/a$. Then (A.2) is equal to

$$\frac{a^d}{m^d} \int_{\mathbb{R}^d} p\left(\frac{\boldsymbol{v} - \boldsymbol{x}'}{m/a}\right) f\left(\frac{p\left(\frac{\boldsymbol{v}-\boldsymbol{x}}{m/a}\right)}{p\left(\frac{\boldsymbol{v}-\boldsymbol{x}'}{m/a}\right)}\right) \mathrm{d}\boldsymbol{v} = \mathsf{C}_f\left(\boldsymbol{x}, \boldsymbol{x}'; \frac{m}{a}\right).$$

Therefore, $\mathsf{C}_f(a\boldsymbol{x} + \boldsymbol{z}, a\boldsymbol{x}' + \boldsymbol{z}; m) = \mathsf{C}_f\left(\boldsymbol{x}, \boldsymbol{x}'; \frac{m}{a}\right)$. By choosing $a = 1$ and $\boldsymbol{z} = -\boldsymbol{x} - \boldsymbol{x}'$, we have $\mathsf{C}_f(-\boldsymbol{x}', -\boldsymbol{x}; m) = \mathsf{C}(\boldsymbol{x}, \boldsymbol{x}'; m)$. □

Next, we derive closed-form expressions (or upper bounds) for $\delta(A, m)$ in Table 3.1.

Proof. The closed-form expression of $\delta(A, m)$ for uniform distribution can be naturally obtained from its definition so we skip the proof. The closed-form expressions for standard multivariate Gaussian distribution and standard univariate Laplace distribution can be found at Polyanskiy and Wu [219] and Asoodeh et al. [19]. Hence, in what follows, we provide an upper bound for $\delta(A, m)$ when N follows a standard multivariate Laplace distribution. For a given positive number A and a random variable N which follows a standard multivariate Laplace distribution, consider the

following optimization problem:

$$\sup_{\|v\|_1 \le A} \mathrm{D}_{\mathrm{TV}}\left(P_N \| P_{v+N}\right) = \sup_{\|v\|_1 \le A} \mathbb{E}\left[\left(1 - \frac{\exp(-\|N-v\|_1)}{\exp(-\|N\|_1)}\right) \mathbb{I}_{\|N-v\|_1 \ge \|N\|_1}\right].$$

By exchanging the supremum and the expectation, we have

$$\sup_{\|v\|_1 \le A} \mathrm{D}_{\mathrm{TV}}\left(P_N \| P_{v+N}\right) \le \mathbb{E}\left[\sup_{\|v\|_1 \le A} \left\{\left(1 - \frac{\exp(-\|N-v\|_1)}{\exp(-\|N\|_1)}\right) \mathbb{I}_{\|N-v\|_1 \ge \|N\|_1}\right\}\right]. \tag{A.3}$$

Note that

$$\begin{aligned}
&\sup_{\|v\|_1 \le A} \left\{\left(1 - \frac{\exp(-\|N-v\|_1)}{\exp(-\|N\|_1)}\right) \mathbb{I}_{\|N-v\|_1 \ge \|N\|_1}\right\} \\
&= 1 - \exp\left(-\sup_{\|v\|_1 \le A} \{\|N-v\|_1 - \|N\|_1\}\right) = 1 - \exp(-A).
\end{aligned}$$

Substituting this equality into (A.3) gives

$$\sup_{\|x-x'\|_1 \le A} \mathrm{D}_{\mathrm{TV}}\left(P_{x+N} \| P_{x'+N}\right) = \sup_{\|v\|_1 \le A} \mathrm{D}_{\mathrm{TV}}\left(P_N \| P_{v+N}\right) \le 1 - \exp(-A),$$

which leads to $\delta(A, 1) \le 1 - \exp(-A)$. Finally, we have

$$\delta(A, m) = \delta\left(\frac{A}{m}, 1\right) \le 1 - \exp\left(-\frac{A}{m}\right).$$

□

Finally, we derive closed-form expressions (or upper bounds) for $\mathsf{C}_f(x, x'; m)$ in Table 3.1.

Proof. By Lemma 23, we have

$$\mathsf{C}_{\mathrm{KL}}(x, x'; m) = \mathsf{C}_{\mathrm{KL}}\left(0, \frac{x'-x}{m}; 1\right).$$

We denote $(x'-x)/m$ by v. Since all the coordinates of $N = (N_1 \cdots, N_d)$ are mutually independent, $P_N = P_{N_1} \cdots P_{N_d}$ and $P_{v+N} = P_{v_1+N_1} \cdots P_{v_d+N_d}$. By the chain rule of KL-divergence, we have

$$\mathsf{C}_{\mathrm{KL}}(x, x'; m) = \mathrm{D}_{\mathrm{KL}}(P_N \| P_{v+N}) = \sum_{i=1}^{d} \mathrm{D}_{\mathrm{KL}}(P_{N_i} \| P_{v_i+N_i}). \tag{A.4}$$

Hence, we only need to calculate $\mathrm{D}_{\mathrm{KL}}(P_N \| P_{v+N})$ for a constant $v \in \mathbb{R}$ and a random variable $N \in \mathbb{R}$.

(1) If N follows a standard Gaussian distribution, then

$$\begin{aligned}\mathrm{D}_{\mathrm{KL}}(P_N \| P_{v+N}) &= \mathbb{E}\left[\log \frac{\exp(-N^2/2)}{\exp(-(N-v)^2/2)}\right] \\ &= \frac{1}{2}\mathbb{E}\left[(N-v)^2 - N^2\right] = \frac{v^2}{2}.\end{aligned}$$

Substituting this equality into (A.4) gives

$$\mathsf{C}_{\mathrm{KL}}(\boldsymbol{x}, \boldsymbol{x}'; m) = \frac{\|\boldsymbol{v}\|_2^2}{2} = \frac{\|\boldsymbol{x} - \boldsymbol{x}'\|_2^2}{2m^2},$$

where the last step is due to the definition of $\boldsymbol{v}$.

(2) If N follows a standard Laplace distribution, then

$$\mathrm{D}_{\mathrm{KL}}(P_N \| P_{v+N}) = \mathbb{E}\left[\log \frac{\exp(-|N|)}{\exp(-|N-v|)}\right] = |v| + \exp(-|v|) - 1.$$

Substituting this equality into (A.4) gives

$$\begin{aligned}\mathsf{C}_{\mathrm{KL}}(\boldsymbol{x}, \boldsymbol{x}'; m) &= \sum_{i=1}^{d} |v_i| + \exp(-|v_i|) - 1 = \frac{\|\boldsymbol{x} - \boldsymbol{x}'\|_1}{m} + \sum_{i=1}^{d}\left(\exp\left(-\frac{|x_i - x_i'|}{m}\right) - 1\right) \\ &\leq \frac{\|\boldsymbol{x} - \boldsymbol{x}'\|_1}{m}.\end{aligned}$$

Similarly, by Lemma 23, we have

$$\mathsf{C}_{\chi^2}(\boldsymbol{x}, \boldsymbol{x}'; m) = \mathsf{C}_{\chi^2}\left(0, \frac{\boldsymbol{x}' - \boldsymbol{x}}{m}; 1\right).$$

We denote $(\boldsymbol{x}' - \boldsymbol{x})/m$ by $\boldsymbol{v}$. By the property of χ^2-divergence [see Section 2.4 in 266], we have

$$\mathsf{C}_{\chi^2}(\boldsymbol{x}, \boldsymbol{x}'; m) = \mathrm{D}_{\chi^2_N}(P_N \| P_{\boldsymbol{v}+N}) = \prod_{i=1}^{d}\left(1 + \mathrm{D}_{\chi^2_N}(P_{N_i} \| P_{v_i+N_i})\right) - 1. \tag{A.5}$$

Hence, we only need to calculate $\mathrm{D}_{\chi^2_N}(P_N \| P_{v+N})$ for $v \in \mathbb{R}$ and $N \in \mathbb{R}$.

(1) If N follows a standard Gaussian distribution, then

$$\begin{aligned}\mathrm{D}_{\chi^2_N}(P_N \| P_{v+N}) &= \mathbb{E}\left[\frac{\exp(-N^2/2)}{\exp(-(N-v)^2/2)}\right] - 1 \\ &= \exp(v^2/2)\mathbb{E}\left[\exp(-vN)\right] - 1 = \exp(v^2) - 1.\end{aligned}$$

Substituting this equality into (A.5) gives

$$\mathsf{C}_{\chi^2}(\boldsymbol{x}, \boldsymbol{x}'; m) = \exp(\|\boldsymbol{v}\|_2^2) - 1 = \exp\left(\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|_2^2}{m^2}\right) - 1.$$

(2) If N follows a standard Laplace distribution, then

$$\begin{aligned}\mathsf{D}_{\chi^2_N}(P_N \| P_{v+N}) &= \mathbb{E}\left[\frac{\exp(-|N|)}{\exp(-|N-v|)}\right] - 1 \\ &= \frac{2}{3}\exp(|v|) + \frac{1}{3}\exp(-2|v|) - 1.\end{aligned}$$

Substituting this equality into (A.5) gives

$$\begin{aligned}\mathsf{C}_{\chi^2}(\boldsymbol{x}, \boldsymbol{x}'; m) &= \prod_{i=1}^{d}\left(\frac{2}{3}\exp\left(\frac{|x_i - x_i'|}{m}\right) + \frac{1}{3}\exp\left(\frac{-2|x_i - x_i'|}{m}\right)\right) - 1 \\ &\leq \exp\left(\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|_1}{m}\right) - 1.\end{aligned}$$

Finally, we use Pinsker's inequality [see Theorem 4.5 in 297, for a proof] for proving an upper bound of $\mathsf{C}_{\mathrm{TV}}(\boldsymbol{x}, \boldsymbol{x}'; m)$:

$$\begin{aligned}\mathsf{C}_{\mathrm{TV}}(\boldsymbol{x}, \boldsymbol{x}'; m) &= \mathsf{D}_{\mathrm{TV}}\left(P_{\boldsymbol{x}+mN} \| P_{\boldsymbol{x}'+mN}\right) \\ &\leq \sqrt{\frac{\mathsf{D}_{\mathrm{KL}}\left(P_{\boldsymbol{x}+mN} \| P_{\boldsymbol{x}'+mN}\right)}{2}} \\ &= \sqrt{\frac{\mathsf{C}_{\mathrm{KL}}(\boldsymbol{x}, \boldsymbol{x}'; m)}{2}}.\end{aligned}$$

Hence, any upper bound of $\mathsf{C}_{\mathrm{KL}}(\boldsymbol{x}, \boldsymbol{x}'; m)$ can be naturally translated into an upper bound for $\mathsf{C}_{\mathrm{TV}}(\boldsymbol{x}, \boldsymbol{x}'; m)$. This is how we obtain the upper bounds of $\mathsf{C}_{\mathrm{TV}}(\boldsymbol{x}, \boldsymbol{x}'; m)$ under Gaussian or Laplace distribution in Table 3.1. On the other hand, if N follows a uniform distribution on $[-1, 1] \subseteq \mathbb{R}$, by Lemma 23 we have

$$\mathsf{C}_{\mathrm{TV}}(x, x'; m) = \mathsf{C}_{\mathrm{TV}}\left(0, \frac{x' - x}{m}; 1\right) = \min\left\{1, \left|\frac{x - x'}{2m}\right|\right\}.$$

Note that in this case $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}$ so we write them as x, x'. □

Remark 20. We used Pinsker's inequality for deriving an upper bound of $\mathsf{C}_{\mathrm{TV}}(\boldsymbol{x}, \boldsymbol{x}'; m)$ in the above proof. One can potentially tighten this bound by exploring other f-divergence inequalities [see e.g., Eq. 4 in 242].

A.2 Proofs for Section 3.5

A.2.1 Proof of Theorem 2

Proof. Combining Lemma 8 and 9 together leads to an upper bound of $I_f(W_T; Z_i)$ for any data point Z_i used at the t-th iteration:

$$I_f(W_T; Z_i) \leq \mathbb{E}\left[\mathsf{C}_f\left(g(W_{t-1}, Z), g(W_{t-1}, \bar{Z}); \frac{m_t b_t}{\eta_t}\right)\right] \cdot \prod_{t'=t+1}^{T} \delta(D + 2\eta_{t'} K, m_{t'}). \tag{A.6}$$

Additionally, if the loss $\ell(\boldsymbol{w}, Z)$ is σ-sub-Gaussian for all $\boldsymbol{w} \in \mathcal{W}$, Lemma 5 and Assumption 1 altogether yield

$$\begin{aligned}\left|\mathbb{E}\left[L_\mu(W_T) - L_S(W_T)\right]\right| &\leq \frac{1}{n}\sum_{i=1}^{n}\sqrt{2\sigma^2 I(W_T; Z_i)} \\ &= \frac{1}{n}\sum_{t=1}^{T}\sum_{i\in\mathcal{B}_t}\sqrt{2\sigma^2 I(W_T; Z_i)}.\end{aligned} \tag{A.7}$$

Substituting (A.6) into (A.7) yields the following upper bound of the expected generalization gap:

$$\begin{aligned}&\frac{\sqrt{2}\sigma}{n}\sum_{t=1}^{T}\sum_{i\in\mathcal{B}_t}\sqrt{\mathbb{E}\left[\mathsf{C}_{\mathrm{KL}}\left(g(W_{t-1}, Z), g(W_{t-1}, \bar{Z}); \frac{m_t b_t}{\eta_t}\right)\right] \cdot \prod_{t'=t+1}^{T} \delta(D + 2\eta_{t'} K, m_{t'})} \\ &= \frac{\sqrt{2}\sigma}{n}\sum_{t=1}^{T} b_t\sqrt{\mathbb{E}\left[\mathsf{C}_{\mathrm{KL}}\left(g(W_{t-1}, Z), g(W_{t-1}, \bar{Z}); \frac{m_t b_t}{\eta_t}\right)\right] \cdot \prod_{t'=t+1}^{T} \delta(D + 2\eta_{t'} K, m_{t'})}.\end{aligned}$$

Similarly, we can obtain another two generalization bounds using Lemma 5 and the upper bound in (A.6). □

A.3 Proofs for Section 3.6

A.3.1 Proof of Proposition 1 and 2

In the setting of DP-SGD, the three generalization bounds in Theorem 2 become

$$\frac{\sqrt{2}\sigma}{n}\sum_{t=1}^{T} b_t\sqrt{\mathbb{E}\left[\mathsf{C}_{\mathrm{KL}}\left(g(W_{t-1}, Z), g(W_{t-1}, \bar{Z}); b_t\right)\right] \cdot \left(\delta(D + 2\eta K, \eta)\right)^{T-t}} \tag{A.8}$$

where σ is the sub-Gaussian constant;

$$\frac{A}{n}\sum_{t=1}^{T} b_t\mathbb{E}\left[\mathsf{C}_{\mathrm{TV}}\left(g(W_{t-1}, Z), g(W_{t-1}, \bar{Z}); b_t\right)\right] \cdot \left(\delta(D + 2\eta K, \eta)\right)^{T-t} \tag{A.9}$$

where A is an upper bound of the loss function; and

$$\frac{\sigma}{n}\sum_{t=1}^{T} b_t \sqrt{\mathbb{E}\left[\mathsf{C}_{\chi^2}\left(g(W_{t-1},Z),g(W_{t-1},\bar{Z});b_t\right)\right]\cdot\left(\delta(D+2\eta K,\eta)\right)^{T-t}} \tag{A.10}$$

where $\sigma \triangleq \sqrt{\mathsf{Var}\left(\ell(W_T;Z)\right)}$.

Proof. We prove Proposition 2 first.

- If the additive noise follows a standard multivariate Gaussian distribution, Table 3.1 shows that

$$\delta(D+2\eta K,\eta) = 1-2\bar{\Phi}\left(\frac{D+2\eta K}{2\eta}\right), \tag{A.11}$$

$$\mathbb{E}\left[\mathsf{C}_{\mathrm{KL}}\left(g(W_{t-1},Z),g(W_{t-1},\bar{Z});b_t\right)\right] = \frac{1}{2b_t^2}\mathbb{E}\left[\|g(W_{t-1},Z)-g(W_{t-1},\bar{Z})\|_2^2\right]. \tag{A.12}$$

We introduce a constant vector $\boldsymbol{e}$ whose value will be specified later. Since $\|\boldsymbol{a}-\boldsymbol{b}\|_2^2 \le 2\|\boldsymbol{a}\|_2^2 + 2\|\boldsymbol{b}\|_2^2$, we have

$$\begin{aligned}
&\frac{1}{2b_t^2}\mathbb{E}\left[\|g(W_{t-1},Z)-g(W_{t-1},\bar{Z})\|_2^2\right]\\
&\le \frac{1}{b_t^2}\left(\mathbb{E}\left[\|g(W_{t-1},Z)-\boldsymbol{e}\|_2^2\right]+\mathbb{E}\left[\|g(W_{t-1},\bar{Z})-\boldsymbol{e}\|_2^2\right]\right)\\
&= \frac{2}{b_t^2}\mathbb{E}\left[\|g(W_{t-1},Z)-\boldsymbol{e}\|_2^2\right],
\end{aligned} \tag{A.13}$$

where the last step is because W_{t-1} is independent of $(Z,\bar{Z})$ and $Z,\bar{Z}$ follow the same distribution. By choosing the constant vector $\boldsymbol{e} = \mathbb{E}\left[g(W_{t-1},Z)\right]$, we have

$$\mathbb{E}\left[\|g(W_{t-1},Z)-\boldsymbol{e}\|_2^2\right] = \mathsf{Var}\left(g(W_{t-1},Z)\right). \tag{A.14}$$

Combining (A.12–A.14) gives

$$\mathbb{E}\left[\mathsf{C}_{\mathrm{KL}}\left(g(W_{t-1},Z),g(W_{t-1},\bar{Z});b_t\right)\right] \le \frac{2}{b_t^2}\mathsf{Var}\left(g(W_{t-1},Z)\right). \tag{A.15}$$

Substituting (A.11), (A.15) into (A.8) leads to the generalization bound in (3.37).

- Similarly, Table 3.1 shows for Gaussian noise

$$\mathbb{E}\left[\mathsf{C}_{\mathrm{TV}}\left(g(W_{t-1},Z),g(W_{t-1},\bar{Z});b_t\right)\right] \le \frac{1}{2b_t}\mathbb{E}\left[\|g(W_{t-1},Z)-g(W_{t-1},\bar{Z})\|_2\right]. \tag{A.16}$$

Furthermore, by the triangle inequality,

$$\begin{aligned}
&\frac{1}{2b_t}\mathbb{E}\left[\|g(W_{t-1},Z)-g(W_{t-1},\bar{Z})\|_2\right] \\
&\leq \frac{1}{2b_t}\left(\mathbb{E}\left[\|g(W_{t-1},Z)-\boldsymbol{e}\|_2\right]+\mathbb{E}\left[\|g(W_{t-1},\bar{Z})-\boldsymbol{e}\|_2\right]\right) \\
&= \frac{1}{b_t}\mathbb{E}\left[\|g(W_{t-1},Z)-\boldsymbol{e}\|_2\right]. \qquad \text{(A.17)}
\end{aligned}$$

By choosing the constant vector $\boldsymbol{e}=\mathbb{E}\left[g(W_{t-1},Z)\right]$ and combining (A.16) with (A.17), we have

$$\mathbb{E}\left[\mathsf{C}_{\mathrm{TV}}\left(g(W_{t-1},Z),g(W_{t-1},\bar{Z});b_t\right)\right]\leq\frac{1}{b_t}\mathbb{E}\left[\|g(W_{t-1},Z)-\boldsymbol{e}\|_2\right]. \qquad \text{(A.18)}$$

Substituting (A.11), (A.18) into (A.9) leads to the generalization bound in (3.38).

- Finally, Table 3.1 shows for Gaussian noise

$$\mathbb{E}\left[\mathsf{C}_{\chi^2}\left(g(W_{t-1},Z),g(W_{t-1},\bar{Z});b_t\right)\right]=\mathbb{E}\left[\exp\left(\frac{\|g(W_{t-1},Z)-g(W_{t-1},\bar{Z})\|_2^2}{b_t^2}\right)\right]-1. \qquad \text{(A.19)}$$

The Cauchy-Schwarz inequality implies that

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\frac{\|g(W_{t-1},Z)-g(W_{t-1},\bar{Z})\|_2^2}{b_t^2}\right)\right] \\
&\leq \mathbb{E}\left[\exp\left(\frac{2\|g(W_{t-1},Z)-\boldsymbol{e}\|_2^2}{b_t^2}\right)\exp\left(\frac{2\|g(W_{t-1},\bar{Z})-\boldsymbol{e}\|_2^2}{b_t^2}\right)\right] \\
&\leq \sqrt{\mathbb{E}\left[\exp\left(\frac{4\|g(W_{t-1},Z)-\boldsymbol{e}\|_2^2}{b_t^2}\right)\right]\mathbb{E}\left[\exp\left(\frac{4\|g(W_{t-1},\bar{Z})-\boldsymbol{e}\|_2^2}{b_t^2}\right)\right]} \\
&= \mathbb{E}\left[\exp\left(\frac{4\|g(W_{t-1},Z)-\boldsymbol{e}\|_2^2}{b_t^2}\right)\right]. \qquad \text{(A.20)}
\end{aligned}$$

By choosing the constant vector $\boldsymbol{e}=\mathbb{E}\left[g(W_{t-1},Z)\right]$ and combining (A.19) with (A.20), we have

$$\mathbb{E}\left[\mathsf{C}_{\chi^2}\left(g(W_{t-1},Z),g(W_{t-1},\bar{Z});b_t\right)\right]\leq\mathbb{E}\left[\exp\left(\frac{4\|g(W_{t-1},Z)-\boldsymbol{e}\|_2^2}{b_t^2}\right)\right]-1. \qquad \text{(A.21)}$$

Since for any $x\geq 0$ and $b\geq 1$,

$$\exp\left(\frac{x}{b}\right)-1\leq\frac{\exp(x)-1}{b},$$

the inequality in (A.21) can be further upper bounded as

$$\mathbb{E}\left[\mathsf{C}_{\chi^2}\left(g(W_{t-1},Z),g(W_{t-1},\bar{Z});b_t\right)\right]\leq\frac{1}{b_t^2}\left(\mathbb{E}\left[\exp\left(4\|g(W_{t-1},Z)-\boldsymbol{e}\|_2^2\right)\right]-1\right). \qquad \text{(A.22)}$$

Substituting (A.11), (A.22) into (A.10) leads to the generalization bound in (3.39).

□

By a similar analysis, we can prove the generalization bounds in Proposition 1 for the Laplace mechanism.

A.3.2 Proof of Proposition 3

Proof. Within the t-th global update, we can rewrite the local updates conducted by the client $k \in \mathcal{S}_t$ as follows. The parameter is initialized by $W_{t,0}^k = W_{t-1}$ and for $j \in [M]$,

$$U_{t,j}^k = W_{t,j-1}^k - \eta \cdot g\left(W_{t,j-1}^k, \{Z_i^k\}_{i\in[b]}\right) \tag{A.23a}$$

$$V_{t,j}^k = U_{t,j}^k + \eta \cdot N \tag{A.23b}$$

$$W_{t,j}^k = \mathsf{Proj}_{\mathcal{W}}\left(V_{t,j}^k\right) \tag{A.23c}$$

where $\{Z_i^k\}_{i\in[b]}$ are drawn independently from the data distribution μ_k and $N \sim N(0, \mathbf{I}_d)$. If a data point Z_i^k is used at the t-th global update, j-th local update, then the following Markov chain holds:

$$\underbrace{Z_i^k \to \{U_{t,j}^k\}_{k\in\mathcal{S}_t} \to \{V_{t,j}^k\}_{k\in\mathcal{S}_t} \to \{W_{t,j}^k\}_{k\in\mathcal{S}_t} \to \cdots \to \{W_{t,M}^k\}_{k\in\mathcal{S}_t}}_{\text{local}} \underbrace{\to W_t}_{\text{global}} \to \cdots \to W_T$$

Hence, following a similar analysis in the proof of Lemma 8, we have

$$\begin{aligned}\mathrm{T}(W_T; Z_i^k) &\le q^{M(T-t)} \cdot \mathrm{T}(W_t; Z_i^k) \\ &\le q^{(M-j)+M(T-t)} \cdot \mathrm{T}(\{W_{t,j}^k\}_{k\in\mathcal{S}_t}; Z_i^k),\end{aligned} \tag{A.24}$$

where the constant q is defined as

$$q \triangleq 1 - 2\bar{\Phi}\left(\frac{\sqrt{C}(D + 2\eta K)}{2\eta}\right).$$

Analogous to the proof of Lemma 9, we have

$$\mathrm{T}(\{W_{t,j}^k\}_{k\in\mathcal{S}_t}; Z_i^k) \le \frac{1}{b}\mathbb{E}\left[\|g(W_{t,j-1}^k, Z^k) - \boldsymbol{e}\|_2\right] \tag{A.25}$$

where $\boldsymbol{e} \triangleq \mathbb{E}\left[g(W_{t,j-1}^k, Z^k)\right]$. Combining (A.24), (A.25) with the T-information bound in Lemma 5 yields the desired generalization bound for the k-th client. □

A.3.3 Proof of Proposition 4

We first present the following lemma whose proof follows by using the technique in Section II. E of Guo et al. [115].

Lemma 24. *Let X be a random variable which is independent of $N \sim N(0, \mathbf{I}_d)$. Then for any $m > 0$ and deterministic function f*

$$I(f(X) + mN; X) \leq \frac{1}{2m^2} \mathsf{Var}\left(f(X)\right). \tag{A.26}$$

More generally, if Z is another random variable which is independent of N, then for any fixed $\boldsymbol{z}$

$$I(f(X) + mN; X \mid Z = \boldsymbol{z}) \leq \frac{1}{2m^2} \mathsf{Var}\left(f(X) \mid Z = \boldsymbol{z}\right). \tag{A.27}$$

Proof. By the property of mutual information [see Theorem 2.3 in 221],

$$I(f(X) + mN; X) = I\left(\frac{f(X) - \boldsymbol{e}}{m} + N; X\right) \tag{A.28}$$

where $\boldsymbol{e} \triangleq \mathbb{E}\left[f(X)\right]$. We denote

$$g(\boldsymbol{x}) \triangleq \frac{f(\boldsymbol{x}) - \boldsymbol{e}}{m}. \tag{A.29}$$

The golden formula [see Theorem 3.3 in 221, for a proof] yields

$$\begin{aligned} I\left(g(X) + N; X\right) &= \mathrm{D}_{\mathrm{KL}}\left(P_{g(X)+N|X} \| P_N | P_X\right) - \mathrm{D}_{\mathrm{KL}}\left(P_{g(X)+N} \| P_N\right) \\ &\leq \mathrm{D}_{\mathrm{KL}}\left(P_{g(X)+N|X} \| P_N | P_X\right). \end{aligned} \tag{A.30}$$

Furthermore, since X and N are independent, we have

$$\mathrm{D}_{\mathrm{KL}}\left(P_{g(X)+N|X=\boldsymbol{x}} \| P_N\right) = \mathrm{D}_{\mathrm{KL}}\left(P_{g(\boldsymbol{x})+N} \| P_N\right) = \frac{\|g(\boldsymbol{x})\|_2^2}{2},$$

where the last step is due to the closed-form expression of the KL-divergence between two Gaussian distributions. Finally, by the definition of conditional divergence, we have

$$\mathrm{D}_{\mathrm{KL}}\left(P_{g(X)+N|X} \| P_N | P_X\right) = \frac{1}{2}\mathbb{E}\left[\|g(X)\|_2^2\right] = \frac{1}{2m^2} \mathsf{Var}\left(f(X)\right), \tag{A.31}$$

where the last step is due to the definition of g in (A.29). Combining (A.28–A.31) leads to the desired conclusion. Finally, it is straightforward to obtain (A.27) by conditioning on $Z = \boldsymbol{z}$ and repeating our above derivations. □

Next, we present the second lemma which will be used for proving Proposition 4.

Lemma 25. *If the loss function $\ell(\boldsymbol{w}, Z)$ is σ-sub-Gaussian under $Z \sim \mu$ for all $\boldsymbol{w} \in \mathcal{W}$, the expected generalization gap of the SGLD algorithm can be upper bounded by*

$$\frac{\sqrt{2}\sigma}{2n} \sum_{j=1}^{m} \sqrt{\sum_{t \in \mathcal{T}_j} \beta_t \eta_t \cdot \mathsf{Var}\left(\nabla_{\boldsymbol{w}} \hat{\ell}(W_{t-1}, \bar{Z}_j)\right)},$$

where the set $\mathcal{T}_j$ contains the indices of iterations in which the mini-batch S_j is used and the variance is over the randomness of $(W_{t-1}, \bar{Z}_j) \sim P_{W_{t-1}, \bar{Z}_j}$ with $\bar{Z}_j$ being any data point in the mini-batch S_j.

Proof. We denote $Z^{(k)} \triangleq (Z_1, \cdots, Z_k)$ for $k \in [n]$ and $W^{(t)} \triangleq (W_1, \cdots, W_t)$ for $t \in [T]$. For simplicity, in what follows we only provide an upper bound for $I(W; Z_n)$. Since W is a function of $W^{(T)} = (W_1, \cdots, W_T)$, the data processing inequality yields

$$I(W; Z_n) \leq I(W^{(T)}; Z_n) \leq I(W^{(T)}, Z^{(n-1)}; Z_n). \tag{A.32}$$

By the chain rule,

$$I(W^{(T)}, Z^{(n-1)}; Z_n) = I(W_T; Z_n \mid W^{(T-1)}, Z^{(n-1)}) + I(W^{(T-1)}, Z^{(n-1)}; Z_n). \tag{A.33}$$

Let $\boldsymbol{w} = (\boldsymbol{w}_1, \cdots, \boldsymbol{w}_{T-1})$ and $\boldsymbol{z} = (\boldsymbol{z}_1, \cdots, \boldsymbol{z}_{n-1})$ be any two vectors. If Z_n is not used at the T-th iteration, without loss of generality we assume that the data points $Z_1, \cdots, Z_b$ are used in this iteration. Then

$$\begin{aligned}
&I(W_T; Z_n \mid W^{(T-1)} = \boldsymbol{w}, Z^{(n-1)} = \boldsymbol{z}) \\
&= I\left(\boldsymbol{w}_{t-1} - \frac{\eta_T}{b} \sum_{i=1}^{b} \nabla_{\boldsymbol{w}} \hat{\ell}(\boldsymbol{w}_{T-1}, \boldsymbol{z}_i) + \sqrt{\frac{2\eta_T}{\beta_T}} N; Z_n \mid W^{(T-1)} = \boldsymbol{w}, Z^{(n-1)} = \boldsymbol{z}\right) \\
&= I\left(N; Z_n \mid W^{(T-1)} = \boldsymbol{w}, Z^{(n-1)} = \boldsymbol{z}\right) \\
&= 0.
\end{aligned} \tag{A.34}$$

On the other hand, if Z_n is used at the T-th iteration, without loss of generality we assume that the

other $b-1$ data points which are also used in this iteration are $Z_1, \cdots, Z_{b-1}$. Then

$$
\begin{aligned}
&I(W_T; Z_n \mid W^{(T-1)} = \boldsymbol{w}, Z^{(n-1)} = \boldsymbol{z}) \\
&= I\left(\boldsymbol{w}_{t-1} - \frac{\eta_T}{b}\left(\sum_{i=1}^{b-1} \nabla_{\boldsymbol{w}} \hat{\ell}(\boldsymbol{w}_{T-1}, \boldsymbol{z}_i) + \nabla_{\boldsymbol{w}} \hat{\ell}(\boldsymbol{w}_{T-1}, Z_n)\right) + \sqrt{\frac{2\eta_T}{\beta_T}} N; Z_n \mid W^{(T-1)} = \boldsymbol{w}, Z^{(n-1)} = \boldsymbol{z}\right) \\
&= I\left(-\frac{\eta_T}{b} \nabla_{\boldsymbol{w}} \hat{\ell}(\boldsymbol{w}_{T-1}, Z_n) + \sqrt{\frac{2\eta_T}{\beta_T}} N; Z_n \mid W^{(T-1)} = \boldsymbol{w}, Z^{(n-1)} = \boldsymbol{z}\right). \qquad \text{(A.35)}
\end{aligned}
$$

By Lemma 24, we have

$$
\begin{aligned}
&I\left(-\frac{\eta_T}{b} \nabla_{\boldsymbol{w}} \hat{\ell}(\boldsymbol{w}_{T-1}, Z_n) + \sqrt{\frac{2\eta_T}{\beta_T}} N; Z_n \mid W^{(T-1)} = \boldsymbol{w}, Z^{(n-1)} = \boldsymbol{z}\right) \\
&\leq \frac{\beta_T \eta_T}{4b^2} \mathsf{Var}\left(\nabla_{\boldsymbol{w}} \hat{\ell}(\boldsymbol{w}_{T-1}, Z_n) \mid W^{(T-1)} = \boldsymbol{w}, Z^{(n-1)} = \boldsymbol{z}\right). \qquad \text{(A.36)}
\end{aligned}
$$

Substituting (A.36) into (A.35) gives

$$
I(W_T; Z_n \mid W^{(T-1)} = \boldsymbol{w}, Z^{(n-1)} = \boldsymbol{z}) \leq \frac{\beta_T \eta_T}{4b^2} \mathsf{Var}\left(\nabla_{\boldsymbol{w}} \hat{\ell}(\boldsymbol{w}_{T-1}, Z_n) \mid W^{(T-1)} = \boldsymbol{w}, Z^{(n-1)} = \boldsymbol{z}\right).
$$

Taking expectation w.r.t. $(W^{(T-1)}, Z^{(n-1)})$ on both sides of the above inequality and using the law of total variance lead to

$$
I(W_T; Z_n \mid W^{(T-1)}, Z^{(n-1)}) \leq \frac{\beta_T \eta_T}{4b^2} \mathsf{Var}\left(\nabla_{\boldsymbol{w}} \hat{\ell}(W_{T-1}, Z_n)\right). \qquad \text{(A.37)}
$$

To summarize, (A.34) and (A.37) can be rewritten as

$$
\begin{aligned}
&I(W_T; Z_n \mid W^{(T-1)}, Z^{(n-1)}) \\
&\leq \begin{cases} \frac{\beta_T \eta_T}{4b^2} \mathsf{Var}\left(\nabla_{\boldsymbol{w}} \hat{\ell}(W_{T-1}, Z_n)\right) & \text{if } Z_n \text{ is used at the } T\text{-th iteration,} \\ 0 & \text{otherwise.} \end{cases} \qquad \text{(A.38)}
\end{aligned}
$$

Assume that the data point Z_n belongs to the j-th mini-batch S_j. Now substituting (A.38) into (A.33) and doing this procedure recursively lead to

$$
I(W^{(T)}, Z^{(n-1)}; Z_n) \leq \sum_{t \in \mathcal{T}_j} \frac{\beta_t \eta_t}{4b^2} \mathsf{Var}\left(\nabla_{\boldsymbol{w}} \hat{\ell}(W_{t-1}, Z_n)\right),
$$

where the set $\mathcal{T}_j$ contains the indices of iterations in which the mini-batch S_j is used. Hence, this upper bound along with (A.32) naturally gives

$$
I(W; Z_n) \leq \sum_{t \in \mathcal{T}_j} \frac{\beta_t \eta_t}{4b^2} \mathsf{Var}\left(\nabla_{\boldsymbol{w}} \hat{\ell}(W_{t-1}, Z_n)\right). \qquad \text{(A.39)}
$$

By symmetry, for any data point in S_j besides Z_n, the mutual information between W and this data point can be upper bound by the right-hand side of (A.39) as well. Finally, recall that Lemma 5 provides an upper bound for the expected generalization gap:

$$\frac{\sqrt{2}\sigma}{n}\sum_{i=1}^{n}\sqrt{I(W_T;Z_i)} = \frac{\sqrt{2}\sigma}{n}\sum_{j=1}^{m}\sum_{Z\in S_j}\sqrt{I(W_T;Z)}. \tag{A.40}$$

By substituting (A.39) into the above expression, we know the expected generalization gap can be further upper bounded by

$$\frac{\sqrt{2}\sigma}{2n}\sum_{j=1}^{m}\sqrt{\sum_{t\in\mathcal{T}_j}\beta_t\eta_t\cdot\mathsf{Var}\left(\nabla_{\boldsymbol{w}}\hat{\ell}(W_{t-1},\bar{Z}_j)\right)},$$

where $\bar{Z}_j$ is any data point in the mini-batch S_j. □

Finally, we are in a position to prove Proposition 4.

Proof. Consider a new loss function and the gradient of a new surrogate loss:

$$\ell(\boldsymbol{w},S_j) \triangleq \frac{1}{b}\sum_{Z\in S_j}\ell(\boldsymbol{w},Z), \quad \nabla_{\boldsymbol{w}}\hat{\ell}(\boldsymbol{w},S_j) \triangleq \frac{1}{b}\sum_{Z\in S_j}\nabla_{\boldsymbol{w}}\hat{\ell}(\boldsymbol{w},Z).$$

Then $\ell(\boldsymbol{w},S_j)$ is $\sigma/\sqrt{b}$-sub-Gaussian under $S_j\sim\mu^{\otimes b}$ for all $\boldsymbol{w}\in\mathcal{W}$. We view each mini-batch S_j as a data point and view $\ell(\boldsymbol{w},S_j)$ as a new loss function. By using Lemma 25, we obtain:

$$\left|\mathbb{E}\left[L_\mu(W)-L_S(W)\right]\right| \le \frac{\sqrt{2}\sigma}{2m\sqrt{b}}\sum_{j=1}^{m}\sqrt{\sum_{t\in\mathcal{T}_j}\beta_t\eta_t\cdot\mathsf{Var}\left(\nabla_{\boldsymbol{w}}\hat{\ell}(W_{t-1},S_j)\right)}. \tag{A.41}$$

Since the dataset contains n data points and is divided into m disjoint mini-batches with size b, we have $n=mb$. Substituting this into (A.41) leads to the desired conclusion. □

A.3.4 Proof of Corollary 1

Proof. The Minkowski inequality implies that for any non-negative x_i, the inequality $\sqrt{\sum_i x_i}\le\sum_i\sqrt{x_i}$ holds. Therefore, we can further upper bound the generalization bound in Lemma 25 by

$$\frac{\sqrt{2}\sigma}{2n}\sum_{j=1}^{m}\sum_{t\in\mathcal{T}_j}\sqrt{\beta_t\eta_t\cdot\mathsf{Var}\left(\nabla_{\boldsymbol{w}}\hat{\ell}(W_{t-1},\bar{Z}_j)\right)} = \frac{\sqrt{2}\sigma}{2n}\sum_{t=1}^{T}\sqrt{\beta_t\eta_t\cdot\mathsf{Var}\left(\nabla_{\boldsymbol{w}}\hat{\ell}(W_{t-1},Z_t^\dagger)\right)}.$$

Alternatively, by Jensen's inequality and $n = mb$, we can further upper bound the generalization bound in Lemma 25 by

$$\frac{\sqrt{2}\sigma}{2}\sqrt{\frac{1}{bn}\sum_{t=1}^{T}\beta_t\eta_t \cdot \mathsf{Var}\left(\nabla_w \hat{\ell}(W_{t-1}, Z_t^{\dagger})\right)}.$$

□

Appendix B

Appendix to Chapter 4

B.1 Examples of f-divergence

We recall some examples of f-divergence [70] here.

- KL-divergence [165]: $f(x) = x\log(x)$,

$$\mathrm{D}_{\mathrm{KL}}(P\|Q) = \int \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)\mathrm{d}P. \tag{B.1}$$

- Total variation distance: $f(x) = |x-1|/2$,

$$\mathrm{D}_{\mathrm{TV}}(P\|Q) = \frac{1}{2}\int \left|\frac{\mathrm{d}P}{\mathrm{d}Q} - 1\right| \mathrm{d}Q. \tag{B.2}$$

- Chi-square divergence [213]: $f(x) = (x-1)^2$ or $f(x) = x^2 - 1$,

$$\mathrm{D}_{\chi^2}(P\|Q) = \int \left(\frac{\mathrm{d}P}{\mathrm{d}Q} - 1\right)^2 \mathrm{d}Q = \int \left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)^2 \mathrm{d}Q - 1. \tag{B.3}$$

- Jensen-Shannon divergence [182]: $f(x) = x\log(x)/2 - (1+x)\log((1+x)/2)/2$,

$$\mathsf{JS}(P\|Q) = \frac{1}{2}\mathrm{D}_{\mathrm{KL}}\left(P\|\tfrac{P+Q}{2}\right) + \frac{1}{2}\mathrm{D}_{\mathrm{KL}}\left(Q\|\tfrac{P+Q}{2}\right). \tag{B.4}$$

  Note that the Jensen-Shannon divergence is defined in a general form in [182] for $\omega \in [0,1]$

$$\mathsf{JS}_\omega(P\|Q) = \omega\mathrm{D}_{\mathrm{KL}}\left(P\|\omega P + (1-\omega)Q\right) + (1-\omega)\mathrm{D}_{\mathrm{KL}}\left(Q\|\omega P + (1-\omega)Q\right). \tag{B.5}$$

- E_γ-divergence (also called hockey-stick divergence) [184, 218, 242, 248]: $f(x) = (x-\gamma)_+$ for

$\gamma \geq 1$ where $(a)_+ \triangleq \max\{a, 0\}$,

$$E_\gamma(P\|Q) = \int \left(\frac{\mathrm{d}P}{\mathrm{d}Q} - \gamma\right)_+ \mathrm{d}Q. \tag{B.6}$$

- DeGroot statistical information [73] of order p: $f(x) = \min\{p, 1-p\} - \min\{p, 1-px\}$ for $p \in (0,1)$,

$$\mathcal{I}_p(P\|Q) = \min\{p, 1-p\} - \int \min\left\{p, 1 - p\frac{\mathrm{d}P}{\mathrm{d}Q}\right\} \mathrm{d}Q. \tag{B.7}$$

- Marton's divergence [192]: $f(x) = (x-1)^2 \mathbb{I}_{[x<1]}$,

$$d_2(P\|Q)^2 = \inf \mathbb{E}\left[\Pr(X \neq Y \mid Y)^2\right] = \int \left(\frac{\mathrm{d}P}{\mathrm{d}Q} - 1\right)^2 \mathbb{I}_{[\frac{\mathrm{d}P}{\mathrm{d}Q}<1]} \mathrm{d}Q, \tag{B.8}$$

  where the infimum is taken over all couplings, i.e., joint distributions $P_{X,Y}$ which have marginals $P_X = P$ and $P_Y = Q$, respectively.

We refer the readers to [223, 242] for more examples of f-divergence and their properties.

B.2 Proofs for Section 4.5

B.2.1 Proof of Lemma 10

The proof of Lemma 10 relies on Ky Fan's min-max theorem [94]. As a reminder, a function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is said to be concave-like on $\mathcal{X}$ if, for any two elements $x_1, x_2 \in \mathcal{X}$ and $\lambda \in [0,1]$, there exists an element $x_0 \in \mathcal{X}$ such that for all $y \in \mathcal{Y}$

$$f(x_0, y) \geq \lambda f(x_1, y) + (1-\lambda) f(x_2, y).$$

Similarly, f is said to be convex-like on $\mathcal{Y}$, if for any two elements $y_1, y_2 \in \mathcal{Y}$ and $\lambda \in [0,1]$, there exists an element $y_0 \in \mathcal{Y}$ such that for all $x \in \mathcal{X}$

$$f(x, y_0) \leq \lambda f(x, y_1) + (1-\lambda) f(x, y_2).$$

A function $g : \mathcal{X} \to \mathbb{R}$ is called upper semicontinuous on a metric space $\mathcal{X}$ if for every point $x_0 \in \mathcal{X}$, $\limsup_{x \to x_0} g(x) \leq g(x_0)$. Next, we recall[1] Ky Fan's min-max theorem [94].

[1] We apply Ky Fan's min-max theorem to the function $-f$ instead of f.

Lemma 26 ([94, Theorem 2]). *Let $\mathcal{X}$ be a compact Hausdorff space and $\mathcal{Y}$ an arbitrary set (not topologized). Let f be a real-valued function on $\mathcal{X} \times \mathcal{Y}$ such that, for every $y \in \mathcal{Y}$, $f(\cdot, y)$ is upper semicontinuous on $\mathcal{X}$. If f is concave-like on $\mathcal{X}$ and convex-like on $\mathcal{Y}$, then*

$$\inf_{y\in\mathcal{Y}} \max_{x\in\mathcal{X}} f(x,y) = \max_{x\in\mathcal{X}} \inf_{y\in\mathcal{Y}} f(x,y). \tag{B.9}$$

Now we are in a position to prove Lemma 10.

Proof. We first introduce a (measurable) loss function $\ell : [0,1] \times [0,1] \to \mathbb{R}^+ \cup \{\infty\}$ and assume that this loss function satisfies: (i) $\ell(a,a) = 0$ for any $a \in [0,1]$ and (ii) for any $a \in [0,1]$, $\ell(a,\cdot)$ is convex and continuous. The benefit-of-splitting in Definition 6 can be written as

$$\inf_{h:\mathcal{X}\to[0,1]} \max_{s\in\{0,1\}} \mathbb{E}\left[\ell(y_s(X), h(X)) \mid S = s\right] - \max_{s\in\{0,1\}} \inf_{h:\mathcal{X}\to[0,1]} \mathbb{E}\left[\ell(y_s(X), h(X)) \mid S = s\right]. \tag{B.10}$$

By taking the ℓ_1 loss $\ell(a,b) = |a-b|$, ℓ_2 loss $\ell(a,b) = (a-b)^2$, and KL loss $\ell(a,b) = \mathrm{D}_{\mathrm{KL}}(a\|b) \triangleq a\log(a/b) + (1-a)\log((1-a)/(1-b))$, respectively, the above quantity becomes $\epsilon_{\mathsf{split},1}$, $\epsilon_{\mathsf{split},2}$, and $\epsilon_{\mathsf{split},\mathsf{KL}}$. These loss functions all satisfy our above two assumptions. In particular, by our first assumption, one can choose $h(x) = y_s(x)$ which leads to

$$\max_{s\in\{0,1\}} \inf_{h:\mathcal{X}\to[0,1]} \mathbb{E}\left[\ell(y_s(X), h(X)) \mid S = s\right] = 0. \tag{B.11}$$

Hence, the problem remains providing equivalent expression for the inf-max term

$$\begin{aligned}
&\inf_{h:\mathcal{X}\to[0,1]} \max_{s\in\{0,1\}} \mathbb{E}\left[\ell(y_s(X), h(X)) \mid S = s\right] \\
&= \inf_{h:\mathcal{X}\to[0,1]} \sup_{\omega\in[0,1]} \omega \cdot \mathbb{E}\left[\ell(y_0(X), h(X)) \mid S = 0\right] + (1-\omega)\cdot \mathbb{E}\left[\ell(y_1(X), h(X)) \mid S = 1\right].
\end{aligned} \tag{B.12}$$

Next, we use Ky Fan's min-max theorem [94] (see Lemma 26) to swap the positions of infimum and supremum in (B.12). We start with verifying the assumptions in Ky Fan's min-max theorem. We denote the set of all measurable functions from $\mathcal{X}$ to $[0,1]$ by $\mathcal{L}(\mathcal{X} \to [0,1])$ and introduce a function $F : [0,1] \times \mathcal{L}(\mathcal{X} \to [0,1]) \to \mathbb{R}$

$$F(\omega, h) \triangleq \omega \cdot \mathbb{E}\left[\ell(y_0(X), h(X)) \mid S = 0\right] + (1-\omega)\cdot \mathbb{E}\left[\ell(y_1(X), h(X)) \mid S = 1\right].$$

For every fixed $h \in \mathcal{L}(\mathcal{X} \to [0,1])$, $F(\cdot, h)$ is a linear function. Consequently, $F(\cdot, h)$ is upper semicontinuous and F is concave-like on $[0,1]$. Furthermore, for any $h_1, h_1 \in \mathcal{L}(\mathcal{X} \to [0,1])$,

$\lambda \in [0,1]$, and $\omega \in [0,1]$, we have $\lambda h_1 + (1-\lambda)h_2 \in \mathcal{L}(\mathcal{X} \to [0,1])$ and

$$F(\omega, \lambda h_1 + (1-\lambda)h_2) \leq \lambda F(\omega, h_1) + (1-\lambda)F(\omega, h_2)$$

by the convexity of $\ell(a, \cdot)$ for any $a \in [0,1]$. Hence, F is convex-like on $\mathcal{L}(\mathcal{X} \to [0,1])$. Therefore, by Ky Fan's min-max theorem, (B.12) is equal to

$$\sup_{\omega \in [0,1]} \inf_{h: \mathcal{X} \to [0,1]} \omega \cdot \mathbb{E}\left[\ell(y_0(X), h(X)) \mid S = 0\right] + (1-\omega) \cdot \mathbb{E}\left[\ell(y_1(X), h(X)) \mid S = 1\right]. \tag{B.13}$$

Now we take any probability distribution P over $\mathcal{X}$ such that P_0 and P_1 are absolutely continuous with respect to P. For example, one can simply choose $\mathrm{d}P = (\mathrm{d}P_0 + \mathrm{d}P_1)/2$. Then (B.13) can be written as

$$\sup_{\omega \in [0,1]} \inf_{h: \mathcal{X} \to [0,1]} \int \left(\omega \cdot \ell(y_0(x), h(x)) \tfrac{\mathrm{d}P_0(x)}{\mathrm{d}P(x)} + (1-\omega) \cdot \ell(y_1(x), h(x)) \tfrac{\mathrm{d}P_1(x)}{\mathrm{d}P(x)}\right) \mathrm{d}P(x). \tag{B.14}$$

Next, we prove that the infimum and the integer in (B.14) can be interchanged. For a fixed $\omega \in [0,1]$, we introduce a function $f : \mathcal{X} \times [0,1] \to \mathbb{R}$

$$f(x, \bar{h}) \triangleq \omega \cdot \ell(y_0(x), \bar{h}) \tfrac{\mathrm{d}P_0(x)}{\mathrm{d}P(x)} + (1-\omega) \cdot \ell(y_1(x), \bar{h}) \tfrac{\mathrm{d}P_1(x)}{\mathrm{d}P(x)}$$

and aim at proving

$$\inf_{h: \mathcal{X} \to [0,1]} \int f(x, h(x)) \mathrm{d}P(x) = \int \inf_{\bar{h} \in [0,1]} f(x, \bar{h}) \mathrm{d}P(x). \tag{B.15}$$

Since $f(\cdot, \bar{h})$ is measurable and $f(x, \cdot)$ is continuous, f is a Carathéodory function [see Section 4.10 in 6]. Hence, by the measurable maximum theorem [see Theorem 18.19 in 6], the mapping

$$x \to \inf_{\bar{h} \in [0,1]} f(x, \bar{h})$$

is measurable and the argmin correspondence (i.e., set-valued function)

$$\mathcal{H}^*(x) \triangleq \left\{\bar{h}^* \in [0,1] \mid f(x, \bar{h}^*) = \inf_{\bar{h} \in [0,1]} f(x, \bar{h})\right\}$$

is also measurable and admits a measurable selector. We denote this selector by $h^* : \mathcal{X} \to [0,1]$ and, by definition, it satisfies $h^*(x) \in \mathcal{H}^*(x)$ for all $x \in \mathcal{X}$. Now we are ready to prove (B.15). One direction LHS $\geq$ RHS can be obtained directly since for any $h : \mathcal{X} \to [0,1]$

$$\int f(x, h(x)) \mathrm{d}P(x) \geq \int \inf_{\bar{h} \in [0,1]} f(x, \bar{h}) \mathrm{d}P(x).$$

By the definition of $h^*(x)$,

$$\text{RHS} = \int f(x, h^*(x)) \mathrm{d}P(x) \geq \inf_{h:\mathcal{X}\to[0,1]} \int f(x, h(x)) \mathrm{d}P(x) = \text{LHS}.$$

Therefore, the equality in (B.15) holds and (B.14) becomes

$$\sup_{\omega\in[0,1]} \int \left(\omega \cdot \ell(y_0(x), h^*(x)) \frac{\mathrm{d}P_0(x)}{\mathrm{d}P(x)} + (1-\omega) \cdot \ell(y_1(x), h^*(x)) \frac{\mathrm{d}P_1(x)}{\mathrm{d}P(x)}\right) \mathrm{d}P(x). \tag{B.16}$$

Hence, our last step is to compute the function h^* for the loss functions of interest. If the loss function is ℓ_1, then

$$\operatorname*{argmin}_{\bar{h}\in[0,1]} f(x, \bar{h}) = \operatorname*{argmin}_{\bar{h}\in[0,1]} \left\{\omega \frac{\mathrm{d}P_0(x)}{\mathrm{d}P(x)} \cdot |\bar{h} - y_0(x)| + (1-\omega) \frac{\mathrm{d}P_1(x)}{\mathrm{d}P(x)} \cdot |\bar{h} - y_1(x)|\right\}.$$

For a fixed $\omega \in [0,1]$, the optimal classifier is

$$h^*(x) = \begin{cases} y_0(x) & \text{if } \frac{\mathrm{d}P_0(x)}{\mathrm{d}P_1(x)} \geq \frac{1-\omega}{\omega} \\ y_1(x) & \text{otherwise.} \end{cases}$$

By substituting the optimal classifier and ℓ_1 loss into (B.16), we get the desired equivalent expression of $\epsilon_{\text{split},1}$:

$$\sup_{\omega\in[0,1]} (1-\omega) \int_{\mathcal{A}_\omega} |y_1(x) - y_0(x)| \mathrm{d}P_1(x) + \omega \int_{\mathcal{A}_\omega^c} |y_1(x) - y_0(x)| \mathrm{d}P_0(x),$$

where $\mathcal{A}_\omega \triangleq \left\{x \mid \frac{\mathrm{d}P_0(x)}{\mathrm{d}P_1(x)} \geq \frac{1-\omega}{\omega}\right\}$. Similarly, when ℓ_2 loss is used, the optimal classifier becomes

$$h^*(x) = \frac{\omega y_0(x) \mathrm{d}P_0(x) + (1-\omega) y_1(x) \mathrm{d}P_1(x)}{\omega \mathrm{d}P_0(x) + (1-\omega) \mathrm{d}P_1(x)}, \tag{B.17}$$

which leads to the equivalent expression of $\epsilon_{\text{split},2}$:

$$\sup_{\omega\in[0,1]} \omega(1-\omega) \int \frac{(y_1(x) - y_0(x))^2 \mathrm{d}P_0(x) \mathrm{d}P_1(x)}{\omega \mathrm{d}P_0(x) + (1-\omega) \mathrm{d}P_1(x)}.$$

When the KL-loss is used, the optimal classifier h^* has expression in (B.17) as well. Consequently, we have the equivalent expression of $\epsilon_{\text{split,KL}}$:

$$\sup_{\omega\in[0,1]} \omega \mathbb{E}\left[\mathrm{D}_{\mathrm{KL}}(y_0(X) \| h^*(X)) \mid S = 0\right] + (1-\omega) \mathbb{E}\left[\mathrm{D}_{\mathrm{KL}}(y_1(X) \| h^*(X)) \mid S = 1\right]. \tag{B.18}$$

This expression can be further simplified by using the chain rule of KL-divergence:

$$\mathrm{D}_{\mathrm{KL}}(Q_{X,Y}\|R_{X,Y}) = \mathrm{D}_{\mathrm{KL}}(Q_{Y|X}\|R_{Y|X} \mid Q_X) + \mathrm{D}_{\mathrm{KL}}(Q_X\|R_X).$$

By taking $\mathrm{d}Q_X = \mathrm{d}P_s$, $\mathrm{d}R_X = w\mathrm{d}P_0 + (1-w)\mathrm{d}P_1$, $Q_{Y|X}(1|x) = y_s(x)$, and $R_{Y|X}(1|x) = h^*(x)$, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{D}_{\mathrm{KL}}(y_s(X)\|h^*(X)) \mid S = s\right] \\
&= \mathrm{D}_{\mathrm{KL}}\left(P_{X,Y|S=s}\|\omega P_{X,Y|S=0} + (1-\omega)P_{X,Y|S=1}\right) - \mathrm{D}_{\mathrm{KL}}\left(P_s\|\omega P_0 + (1-\omega)P_1\right).
\end{aligned}$$

Substituting this into (B.18) gives

$$\epsilon_{\mathsf{split,KL}} = \sup_{\omega\in[0,1]} \mathsf{JS}_\omega(P_{X,Y|S=0}\|P_{X,Y|S=1}) - \mathsf{JS}_\omega(P_0\|P_1),$$

where $\mathsf{JS}_\omega(\cdot\|\cdot)$ is the Jensen-Shannon divergence. □

B.2.2 Proof of Theorem 3

We divide the proof of Theorem 3 into three independent steps. First, we prove the upper bounds for $\epsilon_{\mathsf{split},1}$, $\epsilon_{\mathsf{split},2}$, and $\epsilon_{\mathsf{split,KL}}$ in a unified way. Then we prove the lower bounds for $\epsilon_{\mathsf{split},1}$ and $\epsilon_{\mathsf{split,KL}}$ using Lemma 10. Finally, we prove the lower bound for $\epsilon_{\mathsf{split},2}$ by leveraging the proof techniques of Brown-Low's two-points lower bound [45].

Proof. Note that (B.11) implies in the information-theoretic regime, optimal split classifiers can always achieve perfect performance. Specifically, one can select labeling functions y_0 and y_1 as split classifiers which have zero loss on each group. Hence, the problem remains upper bounding the performance of the optimal group-blind classifier. To achieve this goal, we consider two special group-blind classifiers:

$$h^*(x) = \frac{\mathrm{d}P_0(x)}{2\mathrm{d}P(x)}y_0(x) + \frac{\mathrm{d}P_1(x)}{2\mathrm{d}P(x)}y_1(x), \tag{B.19}$$

$$h^{**}(x) = \frac{1}{2}(y_0(x) + y_1(x)), \tag{B.20}$$

where $\mathrm{d}P = (\mathrm{d}P_0 + \mathrm{d}P_1)/2$. In what follows, we upper bound the performance of the group-blind classifiers in (B.19) and (B.20) and these bounds will be naturally translated into the upper bounds of $\epsilon_{\mathsf{split},1}$, $\epsilon_{\mathsf{split},2}$, and $\epsilon_{\mathsf{split,KL}}$, respectively.

We upper bound $\epsilon_{\text{split},1}$ by using the group-blind classifier h^* in (B.19).

$$\begin{aligned}\epsilon_{\text{split},1} = \inf_{h:\mathcal{X}\to[0,1]} \max_{s\in\{0,1\}} \mathbb{E}\left[|h(X) - y_s(X)| \mid S = s\right] &\le \max_{s\in\{0,1\}} \mathbb{E}\left[|h^*(X) - y_s(X)| \mid S = s\right] \\ &= \int |y_1(x) - y_0(x)| \tfrac{\mathrm{d}P_1(x)}{2\mathrm{d}P(x)} \mathrm{d}P_0(x). \end{aligned} \tag{B.21}$$

By the Cauchy-Schwarz inequality, we can further upper bound (B.21) by

$$\sqrt{\mathbb{E}\left[(y_1(X) - y_0(X))^2 \mid S = 0\right] \cdot \int \left(\tfrac{\mathrm{d}P_1(x)}{2\mathrm{d}P(x)}\right)^2 \mathrm{d}P_0(x)}. \tag{B.22}$$

Furthermore, we have

$$\int \left(\tfrac{\mathrm{d}P_1(x)}{2\mathrm{d}P(x)}\right)^2 \mathrm{d}P_0(x) = \frac{1}{4}\int \left(\tfrac{\mathrm{d}P_1(x)}{\mathrm{d}P(x)}\right)^2 \tfrac{\mathrm{d}P_0(x)}{\mathrm{d}P(x)} \mathrm{d}P(x) = \frac{1}{4}\int \left(\tfrac{\mathrm{d}P_1(x)}{\mathrm{d}P(x)}\right)^2 \left(2 - \tfrac{\mathrm{d}P_1(x)}{\mathrm{d}P(x)}\right) \mathrm{d}P(x). \tag{B.23}$$

Since $\frac{1}{4}x^2(2-x) \le 1 - |x-1|$ holds for any $x \ge 0$, the RHS of (B.23) can be upper bounded by

$$1 - \int \left|\tfrac{\mathrm{d}P_1(x)}{\mathrm{d}P(x)} - 1\right| \mathrm{d}P(x) = 1 - \frac{1}{2}\int |\mathrm{d}P_1(x) - \mathrm{d}P_0(x)| = 1 - \mathrm{D}_{\text{TV}}(P_0 \| P_1). \tag{B.24}$$

Combining (B.21–B.24) gives

$$\epsilon_{\text{split},1} \le \sqrt{\mathbb{E}\left[(y_1(X) - y_0(X))^2 \mid S = 0\right]} \cdot \sqrt{1 - \mathrm{D}_{\text{TV}}(P_0 \| P_1)}.$$

By symmetry, we can further tighten the upper bound

$$\epsilon_{\text{split},1} \le \min_{s\in\{0,1\}} \sqrt{\mathbb{E}\left[(y_1(X) - y_0(X))^2 \mid S = s\right]} \cdot \sqrt{1 - \mathrm{D}_{\text{TV}}(P_0 \| P_1)}.$$

On the other hand, using the classifier h^{**} in (B.20) leads to an alternative upper bound

$$\epsilon_{\text{split},1} \le \frac{1}{2} \max_{s\in\{0,1\}} \mathbb{E}\left[|y_1(X) - y_0(X)| \mid S = s\right].$$

Similarly, we can upper bound $\epsilon_{\text{split},2}$ by using the classifier h^* in (B.19)

$$\begin{aligned}\epsilon_{\text{split},2} &\le \max_{s\in\{0,1\}} \mathbb{E}\left[(h^*(X) - y_s(X))^2 \mid S = s\right] \\ &\le \int (y_1(x) - y_0(x))^2 \left(\tfrac{\mathrm{d}P_1(x)}{2\mathrm{d}P(x)}\right)^2 \mathrm{d}P_0(x) + \int (y_1(x) - y_0(x))^2 \left(\tfrac{\mathrm{d}P_0(x)}{2\mathrm{d}P(x)}\right)^2 \mathrm{d}P_1(x) \\ &= \int (y_1(x) - y_0(x))^2 \tfrac{\mathrm{d}P_1(x)}{2\mathrm{d}P(x)} \mathrm{d}P_0(x), \end{aligned} \tag{B.25}$$

where the second inequality uses the fact that $\max\{a, b\} \le a + b$. By the Cauchy-Schwarz inequality

and (B.23), (B.24), we can further upper bound (B.25) by

$$\sqrt{\mathbb{E}\left[(y_1(X)-y_0(X))^4 \mid S=0\right]}\cdot\sqrt{1-\mathrm{D}_{\mathrm{TV}}(P_0\|P_1)}.$$

By symmetry, we can further tighten this upper bound by replacing it with

$$\min_{s\in\{0,1\}}\sqrt{\mathbb{E}\left[(y_1(X)-y_0(X))^4 \mid S=s\right]}\cdot\sqrt{1-\mathrm{D}_{\mathrm{TV}}(P_0\|P_1)}.$$

On the other hand, using the classifier h^{**} in (B.20) leads to an alternative upper bound for $\epsilon_{\text{split},2}$.

We repeat the same strategy and upper bound $\epsilon_{\text{split,KL}}$ by using the classifier h^* in (B.19)

$$\begin{aligned}\epsilon_{\text{split,KL}} &\le \max_{s\in\{0,1\}}\mathbb{E}\left[\mathrm{D}_{\mathrm{KL}}(y_s(X)\|h^*(X)) \mid S=s\right]\\ &\le \mathbb{E}\left[\mathrm{D}_{\mathrm{KL}}(y_0(X)\|h^*(X)) \mid S=0\right]+\mathbb{E}\left[\mathrm{D}_{\mathrm{KL}}(y_1(X)\|h^*(X)) \mid S=1\right].\end{aligned}$$

Recall the chain rule of KL-divergence

$$\mathrm{D}_{\mathrm{KL}}(Q_{X,Y}\|R_{X,Y})=\mathrm{D}_{\mathrm{KL}}(Q_{Y|X}\|R_{Y|X}\mid Q_X)+\mathrm{D}_{\mathrm{KL}}(Q_X\|R_X).$$

By taking $\mathrm{d}Q_X=\mathrm{d}P_s$, $\mathrm{d}R_X=\mathrm{d}P$, $Q_{Y|X}(1|x)=y_s(x)$, and $R_{Y|X}(1|x)=h^*(x)$ and noticing the definition of h^* in (B.19), we obtain

$$\mathbb{E}\left[\mathrm{D}_{\mathrm{KL}}(y_s(X)\|h^*(X)) \mid S=s\right]=\mathrm{D}_{\mathrm{KL}}\left(P_{X,Y|S=s}\|\tfrac{P_{X,Y|S=0}+P_{X,Y|S=1}}{2}\right)-\mathrm{D}_{\mathrm{KL}}\left(P_s\|\tfrac{P_0+P_1}{2}\right). \tag{B.26}$$

Hence,

$$\epsilon_{\text{split,KL}}\le 2\mathsf{JS}(P_{X,Y|S=0}\|P_{X,Y|S=1})-2\mathsf{JS}(P_0\|P_1),$$

where $\mathsf{JS}(\cdot\|\cdot)$ is the Jensen-Shannon divergence. On the other hand, taking the classifier h^{**} in (B.20) gives an alternative upper bound for $\epsilon_{\text{split,KL}}$. □

We proceed to prove the lower bounds of $\epsilon_{\text{split},1}$ and $\epsilon_{\text{split,KL}}$.

Proof. Recall that $\mathcal{A}_{0.5} \triangleq \left\{x \in \mathcal{X} \mid \frac{\mathrm{d}P_0(x)}{\mathrm{d}P_1(x)} \geq 1\right\}$. By Lemma 10, we have

$$
\begin{aligned}
\epsilon_{\text{split},1} &\geq \frac{1}{2}\left(\int_{\mathcal{A}_{0.5}} |y_1(x) - y_0(x)| \mathrm{d}P_1(x) + \int_{\mathcal{A}_{0.5}^c} |y_1(x) - y_0(x)| \mathrm{d}P_0(x)\right) \\
&= \frac{1}{2}\left(\mathbb{E}\left[|y_1(X) - y_0(X)| \mid S = 1\right] - \int_{\mathcal{A}_{0.5}^c} |y_1(x) - y_0(x)| (\mathrm{d}P_1(x) - \mathrm{d}P_0(x))\right) \\
&= \frac{1}{2}\left(\mathbb{E}\left[|y_1(X) - y_0(X)| \mid S = 1\right] - \int |y_1(x) - y_0(x)| \left(1 - \frac{\mathrm{d}P_0(x)}{\mathrm{d}P_1(x)}\right)_+ \mathrm{d}P_1(x)\right) \\
&\geq \frac{1}{2}\left(\mathbb{E}\left[|y_1(X) - y_0(X)| \mid S = 1\right] - \sqrt{\int (y_1(x) - y_0(x))^2 \mathrm{d}P_1(x) \int \left(1 - \frac{\mathrm{d}P_0(x)}{\mathrm{d}P_1(x)}\right)^2 \mathbb{I}\left[\frac{\mathrm{d}P_0(x)}{\mathrm{d}P_1(x)} \leq 1\right] \mathrm{d}P_1(x)}\right) \\
&= \frac{1}{2}\left(\mathbb{E}\left[|y_1(X) - y_0(X)| \mid S = 1\right] - \sqrt{\mathbb{E}\left[(y_1(X) - y_0(X))^2 \mid S = 1\right]} \cdot d_2(P_0 \| P_1)\right),
\end{aligned}
$$

where $d_2(P_0\|P_1)$ is Marton's divergence. By symmetry, one can obtain

$$
\epsilon_{\text{split},1} \geq \frac{1}{2} \max_{s \in \{0,1\}} \left(\mathbb{E}\left[|y_1(X) - y_0(X)| \mid S = s\right] - \sqrt{\mathbb{E}\left[(y_1(X) - y_0(X))^2 \mid S = s\right]} \cdot d_2(P_{1-s} \| P_s)\right). \tag{B.27}
$$

Finally, the lower bound of $\epsilon_{\text{split,KL}}$ follows directly from Lemma 10. □

Before getting to the lower bound of $\epsilon_{\text{split},2}$, we prove a useful lemma. Here we denote

$$
A_s \triangleq \mathbb{E}\left[|y_1(X) - y_0(X)| \mid S = s\right] \quad \text{for } s \in \{0,1\}, \tag{B.28a}
$$

$$
B \triangleq \sqrt{\mathrm{D}_{\chi^2}(P_1 \| P_0) + 1}. \tag{B.28b}
$$

Lemma 27. *Assume that $A_0 \leq A_1$. For any measurable classifier $h : \mathcal{X} \to [0,1]$ and constant $0 \leq \epsilon < A_0^2/B^2$, if $\mathbb{E}\left[(h(X) - y_0(X))^2 \mid S = 0\right] \leq \epsilon$, then $\mathbb{E}\left[(h(X) - y_1(X))^2 \mid S = 1\right] \geq (A_1 - B\sqrt{\epsilon})^2$.*

Proof. Consider a convex optimization problem

$$
\begin{aligned}
\min_{h: \mathcal{X} \to [0,1]} \ & \int (h(x) - y_1(x))^2 \mathrm{d}P_1(x), \\
\text{s.t.} \ & \int (h(x) - y_0(x))^2 \mathrm{d}P_0(x) \leq \epsilon.
\end{aligned}
$$

Computing the Gateaux derivative of the Lagrange multiplier gives the following optimal conditions [166, Theorem 6.6.1],

$$
(h(x) - y_1(x))\mathrm{d}P_1(x) + \lambda(h(x) - y_0(x))\mathrm{d}P_0(x) = 0, \tag{B.29}
$$

$$
\lambda\left(\int (h(x) - y_0(x))^2 \mathrm{d}P_0(x) - \epsilon\right) = 0, \tag{B.30}
$$

$$
\lambda \geq 0, \tag{B.31}
$$

which provides the optimal classifier

$$h^*(x) = \frac{y_1(x)\mathrm{d}P_1(x) + \lambda y_0(x)\mathrm{d}P_0(x)}{\mathrm{d}P_1(x) + \lambda \mathrm{d}P_0(x)}.$$

We denote $r(x) \triangleq \frac{\mathrm{d}P_1(x)}{\mathrm{d}P_0(x)}$ and simplify the expression of the optimal classifier

$$h^*(x) = \frac{y_1(x)r(x) + \lambda y_0(x)}{r(x) + \lambda}. \tag{B.32}$$

If $\lambda = 0$, then $h^*(x) = y_1(x)$ and, consequently,

$$\mathbb{E}\left[(h^*(X) - y_0(X))^2 \mid S = 0\right] = \mathbb{E}\left[(y_1(X) - y_0(X))^2 \mid S = 0\right].$$

However, this contradicts our assumptions $\mathbb{E}\left[(h^*(X) - y_0(X))^2 \mid S = 0\right] \leq \epsilon$ and

$$\epsilon < \frac{\mathbb{E}\left[|y_1(X) - y_0(X)| \mid S = 0\right]^2}{\mathrm{D}_{\chi^2}(P_1 \| P_0) + 1} \leq \mathbb{E}\left[(y_1(X) - y_0(X))^2 \mid S = 0\right].$$

Hence, we have $\lambda > 0$. In this case, (B.30) and (B.32) imply

$$\int \left(\frac{y_1(x)r(x) + \lambda y_0(x)}{r(x) + \lambda} - y_0(x)\right)^2 \mathrm{d}P_0(x) = \epsilon.$$

We simplify the expression and obtain

$$\int r(x)^2 \left(\frac{y_1(x) - y_0(x)}{r(x) + \lambda}\right)^2 \mathrm{d}P_0(x) = \epsilon. \tag{B.33}$$

Now we consider lower bounding $\mathbb{E}\left[(h^*(X) - y_1(X))^2 \mid S = 1\right]$. By its definition and the expression of the optimal classifier (B.32), we have

$$\begin{aligned}
\mathbb{E}\left[(h^*(X) - y_1(X))^2 \mid S = 1\right] &= \int \left(\frac{y_1(x)r(x) + \lambda y_0(x)}{r(x) + \lambda} - y_1(x)\right)^2 \mathrm{d}P_1(x) \\
&= \int \left(\frac{\lambda(y_1(x) - y_0(x))}{r(x) + \lambda}\right)^2 \mathrm{d}P_1(x) \\
&\geq \left(\int \frac{\lambda|y_1(x) - y_0(x)|}{r(x) + \lambda} \mathrm{d}P_1(x)\right)^2 \\
&= \left(\int |y_1(x) - y_0(x)| \mathrm{d}P_1(x) - \int \frac{r(x)|y_1(x) - y_0(x)|}{r(x) + \lambda} \mathrm{d}P_1(x)\right)^2 \\
&= \left(\mathbb{E}\left[|y_1(X) - y_0(X)| \mid S = 1\right] - \int \frac{r(x)|y_1(x) - y_0(x)|}{r(x) + \lambda} \mathrm{d}P_1(x)\right)^2,
\end{aligned} \tag{B.34}$$

where the only inequality is due to the Cauchy-Schwarz inequality. Furthermore, by the Cauchy-

Schwarz inequality again and (B.33), we have

$$\int \frac{r(x)|y_1(x)-y_0(x)|}{r(x)+\lambda}\mathrm{d}P_1(x) \le \sqrt{\int r(x)^2\left(\frac{y_1(x)-y_0(x)}{r(x)+\lambda}\right)^2 \mathrm{d}P_0(x)\int r(x)\mathrm{d}P_1(x)}$$
$$= \sqrt{\epsilon \mathbb{E}\left[r(X) \mid S=1\right]}. \tag{B.35}$$

Recall that $r(x) = \frac{\mathrm{d}P_1(x)}{\mathrm{d}P_0(x)}$. Hence,

$$\mathbb{E}\left[r(X) \mid S=1\right] = \int \frac{\mathrm{d}P_1(x)}{\mathrm{d}P_0(x)}\mathrm{d}P_1(x) = \int\left[\left(\frac{\mathrm{d}P_1(x)}{\mathrm{d}P_0(x)}\right)^2 - 1\right]\mathrm{d}P_0(x) + 1 = \mathrm{D}_{\chi^2}(P_1\|P_0) + 1. \tag{B.36}$$

By our assumptions,

$$\mathbb{E}\left[|y_1(X)-y_0(X)| \mid S=1\right] - \sqrt{\epsilon(\mathrm{D}_{\chi^2}(P_1\|P_0)+1)}$$
$$\ge \mathbb{E}\left[|y_1(X)-y_0(X)| \mid S=1\right] - \mathbb{E}\left[|y_1(X)-y_0(X)| \mid S=0\right] \ge 0.$$

Combining (B.34), (B.35), and (B.36) together, we conclude that

$$\mathbb{E}\left[(h^*(X)-y_1(X))^2 \mid S=1\right] \ge \left(\mathbb{E}\left[|y_1(X)-y_0(X)| \mid S=1\right] - \sqrt{\epsilon\mathbb{E}\left[r(X) \mid S=1\right]}\right)^2$$
$$= \left(\mathbb{E}\left[|y_1(X)-y_0(X)| \mid S=1\right] - \sqrt{\epsilon\left(\mathrm{D}_{\chi^2}(P_1\|P_0\right)+1)}\right)^2.$$

□

Now we are in a position to prove the lower bound for $\epsilon_{\text{split},2}$.

Proof. By Lemma 27, for any classifier $h : \mathcal{X} \to [0,1]$ and

$$0 \le \epsilon < \frac{A_0^2}{B^2},$$

if $\mathbb{E}\left[(h(X)-y_0(X))^2 \mid S=0\right] \le \epsilon$, then $\mathbb{E}\left[(h(X)-y_1(X))^2 \mid S=1\right] \ge (A_1 - B\sqrt{\epsilon})^2$, where A_0, A_1, and B are defined in (B.28). Now we take $\epsilon = (A_0/(B+1))^2$, which naturally satisfies the assumptions in Lemma 27. As a result, if

$$\mathbb{E}\left[(h(X)-y_0(X))^2 \mid S=0\right] \le \left(\frac{A_0}{B+1}\right)^2,$$

then

$$\mathbb{E}\left[(h(X)-y_1(X))^2 \mid S=1\right] \ge \left(A_1 - B\frac{A_0}{B+1}\right)^2 \ge \left(\frac{A_0}{B+1}\right)^2$$

where the second inequality is because of the assumption $A_1 \ge A_0$. Consequently, for any $h : \mathcal{X} \to$

$[0,1]$,

$$\max_{s\in\{0,1\}} \mathbb{E}\left[(h(X)-y_s(X))^2 \mid S=s\right] \geq \left(\frac{A_0}{B+1}\right)^2 = \left(\frac{\mathbb{E}\left[|y_1(X)-y_0(X)| \mid S=0\right]}{\sqrt{\mathrm{D}_{\chi^2}(P_1\|P_0)+1}+1}\right)^2.$$

□

B.2.3 Proof of Proposition 5

Proof. First, note that $\inf_{\substack{h_s:\mathcal{X}\to[0,1]\\ \text{for } s\in\{0,1\}}} \mathbb{E}\left[|h_S(X)-y_S(X)|\right]=0$ as one can choose $h_s(x)=y_s(x)$. Hence, our focus is on upper and lower bounding

$$\begin{aligned}\epsilon_{\text{split,pop}} &= \inf_{h:\mathcal{X}\to[0,1]} \mathbb{E}\left[|h(X)-y_S(X)|\right] \\ &= \inf_{h:\mathcal{X}\to[0,1]} \Pr(S=0)\int |h(x)-y_0(x)|\mathrm{d}P_0(x) + \Pr(S=1)\int |h(x)-y_1(x)|\mathrm{d}P_1(x). \quad \text{(B.37)}\end{aligned}$$

By the proof of Lemma 10, the optimal classifier of (B.37) is

$$h^*(x)=\begin{cases} y_0(x) & \text{if } \frac{\Pr(S=0)\cdot \mathrm{d}P_0(x)}{\Pr(S=1)\cdot \mathrm{d}P_1(x)} \geq 1 \\ y_1(x) & \text{otherwise.}\end{cases}$$

By plugging the optimal classifier into (B.37), we can write $\epsilon_{\text{split,pop}}$ equivalently as

$$\Pr(S=0)\cdot\mathbb{E}\left[|y_1(X)-y_0(X)| \mid S=0\right] - \int |y_1(x)-y_0(x)|\left(\Pr(S=0)\cdot \mathrm{d}P_0(x) - \Pr(S=1)\cdot \mathrm{d}P_1(x)\right)_+.$$

The desired upper bound can be obtained by dropping the negative term. Now we proceed to prove the lower bound. Since

$$\begin{aligned}&\int |y_1(x)-y_0(x)|\left(\Pr(S=0)\cdot \mathrm{d}P_0(x) - \Pr(S=1)\cdot \mathrm{d}P_1(x)\right)_+ \\ &\leq \Pr(S=0)\int \left(\frac{\mathrm{d}P_0(x)}{\mathrm{d}P_1(x)} - \frac{\Pr(S=1)}{\Pr(S=0)}\right)_+ \mathrm{d}P_1(x) \\ &= \Pr(S=0)\cdot E_{\frac{\Pr(S=1)}{\Pr(S=0)}}(P_0\|P_1),\end{aligned}$$

where $E_{\frac{\Pr(S=1)}{\Pr(S=0)}}(P_0\|P_1)$ is the E_γ-divergence with $\gamma=\Pr(S=1)/\Pr(S=0)$, we have

$$\epsilon_{\text{split,pop}} \geq \Pr(S=0)\left(\mathbb{E}\left[|y_1(X)-y_0(X)| \mid S=0\right] - E_{\frac{\Pr(S=1)}{\Pr(S=0)}}(P_0\|P_1)\right).$$

□

B.3 Proofs for Section 4.6

B.3.1 Proof of Theorem 4

Recall that the false error rate is the maximum between false positive rate and false negative rate

$$\mathsf{FER}_s(h) \triangleq \max\left\{\mathbb{E}\left[h(X) \mid Y=0, S=s\right], \mathbb{E}\left[1-h(X) \mid Y=1, S=s\right]\right\}. \tag{B.38}$$

We prove the following lemma which will be used in the proof of Theorem 4.

Lemma 28. *The false error rate has the following equivalent expressions*

$$\begin{aligned}\mathsf{FER}_s(h) &= \max\left\{\frac{\mathbb{E}\left[h(X)(1-y_s(X)) \mid S=s\right]}{\Pr(Y=0 \mid S=s)}, 1-\frac{\mathbb{E}\left[h(X)y_s(X) \mid S=s\right]}{\Pr(Y=1 \mid S=s)}\right\} \\ &= \max\left\{\frac{\mathbb{E}\left[h(X)(1-y_s(X))f_s(X)\right]}{\Pr(Y=0 \mid S=s)}, 1-\frac{\mathbb{E}\left[h(X)y_s(X)f_s(X)\right]}{\Pr(Y=1 \mid S=s)}\right\}.\end{aligned}$$

where $f_s(x) \triangleq \frac{\Pr(S=s|X=x)}{\Pr(S=s)}$.

Proof. The proof follows directly from Bayes's rule,

$$\mathrm{d}P_{X|Y=0,S=s} = \frac{1-y_s(x)}{\Pr(Y=0 \mid S=s)}\mathrm{d}P_{X|S=s} = \frac{1-y_s(x)}{\Pr(Y=0 \mid S=s)} \cdot f_s(x)\mathrm{d}P_X.$$

□

Now we are in a position to prove Theorem 4.

Proof. By Lemma 28, the quantity $\inf_{h:\mathcal{X}\to[0,1]} \max_{s\in\{0,1\}} \mathsf{FER}_s(h)$ can be equivalently written as

$$\inf_{h:\mathcal{X}\to[0,1]} \max_{s\in\{0,1\}} \left\{\frac{\mathbb{E}\left[h(X)(1-y_s(X))f_s(X)\right]}{\Pr(Y=0 \mid S=s)}, 1-\frac{\mathbb{E}\left[h(X)y_s(X)f_s(X)\right]}{\Pr(Y=1 \mid S=s)}\right\} = \inf_{h:\mathcal{X}\to[0,1]} \max_{\boldsymbol{\mu}\in\Delta_4} G(\boldsymbol{\mu}, h), \tag{B.39}$$

where $\boldsymbol{\mu} \triangleq (\mu_{0,0}, \mu_{0,1}, \mu_{1,0}, \mu_{1,1})$ and

$$\begin{aligned}G(\boldsymbol{\mu}, h) &\triangleq \sum_{s\in\{0,1\}} \left(\mu_{s,0}\frac{\mathbb{E}\left[h(X)(1-y_s(X))f_s(X)\right]}{\Pr(Y=0 \mid S=s)} + \mu_{s,1}\left(1-\frac{\mathbb{E}\left[h(X)y_s(X)f_s(X)\right]}{\Pr(Y=1 \mid S=s)}\right)\right) \\ &= \sum_{s\in\{0,1\}} \mu_{s,1} + \mathbb{E}\left[\sum_{s\in\{0,1\}} \left(\frac{\mu_{s,0}(1-y_s(X))f_s(X)}{\Pr(Y=0 \mid S=s)} - \frac{\mu_{s,1}y_s(X)f_s(X)}{\Pr(Y=1 \mid S=s)}\right) h(X)\right].\end{aligned}$$

By denoting

$$\phi_{s,0}(x) \triangleq \frac{(1-y_s(x))f_s(x)}{\Pr(Y=0 \mid S=s)} \quad \phi_{s,1}(x) \triangleq \frac{-y_s(x)f_s(x)}{\Pr(Y=1 \mid S=s)},$$

we can write

$$G(\boldsymbol{\mu}, h) = \sum_{s\in\{0,1\}} \mu_{s,1} + \mathbb{E}\left[\sum_{s,i\in\{0,1\}} \mu_{s,i}\phi_{s,i}(X)h(X)\right].$$

We next use Ky Fan's min-max theorem [94] (see Lemma 26) to swap the positions of infimum and maximum. First, Δ_4 is a compact set and for any $h : \mathcal{X} \to [0,1]$, $G(\cdot, h)$ is continuous on Δ_4. Furthermore, for any $h : \mathcal{X} \to [0,1]$, $G(\cdot, h)$ is linear over Δ_4; for any $\boldsymbol{\mu} \in \Delta_4$, $G(\boldsymbol{\mu}, \cdot)$ is convex-like over all (measurable) classifiers from $\mathcal{X}$ to $[0,1]$. Hence, we have

$$\inf_{h:\mathcal{X}\to[0,1]} \max_{\boldsymbol{\mu}\in\Delta_4} G(\boldsymbol{\mu}, h) = \max_{\boldsymbol{\mu}\in\Delta_4} \inf_{h:\mathcal{X}\to[0,1]} G(\boldsymbol{\mu}, h). \tag{B.40}$$

Next, we prove that, for any fixed $\boldsymbol{\mu} \in \Delta_4$,

$$\inf_{h:\mathcal{X}\to[0,1]} \mathbb{E}\left[\sum_{s,i\in\{0,1\}} \mu_{s,i}\phi_{s,i}(X)h(X)\right] = \mathbb{E}\left[\inf_{h:\mathcal{X}\to[0,1]} \sum_{s,i\in\{0,1\}} \mu_{s,i}\phi_{s,i}(X)h(X)\right]. \tag{B.41}$$

One direction LHS $\geq$ RHS can be obtained directly since for any $h : \mathcal{X} \to [0,1]$

$$\mathbb{E}\left[\sum_{s,i\in\{0,1\}} \mu_{s,i}\phi_{s,i}(X)h(X)\right] \geq \mathbb{E}\left[\inf_{h:\mathcal{X}\to[0,1]} \sum_{s,i\in\{0,1\}} \mu_{s,i}\phi_{s,i}(X)h(X)\right].$$

Note that the infimum in the RHS of (B.41) is point-wise. For any fixed $x \in \mathcal{X}$, the following optimization problem

$$\inf_{\bar{h}\in[0,1]} \sum_{s,i\in\{0,1\}} \mu_{s,i}\phi_{s,i}(x)\bar{h}$$

has an optimal solution $\bar{h}^* = \mathbb{I}[\sum_{s,i\in\{0,1\}} \mu_{s,i}\phi_{s,i}(x) \leq 0]$. Hence, there is a measurable classifier which can achieve the point-wise infimum inside the expectation of the RHS in (B.41): $h^*(x) = \mathbb{I}[\sum_{s,i\in\{0,1\}} \mu_{s,i}\phi_{s,i}(x) \leq 0]$. Consequently, the RHS of (B.41) can be simplified as

$$\text{RHS} = \mathbb{E}\left[\left(\sum_{s,i\in\{0,1\}} \mu_{s,i}\phi_{s,i}(X)\right)_-\right], \tag{B.42}$$

where for $a \in \mathbb{R}$, $(a)_- \triangleq \min\{a, 0\}$. Since the LHS of (B.41) is an infimum over all measurable classifiers, using the classifier h^* leads to

$$\text{LHS} \leq \mathbb{E}\left[\sum_{s,i\in\{0,1\}} \mu_{s,i}\phi_{s,i}(X)h^*(X)\right] = \mathbb{E}\left[\left(\sum_{s,i\in\{0,1\}} \mu_{s,i}\phi_{s,i}(X)\right)_-\right] = \text{RHS}.$$

Combining (B.39–B.42) together implies

$$\inf_{h:\mathcal{X}\to[0,1]} \max_{s\in\{0,1\}} \mathsf{FER}_s(h) = \max_{\boldsymbol{\mu}\in\Delta_4} \left\{ \sum_{s\in\{0,1\}} \mu_{s,1} + \mathbb{E}\left[\left(\sum_{s,i\in\{0,1\}} \mu_{s,i}\phi_{s,i}(X) \right)_- \right] \right\}.$$

Similarly, one can prove that

$$\max_{s\in\{0,1\}} \inf_{h:\mathcal{X}\to[0,1]} \mathsf{FER}_s(h) = \max_{s\in\{0,1\}} \max_{\boldsymbol{\nu}^{(s)}\in\Delta_2} \left\{ \nu_1^{(s)} + \mathbb{E}\left[\left(\sum_{i\in\{0,1\}} \nu_i^{(s)}\phi_{s,i}(X) \right)_- \right] \right\}.$$

□

B.3.2 Proof of Proposition 6

We start with a useful lemma which will be used in the proof of Proposition 6.

Lemma 29. *Let $f : \mathcal{X} \times \mathbb{R}^k \to \mathbb{R}$ be a bounded measurable function. For a fixed $x \in \mathcal{X}$, if $v(x, w_0) \in \mathbb{R}^k$ is a supergradient of $f(x,\cdot)$ at w_0:*

$$f(x,w) - f(x,w_0) \le v(x,w_0)^T(w - w_0), \tag{B.43}$$

then $\mathbb{E}[v(X,w_0)]$ is a supergradient of $\mathbb{E}[f(X,\cdot)]$ at w_0:

$$\mathbb{E}[f(X,w)] - \mathbb{E}[f(X,w_0)] \le \mathbb{E}[v(X,w_0)]^T(w - w_0). \tag{B.44}$$

The proof of Lemma 29 follows directly by taking expectation on both sides of (B.43). We refer the readers to [232] for a more general result on the interchangeability of subdifferentiation and (conditional) expectation. Now we are in a position to prove Proposition 6.

Proof. Consider a function $g(x,\boldsymbol{\mu}) \triangleq \sum_{s\in\{0,1\}} \mu_{s,1} + \left(\sum_{s,i\in\{0,1\}} \mu_{s,i}\phi_{s,i}(x)\right)_-$. For a fixed x, $g(x,\cdot)$ has a supergradient at $\boldsymbol{\mu} \triangleq (\mu_{0,0}, \mu_{0,1}, \mu_{1,0}, \mu_{1,1})$:

$$\left(i + \phi_{s,i}(x)\mathbb{I}\Big[\sum_{s',i'\in\{0,1\}} \mu_{s',i'}\phi_{s',i'}(x) < 0 \Big] \right)_{s,i\in\{0,1\}}.$$

Therefore, by Lemma 29, g has a supergradient at $\boldsymbol{\mu}$:

$$\partial g(\boldsymbol{\mu}) \ni \left(i + \mathbb{E}\left[\phi_{s,i}(X) \cdot \mathbb{I}\Big[\sum_{s',i'\in\{0,1\}} \mu_{s',i'}\phi_{s',i'}(X) < 0 \Big] \right] \right)_{s,i\in\{0,1\}}.$$

Now we introduce auxiliary functions

$$\psi_{s,i}(x) \triangleq \frac{1 - i - y_s(x)}{\Pr(Y = i \mid S = s)}, \quad s, i \in \{0, 1\}.$$

By Bayes's rule and the definition of $\phi_{s,i}$ (see (4.5)), we have

$$\psi_{s,i}(x) \cdot \mathrm{d}P_{X|S=s}(x) = \phi_{s,i}(x) \cdot \mathrm{d}P_X(x).$$

Hence,

$$\partial g(\boldsymbol{\mu}) \ni \left(i + \mathbb{E}\left[\psi_{s,i}(X) \cdot \mathbb{I}\Big[\sum_{s',i' \in \{0,1\}} \mu_{s',i'} \phi_{s',i'}(X) < 0 \Big] \,\middle|\, S = s \right] \right)_{s,i \in \{0,1\}}.$$

Similarly, one can obtain a closed-form supergradient of $g_s(\boldsymbol{\nu})$. □

B.4 Proofs for Section 4.7

B.4.1 Proof of Theorem 5

We first recall a useful lemma which can be proved by the variational representation [207] of total variation distance.

Lemma 30. *For any measurable and non-negative function $f : \mathcal{X} \to \mathbb{R}^+$,*

$$|\mathbb{E}[f(X_0)] - \mathbb{E}[f(X_1)]| \leq \|f\|_\infty \mathrm{D}_{\mathrm{TV}}(P_0 \| P_1),$$

where $X_0 \sim P_0$ and $X_1 \sim P_1$.

Now we are in a position to prove Theorem 5.

Proof. We prove the upper bound first. Let h_s^* be an optimal classifier for the group $s \in \{0, 1\}$, i.e., $h_s^* \in \operatorname{argmin}_{h \in \mathcal{H}} \mathbb{E}[|h(X) - y_s(X)| \mid S = s]$. Then

$$\inf_{h \in \mathcal{H}} \max_{s \in \{0,1\}} \mathbb{E}[|h(X) - y_s(X)| \mid S = s] \leq \max\{\mathbb{E}[|h_0^*(X) - y_1(X)| \mid S = 1], \mathbb{E}[|h_0^*(X) - y_0(X)| \mid S = 0]\}.$$

By the triangle inequality,

$$\mathbb{E}[|h_0^*(X) - y_1(X)| \mid S = 1] \leq \mathbb{E}[|h_1^*(X) - h_0^*(X)| \mid S = 1] + \mathbb{E}[|h_1^*(X) - y_1(X)| \mid S = 1].$$

Therefore,

$$\inf_{h\in\mathcal{H}}\max_{s\in\{0,1\}}\mathbb{E}\left[|h(X)-y_s(X)|\mid S=s\right]\le\mathbb{E}\left[|h_1^*(X)-h_0^*(X)|\mid S=1\right]+\max_{s\in\{0,1\}}\mathbb{E}\left[|h_s^*(X)-y_s(X)|\mid S=s\right],$$

which implies that

$$\begin{aligned}\epsilon_{\text{split}}^{\mathcal{H}}&=\inf_{h\in\mathcal{H}}\max_{s\in\{0,1\}}\mathbb{E}\left[|h(X)-y_s(X)|\mid S=s\right]-\max_{s\in\{0,1\}}\inf_{h\in\mathcal{H}}\mathbb{E}\left[|h(X)-y_s(X)|\mid S=s\right]\\&\le\mathbb{E}\left[|h_1^*(X)-h_0^*(X)|\mid S=1\right].\end{aligned}$$

By symmetry, we obtain the desired upper bound for $\epsilon_{\text{split}}^{\mathcal{H}}$. Now we proceed to prove the lower bound for $\epsilon_{\text{split}}^{\mathcal{H}}$. By the triangle inequality, we have

$$\begin{aligned}\mathbb{E}\left[|y_1(X)-y_0(X)|\mid S=0\right]\ge\mathbb{E}\left[|h_1^*(X)-h_0^*(X)|\mid S=0\right]-\mathbb{E}\left[|h_0^*(X)-y_0(X)|\mid S=0\right]\\-\mathbb{E}\left[|h_1^*(X)-y_1(X)|\mid S=0\right].\end{aligned}$$

By Lemma 30,

$$\mathbb{E}\left[|h_1^*(X)-y_1(X)|\mid S=0\right]\le\mathbb{E}\left[|h_1^*(X)-y_1(X)|\mid S=1\right]+\mathrm{D}_{\text{TV}}(P_0\|P_1).$$

Therefore,

$$\mathbb{E}\left[|y_1(X)-y_0(X)|\mid S=0\right]\ge\mathbb{E}\left[|h_1^*(X)-h_0^*(X)|\mid S=0\right]-2\lambda-\mathrm{D}_{\text{TV}}(P_0\|P_1).$$

where $\lambda\triangleq\sum_{s\in\{0,1\}}\mathbb{E}\left[|h_s^*(X)-y_s(X)|\mid S=s\right]/2$. Hence,

$$\max_{s\in\{0,1\}}\mathbb{E}\left[|y_1(X)-y_0(X)|\mid S=s\right]\ge\max_{s\in\{0,1\}}\mathbb{E}\left[|h_1^*(X)-h_0^*(X)|\mid S=s\right]-2\lambda-\mathrm{D}_{\text{TV}}(P_0\|P_1).\quad\text{(B.45)}$$

By a slight modification of the proof of Theorem 3, we have

$$\inf_{h:\mathcal{X}\to[0,1]}\max_{s\in\{0,1\}}\mathbb{E}\left[|h(X)-y_s(X)|\mid S=s\right]\ge\frac{1}{2}\left(\max_{s\in\{0,1\}}\mathbb{E}\left[|y_1(X)-y_0(X)|\mid S=s\right]-\mathrm{D}_{\text{TV}}(P_0\|P_1)\right).\quad\text{(B.46)}$$

Substituting (B.45) into (B.46) leads to

$$\begin{aligned}\inf_{h\in\mathcal{H}}\max_{s\in\{0,1\}}\mathbb{E}\left[|h(X)-y_s(X)|\mid S=s\right]&\ge\inf_{h:\mathcal{X}\to[0,1]}\max_{s\in\{0,1\}}\mathbb{E}\left[|h(X)-y_s(X)|\mid S=s\right]\\&\ge\frac{1}{2}\left(\max_{s\in\{0,1\}}\mathbb{E}\left[|h_1^*(X)-h_0^*(X)|\mid S=s\right]-2\lambda-2\mathrm{D}_{\text{TV}}(P_0\|P_1)\right).\end{aligned}\quad\text{(B.47)}$$

Finally, since $\max\{a,b\} \le a+b$ and $\{h_s^*\}_{s\in\{0,1\}}$ is the set of optimal split classifiers, then

$$\max_{s\in\{0,1\}} \inf_{h\in\mathcal{H}} \mathbb{E}\left[|h(X)-y_s(X)| \mid S=s\right] = \max_{s\in\{0,1\}} \mathbb{E}\left[|h_s^*(X)-y_s(X)| \mid S=s\right] \le 2\lambda. \tag{B.48}$$

Combining (B.47) with (B.48) gives

$$\epsilon_{\text{split}}^{\mathcal{H}} \ge \frac{1}{2} \max_{s\in\{0,1\}} \mathbb{E}\left[|h_1^*(X)-h_0^*(X)| \mid S=s\right] - \mathrm{D}_{\text{TV}}(P_0\|P_1) - 3\lambda.$$

□

B.4.2 Proof of Proposition 7

Proof. By the triangle inequality, $\inf_{h\in\mathcal{H}} \max_{s\in\{0,1\}} \mathbb{E}\left[|h(X)-y_s(X)| \mid S=s\right] \le \mathrm{I}+\mathrm{II}$ where

$$\mathrm{I} \triangleq \inf_{h\in\mathcal{H}} \max_{s\in\{0,1\}} \mathbb{E}\left[|h(X)-h^*(X)| \mid S=s\right],$$

$$\mathrm{II} \triangleq \max_{s\in\{0,1\}} \mathbb{E}\left[|h^*(X)-y_s(X)| \mid S=s\right],$$

and h^* is defined in (B.19). Since $\max\{a,b\} \le a+b$, we have $\mathrm{I} \le 2\inf_{h\in\mathcal{H}} \mathbb{E}\left[|h(\bar{X})-h^*(\bar{X})|\right]$ where the random variable $\bar{X}$ follows the probability distribution $(P_0+P_1)/2$. By Barron's approximation bounds [25],

$$\inf_{h\in\mathcal{H}} \mathbb{E}\left[|h(\bar{X})-h^*(\bar{X})|\right] \le \frac{\mathsf{diam}(\mathcal{X})C}{\sqrt{k}}, \tag{B.49}$$

where $C \triangleq \int_{\mathbb{R}^d} \|w\|_2 |\widetilde{h^*}(w)| \mathrm{d}w$, $\widetilde{h^*}(w) \triangleq \frac{1}{(2\pi)^d} \int_{\mathcal{X}} h^*(x) \exp(-iwx) \mathrm{d}x$. Moreover, by the proof of Theorem 3 (see Appendix B.2.2), we have

$$\mathrm{II} \le \min_{s\in\{0,1\}} \sqrt{\mathbb{E}\left[(y_1(X)-y_0(X))^2 \mid S=s\right]} \cdot \sqrt{1-\mathrm{D}_{\text{TV}}(P_0\|P_1)}.$$

To summarize, if the hypothesis class contains feedforward neural network models with one layer of sigmoidal functions, the $\mathcal{H}$-benefit-of-splitting has an upper bound below.

$$\begin{aligned}\epsilon_{\text{split}}^{\mathcal{H}} &= \inf_{h\in\mathcal{H}} \max_{s\in\{0,1\}} \mathbb{E}\left[|h(X)-y_s(X)| \mid S=s\right] - \max_{s\in\{0,1\}} \inf_{h\in\mathcal{H}} \mathbb{E}\left[|h(X)-y_s(X)| \mid S=s\right] \\ &\le \min_{s\in\{0,1\}} \sqrt{\mathbb{E}\left[(y_1(X)-y_0(X))^2 \mid S=s\right]} \cdot \sqrt{1-\mathrm{D}_{\text{TV}}(P_0\|P_1)} + \frac{2\mathsf{diam}(\mathcal{X})C}{\sqrt{k}}.\end{aligned}$$

□

B.4.3 Proof of Corollary 2

We approach Corollary 2 by proving a more general result.

Lemma 31. *For any hypothesis $\mathcal{H}$, there exists a probability distribution $Q_{S,X,Y}$ whose $\mathcal{H}$-benefit-of-splitting is at least*

$$\frac{1}{2}\sup_{\substack{h_1,h_0\in\mathcal{H}\\ x\in\mathcal{X}}}|h_1(x)-h_0(x)|.$$

Proof. For any $\epsilon>0$, there exist two classifiers $h_1^*,h_0^*\in\mathcal{H}$ and $x^*\in\mathcal{X}$ such that

$$|h_1^*(x^*)-h_0^*(x^*)|\geq\sup_{\substack{h_1,h_0\in\mathcal{H}\\ x\in\mathcal{X}}}|h_1(x)-h_0(x)|-\epsilon. \tag{B.50}$$

Now we construct a probability distribution $Q_{S,X,Y}$ with $Q_{Y|X,S}(1|x,s)=h_s^*(x)$, $Q_{X|S=s}(x)=\delta(x-x^*)$, $Q_S(s)=0.5$ for $s\in\{0,1\}$ where $\delta(\cdot)$ is the Dirac delta function. Our lower bound in Theorem 5 implies that $\epsilon_{\text{split}}^{\mathcal{H}}\geq\frac{1}{2}|h_1^*(x^*)-h_0^*(x^*)|$ which, due to (B.50), can be further lower bounded by $\frac{1}{2}(\sup_{h_1,h_0\in\mathcal{H},x\in\mathcal{X}}|h_1(x)-h_0(x)|-\epsilon)$. Since this lower bound of $\epsilon_{\text{split}}^{\mathcal{H}}$ holds for any $\epsilon>0$, one can let ϵ be sufficiently small which leads to the desired conclusion. □

B.4.4 Proof of Theorem 6

We introduce the empirical benefit-of-splitting and bound its difference from the sample-limited-splitting (see Definition 11).

Definition 27. For a given hypothesis class $\mathcal{H}$ and n_s i.i.d. samples $\{(x_{s,i},y_{s,i})\}_{i=1}^{n_s}$ from group $s\in\{0,1\}$, the empirical-splitting is defined as

$$\hat{\epsilon}_{\text{split,emp}}\triangleq\inf_{h\in\mathcal{H}}\max_{s\in\{0,1\}}\frac{\sum_{i=1}^{n_s}|h(x_{s,i})-y_{s,i}|}{n_s}-\max_{s\in\{0,1\}}\inf_{h\in\mathcal{H}}\frac{\sum_{i=1}^{n_s}|h(x_{s,i})-y_{s,i}|}{n_s}. \tag{B.51}$$

Lemma 32. *Let $\mathcal{H}$ be a hypothesis class from $\mathcal{X}$ to $\{0,1\}$ with VC dimension D. Then with probability at least $1-\delta$,*

$$\left|\hat{\epsilon}_{split}-\hat{\epsilon}_{split,emp}\right|\leq 4\max_{s\in\{0,1\}}\sqrt{\frac{2D\log(6n_s)+2\log(16/\delta)}{n_s}}, \tag{B.52}$$

where n_s is the number of samples from group $s\in\{0,1\}$.

Proof. Corollary 3.8 and Theorem 4.3 in [11] together imply that with probability at least $1-\delta$, for

any $s \in \{0,1\}$ and $h \in \mathcal{H}$,

$$\left| \frac{\sum_{i=1}^{n_s} |h(x_{s,i}) - y_{s,i}|}{n_s} - \mathbb{E}\left[|h(X) - y_s(X)| \mid S = s\right] \right| \leq 2\sqrt{\frac{2D\log(6n_s) + 2\log(8/\delta)}{n_s}}.$$

Therefore, for any $h_s \in \mathcal{H}$ with $s \in \{0,1\}$

$$\left| \max_{s\in\{0,1\}} \frac{\sum_{i=1}^{n_s} |h_s(x_{s,i}) - y_{s,i}|}{n_s} - \max_{s\in\{0,1\}} \mathbb{E}\left[|h_s(X) - y_s(X)| \mid S = s\right] \right| \leq 2 \max_{s\in\{0,1\}} \sqrt{\frac{2D\log(6n_s) + 2\log(8/\delta)}{n_s}}. \tag{B.53}$$

Recall that

$$\hat{\epsilon}_{\text{split}} = \max_{s\in\{0,1\}} \mathbb{E}\left[|\hat{h}^*(X) - y_s(X)| \mid S = s\right] - \max_{s\in\{0,1\}} \mathbb{E}\left[|\hat{h}_s^*(X) - y_s(X)| \mid S = s\right].$$

Now by (B.53), we conclude that

$$\left|\hat{\epsilon}_{\text{split}} - \hat{\epsilon}_{\text{split,emp}}\right| \leq 4 \max_{s\in\{0,1\}} \sqrt{\frac{2D\log(6n_s) + 2\log(8/\delta)}{n_s}}.$$

□

Since the upper and lower bounds of $\epsilon_{\text{split}}^{\mathcal{H}}$ (see Theorem 5) hold for any underlying distribution $P_{S,X,Y}$. One can plug in the empirical distribution and obtain the corresponding bounds for $\hat{\epsilon}_{\text{split,emp}}$. Then we obtain the desired bounds for $\hat{\epsilon}_{\text{split}}$ by using Lemma 32 for bounding the difference between $\hat{\epsilon}_{\text{split,emp}}$ and $\hat{\epsilon}_{\text{split}}$.

B.5 Proofs for Section 4.8

B.5.1 Proof of Proposition 9

Proof. First, we define

$$\Delta(f) \triangleq \lim_{\epsilon\to 0} \frac{\mathsf{M}(\widetilde{P}_0) - \mathsf{M}(P_0)}{\epsilon}, \tag{B.54}$$

where $\widetilde{P}_0(\boldsymbol{x})$ is the perturbed distribution defined in (4.20). Then we prove that

$$\Delta(f) = \mathbb{E}\left[f(X)\psi(X)|S = 0\right].$$

Note that an alternative way [see e.g., 129] to define influence functions is in terms of the Gâteaux derivative:

$$\sum_{\boldsymbol{x}\in\mathcal{X}} \psi(\boldsymbol{x})P_0(\boldsymbol{x}) = 0,$$

and

$$\lim_{\epsilon\to 0}\frac{1}{\epsilon}\left(\mathsf{M}\left((1-\epsilon)P_0+\epsilon Q\right)-\mathsf{M}\left(P_0\right)\right) = \sum_{\boldsymbol{x}\in\mathcal{X}} \psi(\boldsymbol{x})Q(\boldsymbol{x}),\ \forall Q\in\mathcal{P}.$$

In particular, we can choose $Q(\boldsymbol{x}) = \left(\frac{1}{M_U}f(\boldsymbol{x})+1\right)P_0(\boldsymbol{x})$, where $M_U \triangleq \sup\{|f(\boldsymbol{x})| \mid \boldsymbol{x}\in\mathcal{X}\}+1$. Then

$$(1-\epsilon)P_0(\boldsymbol{x})+\epsilon Q(\boldsymbol{x}) = P_0(\boldsymbol{x}) + \frac{\epsilon}{M_U}f(\boldsymbol{x})P_0(\boldsymbol{x}).$$

For simplicity, we use $P_0+\epsilon f P_0$ and $P_0+\frac{\epsilon}{M_U}fP_0$ to represent $P_0(\boldsymbol{x})+\epsilon f(\boldsymbol{x})P_0(\boldsymbol{x})$ and $P_0(\boldsymbol{x})+\frac{\epsilon}{M_U}f(\boldsymbol{x})P_0(\boldsymbol{x})$, respectively. Then

$$\begin{aligned}
\Delta(f) &= \lim_{\epsilon\to 0}\frac{1}{\epsilon}\left(\mathsf{M}(P_0+\epsilon f P_0)-\mathsf{M}(P_0)\right)\\
&= \lim_{\epsilon\to 0}\frac{M_U}{\epsilon}\left(\mathsf{M}\left(P_0+\frac{\epsilon}{M_U}fP_0\right)-\mathsf{M}(P_0)\right)\\
&= M_U\lim_{\epsilon\to 0}\frac{1}{\epsilon}\left(\mathsf{M}((1-\epsilon)P_0+\epsilon Q)-\mathsf{M}(P_0)\right)\\
&= M_U\sum_{\boldsymbol{x}\in\mathcal{X}}\psi(\boldsymbol{x})Q(\boldsymbol{x})\\
&= M_U\sum_{\boldsymbol{x}\in\mathcal{X}}\psi(\boldsymbol{x})\left(\frac{1}{M_U}f(\boldsymbol{x})+1\right)P_0(\boldsymbol{x})\\
&= \sum_{\boldsymbol{x}\in\mathcal{X}}\psi(\boldsymbol{x})f(\boldsymbol{x})P_0(\boldsymbol{x})\\
&= \mathbb{E}\left[f(X)\psi(X)|S=0\right].
\end{aligned}$$

Following from Cauchy-Schwarz inequality,

$$\mathbb{E}\left[f(X)\psi(X)|S=0\right] \geq -\sqrt{\mathbb{E}\left[f(X)^2|S=0\right]}\sqrt{\mathbb{E}\left[\psi(X)^2|S=0\right]} = -\sqrt{\mathbb{E}\left[\psi(X)^2|S=0\right]}.$$

Here the equality can be achieved by choosing

$$f(\boldsymbol{x}) = \frac{-\psi(\boldsymbol{x})}{\sqrt{\mathbb{E}\left[\psi(X)^2|S=0\right]}}.$$

□

B.5.2 Proof of Proposition 10

Proof. When the disparity metric is a linear combination of K different disparity metrics:

$$\mathsf{M}(P_0) = \sum_{i=1}^{K} \lambda_i \mathsf{M}_i(P_0),$$

the influence function, following from Definition 14, is

$$\psi(\boldsymbol{x}) = \lim_{\epsilon \to 0} \frac{\mathsf{M}\left((1-\epsilon)P_0 + \epsilon\delta_{\boldsymbol{x}}\right) - \mathsf{M}(P_0)}{\epsilon} \tag{B.55}$$

$$= \sum_{i=1}^{K} \lambda_i \lim_{\epsilon \to 0} \frac{\mathsf{M}_i\left((1-\epsilon)P_0 + \epsilon\delta_{\boldsymbol{x}}\right) - \mathsf{M}_i(P_0)}{\epsilon} \tag{B.56}$$

$$= \sum_{i=1}^{K} \lambda_i \psi_i(\boldsymbol{x}). \tag{B.57}$$

□

B.5.3 Proofs of Proposition 11

We first show that the disparity metrics in Table 4.1 can be expressed in terms of P_0 when $P_{\hat{Y}|X}$, $P_{Y|X,S}$, P_1, and P_S are given. To start with, we express $P_{S,X,Y,\hat{Y}}$ as:

$$P_{S,X,Y,\hat{Y}} = P_{\hat{Y}|X} P_{Y|X,S} P_S P_{X|S}. \tag{B.58}$$

Note that $h(\boldsymbol{x}) = P_{\hat{Y}|X}(1|\boldsymbol{x})$. In what follows, we use these observations to express each of the disparity metrics in Table 1 as $\mathsf{M}(P_0)$ (i.e., a function of P_0).

1. **SP.**

$$\begin{aligned}\Pr(\hat{Y}=0|S=0) - \Pr(\hat{Y}=0|S=1) &= \mathbb{E}\left[(1-h(X))|S=0\right] - \mathbb{E}\left[(1-h(X))|S=1\right] \\ &= -\sum_{\boldsymbol{x}\in\mathcal{X}} h(\boldsymbol{x}) P_0(\boldsymbol{x}) + \sum_{\boldsymbol{x}\in\mathcal{X}} h(\boldsymbol{x}) P_1(\boldsymbol{x}).\end{aligned} \tag{B.59}$$

2. **FDR.**

$$\begin{aligned}&\Pr(Y=0|\hat{Y}=1,S=0) - \Pr(Y=0|\hat{Y}=1,S=1) \\ &= \frac{\Pr(Y=0,\hat{Y}=1,S=0)}{\Pr(\hat{Y}=1,S=0)} - \frac{\Pr(Y=0,\hat{Y}=1,S=1)}{\Pr(\hat{Y}=1,S=1)} \\ &= \frac{\sum_{\boldsymbol{x}\in\mathcal{X}} P_{\hat{Y}|X}(1|\boldsymbol{x}) P_{Y|X,S=0}(0|\boldsymbol{x}) P_0(\boldsymbol{x})}{\sum_{\boldsymbol{x}\in\mathcal{X}} P_{\hat{Y}|X}(1|\boldsymbol{x}) P_0(\boldsymbol{x})} - \frac{\sum_{\boldsymbol{x}\in\mathcal{X}} P_{\hat{Y}|X}(1|\boldsymbol{x}) P_{Y|X,S=1}(0|\boldsymbol{x}) P_1(\boldsymbol{x})}{\sum_{\boldsymbol{x}\in\mathcal{X}} P_{\hat{Y}|X}(1|\boldsymbol{x}) P_1(\boldsymbol{x})}.\end{aligned} \tag{B.60}$$

3. **FNR.**

$$\begin{aligned}&\Pr(\hat{Y}=0|Y=1,S=0)-\Pr(\hat{Y}=0|Y=1,S=1)\\&=\frac{\sum_{\boldsymbol{x}\in\mathcal{X}}P_{\hat{Y}|X}(0|\boldsymbol{x})P_{Y|X,S=0}(1|\boldsymbol{x})P_0(\boldsymbol{x})}{\sum_{\boldsymbol{x}\in\mathcal{X}}P_{Y|X,S=0}(1|\boldsymbol{x})P_0(\boldsymbol{x})}-\frac{\sum_{\boldsymbol{x}\in\mathcal{X}}P_{\hat{Y}|X}(0|\boldsymbol{x})P_{Y|X,S=1}(1|\boldsymbol{x})P_1(\boldsymbol{x})}{\sum_{\boldsymbol{x}\in\mathcal{X}}P_{Y|X,S=1}(1|\boldsymbol{x})P_1(\boldsymbol{x})}.\end{aligned}\tag{B.61}$$

4. **FPR.**

$$\begin{aligned}&\Pr(\hat{Y}=1|Y=0,S=0)-\Pr(\hat{Y}=1|Y=0,S=1)\\&=\frac{\sum_{\boldsymbol{x}\in\mathcal{X}}P_{\hat{Y}|X}(1|\boldsymbol{x})P_{Y|X,S=0}(0|\boldsymbol{x})P_0(\boldsymbol{x})}{\sum_{\boldsymbol{x}\in\mathcal{X}}P_{Y|X,S=0}(0|\boldsymbol{x})P_0(\boldsymbol{x})}-\frac{\sum_{\boldsymbol{x}\in\mathcal{X}}P_{\hat{Y}|X}(1|\boldsymbol{x})P_{Y|X,S=1}(0|\boldsymbol{x})P_1(\boldsymbol{x})}{\sum_{\boldsymbol{x}\in\mathcal{X}}P_{Y|X,S=1}(0|\boldsymbol{x})P_1(\boldsymbol{x})}.\end{aligned}\tag{B.62}$$

Now we are in a position to prove Proposition 11.

Proof. **Influence function for SP.** Recall that

$$\Pr(\hat{Y}=0|S=0)=1-\sum_{\boldsymbol{x}\in\mathcal{X}}h(\boldsymbol{x})P_0(\boldsymbol{x}).$$

When we perturb the distribution P_0, the classifier $h(\boldsymbol{x})$ and $\Pr(\hat{Y}=1|S=1)$ do not change. Therefore,

$$\begin{aligned}\psi(\boldsymbol{x})&=\lim_{\epsilon\to0}-\frac{1}{\epsilon}\left(\sum_{\boldsymbol{x}'\in\mathcal{X}}h(\boldsymbol{x}')((1-\epsilon)P_0(\boldsymbol{x}')+\epsilon\delta_{\boldsymbol{x}}(\boldsymbol{x}'))-\sum_{\boldsymbol{x}'\in\mathcal{X}}h(\boldsymbol{x}')P_0(\boldsymbol{x}')\right)\\&=-h(\boldsymbol{x})+\Pr(\hat{Y}=1|S=0).\end{aligned}$$

Influence function for FNR. Next, we compute the influence function of FNR. Similar analysis holds for FPR and FDR. Due to the factorization of the joint distribution, we have

$$\Pr(\hat{Y}=0|Y=1,S=0)=\frac{\sum_{\boldsymbol{x}'\in\mathcal{X}}P_{\hat{Y}|X}(0|\boldsymbol{x}')P_{Y|X,S=0}(1|\boldsymbol{x}')P_0(\boldsymbol{x}')}{\sum_{\boldsymbol{x}'\in\mathcal{X}}P_{Y|X,S=0}(1|\boldsymbol{x}')P_0(\boldsymbol{x}')}.$$

We denote $r_1(\boldsymbol{x})\triangleq P_{\hat{Y}|X}(0|\boldsymbol{x})P_{Y|X,S=0}(1|\boldsymbol{x})$ and $r_2(\boldsymbol{x})\triangleq P_{Y|X,S=0}(1|\boldsymbol{x})$. Then

$$\Pr(\hat{Y}=0|Y=1,S=0)=\frac{\sum_{\boldsymbol{x}'\in\mathcal{X}}r_1(\boldsymbol{x}')P_0(\boldsymbol{x}')}{\sum_{\boldsymbol{x}'\in\mathcal{X}}r_2(\boldsymbol{x}')P_0(\boldsymbol{x}')}=\frac{\mathbb{E}\left[r_1(X)|S=0\right]}{\mathbb{E}\left[r_2(X)|S=0\right]},$$

which implies

$$
\begin{aligned}
&\mathsf{M}((1-\epsilon)P_0 + \epsilon\delta_{\boldsymbol{x}}) \\
&= \frac{\sum_{\boldsymbol{x}' \in \mathcal{X}} r_1(\boldsymbol{x}')((1-\epsilon)P_0(\boldsymbol{x}') + \epsilon\delta_{\boldsymbol{x}}(\boldsymbol{x}'))}{\sum_{\boldsymbol{x}' \in \mathcal{X}} r_2(\boldsymbol{x}')((1-\epsilon)P_0(\boldsymbol{x}') + \epsilon\delta_{\boldsymbol{x}}(\boldsymbol{x}'))} - \Pr(\hat{Y} = 0 | Y = 1, S = 1) \\
&= \frac{\mathbb{E}\left[r_1(X)|S=0\right] + \epsilon\left(r_1(\boldsymbol{x}) - \mathbb{E}\left[r_1(X)|S=0\right]\right)}{\mathbb{E}\left[r_2(X)|S=0\right] + \epsilon\left(r_2(\boldsymbol{x}) - \mathbb{E}\left[r_2(X)|S=0\right]\right)} - \Pr(\hat{Y} = 0 | Y = 1, S = 1).
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\psi(\boldsymbol{x}) &= \lim_{\epsilon \to 0} \frac{1}{\epsilon}\left(\mathsf{M}((1-\epsilon)P_0 + \epsilon\delta_{\boldsymbol{x}}) - \mathsf{M}(P_0)\right) \\
&= \frac{\mathbb{E}\left[r_2(X)|S=0\right] r_1(\boldsymbol{x}) - \mathbb{E}\left[r_1(X)|S=0\right] r_2(\boldsymbol{x})}{\mathbb{E}\left[r_2(X)|S=0\right]^2} \\
&= \frac{\Pr(Y=1|S=0) r_1(\boldsymbol{x}) - \Pr(\hat{Y}=0, Y=1|S=0) r_2(\boldsymbol{x})}{\Pr(Y=1|S=0)^2} \\
&= \frac{P_{\hat{Y}|X}(0|\boldsymbol{x}) P_{Y|X,S=0}(1|\boldsymbol{x}) - \Pr(\hat{Y}=0|Y=1,S=0) P_{Y|X,S=0}(1|\boldsymbol{x})}{\Pr(Y=1|S=0)}.
\end{aligned}
$$

□

Appendix C

Appendix to Chapter 5

C.1 Proofs for Section 5.4

C.1.1 Proof of Lemma 11

Proof. By the definition of χ^2-information,

$$\chi^2(X;Y)+1=\sum_{x=1}^{|\mathcal{X}|}\sum_{y=1}^{|\mathcal{Y}|}\frac{P_{X,Y}(x,y)}{P_X(x)P_Y(y)}P_{X,Y}(x,y).$$

Note that $\mathbf{Q}_{X,Y}=\mathbf{D}_X^{-\frac{1}{2}}\mathbf{P}_{X,Y}\mathbf{D}_Y^{-\frac{1}{2}}$ which implies

$$\begin{aligned}\mathsf{tr}(\mathbf{Q}_{X,Y}\mathbf{Q}_{X,Y}^T)&=\mathsf{tr}(\mathbf{D}_X^{-\frac{1}{2}}\mathbf{P}_{X,Y}\mathbf{D}_Y^{-1}\mathbf{P}_{X,Y}^T\mathbf{D}_X^{-\frac{1}{2}})=\mathsf{tr}(\mathbf{D}_X^{-1}\mathbf{P}_{X,Y}\mathbf{D}_Y^{-1}\mathbf{P}_{X,Y}^T)\\&=\sum_{x=1}^{|\mathcal{X}|}\sum_{y=1}^{|\mathcal{Y}|}\frac{P_{X,Y}(x,y)}{P_X(x)}\frac{P_{X,Y}(x,y)}{P_Y(y)}.\end{aligned}$$

Therefore,

$$\chi^2(X;Y)=\mathsf{tr}(\mathbf{Q}_{X,Y}\mathbf{Q}_{X,Y}^T)-1=\mathsf{tr}(\mathbf{A})-1.$$

Since

$$\mathbf{Q}_{S,Y}=\mathbf{D}_S^{-\frac{1}{2}}\mathbf{P}_{S,Y}\mathbf{D}_Y^{-\frac{1}{2}}=\mathbf{D}_S^{-\frac{1}{2}}\mathbf{P}_{S,X}\mathbf{D}_X^{-\frac{1}{2}}\mathbf{D}_X^{-\frac{1}{2}}\mathbf{P}_{X,Y}\mathbf{D}_Y^{-\frac{1}{2}}=\mathbf{Q}_{S,X}\mathbf{Q}_{X,Y},$$

then

$$\chi^2(S;Y)=\mathsf{tr}(\mathbf{Q}_{S,Y}\mathbf{Q}_{S,Y}^T)-1=\mathsf{tr}(\mathbf{Q}_{S,X}\mathbf{Q}_{X,Y}\mathbf{Q}_{X,Y}^T\mathbf{Q}_{S,X}^T)-1=\mathsf{tr}(\mathbf{BA})-1.$$

□

C.1.2 Proof of Lemma 12

Proof. For $0 \leq \epsilon_1 < \epsilon_2 < \epsilon_3 \leq \chi^2(S;X)$, it suffices to show that

$$\frac{F_{\chi^2}(\epsilon_3; P_{S,X}) - F_{\chi^2}(\epsilon_1; P_{S,X})}{\epsilon_3 - \epsilon_1} \leq \frac{F_{\chi^2}(\epsilon_2; P_{S,X}) - F_{\chi^2}(\epsilon_1; P_{S,X})}{\epsilon_2 - \epsilon_1},$$

which is equivalent to

$$\frac{\epsilon_2 - \epsilon_1}{\epsilon_3 - \epsilon_1} F_{\chi^2}(\epsilon_3; P_{S,X}) + \frac{\epsilon_3 - \epsilon_2}{\epsilon_3 - \epsilon_1} F_{\chi^2}(\epsilon_1; P_{S,X}) \leq F_{\chi^2}(\epsilon_2; P_{S,X}). \tag{C.1}$$

Let $P_{Y_1|X}$ and $P_{Y_3|X}$ be two optimal solutions in $\mathcal{D}(\epsilon_1; P_{S,X})$ and $\mathcal{D}(\epsilon_3; P_{S,X})$, respectively. Assume that Y_1 and Y_3 take values in $[m_1]$ and $[m_3]$, respectively. Furthermore, we denote $\lambda \triangleq \frac{\epsilon_2 - \epsilon_1}{\epsilon_3 - \epsilon_1}$. Next, we introduce a new privacy mechanism defined as

$$P_{Y_\lambda|X}(y|x) \triangleq \begin{cases} \lambda P_{Y_3|X}(y|x) & \text{if } y \in [m_3], \\ (1-\lambda) P_{Y_1|X}(y - m_3|x) & \text{if } y - m_3 \in [m_1]. \end{cases} \tag{C.2}$$

Consequently, we have

$$P_{Y_\lambda}(y) = \begin{cases} \lambda P_{Y_3}(y) & \text{if } y \in [m_3], \\ (1-\lambda) P_{Y_1}(y - m_3) & \text{if } y - m_3 \in [m_1]. \end{cases}$$

Then

$$\begin{aligned}
\chi^2(X;Y_\lambda) &= \mathbb{E}\left[\frac{P_{X,Y_\lambda}(X,Y_\lambda)}{P_X(X)P_{Y_\lambda}(Y_\lambda)}\right] - 1 \\
&= \sum_{y \in [m_3]} \sum_{x=1}^{|\mathcal{X}|} \frac{P_{X,Y_\lambda}(x,y)^2}{P_X(x)P_{Y_\lambda}(y)} + \sum_{y - m_3 \in [m_1]} \sum_{x=1}^{|\mathcal{X}|} \frac{P_{X,Y_\lambda}(x,y)^2}{P_X(x)P_{Y_\lambda}(y)} - 1 \\
&= \sum_{y \in [m_3]} \sum_{x=1}^{|\mathcal{X}|} \frac{P_{Y_\lambda|X}(y|x)^2 P_X(x)}{P_{Y_\lambda}(y)} + \sum_{y - m_3 \in [m_1]} \sum_{x=1}^{|\mathcal{X}|} \frac{P_{Y_\lambda|X}(y|x)^2 P_X(x)}{P_{Y_\lambda}(y)} - 1 \\
&= \sum_{y \in [m_3]} \sum_{x=1}^{|\mathcal{X}|} \frac{\lambda^2 P_{Y_3|X}(y|x)^2 P_X(x)}{\lambda P_{Y_3}(y)} + \sum_{y \in [m_1]} \sum_{x=1}^{|\mathcal{X}|} \frac{(1-\lambda)^2 P_{Y_1|X}(y|x)^2 P_X(x)}{(1-\lambda) P_{Y_1}(y)} - 1 \\
&= \lambda \chi^2(X;Y_3) + (1-\lambda)\chi^2(X;Y_1).
\end{aligned}$$

Similarly, we have

$$\chi^2(S;Y_\lambda) = \lambda \chi^2(S;Y_3) + (1-\lambda)\chi^2(S;Y_1) \leq \epsilon_2, \tag{C.3}$$

which implies that $P_{Y_\lambda|X} \in \mathcal{D}(\epsilon_2; P_{S,X})$. Therefore,

$$\begin{aligned} F_{\chi^2}(\epsilon_2; P_{S,X}) &\geq \chi^2(X; Y_\lambda) \\ &= \lambda\chi^2(X; Y_3) + (1-\lambda)\chi^2(X; Y_1) \\ &= \frac{\epsilon_2 - \epsilon_1}{\epsilon_3 - \epsilon_1} F_{\chi^2}(\epsilon_3; P_{S,X}) + \frac{\epsilon_3 - \epsilon_2}{\epsilon_3 - \epsilon_1} F_{\chi^2}(\epsilon_1; P_{S,X}), \end{aligned} \tag{C.4}$$

which implies that (C.1) is true, so $F_{\chi^2}(\epsilon; P_{S,X})$ is a concave function. Furthermore, $\epsilon \to \frac{1}{\epsilon} F_{\chi^2}(\epsilon; P_{S,X})$ is non-increasing since $F_{\chi^2}(\epsilon; P_{S,X})$ is non-negative and concave. □

C.1.3 Proof of Theorem 7

Proof. The lower bound for $F_{\chi^2}(\epsilon; P_{S,X})$ follows immediately from the concavity of $F_{\chi^2}(\epsilon; P_{S,X})$ and

$$\begin{aligned} F_{\chi^2}(0; P_{S,X}) &\geq 0, \\ F_{\chi^2}\left(\chi^2(S; X); P_{S,X}\right) &= \chi^2(X; X) = |\mathcal{X}| - 1. \end{aligned}$$

Using Lemma 11, the χ^2-privacy-utility function can be simplified as

$$\begin{aligned} F_{\chi^2}(\epsilon; P_{S,X}) &= \max_{P_{Y|X} \in \mathcal{D}(\epsilon; P_{S,X})} \mathsf{tr}(\mathbf{A}) - 1, \\ \mathcal{D}(\epsilon; P_{S,X}) &= \{P_{Y|X} \mid S \to X \to Y, \mathsf{tr}(\mathbf{BA}) - 1 \leq \epsilon\}, \end{aligned}$$

where

$$\mathbf{A} = \mathbf{Q}_{X,Y}\mathbf{Q}_{X,Y}^T, \mathbf{B} = \mathbf{Q}_{S,X}^T\mathbf{Q}_{S,X}.$$

We denote the singular value decomposition of $\mathbf{Q}_{S,X}$ and $\mathbf{Q}_{X,Y}$ by $\mathbf{Q}_{S,X} = \mathbf{W}\mathbf{\Sigma}_1\mathbf{U}^T$ and $\mathbf{Q}_{X,Y} = \mathbf{V}\mathbf{\Sigma}_2\mathbf{M}^T$, respectively. Then $\mathbf{B} = \mathbf{U}\mathbf{\Sigma}_1^T\mathbf{\Sigma}_1\mathbf{U}^T = \mathbf{U}\mathbf{\Sigma}_B\mathbf{U}^T$, $\mathbf{A} = \mathbf{V}\mathbf{\Sigma}_2\mathbf{\Sigma}_2^T\mathbf{V}^T = \mathbf{V}\mathbf{\Sigma}_A\mathbf{V}^T$ where $\mathbf{\Sigma}_B \triangleq \mathbf{\Sigma}_1^T\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_A \triangleq \mathbf{\Sigma}_2\mathbf{\Sigma}_2^T$.

Let $\mathbf{A}_1 = \mathbf{U}^T\mathbf{A}\mathbf{U} = \mathbf{L}\mathbf{\Sigma}_A\mathbf{L}^T$ where $\mathbf{L} \triangleq \mathbf{U}^T\mathbf{V}$. Suppose the diagonal elements of $\mathbf{A}_1$ are $a_1, ..., a_{|\mathcal{X}|}$. Then, from characterization 3 in Theorem 1, we have

$$\mathsf{tr}(\mathbf{BA}) - 1 = a_1 - 1 + \sum_{i=2}^{d+1} \lambda_{i-1}(S; X) a_i. \tag{C.5}$$

Suppose the i-th row of $\mathbf{L}$ is $\mathbf{l}_i = (l_{i,1}, ..., l_{i,|\mathcal{X}|})$, the i-th column of $\mathbf{U}$ is $\mathbf{u}_i^T$ and $\mathbf{\Sigma}_A = \mathsf{diag}(\sigma_1, ..., \sigma_{|\mathcal{X}|})$. By characterization 3 in Theorem 1, $\sigma_1 = 1$, $\sigma_{j+1} = \lambda_j(X; Y)$ for $j = 1, ..., d$ and $\sigma_{j+1} = 0$ for

$j = d+1, ..., |\mathcal{X}|-1$. Then, for $\forall i \in [|\mathcal{X}|]$,

$$0 \le a_i = \sum_{j=1}^{|\mathcal{X}|} \sigma_j l_{i,j}^2 \le \sum_{j=1}^{|\mathcal{X}|} l_{ij}^2 = 1.$$

Since, following from characterization 3 in Theorem 1, the first column of $\mathbf{U}$ and that of $\mathbf{V}$ are both $(\sqrt{P_X(1)}, ..., \sqrt{P_X(|\mathcal{X}|)})^T$, then $\mathbf{l}_1 = \mathbf{u}_1\mathbf{V} = (1, 0, ..., 0)$. Therefore, $a_1 = \sigma_1 = 1$. If $P_{Y|X} \in \mathcal{D}(\epsilon; P_{S,X})$, then (C.5) shows

$$\mathrm{tr}(\mathbf{BA}) - 1 = \sum_{i=2}^{d+1} \lambda_{i-1}(S;X) a_i \le \epsilon,$$

which implies that

$$(a_2, ..., a_{|\mathcal{X}|}) \in \mathcal{D}_\epsilon^{|\mathcal{X}|-1}(\lambda_1(S;X), ..., \lambda_d(S;X)).$$

Thus,

$$\begin{aligned} F_{\chi^2}(\epsilon; P_{S,X}) &\le \max_{\substack{(a_2,...,a_{|\mathcal{X}|}) \\ \in \mathcal{D}_\epsilon^{|\mathcal{X}|-1}(\lambda_1(S;X),...,\lambda_d(S;X))}} \sum_{i=2}^{|\mathcal{X}|} a_i \\ &= G_\epsilon^{|\mathcal{X}|-1}(\lambda_1(S;X), ..., \lambda_d(S;X)), \end{aligned}$$

as required. □

Next, we derive a closed-form expression of $G_\epsilon^m(t_1, ..., t_n)$. Recall that, for $t_i \in [0,1]$ $(i \in [n])$, $0 \le \epsilon \le \sum_{i\in[n]} t_i$, and $n \le m$, $G_\epsilon^m(t_1, ..., t_n)$ is defined as:

$$G_\epsilon^m(t_1, ..., t_n) \triangleq \max\left\{ \sum_{i=1}^m x_i \;\middle|\; (x_1, ..., x_m) \in \mathcal{D}_\epsilon^m(t_1, ..., t_n) \right\},$$

where

$$\mathcal{D}_\epsilon^m(t_1, ..., t_n) \triangleq \left\{ (x_1, ..., x_m) \;\middle|\; \sum_{i=1}^n t_i x_i \le \epsilon, x_i \in [0,1], i \in [m] \right\}.$$

We assume $1 \ge t_1 \ge ... \ge t_{n-s} > t_{n-s+1} = ... = t_n = 0$ without loss of generality. Then we divide $[0, \sum_{i=1}^n t_i]$ into $n-s$ intervals:

$$\left[0, \sum_{i=1}^n t_i\right] = \bigcup_{j=0}^{n-1-s} \left[\sum_{i=n-s-j+1}^{n-s} t_i, \sum_{i=n-s-j}^{n-s} t_i\right].$$

If $\epsilon \in \left[\sum_{i=n-s-j+1}^{n-s} t_i, \sum_{i=n-s-j}^{n-s} t_i\right]$, then

$$G_\epsilon^m(t_1, ..., t_n) = s + (m-n) + j + \frac{\epsilon - \sum_{i=n-s-j+1}^{n-s} t_i}{t_{n-s-j}}, \tag{C.6}$$

and it can be achieved by setting

$$x_i = 1, \text{ for } i = n-s-j+1, ..., m,$$
$$x_{n-s-j} = \frac{\epsilon - \sum_{i=n-s-j+1}^{n-s} t_i}{t_{n-s-j}},$$
$$x_i = 0, \text{ for } i = 1, ..., n-s-j-1.$$

C.1.4 Proof of Corollary 3

Proof. First, $F_{\chi^2}(\epsilon; P_{S,X})$ is non-decreasing since, for any $0 \leq \epsilon_1 < \epsilon_2 \leq \chi^2(S;X)$, we have $\mathcal{D}(\epsilon_1; P_{S,X}) \subseteq \mathcal{D}(\epsilon_2; P_{S,X})$. Now suppose there exist ϵ_1 and ϵ_2, such that $F_{\chi^2}(\epsilon_1; P_{S,X}) = F_{\chi^2}(\epsilon_2; P_{S,X})$. We denote $\chi^2(S;X)$ by ϵ_0. Since $F_{\chi^2}(\epsilon; P_{S,X})$ is a concave and non-decreasing function, then for any $\epsilon > \epsilon_1$, $F_{\chi^2}(\epsilon; P_{S,X}) = F_{\chi^2}(\epsilon_1; P_{S,X})$. In particular, $F_{\chi^2}(\epsilon_1; P_{S,X}) = F_{\chi^2}(\epsilon_0; P_{S,X}) = |\mathcal{X}| - 1$. This contradicts the upper bound of χ^2-privacy-utility function in Theorem 7 since the upper bound implies that $F_{\chi^2}(\epsilon; P_{S,X}) < |\mathcal{X}| - 1$ when $\epsilon < \epsilon_0$. □

C.1.5 Proof of Theorem 8

Proof. Following from characterization 2 in Theorem 1, there exists $f \in \mathcal{L}_2(P_X)$ such that $\|f(X)\|_2 = 1$, $\mathbb{E}[f(X)] = 0$ and $\|\mathbb{E}[f(X)|S]\|_2^2 = \lambda_{\min}(S;X)$.

Fix $\mathcal{Y} = \{1, 2\}$ and the privacy mechanism is defined as

$$P_{Y|X}(y|x) = \frac{1}{2} + (-1)^y \frac{\sqrt{P_{X\min}} f(x)}{2}. \tag{C.7}$$

Since

$$1 = ||f(X)||_2^2 = \sum_{x=1}^{|\mathcal{X}|} f(x)^2 P_X(x) \geq f(x)^2 P_X(x),$$

for any $x \in [|\mathcal{X}|]$

$$|f(x)| \leq \frac{1}{\sqrt{P_X(x)}} \leq \frac{1}{\sqrt{P_{X\min}}}.$$

Therefore, $\left|\frac{\sqrt{P_{X\min}} f(x)}{2}\right| \leq \frac{1}{2}$, which implies that $P_{Y|X}(y|x)$ is feasible. Furthermore, $P_Y(y) = \frac{1}{2}$ because of $\mathbb{E}[f(X)] = 0$.

$$\chi^2(X;Y) = \sum_{x=1}^{|\mathcal{X}|} \sum_{y=1}^{|\mathcal{Y}|} \frac{P_{Y|X}(y|x)^2 P_X(x)}{P_Y(y)} - 1 = \sum_{x=1}^{|\mathcal{X}|} (P_X(x) + P_{X\min} f(x)^2 P_X(x)) - 1 = P_{X\min}.$$

Since

$$P_{Y|S}(y|s) = \sum_{x=1}^{|\mathcal{X}|} P_{Y|X}(y|x)P_{X|S}(x|s) = \sum_{x=1}^{|\mathcal{X}|} \left(\frac{1}{2} + (-1)^y \frac{\sqrt{P_{X\min}} f(x)}{2}\right) P_{X|S}(x|s)$$
$$= \frac{1}{2} + (-1)^y \frac{\sqrt{P_{X\min}}}{2} \mathbb{E}\left[f(X)|S=s\right],$$

then

$$\chi^2(S;Y) = \sum_{s=1}^{|\mathcal{S}|} \sum_{y=1}^{|\mathcal{Y}|} \frac{P_{Y|S}(y|s)^2 P_S(s)}{P_Y(y)} - 1 = \sum_{s=1}^{|\mathcal{S}|} \left(P_S(s) + P_{X\min}\mathbb{E}\left[f(X)|S=s\right]^2 P_S(s)\right) - 1$$
$$= P_{X\min}\lambda_{\min}(S;X).$$

Hence, this Y satisfies $\chi^2(X;Y) = P_{X\min}$ and $\chi^2(S;Y) = P_{X\min}\lambda_{\min}(S;X)$. □

C.2 Proofs for Section 5.5

C.2.1 Proof of Lemma 13

Proof. Suppose $||f(S)||_2 = 1$ without loss of generality. Observe that

$$\mathbb{E}\left[\mathbb{E}\left[f(S)|X\right]\right] = \mathbb{E}\left[f(S)\right] = 0.$$

Since $f(S) \to X \to Y$, then $\mathbb{E}\left[f(S)|X\right] = \mathbb{E}\left[f(S)|X,Y\right]$. Therefore,

$$\mathsf{mmse}\left(\frac{\mathbb{E}\left[f(S)|X\right]}{||\mathbb{E}\left[f(S)|X\right]||_2}\middle|Y\right) = \frac{\mathbb{E}\left[\mathbb{E}\left[f(S)|X\right]^2\right] - \mathbb{E}\left[\mathbb{E}\left[\mathbb{E}\left[f(S)|X\right]|Y\right]^2\right]}{||\mathbb{E}\left[f(S)|X\right]||_2^2}$$
$$= \frac{\mathbb{E}\left[\mathbb{E}\left[f(S)|X\right]^2\right] - \mathbb{E}\left[\mathbb{E}\left[\mathbb{E}\left[f(S)|X,Y\right]|Y\right]^2\right]}{||\mathbb{E}\left[f(S)|X\right]||_2^2}$$
$$= 1 - \frac{\mathbb{E}\left[\mathbb{E}\left[f(S)|Y\right]^2\right]}{||\mathbb{E}\left[f(S)|X\right]||_2^2}$$
$$= \mathbb{E}\left[f(S)^2\right] - \frac{\mathbb{E}\left[\mathbb{E}\left[f(S)|Y\right]^2\right]}{||\mathbb{E}\left[f(S)|X\right]||_2^2}$$
$$\leq \mathbb{E}\left[f(S)^2\right] - \mathbb{E}\left[\mathbb{E}\left[f(S)|Y\right]^2\right]$$
$$= \mathsf{mmse}(f(S)|Y),$$

where the last inequality follows from Jensen's inequality:

$$1 = \mathbb{E}\left[f(S)^2\right] = \mathbb{E}\left[\mathbb{E}\left[f(S)^2|X\right]\right] \geq \mathbb{E}\left[\mathbb{E}\left[f(S)|X\right]^2\right] = ||\mathbb{E}\left[f(S)|X\right]||_2^2.$$

□

C.3 Proofs for Section 5.6

C.3.1 Proof of Lemma 14

Proof. For fixed $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$ where $a_i > 0$ and $b_i \geq 0$, let $L_P : \mathbb{R}^n \to \mathbb{R}$ and $L_D : \mathbb{R}^n \to \mathbb{R}$ be given by

$$L_P(\boldsymbol{y}) \triangleq \boldsymbol{a}^T \boldsymbol{y},$$
$$L_D(\boldsymbol{u}) \triangleq \boldsymbol{a}^T \boldsymbol{b} + \boldsymbol{u}^T \boldsymbol{b} + \|\boldsymbol{u}\|_2.$$

Furthermore, we define $\mathcal{A}(\boldsymbol{a}) \triangleq \{\boldsymbol{u} \in \mathbb{R}^n \mid \boldsymbol{u} \geq -\boldsymbol{a}\}$ and $\mathcal{B}(\boldsymbol{b}) \triangleq \{\boldsymbol{y} \in \mathbb{R}^n \mid \|\boldsymbol{y}\|_2 \leq 1, \boldsymbol{y} \leq \boldsymbol{b}\}$.

Assume, without loss of generality, that $b_1/a_1 \leq b_2/a_2 \leq \cdots \leq b_n/a_n$, and let k^* be defined in (5.20). Note that $b_1 \leq 1$ and for $k \in [k^*]$

$$\sum_{i=1}^{k} b_i^2 \leq \frac{a_k^2}{\sum_{i=k}^{n} a_i^2} \left(1 - \sum_{i=1}^{k-1} b_i^2\right)^+ + \sum_{i=1}^{k-1} b_i^2,$$

so $\sum_{i=1}^{k} b_i^2 \leq 1$. Especially, $\sum_{i=1}^{k^*} b_i^2 \leq 1$. For $c_j \triangleq \sqrt{\frac{\left(1-\sum_{i=1}^{j} b_i^2\right)}{\|\boldsymbol{a}\|_2^2 - \sum_{i=1}^{j} a_i^2}}$, let

$$\boldsymbol{y}^* = (b_1, \ldots, b_{k^*}, a_{k^*+1} c_{k^*}, \ldots, a_n c_{k^*})$$

and

$$\boldsymbol{u}^* = (-b_1/c_{k^*}, \ldots, -b_{k^*}/c_{k^*}, -a_{k^*+1}, \ldots, -a_n).$$

From the definition of k^*, $\boldsymbol{y}^* \in \mathcal{B}(\boldsymbol{b})$ and $\boldsymbol{u}^* \in \mathcal{A}(\boldsymbol{a})$. Furthermore,

$$\begin{aligned} L_P(\boldsymbol{y}^*) &= \boldsymbol{a}^T \boldsymbol{y}^* \\ &= \sum_{i=1}^{k^*} a_i b_i + \sum_{i=k^*+1}^{n} c_{k^*} a_i^2 \\ &= \sum_{i=1}^{k^*} a_i b_i + \sqrt{\left(\|\boldsymbol{a}\|_2^2 - \sum_{i=1}^{k^*} a_i^2\right)\left(1 - \sum_{i=1}^{k^*} b_i^2\right)}, \end{aligned} \tag{C.8}$$

and

$$\begin{aligned} L_D(\boldsymbol{u}^*) &= \boldsymbol{a}^T\boldsymbol{b} + \boldsymbol{u}^{*T}\boldsymbol{b} + \|\boldsymbol{u}^*\|_2 \\ &= \sum_{i=1}^{k^*}\left(a_i b_i - \frac{b_i^2}{c_k^*}\right) + c_{k^*}^{-1}\sqrt{\sum_{i=1}^{k^*} b_i^2 + c_{k^*}^2\left(\|\boldsymbol{a}\|_2^2 - \sum_{i=1}^{k^*} a_i^2\right)} \\ &= \sum_{i=1}^{k^*} a_i b_i + c_{k^*}^{-1}\left(1 - \sum_{i=1}^{k^*} b_i^2\right) \\ &= \sum_{i=1}^{k^*} a_i b_i + \sqrt{\left(\|\boldsymbol{a}\|_2^2 - \sum_{i=1}^{k^*} a_i^2\right)\left(1 - \sum_{i=1}^{k^*} b_i^2\right)} = L_P(\boldsymbol{y}^*). \end{aligned}$$

Since both the primal and the dual achieve the same value at $\boldsymbol{y}^*$ and $\boldsymbol{u}^*$, respectively, it follows that the value $L_P(\boldsymbol{y}^*)$ given in (C.8) is optimal. □

C.3.2 Proof of Theorem 9

Proof. Let

$$h(x) \triangleq \begin{cases} \rho_0^{-1}(\phi(x) - \sum_{i=1}^m \rho_i \phi_i(x)) & \text{if } \rho_0 > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{C.9}$$

Note that when $\rho_0 > 0$, we have $\|h(X)\|_2 = 1$. Then for any $\psi \in \mathcal{L}_2(P_Y)$ and $\|\psi(Y)\|_2 = 1$,

$$\begin{aligned} |\mathbb{E}\left[\phi(X)\psi(Y)\right]| &= \left|\rho_0 \mathbb{E}\left[h(X)\psi(Y)\right] + \sum_{i=1}^m \rho_i \mathbb{E}\left[\phi_i(X)\psi(Y)\right]\right| \\ &\le \rho_0 \left|\mathbb{E}\left[h(X)\psi(Y)\right]\right| + \sum_{i=1}^m \left|\rho_i \mathbb{E}\left[\phi_i(X)\psi(Y)\right]\right| \\ &= \rho_0 \left|\mathbb{E}\left[h(X)(T_{Y|X}\psi)(X)\right]\right| + \sum_{i=1}^m \left|\rho_i \mathbb{E}\left[\phi_i(X)(T_{Y|X}\psi)(X)\right]\right|, \end{aligned}$$

where $T_{Y|X}$ is defined in Section 5.3. Denoting $|\mathbb{E}\left[h(X)(T_{Y|X}\psi)(X)\right]| \triangleq x_0$, $|\mathbb{E}\left[\phi_i(X)(T_{Y|X}\psi)(X)\right]| \triangleq x_i$, $\boldsymbol{x} \triangleq (x_0, x_1, \ldots, x_m)$, the last inequality can be rewritten as

$$|\mathbb{E}\left[\phi(X)\psi(Y)\right]| \le \boldsymbol{\rho}_0^T \boldsymbol{x}. \tag{C.10}$$

Observe that $\|\boldsymbol{x}\|_2 \le 1$ and $x_i \le \nu_i$ for $i \in [m]$, and the right hand side of (C.10) can be maximized over all values of $\boldsymbol{x}$ that satisfy these constraints. We assume, without loss of generality, that $\rho_0 > 0$ (otherwise set $x_0 = 0$). The left-hand side of (C.10) can be further bounded by

$$|\mathbb{E}\left[\phi(X)\psi(Y)\right]| \le L_{m+1}(\boldsymbol{\rho}_0, \boldsymbol{\nu}_0), \tag{C.11}$$

where $\boldsymbol{\nu}_0 = (1, \nu_1, \ldots, \nu_m)$ and L_{m+1} is defined in (5.19). The result follows directly from Lemma 14 and noting that $\max_{\psi \in \mathcal{L}_2(P_Y)} \mathbb{E}\left[\phi(X)\psi(Y)\right] = \|\mathbb{E}\left[\phi(X)|Y\right]\|_2$. □

C.3.3 Proof of Theorem 10

Proof. For any $\psi \in \mathcal{L}_2(P_Y)$ with $\|\psi(Y)\|_2 = 1$, let $\alpha_i \triangleq \mathbb{E}\left[\psi(Y)\psi_i(Y)\right]$, $\alpha_0 \triangleq \sqrt{1 - \sum_{i=i}^{t} \alpha_i^2}$ and $\psi_0(Y) \triangleq \alpha_0^{-1}(\psi(Y) - \sum_{i=1}^{t} \alpha_i \psi_i(Y))$ if $\alpha_0 > 0$, otherwise $\psi_0(Y) \triangleq 0$. Observe that $\|\psi_0(Y)\|_2 = 1$ when $\alpha_0 > 0$. Also, $\mathbb{E}\left[\phi_i(X)\psi_j(Y)\right] = 0$ for $i \neq j$, $i \in \{0, \ldots, m\}$, $j \in [t]$. Consequently,

$$\begin{aligned}
\mathbb{E}\left[\phi(X)\psi(Y)\right] &= \mathbb{E}\left[\left(\sum_{i=0}^{m} \rho_i \phi_i(X)\right)\left(\sum_{j=0}^{t} \alpha_j \psi_j(Y)\right)\right] \\
&= \sum_{i=0}^{m}\sum_{j=0}^{t} \rho_i \alpha_j \mathbb{E}\left[\phi_i(X)\psi_j(Y)\right] \\
&\leq \left|\alpha_0 \sum_{i=0, i\notin[t]}^{m} \rho_i \mathbb{E}\left[\phi_i(X)\psi_0(Y)\right]\right| + \sum_{i=1}^{t} |\nu_i \rho_i \alpha_i| \\
&\leq |\alpha_0| B_{m-t}\left(\widetilde{\boldsymbol{\rho}}, \widetilde{\boldsymbol{\nu}}\right) + \sum_{i=1}^{t} |\nu_i \rho_i \alpha_i| \qquad \text{(C.12)} \\
&\leq \sqrt{\sum_{k=1}^{t} \nu_i^2 \rho_i^2 + B_{m-t}\left(\widetilde{\boldsymbol{\rho}}, \widetilde{\boldsymbol{\nu}}\right)^2}. \qquad \text{(C.13)}
\end{aligned}$$

Inequality (C.12) follows from the similar proof of Theorem 9, and (C.13) follows by observing that $\sum_{i=0}^{t} \alpha_i^2 = 1$ and applying Cauchy-Schwarz inequality.

Finally, when $\rho_0 = 0$ and $t = m$, (C.13) can be achieved with equality by taking

$$\psi(Y) = \frac{\sum_{i=1}^{m} \nu_i \rho_i \psi_i(Y)}{\sqrt{\sum_{i=1}^{m} \nu_i^2 \rho_i^2}}.$$

□

Appendix D

Appendix to Chapter 6

D.1 Proofs for Section 6.4

D.1.1 Proof of Lemma 15

Consider the following elementary lemma whose proof is included for the sake of completeness.

Lemma 33. *Let $Q_1, Q_2 \in \mathcal{P}$ and $W \in \mathcal{W}$ be given. For each $i \in \{1,2\}$, let $S_i \to X_i \to Y_i$ be such that $P_{S_i,X_i} = Q_i$ and $P_{Y_i|X_i} = W$. Then,*

$$\|P_{S_1,Y_1} - P_{S_2,Y_2}\|_1 \leq \|Q_1 - Q_2\|_1. \tag{D.1}$$

Proof. By the definition of $\|\cdot\|_1$, we have that

$$\|P_{S_1,Y_1} - P_{S_2,Y_2}\|_1 = \sum_{s\in\mathcal{S}} \sum_{y\in\mathcal{Y}} |P_{S_1,Y_1}(s,y) - P_{S_2,Y_2}(s,y)|. \tag{D.2}$$

By assumption, for each $i \in \{1,2\}$, $P_{S_i,Y_i}(s,y) = \sum_x Q_i(s,x)W(x,y)$. Hence, by the triangle inequality,

$$\|P_{S_1,Y_1} - P_{S_2,Y_2}\|_1 \leq \sum_{y\in\mathcal{Y}} \sum_{s\in\mathcal{S}} \sum_{x\in\mathcal{X}} |Q_1(s,x) - Q_2(s,x)|W(x,y) \tag{D.3}$$

$$= \|Q_1 - Q_2\|_1, \tag{D.4}$$

where we used the fact that $\sum_y W(x,y) = 1$ for all $x \in \mathcal{X}$. □

The proof of Lemma 15 relies on the following elementary observation: Let $n \in \mathbb{N}$. If $a_i, b_i \in \mathbb{R}$

for all $i = 1, \ldots, n$, then

$$\left| \max_{i=1,\ldots,n} a_i - \max_{i=1,\ldots,n} b_i \right| \leq \max_{i=1,\ldots,n} |a_i - b_i| \leq \sum_{i=1}^{n} |a_i - b_i|. \tag{D.5}$$

Proof. For ease of notation, let $\Delta_L \triangleq |\mathcal{L}_c(Q_1, W) - \mathcal{L}_c(Q_2, W)|$. For each $i \in \{1, 2\}$, let $S_i \to X_i \to Y_i$ be such that $P_{S_i, X_i} = Q_i$ and $P_{Y_i|X_i} = W$. With this notation, $\mathcal{L}_c(Q_i, W) = P_c(S_i|Y_i)$ where

$$P_c(S_i|Y_i) = \sum_{y \in \mathcal{Y}} \max_{s \in \mathcal{S}} P_{S_i, Y_i}(s, y). \tag{D.6}$$

By the triangle inequality, we have that

$$\Delta_L \leq \sum_{y \in \mathcal{Y}} \left| \max_{s \in \mathcal{S}} P_{S_1, Y_1}(s, y) - \max_{s \in \mathcal{S}} P_{S_2, Y_2}(s, y) \right|. \tag{D.7}$$

An immediate application of (D.5) leads to

$$\Delta_L \leq \sum_{y \in \mathcal{Y}} \sum_{s \in \mathcal{S}} |P_{S_1, Y_1}(s, y) - P_{S_2, Y_2}(s, y)| \tag{D.8}$$

$$= \|P_{S_1, Y_1} - P_{S_2, Y_2}\|_1 \tag{D.9}$$

$$\leq \|Q_1 - Q_2\|_1, \tag{D.10}$$

where the last inequality follows from Lemma 33. Mutatis mutandis, it can be shown that

$$|\mathcal{U}_c(Q_1, W) - \mathcal{U}_c(Q_2, W)| \leq \|Q_1 - Q_2\|_1, \tag{D.11}$$

as required. □

D.1.2 Proof of Lemma 16

The proof of the following lemma is similar to the one of Lemma 33. The details are left to the reader.

Lemma 34. *Let $Q_1, Q_2 \in \mathcal{P}$ and $W \in \mathcal{W}$ be given. For each $i \in \{1, 2\}$, let $S_i \to X_i \to Y_i$ be such that $P_{S_i, X_i} = Q_i$ and $P_{Y_i|X_i} = W$. Then,*

$$\|P_{Y_1} - P_{Y_2}\|_1 \leq \max\{\|P_{S_1} - P_{S_2}\|_1, \|P_{X_1} - P_{X_2}\|_1\} \leq \|Q_1 - Q_2\|_1. \tag{D.12}$$

Now we are in position to prove Lemma 16.

Proof. For ease of notation, let $\Delta_L \triangleq |\mathcal{L}_f(Q_1, W) - \mathcal{L}_f(Q_2, W)|$. For each $i \in \{1, 2\}$, let $S_i \to X_i \to Y_i$ be such that $P_{S_i, X_i} = Q_i$ and $P_{Y_i|X_i} = W$. With this notation, $\mathcal{L}_f(Q_i, W) = I_f(P_{S_i, Y_i})$. By the definition

of f-information, we have that

$$\Delta_L = \left| \sum_{s\in\mathcal{S}} \sum_{y\in\mathcal{Y}} \left(P_{S_1}(s)P_{Y_1}(y) f\left(\frac{P_{S_1,Y_1}(s,y)}{P_{S_1}(s)P_{Y_1}(y)} \right) - P_{S_2}(s)P_{Y_2}(y) f\left(\frac{P_{S_2,Y_2}(s,y)}{P_{S_2}(s)P_{Y_2}(y)} \right) \right) \right|. \tag{D.13}$$

An application of the triangle inequality leads to $\Delta_L \leq \mathrm{I} + \mathrm{II}$, where

$$\mathrm{I} = \sum_{s\in\mathcal{S}} \sum_{y\in\mathcal{Y}} |P_{S_1}(s)P_{Y_1}(y) - P_{S_2}(s)P_{Y_2}(y)| \left| f\left(\frac{P_{S_1,Y_1}(s,y)}{P_{S_1}(s)P_{Y_1}(y)} \right) \right|, \tag{D.14}$$

$$\mathrm{II} = \sum_{s\in\mathcal{S}} \sum_{y\in\mathcal{Y}} P_{S_2}(s)P_{Y_2}(y) \left| f\left(\frac{P_{S_1,Y_1}(s,y)}{P_{S_1}(s)P_{Y_1}(y)} \right) - f\left(\frac{P_{S_2,Y_2}(s,y)}{P_{S_2}(s)P_{Y_2}(y)} \right) \right|. \tag{D.15}$$

First we provide an upper bound for I. Recall the definition of the supremum norm in (6.40). Observe that $P_{S_i,Y_i}(s,y) \leq P_{Y_i}(y)$ for all $i \in \{1,2\}$, $s \in \mathcal{S}$, and $y \in \mathcal{Y}$. Hence, for all $s \in \mathcal{S}$ and $y \in \mathcal{Y}$,

$$\max\left\{ \frac{P_{S_1,Y_1}(s,y)}{P_{S_1}(s)P_{Y_1}(y)}, \frac{P_{S_2,Y_2}(s,y)}{P_{S_2}(s)P_{Y_2}(y)} \right\} \leq \max\left\{ \frac{1}{P_{S_1}(s)}, \frac{1}{P_{S_2}(s)} \right\} \leq m_S^{-1}, \tag{D.16}$$

and thus $\left| f\left(\frac{P_{S_1,Y_1}(s,y)}{P_{S_1}(s)P_{Y_1}(y)} \right) \right| \leq K_{f,m_S^{-1}}$, as $|f(t)| \leq K_{f,m_S^{-1}}$ for all $t \in [0, m_S^{-1}]$. As a consequence,

$$\mathrm{I} \leq K_{f,m_S^{-1}} \sum_{s\in\mathcal{S}} \sum_{y\in\mathcal{Y}} |P_{S_1}(s)P_{Y_1}(y) - P_{S_2}(s)P_{Y_2}(y)| \tag{D.17}$$

$$\leq K_{f,m_S^{-1}} \sum_{s\in\mathcal{S}} \sum_{y\in\mathcal{Y}} \left[P_{Y_1}(y)|P_{S_1}(s) - P_{S_2}(s)| + P_{S_2}(s)|P_{Y_1}(y) - P_{Y_2}(y)| \right] \tag{D.18}$$

$$= K_{f,m_S^{-1}} \left(\|P_{S_1} - P_{S_2}\|_1 + \|P_{Y_1} - P_{Y_2}\|_1 \right). \tag{D.19}$$

By Lemma 34, we conclude that

$$\mathrm{I} \leq 2K_{f,m_S^{-1}} \|Q_1 - Q_2\|_1. \tag{D.20}$$

Now we focus on II. Recall that f is locally Lipschitz by assumption, and thus it is Lipschitz on $[0, m_S^{-1}]$. As defined in (6.41), let $L_{f,m_S^{-1}}$ be the Lipschitz constant of f on the latter interval. Hence, (D.16) implies that

$$\left| f\left(\frac{P_{S_1,Y_1}(s,y)}{P_{S_1}(s)P_{Y_1}(y)} \right) - f\left(\frac{P_{S_2,Y_2}(s,y)}{P_{S_2}(s)P_{Y_2}(y)} \right) \right| \leq L_{f,m_S^{-1}} \left| \frac{P_{S_1,Y_1}(s,y)}{P_{S_1}(s)P_{Y_1}(y)} - \frac{P_{S_2,Y_2}(s,y)}{P_{S_2}(s)P_{Y_2}(y)} \right|. \tag{D.21}$$

As a result, we have that

$$\mathrm{II} \leq L_{f,m_S^{-1}} \sum_{s\in\mathcal{S}} \sum_{y\in\mathcal{Y}} \frac{|P_{S_2}(s)P_{Y_2}(y)P_{S_1,Y_1}(s,y) - P_{S_1}(s)P_{Y_1}(y)P_{S_2,Y_2}(s,y)|}{P_{S_1}(s)P_{Y_1}(y)}. \tag{D.22}$$

By the triangle inequality, the numerator of the quotient in (D.22) is upper bounded by

$$P_{S_1,Y_1}(s,y)|P_{S_2}(s)P_{Y_2}(y) - P_{S_1}(s)P_{Y_1}(y)| + P_{S_1}(s)P_{Y_1}(y)|P_{S_1,Y_1}(s,y) - P_{S_2,Y_2}(s,y)|. \tag{D.23}$$

In particular,

$$\mathrm{II} \leq L_{f,m_S^{-1}} \sum_{s\in\mathcal{S}} \sum_{y\in\mathcal{Y}} \frac{P_{S_1,Y_1}(s,y)}{P_{S_1}(s)P_{Y_1}(y)} |P_{S_2}(s)P_{Y_2}(y) - P_{S_1}(s)P_{Y_1}(y)| \tag{D.24}$$

$$+ L_{f,m_S^{-1}} \sum_{s\in\mathcal{S}} \sum_{y\in\mathcal{Y}} |P_{S_1,Y_1}(s,y) - P_{S_2,Y_2}(s,y)| \tag{D.25}$$

$$\leq m_S^{-1} L_{f,m_S^{-1}} \Big(\|P_{Y_1} - P_{Y_2}\|_1 + \|P_{S_1} - P_{S_2}\|_1 \Big) + L_{f,m_S^{-1}} \|P_{S_1,Y_1} - P_{S_2,Y_2}\|_1, \tag{D.26}$$

where the last inequality follows from (D.16) and the argument used in (D.19). Hence, Lemma 33 and Lemma 34 imply that

$$\mathrm{II} \leq \left(2m_S^{-1} + 1\right) L_{f,m_S^{-1}} \|Q_1 - Q_2\|_1. \tag{D.27}$$

Since $\Delta_L \leq \mathrm{I} + \mathrm{II}$, we conclude that

$$\Delta_L \leq \left(2K_{f,m_S^{-1}} + \left(2m_S^{-1} + 1\right) L_{f,m_S^{-1}}\right) \|Q_1 - Q_2\|_1. \tag{D.28}$$

Mutatis mutandis, it can be shown that

$$|\mathcal{U}_f(Q_1, W) - \mathcal{U}_f(Q_2, W)| \leq \left(2K_{f,m_X^{-1}} + \left(2m_X^{-1} + 1\right) L_{f,m_X^{-1}}\right) \|Q_1 - Q_2\|_1. \tag{D.29}$$

The proof is complete. □

D.1.3 Proof of Theorem 13

Proof. By choosing $Q_1 = \hat{P}_n$ and $Q_2 = P$ in Lemma 16, we have

$$|\mathcal{L}_f(\hat{P}_n, W) - \mathcal{L}_f(P, W)| \leq C_{f,m_S} \|\hat{P}_n - P\|_1, \tag{D.30}$$

where

$$m_S = \min\left\{ \left\{ \sum_{x\in\mathcal{X}} \hat{P}_n(s,x) : s \in \mathcal{S} \right\} \cup \left\{ \sum_{x\in\mathcal{X}} P(s,x) : s \in \mathcal{S} \right\} \right\}. \tag{D.31}$$

Observe that, for all $s \in \mathcal{S}$,

$$\sum_{x \in \mathcal{X}} P(s,x) \geq \sum_{x \in \mathcal{X}} \hat{P}_n(s,x) - \sum_{x \in \mathcal{X}} |\hat{P}_n(s,x) - P(s,x)| \tag{D.32}$$

$$\geq \sum_{x \in \mathcal{X}} \hat{P}_n(s,x) - \|\hat{P}_n - P\|_1. \tag{D.33}$$

In particular, we have that

$$m_S \geq \left(\min \left\{ \sum_{x \in \mathcal{X}} \hat{P}_n(s,x) : s \in \mathcal{S} \right\} - \|\hat{P}_n - P\|_1 \right)_+. \tag{D.34}$$

Inequality (6.27) shows that, with probability at least $1 - \beta$,

$$\|\hat{P}_n - P\|_1 \leq \sqrt{\frac{2}{n}\left(|\mathcal{S}| \cdot |\mathcal{X}| - \log \beta\right)}, \tag{D.35}$$

and, thus, $m_S \geq \overline{m}_S$ whenever (D.35) holds true. Recall the definitions of $K_{g,u}$ and $L_{g,u}$ in (6.40) and (6.41), respectively. It is straightforward to verify that the mappings $u \mapsto K_{f,u}$ and $u \mapsto L_{f,u}$ are non-decreasing. Since the mapping $u \mapsto u^{-1}$ is non-increasing, we conclude that

$$u \mapsto C_{f,u} = 2K_{f,u^{-1}} + (2u^{-1} + 1)L_{f,u^{-1}} \tag{D.36}$$

is non-increasing. Therefore, under (D.35), we have

$$C_{f,m_S} \leq C_{f,\overline{m}_S}. \tag{D.37}$$

Combining (D.30), (D.35) and (D.37), we have that, with probability at least $1 - \beta$,

$$|\mathcal{L}_f(\hat{P}_n, W) - \mathcal{L}_f(P, W)| \leq C_{f,\overline{m}_S} \sqrt{\frac{2}{n}\left(|\mathcal{S}| \cdot |\mathcal{X}| - \log \beta\right)}. \tag{D.38}$$

Mutatis mutandis, it can be shown that

$$|\mathcal{U}_f(\hat{P}_n, W) - \mathcal{U}_f(P, W)| \leq C_{f,\overline{m}_X} \sqrt{\frac{2}{n}\left(|\mathcal{S}| \cdot |\mathcal{X}| - \log \beta\right)}, \tag{D.39}$$

whenever (D.35) holds true. □

D.1.4 Proof of Proposition 15

Proof. Assume that $\widetilde{S}$, $\widetilde{X}$ and $\widetilde{X}_0$ satisfy that $P_{\widetilde{S},\widetilde{X}} = \hat{P}_n$ and $\widetilde{X}_0 = \Pi_\gamma(\widetilde{X})$. With this notation, it can be verified that $\widetilde{S} \to \widetilde{X} \to \widetilde{X}_0$ and

$$P_{\widetilde{S},\widetilde{X}_0} = \hat{P}_n P_{\widetilde{X}_0|\widetilde{X}} = \hat{P}_\gamma, \tag{D.40}$$

where, by definition,

$$\left(\hat{P}_n P_{\widetilde{X}_0|\widetilde{X}}\right)(s, x_0) = \sum_{x \in \mathcal{X}} \hat{P}_n(s, x) P_{\widetilde{X}_0|\widetilde{X}}(x_0|x). \tag{D.41}$$

By assumption, $\mathcal{W}^*(\hat{P}_\gamma; \epsilon) \neq \emptyset$. Let $\widetilde{W} \in \mathcal{W}^*(\hat{P}_\gamma; \epsilon)$, i.e.,

$$\mathcal{L}_f(\hat{P}_\gamma, \widetilde{W}) \leq \epsilon \quad \text{and} \quad \mathcal{U}_f(\hat{P}_\gamma, \widetilde{W}) = \mathsf{H}_f(\hat{P}_\gamma; \epsilon). \tag{D.42}$$

Let $\widetilde{Y}_0$ be such that $\widetilde{S} \to \widetilde{X} \to \widetilde{X}_0 \to \widetilde{Y}_0$ and $P_{\widetilde{Y}_0|\widetilde{X}_0} = \widetilde{W}$. Observe that $P_{\widetilde{Y}_0|\widetilde{X}} = P_{\widetilde{X}_0|\widetilde{X}}\widetilde{W}$. In particular,

$$P_{\widetilde{S},\widetilde{Y}_0} = \hat{P}_n P_{\widetilde{Y}_0|\widetilde{X}} = \hat{P}_n P_{\widetilde{X}_0|\widetilde{X}}\widetilde{W} = \hat{P}_\gamma \widetilde{W}, \tag{D.43}$$

where the last equality follows from (D.40). Therefore,

$$\mathcal{L}_f(\hat{P}_\gamma, \widetilde{W}) = I_f(P_{\widetilde{S},\widetilde{Y}_0}) = \mathcal{L}_f(\hat{P}_n, P_{\widetilde{Y}_0|\widetilde{X}}). \tag{D.44}$$

Furthermore, Lemma 17 implies that

$$\mathcal{U}_f(\hat{P}_\gamma, \widetilde{W}) = I_f(P_{\widetilde{X}_0,\widetilde{Y}_0}) = I_f(P_{\widetilde{X},\widetilde{Y}_0}) = \mathcal{U}_f(\hat{P}_n, P_{\widetilde{Y}_0|\widetilde{X}}). \tag{D.45}$$

By (D.42), (D.44) and (D.45), we have

$$\mathcal{L}_f(\hat{P}_n, P_{\widetilde{Y}_0|\widetilde{X}}) \leq \epsilon \quad \text{and} \quad \mathcal{U}_f(\hat{P}_n, P_{\widetilde{Y}_0|\widetilde{X}}) = \mathsf{H}_f(\hat{P}_\gamma; \epsilon). \tag{D.46}$$

Recall that, by definition, $\mathsf{H}_f(\hat{P}_n; \epsilon) = \sup_W \mathcal{U}_f(\hat{P}_n, W)$ where the supremum is taken over all $W \in \mathcal{W}$ such that $\mathcal{L}_f(\hat{P}_n, W) \leq \epsilon$. Thus, (D.46) implies that

$$\mathsf{H}_f(\hat{P}_n; \epsilon) \geq \mathcal{U}_f(\hat{P}_n, P_{\widetilde{Y}_0|\widetilde{X}}) = \mathsf{H}_f(\hat{P}_\gamma; \epsilon), \tag{D.47}$$

as we wanted to prove. □

D.1.5 Proof of Lemma 18

Proof. For ease of notation, let $\Delta_L \triangleq |\mathcal{L}_\alpha^{\mathrm{A}}(Q_1, W) - \mathcal{L}_\alpha^{\mathrm{A}}(Q_2, W)|$. For each $i \in \{1, 2\}$, let $S_i \to X_i \to Y_i$ be such that $P_{S_i,X_i} = Q_i$ and $P_{Y_i|X_i} = W$. With this notation,

$$\Delta_L = |I_\alpha^{\mathrm{A}}(P_{S_1,Y_1}) - I_\alpha^{\mathrm{A}}(P_{S_2,Y_2})| \tag{D.48}$$

$$= \frac{\alpha}{\alpha - 1} \left| \log \frac{\sum_y \|P_{S_1,Y_1}(\cdot, y)\|_\alpha}{\sum_y \|P_{S_2,Y_2}(\cdot, y)\|_\alpha} - \log \frac{\|P_{S_1}(\cdot)\|_\alpha}{\|P_{S_2}(\cdot)\|_\alpha} \right|, \tag{D.49}$$

where the second equality follows directly from the definition of Arimoto's mutual information. By the triangle inequality, we obtain that $\frac{\alpha-1}{\alpha}\Delta_L \leq \mathrm{I} + \mathrm{II}$ where

$$\mathrm{I} \triangleq \left| \log \frac{\sum_y \|P_{S_1,Y_1}(\cdot,y)\|_\alpha}{\sum_y \|P_{S_2,Y_2}(\cdot,y)\|_\alpha} \right|, \tag{D.50}$$

$$\mathrm{II} \triangleq \left| \log \frac{\|P_{S_1}(\cdot)\|_\alpha}{\|P_{S_2}(\cdot)\|_\alpha} \right|. \tag{D.51}$$

Notice that if $a, b > 0$, then $\left|\log \frac{a}{b}\right| \leq \frac{|a-b|}{\min\{a,b\}}$. Hence,

$$\mathrm{I} \leq \frac{\left|\sum_y \|P_{S_1,Y_1}(\cdot,y)\|_\alpha - \sum_y \|P_{S_2,Y_2}(\cdot,y)\|_\alpha\right|}{\min\{\sum_y \|P_{S_1,Y_1}(\cdot,y)\|_\alpha, \sum_y \|P_{S_2,Y_2}(\cdot,y)\|_\alpha\}}. \tag{D.52}$$

Observe that, by the triangle inequality,

$$\left| \sum_{y\in\mathcal{Y}} \|P_{S_1,Y_1}(\cdot,y)\|_\alpha - \sum_{y\in\mathcal{Y}} \|P_{S_2,Y_2}(\cdot,y)\|_\alpha \right| \leq \sum_{y\in\mathcal{Y}} \left| \|P_{S_1,Y_1}(\cdot,y)\|_\alpha - \|P_{S_2,Y_2}(\cdot,y)\|_\alpha \right|. \tag{D.53}$$

By Minkowski's inequality, $|\|a\|_\alpha - \|b\|_\alpha| \leq \|a-b\|_\alpha$ for every $a, b \in \mathbb{R}^n$ and $\alpha \geq 1$. Therefore,

$$\left| \sum_{y\in\mathcal{Y}} \|P_{S_1,Y_1}(\cdot,y)\|_\alpha - \sum_{y\in\mathcal{Y}} \|P_{S_2,Y_2}(\cdot,y)\|_\alpha \right| \leq \sum_{y\in\mathcal{Y}} \|P_{S_1,Y_1}(\cdot,y) - P_{S_2,Y_2}(\cdot,y)\|_\alpha \tag{D.54}$$

$$\leq \sum_{y\in\mathcal{Y}} \|P_{S_1,Y_1}(\cdot,y) - P_{S_2,Y_2}(\cdot,y)\|_1, \tag{D.55}$$

$$= \|P_{S_1,Y_1} - P_{S_2,Y_2}\|_1, \tag{D.56}$$

where the second inequality follows from the fact that $\|a\|_\alpha \leq \|a\|_1$ for all $a \in \mathbb{R}^n$ and $\alpha \geq 1$. By assumption, we have that $P_{Y_1|X_1} = W = P_{Y_2|X_2}$. Hence, Lemma 33 implies that

$$\|P_{S_1,Y_1} - P_{S_2,Y_2}\|_1 \leq \|Q_1 - Q_2\|_1. \tag{D.57}$$

This leads to

$$\left| \sum_{y\in\mathcal{Y}} \|P_{S_1,Y_1}(\cdot,y)\|_\alpha - \sum_{y\in\mathcal{Y}} \|P_{S_2,Y_2}(\cdot,y)\|_\alpha \right| \leq \|Q_1 - Q_2\|_1. \tag{D.58}$$

For $i \in \{1,2\}$, Minkowski's inequality implies that

$$\sum_{y\in\mathcal{Y}} \|P_{S_i,Y_i}(\cdot,y)\|_\alpha \geq \left\| \sum_{y\in\mathcal{Y}} P_{S_i,Y_i}(\cdot,y) \right\|_\alpha = \|P_{S_i}(\cdot)\|_\alpha. \tag{D.59}$$

The generalized mean inequality implies that

$$\|P_{S_i}(\cdot)\|_\alpha \geq |\mathcal{S}|^{1/\alpha-1}\|P_{S_i}(\cdot)\|_1 = |\mathcal{S}|^{1/\alpha-1}, \tag{D.60}$$

as $\|P_{S_i}(\cdot)\|_1 = \sum_s P_{S_i}(s) = 1$. Hence,

$$\sum_{y\in\mathcal{Y}} \|P_{S_i,Y_i}(\cdot,y)\|_\alpha \geq |\mathcal{S}|^{1/\alpha-1}. \tag{D.61}$$

Therefore, by plugging (D.58) and (D.61) in (D.52),

$$\mathrm{I} \leq |\mathcal{S}|^{1-1/\alpha}\|Q_1 - Q_2\|_1. \tag{D.62}$$

Similarly, we have that

$$\mathrm{II} \leq \frac{|\|P_{S_2}(\cdot)\|_\alpha - \|P_{S_1}(\cdot)\|_\alpha|}{\min\{\|P_{S_2}(\cdot)\|_\alpha, \|P_{S_1}(\cdot)\|_\alpha\}}. \tag{D.63}$$

As in (D.56) and (D.57),

$$|\|P_{S_2}(\cdot)\|_\alpha - \|P_{S_1}(\cdot)\|_\alpha| \leq \|P_{S_1} - P_{S_2}\|_1 \leq \|Q_1 - Q_2\|_1. \tag{D.64}$$

As established in (D.61), for all $i \in \{1,2\}$,

$$\|P_{S_i}(\cdot)\|_\alpha \geq |\mathcal{S}|^{1/\alpha-1}. \tag{D.65}$$

Thus, $\mathrm{II} \leq |\mathcal{S}|^{1-1/\alpha}\|Q_1 - Q_2\|_1$. Since $\frac{\alpha-1}{\alpha}\Delta_L \leq \mathrm{I} + \mathrm{II}$, we conclude that

$$|\mathcal{L}_\alpha^{\mathrm{A}}(Q_1,W) - \mathcal{L}_\alpha^{\mathrm{A}}(Q_2,W)| \leq \frac{2\alpha}{\alpha-1}|\mathcal{S}|^{1-1/\alpha}\|Q_1 - Q_2\|_1. \tag{D.66}$$

Mutatis mutandis, one can obtain that

$$|\mathcal{U}_\alpha^{\mathrm{A}}(Q_1,W) - \mathcal{U}_\alpha^{\mathrm{A}}(Q_2,W)| \leq \frac{2\alpha}{\alpha-1}|\mathcal{X}|^{1-1/\alpha}\|Q_1 - Q_2\|_1, \tag{D.67}$$

as required □

D.1.6 Proof of Lemma 19

Consider the following lemma.

Lemma 35. *If S_1, S_2 and X_1, X_2 are random variables supported over $\mathcal{S}$ and $\mathcal{X}$ respectively, then,*

$$\|P_{S_1}\cdot P_{X_1|S_1} - P_{S_1}\cdot P_{X_2|S_2}\|_1 \leq 2\|P_{S_1,X_1} - P_{S_2,X_2}\|_1. \tag{D.68}$$

Proof. For ease of notation, let $\Delta \triangleq \|P_{S_1} \cdot P_{X_1|S_1} - P_{S_1} \cdot P_{X_2|S_2}\|_1$. By the triangle inequality, we have that

$$\Delta \leq \|P_{S_1} \cdot P_{X_1|S_1} - P_{S_2} \cdot P_{X_2|S_2}\|_1 + \|P_{S_2} \cdot P_{X_2|S_2} - P_{S_1} \cdot P_{X_2|S_2}\|_1. \tag{D.69}$$

By Lemma 34, and the fact that $P_{S_i} \cdot P_{X_i|S_i} = P_{S_i,X_i}$ for every $i \in \{1,2\}$,

$$\Delta \leq \|P_{S_1,X_1} - P_{S_2,X_2}\|_1 + \|P_{S_1} - P_{S_2}\|_1 \leq 2\|P_{S_1,X_1} - P_{S_2,X_2}\|_1, \tag{D.70}$$

as required. □

Now we are in position to prove Lemma 19.

Proof. For ease of notation, let $\Delta_L \triangleq |\mathcal{L}_\alpha^{\mathsf{S}}(Q_1, W) - \mathcal{L}_\alpha^{\mathsf{S}}(Q_2, W)|$. For each $i \in \{1,2\}$, let $S_i \to X_i \to Y_i$ be such that $P_{S_i,X_i} = Q_i$ and $P_{Y_i|X_i} = W$. Assume that $\alpha \in (1,\infty)$. Then

$$\Delta_L = |I_\alpha^{\mathsf{S}}(P_{S_1,Y_1}) - I_\alpha^{\mathsf{S}}(P_{S_2,Y_2})| \tag{D.71}$$

$$= \frac{\alpha}{\alpha - 1} \left| \log \frac{\sum_y \|P_{S_1}(\cdot)^{1/\alpha} P_{Y_1|S_1}(y|\cdot)\|_\alpha}{\sum_y \|P_{S_2}(\cdot)^{1/\alpha} P_{Y_2|S_2}(y|\cdot)\|_\alpha} \right|, \tag{D.72}$$

where the second equality follows directly from the definition of Sibson's mutual information. Notice that if $a, b > 0$, then $\left|\log \frac{a}{b}\right| \leq \frac{|a-b|}{\min\{a,b\}}$. Hence,

$$\Delta_L \leq \frac{\alpha}{\alpha - 1} \frac{|\sum_y \|P_{S_1}(\cdot)^{1/\alpha} P_{Y_1|S_1}(y|\cdot)\|_\alpha - \sum_y \|P_{S_2}(\cdot)^{1/\alpha} P_{Y_2|S_2}(y|\cdot)\|_\alpha|}{\min\{\sum_y \|P_{S_1}(\cdot)^{1/\alpha} P_{Y_1|S_1}(y|\cdot)\|_\alpha, \sum_y \|P_{S_2}(\cdot)^{1/\alpha} P_{Y_2|S_2}(y|\cdot)\|_\alpha\}}. \tag{D.73}$$

By Minkowski's inequality, we have that

$$\sum_{y \in \mathcal{Y}} \|P_{S_i}(\cdot)^{1/\alpha} P_{Y_i|S_i}(y|\cdot)\|_\alpha \geq \|P_{S_i}(\cdot)^{1/\alpha}\|_\alpha = 1, \tag{D.74}$$

where we used the fact that $\sum_y P_{Y_i|S_i}(y|s) = 1$ for every $s \in \mathcal{S}$ and $i \in \{1,2\}$. Hence, (D.73) becomes

$$\Delta_L \leq \frac{\alpha}{\alpha - 1} \sum_{y \in \mathcal{Y}} \left| \|P_{S_1}(\cdot)^{1/\alpha} P_{Y_1|S_1}(y|\cdot)\|_\alpha - \|P_{S_2}(\cdot)^{1/\alpha} P_{Y_2|S_2}(y|\cdot)\|_\alpha \right| \tag{D.75}$$

$$\leq \frac{\alpha}{\alpha - 1} \sum_{y \in \mathcal{Y}} \|P_{S_1}(\cdot)^{1/\alpha} P_{Y_1|S_1}(y|\cdot) - P_{S_2}(\cdot)^{1/\alpha} P_{Y_2|S_2}(y|\cdot)\|_\alpha, \tag{D.76}$$

where the last inequality follows from another application of Minkowski's inequality. Since $\|a\|_\alpha \leq \|a\|_1$ for all $a \in \mathbb{R}^n$ and $\alpha > 1$, we obtain that

$$\Delta_L \leq \frac{\alpha}{\alpha - 1} \sum_{y \in \mathcal{Y}} \|P_{S_1}(\cdot)^{1/\alpha} P_{Y_1|S_1}(y|\cdot) - P_{S_2}(\cdot)^{1/\alpha} P_{Y_2|S_2}(y|\cdot)\|_1. \tag{D.77}$$

Observe that, for each $i \in \{1,2\}$,

$$P_{S_i}(s)^{1/\alpha} P_{Y_i|S_i}(y|s) = \sum_{x\in\mathcal{X}} P_{S_i}(s)^{1/\alpha} P_{X_i|S_i}(x|s) W(x,y). \tag{D.78}$$

Thus, a straightforward manipulation leads to

$$\Delta_L \leq \frac{\alpha}{\alpha-1} \sum_{y\in\mathcal{Y}} \sum_{x\in\mathcal{X}} \|P_{S_1}(\cdot)^{1/\alpha} P_{X_1|S_1}(x|\cdot) - P_{S_2}(\cdot)^{1/\alpha} P_{X_2|S_2}(x|\cdot)\|_1 W(x,y) \tag{D.79}$$

$$= \frac{\alpha}{\alpha-1} \sum_{x\in\mathcal{X}} \|P_{S_1}(\cdot)^{1/\alpha} P_{X_1|S_1}(x|\cdot) - P_{S_2}(\cdot)^{1/\alpha} P_{X_2|S_2}(x|\cdot)\|_1. \tag{D.80}$$

By adding and subtracting the term $P_{S_1}(\cdot)^{1/\alpha} P_{X_2|S_2}(x|\cdot)$ inside the norm in (D.80), Minkowski's inequality implies that $\frac{\alpha-1}{\alpha}\Delta_L \leq \sum_x \mathrm{I}_x + \sum_x \mathrm{II}_x$ where, for each $x \in \mathcal{X}$,

$$\mathrm{I}_x \triangleq \|P_{S_1}(\cdot)^{1/\alpha}(P_{X_1|S_1}(x|\cdot) - P_{X_2|S_2}(x|\cdot))\|_1, \tag{D.81}$$

$$\mathrm{II}_x \triangleq \|(P_{S_1}(\cdot)^{1/\alpha} - P_{S_2}(\cdot)^{1/\alpha}) P_{X_2|S_2}(x|\cdot)\|_1. \tag{D.82}$$

Observe that, for each $x \in \mathcal{X}$,

$$\mathrm{I}_x = \sum_{s\in\mathcal{S}} P_{S_1}(s)^{1/\alpha} |P_{X_1|S_1}(x|s) - P_{X_2|S_2}(x|s)| \tag{D.83}$$

$$= \sum_{s\in\mathcal{S}} \frac{P_{S_1}(s)}{P_{S_1}(s)^{1-1/\alpha}} |P_{X_1|S_1}(x|s) - P_{X_2|S_2}(x|s)|. \tag{D.84}$$

Since $P_{S_1}(s) \geq m_S$ for all $s \in \mathcal{S}$, we obtain that

$$\sum_{x\in\mathcal{X}} \mathrm{I}_x \leq \frac{1}{m_S^{1-1/\alpha}} \sum_{x\in\mathcal{X}} \sum_{s\in\mathcal{S}} P_{S_1}(s) |P_{X_1|S_1}(x|s) - P_{X_2|S_2}(x|s)| \tag{D.85}$$

$$= \frac{\|P_{S_1}\cdot P_{X_1|S_1} - P_{S_1}\cdot P_{X_2|S_2}\|_1}{m_S^{1-1/\alpha}}. \tag{D.86}$$

Therefore, Lemma 35 leads to

$$\sum_{x\in\mathcal{X}} \mathrm{I}_x \leq \frac{2\|Q_1 - Q_2\|_1}{m_S^{1-1/\alpha}}. \tag{D.87}$$

By definition, $\mathrm{II}_x = \sum_s |P_{S_1}(s)^{1/\alpha} - P_{S_2}(s)^{1/\alpha}| P_{X_2|S_2}(x|s)$ for every $x \in \mathcal{X}$. Hence,

$$\sum_{x\in\mathcal{X}} \mathrm{II}_x = \sum_{s\in\mathcal{S}} |P_{S_1}(\cdot)^{1/\alpha} - P_{S_2}(\cdot)^{1/\alpha}|. \tag{D.88}$$

Since the function $t \mapsto t^{1/\alpha}$ is Lipschitz continuous on $[m_S, 1]$ with Lipschitz constant $(\alpha m_S^{1-1/\alpha})^{-1}$,

$$\sum_{x\in\mathcal{X}} \mathrm{II}_x \leq \frac{\|P_{S_1} - P_{S_2}\|_1}{\alpha m_S^{1-1/\alpha}} \leq \frac{\|Q_1 - Q_2\|_1}{\alpha m_S^{1-1/\alpha}}, \tag{D.89}$$

where the last inequality follows from Lemma 34. By (D.87) and (D.89), we conclude that

$$\Delta_L \leq \frac{2\alpha+1}{\alpha-1}\frac{\|Q_1-Q_2\|_1}{m_S^{1-1/\alpha}}. \tag{D.90}$$

Mutatis mutandis, it can be shown that

$$|\mathcal{U}_\alpha^{\mathrm{S}}(Q_1,W)-\mathcal{U}_\alpha^{\mathrm{S}}(Q_2,W)| \leq \frac{1}{\alpha-1}\frac{\|Q_1-Q_2\|_1}{m_X^{1-1/\alpha}}. \tag{D.91}$$

Recall that $\lim_{\alpha\to\infty} I_\alpha^{\mathrm{S}}(P_{X,Y}) = I_\infty^{\mathrm{S}}(P_{X,Y})$ [276]. Therefore, (6.70) follows after taking the limit $\alpha\to\infty$ in both sides of (D.91). Using a similar argument as above, it can be shown that

$$\begin{aligned}
\Delta_L &\leq \sum_{s\in\mathcal{S}}\sum_{x\in\mathcal{X}} |P_{X_1|S_1}(x|s)-P_{X_2|S_2}(x|s)| && \text{(D.92)}\\
&\leq \frac{1}{\min_{s'} P_{S_1}(s')}\sum_{s\in\mathcal{S}}\sum_{x\in\mathcal{X}} P_{S_1}(s)|P_{X_1|S_1}(x|s)-P_{X_2|S_2}(x|s)| && \text{(D.93)}\\
&= \frac{\|P_{S_1}\cdot P_{X_1|S_1}-P_{S_1}\cdot P_{X_2|S_2}\|_1}{\min_{s'} P_{S_1}(s')}. && \text{(D.94)}
\end{aligned}$$

By applying Lemma 35 and noting that $\min_{s'} P_{S_1}(s') = \min_{s'}\sum_x Q_1(s',x)$, the result follows. □

D.1.7 Proof of Lemma 20

Proof. For ease of notation, let $\Delta_L \triangleq |\mathcal{L}_\alpha^{\max}(Q_1,W)-\mathcal{L}_\alpha^{\max}(Q_2,W)|$. For each $i\in\{1,2\}$, let $S_i\to X_i\to Y_i$ be such that $P_{S_i,X_i}=Q_i$ and $P_{Y_i|X_i}=W$. Observe that $P_{Y_i|S_i}=P_{X_i|S_i}W$ where, by definition,

$$\left(P_{X_i|S_i}W\right)(y|s) = \sum_{x\in\mathcal{X}} P_{X_i|S_i}(x|s)W(x,y). \tag{D.95}$$

In particular,

$$\begin{aligned}
\mathcal{L}_\alpha^{\max}(Q_i,W) &\triangleq \sup_{P_{\widetilde{S}}} I_\alpha^{\mathrm{A}}(P_{\widetilde{S}}\cdot P_{Y_i|S_i}) && \text{(D.96)}\\
&= \sup_{P_{\widetilde{S}}} I_\alpha^{\mathrm{A}}(P_{\widetilde{S}}\cdot P_{X_i|S_i}W). && \text{(D.97)}
\end{aligned}$$

By the definition of $\mathcal{L}_\alpha^{\mathrm{A}}(Q,W)$, we have that

$$\mathcal{L}_\alpha^{\max}(Q_i,W) = \sup_{P_{\widetilde{S}}} \mathcal{L}_\alpha^{\mathrm{A}}(P_{\widetilde{S}}\cdot P_{X_i|S_i},W). \tag{D.98}$$

It is easy to verify that if $\mathcal{I}$ is an arbitrary index set and $a_\iota, b_\iota\in\mathbb{R}$ for all $\iota\in\mathcal{I}$, then

$$\left|\sup_{\iota\in\mathcal{I}} a_\iota - \sup_{\iota\in\mathcal{I}} b_\iota\right| \leq \sup_{\iota\in\mathcal{I}} |a_\iota-b_\iota|. \tag{D.99}$$

Therefore, (D.98) implies that

$$\Delta_L = \left| \sup_{P_{\widetilde{S}}} \mathcal{L}_\alpha^{\mathrm{A}}(P_{\widetilde{S}} \cdot P_{X_1|S_1}, W) - \sup_{P_{\widetilde{S}}} \mathcal{L}_\alpha^{\mathrm{A}}(P_{\widetilde{S}} \cdot P_{X_2|S_2}, W) \right| \tag{D.100}$$

$$\leq \sup_{P_{\widetilde{S}}} \left| \mathcal{L}_\alpha^{\mathrm{A}}(P_{\widetilde{S}} \cdot P_{X_1|S_1}, W) - \mathcal{L}_\alpha^{\mathrm{A}}(P_{\widetilde{S}} \cdot P_{X_2|S_2}, W) \right|. \tag{D.101}$$

By Lemma 18, we obtain that

$$\Delta_L \leq \frac{2\alpha}{\alpha - 1} |\mathcal{S}|^{1-1/\alpha} \sup_{P_{\widetilde{S}}} \|P_{\widetilde{S}} \cdot P_{X_1|S_1} - P_{\widetilde{S}} \cdot P_{X_2|S_2}\|_1. \tag{D.102}$$

A straightforward computation shows that, for any $P_{\widetilde{S}}$,

$$\|P_{\widetilde{S}} \cdot P_{X_1|S_1} - P_{\widetilde{S}} \cdot P_{X_2|S_2}\|_1 = \sum_{s \in \mathcal{S}} P_{\widetilde{S}}(s) \|P_{X_1|S_1}(\cdot|s) - P_{X_2|S_2}(\cdot|s)\|_1 \tag{D.103}$$

$$\leq \sum_{s \in \mathcal{S}} \|P_{X_1|S_1}(\cdot|s) - P_{X_2|S_2}(\cdot|s)\|_1 \tag{D.104}$$

$$\leq \frac{1}{\min_{s'} P_{S_1}(s')} \sum_{s \in \mathcal{S}} P_{S_1}(s) \|P_{X_1|S_1}(\cdot|s) - P_{X_2|S_2}(\cdot|s)\|_1 \tag{D.105}$$

$$= \frac{\|P_{S_1} \cdot P_{X_1|S_1} - P_{S_1} \cdot P_{X_2|S_2}\|_1}{\min_{s'} P_{S_1}(s')}. \tag{D.106}$$

Therefore, Lemma 35 implies that

$$\sup_{P_{\widetilde{S}}} \|P_{\widetilde{S}} \cdot P_{X_1|S_1} - P_{\widetilde{S}} \cdot P_{X_2|S_2}\|_1 \leq \frac{2\|Q_1 - Q_2\|}{\min_{s'} P_{S_1}(s')}. \tag{D.107}$$

By (D.102) and (D.107), we conclude that

$$\Delta_L \leq \frac{4\alpha}{\alpha - 1} \frac{|\mathcal{S}|^{1-1/\alpha}}{\min_{s'} P_{S_1}(s')} \|Q_1 - Q_2\|_1. \tag{D.108}$$

By noting that $\min_{s'} P_{S_1}(s') = \min_{s'} \sum_x Q_1(s', x)$, (6.74) follows. Finally, note that $\mathcal{U}_\alpha^{\max}(Q, W)$ only depends on W. Therefore, $\mathcal{U}_\alpha^{\max}(Q_1, W) = \mathcal{U}_\alpha^{\max}(Q_2, W)$ which trivially leads to (6.75). □

D.2 Proofs for Section 6.5

D.2.1 Proof of Lemma 21

Lemma 21 is an immediate consequence of the following standard result whose proof is included for the sake of completeness.

Lemma 36. *Let $(\mathcal{A}, d_{\mathcal{A}})$ and $(\mathcal{B}, d_{\mathcal{B}})$ be two metric spaces. If $f : \mathcal{A} \times \mathcal{B} \to \mathbb{R}$ satisfies that*

(i) $|f(a_1, b) - f(a_2, b)| \leq d_{\mathcal{A}}(a_1, a_2)$ *for all* $a_1, a_2 \in \mathcal{A}$ *and* $b \in \mathcal{B}$;

(ii) $f(a, \cdot)$ *is continuous over* $\mathcal{B}$ *for all* $a \in \mathcal{A}$;

then $f(\cdot, \cdot)$ *is continuous over* $\mathcal{A} \times \mathcal{B}$.

Proof. In order to prove the continuity of $f(\cdot, \cdot)$ over $\mathcal{A} \times \mathcal{B}$, we will show that for any sequence $\{(a_n, b_n)\}_{n=0}^{\infty}$ such that $\lim_n (a_n, b_n) = (a_0, b_0)$, it holds true that $\lim_n f(a_n, b_n) = f(a_0, b_0)$.

Let $\{(a_n, b_n)\}_{n=0}^{\infty}$ be such that $\lim_n (a_n, b_n) = (a_0, b_0)$. By the triangle inequality, for every $n \in \mathbb{N}$,

$$|f(a_n, b_n) - f(a_0, b_0)| \leq |f(a_n, b_n) - f(a_0, b_n)| + |f(a_0, b_n) - f(a_0, b_0)| \tag{D.109}$$

$$\leq d_{\mathcal{A}}(a_n, a_0) + |f(a_0, b_n) - f(a_0, b_0)|, \tag{D.110}$$

where the last inequality follows from assumption (i). By assumption (ii), $f(a_0, \cdot)$ is continuous over $\mathcal{B}$. Since $\lim_n (a_n, b_n) = (a_0, b_0)$ is equivalent to $\lim_n a_n = a_0$ and $\lim_n b_n = b_0$, we conclude that

$$\lim_{n \to \infty} |f(a_n, b_n) - f(a_0, b_0)| = 0, \tag{D.111}$$

as required. □

D.2.2 Proof of Proposition 16

We start recalling an elementary fact about convergent sequences in metric spaces, see, e.g., [233, Exercise 2.4.11].

Theorem 20. *Let* $\mathcal{M}$ *be a metric space. A sequence* $(a_n)_{n=1}^{\infty} \subset \mathcal{M}$ *converges to* $a \in \mathcal{M}$ *if and only if every subsequence of* $(a_n)_{n=1}^{\infty}$ *has a further subsequence which converges to* a.

We proceed with the proof of Proposition 16.

Proof. In order to prove the continuity of $\mathsf{H}(\cdot; \epsilon)$ over $\{Q \in \mathcal{Q} : \epsilon_{\min}(Q) < \epsilon\}$, we will show that for any sequence $(Q_n)_{n=0}^{\infty} \subset \{Q \in \mathcal{Q} : \epsilon_{\min}(Q) < \epsilon\}$ such that

$$\lim_{n \to \infty} \|Q_n - Q_0\|_1 = 0, \tag{D.112}$$

it holds true that

$$\lim_{n \to \infty} \mathsf{H}(Q_n; \epsilon) = \mathsf{H}(Q_0; \epsilon). \tag{D.113}$$

In order to prove (D.113), Theorem 20 implies that it is enough to show that any subsequence $(\mathsf{H}(Q_{n_k}; \epsilon))_{k=1}^{\infty}$ has a further subsequence converging to $\mathsf{H}(Q_0; \epsilon)$.

Let $(n_k)_{k=1}^{\infty} \subset \mathbb{N}$ be given. By Remark 14, there exists $(W_{n_k}^*)_{k=1}^{\infty} \subset \mathcal{W}_N$ such that, for each $k \geq 1$, we have that $\mathcal{L}(Q_{n_k}, W_{n_k}^*) \leq \epsilon$ and

$$\mathsf{H}(Q_{n_k}; \epsilon) = \mathcal{U}(Q_{n_k}, W_{n_k}^*). \tag{D.114}$$

Since the space $\mathcal{W}_N$ is compact, there exists a further subsequence $(n_{k_j})_{j=1}^{\infty}$ such that

$$\lim_{j \to \infty} W_{n_{k_j}}^* = W_0, \tag{D.115}$$

for some $W_0 \in \mathcal{W}_N$. Lemma 21 shows that $\mathcal{U}(\cdot, \cdot)$ is continuous over $\mathcal{Q} \times \mathcal{W}_N$. Then (D.112) and (D.115) imply that

$$\mathcal{U}(Q_0, W_0) = \lim_{j \to \infty} \mathcal{U}(Q_{n_{k_j}}, W_{n_{k_j}}^*) = \lim_{j \to \infty} \mathsf{H}(Q_{n_{k_j}}; \epsilon), \tag{D.116}$$

where the second equality follows from (D.114). A similar reasoning shows that

$$\mathcal{L}(Q_0, W_0) = \lim_{j \to \infty} \mathcal{L}(Q_{n_{k_j}}, W_{n_{k_j}}^*) \leq \epsilon. \tag{D.117}$$

Therefore, by the maximality of $\mathsf{H}(Q_0; \epsilon)$,

$$\lim_{j \to \infty} \mathsf{H}(Q_{n_{k_j}}; \epsilon) = \mathcal{U}(Q_0, W_0) \leq \mathsf{H}(Q_0; \epsilon). \tag{D.118}$$

Next, we show that $\lim_j \mathsf{H}(Q_{n_{k_j}}; \epsilon) \geq \mathsf{H}(Q_0; \epsilon)$. Recall that, by assumption, $\epsilon_{\min}(Q_0) < \epsilon$. Let $\delta > 0$ be such that $\epsilon_{\min}(Q_0) \leq \epsilon - \delta$. By condition (**C.**3), there exists $W_0' \in \mathcal{W}_N$ such that $\mathcal{L}(Q_0, W_0') \leq \epsilon - \delta$ and $\mathcal{U}(Q_0, W_0') = \mathsf{H}(Q_0; \epsilon - \delta)$. By the Lipschitz continuity given in condition (**C.**2) and (D.112),

$$\lim_{j \to \infty} \mathcal{L}(Q_{n_{k_j}}, W_0') = \mathcal{L}(Q_0, W_0') \leq \epsilon - \delta, \tag{D.119}$$

$$\lim_{j \to \infty} \mathcal{U}(Q_{n_{k_j}}, W_0') = \mathcal{U}(Q_0, W_0') = \mathsf{H}(Q_0; \epsilon - \delta). \tag{D.120}$$

In particular, for j large enough we have that $\mathcal{L}(Q_{n_{k_j}}, W_0') \leq \epsilon$ and, by the maximality of $\mathsf{H}(Q_{n_{k_j}}; \epsilon)$,

$$\lim_{j \to \infty} \mathsf{H}(Q_{n_{k_j}}; \epsilon) \geq \lim_{j \to \infty} \mathcal{U}(Q_{n_{k_j}}, W_0') = \mathsf{H}(Q_0; \epsilon - \delta), \tag{D.121}$$

where the last equality follows from (D.120). Recall that, by condition (**C.**1), the mapping $\mathsf{H}(Q_0; \cdot)$ is continuous over $[\epsilon_{\min}(Q_0), \infty)$. Hence, by taking limits in (D.121),

$$\lim_{j \to \infty} \mathsf{H}(Q_{n_{k_j}}; \epsilon) \geq \lim_{\delta \downarrow 0} \mathsf{H}(Q_0; \epsilon - \delta) = \mathsf{H}(Q_0; \epsilon). \tag{D.122}$$

By combining (D.118) and (D.122), we conclude that

$$\lim_{j\to\infty} \mathsf{H}(Q_{n_{k_j}};\epsilon) = \mathsf{H}(Q_0;\epsilon), \tag{D.123}$$

as required. □

D.2.3 Proof of Corollary 5

Before we prove Corollary 5, let us recall two standard results from analysis on metric spaces, see, e.g., Chapter 0.1 in [231] and Lemma 14 in [261], respectively.

Lemma 37. *Let $f_n : [a,b] \to \mathbb{R}$ be a non-decreasing function for each $n \in \mathbb{N}$. If there exists a continuous function $f : [a,b] \to \mathbb{R}$ such that $\lim_{n\to\infty} f_n(x) = f(x)$ for all $x \in [a,b]$, then f_n converges uniformly to f.*

Lemma 38. *Let $\mathcal{A}$ and $\mathcal{B}$ be two metric spaces with $\mathcal{B}$ compact. If $f : \mathcal{A} \times \mathcal{B} \to \mathbb{R}$ is continuous, then the function $g : \mathcal{A} \to \mathbb{R}$ defined by $g(a) \triangleq \inf_{b\in\mathcal{B}} f(a,b)$ is also continuous.*

Observe that the infimum in the previous lemma can be replaced by a supremum. We proceed with the proof of Corollary 5.

Proof. We first introduce

$$\epsilon_{\max}(Q) \triangleq \max\{\mathcal{L}(Q,W) :\ W \in \mathcal{W}_N\}. \tag{D.124}$$

Observe that, by the maximality of $\epsilon_{\max}(Q)$, for all $\epsilon \geq \epsilon_{\max}(Q)$,

$$\mathsf{H}(Q;\epsilon) = \mathsf{H}(Q;\epsilon_{\max}(Q)). \tag{D.125}$$

By Lemma 21 the function $\mathcal{L}(\cdot,\cdot)$ is continuous over $\mathcal{Q} \times \mathcal{W}_N$. Since $\mathcal{W}_N$ is compact, Lemma 38 implies that $\epsilon_{\max}(\cdot)$ is continuous over $\mathcal{Q}$. By hypothesis $\lim_n P_n = P$, thus $\lim_n \epsilon_{\max}(P_n) = \epsilon_{\max}(P)$ and, in particular, there exists $\epsilon_1 > \epsilon_0$ such that $\epsilon_{\max}(P_n) \leq \epsilon_1$ for all $n \in \mathbb{N}$. By (D.125), for all $\epsilon \geq \epsilon_1$ and all $n \in \mathbb{N}$,

$$\mathsf{H}(P_n;\epsilon) = \mathsf{H}(P_n;\epsilon_{\max}(P_n)) = \mathsf{H}(P_n;\epsilon_1). \tag{D.126}$$

Observe that an analogous equality holds for P. Hence, we obtain that, for all $n \in \mathbb{N}$,

$$\sup_{\epsilon\geq\epsilon_1} |\mathsf{H}(P_n;\epsilon) - \mathsf{H}(P;\epsilon)| = |\mathsf{H}(P_n;\epsilon_1) - \mathsf{H}(P_n;\epsilon_1)|. \tag{D.127}$$

As established by Proposition 16, the right hand side of (D.127) vanishes as n goes to infinity. Hence, we conclude that $\mathsf{H}(P_n;\cdot)$ converge uniformly to $\mathsf{H}(P;\cdot)$ over $[\epsilon_1,\infty)$. Finally, recall that the

function $\mathsf{H}(P_n;\cdot)$ is non-decreasing over $[\epsilon_0,\epsilon_1]$ for each $n\in\mathbb{N}$. Furthermore, by condition (**C.**1), the function $\mathsf{H}(P;\cdot)$ is continuous over $[\epsilon_0,\epsilon_1]$. Therefore, Proposition 16 and Lemma 37 imply that $\mathsf{H}(P_n;\cdot)$ converges uniformly to $\mathsf{H}(P;\cdot)$ over $[\epsilon_0,\epsilon_1]$. This shows that uniform convergence holds over $[\epsilon_0,\infty)$. □

D.2.4 Proof of Theorem 16

The following lemma follows directly from Lemma 38 by noticing that $\mathcal{L}(\cdot,\cdot)$ is continuous over $\mathcal{Q}\times\mathcal{W}_N$ and $\mathcal{W}_N$ is compact.

Lemma 39. *Assume that conditions (**C.**1–3) hold true for a given closed set $\mathcal{Q}\subseteq\mathcal{P}$ and a given $N\in\mathbb{N}$. Then $\epsilon_{\min}(\cdot)$, as given in (6.85), is continuous over $\mathcal{Q}$.*

We prove the following lemma which will be used in the proof of Theorem 16.

Lemma 40. *Assume that conditions (**C.**1) and (**C.**3) hold true for a given closed set $\mathcal{Q}\subseteq\mathcal{P}$ and a given $N\in\mathbb{N}$. Then the set $\mathcal{W}_N^*(P;\epsilon)$, defined in (6.84), is compact.*

Proof. Observe that

$$\mathcal{W}_N^*(P;\epsilon)=\{W\in\mathcal{W}_N:\mathcal{L}(P,W)\le\epsilon\}\cap\{W\in\mathcal{W}_N:\mathcal{U}(P,W)=\mathsf{H}(P;\epsilon)\}. \tag{D.128}$$

Recall that, by condition (**C.**1), the mappings $\mathcal{L}(P,\cdot)$ and $\mathcal{U}(P,\cdot)$ are continuous. Since both sets in the RHS of (D.128) are the preimage of a closed set under a continuous function, they are closed. Therefore, $\mathcal{W}_N^*(P;\epsilon)$ is closed as the finite intersection of closed sets is closed. Finally, we conclude that $\mathcal{W}_N^*(P;\epsilon)$ is compact since it is a closed subset of the compact set $\mathcal{W}_N$. □

Observe that, by Lemma 39, $\lim_n\epsilon_{\min}(P_n)=\epsilon_{\min}(P)$ whenever $\lim_n P_n=P$. Hence, if $\epsilon_{\min}(P)<\epsilon$, then $\epsilon_{\min}(P_n)<\epsilon$ for n large enough. Therefore, under the assumptions of Theorem 16, for n large enough we have that the set $\mathcal{W}^*(P_n;\epsilon)$ is non-empty and, by condition (**C.**3), $\mathcal{W}_N^*(P_n;\epsilon)$ is non-empty as well. Now we are in position to prove Theorem 16.

Proof. We assume, without loss of generality, that $\mathcal{W}_N^*(P_n;\epsilon)$ is non-empty for all $n\in\mathbb{N}$. In order to reach contradiction, assume that (6.90) does not hold true, i.e., there exists $W_n^*\in\mathcal{W}_N^*(P_n;\epsilon)$ such that

$$\limsup_{n\to\infty}\mathsf{dist}(W_n,\mathcal{W}_N^*(P;\epsilon))>0. \tag{D.129}$$

Since $\mathcal{W}_N$ and $\mathcal{W}_N^*(P;\epsilon)$ are compact from Lemma 40, a routine argument shows that there exist a subsequence $(W_{n_k}^*)_{k=1}^{\infty}$ and $W_0 \in \mathcal{W}_N$ such that $\lim_k W_{n_k}^* = W_0$ and $W_0 \notin \mathcal{W}_N^*(P;\epsilon)$. Lemma 21 shows that $\mathcal{L}(\cdot,\cdot)$ is continuous over $\mathcal{Q} \times \mathcal{W}_N$. Therefore, we have

$$\mathcal{L}(P, W_0) = \lim_{k\to\infty} \mathcal{L}(P_{n_k}, W_{n_k}^*) \leq \epsilon. \tag{D.130}$$

Hence, by the maximality of the privacy-utility function, $\mathcal{U}(P, W_0) \leq \mathsf{H}(P;\epsilon)$. Using a similar argument, one can show that

$$\mathcal{U}(P, W_0) = \lim_{k\to\infty} \mathcal{U}(P_{n_k}, W_{n_k}^*) = \lim_{k\to\infty} \mathsf{H}(P_{n_k}, \epsilon) = \mathsf{H}(P;\epsilon), \tag{D.131}$$

where the last equality follows from the fact that the mapping $\mathsf{H}(\cdot;\epsilon)$ is continuous, as established in Proposition 16. Altogether,

$$\mathcal{L}(P, W_0) \leq \epsilon \quad \text{and} \quad \mathcal{U}(P, W_0) = \mathsf{H}(P;\epsilon). \tag{D.132}$$

Therefore, $W_0 \in \mathcal{W}_N^*(P;\epsilon)$ which contradicts the fact that $W_0 \notin \mathcal{W}_N^*(P;\epsilon)$.

Furthermore, by the strong law of large numbers, $(\hat{P}_n)_{n=1}^{\infty}$ converges almost surely to P as n goes to infinity. Therefore, if $P_n = \hat{P}_n$, then

$$\Pr\left(\lim_{n\to\infty} \mathsf{dist}(W_n^*, \mathcal{W}_N^*(P;\epsilon)) = 0\right) = 1, \tag{D.133}$$

as claimed. □

Note that in the previous proof we implicitly assume that, for each $n \in \mathbb{N}$, the (possibly random) privacy mechanism $W_n^* \in \mathcal{W}_N^*(\hat{P}_n;\epsilon)$ is a random variable, i.e., a measurable function. This is a minor subtlety given the rareness of non-measurable sets (functions) [212].

D.2.5 Proof of Theorem 17

Before we prove Theorem 17, let us recall some definitions and theorems from set-valued analysis [20].

Given two metric spaces $\mathcal{A}$ and $\mathcal{B}$, a set-valued mapping F from $\mathcal{A}$ to $\mathcal{B}$, denoted by $F : \mathcal{A} \rightsquigarrow \mathcal{B}$, is a function from $\mathcal{A}$ to the subsets of $\mathcal{B}$. The domain of a set-valued mapping $F : \mathcal{A} \rightsquigarrow \mathcal{B}$ is defined as

$$\mathrm{Dom}(F) \triangleq \{a \in \mathcal{A} : F(a) \neq \emptyset\}. \tag{D.134}$$

For $r > 0$, $\mathcal{A}_r(a) \triangleq \{a' \in \mathcal{A} : d_{\mathcal{A}}(a, a') < r\}$ where $d_{\mathcal{A}}$ denotes the metric associated to $\mathcal{A}$.

Definition 28 (Definition 1.4.1, [20])**.** A set-valued mapping $F : \mathcal{A} \rightsquigarrow \mathcal{B}$ is called upper semicontinuous at $a \in \mathrm{Dom}(F)$ if and only if for every neighborhood $\mathcal{N}$ of $F(a)$ there exists $r > 0$ such that $F(a') \subset \mathcal{N}$ for all $a' \in \mathcal{A}_r(a)$. A set-valued mapping is said to be upper semicontinuous if and only if it is upper semicontinuous at every point of its domain.

Definition 29 (Definition 1.4.2, [20])**.** A set-valued mapping $F : \mathcal{A} \rightsquigarrow \mathcal{B}$ is called lower semicontinuous at $a \in \mathrm{Dom}(F)$ if and only if for every $b \in F(a)$ and every sequence $(a_n)_{n=1}^{\infty} \subset \mathrm{Dom}(F)$ with $\lim_n a_n = a$, there exists a sequence $(b_n)_{n=1}^{\infty}$ such that $b_n \in F(a_n)$ for each $n \in \mathbb{N}$ and $\lim_n b_n = b$. A set-valued mapping is said to be lower semicontinuous if and only if it is lower semicontinuous at every point of its domain.

By definition, a the set-valued mapping $F : \mathcal{A} \rightsquigarrow \mathcal{B}$ is continuous at $a \in \mathcal{A}$ if and only if it is both upper and lower semicontinuous at a.

Theorem 21 (Theorem 1.4.13, [20])**.** *Let F be a set-valued map from a complete metric space $\mathcal{A}$ to a complete separable metric space $\mathcal{B}$. If F is upper semicontinuous, then it is continuous on a residual set of $\mathcal{A}$.*

The proof of Theorem 17 relies on the following lemma.

Lemma 41. *Assume that conditions (**C**.1–3) hold true for a given closed set $\mathcal{Q} \subseteq \mathcal{P}$ and a given $N \in \mathbb{N}$. For any $\epsilon \in \mathbb{R}$, the set-valued mapping $\mathcal{W}_N^*(\cdot; \epsilon)$ is upper semicontinuous over $\{Q \in \mathcal{Q} : \epsilon_{\min}(Q) < \epsilon\}$.*

Proof. For ease of notation, let $F : \mathcal{Q} \rightsquigarrow \mathcal{W}_N$ be the set-valued mapping defined by $F(Q) \triangleq \mathcal{W}_N^*(Q; \epsilon)$. Observe that $\mathrm{Dom}(F) = \{Q \in \mathcal{Q} : \epsilon_{\min}(Q) \leq \epsilon\}$. In order to reach contradiction, assume that the set-valued mapping F is not upper semicontinuous over $\{Q \in \mathcal{Q} : \epsilon_{\min}(Q) < \epsilon\}$. In this case, there exist a joint distribution $Q \in \mathcal{Q}$ with $\epsilon_{\min}(Q) < \epsilon$ and a neighborhood $\mathcal{N}$ of $F(Q)$ such that for every $r > 0$ there exists $Q_r \in \mathcal{Q}_r(Q)$ with $F(Q_r) \not\subset \mathcal{N}$. For each $n \in \mathbb{N}$, let W_n^* be such that $W_n^* \in F(Q_{1/n}) \cap \mathcal{N}^c$, where $\mathcal{N}^c$ denotes the complement of $\mathcal{N}$ w.r.t. $\mathcal{W}_N$. Since $\mathcal{N}^c$ is closed and $\mathcal{W}_N$ is compact, we have that $\mathcal{N}^c$ is also compact. Hence, there exists a subsequence $(W_{n_k}^*)_{k=1}^{\infty}$ converging to some $W \in \mathcal{N}^c$. Since $F(Q) \subset \mathcal{N}$, we have that $W \notin F(Q)$. Recall that $\lim_k Q_{1/n_k} = Q$ by construction, and thus Corollary 6 implies that $W \in \mathcal{W}_N^*(Q; \epsilon)$. Since this contradicts the fact that $W \notin \mathcal{W}_N^*(Q; \epsilon)$, we conclude that F is necessarily upper semicontinuous over $\{Q \in \mathcal{Q} : \epsilon_{\min}(Q) < \epsilon\}$. □

Now we are in position to prove Theorem 17.

Proof. For ease of notation, let

$$\mathcal{Q}^{(\delta)} \triangleq \{Q \in \mathcal{Q} : \epsilon_{\min}(Q) + \delta \leq \epsilon\}. \tag{D.135}$$

Observe that $\mathcal{Q}^{(\delta)} \subset \{Q \in \mathcal{Q} : \epsilon_{\min}(Q) < \epsilon\}$. Thus, Lemma 41 implies that the mapping $\mathcal{W}_N^*(\cdot;\epsilon)$ is upper semicontinuous over $\mathcal{Q}^{(\delta)}$. Observe that both $\mathcal{Q}^{(\delta)}$ and $\mathcal{W}_N$ are compact metric spaces. Therefore, Theorem 21 establishes the lower semicontinuity of the mapping $\mathcal{W}_N^*(\cdot;\epsilon)$ on a residual set of $\mathcal{Q}^{(\delta)}$. Theorem 17 follows immediately from this fact and Definition 29. □

D.3 Proofs for Section 6.6

D.3.1 Proof of Lemma 22

Proof. (i) Let $W \in \mathcal{W}_N$ be fixed. By condition (**C.**2), the mapping $\mathcal{U}(\cdot, W)$ is (Lipschitz) continuous over $\mathcal{Q}_r(\hat{P})$. Since $\mathcal{Q}_r(\hat{P})$ is compact, the mapping $\mathcal{U}(\cdot, W)$ attains its infimum over $\mathcal{Q}_r(\hat{P})$.

(ii) Assume that $\mathcal{D}_{\mathcal{Q},N}(\hat{P};\epsilon,r)$ is non-empty. By (i), we have that

$$\mathcal{U}_r(\hat{P}, W) = \min_{Q \in \mathcal{Q}_r(\hat{P})} \mathcal{U}(Q, W) \tag{D.136}$$

is well defined. As established in Lemma 21, the mapping $\mathcal{U}(\cdot,\cdot)$ is continuous over $\mathcal{Q} \times \mathcal{W}_N$ and in particular over $\mathcal{Q}_r(\hat{P}) \times \mathcal{W}_N$. Hence, Lemma 38 implies that $\mathcal{U}_r(\hat{P},\cdot)$ is continuous over $\mathcal{W}_N$.

Note that, by definition,

$$\mathcal{D}_{\mathcal{Q},N}(\hat{P};\epsilon,r) = \bigcap_{Q \in \mathcal{Q}_r(\hat{P})} \mathcal{D}_N(Q;\epsilon), \tag{D.137}$$

where $\mathcal{D}_N(Q;\epsilon) = \{W \in \mathcal{W}_N : \mathcal{L}(Q,W) \leq \epsilon\}$. Condition (**C.**1) implies that, for every $Q \in \mathcal{Q}$, the mapping $\mathcal{L}(Q,\cdot)$ is continuous over $\mathcal{W}_N$. Thus, we have that $\mathcal{D}_N(Q;\epsilon)$ is compact as $\mathcal{W}_N$ is compact. Since the arbitrary intersection of compact sets is also compact, (D.137) readily shows that $\mathcal{D}_{\mathcal{Q},N}(\hat{P};\epsilon,r)$ is compact. Therefore, the (continuous) mapping $\mathcal{U}_r(\hat{P},\cdot)$ attains its supremum over $\mathcal{D}_{\mathcal{Q},N}(\hat{P};\epsilon,r)$, as required. □

D.3.2 Proof of Theorem 18

Consider the following lemma.

Lemma 42. *Assume that conditions (**C.**1–3) hold true for a given closed set $\mathcal{Q} \subseteq \mathcal{P}$ and a given $N \in \mathbb{N}$. Let*

$\epsilon \in \mathbb{R}$ *and* $r \geq 0$ *be given. If* $P \in \mathcal{Q}_r(\hat{P})$ *and* $\epsilon - C_L r \geq \epsilon_{\min}(\hat{P})$, *then*

$$\mathsf{H}(P;\epsilon) \leq \mathsf{H}(\hat{P};\epsilon + C_L r) + C_U r. \tag{D.138}$$

Proof. We start proving that $\epsilon \geq \epsilon_{\min}(P)$. Remark 14 implies that there exists $W' \in \mathcal{W}_N$ such that $\mathcal{L}(\hat{P}, W') = \epsilon_{\min}(\hat{P})$. By the Lipschitz continuity given in condition (**C.**2), we have that

$$\mathcal{L}(P, W') \leq \mathcal{L}(\hat{P}, W') + |\mathcal{L}(\hat{P}, W') - \mathcal{L}(P, W')| \tag{D.139}$$

$$\leq \epsilon_{\min}(\hat{P}) + C_L \|\hat{P} - P\|_1 \tag{D.140}$$

$$\leq \epsilon_{\min}(\hat{P}) + C_L r \leq \epsilon, \tag{D.141}$$

where the last inequality follows from the assumption $\epsilon - C_L r \geq \epsilon_{\min}(\hat{P})$. By the minimality of $\epsilon_{\min}(P)$, we conclude that $\epsilon_{\min}(P) \leq \epsilon$. Then, Remark 14 implies that there exists $W \in \mathcal{W}_N$ such that $\mathcal{L}(P, W) \leq \epsilon$ and

$$\mathsf{H}(P;\epsilon) = \mathcal{U}(P, W). \tag{D.142}$$

By the Lipschitz continuity given in condition (**C.**2), we have that

$$|\mathcal{U}(\hat{P}, W) - \mathcal{U}(P, W)| \leq C_U \|\hat{P} - P\|_1 \leq C_U r. \tag{D.143}$$

In particular,

$$\mathsf{H}(P;\epsilon) \leq \mathcal{U}(\hat{P}, W) + C_U r. \tag{D.144}$$

Similarly, since

$$|\mathcal{L}(\hat{P}, W) - \mathcal{L}(P, W)| \leq C_L \|\hat{P} - P\|_1 \leq C_L r, \tag{D.145}$$

we have $\mathcal{L}(\hat{P}, W) \leq \mathcal{L}(P, W) + C_L r \leq \epsilon + C_L r$. Therefore, from inequality (D.144) and the maximality of the privacy-utility function, we conclude that

$$\mathsf{H}(P;\epsilon) \leq \mathsf{H}(\hat{P};\epsilon + C_L r) + C_U r, \tag{D.146}$$

as we wanted to prove. □

Now we are in position to prove Theorem 18.

Proof. For any $W \in \mathcal{D}_N(\hat{P};\epsilon - C_L r)$, we have $\mathcal{L}(\hat{P}, W) + C_L r \leq \epsilon$. By the Lipschitz continuity given in condition (**C.**2), for every $Q \in \mathcal{Q}_r(\hat{P})$,

$$|\mathcal{L}(\hat{P}, W) - \mathcal{L}(Q, W)| \leq C_L \|\hat{P} - Q\|_1 \leq C_L r. \tag{D.147}$$

Hence, $\mathcal{L}(Q,W) \leq \epsilon$ and $W \in \mathcal{D}_N(Q;\epsilon)$ for every $Q \in \mathcal{Q}_r(\hat{P})$. By (6.110), it follows that $W \in \mathcal{D}_{\mathcal{Q},N}(\hat{P};\epsilon,r)$.

Let $W^* \in \mathcal{W}_N^*(\hat{P};\epsilon - C_L r)$ and $W^\dagger \in \mathcal{W}_{\mathcal{Q},N}^\dagger(\hat{P};\epsilon,r)$. By definition,

$$\mathcal{U}(\hat{P},W^*) = \mathsf{H}(\hat{P};\epsilon - C_L r). \tag{D.148}$$

By the Lipschitz continuity given in condition (**C.**2), we have that

$$|\mathcal{U}(\hat{P},W^*) - \mathcal{U}(P,W^*)| \leq C_U \|\hat{P} - P\|_1 \leq C_U r. \tag{D.149}$$

Combining (D.148) and (D.149) together, we have

$$\mathcal{U}(P,W^*) \geq \mathsf{H}(\hat{P};\epsilon - C_L r) - C_U r. \tag{D.150}$$

Since $W^\dagger \in \mathcal{W}_{\mathcal{Q},N}^\dagger(\hat{P};\epsilon,r) \subset \mathcal{D}_{\mathcal{Q},N}(\hat{P};\epsilon,r)$ and $P \in \mathcal{Q}_r(\hat{P})$, (6.110) implies that $W^\dagger \in \mathcal{D}_N(P;\epsilon)$. Thus, by the maximality of $\mathsf{H}(P;\epsilon)$, we have

$$\mathcal{U}(P,W^\dagger) \leq \mathsf{H}(P;\epsilon). \tag{D.151}$$

An immediate application of Lemma 42 shows that

$$\mathcal{U}(P,W^\dagger) \leq \mathsf{H}(P;\epsilon) \leq \mathsf{H}(\hat{P};\epsilon + C_L r) + C_U r. \tag{D.152}$$

Therefore, combining (D.150) and (D.152) together,

$$\mathcal{U}(P,W^*) - \mathcal{U}(P,W^\dagger) \geq -\left(\mathsf{H}(\hat{P};\epsilon + C_L r) - \mathsf{H}(\hat{P};\epsilon - C_L r) + 2C_U r\right), \tag{D.153}$$

which implied the desired conclusion. □

D.3.3 Proof of Theorem 19

Proof. Recall that, by Lemma 39, the mapping $\epsilon_{\min}(\cdot)$ is continuous over $\mathcal{Q}$. In particular, we have that $\lim_n \epsilon_{\min}(P_n) = \epsilon_{\min}(P)$ whenever $\lim_n P_n = P$. Since $\epsilon_{\min}(P) < \epsilon$, $\lim_n P_n = P$, and $\lim_n r_n = 0$, we have that $\epsilon_{\min}(P_n) + C_L r_n < \epsilon$ for n large enough. Therefore, for n large enough, the first part of Theorem 18 establishes that the set $\mathcal{D}_{\mathcal{Q},N}(P_n;\epsilon,r_n)$ is non-empty. Furthermore, part (ii) of Lemma 22 implies that $\mathcal{W}_{\mathcal{Q},N}^\dagger(P_n;\epsilon,r_n)$ is non-empty as well. Without loss of generality, we assume that $\mathcal{W}_{\mathcal{Q},N}^\dagger(P_n;\epsilon,r_n) \neq \emptyset$ for all $n \geq 1$.

In order to reach contradiction, we assume that the conclusion does not hold true, i.e., that there

exists a sequence $(W_n^\dagger)_{n=1}^\infty$ such that $W_n^\dagger \in \mathcal{W}_{\mathcal{Q},N}^\dagger(P_n;\epsilon,r_n)$ for all $n \geq 1$ and

$$\limsup_{n\to\infty} \mathsf{dist}(W_n^\dagger, \mathcal{W}_N^*(P;\epsilon)) > 0. \tag{D.154}$$

Since both $\mathcal{W}_N$ and $\mathcal{W}_N^*(P;\epsilon)$ are compact (Lemma 40), there exists a subsequence of $(W_n^\dagger)_{n=1}^\infty$ converging to some $W_0 \in \mathcal{W}_N$ with $W_0 \notin \mathcal{W}_N^*(P;\epsilon)$. Without loss of generality, we assume that $\lim_n W_n^\dagger = W_0$. By the continuity of $\mathcal{L}(\cdot,\cdot)$ established in Lemma 21,

$$\mathcal{L}(P,W_0) = \lim_{n\to\infty} \mathcal{L}(P_n, W_n^\dagger) \leq \epsilon. \tag{D.155}$$

The last inequality implies that $W_0 \in \mathcal{D}_N(P;\epsilon)$ and hence $\mathcal{U}(P,W_0) \leq \mathsf{H}(P;\epsilon)$. Assume that we have already proved that $\mathcal{U}(P,W_0) \geq \mathsf{H}(P;\epsilon)$. In particular, we would have that $\mathcal{U}(P,W_0) = \mathsf{H}(P;\epsilon)$ and thus $W_0 \in \mathcal{W}_N^*(P;\epsilon)$. Since this conclusion contradicts the fact that $W_0 \notin \mathcal{W}_N^*(P;\epsilon)$, the result would follow. Now we focus on proving that $\mathcal{U}(P,W_0) \geq \mathsf{H}(P;\epsilon)$.

Since $\epsilon > \epsilon_{\min}(P)$ and $\lim_n r_n = 0$, we have that $\epsilon - 2C_L r_n > \epsilon_{\min}(P)$ for n large enough. Without loss of generality, we assume that $\epsilon - 2C_L r_n > \epsilon_{\min}(P)$ for all $n \geq 1$. For a given $n \geq 1$, let $W_0^* \in \mathcal{W}_N$ be such that $\mathcal{L}(P,W_0^*) \leq \epsilon - 2C_L r_n$ and $\mathcal{U}(P,W_0^*) = \mathsf{H}(P;\epsilon - 2C_L r_n)$. By condition (**C.**2), for any $Q \in \mathcal{Q}_{r_n}(P_n)$,

$$|\mathcal{L}(P,W_0^*) - \mathcal{L}(Q,W_0^*)| \leq C_L \|P-Q\|_1 \leq 2C_L r_n, \tag{D.156}$$

where the last inequality follows from the triangle inequality and the fact that $P, Q \in \mathcal{Q}_{r_n}(P_n)$. Thus,

$$\mathcal{L}(Q,W_0^*) \leq \mathcal{L}(P,W_0^*) + 2C_L r_n \leq \epsilon, \tag{D.157}$$

for any $Q \in \mathcal{Q}_{r_n}(P_n)$. Hence, $W_0^* \in \mathcal{D}_{\mathcal{Q},N}(P_n;\epsilon,r_n)$ and, by the definition of $\mathcal{W}_{\mathcal{Q},N}^\dagger(P_n;\epsilon,r_n)$,

$$\min_{Q\in\mathcal{Q}_{r_n}(P_n)} \mathcal{U}(Q,W_0^*) \leq \min_{Q\in\mathcal{Q}_{r_n}(P_n)} \mathcal{U}(Q,W_n^\dagger) \leq \mathcal{U}(P,W_n^\dagger). \tag{D.158}$$

As in (D.156), we have that

$$|\mathcal{U}(P,W_0^*) - \mathcal{U}(Q,W_0^*)| \leq 2C_U r_n. \tag{D.159}$$

Therefore, for any $Q \in \mathcal{Q}_{r_n}(P_n)$,

$$\mathcal{U}(Q,W_0^*) \geq \mathcal{U}(P,W_0^*) - 2C_U r_n \tag{D.160}$$

$$= \mathsf{H}(P;\epsilon - 2C_L r_n) - 2C_U r_n. \tag{D.161}$$

In particular, this implies that

$$\min_{Q \in \mathcal{Q}_{r_n}(P_n)} \mathcal{U}(Q, W_0^*) \geq \mathsf{H}(P; \epsilon - 2C_L r_n) - 2C_U r_n. \tag{D.162}$$

Combining (D.158) and (D.162) together, we have

$$\mathcal{U}(P, W_n^\dagger) \geq \min_{Q \in \mathcal{Q}_{r_n}(P_n)} \mathcal{U}(Q, W_0^*) \tag{D.163}$$

$$\geq \mathsf{H}(P; \epsilon - 2C_L r_n) - 2C_U r_n. \tag{D.164}$$

Therefore, by the continuity of $\mathcal{U}(P, \cdot)$ and $\mathsf{H}(P; \cdot)$ assumed in condition (**C.**1),

$$\mathcal{U}(P, W_0) = \lim_{n \to \infty} \mathcal{U}(P, W_n^\dagger) \tag{D.165}$$

$$\geq \lim_{n \to \infty} \mathsf{H}(P; \epsilon - 2C_L r_n) - 2C_U r_n \tag{D.166}$$

$$= \mathsf{H}(P; \epsilon). \tag{D.167}$$

This concludes the proof of the first part of the theorem.

Now we prove the second part of the theorem. Recall that in this case, for all $n \geq 1$, $P_n = \hat{P}_n$ and $r_n \geq (2p \log(n)/n)^{1/2}$ for some $p > 1$. Thus,

$$\sum_{n=1}^{\infty} \Pr(\|\hat{P}_n - P\|_1 \geq r_n) \leq \sum_{n=1}^{\infty} \Pr\left(\|\hat{P}_n - P\|_1 \geq \sqrt{\frac{2p \log(n)}{n}} \right) \tag{D.168}$$

$$\leq \exp(|\mathcal{S}| \cdot |\mathcal{X}|) \sum_{n=1}^{\infty} \frac{1}{n^p}, \tag{D.169}$$

where (D.169) follows directly from (6.26). Since $p > 1$ by assumption, we have that $\sum_n 1/n^p$ is finite. Thus, by a routine application of the Borel-Cantelli Lemma, see, e.g., [116, Sec. 2.18],

$$\Pr\left(\left\{ \omega : \|\hat{P}_n(\omega) - P\|_1 < r_n \text{ for all } n \geq N = N(\omega) \right\} \right) = 1. \tag{D.170}$$

Observe that, by the last equality, the hypotheses of the first part of this theorem are satisfied almost surely. A direct application of the former part of this theorem leads to

$$\Pr\left(\lim_{n \to \infty} \mathsf{dist}(W_n^\dagger, \mathcal{W}_N^*(P; \epsilon)) = 0 \right) = 1. \tag{D.171}$$

□