Statistical Privacy and Security

by

Salman Salamatian

B.S., École Polytechnique Féderale de Lausanne (2012)
M.S., École Polytechnique Féderale de Lausanne (2014)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2020

Author .....
Department of Electrical Engineering and Computer Science
August 28, 2020

Certified by .....
Muriel Médard
Cecil H. Green Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by .....
Leslie A. Kolodziejski
Professor of Electrical Engineering
Chair, Department Committee on Graduate Students

Statistical Privacy and Security

by

Salman Salamatian

Submitted to the Department of Electrical Engineering and Computer Science
on August 28, 2020, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

The tremendous increase of personal data being shared online, along with the rapid development of data mining techniques is a serious threats to privacy and security, as evidenced by the numerous privacy and security scandals of the past several years. At their core, the new privacy and security challenges that the big data revolution poses are due to the unclear boundary between data shared willingly, which is deemed not-sensitive, and the sensitive data that one wants to keep private.

Traditional tools in security and privacy provide protection by encrypting personal data, but this method is not sustainable when it is unclear whether, or how much, the data is sensitive to begin with. The premise of this thesis is that information theoretic tools and insights are useful to identify how releasing personal data can impact privacy and security, and can serve as a design driver for building privacy preserving, and security enhancing systems.

In particular, we will be focused on two types of attacks. In the first, we consider how a user may release some personal data (e.g. movie ratings) in exchange for a service (e.g. movie recommendations), while simultaneously not leaking information about a sensitive attribute correlated with the personal data (e.g. political orientation). To this end, we design a privacy framework which captures the inference threat of releasing data, and use the latter to find optimal privacy-preserving mechanisms, which allows the user to trade utility for privacy. In the second part, we look at brute-force attacks where an adversary attempts to breach into a password secured system by querying potential passwords. Users of such systems are likely to generate poor passwords, re-use passwords across systems, and especially susceptible to targeted attacks if their password is correlated with personal data that is available online. We consider various setups under which Brute-force attacks occur, and analyze the security guarantees one obtain via Guesswork – an information theoretic quantity that is a surrogate for the computational effort than the attacker has to perform.

The analysis of both attacks reveals that data is a precious commodity which should be handled with care, and how the entire data acquisition and communication pipeline can come under attack. Additionally, Information Theory and Statistics offers a dimension of tools which is complementary to the existing ones, while still capturing the fundamentals of the security and privacy threats in the digital age.

Thesis Supervisor: Muriel Médard
Title: Cecil H. Green Professor of Electrical Engineering and Computer Science

Acknowledgments

This thesis would not have been possible without the support and kindness of many.

First and foremost, I would like to express my deep gratitude to my research advisor Muriel Médard, for her patient guidance, enthusiastic support, and immense kindness throughout my journey at MIT. I was fortunate to have Muriel co-supervise my Masters thesis in 2014 when I spent the summer in the *Network Coding and Reliable Communications* (NCRC) group, before I eventually joined the group for my PhD. Since that first meeting, Muriel has been an never-ending source of inspiration, mentorship, and all-around support. Muriel's vision and insight, her ability to articulate and address important problems, and her deep and wide knowledge of many technical fields have heavily influenced my research. However, Muriel's impact in my life goes beyond the academic. No matter the inevitable hardships of a graduate journey, Muriel's door is always open. Muriel also leads by the example. Among other many contributions, the significant strides she has made towards increasing diversity in STEM at MIT, and in the various communities she is part of, is a source of constant inspiration and respect; we will benefit from her efforts for many years to come. Muriel – Merci pour tout.

I am also extremely grateful to the members of this committee: Flavio Du Pin Calmon, and Vinod Vaikuntanathan. I trace back my very first steps in research to a serendipitous internship where I met Flavio, in 2012. This meeting single-handedly shaped so many of my research interests at the intersection of privacy, information theory, and machine learning. He exemplifies the best qualities of an Information Theorist: the marriage between a mathematician and an engineer. Working on a research problem with Flavio is always exciting, delving deep into the equations one moment, and discussing practical applications the next. Many of the results and ideas in this thesis are the product of such discussions. Beyond research, Flavio has been an amazing mentor and role-model. His guidance, support, and help, in research and in more, have gone beyond what is expected of a mentor, colleague, or friend. Flavio – I cannot thank you enough.

I would also like to thank Vinod Vaikuntanathan for being on my committee. While my first discussion with Vinod took place later in my academic carreer, this thesis bears the mark of this meeting heavily. In every interaction we have had, Vinod's open-mindedness, and his exceptionally sharp questions, have steered the discussion in creative and fruitful directions. Thank you Vinod, for positively influencing this thesis.

I would like to thank all the past and present members of the NCRC group. Up until 2020, these few rooms in building 36 have been like another home. This feeling was fueled not only by the magical supply of chocolate and coffee, but by the camaraderie and friendships. Whether it is a deadline week, or a particularly long Wednesday, there is always someone willing to discuss about an interesting math brain-teasers, play soccer in the office, engage in a round or two of Geo-guessr, or take part in a debate on what topological properties differentiate sandwiches from tacos. Being with such a fun, yet brilliant and passionate group of people has been an incredibly rewarding experience. I am thankful to the many students and friends that fostered this environment, including Amit Solomon, Litian Liu, Kathleen Yang, Diana González, Derya Malak, Ahmad Beirami, Wasim Huleihel, Alejandro Cohen, Rafael D'Oliveira, Weifei Zeng, Arman Rezaee, Homa Esfahanizadeh, Georgios Angelopoulos, Jason Cloud, Soheil Feizi, Jiange Li, Kerim Fouli, Vitaly Abdrashitov, Jinfeng Du, and Ali Makhdoumi. I am particularly thankful to Soheil Feizi, for hosting me in his apartment when I first joined MIT, and to Arman Rezaee for making sure I never broke

anything while playing soccer in the office. Perhaps the most important person in our group is Molly Kruko. Not only is Molly essential to the functioning of the group, she is a ray of sunshine in the often gray and cold Cambridge. Molly, thank you for your kindness and support.

I am also thankful to the many colleagues and co-authors, without whom this thesis would not have been what it is. I was fortunate to interact and collaborate with several amazing mentors, who shaped my research and my interest, and guided me in my academic life and beyond. In particular, I would like to thank Nadia Fawaz, Brano Kveton, and Nina Taft for believing in me and offering me my first internship, Emre Telatar, for his constant mentorship and guidance throughout the years, Asaf Cohen, for being an exceptional mentor during my first year at MIT, Ken Duffy, whose advice and guidance I benefited from greatly, Yonina Eldar, whose short stay at MIT left a mark on many of my research interest, the postdocs in the NCRC group for being such reliable sources of help and guidance, Hsiang Hsu, for welcoming me into his group at Harvard, and to the many friends and colleagues in the Information Theory Community.

I am also thankful for the support and friendship of the many people who have made Boston feel like home. I am thankful to my *Cambridge family* for, among other things, feeding me countless times. I am also thankful to the many friends for making my time here special. I am particularly thankful for my friendship with Saviz Mowlavi – who has followed me to Boston all the way from high-school. I would also like the thank my friends back home (*l'ancienneté*). It is a long way from Ferney-Voltaire to Boston, but you all make me a better person.

I would like to thank Maryam Khalid for filling my heart with joy, day after day. Her support, love, and kindness have no bounds, and I could not hope to summarize all the ways she has impacted this thesis, and more importantly, my life, in a few lines. In addition to being incredibly smart, driven, and inspiring, she is a constant source of surprise, laugh, and love. Maryam, thank you for being my partner on this journey, and I am looking forward our next adventure.

Finally, I would like to thank my family, my brother Loqman, my father Kavé, and my mother Venus. I would not have been able to make it, without your constant love, support and sacrifice. From my brother, I learned to enjoy the small things in life, and that love knows no distance. From my father, I learned the essence of hard-work and to always aspire to better myself. From my mother, I learned to not give up, be proud of who I am, and to never quit learning. Thank you, for making me who I am.

Contents

List of Figures

To my parents.

Chapter 1

Introduction

The past decade has showcased the tremendous potential of data-driven methods in a variety of domains from engineering, medicine, entertainment, and more. At the origin of this *big data* wave is access to cheap and massive amounts of data, which is a necessary ingredient in the success of current learning models. Thus, we live in an age in which our personal data is collected, stored, and heavily processed. This has repercussions – our modern society has already been affected by several privacy and security scandals including personal data on social networks being used for political purposes [120, 7], passwords being guessed by abusing security questions [111], and patient medical records being leaked and mined in unlawful ways [8] – all showcasing how fragile our privacy is today. But, Is Privacy Dead [115]?

In this thesis, we leverage tools from Information Theory and Statistical Learning to provide insights into this question. While data security is a well-studied subject, we argue that some new challenges of the big data era cannot be well understood from the lens of traditional security primitives such as cryptography or differential privacy. Privacy leaks sometimes happen in unexpected ways. In 2012, the father of a teenager learned via the personalized coupons he received at home that his daughter was pregnant – before she had a chance to tell him herself [3]. This example highlights how unclear the boundary between private and public data is: sensitive data (pregnancy status of a teenager) could be leaked from the correlation with data, which at first appears to be not sensitive (purchases at the supermarket). The boundaries between sensitive data and publicly shared data are unclear, resulting in a need for an understanding that is complementary to the traditional methods

in data security. Therefore, our goal is to develop new ways to:

- Quantify privacy and security threats that arise from sharing personal data.
- Explore the fundamental trade-offs between the amount of data shared and the performance of a service .
- Devise tools and methods to share/collect less data altogether.

The privacy and security issues highlight fundamental flaws in the entirety of our data acquisition and communication pipeline. From the perspective of the user, personal data is collected and can potentially be misused without having much control over it. From the perspective of the data collectors, this massive amount of data creates tremendous challenges in communication overhead, computational costs, and, of course, additional security concerns. For both parties, there is interest in understanding what in the data is truly needed, how to effectively represent, communicate, and process it. In other words, how can we be parsimonious in what we share? In this thesis, we will discuss several problems which are relevant to the big-data era:

1. **Privacy against Inference:** A user wishes to share some information and, in exchange, receives a service. For example, the information shared could be movie ratings and the service – movie recommendations. The issue lies in the correlation between personal data, and some sensitive information (e.g. political or sexual orientation) that the user does not want to disclose. We will introduce novel tools and frameworks to formally analyze this problem and quantify the privacy threat of an adversary performing an inference attack on the sensitive information. At the core of this problem lies an inherent trade-off between how private one is, and the quality of the service one receives. This trade-off can be captured by the Privacy Funnel – a method to find privacy mechanisms with strong guarantees that introduces as little noise in the data as possible. The Privacy Funnel and related formulations we will discuss have several generalizations and applications.
2. **Brute-Force Security & Guesswork:** Passwords are among the main ways we provide security online. What happens when humans generate keys? We seldom make good random number generators, and in fact, in many cases we are very poor at creating strong passwords. Human generated passwords are predictable and re-used

across many platforms. More importantly, personal information is often part of the password itself (e.g. date of birth, name of family members or pets, etc.), which may be already shared online. As a result of this, brute-force attacks – which consist in querying password secured systems until the attacker guesses the correct password – are a grave concern. We will introduce various settings in which a system undergoes a brute-force attack.

3. **Parsimonious Data Representation, beyond Privacy and Security:** At a high level, the two previous subjects share the same take-away: data is a precious commodity and should be shared and collected only when necessary. We will briefly explain how to collect/share less data altogether, and re-think the entirety of our data acquisition and communication pipeline. The focus should be on obtaining quality data as opposed to sheer quantity.

1.1 A new paradigm for privacy and security

The general premise of this thesis is that information theoretic tools and insights can help us tackle some of the major privacy and security challenges in the era of big data. The interplay between Information Theory and security dates back to Shannon himself. In 1945, Shannon publishes a classified report titled *"A Mathematical Theory of Cryptography"*, while working at Bell Labs. The paper is eventually published for the public in 1949 [130], under a different title, but the earlier version predates the publication of his *"A Mathematical Theory of Communications"* which appeared in 1948. The relationship between the two fields goes beyond the anecdote: It is undeniable that the perspective that Shannon obtained from working on cryptanalysis helped him gain a completely revolutionary perspective on the problem of communication – and vice-versa.

1.1.1 Privacy against Inference

One of the central problems in managing privacy on the Internet lies in the simultaneous management of both private and public data. Many users are willing to release some information about themselves, such as their movie watching history, or their gender; they do so because this data enables useful services, and is often not deemed sensitive. However, users also have data that they consider private – their political or sexual orientation, income

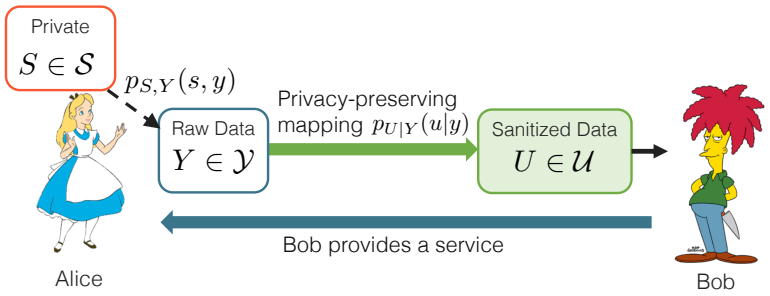

Figure 1-1: Setup for Privacy against Inference. Alice wishes to share some data in exchange for a service, but first sanitizes it using a privacy-preserving mapping so that Bob cannot infer her private attributes. The probability distributions are known to both Alice and Bob.

levels, health status, etc. When releasing public data, users are thus susceptible to inference attacks, where an adversary attempts to infer their private attributes from the public observations. Thus, instead of releasing her raw data, a user might want to modify the latter as to sanitize it, see Figure 1-1. This is done via a privacy-preserving mapping. A main contribution of this thesis, is an in-depth study of the privacy-utility trade-off introduced in [60]. The framework can be characterized as an optimization problem, where a privacy mapping must be found such that privacy is preserved, while the utility is above a specified threshold. One instantiation of this privacy-utility trade-off is the the privacy distortion trade-off

$$\begin{aligned} \min_{P_{U|Y}} \quad & I(S;U) \\ \text{s.t.} \quad & \mathbb{E}[d(Y;U)] \leq \Delta, \end{aligned} \tag{1.1}$$

where on the one hand, privacy is captured by the mutual information $I(S;U)$ between the private attribute S, and the sanitized data U, and on the other hand a measure of utility or distortion $d : \mathcal{Y} \times \mathcal{U} \to \mathbb{R}^+$ is specified between U and the original data Y. When the utility metric is also a mutual information, the resulting optimization is called the Privacy-Funnel [96], and turns out to have deep connections with various subjects in Information Theory (Mrs Gerber's Lemma [150], Strong Data Processing Inequalities[112]), as well as in Statistical Learning (Information Bottleneck[139]). In chapter 2 of this thesis, we will discuss various instantiations of privacy-utility trade-offs, with an emphasis on algorithms and insights to solve the resulting optimization problems either exactly or approximately. Our contributions are three-fold (1) theoretical guarantees of the Privacy Funnel, and the Privacy-distortion formulations, along with connections with information theoretic concepts, and in the privacy

literature (e.g. differential privacy); (2) How does one solve the optimizations and obtain the optimal privacy-preserving mappings efficiently; (3) Practical considerations when looking at real-world settings. The methods we develop are rooted in theory, but are adaptable to the constraints of real-world data, as evidenced by several experiments showcased in the thesis.

1.1.2 Guesswork and Brute-force Security

As previously stated, human-generated passwords are often far from random. Numerous studies based on large password leaks reveal that some passwords are widely more popular than others [33, 142, 140], that passwords are re-used across platforms [57], and that they often contain personal information [143]. These three observations indicate that guessing one's password might be easier than expected, and that the consequences of a breach may be dramatic (in the case of password reuse across many platforms). This means that a brute-force attack is a great threat which should not be underestimated. To quantify this threat, we use Guesswork as a surrogate for the computational cost an adversary has to pay to breach the system. More precisely, a password is modeled as the realization of a random variable X, drawn according to a probability distribution. When this probability distribution is far from uniform, an adversary can more easily guess the password. The guesswork precisely captures this intuition, where $G_P(X)$ is the position of the password X in the list of potential passwords, sorted from most likely to least likely according to the distribution P. Studying this quantity gives insight on the risks associated with poor passwords. An interesting characterization of guesswork is possible when looking at the specific case of passwords of increasing lengths. In this setup, one can make use of mathematical tools from Information Theory, to circumvent the combinatorial nature of the problem, and obtain the asymptotic behavior of guesswork.

In this thesis, we will consider various brute-force attack scenarios. We will discuss attacks performed via a botnet, where distributed machines which are completely uncoordinated, attempt to query passwords to breach into a system. In this case, it is impossible to construct the optimal list and query passwords one after the other, as the attacking agents cannot coordinate. Despite this, we show that the asymptotic performance of the attackers does not change, as they can employ a randomized strategy, and draw their guesses according to a distribution. Perhaps surprisingly, the best guessing distribution is not, in fact, the

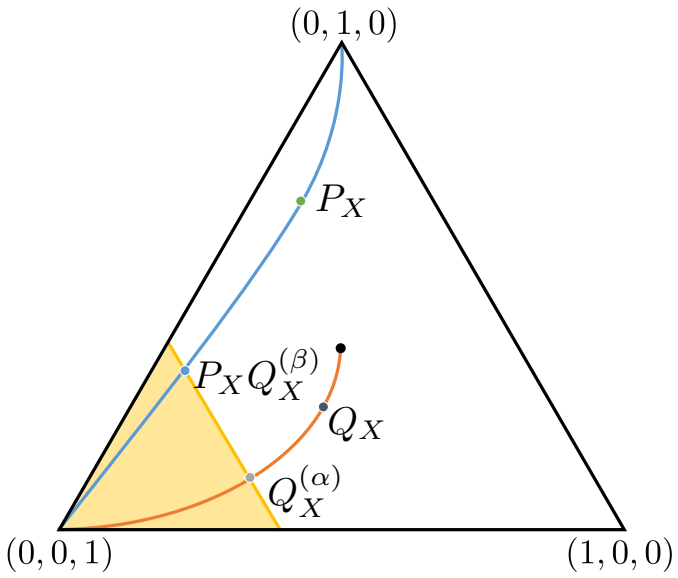

Figure 1-2: Figure 4: Representation of the 3-dimensional simplex. Each point in the triangle corresponds to a distribution over 3 elements. The blue and orange line correspond to the tilted family that govern the matched (orange) and mismatched (blue) guessworks, when P is the true distribution and Q the mismatched.

distribution which was used to generate the passwords, but rather a tilted version of this distribution. We also study a scenario where several attacking agents have access to some side-information about the password. This could be personal details about the user (targeted attack), or knowledge of some previously leaked password (password reuse). There, we ask whether it is best to have a lot of side-information, or a lot of attacking agents – it turns out that side-information is key, and in many setups better side-information is more valuable than any finite number of agents, asymptotically. We will also discuss the case of mismatch in the adversary's knowledge of the password generating distribution P, and derive the resulting guesswork when using this mismatched distribution to construct the list. An interesting consequence of this setup is that an adversary with a mismatch may still perform exactly as well as if there was no mismatch at all, if the mismatched distribution lies on a specific family of distributions. A main contribution of our work is also in the technical tools we develop as proof techniques of the analysis of guesswork. In Figure1-2, an overview of some of the results are shown using a geometric perspective which we developed. This geometric viewpoint showcases an elegant structure in the problem of guessing and is of independent interest in Information Theory, both as a proof and an interpretation tool.

1.2 A brief overview of existing solutions

The goal of this section is to introduce some of the existing solutions in security and privacy. The selected references and notions do not aim at drawing a complete picture of the literature in security and privacy – such a survey would be out of scope for this thesis. However, we will provide hints of the current technical landscape in the area, and discuss why there is a need for novel solutions to address some of the new challenges that arise in the big-data era. Where relevant, we will provide survey papers for the interested reader, and focus the discussion on the high level shortcomings.

1.2.1 Cryptographic security

Cryptographic security[1] is the one major toolset available to the practitioner interested in securing a computer system, be it online or offline. Originally, cryptography was synonymous with secure encryption – a method to hide private messages such that unauthorized users (referred as eavesdroppers) are unable to recover the content of the message, while the authorized users can decode and communicate efficiently. Nowadays, modern cryptographic techniques are at the base of numerous applications which impact most of us on a daily basis. The advent of e-commerce, one of the fastest growing industries of the past decade, relies on cryptographic security to provide authentication, perform secure payments online, and guarantee privacy. Recently, digital currencies such as BitCoin have also originated via the use of cryptographic primitives. Needless to say that there are many more applications of cryptography, new and old, and there is no doubt that cryptographic techniques are powerful, and essential, in building tomorrow's information age.

Perhaps the most relevant application of cryptography to the topic of privacy against inference is given by the recent field of functional encryption (FE) [32, 76, 77] – a scheme that allows function computation on ciphertext. More precisely, letting y be Alice's data, and $f(y)$ be the desired function to be computed, under FE Alice releases an encrypted ciphertext $\mathrm{Enc}(y, k)$ with k being a secret key, such that there exist an efficient function g such that $f(y) = g(\mathrm{Enc}(y), k)$. In other words, Alice may intentionally reveal part of her encrypted data (the function output) to Bob if he is an authorized entity. With a FE scheme, Alice may disclose her personal data Y for the purpose of receiving a specific service

[1] In this thesis, we refer any scheme with computational hardness assumptions on the adversary as cryptographic security.

$f(Y)$– and nothing more.

Similarly, we mention secure multi-party computation systems (MPC), see [59] and references therein for a survey of results. In MPC, the goal is for a group of users to jointly compute a function of private data, while keeping their own data cryptographically secure. More precisely, m participants have access to private data $y_1, \ldots, y_m$, and wish to jointly compute a function $f(y_1, \ldots, y_m)$, while keeping their own private data y_i hidden from the other participants. In principle, MPC also provides a solution to disclosure of personal data online, via the means of encryption.

The final approach we mention is given via the means of fully homomorphic encryption (FHE), see [11]. FHE allows operations to be performed on the cyphertext, such that the result of the decoding process is the desired operation performed on the raw data. In other words, Bob may perform operations on the ciphertext $\mathrm{Enc}(y)$, such that the results of the decoding output is the desired function. Note that, while in in FE, Bob has access to the function $f(y)$ in *plaintext*, in FHE, decoding must happen at Alice's end, i.e., Bob may only perform the computation but may not observe the result. This technique can thus be used to remove privacy barriers in several applications, by allowing operations on the personal data of Alice without ever revealing the content of the data itself.

There are, however several main differences in the threats models which ought to be emphasized:

- **Computation hardness assumptions:** In MPC and related methods, it is assumed that the adversary has some computational restrictions. The methods we will describe in this thesis regarding privacy against inference make no such assumption, and rather provide fundamental information theoretic guarantees on the expected performance of an inference attack.

- **Exactness of the function computation:** Traditionally, the function to be computed is known in advance at all parties [2], and is recovered perfectly following the decoding. Instead, the solutions we propose will introduce statistical noise, i.e., the recovered function will be only approximately and statistically close to the desired function.

- **Key generation:** MPC protocols require keys, whose length depends on the length of

[2]Note that this can be generalized, via fully homomorphic encryption.

the data, and the computational guarantee that one wishes to provide. In the methods we will discuss, no key generation step is assumed.

- **Probabilistic assumption:** In the cryptographic setup, there is no major assumption on the probabilistic distribution of the data d. Instead, the inference setup we introduce considers a probabilistic data generation process for both the data $Y \in \mathcal{Y}$ and the sensitive data $S \in \mathcal{S}$.

- **Knowledge of sensitive data:** Finally, in the MPC setup, the goal is to encrypt the data Y itself – there is no sensitive variable S which is related to the data itself. Instead, the privacy setup we consider is an inference setup in which the sensitive parameters are known in advance, and the schemes depend on the distribution $P_{S,Y}$.

The differences listed above are significant, but we believe that the information theoretic point of view brings an additional dimension to the problem of privacy against inference, and is relevant in practice, as will be seen through the various applications showcased in this thesis. Note that the methods presented can also be complimentary, i.e., we expect that a realistic privacy preserving system should make use of both statistical tools to provide privacy, while also relying on cryptographic principles.

1.2.2 Differential-Privacy

Since its inception in 2006, Differential Privacy[68] (DP) has emerged as one of the main frameworks to design, evaluate and implement privacy preserving data analytics. Privacy systems based on DP, and its numerous generalizations (see e.g. [64, 69] for a survey of results), have been used successfully deployed in the context of statistical databases, deferentially private learning, privacy preserving surveying, and much more. Some highlighted applications of DP are Apple's large-scale private learning of users preferences and behaviors [6], and the 2020 United States Census' privatization method to provide data privacy protection [9], each impacting millions of individuals. In a few words, DP guarantees that the answer of a query be statistically indistinguishable whether an individual participates in the database or not. In other words, even if an adversary had background knowledge of all records in the database before the participation of the an individual, he/she would be unable to infer the private record of the latter from the output of the query. More precisely,

a randomized mapping $\mathcal{M}(y) \to \mathbb{R}$ taking inputs $y \in \mathcal{Y}$ [3] is (ϵ, δ)-DP, if for any pair of *neighboring* y and y', and any measurable set S,

$$\mathbb{P}[\mathcal{M}(y) \in S] \leq e^{\epsilon}\mathbb{P}[\mathcal{M}(y') \in S] + \delta. \tag{1.2}$$

Note that by denoting the output of the randomized mapping $\mathcal{M}(y)$ by U, we can equivalently represent the statistics of the randomized mapping via a channel $P_{U|Y}$, and thus, DP can be seen as a stability property of the distributions $P_{U|Y=y}$ for varying y.

Similar to the methods we will study, the approximately indistinguishable outputs in DP are usually obtained via randomization, that is some form of statistical mechanism is used to obfuscate the individual's personal record. How private a given statistical mechanism is, is captured by the parameters ϵ and δ. The smaller ϵ and δ, the more privacy can be guaranteed, which captures how little the adversary can infer about the individual's record, or equivalently, how many queries he would require to learn a fixed amount about this record. However, mechanisms with high privacy requirements also suffer from loss in utility as a highly private scheme relies on additional noise. Thus, there is also an inherent trade-off between privacy and utility, which in DP manifests itself in the trade-off between the utility of a mechanism, and the privacy parameters (ϵ, δ).

Despite the similarities, we note the following main differences between DP, and in particular local-DP, and the privacy against inference framework we propose:

- **Probabilistic assumption:** In DP, there is no assumption on the data generation process of the data y – the condition (1.2) holds for all neighboring pairs in $\mathcal{Y}$. Instead, the methods we will study in this thesis make an assumption on the probabilistic generation of the data, i.e. $y \sim P_Y$, and provide an expected guarantee over this distribution.

- **Knowledge of the sensitive data:** Note that in DP, there is no distinction between the personal data Y and the sensitive data S. As such, DP schemes are *universal* to the sensitive data S. However, this universality comes at a cost – the DP condition fails to provide information theoretic guarantee on the amount of knowledge an adversary can gain from the observation U, as was shown in [60]. On the other hand, the privacy

[3] The setup we describe here is akin to the *local*-DP setup [85], in which Alice disorts her data before releasing it to the central authority Bob.

against inference framework we will discuss depends on the sensitive data S, and the probabilistic relationship between Y and S.

- **Knowledge of the computed function:** DP schemes are often designed hand-in-hand with the desired query $f(y)$. Thus, they often require cooperation from the service provider Bob in that the query that is sent needs to be known in advance. In the framework we propose, we will modify the data Y itself, in a way which can be made transparent from Bob's point of view.

Despite these differences, local DP and Information theoretic measures of privacy are related as can be seen in [95]. Once again, the methods we propose are complimentary to DP, and can be used hand in hand. In particular, DP can be used as an incentivization mechanism for users to participate in a data collection system, while the privacy against inference framework can be used to provide strong statistical guarantees on the adversary's inference.

Chapter 2

Privacy against Inference

Perhaps the most illuminating example of the modern-age privacy puzzle is given by the Netflix privacy-lawsuit. In 2006, the streaming company launched a public contest with a massive price pool of one million dollars to improve its recommender system. To this end, a large dataset of user data was released to the public – 100 millions movie ratings, from about 480 thousands Netflix users. Researchers from all over the world could use this data to design recommender systems, potentially significantly improving upon Netflix's system of the time. The data was sanitized for privacy, the names of the users were erased and replaced with unique IDs, along with some other mild forms of anonymization. Despite this, it only took several weeks for two researchers from University of Texas, Arvind Narayanan and Vitaly Shmatikov, to de-anonymize several Netflix users, using publicly available information from another movie rating aggregation website. Their publication [106] has since been cited more than two thousand times and is a seminal work on the subject of data-anonymization. But is it really a big privacy breach to learn about someone's movie taste – enough to warrant Netflix ultimately settling the lawsuit at the cost of nine millions dollars? The answer to this question boils down to the so-called *Brokeback Mountain factor*. The major privacy leak was not limited to movie taste, but rather a personal and sensitive attribute of some users: in the case of the Netflix lawsuit, sexual orientation. This is why a closeted lesbian mother joined the lawsuit, because she believed that "were her sexual orientation public knowledge, it would negatively affect her ability to pursue her livelihood and support her family and would hinder her and her children's ability to live peaceful lives" [121]. In other words, the privacy risk is in the inference threat – from the movie ratings, what can be deduced

about a person's sexual orientation. In fact, the issue goes beyond sexual orientation, as it was shown that political leaning, race, gender, social status, and other potentially sensitive attributes can be inferred from movie ratings.

This simultaneous management of data that is publicly shared, e.g. movie ratings, and data that is sensitive, e.g. sexual orientation, is really at the core of the privacy conundrum. Privacy leaks also happen in unexpected ways, especially when the boundary between sensitive data and public data is unclear. For example, while some TV channels and programs are clearly indicative of a person's political leaning, there is also more indirect correlation, e.g., fans of the NBA tend to be more liberal, and fans of the NFL more conservative. Without a systematic and global understanding of the inference threat, users are bound to inadvertently leak information about their private attributes. The object of this chapter is to provide this theoretical foundation to quantify the privacy threat, and subsequently, to design privacy-enhancing solutions. A critical component of our discussion is related to the fundamental privacy-utility trade-off. As shall be made formal in the following sections, there is an inherent tension between leakage of personal information, and quality of the service one receives in exchange from personal data. An instance of this trade-off is illustrated in the movie rating problem: sharing all your ratings improves the quality of your recommendation, but at the expense of additional personal data leakage. Privacy is often thought of as an all-or-nothing issue, but in our work, we provide nuance by discussing operating points in between the complete privacy and the complete disregard for privacy cases. In fact, we argue that this viewpoint is essential in tackling some of the important applications in the era of big data. In certain domains such as genomic data, privacy is a critical concern – because of ethical considerations, and heavy regulations. Laboratories which collect such data may, in fact, not have an option, and are bound by law to provide strong privacy guarantees to their patients. On the other hand, a movie streaming company, a large online social network platform, or an online advertising group are less likely to provide complete privacy if it jeopardizes their revenue significantly. Similarly, users of such services might also be willing to give up part of their privacy in exchange for a better service. Therefore, for data aggregating entities, it is essential to capture the best privacy that can be achieved within a given accuracy budget. The dual problem of obtaining the best service within a privacy budget is relevant for users. Characterizing precisely this optimal trade-off is a major goal of our work. On a technical level, we build upon the privacy versus inference

formulation which was introduced in [60], and expand this setup on its properties in several ways, as explained in the contributions to follow.

Main Contributions and Organization of this Chapter:

To address the issues discussed above, we have organized this chapter as follows. First, in Section 2.1, we will provide the mathematical formulation of the privacy versus inference problem, starting with tools from statistical learning theory. Using this formulation, we will develop in Section 2.2 a privacy distortion trade-off problem which will be characterized as an optimization problem, and discuss its properties. Section 2.3 is devoted to the practical issues around the design of privacy mappings, and we will develop several tools which allows to efficiently solve the large scale optimization induced by the privacy-distortion formulation, and discuss the case of uncertainty in the knowledge of the prior distribution $P_{S,X}$. Finally, in Section 2.4, we introduce the Privacy Funnel as an optimization which captures the fundamental information theoretic trade-off between privacy and utility. Despite the non-convexity of this problem, we discuss some applications and approximations to the Privacy Funnel method. Our novel key contributions are as follows:

1. Establish the universality of the log-loss as privacy metric by bounding the loss of any bounded loss by $O(\sqrt{I(Y;S)})$. In other words, guaranteeing that the log-loss is small is sufficient to guarantee that any such loss is also bounded.

2. Develop Algorithms to efficiently solve the Privacy-Distortion Optimization by leveraging the structure of the optimization (via sparse mappings), and/or the structure of the prior distribution via quantization.

3. Establish stability results that guarantee small errors in the estimate of the prior distribution $P_{S,X}$ lead to almost-optimal privacy-mappings.

4. Introduce the Privacy Funnel, as an information theoretic formulation of the privacy-utility trade-off, provide a close-form solution of the Privacy-Funnel for the Binary Symmetric Sources, and propose a greedy algorithm to approximately solve the Privacy Funnel for general discrete sources.

Related Work:

Privacy-utility tradeoffs have been studied under either a local privacy setting, or a centralized privacy setting. In the local privacy setting, users do not trust the entity aggregating data. Thus, each user holds her data locally, and processes it according to a privacy-preserving mechanism before releasing it to the aggregator. Local privacy dates back to randomized response in surveys [148], and has been considered in privacy for data mining and statistics [12, 104, 72, 116, 85, 22, 61]. The setup we consider falls under the local privacy setting, since the analyst is assumed to be untrusted, and users wish to protect against statistical inference of private information from data they release to the analyst. In contrast, the framework we study models non-asymptotic privacy guarantees in terms of the inference cost gain that an adversary achieves by observing the released output. Local privacy has also been considered in the differential privacy [62, 63] corpus, e.g. for learning concept classes [85], training clustering algorithms [22], and statistical parameter estimation [61], from data distorted locally by users. These works are concerned with the problem of learning aggregate statistical properties from the data of several users. In contrast, we focus on providing utility to an individual user while maintaining the privacy of this individual user's attributes.

In the centralized privacy setting, a trusted entity aggregates data from users in a database, while an untrusted analyst queries the database. The trusted aggregator jointly processes data from multiple users according to a centralized privacy-preserving mechanism to produce a privatized answer to the query, that is released to the analyst. The centralized privacy setting is less stringent than the local privacy setting. Information theoretic frameworks have been used to analyze privacy-utility tradeoffs in the centralized database setting. One line of work [118, 152, 127] focuses mainly on collective privacy for all or subsets of the entries of a data base, and provide fundamental and asymptotic results on the rate-distortion-equivocation region as the number of data samples grows arbitrarily large. Traditionally, many differential privacy works assumed a centralized setting with a trusted database owner, and focused on making the output of an application running on the database differentially private, e.g. data mining [73], social recommendations [94], recommender systems [99], as well as algorithms for statistical estimators [132, 67], classifiers [46, 119], principal component analysis [47], etc. More specifically, [99] considers the case of

a trusted recommender system who has access to ratings from privacy-conscious users, and addresses the challenge of training a differentially-private recommendation algorithm based on these original ratings. In contrast, we study a local privacy setup where the analyst is not trusted by privacy-conscious users, who wish to protect against statistical inference of private information from data they release to the analyst.

This chapter of the thesis relates to a vast literature on the study of differential privacy introduced in [62, 63]. Differential privacy is studied in many contexts including mechanism design [100, 75], learning theory [70, 30, 61], and data mining [22, 66, 65] (see [69] for a survey of results). Moreover, [12, 98] study the class of adding distortion to the public data to protect privacy and [137, 146] study the use of $k-$anonymity to mask private information in classification.

Several approaches rely on information-theoretic tools to model privacy-utility trade-offs, such as [117, 151, 72, 127]. Indeed, information theory, and more specifically rate-distortion theory, appear as natural frameworks to analyze the privacy-utility trade-off resulting from the distortion of correlated data. Although the approach we introduce in this thesis involves information theoretic metrics, it is fundamentally different from previous information theoretic privacy models. Indeed, traditional information theoretic privacy models, such as [151, 127, 126], focus on collective privacy for all or subsets of the entries of a database, and provide asymptotic guarantees on the average remaining uncertainty per database entry – or equivocation per input variable – after the output release. More precisely, the average equivocation per entry is modeled as the conditional entropy of the input variables given the released output, normalized by the number of input variables. In contrast, the general framework introduced in this thesis provides privacy guarantees in terms of bounds on the inference cost gain that an adversary achieves by observing the released output. The use of a self-information cost yields a non-asymptotic information theoretic framework modeling the privacy risk in terms of information leakage. This framework, in turn, can be used to design practical privacy preserving mappings.

Finally, mutual information as a measure of privacy has been used in the literature (see, e.g., [44, 154, 45]), mostly under the context of quantitative information flow and anonymity systems. The connections between different privacy notions have been studied recently, e.g., [13, 103, 147]. Several works have studied a rate-distortion approach to privacy including [128, 16, 105, 23, 17, 34]. More recently, generalizations to the privacy-utility trade-offs

have been considered, e.g. [113] measures the privacy leakage in terms of total variation; [18, 107] consider privacy against guessing attacks; [90, 89] study privacy guarantees under α-maximum leakage; [91, 88, 134] are concerned with privacy against an adversary performing a hypothesis test; the estimation formulations of the privacy utility trade-offs have also been extensively considered in [19, 144, 145]. We also mention [41, 114] which study the fundamental limits of perfect privacy. Finally, [82] takes a data-driven approach to the privacy funnel problem.

2.1 Preliminaries

We start by reviewing the general threat model which was introduced in [60]. We assume that there are two parties that communicate over a noiseless channel, Alice and Bob. Alice has access to a set of measurement points, represented by the r.v. $X \in \mathcal{X}$, that she wishes to transmit to Bob. Simultaneously, Alice also requires that a set of variables $S \in \mathcal{S}$ should remain private, where S is jointly distributed with X according to the distribution $(X, S) \sim P_{X,S}(x, s)$, $(x, s) \in \mathcal{X} \times \mathcal{S}$. Depending on the considered setting, the variable S can be either directly accessible to Alice or inferred from X. If no privacy mechanism was in place, Alice would simply transmit X to Bob.

Bob has a utility requirement for the information sent by Alice. Furthermore, Bob will try to learn S from Alice's transmission. Alice's goal is thus to find and transmit a sanitized version of X, denoted by $Y \in \mathcal{Y}$, such that Y satisfies a target utility constraint, but "protects" (in a sense made more precise later) the private variable S. In the settings we will consider in this thesis, Bob is passive but computationally unbounded, and will try to infer S based on Y. This setting is also known as honest but curious adversary.

We consider, without loss of generality, that $S \to X \to Y$. Note that this model can capture the case where S is directly accessible by Alice by appropriately adjusting the alphabet $\mathcal{X}$. For example, this can be done by representing $S \to Y$ as an injective mapping or allowing $\mathcal{S} \subset \mathcal{X}$. In other words, even though the privacy mechanism is designed as a mapping from $\mathcal{X}$ to $\mathcal{Y}$, it is not limited to an output perturbation, and it encompasses input perturbation settings.

Definition 1. A privacy-preserving mapping is a transition probability $P_{Y|X}(y|x)$, $x \in \mathcal{X}$, $y \in \mathcal{Y}$. A distortion, or utility measure, is a function $d : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$. We say a privacy

mapping $P_{Y|X}$ has δ-distortion for some $\delta \geq 0$, if $\mathbb{E}[d(X,Y)] \leq \delta$ when $(X,Y) \sim P_X P_{Y|X}$.

Throughout this preliminaries section, we make the following assumptions:

1. Alice and Bob know the prior distribution of $P_{X,S}(\cdot)$. This represents the side information that an adversary has. In Section 2.3.4, we relax this assumption to the case where only P_X is known.

2. Bob has complete knowledge of the privacy preserving mapping, i.e., g and $P_{Y|X}(\cdot)$ are known.

Note that this represents the *worst-case* statistical side information that an adversary can have about the input. In Section 2.3.4 we will discuss the case where the knowledge of $P_{S,X}$ is inexact.

2.1.1 Threat model

We assume that Bob selects a revised distribution $q \in \mathcal{P}_S$, where $\mathcal{P}_S$ is the set of all probability distributions over $\mathcal{S}$, in order to minimize an expected cost $C(S,q)$. The cost $C : \mathcal{S} \times \mathcal{P}_S \to \mathbb{R}^+$ models the statistical risk or cost, of picking an estimator q to estimate the random variable S. In other words, the adversary chooses q as the solution of the minimization

$$c_0^* = \min_{q \in \mathcal{P}_S} \mathbb{E}_S[C(S,q)]$$

prior to observing Y, and

$$c_y^* = \min_{q \in \mathcal{P}_S} \mathbb{E}_{S|Y}[C(S,q)|Y=y]$$

after observing the output Y. Note that this restriction on Bob models a very broad class of adversaries that perform statistical inference, capturing how an adversary acts in order to infer a revised belief distribution over the private variables S when observing Y. After choosing this distribution, the adversary can perform an estimate of the input distribution (e.g. using a MAP estimator). However, the quality of the inference is inherently tied to the revised distribution q.

The average cost gain by an adversary after observing the output is

$$\Delta C = c_0^* - \mathbb{E}_Y[c_y^*]. \tag{2.1}$$

We also mention that one can represent similarly the maximum cost gain by an adversary in terms of the most informative output (i.e. the output that give the largest gain in cost), via:

$$\Delta C^* = c_0^* - \min_{y \in \mathcal{Y}} c_y^*. \tag{2.2}$$

In the next section we present a formulation for the privacy-utility tradeoff based on this general setting.

2.2 The Privacy-Distortion Trade-off

Our goal is to design privacy preserving mappings that minimize ΔC for a given distortion level D, characterizing the fundamental privacy-utility tradeoff. More precisely, our focus is to solve optimization problems over $P_{Y|X} \in \mathcal{P}_{Y|X}$ of the form

$$\begin{aligned} &\min \ \Delta C \\ &\text{s.t.} \ \ \mathbb{E}_{X,Y}[d(X,Y)] \leq \delta \ , \end{aligned} \tag{2.3}$$

where $\mathcal{P}_{Y|X}$ is the set of all conditional probability distributions of Y given X.

Remark 1. In the remainder of the chapter we consider only one distortion constraint. However, it is straightforward to generalize the formulation and the subsequent optimization problems to multiple distinct distortion constraints

$$\mathbb{E}_{X,Y}[d_1(X,Y)] \leq \delta_1, \ldots, \mathbb{E}_{X,Y}[d_n(X,Y)] \leq \delta_n.$$

This can be done by simply adding linear constraints to the optimization problem.

In principle, the formulation introduced above is general and can be applied to different cost functions. Throughout the chapter, we specialize this formulation to the log-loss, or self-information cost. We will show subsequently how the log-loss can be used to bound

any other loss function. In addition to its generality, the log-loss has additional convenient advantages. Namely, it is a local, proper and differentiable loss, which will, as we will see, lead to a convex optimization formulation for privacy-utility trade-offs. For an overview of the central role of the self-information cost function in prediction, we refer the reader to [102]. Nevertheless, it is important to emphasize that many of the results presented in this chapter hold for more general loss functions, at the expense of additional notation.

The *self information* (or *log-loss*) cost function is given by

$$C(S, q) = -\log q(S).$$

It is straightforward to show that for the log-loss function $c_0^* = H(S)$ and, consequently, $c_y^* = H(S|Y = y)$, and, therefore

$$\Delta C = I(S;Y) = \mathbb{E}_Y[D(P_{S|Y}||P_S)],$$

From this definition, the optimal privacy-preserving mapping (the one with privacy $G_d(\delta, P_{S,X})$) is the solution of the minimization

$$\begin{aligned} \min_{P_{Y|X}} \quad & I(S;Y) \\ \text{s.t.} \quad & \mathbb{E}_{X,Y}[d(X,Y)] \leq \delta \ . \end{aligned} \tag{2.4}$$

In extreme cases, we say a privacy-mapping has full privacy if $I(S;Y) = 0$ (which implies the released random variable, Y, is independent from the private random variable, S), and no privacy if $I(S;Y) = H(S)$ (implies that S is fully recoverable from Y).

Observe that finding the mapping $P_{Y|X}(y|x)$ that provides the minimum information leakage is a modified rate-distortion problem. Alternatively, we can rewrite this optimization as

$$\begin{aligned} \min_{P_{Y|X}} \quad & \mathbb{E}_Y[D(P_{S|Y}||P_S)] \\ \text{s.t.} \quad & \mathbb{E}_{X,Y}[d(X,Y)] \leq \delta \ . \end{aligned} \tag{2.5}$$

The minimization (2.5) has an interesting and intuitive interpretation. If we consider

KL-divergence as a metric for the distance between two distributions, (2.5) states that the revised distribution after observing Y should be as close as possible to the a priori distribution.

We are now ready to define the privacy-utility region.

Definition 2. For $D \geq 0$, distortion measure $d : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$, and a joint distribution $P_{S,X}$ over $\mathcal{S} \times \mathcal{X}$, we define the optimal *privacy-utility function* $G_d(D, P_{S,X})$ as

$$G_d(D, P_{S,X}) \triangleq \inf \{I(S;Y) : \mathbb{E}[d(X,Y)] \leq D, S \to X \to Y\}, \tag{2.6}$$

where the infimum is over all mappings $P_{Y|X}$ such that $\mathcal{Y}$ is finite. For a fixed $P_{S,X}$ and $D \geq 0$, the set of pairs $\{(D, G_d(D, P_{S,X}))\}$ is called the *privacy-utility region* of $P_{S,X}$.

We next characterize a property of the optimal privacy-preserving mapping which will be useful in Section 2.3 to construct solutions to the optimization problem 2.4. In particular, the next lemma suggests that the size of the output alphabets $|\mathcal{Y}|$ one need to consider is bounded by $|\mathcal{X}| + 1$. This lemma will be used in Section 2.3 when we find to design algorithms to find the optimum privacy-preserving mapping.

Lemma 1. *We have*

$$G_d(D, P_{S,X}) = \min_{P_{Y|X}} \{I(S;Y) : \mathbb{E}[d(X,Y)] \leq D, \\ S \to X \to Y, |\mathcal{Y}| \leq |\mathcal{X}| + 1\}.$$

Proof. Let $P_{S,X}$ and $P_{Y|X}$ be given, with $S \to X \to Y$. Denote by $\mathbf{w}_i$ the vector in the $|\mathcal{X}|$-simplex with entries $P_{X|Y}(\cdot|i)$. Furthermore, let $a_i \triangleq \mathbb{E}[d(X,Y)|Y = i]$, and $b_i \triangleq H(S) - H(S|Y = i)$. Therefore

$$\sum_{i=1}^{|\mathcal{Y}|} P_Y(i) [\mathbf{w}_i, a_i, b_i] = [P_X, \mathbb{E}[d(X,Y)], I(S;Y)]. \tag{2.7}$$

Since $\mathbf{w}_i$ belongs to the $|\mathcal{X}|$-simplex, the vector $[\mathbf{w}_i, a_i, b_i]$ is taken from a connected, compact $|\mathcal{X}| + 1$ dimensional space. Then, from Fenchel-Eggleston strengthening of Carathéodory's theorem [71, Theorem 18, pg. 35], the point $[\mathbf{p}_X, \mathbb{E}[d(X,Y)], \Delta C]$ can also be achieved by at most $|\mathcal{X}| + 1$ non-zero values of $P_Y(i)$. It follows directly that it is sufficient to consider

$|\mathcal{Y}| \leq |\mathcal{X}| + 1$ for the infimum (2.43). The set of all mappings $P_{Y|X}$ for $|\mathcal{Y}| \leq |\mathcal{X}| + 1$ is compact, and both $P_{Y|X} \to I(S;Y)$ and $P_{Y|X} \to \mathbb{E}[d(X,Y)]$ are continuous and bounded when S, X and Y have finite support. Consequently, the infimum in (2.43) is attainable. ■

Next, we give an example of the optimization given in (2.5) and its solution.

Example 1. Let S be a Bernoulli($\frac{1}{2}$) distribution and X be the result of S passing through a BSC(p) channel where $p \leq \frac{1}{2}$. Suppose the distortion measure is hamming distortion, i.e. $\mathbb{E}[d(X,Y)] = \mathbb{P}[X \neq Y]$, and consider the log-loss. We claim that in this setting for a given $\delta \in (0,1)$, we have

$$G_d(\delta, P_{S,X}) = 1 - h_b(p * \delta),$$

where $p * \delta = p(1-\delta) + (1-p)\delta$. First, note that using the privacy-preserving mapping, $P_{Y|X}$, given by $Y = X \oplus Z$, where Z has a Bern(δ) distribution, we have $\mathbb{E}[d(X,Y)] \leq \delta$ and $I(S;Y) = 1 - h(p * \delta)$. This shows that $G_d(\delta, P_{S,X}) \leq 1 - h_b(p * \delta)$. Next, we show that $G_d(\delta, P_{S,X}) \geq 1 - h_b(p * \delta)$. We have $I(S;Y) = H(S) - H(S|Y) = 1 - H(S \oplus Y|Y) \geq 1 - H(S \oplus Y)$. Using Markov property, it follows that $\mathbb{P}[S \oplus Y = 1] \leq p * \delta$, which completes the proof of the claim. Now suppose we want to have full privacy. Given $G_d(D, P_{S,X}) = 1 - h_b(p * D)$, full privacy is possible only in the following two cases:

1. $p = \frac{1}{2}$, implying X is independent from S. In this case, there is no privacy problem to begin with.

2. $\delta = \frac{1}{2}$, implying Y is independent from X. In this case, full privacy implies no utility is preserved in the released data.

2.2.1 Generality of log-loss as a privacy metric

In this section, we focus on the threat model under the log-loss cost function and show its generality. In particular, we establish that for any bounded cost function $C(S,q)$, the associated inference cost gain ΔC can be upperbounded by an explicit constant factor of $\sqrt{I(S;Y)}$. Thus, controlling the cost gain under the log-loss, so that it does not exceed a target privacy level, is sufficient to ensure that the privacy threat under a different bounded cost function would also be controlled. Therefore, the design of the privacy mapping can be focused on minimizing the privacy leakage as measured by $I(S;Y)$.

Theorem 1. *Let* $L = \sup_{s\in\mathcal{S},q\in\mathcal{P}_S} |C(s,q)| < \infty$. *We have* $\Delta C = c_0^* - \mathbb{E}_{P_Y}[c_Y^*] \leq 2\sqrt{2}L\sqrt{I(S;Y)}$.

The proof of Theorem 1 requires the following lemma.

Lemma 2. *Let* $C(s,q)$ *be a bounded cost function such that* $L = \sup_{s\in\mathcal{S},q\in\mathcal{P}_S} |C(s,q)| < \infty$. *For any given* $y \in \mathcal{Y}$,

$$\mathbb{E}_{P_{S|Y}}[C(S,q_0^*) - C(S,q_y^*)|Y=y] \leq 2\sqrt{2}L\sqrt{D(P_{S|Y=y}||P_S)},$$

where q_0^* *and* q_y^* *are the maximizing distributions for* c_0^* *and* c_y^* *as defined in Section 2.1.1, respectively.*

Proof. We expand $\mathbb{E}_{P_{S|Y}}[C(S,q_0^*) - C(S,q_y^*)|Y=y]$ and have:

$$\begin{aligned}
&\sum_s p(s|y)[C(s,q_0^*) - C(s,q_y^*)] \\
&= \sum_s (P_{S|Y}(s|y) - P_S(s) + P_S(s))[C(s,c_0^*) - C(s,q_y^*)] \\
&= \sum_s (P_{S|Y}(s|y) - P_S(s))[C(s,q_0^*) - C(s,q_y^*)] \\
&\qquad + \sum_s p(s)[C(s,q_0^*) - C(s,q_y^*)] \\
&\leq 2L\sum_s |p(s|y) - p(s)| + (\mathbb{E}_{P_S}[C(S,q_0^*)] - \mathbb{E}_{P_S}[C(S,q_y^*)]), \\
&\leq 2L\sum_s |P_{S|Y}(s|y) - P_S(s)| \\
&= 4L||P_{S|Y=y} - P_S||_{TV} \\
&\leq 4L\sqrt{\frac{1}{2}D(P_{S|Y=y}||P_S)},
\end{aligned}$$

where we used that $C(s,q_0^*) - C(s,q_y^*) \leq 2L$ and $\mathbb{E}_{P_S}[C(S,q_0^*)] - \mathbb{E}_{P_S}[C(S,q_y^*)] \leq 0$. And the last inequality follows from using Pinsker's inequality [54, Problem 3.18] (where the log in the definition of divergence is natural log). ■

We now prove Theorem 1.

proof of Theorem 1. We have

$$\Delta C = \mathbb{E}_{P_S}[C(S,q_0^*)] - \mathbb{E}_{P_Y}\left[\mathbb{E}_{P_{S|Y}}[C(S,q_y^*)|Y=y]\right]$$

$$= \mathbb{E}_{P_Y}\left[\mathbb{E}_{P_{S|Y}}[C(S,q_0^*) - C(S,q_y^*)|Y=y]\right]$$
$$\leq 2\sqrt{2}L\mathbb{E}_{P_Y}\left[D(P_{S|Y=y}||P_S)\right] \leq 2\sqrt{2}L\sqrt{I(S;Y)},$$

where the last step follows from concavity of square root function and the one before that follows from Lemma 2. ■

Another important property of the log-loss is that it is a *proper loss* function, i.e., for any $S \sim P_S$, $\min_q \mathbb{E}[C(q,S)] = \mathbb{E}[C(P_S,S)]$. In other words, a proper loss function can be minimized by using the true distribution P_S. The next proposition shows that, under some regularity conditions, the log-loss $C(q,S) = -\log q(S)$ is in fact the unique proper loss-function.

Proposition 1. *Let $C(q,S)$ be smooth and differentiable in q, and assume that it takes the form $C(S,q) = F(q(S))$ for some function F. If $\operatorname{argmin}_q \mathbb{E}[C(S,q)] = P_S$, then $C(q,S) = -A\log q(S) + B$ for some constants $A, B \in \mathbb{R}$ with $A > 0$.*

Proof. This can be proved in several ways, see e.g. [10]. The proof sketch below reduces the problem to a differential equation, whose solution is given by the log-loss functional.

Note that since $C(q,S) = F(q(S))$ is differentiable in q, we have that the functional $J(q) = \nabla_q \mathbb{E}[F(q(S))] + \lambda$ must evaluate to zero at $q = P_S$, where $\lambda > 0$ is a Lagrange multiplier which enforces that $\sum_{s\in\mathcal{S}} q(s) = 1$. Denoting $q_i \triangleq q(s_i)$, for s_i, $i = 1,\ldots,|\mathcal{S}|$ an indexing of $\mathcal{S}$, we have:

$$[J(q)]_i = P_S(i)\frac{\partial}{\partial q_i}F(q_i) + \lambda, \tag{2.8}$$

for $i = 1,\ldots,|\mathcal{S}|$. Now evaluating (2.8) at $q_i = P_S(i)$ and equaling to zero, and noting that this is true for any $P_S(i)$, we have that any proper cost function must satisfy a differential equation of the form, where we have used the change of variables $x = P_S(i)$, and with $F'(x) = \frac{\partial}{\partial x}F(x)$ and $A = \lambda$:

$$xF'(x) + A = 0, \tag{2.9}$$

The differential equation in (2.9) can be solved directly and has solutions of the type $F(x) = -A\log x + B$, for constant $B \in \mathbb{R}$ and $A > 0$, which concludes the proof. ■

2.2.2 Inference Defeat through Privacy

One natural and related question is whether a privacy mapping which is designed to minimize average information leakage, privacy, by solving problem (2.4), also provides guarantees on the probability of correctly inferring S from the observation of Y, using any inference algorithm. Next, we show a lower bound on the error probability in inferring S from Y, based on a bound on privacy, using Fano's inequality.

Proposition 2. *Assume $|\mathcal{S}| > 2$ and $I(S;Y) \leq \epsilon H(S)$, for some $\epsilon \in [0,1]$. Let $\hat{S}$ be an estimator of S based on the observation Y (possibly randomized). We have*

$$P_e \triangleq \mathbb{P}[\hat{S}(Y) \neq S] \geq \frac{(1-\epsilon)H(S) - 1}{\log(|\mathcal{S}| - 1)}.$$

For $|\mathcal{S}| = 2$, we have $h(P_e) \geq (1-\epsilon)H(S)$.

Proof. Denote $P_e = \mathbb{P}[\hat{S}(Y) \neq S]$. From Fano's inequality [53], Theorem 2.10.1, we have

$$P_e \left(\log(|\mathcal{S}| - 1)\right) \geq H(S|Y) - h(P_e).$$

Since $I(Y;S) = H(S) - H(S|Y) \leq \epsilon H(S)$, we have $H(S|Y) \geq (1-\epsilon)H(S)$. Therefore,

$$P_e \geq \frac{(1-\epsilon)H(S) - h(P_e)}{\log(|\mathcal{S}| - 1)} \geq \frac{(1-\epsilon)H(S) - 1}{\log(|\mathcal{S}| - 1)}.$$

The proof when $|\mathcal{S}| = 2$ is similar. ■

Note that one can obtain tighter bounds than the one in Proposition 2 by considering β-conditional entropies as the privacy metric, as shown in [129]. In particular, as β goes to ∞, the bound becomes tight as the loss considered becomes the 0-1 loss.

2.2.3 Application examples

We illustrate next how the proposed model can be cast in terms of privacy preserving queries and hiding features within data sets.

Privacy-preserving queries to a database

The framework described above can be applied to database privacy problems, such as those considered in differential privacy. In this case we denote the private variable as a vector

$\mathbf{S} = S_1, \ldots, S_n$, where $S_j \in \mathcal{S}$, $1 \leq j \leq n$ and $S_1, \ldots, S_n$ are discrete entries of a database that represent, for example, the entries of n users. A (not necessarily deterministic) function $f : \mathcal{S}^n \to \mathcal{X}$ is calculated over the database with output X such that $X = f(S_1, \ldots, S_n)$. The goal of the privacy preserving mapping is to present a query output Y such that the individual entries $S_1, \ldots, S_n$ are "hidden", i.e. the estimation cost gain of an adversary is minimized according to the previous discussion, while still preserving the utility of the query in terms of the target distortion constraint. We illustrate this case with the counting query, which will be a recurring example throughout the rest of this chapter.

Example 2 (Counting query). Let $S = (S_1, \ldots, S_n)$, where S_i's are the entries in a database, and define:

$$X = f(S_1, \ldots, S_n) = \sum_{i=1}^{n} \mathbf{1}_A(S_i), \tag{2.10}$$

where

$$\mathbf{1}_A(z) = \begin{cases} 1 & \text{if } z \text{ has property } A, \\ 0 & \text{otherwise.} \end{cases}$$

In this case there are two possible approaches: (i) output perturbation, where X is distorted directly to produce Y, and (ii) input perturbation, where each individual entry S_i is distorted directly, resulting in a new query output Y. In particular, if each database input S_i, $1 \leq i \leq n$ satisfies $\mathbb{P}[\mathbf{1}_A(S_i) = 1] = p$ and are independent and identically distributed. Then X is a binomial random variable with parameter (n, p). It follows that $H(S|X = x) = \log \binom{n}{x}$. Consequently, the optimal privacy preserving mapping will be the one that results in a posterior probability $P_{X|Y}(x|y)$ that is proportional to the size of the pre-image of x, i.e. $P_{X|Y}(x|y) \propto |f^{-1}(x)| = \binom{n}{x}$.

Hiding dataset features

Another important particularization of the proposed framework is the obfuscation of a set of features S by distorting the entries of a data set X. In this case $|\mathcal{S}| \ll |\mathcal{X}|$, and S represents a set of features that might be inferred from the data X, such as age group or salary. The distortion can be defined according to the the utility of a given statistical learning algorithm (e.g. a recommendation system) used by Bob.

2.3 Design of Privacy Preserving Mappings

In this section, we consider the problem of finding optimal privacy mapping by solving (2.4). We will discuss several dimensions which are relevant when designing privacy mappings: 1) Solving the optimization problems in (2.4) efficiently, and 2) Reliance on the knowledge of the joint distribution $P_{S,X}$. We first discuss the optimization itself.

2.3.1 The Privacy-Distortion Optimization

Consider the optimization given in (2.4). The following theorem shows that the problem can be expressed as a convex optimization problem. We note that this optimization is solved in terms of the unknowns $P_{Y|X}(\cdot|\cdot)$ and $P_{Y|S}(\cdot|\cdot)$, which are coupled together through a linear equality constraint.

Proposition 3. *Given $P_{S,X}(\cdot,\cdot)$, a distortion function $d(\cdot,\cdot)$ and a distortion constraint D, the mapping $P_{Y|X}(\cdot|\cdot)$ that minimizes the average information leakage can be found by solving the following convex optimization (sssuming the usual simplex constraints on the probability distributions):*

$$\min_{P_{Y|X},\ P_{Y|S},\ |\mathcal{Y}|\leq|\mathcal{X}|+1} \sum_{y\in\mathcal{Y}}\sum_{s\in\mathcal{S}} P_{Y|S}(y|s)P_S(s)\log\left(\frac{P_{Y|S}(y|s)}{P_Y(y)}\right) \tag{2.11}$$

$$\text{s.t.} \quad \sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} P_{Y|X}(y|x)P_X(x)d(y,x) \leq D, \tag{2.12}$$

$$\sum_{x\in\mathcal{X}} P_{X|S}(x|s)P_{Y|X}(y|x) = P_{Y|S}(y|s)\ \forall y,s, \tag{2.13}$$

$$\sum_{s\in\mathcal{S}} P_{Y|S}(y|s)P_S(s) = P_Y(y)\ \forall y. \tag{2.14}$$

Proof. Clearly the previous optimization is the same as (2.4). To prove the convexity of the objective function, note that $h(x,a) = ax\log x$ is convex for a fixed $a \geq 0$ and $x \geq 0$, and, therefore, the perspective of $g_1(x,z,a) = ax\log(x/z)$ is also convex in x and z for $z > 0, a \geq 0$ ([36]). Since the objective function (2.11) can be written as

$$\sum_{y\in\mathcal{Y}}\sum_{s\in\mathcal{S}} g(P_{Y|S}(y|s), P_Y(y), P_S(s)),$$

it follows the optimization is convex. In addition, since $p(y) \to 0 \Leftrightarrow p(y|s) \to 0\ \forall y$, the

minimization is well defined over the probability simplex. Finally, the constraint $|\mathcal{Y}| \leq |\mathcal{X}|+1$ follows from Lemma 1. ■

In the particular case where X is a deterministic function of S, the optimization takes a simpler form, as shown in the corollary below.

Corollary 1. *If X is a deterministic function of S and $S \to X \to Y$ then the minimization in* (2.4) *can be simplified to a rate-distortion problem:*

$$\begin{aligned} &\min_{P_{Y|X}} \; I(X;Y) \\ &\text{s.t.} \;\; \mathbb{E}_{X,U}[d(X,Y)] \leq D \; . \end{aligned}$$

Furthermore, by restricting $Y = X + Z$ and $d(X,Y) = d(X-Y)$, the optimization reduces to

$$\begin{aligned} &\max_{P_Z} \; H(Z) \\ &\text{s.t.} \;\; \mathbb{E}_Z[d(Z)] \leq D \; . \end{aligned}$$

Proof. Since X s a deterministic function of S and $S \to X \to Y$, then

$$\begin{aligned} I(S;Y) &= I(S,X;Y) - I(X;Y|S) \\ &= I(X;Y) + I(S;Y|X) - I(X;Y|S) \\ &= I(X;Y), \end{aligned} \tag{2.15}$$

where (2.15) follows from the fact that X is a deterministic function of S ($I(X;Y|S) = 0$) and $S \to X \to Y$ ($I(S;Y|X) = 0$). For the additive noise case, the result follows by observing that $H(X|Y) = H(Z)$. ■

The above formulation allows the use of efficient algorithms for solving convex problems, such as interior-point methods. However, it can also be solved using a dual minimization procedure analogous to the Arimoto-Blahut algorithm [53] by starting at a fixed marginal probability $P_Y(y)$, solving a convex minimization at each step (with an added linear constraint compared to the original algorithm) and updating the marginal distribution.

In either cases, the number of free parameters which need to be optimized is $|\mathcal{X}| \times |\mathcal{Y}|$, as we are optimizing upon the mapping $P_{Y|X}$. This number of parameters is undesirable when $|\mathcal{X}|$ and $|\mathcal{Y}|$ are even moderately large, and quickly becomes intractable in high-dimensional settings. For example, taking X to be a vector of movie ratings from the Movielens dataset [80], the size $|\mathcal{X}|$ is 5^{2800}, as there are 2800 movies which can all be rated from 1 to 5. By Lemma 1, the size of $\mathcal{Y}$ needs to also be comparable to obtain the best trade-off. In other words, despite the optimization being convex, the staggering number of parameters in the optimization does not suit itself to a simple solving, and it is necessary to come up with alternative solutions. In the rest of this section, we propose three strategies to handle this dimensionality issue: sparse mappings, unsupervised clustering, and noise mechanisms. Note that, while we present these techniques as separate, they can be used together and can complement each other. We start by leveraging the specific structure of the solution of the optimization.

2.3.2 Sparse Privacy Preserving Mappings

In this section, we introduce an optimization technique which reduces the number of free parameters in the optimization (2.4). The main idea is based on the following heuristic. While theoretically, the number of parameters in $P_{Y|X}$ is large, we make the assumption that the matrix $\mathbb{P}_{Y|X}$ is in fact sparse, i.e. most entries of $\mathbf{P}_{Y|X}$ are zero. This heuristic turns out to be empirically verified in the *low-privacy* regime, i.e., when D is small, and can be made formal in the limit of $D \to 0$. Indeed, in this case, let $P^*_{Y|X}$ be the optimal privacy preserving mapping, and consider the set $\mathcal{S} = \{(x, y) : P^*_{Y|X}(y|x) > \epsilon, d(x, y) > \delta_{\min}\}$ for some $\epsilon > 0$, and with $\delta_{\min} = \inf\{d(x, y) : d(x, y) > 0, (x, y) \in \mathcal{X} \times \mathcal{Y}\}$. Then, the expected distortion $\mathbb{E}[d(X; Y)]$ can be bounded as:

$$D = \mathbb{E}[d(X;Y)] = \sum_{(x,y)\in\mathcal{S}} P_X(x) P^*_{Y|X}(y|x) d(x, y) \tag{2.16}$$

$$> \epsilon \cdot \delta \sum_{(x,y)\in\mathcal{S}} P_X(x). \tag{2.17}$$

Therefore, for a fixed ϵ and δ, letting $D \to 0$, the bound above implies that $\sum_{(x,y)\in S} P_X(x)$ must go to zero as well. Since $P_X(x)$ is fixed, and cannot be changed by the choice of the privacy-mapping $P_{Y|X}$, it must be the case that $|\mathcal{S}| \to 0$.

The assumption that $P^*_{Y|X}$ is sparse can be leveraged to use more efficient optimization techniques. One such solution, which we will develop further in the rest of this subsection is known as the Dantzig-Wolfe Decomposition, see e.g. [27, Section 6.4].

Before we describe our algorithm, we rewrite Optimization (2.4) compactly. Let $\mathbf{X}$ be a $n \times n$ matrix of optimized variables, whose entries are defined as $x_{i,j} = P_{Y|X}(y_i \mid x_j)$, and let $\mathbf{X}_j$ be the j-th column of $\mathbf{X}$. To highlight the optimization aspect of our problem, we write the objective function $J(P_{S,X}, P_{Y|X})$ as a function $f(\mathbf{X})$, with the understanding that f depends on $P_{S,X}$, which is not optimized, and on $\mathbf{X}$, which is optimized. Similarly the distortion constraint can be written as $\sum_{j=1}^{n} \boldsymbol{d}_j^T \mathbf{X}_j \leq \Delta$, where each $\boldsymbol{d}_j = P_X(x_j)(d(y_1, x_j),\ d(y_2, x_j),\ \ldots,\ d(y_n, x_j))^T$ is a vector of length n that represents the distortion metric scaled by the probability of the corresponding symbol x_j. The marginal of X is computed as $P_X(x_j) = \sum_s P_{S,X}(s, x_j)$. Finally, the simplex constrain can be written as $\mathbf{1}_n \mathbf{X}_j = 1$ for all j, where $\mathbf{1}_n$ is an all-ones vector of length n. Given the new notation, our original problem (2.4) can be written compactly as:

$$\begin{aligned} \underset{\mathbf{X}}{\text{minimize}} \quad & f(\mathbf{X}) \\ \text{subject to} \quad & \sum_{j=1}^{n} \boldsymbol{d}_j^T \mathbf{X}_j \leq \Delta \\ & \mathbf{1}_n^T \mathbf{X}_j = 1 \quad \forall j = 1, \ldots, n \\ & \mathbf{X} \geq 0 \end{aligned} \tag{2.18}$$

where $\mathbf{X} \geq 0$ is an entry-wise inequality.

Franke-Wolfe Linearization

The optimization problem (2.4) has linear constraints but its objective function is non-linear. We propose to solve the problem as a sequence of linear programs, also known as the *Frank-Wolfe method*. Each iteration ℓ of the method consists of three major steps. First, we compute the gradient $\nabla_{\mathbf{X}} f(\mathbf{X}_{\ell-1})$ at the solution from the previous step $\mathbf{X}_{\ell-1}$. The gradient is a $n \times n$ matrix $\mathbf{C}$, where $c_{i,j} = \frac{\partial}{\partial x_{i,j}} f(\mathbf{X}_{\ell-1})$ is a partial derivative of the objective function with respect to the variable $x_{i,j}$. Second, we find a feasible solution $\mathbf{X}'$ in the direction of the gradient. This problem is solved as a linear program with the same constraint as the

original problem:

$$\begin{aligned} \underset{\mathbf{X}}{\text{minimize}} \quad & \sum_{j=1}^{n} \boldsymbol{c}_j^T \mathbf{X}_j \\ \text{subject to} \quad & \sum_{j=1}^{n} \boldsymbol{d}_j^T \mathbf{X}_j \leq \Delta \\ & \mathbf{1}_n^T \mathbf{X}_j = 1 \quad \forall j = 1, \ldots, n \\ & \mathbf{X} \geq 0 \end{aligned} \tag{2.19}$$

where $\boldsymbol{c}_j$ is the j-th column of $\boldsymbol{C}$. Finally, we find the minimum of f between $\mathbf{X}_{\ell-1}$ and $\mathbf{X}'$, $\mathbf{X}_\ell$, and make it the current solution. Since f is convex, this minimum can be found efficiently by ternary search. The minimum is also feasible because the feasible region is convex, and both $\mathbf{X}'$ and $\mathbf{X}_{\ell-1}$ are feasible.

Sparse Approximation

The linear program (2.19) has n^2 variables and therefore is hard to solve when n is large. In this section, we propose an incremental solution to this problem, which is defined only on a subset of *active variables* $\mathcal{V} \subseteq \{1, 2, \ldots, n\} \times \{1, 2, \ldots, n\}$. The active variables are the non-zero variables in the solution to the problem (2.19). Therefore, solving (2.19) on active variables $\mathcal{V}$ is equivalent to restricting all inactive variables to zero. The corresponding linear program is shown in (2.20) in Algorithm 1. This linear program has only $|\mathcal{V}|$ variables. Now the challenge is in finding a good set of active variables $\mathcal{V}$. This set should be small, and such that the solutions of (2.19) and (2.20) are close.

We grow the set $\mathcal{V}$ greedily using the dual linear program of (2.20). In particular, we incrementally solve the dual by adding most violated constraints, which corresponds to adding most beneficial variables in the primal. The dual of (2.20) is (2.22) in Algorithm 2, where $\lambda \in \mathbb{R}$ is a variable associated with the distortion constraint and $\mu \in \mathbb{R}^n$ is vector of n variables associated with the simplex constraints. Given a solution (λ^*, μ^*) to the dual, the most violated constraint for a given j is the one that minimizes:

$$c_{i,j} - \lambda^* d_{i,j} - \mu_j^*. \tag{2.24}$$

This quantity, called the *reduced cost*, has an intuitive interpretation. We choose an example

Algorithm 1 SPPM: Sparse privacy preserving maps

Input: Starting point $\mathbf{X}_0$, number of steps L

for all $\ell = 1, 2, \ldots, L$ **do**

$\boldsymbol{C} \leftarrow \nabla_{\mathbf{X}} f(\mathbf{X}_{\ell-1})$

$\mathcal{V} \leftarrow$ DWD

Find a feasible solution $\mathbf{X}'$ in the direction of the gradient $\boldsymbol{C}$:

$$\begin{aligned} \underset{\mathbf{X}}{\text{minimize}} \quad & \sum_{j=1}^{n} \boldsymbol{c}_j^T \mathbf{X}_j \\ \text{subject to} \quad & \sum_{j=1}^{n} \boldsymbol{d}_j^T \mathbf{X}_j \leq \Delta \\ & \mathbf{1}_n^T \mathbf{X}_j = 1 \quad \forall j = 1, \ldots, n \\ & \mathbf{X} \geq 0 \\ & x_{i,j} = 0 \quad \forall (i,j) \notin \mathcal{V} \end{aligned} \tag{2.20}$$

Find the minimum of f between $\mathbf{X}_{\ell-1}$ and $\mathbf{X}'$:

$$\gamma^* \leftarrow \underset{\gamma \in [0,1]}{\operatorname{argmin}} f((1-\gamma)\mathbf{X}_{\ell-1} + \gamma \mathbf{X}') \tag{2.21}$$

$\mathbf{X}_\ell \leftarrow (1-\gamma^*)\mathbf{X}_{\ell-1} + \gamma^* \mathbf{X}'$

end for

Output: Suboptimal feasible solution $\mathbf{X}_L$

i in the direction of the steepest gradient of $f(\mathbf{X})$, so $c_{i,j}$ is small; which is close to j, so $d_{i,j}$ is close to zero (ss $\lambda^* \leq 0$). The search for the most violated constraint leverages the problem structure. Therefore, our approach can be viewed as an instance of *Dantzig-Wolfe decomposition*.

The pseudocode of our search procedure is in Algorithm 2. This is an iterative algorithm, where each iteration consists of three steps. First, we solve the reduced dual linear program (2.22) on active variables. Second, for each point j, we identify a point i^* that minimize the reduced cost. Finally, if the pair (i^*, j) corresponds to a violated constraint, we add it to the set of active variables $\mathcal{V}$.

The pseudocode of our final solution is in Algorithm 1. We refer to Algorithm 1 as *Sparse Privacy Preserving Mappings (*SPPM*)*, because of the mappings learned by the algorithm. Algorithm 2 is a subroutine of Algorithm 1, which identifies the set of active variables $\mathcal{V}$. SPPM is parameterized by the number of iterations L.

Algorithm 2 DWD: Dantzig-Wolfe decomposition

Initialize the set of active variables:
$\mathcal{V} \leftarrow \{(1,1),(2,2),\ldots,(n,n)\}$

while the set $\mathcal{V}$ grows **do**
Solve the master problem for λ^* and μ^*:

$$\begin{aligned} \underset{\lambda,\mu}{\text{maximize}} \quad & \lambda\Delta + \sum_{j=1}^{n} \mu_j \\ \text{subject to} \quad & \lambda \leq 0 \\ & \lambda d_{i,j} + \mu_j \leq c_{i,j} \quad \forall (i,j) \in \mathcal{V} \end{aligned} \tag{2.22}$$

for all $j = 1, 2, \ldots, n$ **do**
Find the most violated constraint in the master problem for fixed j:

$$i^* = \arg\min_i [c_{i,j} - \lambda d_{i,j} - \mu_j] \tag{2.23}$$

if $(c_{i^*,j} - \lambda d_{i^*,j} - \mu_j < 0)$ **then**
$\mathcal{V} \leftarrow \mathcal{V} \cup \{(i^*, j)\}$
end if
end for
end while

Output: Active variables $\mathcal{V}$

Convergence

Algorithm SPPM is a gradient descent method. In each iteration ℓ, we find a solution $\mathbf{X}'$ in the direction of the gradient at the current solution $\mathbf{X}_{\ell-1}$. Then we find the minimum of f between $\mathbf{X}_{\ell-1}$ and $\mathbf{X}'$, and make it the next solution $\mathbf{X}_\ell$. By assumption, the initial solution $\mathbf{X}_0$ is feasible in the original problem (2.4). The solution $\mathbf{X}'$ to the LP (2.20) is always feasible in (2.4), because it satisfies all constraints in (2.4), and some additional constraints $x_{i,j} = 0$ on inactive variables. After the first iteration of SPPM, $\mathbf{X}_1$ is a convex combination of $\mathbf{X}_0$ and $\mathbf{X}'$. Since the feasible region is convex, and both $\mathbf{X}_0$ and $\mathbf{X}'$ are feasible, $\mathbf{X}_1$ is also feasible. By induction, all solutions $\mathbf{X}_\ell$ are feasible.

The value of $f(X_\ell)$ is guaranteed to monotonically decrease with ℓ. When the method converges, $f(X_\ell) = f(X_{\ell-1})$. The convergence rate of the Frank-Wolfe algorithm is $O(1/L)$ in the worst case [27].

Computational Efficiency

The computation time of our method is dominated by the search for n^2 violated constraints in Algorithm 2. To search efficiently, we implement the following speedup in the computation of the gradients $c_{i,j}$. The marginal and conditional distributions:

$$P_Y(y) = \sum_{s,x} P_{S,X}(s,x) P_{Y|X}(y \mid x) \tag{2.25}$$

$$P_{Y|S}(y \mid s) = \frac{\sum_x P_{S,X}(s,x) P_{Y|X}(y \mid x)}{\sum_x P_{S,X}(s,x)} \tag{2.26}$$

are precomputed, because these terms are common for all elements of $\boldsymbol{C}$. Then each gradient is computed as:

$$\begin{aligned}\frac{\partial}{\partial p(y_i \mid x_j)} J(P_{S,X}, P_{Y|X}) = \sum_s p(s,x_j) \log \frac{p(y_i \mid x)}{p(y_i)} \\ + \sum_s p(s,y) \left(\frac{p(x_j \mid s)}{p(y_i \mid s)} - \frac{p(s,x_j)}{p(y_i)} \right).\end{aligned}$$

Since all marginals and conditionals are precomputed, each gradient can be computed in $O(|\mathcal{S}|)$ time. The space complexity of our method is $O(|\mathcal{V}|)$, because we operate only on active variables $\mathcal{V}$. We point out that the complexity of the algorithm is closely linked to the sparsity of the optimal solution, which itself is related to the value of the distortion constraint Δ. This means that some distortion regimes may not be achievable with a given computational budget. Therefore one has to reduce Δ in order to have a sparser solution [1].

While SPPM is an attractive solution to solve privacy-distortion trade-off problems of large scale, it still falls short at tackling high-dimensional setups because of the exponential growth of the number of parameters, as the dimension grows. The next section is devoted to tackling this issue via an unsupervised Quantization (or clustering) method, to reduce the search space considerably.

2.3.3 Dimensionality Reduction via Quantization

As mentioned before, in real-world datasets, the alphabet $\mathcal{X}$ is often large. In particular, the number of symbols in the alphabet $\mathcal{X}$ observed in the available dataset may be $\theta(n)$,

[1] In several practical experiments however, we did not run into problems of the sort, and were able to generate mappings efficiently even when high distortion was needed to drive the mutual information close to 0.

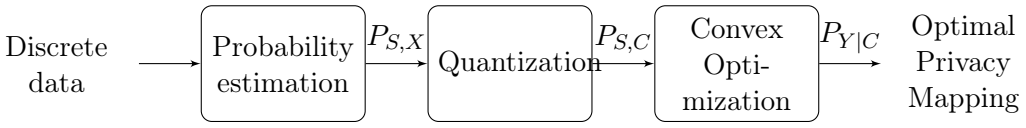

Figure 2-1: The quantization approach for large alphabets

linear in the number of samples n in the dataset. Suppose that $\mathcal{Y} = \mathcal{X}$, for simplicity of notation. Then the number of optimized variables in Problem (2.4) is $\theta(n^2)$. Note that the distortion constraint is linear in $P_{Y|X}(y \mid x)$, but the objective function is neither linear nor quadratic. As a result, Optimization (2.4) cannot be solved using fast linear or quadratic programming solvers. In general, the problem is hard to solve when the size of alphabet $\mathcal{X}$ exceeds a few hundreds symbols.

However, in many problems of interest, data lies on a low-dimensional manifold. For instance, in recommender systems, the ratings of a user can be viewed as a low-dimensional vector in the so-called latent space, whose length is the number of latent factors[86]. In such cases, quantization is guaranteed to reduce the dimensionality of the problem. In particular, let the data lie in a compact d-dimensional latent space where d is small. Then based on a standard sphere packing argument [51], this space can be covered by k representative points such the maximum distance of any point from the closest representative point is $\theta(k^{-1/d})$. In other words, to guarantee that the maximum distance is δ, $\theta((1/\delta)^d)$ representative points are necessary. Note that this quantity is independent of the number of data samples n.

We leverage this observation to propose an approach to reduce the number of optimization variables. Our method comprises three steps. First, a quantization [74] step maps the symbols in alphabet $\mathcal{X}$ to $|\mathcal{C}|$ representative examples in a smaller alphabet $\mathcal{C}$. Second, we learn a privacy-preserving mapping $P_{Y|C}$ on the new alphabet, where $\mathcal{Y} = \mathcal{C}$. Third, the symbols in $\mathcal{X}$ are mapped to the representative examples $\mathcal{Y}$ based on the learned mapping $P_{Y|C}$. This approach is summarized in Diagram 2-1.

This solution has several notable properties. To begin with, the privacy-preserving mapping $P_{Y|C}$ is learned on the reduced alphabet $\mathcal{C}$. Thus, we need to solve the convex optimization (2.4) for only $|\mathcal{C}||\hat{\mathcal{C}}|$ variables instead of $|\mathcal{X}||\mathcal{Y}|$. In practice, $|\mathcal{C}| \ll |\mathcal{B}|$ and this results in major computational savings. Second, quantization and privacy-preserving optimization are done separately. Therefore, any quantization method can be easily combined with our approach. In particular, we can minimize the quantization error in the quantization step, and

then our privacy mechanism guarantees the optimal mapping in terms of additional distortion. It should be noted that the distance used in the quantization phase and the distortion function in the constraint of the privacy mapping optimization need not be the same. In the case where they differ, the end-to-end distortion can be obtained by first computing the value of the distortion function for the representative points resulting from quantization, and then adding this value to the distortion generated by the privacy mapping. Finally, quantization obviously yields a suboptimal privacy-accuracy tradeoff, since the quantization step is an additional source of distortion. However, in Theorem 2, we quantify how quantization affects the privacy-accuracy tradeoff, and show that the levels of privacy that can be achieved are not affected, but come at the expense of a bounded amount of distortion.

In the rest of this section the following variant of problem (2.4):

$$\begin{aligned} \underset{P_{Y|C}}{\text{minimize}} \quad & I(S;Y) && (2.27) \\ \text{subject to:} \quad & \mathbb{E}_{P_{C,Y}} d(C,Y) \leq \Delta \\ & P_{Y|C} \in \text{Simplex} \\ & S \to C \to Y && (2.28) \end{aligned}$$

where the joint distribution over S and C is defined as

$$P_{S,C}(s,c) = \sum_{x \sim c} P_{S,X}(s,x), \tag{2.29}$$

where $x \sim c$ means that the symbol x is in the cluster represented by center c. The above equation aggregates the probability mass of all symbols in the cluster in its center. The symbols in $\mathcal{X}$ are mapped to $\mathcal{Y}$ according to

$$P_{Y|X}(y \mid x) = P_{Y|C}(y \mid \psi(x)), \tag{2.30}$$

where $\psi : \mathcal{X} \to \mathcal{C}$ is a function that maps a symbol in $\mathcal{X}$ to a cluster center in $\mathcal{C}$. Finally, we use the notation $J(P_{S,X}, P_{Y|X}) \triangleq I(X;Y)$, which explicitly show how the privacy leakage depends on the prior $P_{S,X}$, and on the privacy-mapping $P_{Y|X}$. We now prove our main claim.

Theorem 2. *Let $Q_{Y|C}$ be a solution to problem (2.27) and $P_{Y|X}$ be the corresponding*

mapping from $\mathcal{X}$ (Equation 2.30). Moreover, let $\mathcal{C}$ be an alphabet such that $\max_{x\in\mathcal{X}} \min_{c\in\mathcal{C}} d(x,c) \leq r$. Then the privacy leakage $J(P_{S,X}, P_{Y|X})$ of the mapping $P_{Y|X}$ is equal to the value of the objective function of (2.27):

$$J(P_{S,X}, P_{Y|X}) = J(Q_{S,C}, Q_{Y|C}),$$

and its total distortion rate is no more than r larger than the target Δ:

$$\mathbb{E}_{P_{X,Y}}[d(X,Y)] \leq \Delta + r.$$

Proof. The information-leakage equality can be proved as follows. First, both $J(P_{S,X}, Q_{Y|X})$ and $J(Q_{A,C}, Q_{Y|C})$ can be rewritten as

$$J(P_{S,X}, Q_{Y|X}) = H(P_A) + H(P_Y) - H(P_{A,Y}) \tag{2.31}$$

$$J(Q_{A,C}, Q_{Y|C}) = H(Q_S) + H(Q_Y) - H(Q_{A,Y}), \tag{2.32}$$

where

$$P(s,y) = \sum_b Q(y|\psi(x))P(s,x) \tag{2.33}$$

$$Q(s,y) = \sum_c Q(y|c)Q(s,c). \tag{2.34}$$

Second, note that

$$\begin{aligned} P(s,y) &= \sum_b Q(y|\psi(x))P(s,x) \\ &= \sum_c Q(y|c) \sum_{x\sim c} P(s,x) \\ &= \sum_c Q(y|c)Q(s,c) \\ &= Q(s,y). \end{aligned} \tag{2.35}$$

The two distributions are identical, thus $H(P_{A,Y}) = H(Q_{A,Y})$. An analogous result holds for the entropies of the marginals. As a result, the privacy leakage of the mapping $Q_{Y|X}$ on $\mathcal{X}$ is equal to the privacy leakage of the mapping $Q_{Y|C}$ on $\mathcal{C}$.

The distortion inequality is proved as follows. (2.30) implies

$$\begin{aligned} Q_{B,Y}(x,y) &= \sum_a Q_{Y|X}(y|x)P_{S,X}(s,x) \\ &= \sum_a Q_{Y|C}(y|\psi(x))P_{S,X}(s,x). \end{aligned} \tag{2.36}$$

Based on this equality, we can bound the distortion as

$$\begin{aligned} \mathbb{E}Q_{X,Y}d(X,Y) &= \sum_{x,y} Q(x,y)d(x,y) \\ &= \sum_{a,x,y} Q(y|\psi(x))P(s,x)d(x,y) \\ &= \sum_{a,c,y} Q(y|c) \sum_{x\sim c} P(s,x)d(x,y) \\ &\leq \sum_{a,c,y} Q(y|c) \sum_{x\sim c} P(s,x)[d(x,c)+d(c,y)] \\ &= \sum_{a,c,y} Q(y|c) \underbrace{\sum_{x\sim c} P(s,x)}_{Q(s,c)} d(c,y) + \\ &\quad \sum_{a,c} \underbrace{\sum_{y} Q(y|c)}_{1} \sum_{x\sim c} P(s,x)d(x,\psi(x)) \\ &\leq \mathbb{E}Q_{C,Y}d(C,Y) + r\sum_{a,x} P(s,x) \\ &\leq \Delta + r. \end{aligned}$$

■

Theorem 2 states that the information leakage of the mapping $P_{Y|X}$ is the same as that of the optimized mapping $Q_{Y|C}$. So we optimize the quantity of interest $J(P_{S,X}, P_{Y|X})$ in a time which is independent of the size of the input alphabet $\mathcal{X}$. The total distortion increases due to quantization, linearly with the maximum distance r between any example x and its closest representative example $\psi(x)$.

The maximum distance r can be minimized by existing quantization techniques, e.g. online k-center clustering [43] and cover trees [28]. Both methods quantize data nearly optimally. In particular, if the minimum quantization error by $|\mathcal{C}|$ examples is r^*, then the

maximum error produced by these methods is $8r^*$. Note that finding $|\mathcal{C}|$ examples that minimize the quantization error is NP hard.

In the rest of this section, we discuss the case of uncertainty in the knowledge of the distribution $P_{S,X}$, which is an input in all of the methods we have discussed so-far.

2.3.4 Uncertainty in the prior distribution $P_{S,X}$

In practice, we may not have access to the probability of the underlying variable X and S. The distribution $P_{S,X}$ is however a main input in all the algorithms and methods we have discussed so far. We discuss two strategies to handle this issue, namely a min-max formulation, and a stability result.

First, we look at a *worst-case* setup, in which P_X is assumed to be known, and a mapping $P_{Y|X}$ aims to optimize a min-max loss. This setup is very relevant considering the nature of the r.v. S, which is assumed to be sensitive. Therefore, in practical setups, in might be challenging to gather data to make a reliable estimate of the relationship between S and X, as it would require obtaining the sensitive data S. On the other hand, the data X is not deemed sensitive, and is often easier to gather. In any case, the distribution $P_{S,X}$ is assumed to be unavailable. Consequently, finding the exact solution of problem (2.4) may not be possible. This raises the question of the design of privacy-preserving mappings under this partial knowledge on the priors, i.e. suppose P_X is known, but $P_{S|X}$ is unknown. We consider the privacy-preserving mapping which minimizes the worst-case privacy over all possible $P_{S|X}$ while satisfying the utility constraint. That is, the optimal privacy-preserving mapping under this partial knowledge is

$$\begin{aligned} &\min_{P_{Y|X}} \ \max_{P_{S|X}} I(S;Y), \\ &\text{s.t.} \ \ \mathbb{E}_{X,Y}[d(X,Y)] \leq D \ . \end{aligned} \tag{2.37}$$

The proposition which follows shows that this can actually be solved and reduces the problem to a traditional rate-distortion formulation.

Proposition 4. *The problem in* (2.37) *is equivalent to the following rate distortion problem.*

$$\min_{P_{Y|X}:\ |\mathcal{Y}|\leq|\mathcal{X}|+1} I(X;Y),$$

$$s.t. \quad \mathbb{E}_{X,Y}[d(X,Y)] \leq D \tag{2.38}$$

Proof. First note that by letting $S = X$, we obtain $I(X;Y) \leq \max_{P_{S|X}} I(S;Y)$ which then result in

$$\min_{P_{Y|X}} \; I(X;Y) \leq \min_{P_{Y|X}} \; \max_{P_{S|X}} I(S;Y).$$

The other direction follows from the Markov chain property, i.e. $S \to X \to Y$. In particular, for any $P_{S|X}$ we have $I(X;Y) \geq I(S;Y)$ which results in $I(X;Y) \geq \max_{P_{S|X}} I(S;Y)$ for any $P_{Y|X}$. Therefore, we have

$$\min_{P_{Y|X}} \; I(X;Y) \geq \min_{P_{Y|X}} \; \max_{P_{S|X}} I(S;Y).$$

Also, note that the constraint $|\mathcal{Y}| \leq |\mathcal{X}| + 1$ follows from the same argument as in Lemma 1, which completes the proof. ■

Proposition 4 shows that optimization (2.38) can be solved by using any convex solver. Also, note that, once again, the optimization (2.38) can be solved using an Expectation-Minimization (EM) algorithm such as Arimoto-Blahut algorithm [53].

Next, we discuss the case where an estimate $Q_{S,X}$ of the true prior $P_{S,X}$ is made, and used as an input in the optimization. Suppose that we do not have perfect knowledge of the true prior distribution $P_{S,X}$ but that we have its estimate $Q_{S,X}$. Let the $|P_{S,X} - Q_{S,X}|_{\ell_1}$ represent the mismatch between the true prior $P_{S,X}$ and the estimate $Q_{S,X}$. Also denote by $P^*_{Y|X}$ the optimal privacy mapping obtained when $P_{S,X}$ is fed as an input to the optimization (2.4), and let $Q^*_{Y|X}$ denote the solution obtained when feeding the mismatched distribution $Q_{S,X}$ as an input to the optimization (2.4). A useful notation will be to denote the privacy leakage with $P_{S,X}$ and privacy-mapping $P_{Y|X}$ by $J(P_{S,X}, P_{Y|X}) \triangleq I(S;Y)$, where $P_{S,X,Y} = P_{S,X} \cdot P_{Y|X}$. Then, if $Q_{S,X}$ is a *good* estimate of $P_{S,X}$ (low mismatch), then $Q^*_{Y|X}$ should be close to $P^*_{Y|X}$. In particular, we distinguish between two desirable properties:

- **Consistency:** As the true prior is $P_{S,X}$, the *actual* privacy leakage when using privacy mappings $Q^*_{Y|X}$ is given by $J(P_{S,X}, Q^*_{Y|X})$, and not by the quantity $I(Q_{S,X}, Q^*_{Y|X})$ that is optimized when the estimate $Q_{S,X}$ is fed as an input to the optimization. By consistency, we mean that the privacy mappings $Q^*_{Y|X}$ obtained using the estimate

$Q_{S,X}$ should have a good performance, both in terms of actual privacy leakage and distortion, when used under the true prior $P_{S,X}$. Theorem 3 expresses the difference in privacy leakage $|J(P_{S,X}, Q^*_{Y|X}) - J(Q_{S,X}, Q^*_{Y|X})|$ in terms of the mismatch $|P_{S,X} - Q_{S,X}|\ell_1$.

- **Near-Optimality:** For near-optimality, we wish that the performance of the privacy mappings $Q^*_{Y|X}$ be close to that of the optimal mappings $Q^*_{Y|X}$. Theorem 4 expresses the difference in privacy leakage $|J(Q_{S,X}, Q^*_{Y|X}) - J(P_{S,X}, P^*_{X|Y})|$ in terms of the mismatch $|P_{S,X} - Q_{S,X}|\ell_1$.

Theorem 3 (Consistency)**.** *Let $Q^*_{Y|X}$ be a solution to the optimization problem* (2.4) *with $Q_{S,X}$ as input. Then:*

$$\left|J(P_{S,X}, Q^*_{Y|X}) - J(Q_{S,X}, Q^*_{Y|X})\right| \leq 3|P_{S,X} - Q_{S,X}|_{\ell_1} \log \frac{|\mathcal{S}|\,|\mathcal{X}|}{|P_{S,X} - Q_{S,X}|_{\ell_1}}$$
$$\mathbb{E}P_{Y,X} d(X;Y) \leq \Delta + d_{\max}|P_{S,X} - Q_{S,X}|_{\ell_1}$$

where $d_{\max} = \max_{y,x} d(y,x)$ is the maximum distance in the feature space and $\mathbb{E}_{P_{Y,X}}$ is the expectation over $P_{X,Y}(x,y) = \sum_s P_{S,X}(s,x) Q^*_{Y|X}(y|x)$.

Theorem 3 can be interpreted as a consistency result. Indeed, the optimized privacy leakage $J(Q_{S,X}, Q^*_{Y|X})$ and the actual leakage $J(P_{S,X}, Q^*_{Y|X})$ are close if the priors are close. Note, however, that there is no mention of the true optimal leakage $J(P_{S,X}, P^*_{Y|X})$. Theorem 4 bounds this loss.

Theorem 4 (Near-optimality)**.** *Let $Q^*_{Y|X}$ and $P^*_{Y|X}$ be the solutions of the optimization problem* (2.4) *respectively with $Q_{S,X}$ and $P_{S,X}$ as inputs and distortion constraint Δ. Then,*

$$\begin{aligned} &|J(P_{S,X}, P^*_{Y|X}) - J(Q_{S,X}, Q^*_{Y|X})| \\ &\quad \leq 7\|P_{S,X} - Q_{S,X}\|_1 \frac{d_{max}}{d_{min}} \log \frac{|\mathcal{S}||\mathcal{X}|}{\|P_{S,X} - Q_{S,X}\|_1} \end{aligned} \tag{2.39}$$

with d_{max} defined as in Thm. 3, and d_{min} the smallest non-zero value of the distortion, i.e.*, $d_{min} = \min_{x,y,s.t.,d(x,y)>0} d(x,y)$.*

Theorem 3 and Theorem 4 can be combined using the triangle inequality to give a bound on the difference between the actual leakage when having $Q^*_{Y|X}$,*i.e.*, $J(P_{S,X}, Q^*_{Y|X})$ and the

optimal leakage $J(P_{S,X}, P^*_{Y|X})$. The proof of these theorems are inspired by existing results in Information Theory regarding uniform continuity of information theoretic measures such as [153], [21], and methods for the proof of Theorem 4 can be found in [109]. The results can be tighten by using a tighter version of Lemma 17 in the appendix, as in [54, Problem 3.10], but the order of the error stays the same.

This set of results allows us to construct mappings $Q^*_{Y|X}$ that have close to optimal performance, even though the mapping is not perfectly known. The error grows in the order of $O(-|P_{S,X} - Q_{S,X}|_{\ell_1} \log |P_{S,X} - Q_{S,X}|_{\ell_1})$ with the mismatch. Note that only this distance is necessary to compute the bounds, and not the true prior itself. In Prop. 5 below, we provide a bound on the probability of $|P_{S,X} - Q_{S,X}|_{\ell_1}$ being large, when $Q_{S,X}$ is simply the empirical distribution obtained from counting on n samples.

Proposition 5. *Let $Q_{S,X} = \frac{\#\{s_i=s, x_i=x\}}{n}$ be the empirical distribution of $P_{S,X}$, where n is the total number of samples, and $\#\{s_i = s, x_i = x\}$ is the number of examples where $S = s$ and $X = x$. Then,*

$$\mathbb{P}(\|Q_{S,X} - P_{S,X}\|_1 \geq \epsilon) \leq (n+1)^{|\mathcal{S}||\mathcal{X}|} 2^{-2n\epsilon^2}$$

The proof of Prop. 5 follows from Sanov's theorem [53, Thm 12.4.1] and Pinsker's Inequality[54, Problem 3.18].

Therefore, as the sample size n increases, the probability of having a poor empirical estimator of the true distribution in terms of ℓ_1-norm decreases with rate $(n+1)^{|\mathcal{S}||\mathcal{X}|} 2^{-2n\epsilon^2}$. This proposition allows us to formulate corollaries of the following form, here by combining it with Theorem 3:

Corollary 2. *Le $Q_{S,X}$ be the empirical distribution over n samples, and let $0 < \epsilon \leq \frac{1}{2}$. Then,*

$$\left| J(P_{S,X}, p^*_{\hat{B}|B}) - J(Q_{S,X}, P^*_{Y|X}) \right| \leq 3\epsilon \log \frac{|\mathcal{S}||\mathcal{X}|}{\epsilon} \tag{2.40}$$

$$\mathbb{E}P_{X,Y} d(X;Y) \leq \Delta + d_{\max}\epsilon \tag{2.41}$$

with probability $(n+1)^{|\mathcal{A}||\mathcal{B}|} 2^{-2n\epsilon^2}$

This corollary shows the impact on the privacy-accuracy tradeoff of the number of samples available to estimate the distribution and the size of the alphabets.

In the next section, we slightly switch gears and discuss the privacy funnel optimization, which arises when the log-loss is selected as the loss between X and Y.

2.4 Log-loss Distortion and Privacy Funnel

The log-loss distortion is defined as $d(x, y) = -\log P_{X|Y}(x|y)$. Note that this distortion (unlike the one in Definition 1) is a function of x and y as well as $P_{Y|X}$. Using log-loss, the average distortion becomes $\mathbb{E}[d(X, Y)] = \mathbb{E}_{P_{X,Y}}[-\log P_{X|Y}] = H(X|Y)$. Therefore, for a given $D \geq 0$, the distortion bound $H(X|Y) \leq D$ is equivalent to $I(X;Y) \geq t$, where $t = H(X) - D$. It should be noted that the average distortion under the log-loss is not linear in $P_{Y|X}$ (unlike the one in Definition 1).

2.4.1 Privacy-Utility Trade-off under Log-loss

Using log-loss distortion the tradeoff between between utility and privacy becomes minimizing $I(S;Y)$ while $I(X;Y) \geq t$ for some $t \geq 0$. Therefore, the trade-off between utility and privacy in the design of the privacy-preserving mapping is represented by the following optimization, that we refer to as the *Privacy Funnel*:

$$\begin{aligned} &\min\ I(S;Y) \\ &\quad P_{Y|X}:\ I(X;Y) \geq t. \end{aligned} \tag{2.42}$$

For a given utility level t, among all feasible privacy mappings $P_{Y|X}$ satisfying $I(X;Y) \geq t$, the privacy funnel selects the one that minimizes $I(S;Y)$.

Similar to Definition 2 We define next the privacy funnel function, which captures the smallest amount of disclosed private information for a given threshold on the amount of disclosed useful information. We then characterize properties of the privacy funnel function in the rest of this section.

Definition 3. For $0 \leq t \leq H(X)$ and a joint distribution $P_{S,X}$ over $\mathcal{S} \times \mathcal{X}$, we define the *privacy funnel function* $G_I(t, P_{S,X})$ as

$$G_I(t, P_{S,X}) \triangleq \inf \{I(S;Y):\ I(X;Y) \geq t, S \to X \to Y\}, \tag{2.43}$$

where the infimum is over all mappings $P_{Y|X}$ such that $\mathcal{Y}$ is finite. For a fixed $P_{S,X}$ and

$t \geq 0$, the set of pairs $\{(t, G_I(t, P_{S,X}))\}$ is called the *privacy funnel region* of $P_{S,X}$.

Before we proceed to the rest of the discussion, we can prove the counterpart of Lemma 1 for this setting, which gives a bound on the size of the alphabet $\mathcal{Y}$ one needs to consider.

Lemma 3. *We have*

$$G_I(t, P_{S,X}) = \min_{P_{Y|X}} \{I(S;Y) \ : I(X;Y) \leq t, \\ S \to X \to Y, \ |\mathcal{Y}| \leq |\mathcal{X}| + 1\}. \tag{2.44}$$

Proof. In the Proof of Lemma 1, we let $a_i \triangleq H(X) - H(X|Y = i)$ and the rest of the proof is identical to that of Lemma 1. ■

We now prove a few useful properties of $G_I(t, P_{S,X})$ and the privacy region.

Lemma 4. *For $0 \leq t \leq H(X)$, we have*

$$\max\{t - H(X|S), 0\} \leq G_I(t, P_{S,X}) \leq \frac{tI(X;S)}{H(X)}. \tag{2.45}$$

Proof. Observe that $G_I(H(X), P_{S,X}) = I(X;S)$, since $I(X;Y) = H(X)$ implies that $P_{Y|X}$ is a one-to-one mapping of X. The upper bound then follows from Lemma 3 as follows. For $0 < t \leq H(X)$ and $P_{S,X}$ fixed, let $G_I(t, P_{S,X}) = \alpha$. From the discussion above, there exists $P_{Y|X}$ that achieves $I(S;Y) = \alpha$ for $I(X;Y) \geq t$. Now consider $P_{\tilde{Y}|X}$ where $\tilde{\mathcal{Y}} = [|\mathcal{Y}| + 1]$ and, for $0 < \lambda \leq 1$,

$$P_{\tilde{Y}|X}(y|x) = (1 - \lambda)\mathbf{1}_{\{y=|\mathcal{Y}|+1\}} + \lambda \mathbf{1}_{\{y \neq |\mathcal{Y}|+1\}} P_{Y|X}(y|x).$$

Intuitively, $\tilde{Y}$ is an "erased" version of Y, with the erasure symbol being $|\mathcal{Y}| + 1$. It follows directly that $I(S;\tilde{Y}) = \lambda I(S;Y) = \lambda\alpha$, $I(X;\tilde{Y}) = \lambda I(X;Y) \geq \lambda t$, and

$$\frac{G_I(\lambda t, P_{S,X})}{\lambda t} \leq \frac{\lambda I(S;Y)}{\lambda t} = \frac{G_I(t, P_{S,X})}{t}.$$

Since this holds for any $0 < \lambda \leq 1$, then $\frac{G_I(t,P_{S,X})}{t}$ is non-decreasing in t. Finally, for a fixed $P_{S,X}$, the set of points $(\mathbf{w}_i, a_i, b_i) \in \mathbb{R}^{|\mathcal{X}|+2}$ that satisfies (2.7) is convex, and thus, for a fixed P_X, it's lower-boundary, which corresponds to the graph of $(t, G_I(t, P_{S,X}))$, is convex.

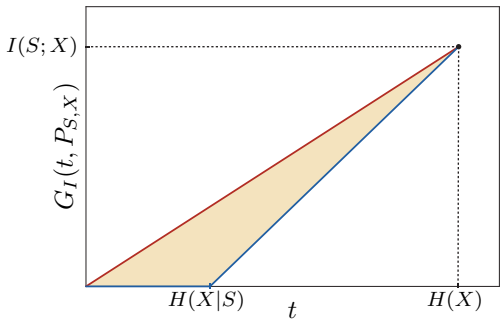

Figure 2-2: For a fixed $P_{S,X}$, the privacy region is contained within the shaded area. The red and the blue lines correspond, respectively, to the upper and lower bounds presented in Lemma 4.

Clearly $G_I(t, P_{S,X}) \geq 0$. In addition, for any $P_{Y|X}$,

$$\begin{aligned} I(S;Y) &= I(X;Y) - I(X;Y|S) \\ &\geq I(X;Y) - H(X|S) \\ &\geq t - H(X|S), \end{aligned}$$

proving the lower bound. ■

Figure 2-2 illustrates the bounds from Lemma 4. The privacy region is contained withing the shaded area. The next two examples illustrate that both the upper bound (red line) and the lower bound (blue line) of the privacy region can be achieved for particular instances of $P_{S,X}$.

Example 3.

- Let $X = (S, W)$, where $W \perp\!\!\!\perp S$. Then by setting $Y = W$, we have $I(S;Y) = 0$ and $I(X;Y) = H(W) = H(X|S)$. Consequently, from Lemmas 3 and 4, $G_I(t, P_{S,X}) = 0$ for $t \in [0, H(X|S)]$. By letting $Y = W$ with probability λ and $Y = (S, W)$ with probability $1 - \lambda$ for $\lambda \in [0, 1]$, the lower-bound $G_I(t, P_{S,X}) = t - H(X|S)$ can be achieved for $H(X|S) = H(W) \leq t \leq H(X)$. Consequently, the lower bound in (2.45) is sharp.

- Now let $X = f(S)$. Then $I(X;S) = H(X)$ and

$$I(S;Y) = I(X;Y) - I(X;Y|S) = I(X;Y).$$

Consequently, $G_I(t, P_{S,X}) = t$, and the upper bound in (2.45) is sharp.

2.4.2 Connections to the Information Bottleneck Method

The information bottleneck method, introduced in [139], considers the setting where a variable X is to be compressed, while maintaining the information it bears about another correlated variable S. The information bottleneck method is a technique generalizing rate-distortion, as it seeks to optimize the tradeoff between the compression length of X and the accuracy of the information preserved about S in the compressed output Y. The information bottleneck optimization [139] is

$$\begin{aligned} &\min\ I(X;Y) \\ &\quad P_{Y|X}:\ I(S;Y) \geq C \end{aligned} \tag{2.46}$$

for some constant C. In the information bottleneck, the compression mapping $P_{Y|X}$ is designed to make X and Y as far as possible from each other (minimizes $I(X;Y)$) while guaranteeing that S and Y are close to each other. In other words, in the information botteleneck the mapping $P_{Y|S}$ is designed to make $I(S;Y)$ large and $I(X;Y)$ small. The information bottleneck optimization (2.46) bears some resemblance to the privacy funnel (2.42), but is actually the opposite optimization. Indeed, in the privacy funnel, the privacy mapping is designed to make $I(S;Y)$ small and $I(X;Y)$ large.

Several techniques were developed to solve the information bottleneck problem such as alternating iteration [139] and agglomerative information bottleneck [131]. This connection is harvested in Section 2.4.4 to design algorithms for the privacy funnel optimization inspired by the existing litterature on the information bottleneck optimization. Before we proceed to the optimization, we provide another connection with a seminal result in Information Theory colloquially referred to as Mrs Gerber's Lemma.

2.4.3 Connections with Mrs Gerber's Lemma

In this section, we explore connections between the privacy funnel optimization, and several fundamental results in Information Theory. In 1973, Wyner and Ziv published a lemma which will be known as Mrs Gerber's Lemma (MGL) [150]. The lemma discusses an extramal property of the entropy for binary sequences of random variables which are distorted by

independent symmetric noise. The result is a simple consequence a scalar form of the lemma, which is the one we refer to by MGL in this section:

Theorem 5 ([150]). *Let Y be any r.v. such that $X|Y = y \sim \text{Ber}(p_y)$, and let $Z \sim \text{Ber}(\alpha)$ be independent of (X, Y), then:*

$$H(X \oplus Z|Y) \geq h(\alpha \star h^{-1}(H(X|Y))), \tag{2.47}$$

where $h(\cdot)$ is the binary entropy function, $a \star x \triangleq a(1-x)+x(1-a)$, and $h^{-1}(\cdot)$ is the inverse binary function restricted to $[0, 1/2]$. Furthermore, the equality is achieved when $P_{Y|X}$ is a binary symmetric channel.

In what follows, we will show that this result essentially provides a closed-form solution for the Information Bottleneck optimization, when (S, X) is a binary symmetric source, i.e., $X \sim \text{Ber}(p)$ and $S = X \oplus Z$, where $Z \sim \text{Ber}(\alpha)$ is independent of X. Indeed, the result in Theorem 5 states that $H(S|Y) \geq h(\alpha \star h^{-1}(H(X|Y)))$. We thus get:

$$I(S;Y) \leq H(S) - h(\alpha \star h^{-1}(H(X|Y))). \tag{2.48}$$

Noting that $H(X|Y) = H(X) - I(X;Y)$, the Information Bottleneck optimization can be equivalently written as:

$$\begin{aligned} &\max_{P_{Y|X}} I(S;Y) \\ &\text{such that} \quad H(X|Y) \geq H(X) - C \triangleq \tilde{C} \end{aligned} \tag{2.49}$$

Finally, noting that the inequality $H(X|Y) \geq \tilde{C}$ must be tight when $C \leq H(X)$, and using the inequality (2.48), the information bottleneck is thus solved by $P_{Y|X} = \text{BSC}(h^{-1}(\tilde{C}))$, with value $H(S) - h(\alpha \star h^{-1}(\tilde{C}))$.

The privacy funnel on the other hand, does not follow directly from MGL, but rather from a *dual* of the MGL. This dual result is referred to as Mr Gerber's Lemma in [83] is proved using tools introduced in [149]. While the proof techniques and arguments are interesting, a complete derivation of these results would detract from the main object of this thesis, and we refer the interested reader to [83]. We simply state Mr Gerber's Lemma, and end this section by summarizing the results on the Information Bottleneck and Privacy

Funnel for binary symmetric sources (S, X) in a theorem.

Lemma 5 (Mr Gerber's Lemma [83])**.** *Let Y be any r.v. such that $X|Y = y \sim \mathrm{Ber}(p_y)$, and let $Z \sim \mathrm{Ber}(\alpha)$ be independent of (X, Y), then:*

$$H(X \oplus Z|Y) \leq \lambda h\left(\alpha \star \frac{q}{z}\right) + (1-\lambda)h(\alpha) \tag{2.50}$$

where $q = P_X(0) \leq 1/2$, $z = \max(\alpha, 2q)$, and $\alpha \in [0,1]$ satisfies $H(X|Y) = \alpha h(q/z)$.

Using this lemma, and essentially the same steps as in the Information Bottleneck case, we may prove the following theorem.

Theorem 6. *Let (S, X) be a binary symmetric source with $P_X(0) = q \leq 1/2$, and $P_{S|X} = \mathrm{BSC}(\alpha)$. Then, the Information Bottleneck and Privacy Funnel trade-offs are characterized implicitly by the pairs:*

$$\begin{aligned} \textit{Information Bottleneck:} \quad & I(X;Y) = h(q) - x \\ & I(S;Y) = h(q \star \alpha) - L(q,x) \\ \textit{Privacy Funnel:} \quad & I(X;Y) = h(q) - x \\ & I(S;Y) = h(q \star \alpha) - U(q,x), \end{aligned} \tag{2.51}$$

where $L(q,x) = h(\alpha \star h^{-1}(x))$ and $U(q,x) = h\left(\alpha \star \frac{q}{z}\right) + (1-\lambda)h(\alpha)$, with λ and z as defined in Lemma 5.

It should be noted that the techniques introduced in [149], and in [83] are not restricted to binary symmetric sources. However, the technique fail to be tractable for larger alphabets, as they involve finding the convex envelope to a polytope whose numbers of vertices grows exponentially with the size of the alphabet. In the next section, we take an algorithmic approach and look at approximate solutions to these problems instead.

2.4.4 Algorithm for Privacy Funnel

The alternating iteration algorithm [139] finds a stationary point of the Lagrangian of information bottleneck optimization (2.46) defined as $\mathcal{L} = I(X;Y) - \beta I(S;Y)$ for some β. The stationary point can be a local minimum, which addresses the information bottleneck, or a local maximum in which case it addresses the privacy funnel. However, there is no

guarantee on the convergence of this alternating algorithm to either a local minimum or a local maximum.

$$\min I(S;Y)$$
$$P_{Y|X}: \ I(X;Y) \geq t.$$

For a given utility level t, among all feasible privacy mappings $P_{Y|X}$ satisfying $I(X;Y) \geq t$, the privacy funnel selects the one that minimizes $I(S;Y)$.

Note that $I(X;Y)$ is convex in $P_{Y|X}$ and since $P_{Y|S}$ is linear in $P_{Y|X}$ and $I(S;Y)$ is convex in $P_{Y|S}$, the objective function $I(S;Y)$ is convex in $P_{Y|X}$. However, because of the constraint $I(X;Y) \geq t$, the Privacy Funnel (2.42) is not a convex optimization [36, Chap. 4]. As mentioned previously, the Privacy Funnel (2.42) is not a convex optimization. In this section, we provide a greedy algorithm and an alternating iteration algorithm to solve optimization (2.42), and we evaluate them on simulated data.

We use a greedy algorithm to find a privacy mapping as described next. Assume $I(X;Y) \geq t$ is given and we are looking for $P_{Y|X}$ that minimizes $I(S;Y)$. Note that for $\mathcal{Y} = \mathcal{X}$ and $P_{Y|X}(y|x) = \mathbf{1}\{x = y\}$ (where $\mathbf{1}\{x = y\} = 1$ if and only if $x = y$), the condition $I(X;Y) \geq t$ is satisfied, but, $I(S;Y)$ might be too large. The idea is to merge two elements of $\mathcal{Y}$ to make $I(S;Y)$ smaller, while satisfying $I(X;Y) \geq t$. This method is motivated by agglomerative information method introduced in [131]. We merge y_i and y_j and denote the merged element by y_{ij}. We then update $P_{Y|X}$ as $p(y_{ij}|x) = p(y_i|x) + p(y_j|x)$, for all $x \in \mathcal{X}$. After merging, we also have $p(y_{ij}) = p(y_i) + p(y_j)$. Let $Y^{(i,j)}$ be the resulting Y from merging i and j. Algorithm 1 is a greedy algorithm to solve optimization (2.42). Proposition 6 shows that, there is an efficient way to calculate $I(S;Y) - I(S;Y^{(i,j)})$ and $I(X;Y) - I(X;Y^{(i,j)})$ at each iteration of algorithm 1.

Proposition 6. *For a given joint distribution $P_{S,X,Y} = P_{S,X}P_{Y|X}$, we have $I(S;Y) - I(S;Y^{(i,j)}) =$*

$$p(y_{ij})H\left(\frac{p(y_i)P_{S|Y=y_j} + p(y_j)P_{S|Y=y_j}}{p(y_{ij})}\right) - \Big(p(y_i)H(P_{S|Y=y_i}) + p(y_j)H(P_{S|Y=y_j})\Big).$$

We also have $I(X;Y) - I(X;Y^{(i,j)}) =$

Algorithm 3 Greedy algorithm-privacy funnel

Input: t, $P_{S,X}$
Initialization: $\mathcal{Y} = \mathcal{X}$, $P_{Y|X}(y|x) = \mathbf{1}\{y = x\}$.
while there exists $i', j' \in \mathcal{Y}$ such that $I(X;Y^{(i',j')}) \geq t$ **do**
  among those i', j', let
  $\{y_i, y_j\} = \text{argmax}_{y_{i'}, y_{j'} \in \mathcal{Y}} I(S;Y) - I(S;Y^{(i',j')})$
  merge: $\{y_i, y_j\} \to y_{ij}$
  update: $\mathcal{Y} = \{\mathcal{Y} \setminus \{y_i, y_j\}\} \cup \{y_{ij}\}$ and $P_{Y|X}$
end while
Output: $P_{Y|X}$

Algorithm 4 Greedy algorithm-information bottleneck

Input: Δ, $P_{S,X}$
Initialization: $\mathcal{Y} = \mathcal{X}$, $P_{Y|X}(y|x) = \mathbf{1}\{y = x\}$
while there exists $i', j' \in \mathcal{Y}$ such that $I(S;Y^{(i',j')}) \geq \Delta$ **do**
  among those i', j', let
  $\{y_i, y_j\} = \text{argmax}_{y_{i'}, y_{j'} \in \mathcal{Y}} I(X;Y) - I(X;Y^{(i',j')})$
  merge: $\{y_i, y_j\} \to y_{ij}$
  update: $\mathcal{Y} = \{\mathcal{Y} \setminus \{y_i, y_j\}\} \cup \{y_{ij}\}$ and $P_{Y|X}$
end while
Output: $P_{Y|X}$

$$p(y_{ij}) H\left(\frac{p(y_i)P_{X|Y=y_j} + p(y_j)P_{X|Y=y_j}}{p(y_{ij})}\right) - \Big(p(y_i)H(P_{X|Y=y_i}) + p(y_j)H(P_{X|Y=y_j})\Big).$$

Proof. After merging y_i and y_j, we have

$$p(s|y_{ij}) = \frac{p(y_i)}{p(y_{ij})}p(s|y_i) + \frac{p(y_j)}{p(y_{ij})}p(s|y_j), \text{ for all } s \in \mathcal{S},$$
$$p(x|y_{ij}) = \frac{p(y_i)}{p(y_{ij})}p(x|y_i) + \frac{p(y_j)}{p(y_{ij})}p(x|y_j), \text{ for all } x \in \mathcal{X}.$$

The proof follows from writing $I(S;Y) - I(S;Y^{(i,j)}) = H(S|Y^{(i,j)}) - H(S|Y)$ and $I(X;Y) - I(X;Y^{(i,j)}) = H(X|Y^{(i,j)}) - H(X|Y)$. ∎

Note that the greedy algorithm is locally optimal at every step since we minimize $I(S;Y)$. However, there is no guarantee that such a greedy algorithm induces a global optimal privacy mapping.

Remark 2. The minimum of $I(S;Y)$ in (2.42) is a decreasing function of $I(X;Y)$ and is achieved for a mapping $P_{Y|X}$ that satisfies $I(X;Y) = t$ (if possible due to discrete alphabets).

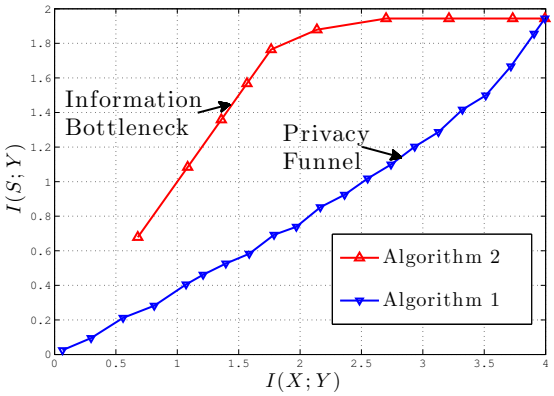

Figure 2-3: Maximum and minimum of $I(S;Y)$ for a given $I(X;Y)$: using greedy algorithms.

For a given mutual information, t, there are many conditional probability distributions, $P_{Y|X}$, achieving $I(X;Y) = t$. Among which there is one that gives the minimum $I(S;Y)$ and one that gives the maximum $I(S;Y)$. We can modify the greedy algorithm so that it converges to a local maximum of $I(S;Y)$ for a given $I(X;Y) = t$. The algorithm which we call *greedy algorithm-information bottleneck* is given in Algorithm (2). Algorithm (1) and Algorithm (2) allow us to approximately characterize the range of values $I(S;Y)$ can take for a given value of $I(X;Y)$ as being those between the local minimum and the local maximum. Interestingly, by observing the gap between the local maximum and the local minimum, we have a relative idea on the effectiveness of the Greedy algorithm, i.e., if the difference is significant it means a negligent mapping may lie anywhere between those values, possibly leading to a much higher privacy threat.

Example 4 (Numerical Example).

Data Set: The US 1994 Census dataset [20] is a well-known dataset in the machine learning community, which is a sample of the US population from 1994. For each of the entries, it contains features such age, work-class, education, gender, and native country, as well as an income category. The income level is a binary variable which determines whether the income is above or below USD 50000, gender is a binary variable, education level is a variable with four categories, age is a variable divided into seven categories. For our purposes, we consider the private attributes $S =$ (age, income level) and the attributes to be released as $X =$ (age, gender, education level). The goal of the privacy mapping is to release a modified version of attributes Y which is informative about X but that renders the inference of S based on Y hard.

Numerical illustration In Fig. 2-3, we plot the minimum and maximum of $I(S;Y)$ for a given $I(X;Y)$. This figure is based on US 1994 census data set described before. The top curve shows the maximum of $I(S;Y)$ versus $I(X;Y)$, using Algorithm (2). The bottom curve shows the minimum of $I(S;Y)$ versus $I(X;Y)$, using Algorithm (1). The area between the two curves shows the possible pairs of $(I(X;Y), I(S;Y))$ as $P_{Y|X}$ varies (s subset of possible pairs, since the algorithms are sub-optimal). Indeed, we will design the mapping to lie on the bottom curve. For a given t, if we design the mapping negligently, we may have $I(S;Y)$ on the top curve instead of the bottom curve.

Chapter 3

Guessing passwords

In the early 1960s, researchers at the Massachusetts Institute of Technology (MIT) were using a rudimentary form of time-sharing computer system which allowed multiple users access to computational resources. The system was known as the Compatible Time-Sharing System (CTSS). They were faced with one problematic feature of the CTSS – users could interrupt each-other's activity on the machine, as well as access all files. It was to handle this shortcoming of the time-sharing system that Professor Fernando Corbató designed the *computer password*. While Prof. Corbató's legacy in computer science is renowned – he received the Turing Award in 1990 for his pioneering contributions – it was hard to envision back then the major role that password secured systems would now occupy in our lives. From purchases, to social networks, password secured systems are ubiquitous, and the backbone of the digital revolution. While passwords have allowed what was only a dream in the 1960s, in recent years the dream has evolved into a "kind of a nightmare" – the words of Prof Corbató himself. This refers to a multitude of unforeseen effects of passwords, and how prevalent they are in our lives.

First and foremost, passwords generated by humans are generally weak. When choosing a secret string of length n in an alphabet of size $|\mathcal{X}|$, the best password is picked by choosing one of the $|\mathcal{X}|^n$ string uniformly at random. However, humans tend to generate strings using a very different distribution, see [33]. Several models for the real-life distribution of passwords have been proposed in the literature, and we refer the reader to the related works section [33, 142, 140] for an in-depth review of such models. Needless to say though, that human generated secret strings are strictly worse than the ideal uniform ones, and thus

adversaries can leverage this knowledge to perform attacks on the secret string.

Next, there are now an unmanageable amount of passwords to keep track of. According to a recent survey [93], the average user has more than 90 password-secured accounts. This leads to *password reuse*. Suppose Alice has several accounts, each requiring a password. From the security standpoint, Alice should generate each password independently from each-other, in which case one password being compromised does not give away any information on any of the other passwords. On the other extreme, Alice may decide to use one identical password for all of her accounts, where the compromise of one of the accounts puts in peril all of her accounts. Most users fall somewhere on this spectrum of password reuse. In either case, this is knowledge an adversary can utilize to its benefit, as one compromised account may actually reveal significant amounts of information and open the door to a person's digital life with catastrophic consequences. Since "a chain is as strong as its weakest link", the security of a password secured system cannot be evaluated in a vacuum.

Finally, there is more readily available public information about individuals online. This data, that is often shared willingly, may seem inauspicious but often users neglect to realize that it reveals a large amount of side-information to adversaries when it comes to guessing passwords. Indeed, it is quite common to generate passwords using some personal information, e.g. date of birth, name of family member or pets, etc. This information is often not deemed sensitive to users, who willingly share it online, for example on social networks. This creates a huge vulnerability against *targeted attacks*. In a targeted attack, the adversary leverages personal information about the user to generate password guesses which are targeted to that specific user. As it becomes easier to collect this personal data, targeted attacks are increasingly powerful, and dangerous.

For these reasons, *brute-force attacks* represent a significant portion of cyber-attacks [5]. They target password-secured systems and consist in querying tentative passwords until the correct one is found. This can occur online, where the adversary connects to a host server, sends her password queries, and receives notification of her success or failure after each guess. More often though, these attacks happen offline. In this case, the adversary has previously gained access to a collection of hashed passwords through another breach, and queries tentative passwords by comparing them to a hash. In either case, these attacks turn out to be surprisingly efficient. From online banking [81] and bitcoin wallets [1], to secure shell (SSH), file transfer protocol(ftp), and telnet servers [108], and passing by governmental

institutions [2], brute-force attacks have shown to be one of the major threats to network security. In this chapter of the thesis, we will present a mathematical model for brute-force attacks, and discuss the security of password secured systems under several types of attack scenarios. In all these cases, we identify the number of queries—or guesses—as a surrogate for the computational effort the adversary has to accomplish to breach the system. As such, understanding quantities such as the average number of guesses before the correct password is found, are useful in assessing the security risks of a system against brute-force attacks. The resulting mathematical formulation is based on a quantity denoted by guesswork, which also has applications in other areas of engineering, information theory, and statistics.

Main contributions and Organization of this chapter:

This chapter is organized as follows. First, we will introduce the mathematical foundation for brute-force attacks, which is based on Guesswork, in 3.1, with a focus on the geometric properties of this quantity. The rest of the chapter is devoted to the evaluation of the asymptotic security of password secured systems under various brute-force attacks. In Section 3.2, we study how targeted attacks and password reuse can be modeled by guesswork with side-information, and also quantify the impact of centralized versus decentralized attacks. In Section 3.3, we explore password guessing under a distributional mismatch, where the password distribution and the adversary's knowledge of the latter are not identical. In Section 3.4, we investigate attacks performed by asynchronous botnets by deriving a series of results in guesswork without memory. Finally, we refer to Appendix B for additional lemmas on Guesswork, and to Appendix C for an exploration of connections between Guesswork and lossless source coding. Our key novel contributions are as follows:

1. Derive a closed-form solution for the decentralized guessing with independent side-information generated from the same channel $P_{Y|X}$.

2. We prove a novel Large Deviations Principle result for Mismatched Guesswork, a guessing setup where the guessing list is made according to the wrong distribution. We also characterize the rate function in terms of information theoretic quantities, which can be interpreted via tools from Information Geometry.

3. Characterize the asymptotic performance, both in terms of averge number of guesses, and in terms of probability of success for a fixed number of guesses, of asynchronous

brute-force attackers by establishing a connection with a guessing problem with no-memory.

Related Work: The problem of a cipher with a guessing wiretapper was considered in [101]. The problem of guessing subject to distortion and constrained Shannon entropy were investigated in [15] and [26], respectively. The above results have been generalized to ergodic Markov chains [135] and a wide range of stationary sources [110]. The problem of guessing under source uncertainty was investigated in [136]. The analysis of the guessing exponents, using large deviations theory, was considered in [79]. In [48] it was shown that the guesswork satisfies a large deviation property and the rate function was characterized. They also provided an approximation to the distribution of guesswork using the large deviation property. Guessing a sequence given an erased version of the sequence was studied in [49], where the interplay between the large-deviations of the erasure process, and of the sequence generation were characterized. A brute-force attack where adversaries are interested in multiple passwords is discussed in [50]. A distributed attack model based on password hints was proposed in [39] and evaluated under guesswork metrics, and a wiretap system under guessing guarantees was studied in [101]. A geometric characterization of the guesswork was established in [24] and expanded in [25]. Finally, applications of guesswork [38] to cryptographic guessing was studied in [38], where oblivious or memoryless guessers were studied. The results of [38] are non-asymptotic, but very much related to our setting, as optimal i.i.d. guessing strategies both in terms of number of guesses and in terms of probability of success are studied, and a distributed attack scenario is also envisaged.

The statistics of password generation were studied in [142, 140, 29, 33]. Password frequencies have been shown to follow closely variants of the Zipf's law distribution. In particular, the so-called *CDF-Zipf's law* model introduced in [142, 140] is a modification of the Zipf's law which captures the frequencies of passwords, both for very frequent passwords, and the tails, as exhibited by the close empirical fit to multiple password datasets (see [140, 142, 29]). Note than an adversary can benefit greatly from the the non-uniformity of these distributions to design more powerful brute-force attacks. Indeed, Guesswork, and other related notions of security related to brute-force attacks are also studied in [142, 29, 33]. A special case of brute-force attack is given by *targeted attacks*, in which the adversary uses the personal information of an user in his guessing strategy, see e.g. [143]. Works such as [141, 57] empirically demonstrate the threat of these targeted attacks, as most users chose

their passwords according to some personal information which an adversary might have easy access to (e.g. birthdays, names of family members, locations, or simply password reuse) .

3.1 Guesswork: A mathematical model for brute-force attacks

In this section, and throuhgout this thesis, we have made several modeling assumptions, both on the password generation process, and the brute-force attack itself. In particular, we assume the following.

1. Passwords are assumed to be strings of given length n, which is known. Note that in some applications, the brute-force attack takes place on private key of some fixed size, in which case the length of the secret key is often known.

2. Passwords are assumed to be strings whose characters are generated i.i.d. from a distribution P_X.[1] In some cases, we prove non-asymptotic results, which hold true for an arbitrary alphabet $\mathcal{X}$, and a distribution P.

3. The common goal of the agents is to guess one given password, or string. In practice, there might be multiple accounts which undergo attacks simultaneously.

We believe that some of these assumptions could be relaxed and generalized using techniques from the literature, as discussed in the related works section. In addition, the i.i.d. setting, and the resulting asymptotic results, can be used as guidelines in designing systems even if the real system violates the memoryless assumption. For example, such results can be used to choose the minimal length of a password to secure a system. Despite these assumptions, the insights gained from the model we study shed light on the robustness of brute-force attacks to the various setups we consider.

3.1.1 Moments of Guesswork

The goal of this section is to introduce guesswork as a surrogate for the computation burden that an adversary has to commit before breaching into a password secured system. The guesswork measures the number of queries needed before *guessing* correctly a discrete random variable X with probability mass function (pmf) P. More precisely, let Alice select

[1] We briefly mention generalizations to passwords generated according to an irreducible stationary Markov Chain in several remarks throughout the thesis, e.g. Remark 6 in Section 3.4.1.

a secret password X from an alphabet $\mathcal{X}$, which is assumed finite, i.e. $|\mathcal{X}| < \infty$. Bob is assumed to know P, but has no access to X when designing his guesses. To this end, he employs a guessing strategy, defined as a sequence $\hat{X}_1^\infty = \{\hat{X}_k(P) : k \geq 1\}$, where $\hat{X}_k(P) \in \mathcal{X}$ is independent of the realization X but may depend on the password distribution P. In other words, $\hat{X}_1^\infty$ is the list of guesses the attacker will use, one after the other, when trying to guess X. The corresponding guessing function $G(X, \hat{X}_1^\infty)$, defined as

$$G(X, \hat{X}_1^\infty) = \inf\{k \geq 1 : \hat{X}_k(P) = X\}, \tag{3.1}$$

represents the number of guesses Bob has to perform before correctly guessing X. This allows to define the guesswork, as such:

Definition 4 (Guesswork)**.** Let $X \sim P$, with $X \in \mathcal{X}$ finite, and let $\rho > 0$. Then, the ρ-th moment of guesswork is defined as:

$$\min_{\hat{X}_1^\infty} \mathbb{E}[G(X, \hat{X}_1^\infty)^\rho], \tag{3.2}$$

where the minimization is over all guessing strategies $\hat{X}_1^\infty$.

Note that for this definition to be valid, there must be an optimal guessing strategy which achieves the minimization in (3.2). The following guarantees the existence of such optimal strategy and characterizes it.

Proposition 7 (Optimal Guessing Strategy)**.** *Consider list of symbols in $\mathcal{X}$, with ties broken arbitrarily*[2] *, that is $\{x_1, \ldots, x_{|\mathcal{X}|}\}$, where $P(x_1) \geq P(x_2) \geq \ldots P(x_{|\mathcal{X}|})$. Then, for any $\rho > 0$ and any guessing strategy $\hat{X}_1^\infty$, we have:*

$$\mathbb{E}\left[G(X, \{x_1, \ldots, x_{|\mathcal{X}|}\})^\rho\right] \leq \mathbb{E}[G(X, \hat{X}_1^\infty)^\rho]. \tag{3.3}$$

The optimal guessing function is denoted by $G_P(X)$, and thus corresponds to the position of X in the list of symbols ordered from most likely to least likely.

Remark 3 (Notation). While the formalism of the guessing function $G(X, \hat{X}_1^\infty)$ is necessary to consider randomized guesses, which will be of interest to us in Section 3.4 to come,

[2]For convenience, we let the ties be broken by lexicographical ordering when applicable throughout this thesis.

throughout several parts of this chapter, we will focus on deterministic guesses. Particularly, the optimal guessing function G_P is a bijection from $\mathcal{X} \to [|\mathcal{X}|]$, and it is optimal under stronged notions than the moment condition (3.3). For example, not only does it minimize simultaneously all moments ρ of Guesswork, it is also the optimal strategy for a fixed number of guesses, i.e., for any guessing strategy $\hat{X}_1^\infty$, and any natural number k,

$$\mathbb{P}\left[G(X, \{x_1, \ldots, x_{|\mathcal{X}|}\}) \leq k\right] \geq \mathbb{P}[G(X, \hat{X}_1^\infty) \leq k]. \tag{3.4}$$

Therefore, unless specified otherwise, the guessing strategy will often be implicit, and we will focus on the quantity $G_P(X)$. Finally, when looking at sequences of random variables of length n, we may use exchangeably the notation X^n, or the bold font $\boldsymbol{X}$, when the index n is clear from the context.

Guessing functions, and in particular the moments of guesswork, were studied by Massey [97] where it was shown that the average guesswork $\mathbb{E}[G_P(X)]$ could not be bounded by the entropy $H(P)$. This fact was then revisited by Arikan [14], where the following result is established, here depicted as a lemma:

Lemma 6 ([14][Theorem 1)**.** *For any $\rho \geq 0$, the optimal guessing function $G_P(X)$ satisfies:*

$$(1 + \log|\mathcal{X}|)^{-\rho} \left[\sum_{x \in \mathcal{X}} P(x)^{\frac{1}{1+\rho}}\right]^{1+\rho} \leq \mathbb{E}[G_P(X)] \leq \left[\sum_{x \in \mathcal{X}} P(x)^{\frac{1}{1+\rho}}\right]^{1+\rho}. \tag{3.5}$$

When looking at iid sequences of random variables $X_1, \ldots, X_n \sim_{i.i.d.} P$, the lemma admits a powerful corollary, which characterizes the asymptotics of guesswork in terms of the Rényi entropy.

Corollary 3 (Asymptotics of Guesswork)**.** *Let $X_1, \ldots, X_n \in \mathcal{X}$ be generated iid from P, and let $\rho > 0$. Then:*

$$E_\rho(P) \triangleq \lim_{n \to \infty} \frac{1}{n} \log \mathbb{E}[G_{P^{(n)}}(X^n)] = \rho \cdot H_{1/1+\rho}(P), \tag{3.6}$$

where $H_\alpha(P)$ is the Rényi entropy of order α ($\alpha > 0$, $\alpha \neq 1$), defined as

$$H_\alpha(X) \triangleq \frac{1}{1-\alpha} \log \left[\sum_{x \in \mathcal{X}} P(x)^\alpha\right]. \tag{3.7}$$

Before we proceed to the proof, let us briefly discuss this result. First of all, $E_\rho(P)$ is the asymptotic *exponent* of growth as a function of n, or in other words, as the length n of a password grows, the average ρ-moment of the number of required guesses grows exponentially with an exponent given by $E\rho(P)$. Secondly, $E_\rho(P)$ can be seen as an operational interpretation of the Rényi entropy H_α, for $\alpha > 1$.

Proof. There are several proofs of this result, which we will revisit at times throughout this thesis. The first proof, due to Arikan, follows from Lemma 6, and is described below for completeness. We start by expanding the lower-bound, the upper bound follows from similar steps:

$$\mathbb{E}[G_{P^{(n)}}(X^n)^\rho] \geq (1+\log|\mathcal{X}|^n)^{-\rho}\left[\sum_{x^n\in\mathcal{X}^n} P^{(n)}(x^n)^{\frac{1}{1+\rho}}\right]^{1+\rho} \tag{3.8}$$

$$\implies \frac{1}{n}\log\mathbb{E}[G_{P^{(n)}}(X^n)^\rho] \geq \frac{1}{n}\log\left[\sum_{x^n\in\mathcal{X}^n} P^{(n)}(x^n)^{\frac{1}{1+\rho}}\right]^{1+\rho} + o(1) \tag{3.9}$$

$$= \frac{1}{n}\log\left[\sum_{x^n\in\mathcal{X}^n}\prod_{i=1}^{n} P(x_i)^{\frac{1}{1+\rho}}\right]^{1+\rho} + o(1) \tag{3.10}$$

$$= \frac{1}{n}\log\left[\prod_{i=1}^{n}\sum_{y\in\mathcal{X}} P(y)^{\frac{1}{1+\rho}}\right]^{1+\rho} + o(1) \tag{3.11}$$

$$= \frac{1}{n}\cdot n\cdot\rho H_{1/1+\rho}(P) + o(1). \tag{3.12}$$

Evaluating the upper bound, and letting $n\to\infty$ yields the desired result via the squeeze theorem. ∎

Guesswork can also be studied when side-information is available. This models targeted attacks, where Bob has access to additional information about Alice, which is modeled by a random variable Y. More precisely, we let $Y\in\mathcal{Y}$ be the output of X through a discrete memoryless channel (DMC) with transition probability $P_{Y|X}$. Upon receiving a realization $y\in\mathcal{Y}$, Bob updates his belief on the distribution of X by ordering the candidate strings in decreasing order with respect to the posterior $P_{X|Y}(\cdot|y)$, resulting in the optimal guessing function $G_{P_{X|Y}}(X|Y=y)\triangleq G_{P_{X|Y=y}}(X)$. The ρ-th moment of the *conditional guesswork*

$G_{P_{X|Y}}(X|Y)$ be defined as the average:

$$\mathbb{E}\left[G_P(X|Y)^\rho\right] \triangleq \sum_{y\in\mathcal{Y}} P_Y(y)\mathbb{E}[G_P(X|Y=y)^\rho]. \tag{3.13}$$

Similarly, the asymptotic exponent of the conditional guesswork is defined as

$$E_\rho(P, P_{Y|X}) \triangleq \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}\left[G_{P_{X^n|Y^n}}(X^n|Y^n)\right]. \tag{3.14}$$

Finally, it was shown in [14], using essentially identical tools as the no side-information case, that

$$E_\rho(P, P_{Y|X}) = \rho \cdot H_{\frac{1}{1+\rho}}(X|Y) \tag{3.15}$$

$$= \rho \cdot \sum_{y\in\mathcal{Y}} P_Y(y) H_{\frac{1}{1+\rho}}(X|Y=y). \tag{3.16}$$

3.1.2 Geometry and Large Deviation Principle

In this section, we will introduce a geometric perspective on guesswork, which will be essential, both as a proof technique, and an interpretation of the asymptotics of guesswork. Along the way, we will discuss strengthening of the results on guesswork from moments, to a large deviation principle (LDP). We make the following two assumptions on all the probability distributions that we study in the rest of this section:

Assumption 1. *We say P is unambiguous if it satisfies the following:*

1. $P(x) > 0$ *for all* $x \in \mathcal{X}$.

2. $\operatorname{argmin}_{x\in\mathcal{X}} P(x)$ *and* $\operatorname{argmax}_{x\in\mathcal{X}} P(x)$ *are unique.*

Note that the set of distributions which are not unambiguous forms a set of Lebesgue measure zero in the set of all distributions, which can be seen by the fact that the non-umanbiguous distributions are contained in a finite union of lower dimensional sets. See [25] for the implications of this assumption.

We are ready to define the tilt operator.

Definition 5 (Distribution Tilting)**.** Let $\alpha \in \mathbb{R}$ and Q be unambiguous. We denote by

$T(P, Q, \alpha)$ the mismatched tilted distribution of order α of Q with respect to P, defined as

$$[T(Q, P, \alpha)](x_i) \triangleq \frac{P(x_i) \cdot Q(x_i)^\alpha}{\sum_{x \in \mathcal{X}} P(x) \cdot Q(x)^\alpha}. \tag{3.17}$$

We further define the the mismatched tilted family of Q with respect to P as

$$\mathcal{T}_{Q,P} \triangleq \{T(Q, P, \alpha) : \alpha \in \mathbb{R}\}. \tag{3.18}$$

By taking limits, we define :

$$[T(Q, P, \infty)](x) = \begin{cases} 1 & \text{if } x = \text{argmax}_{x \in \mathcal{X}}\, Q(x), \\ 0 & \text{otherwise} \end{cases}, \tag{3.19}$$

$$[T(Q, P, -\infty)](x) = \begin{cases} 1 & \text{if } x = \text{argmin}_{x \in \mathcal{X}}\, Q(x), \\ 0 & \text{otherwise} \end{cases}, \tag{3.20}$$

$$T(Q, P, 0) = P. \tag{3.21}$$

This definition of mismatched tilt generalizes the tilt defined in [25, Definition 13], and recovers it when P is the uniform distribution. The tilted family $\mathcal{T}_{Q,\mathbf{u}_\mathcal{X}}$, is denoted by $\mathcal{T}_Q$, and $T(Q, \mathbf{u}_\mathcal{X}, \alpha)$ is denoted by $T(Q, \alpha)$. Further, define $\mathcal{T}_Q^+ = \{T(Q, \alpha) : \alpha > 0\}$ as the positive tilted family, and $\mathcal{T}_Q^- = \{T(Q, \alpha) : \alpha < 0\}$ as the negative tilted family. Note that $\mathcal{T}_Q = \mathcal{T}_Q^+ \cup \mathcal{T}_Q^- \cup \mathbf{u}_\mathcal{X}$.

Lemma 7 (closure of the tilted family under tilt operation)**.** *For any $\alpha > 0$, the following holds:*

$$\mathcal{T}_{Q,P} = \mathcal{T}_{T(Q,\alpha),P}. \tag{3.22}$$

Proof. The proof follows from the definition of $T(Q, \alpha)$ and from (3.17). ■

We now define a collection of linear families.

Definition 6 (linear family)**.** We denote by $\mathcal{L}(Q, \alpha)$ the linear family of Q of order α, defined as

$$\mathcal{L}(Q, \alpha) \triangleq \{\gamma \in \Delta_\mathcal{X} : H(\gamma \| Q) = H(T(Q, \alpha) \| Q))\} \tag{3.23}$$

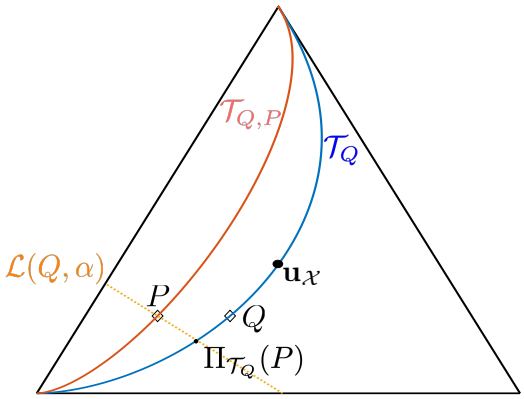

Figure 3-1: Representation of the 3-dimensional simplex, each point in the triangle represents a distribution over $|\mathcal{X}| = 3$. The corners of the triangle correspond to the distribution where all the mass is on a single symbol. The exponential family $\mathcal{T}_Q$ goes through $\mathbf{u}_\mathcal{X}$ and Q. The exponential family $\mathcal{T}_{Q,P}$ goes through P. $\mathcal{L}(Q,\alpha^\star)$ is the linear family of Q of order $\alpha^\star$ which passes through P. The distribution $\Pi_{\mathcal{T}_Q}(P)$ is the projection of P onto $\mathcal{T}_Q$. Of particular interest for lossless coding will be the divergences $D(P\|Q)$ and $D(P\|\Pi_{\mathcal{T}_Q}(P))$.

Intuitively, the mismatched tilted family $\mathcal{T}_{Q,P}$ and the tilted family $\mathcal{T}_Q$, correspond to the curves that are *orthogonal* to the linear families $\mathcal{L}(Q,\alpha)$, and pass through P and $\mathbf{u}_\mathcal{X}$, respectively. We refer the interested reader to [55, Section 3] for an overview of the duality between linear and exponential families, and their applications in statistics, information theory, and large deviations theory.

For a distribution P, we can also define projections on a tilted family $\mathcal{T}_Q$ in the following way:

Definition 7 (projection on a tilted family)**.** We say $\Pi_{\mathcal{T}_Q}(P)$ is the projection of P on $\mathcal{T}_Q$ and define it as

$$\Pi_{\mathcal{T}_Q}(P) \triangleq \arg_{\gamma\in\mathcal{T}_Q}\left\{H(\gamma\|Q) = H(P\|Q)\right\}. \tag{3.24}$$

Note that $\Pi_{\mathcal{T}_Q}(P) = T_Q \cap \mathcal{L}(Q,\alpha^\star) = T(Q,\alpha^\star)$ with $\alpha^\star$ selected such that $H(T(Q,\alpha^\star)\|Q) = H(P\|Q)$.

The following lemma guarantees existence and uniqueness of the projection operator.

Lemma 8. *Let P and Q be umambiguous, then $\Pi_{\mathcal{T}_Q}(P)$ exists and is unique. Further, $\Pi_{\mathcal{T}_Q}(P) = P$ iff $P \in \mathcal{T}_Q$.*

Proof. Note that for an unambiguous Q, $H(T(Q,\beta)\|Q)$ is a strictly decreasing continuous function in β [25]. Further, $H(T(Q,\infty)\|Q) < H(P\|Q) < H(T(Q,-\infty)\|Q)$, thus the

projection must exist and is unique, by the intermediate value theorem. The second part of the claim follows by definition of $\Pi_{\mathcal{T}_Q}(P)$. ■

The definitions above are summarized in Figure 3-1. These geometric quantities satisfy various useful properties, which will be of use in the rest of this paper. We will review some of those in the rest of this section. We start with the I-Projection Pythagorean theorem (see for example [55, Theorem 3.2]).

Lemma 9 (I-Pythagoerean theorem)**.** *Let $\gamma \in \mathcal{T}_Q$, then*

$$D(P\|\gamma) = D(P\|\Pi_{\mathcal{T}_Q}(P)) + D(\Pi_{\mathcal{T}_Q}(P)\|\gamma). \tag{3.25}$$

The next two lemma characterize properties of the projection in terms of entropy and reletive entropy (KL divergence).

Lemma 10 (Projection does not decrease entropy)**.** *Let $\Pi_{\mathcal{T}_Q}(P) \in \mathcal{T}_Q^+$, then*

$$H(\Pi_{\mathcal{T}_Q}(P)) = H(P\|\Pi_{\mathcal{T}_Q}(P)) \geq H(P) \tag{3.26}$$

with equality iff $P \in \mathcal{T}_Q$.

Proof. We first use the identity $H(P) = \log|\mathcal{X}| - D(P\|\mathbf{u}_\mathcal{X})$. By Theorem 9, we have $D(P\|\mathbf{u}_\mathcal{X}) = D(P\|\Pi_{\mathcal{T}_Q}(P)) + D(\Pi_{\mathcal{T}_Q}(P)\|\mathbf{u}_\mathcal{X})$. Thus,

$$\begin{aligned} H(P) &= \log|\mathcal{X}| - D(P\|\Pi_{\mathcal{T}_Q}(P)) - D(\Pi_{\mathcal{T}_Q}(P)\|\mathbf{u}_\mathcal{X}) && (3.27)\\ &\leq \log|\mathcal{X}| - D(\Pi_{\mathcal{T}_Q}(P)\|\mathbf{u}_\mathcal{X}) && (3.28)\\ &= H(\Pi_{\mathcal{T}_Q}(P)) && (3.29) \end{aligned}$$

■

This yields directly the following lemma.

Lemma 11 (Projection does not increase relative entropy)**.** *We have*

$$D(\Pi_{\mathcal{T}_Q}(P)\|Q) = D(P\|Q) + H(P) - H(\Pi_{\mathcal{T}_Q}(P)) \leq D(P\|Q) \tag{3.30}$$

with equality iff $P \in \mathcal{T}_Q$.

Proof. By definition of $\Pi_{\mathcal{T}_Q}$, we have $H(\Pi_{\mathcal{T}_Q}(P)\|Q) = H(P\|Q)$, or equivalently that

$$H(\Pi_{\mathcal{T}_Q}(P)) + D(\Pi_{\mathcal{T}_Q}(P)\|Q) = H(P) + D(P\|Q) \tag{3.31}$$

The proof follows from using Lemma 10. ∎

We are now finally ready to revisit some of the results on guesswork. It was shown in [48] that, under some mild conditions, the logarithm of guesswork satisfies a large deviation principle (LDP), and the rate function was further given in terms of information theoretic quantities in [25, Theorem 5].

Theorem 7 (LDP for matched guesswork)**.** *For any unambiguous P, the sequence $\{\frac{1}{n}\log G_P(X^n)\}_{n\in\mathbb{N}^+}$ satisfies a LDP, with rate function $J(t)$ defined implicitly by*

$$J(t) = D(T(P,\alpha(t))\|P), \tag{3.32}$$

where $\alpha(t) = \arg_{\alpha\geq 0}\{H(T(P,\alpha)) = t\}$.

LDP implies many of the results on the average growth rate of the moments $E_\rho(P)$, via Varadhan's lemma [58, Theorem 4.3.1], which is in essence Laplace's method extended to infinite dimensional spaces. Therefore, an alternative proof of Corollary 3 can be obtained through Theorem 7, via Varadhan's lemma:

Corollary 4. *We have,*

$$\frac{1}{\rho}\cdot E_\rho(P) = \max_{\phi\in\mathcal{T}_\mu^+}\left\{H(\phi) - \frac{1}{\rho}D(\phi\|P)\right\}. \tag{3.33}$$

As expected, It is possible to express the solution for this optimization in (3.33) in terms of Rényi entropies. Indeed, remarking that the optimization (3.33) can be equivalently written as an optimization over the tilt parameter, we have that

$$\frac{1}{\rho}\cdot E_\rho(P) = \max_{\alpha\in\mathbb{R}^+}\left\{H(T(P,\alpha)) - \frac{1}{\rho}D(T(P,\alpha)\|P)\right\}, \tag{3.34}$$

which is maximized by $\alpha = 1/(1+\rho)$, and so $E_\rho(P) = \rho\cdot H_{1/1+\rho}(P)$.

This closes this introductory section on guesswork. Next, we will study targeted attacks, where the side-information is distributed.

3.2 Attacks with Distributed Side-Information

In this section, we study a distributed attack scenario, where m adversarial agents receive additional side-information Y about the password X, and perform a so-called targeted attack [143, 57, 141]. As mentioned before, in the presence of side-information, the agents construct an updated list of password strings, this time, ordered with respect to $P_{X|Y}(\cdot|Y)$, that is they update their belief on the password distribution by taking into account the side-information they have observed. In its most general form, the side-information can model complex additional information that the adversary may have acquired on the choice of the password, ranging from background search on the user who chooses the password, to behind the back attacks in which an illegitimate person observes parts of the password. This setting can also indirectly model adversaries and users over multiple accounts, some of which have been compromised. Suppose Alice has several accounts, each requiring a password. She may decide to use one identical password for all of her accounts, where the compromise of one of the accounts puts in peril all of her accounts. On the other extreme, she may decide to use completely independent passwords for each of the accounts, in which case one password being compromised does not give away any information on any of the other passwords. In practice, most users settle for a solution in between these two extremes. For example, Alice may choose to slightly tweak her passwords from one account to another as to avoid the disastrous consequences of one account being compromised, while still maintaining some convenience. In this case, if one password is compromised, an adversary gains some side-information about the rest of the passwords, see, e.g., [141].

We say that agents are coordinated if they know the guessing strategies of each other, and in this context it means that the agents are able to communicate about their knowledge of the side-information on the password. In this section, we will contrast two strategies illustrated in Fig. 3-2. The first is a decentralized approach in which the agents do not communicate at all, representing the case where agents are fully uncoordinated. The second is a centralized approach in which the side-information is pooled and a central authority provides the optimal lists to the agents, representing a coordinated attack. We show that in the case of an uncoordinated attack, having even a finite number of independent sources of side-information reduces the number of queries exponentially. However, coordination is very powerful, as complete knowledge of all the side-information can potentially reduce

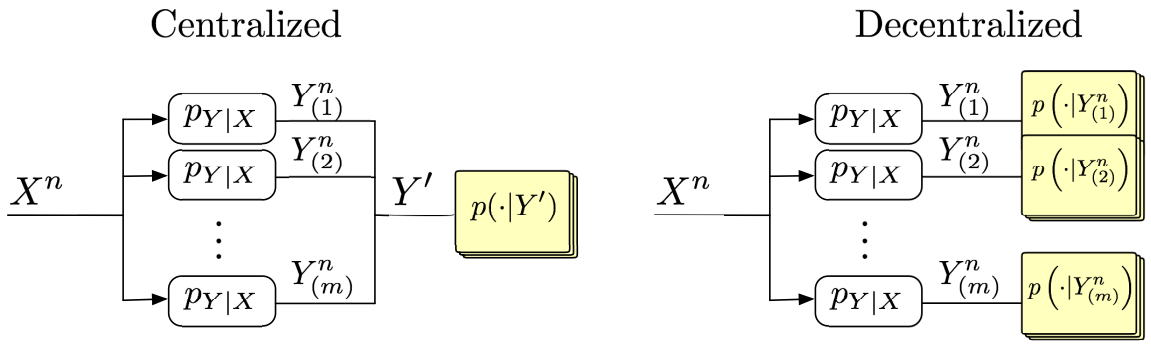

Figure 3-2: In a coordinated attack, a single list is constructed by collecting all the side-information. In the uncoordinated setting, each agent constructs a separate list.

the computational burden on the adversaries by an even bigger exponent. This should be contrasted with the case where side-information is unavailable, as the lack of coordination there does not change the computational burden asymptotically.

We will assume that a finite number m of sources of side information are available. Precisely, for each of the m agents, we consider an independent realization of a side information $Y^n_{(i)}, i = 1, \ldots, m$, where $Y^n_{(i)}$ is the output of the password sequence X^n through a discrete memory-less channel $P_{Y|X}$. It follows that the $Y^n_{(i)}$ are identically distributed and independent given X^n. Recall that coordination refers to the knowledge of the guessing strategies of the other adversaries. Because the optimal guessing strategy of an agent $1 \leq j \leq m$ depends only on the side information $Y^n_{(j)}$, coordination is equivalent to sharing the side information. In other words, if no side information is shared, then the adversaries are uncoordinated, and if all the side-information are pooled and shared among all of the m agents, then the adversaries are perfectly coordinated. We consider two strategies the m adversaries may adopt, reflecting two extremes of coordination c.f. Fig. 3-2.

Centralized: The agents share their observations $Y^n_{(i)}, i = 1, \ldots, m$, with a central authority which collapses the side information and constructs an optimal list based on $P_{X|Y_{(1)},\ldots,Y_{(m)}}$. The ρ-th moment of the guesswork in this strategy is thus

$$\mathbb{E}\left[G(X^n|Y^n_{(1)}, \ldots, Y^n_{(m)})^\rho\right], \tag{3.35}$$

where $P_{Y_1,\ldots,Y_m|X}(y_1, \ldots, y_m|x) = \prod_{i=1}^m P_{Y|X}(y_i|x)$. This corresponds to a completely coor-

dinated attack. Finally, we define,

$$E_\rho^{(c)}(P_{Y|X}, m) \triangleq \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}\left[G(X^n|Y_{(1)}^n, \ldots, Y_{(m)}^n)^\rho\right]. \tag{3.36}$$

Decentralized Mechanism: Each of the m agents tries to guess X^n based on its own observation $Y_{(i)}^n$. The process ends when at least one of the agents correctly guesses X^n. The ρ-th moment of the guesswork for this strategy is thus,

$$\mathbb{E}\left[\min_{i=1,\ldots,m} G(X^n|Y_{(i)}^n)^\rho\right], \tag{3.37}$$

where $G(X^n|Y_{(i)}^n)$ is the optimal guessing function given $Y_{(i)}^n$, that is the position of X^n in the ordered list according to $P_{X^n|Y^n}(\cdot|Y^n = Y_{(i)}^n)$. This corresponds to a completely uncoordinated attack. As before, we define

$$E_\rho^{(d)}(p_{Y|X}, m) \triangleq \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}\left[\min_{i=1,\ldots,m} G(X^n|Y_{(i)}^n)^\rho\right]. \tag{3.38}$$

In the sequel, we shall provide closed-form formulas for (3.36) and (3.38). It has to be noted that we are studying guesswork behaviors for fixed m, that is m may not grow with the block-length n. We may take the limit when $m \to \infty$, but it should be clear that the order of limits is crucial and an interchange of limits is not possible here.

Remark 4. The decentralized strategy we consider above is one in which each agent produce an optimal list regardless of the list produced by the other agents. In particular, it is not clear that this list should be the joint optimal list strategy. More precisely, it is clear that

$$\min_{G_{(j)}, j=1,\ldots,m} \mathbb{E}\left[\min_{j=1,\ldots,m} G_j(X^n|Y_{(i)}^n)^\rho\right] \leq \mathbb{E}\left[\min_{i=1,\ldots,m} G(X^n|Y_{(i)}^n)^\rho\right], \tag{3.39}$$

where the optimization on the left hand size is over all valid guessing functions $G_{(j)}$. While the inequality above holds by definition, it is unclear whether equality should hold. In fact, the right-hand side corresponds to an uncoordinated case as we defined it previously, the left-hand side corresponds to a case in which the agents can coordinate in advance to choose their strategies but no more after the side-information is revealed. We shall address this difference when analyzing the performance of the decentralized scheme under some specific side-information channels for which it is possible to characterize the left-hand side, and shall

$\mathbf{X}$	$\mathbf{Y}_{(1)}$	$\mathbf{Y}_{(2)}$	$\mathbf{Y}_{(3)}$	Pooled SI
password	wasswgrd	phssyotd	password	password
iloveyou	inoieyou	izoveyou	iloviybv	i?oveyou
princess	prinpess	pghjcxys	wrihness	pri??ess
rockyou	rockyeu	rockyou	hozkyxu	rocky?u
nicole	nicoie	nbhole	zocole	n?cole

Figure 3-3: Top 5 lowercase case passwords in the RockYou data. The sister passwords $Y_{(i)}^n$ are generated by changing each letter with probability .3 to any lowercase character. The pooled password Side-Information is obtained by taking the letter that appears in more than 50% of the sister passwords, and putting an erasure ('?') if no such letter exists..

show that they are, at least under these side-information channels, asymptotically identical

.

To illustrate the centralized and decentralized mechanisms, we consider the following toy example, which is based on the RockYou leaked password dataset.

Toy Example: We extract the top 1000 most likely passwords from the *RockYou* dataset (see [56] for a description of the dataset), and limit the scope to passwords with only lowercase letters for convenience. For each such password, we also generate $m = 3$ *sister passwords* synthetically by randomly changing letters, where each letter is changed to any other lowercase letter with a probability of 50%. Those sister passwords model the effect of password reuse, and corresponds to the side-information $Y_{(i)}^n$, that agent $i = 1, \ldots, m$ has access to. We refer to [143] for an empirical study of the statistics of password reuse, which indicate that many users have a sister password with a small Levenshtein distance. Examples of passwords along with the synthetic sister passwords are shown in Figure 3-3.

For the sake of exposition, we assume that all letters are equally likely, which is a sub-optimal but illustrative assumption for the purpose of this toy example. Under this assumption, the optimal strategy of an adversary with side information is to modify the sister password one letter at a time, until the correct password is found. Note that, by making use of prior information such as letter frequency, the adversary can improve his guessing strategy drastically – we refer once again to [143] for an implementation of such guessing strategies. When considering the computational effort (in terms of number of guesses), to recover the password, we can look at two separate scenarios:

- A decentralized mechanism, where each agent makes guesses based on its own sister password $Y_{(i)}^n$, and the first one to finish determines the computational cost.

- A centralized approach, where the sister passwords are pooled. In this case, we assume that any letter that is common in at least 50% of sister passwords is also in the correct password. Again, this is a sub-optimal guessing strategy, but serves as an illustration. Example of this pooled side-information are shown in Figure 3-3.

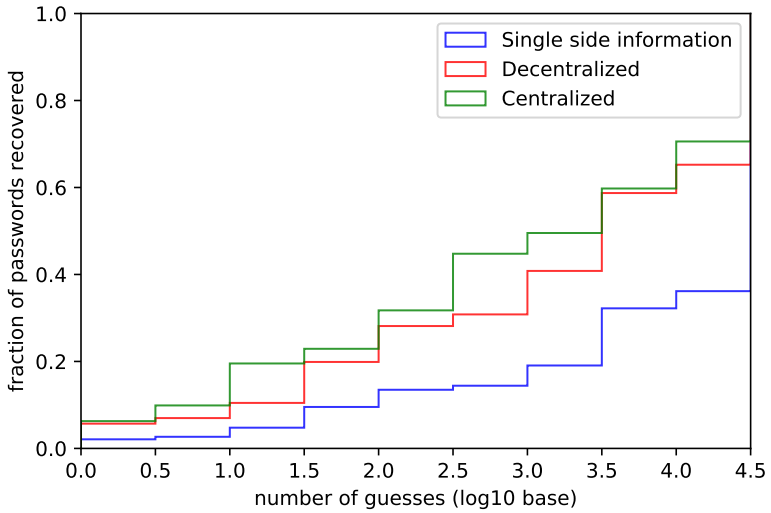

Figure 3-4: With a centralized mechanism, it takes about 300 guesses to recover 50% of the passwords. With a decentralized mechanism, it takes several thousand guesses to reach the same performance. Note that an agent with a single side-information, i.e. with a single sister password, recovers only 40% of the passwords after 30k guesses.

In the centralized approach, the quality of the side information is much better, i.e., many of the letters are already correctly recovered, and the remaining sequence to find are only the *erased* symbols. In the decentralized scenario, the side-information is weaker but there is a benefit in having multiple sources of side-information, as the performance is dominated by the best side-information. The results are showcased in Figure 3-4, and showcase some of the take-aways from the theoretical analysis to follows. Namely, we see that (1) the presence of sister passwords allows for a greatly reduced computational cost (2) a decentralized approach performs better than a single sister password – in fact, we will show that this gain is exponential in the analysis that follows, and (3) the centralized approach allows to essentially improve the quality of the side-information, which proves to be a very potent effect. In the rest of the paper, we will show analytically, that for several sources of side-information, a centralized approach with two agents performs asymptotically better than a decentralized approach with any finite number of agents, suggesting that improving

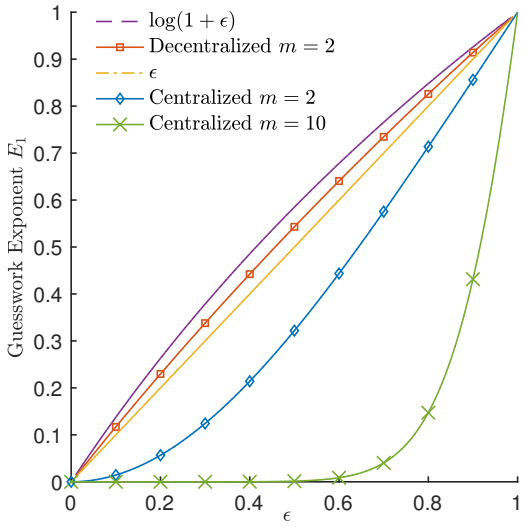

Figure 3-5: BEC(ϵ): Exponents of the average guesswork (i.e. $\rho = 1$) for various m, and under centralized and decentralized strategies. Note that two cooperating agents have a convex exponent, which is better than any number of non-cooperating agents.

the quality of side-information is crucial.

In the rest of this section, we will study the centralized and decentralized mechanisms in detail, and provide closed-form solutions for the the asymptotics of the moments of guesswork under a BSC and BEC side-information channels. The results of Theorems 9,11 and 12 to follow are illustrated in Figures 3-5 and 3-6.

3.2.1 Centralized Mechanism

We illustrate the performance of centralized mechanisms over two side-information channels. First, let X^n be a uniformly distributed sequence of binary digits, i.e., X^n i.i.d. generated from Bern(1/2) [3]. We will contrast two types of side-information channels, namely a binary erasure channel (BEC) with parameter ϵ denoted BEC(ϵ), and a binary symmetric channel (BSC) with parameter δ, denoted BSC(δ).

We start with the BEC channel. Erasures channels have been studied in [49], where the large deviation principle for the guesswork with erasure side-information was characterized. This case is simple to analyze because collapsing information is tractable. In particular, the k-th entry X_k of X^n is erased in all received signals $Y^n_{(i)}$, $i = 1, \ldots, m$, with probability ϵ^m.

[3] Note that the choice of binary inputs is made for the sake of exposition, and those results can be easily generalized to arbitrary discrete sources.

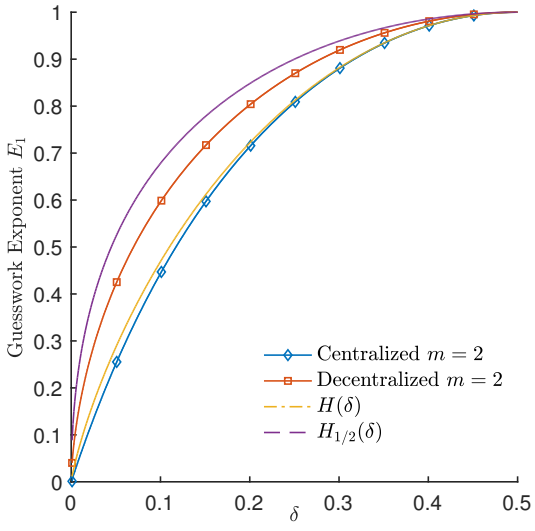

Figure 3-6: BSC(δ): Exponents of the average guesswork (i.e. $\rho = 1$) for $m = 2$ and as $m \to \infty$, and under centralized and decentralized strategies. Again, two cooperating agents have a better exponent than any number of non-cooperating agents.

Therefore, the resulting collapsed random variable $(Y^n_{(1)}, \ldots, Y^n_{(m)})$ is equivalently described by $\tilde{Y}^n$, where $\tilde{Y}^n$ is the output of X^n through a BEC with erasure probability ϵ^m. We have the following result.

Theorem 8 ([49]). *For* BEC(ϵ)*, and* m *agents,*

$$E_\rho^{(c)}(\mathrm{BEC}(\epsilon), m) = \max_{\lambda \in [0,1]} \left[\rho\lambda - D(\lambda \| \epsilon^m)\right]. \tag{3.40}$$

Carrying out the maximization for $\rho = 1$, we get the following immediate result.

Corollary 5. *For* $\rho = 1$*,*

$$E_1^{(c)}(\mathrm{BEC}(\epsilon), m) = \log\left(1 + \epsilon^m\right). \tag{3.41}$$

Remark 5. The function $f(x) = \log(1 + x^m)$ over $x \in [0, 1]$, is convex for any $m \geq 2$. Moreover, as the number of agents increases, the exponent tends towards a flat function $E_\rho^{(c)} = 0$, with a discontinuity at $\epsilon = 1$. Finally, since the first derivative (when $\rho = 1$) is $m\frac{\epsilon^{m-1}}{1+\epsilon^m}$ for any $m \geq 2$, the centralized curve starts flat with a negligible exponent for small ϵ.

For the BSC, the centralized mechanism is more involved to analyze. Indeed, we cannot

describe the channel resulting from collapsing multiple BSC's in terms of a single BSC anymore, since one has m noisy measurements per password-bit. Nevertheless, for $m = 2$, we can characterize precisely this channel by considering the $2^m = 4$ cases. We shall then discuss how to generalize this result to arbitrary $m > 2$.

Theorem 9. *For* BSC(δ)*, and* $m = 2$*,*

$$E_\rho^{(c)}(\mathrm{BSC}(\delta), 2) = \sup_{\lambda \in [0,1]} \left\{ \rho\lambda H_{1/1+\rho}\left(\frac{\delta^2}{1-2\delta(1-\delta)}\right) + \rho(1-\lambda) - D\left(\lambda \| 2\delta(1-\delta)\right) \right\}.$$

Corollary 6. *For* $\rho = 1$*,*

$$E_1^{(c)}(\mathrm{BSC}(\delta), 2) = \log(4\delta(1-\delta) + 1). \tag{3.42}$$

Proof of Theorem 9. Denote by $Y_{(1)}^n$ and $Y_{(2)}^n$ the sequence of side information observed by each agent. For each bit position, there are two cases: either $Y_{(1)}^n$ and $Y_{(2)}^n$ agree and have the same value at that position, or they disagree. Without loss of generality, we assume that all agreements appear consecutively with the disagreements following. In the first part, $Y_{(1)}^n$ and $Y_{(2)}^n$ agree and have the same bit in every position. A simple application of Bayes' rule yields

$$P_{X|Y_1,Y_2}(0|(0,0)) = P_{X|Y_1,Y_2}(1|(1,1)) = \frac{(1-\delta)^2}{\delta^2 + (1-\delta)^2}, \tag{3.43}$$

$$P_{X|Y_1,Y_2}(1|(0,0)) = P_{X|Y_1,Y_2}(0|(1,1)) = \frac{\delta^2}{\delta^2 + (1-\delta)^2}. \tag{3.44}$$

that is on this subsequence, the joint side-information $(Y_{(1)}^n, Y_{(2)}^n)$ can be equivalently represented by a binary vector $\tilde{\mathbf{Y}}$ which is the result of a BSC with parameter $\delta^2/(1-2\delta(1-\delta))$.

In the second part, $Y_{(1)}^n$ and $Y_{(2)}^n$ disagree and have contradicting bits in every position. We then have

$$P_{X|Y_1,Y_2}(0|(0,1)) = P_{X|Y_1,Y_2}(1|(0,1)) = \frac{1}{2}, \tag{3.45}$$

and,

$$P_{X|Y_1,Y_2}(0|(1,0)) = P_{X|Y_1,Y_2}(1|(1,0)) = \frac{1}{2}, \tag{3.46}$$

which is essentially an erasure, since both values of X are equally likely. We let $\lambda \in [0,1]$ be the fraction of bits over which $Y_{(1)}^n$ and $Y_{(2)}^n$ agree, *i.e.*, λn is the size of the first subsequence defined above. Therefore, the central authority has to guess a sequence of the type $\tilde{X}^n = (\tilde{U}^{n(1-\lambda)}, \tilde{Z}^{n\lambda})$, where $\tilde{U}^{n(1-\lambda)}$ is an i.i.d. sequence of uniform Bernoulli random variables that correspond to the erasures, and $Z^{n\lambda}$ is an i.i.d. sequence of Bernoulli random variables with parameter $\tilde{\delta} \triangleq \delta^2/(1-2\delta(1-\delta))$ which corresponds to the bit-flips. By Lemma 20 in the Appendix, we have that

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}[G(\tilde{\mathbf{X}})^\alpha] = \lambda\alpha + (1-\lambda)\alpha H_{1/1+\alpha}(\tilde{\delta}). \tag{3.47}$$

Noting that the probability of the subsequence of agreements of length λn is (up to polynomial factors) $\exp\{-nD(\lambda\|2\delta(1-\delta))\}$, we get the desired optimization. ■

The previous theorem only treats the case of $m = 2$ agents, although a similar technique can be used to tackle any $m \geq 2$ number of agents. Unfortunately, this method is intractable for large m. However, the following result allows us to compute the limit as the number of agents grows to infinity.

Lemma 12. *Assume $\delta \neq \frac{1}{2}$. Then:*

$$\lim_{m\to\infty} E_\rho^{(c)}(\mathrm{BSC}(\delta), m) = 0. \tag{3.48}$$

Proof. Without loss of generality, let $\delta < 1/2$. For a fixed n and m, we do a deterministic pre-processing on the sequences $Y_{(1)}^n, \ldots, Y_{(m)}^n$, which can only increase the guesswork, by definition. We let $\hat{Y}_k$ be defined as the majority bit among the received side information sequences at index k, that is,

$$\hat{Y}_k \triangleq \begin{cases} 0, & \text{if } N_k(0) \geq N_k(1) \\ 1, & \text{if } N_k(0) < N_k(1) \end{cases} \tag{3.49}$$

where $N_k(0) \triangleq \sum_{j=1}^m Y_{(j),k}$, $Y_{(j),k}$ is the k-th bit of the sequence $Y_{(j)}^n$, and $N_k(1) \triangleq n - N_k(0)$.

Then, it is easy to see that the sequence $\hat{Y}^n \triangleq (\hat{Y}_1, \ldots, \hat{Y}_n)$ is the output of X^n through a BSC with parameter δ_m, such that $\delta_m \to 0$ as $m \to \infty$, for any $\delta < 1/2$ [4]. Therefore, for any n and, fixed m, the following equations hold:

$$\mathbb{E}[G(X^n|\tilde{Y^n})^\rho] \leq \mathbb{E}[G(\mathbf{X}|\hat{\mathbf{Y}})^\rho], \tag{3.50}$$

$$\implies E_\rho^{(\mathrm{c})}(\mathrm{BSC}(\delta), m) \leq E_\rho(\mathrm{BSC}(\delta_m)), \tag{3.51}$$

$$\implies \lim_{m\to\infty} E_\rho^{(\mathrm{c})}(\mathrm{BSC}(\delta), m) \leq \lim_{m\to\infty} E_\rho(\mathrm{BSC}(\delta_m)). \tag{3.52}$$

Since the right hand side of the last inequality converges to 0, for any $\delta < \frac{1}{2}$, we obtain the desired result. ■

In other words, when m is large enough, one can *estimate* each bit of the password based on the noisy observations. We now move to the decentralized setup, and contrast some of those results.

3.2.2 Decentralized Mechanism

We now study the number of guesses per adversary under the decentralized approach. Our main result, presented below, gives an asymptotic single letter formula for (3.38).

Theorem 10. *Let X^n be generated i.i.d. from P. Then,*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}\left[\min_{i=1,\ldots,m} G(\mathbf{X}|\mathbf{Y}_{(i)})^\rho\right] =$$

$$\sup_{\alpha\in[0,1]} \sup_{\substack{\hat{P}_{X,Y} \\ \textit{subject to } \hat{P}_{X|Y} \notin \mathcal{Q}(\alpha,\hat{P}_Y)}} \rho\cdot\alpha - D(\hat{P}_X||P_X) - mD(\hat{P}_{Y|X}||P_{Y|X}|\hat{P}_X) \tag{3.53}$$

where $\mathcal{Q}(\alpha, \hat{P}_{\mathbf{y}})$ is defined as

$$\mathcal{Q}(\alpha, \hat{P}_Y) \triangleq \Big\{ Q_{X|Y} : D(Q_{X|Y}\|P_{X|Y}|\hat{P}_Y) + H(Q_{X|Y}|\hat{P}_Y) < D(Q^*_{X|Y}\|P_{X|Y}|\hat{P}_Y) + H(Q^*_{X|Y}|\hat{P}_Y) \Big\},$$

[4] A bound on δ_m can be obtained by an application of Chernoff bound, i.e., $\delta_m < e^{-nD(1/2\|\delta)}$

with $Q^*_{X|Y}$ being the solution of the optimization problem

$$\begin{aligned} \underset{Q_{X|Y}}{\textit{minimize}} \quad & D(Q_{X|Y}\|P_{X|Y}|\hat{P}_Y) + H(Q_{X|Y}|\hat{P}_Y) \\ \textit{subject to} \quad & H(Q_{X|Y}|\hat{P}_Y) \geq \alpha. \end{aligned} \tag{3.54}$$

Proof of Theorem 10. We consider the case of $\rho = 1$. The generalization for any $\rho \geq 0$ is immediate. We start by conditioning on $\mathbf{X}$,

$$\mathbb{E}\left[\min_{j=1,\ldots,m} G^*\left(\mathbf{X}|\mathbf{Y}_{(j)}\right)\right] = \mathbb{E}\left[\mathbb{E}\left[\min_{j=1,\ldots,m} G^*(\mathbf{X}|\mathbf{Y}_{(j)})\middle|\mathbf{X}\right]\right]. \tag{3.55}$$

Since $\min_{j=1,\ldots,m} G^*(\mathbf{X}|\mathbf{Y})$ is non-negative, and recalling that $\mathbb{E}[X] = \sum_{i\geq 0}\mathbb{P}(X \geq i)$ for a non-negative random variable X, we have that the inner expectation on the right hand side evaluates to

$$\sum_{i=1}^{|\mathcal{X}|^n} \mathbb{P}\left\{\min_{j=1,\ldots,m} G^*\left(\mathbf{X}|\mathbf{Y}_{(j)}\right) \geq i\middle|\mathbf{X} = \mathbf{x}\right\}. \tag{3.56}$$

For a fixed i and $\mathbf{x} \in \mathcal{X}^n$, note that $\mathbf{Y}_{(j)}$ are independent given $\mathbf{X}$, and thus $G^*(\mathbf{X}|\mathbf{Y}_{(j)})$ are independent and identically distributed given $\mathbf{X}$. We then have

$$\begin{aligned} &\mathbb{P}\left\{\min_{j=1,\ldots,m} G^*\left(\mathbf{X}|\mathbf{Y}_{(j)}\right) \geq i\middle|\mathbf{X} = \mathbf{x}\right\} && (3.57)\\ &\quad= \prod_{j=1}^{m} \mathbb{P}\left\{G^*\left(\mathbf{X}|\mathbf{Y}_{(j)}\right) \geq i\middle|\mathbf{X} = \mathbf{x}\right\} && (3.58)\\ &\quad= \left[\mathbb{P}\left\{G^*\left(\mathbf{X}|\mathbf{Y}_{(1)}\right) \geq i\middle|\mathbf{X} = \mathbf{x}\right\}\right]^m, && (3.59) \end{aligned}$$

where we have used independence in (3.58). Next, we have,

$$\begin{aligned} &\mathbb{P}\left\{G^*\left(\mathbf{X}|\mathbf{Y}_{(1)}\right) \geq i\middle|\mathbf{X} = \mathbf{x}\right\} && (3.60)\\ &= \sum_{\mathbf{y}:G^*(\mathbf{x}|\mathbf{y})\geq i} P_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) && (3.61)\\ &= \sum_{\mathbf{y}\in\mathcal{L}_i(\mathbf{x})} \exp\left\{-n\left[D\left(\hat{P}_{\mathbf{y}|\mathbf{x}}\|P_{Y|X}\middle|\hat{P}_{\mathbf{x}}\right) + H\left(\hat{P}_{\mathbf{y}|\mathbf{x}}\middle|\hat{P}_{\mathbf{x}}\right)\right]\right\}, && (3.62) \end{aligned}$$

where $\mathcal{L}_i(\mathbf{x})$ corresponds to the set $\mathcal{L}_i(\mathbf{x}) \triangleq \{\mathbf{y} \in \mathcal{Y}^n : G(\mathbf{x}|\mathbf{y}) \geq i\}$, and $\hat{P}_{\mathbf{x}}$ and $\hat{P}_{\mathbf{y}|\mathbf{x}}$ correspond to the empirical distribution (type) of $\mathbf{x}$, and $\mathbf{y}$ given $\mathbf{x}$, respectively (see [54,

Lemma 2.6]). A given sequence $\mathbf{y}$ with conditional type $\hat{P}_{\mathbf{x}|\mathbf{y}}$ induces a *reverse channel* $\hat{P}_{\mathbf{x}|\mathbf{y}} = \frac{\hat{P}_{\mathbf{x}|\mathbf{y}}\hat{P}_{\mathbf{x}}}{\hat{P}_{\mathbf{y}}}$. The condition $\mathbf{y} \in \mathcal{L}_i(\mathbf{x})$ can then be expressed in terms of this reverse channel, as the position of $\mathbf{x}$ in the optimal list constructed according to $P_{X|Y}$ is essentially a function of the types $\hat{P}_{\mathbf{y}}$ and $\hat{P}_{\mathbf{x}|\mathbf{y}}$, and the value of $\alpha \triangleq \log i$, as shown in Lemma 23 in the Appendix. Thus, using the method of types [54, Chapter 2], we may rewrite (3.62) as follows

$$\begin{aligned}
&\mathbb{P}\left\{G^*\left(\mathbf{X}|\mathbf{Y}_{(1)}\right) \geq i \middle| \mathbf{X}=\mathbf{x}\right\} && (3.63)\\
&= \sum_{\hat{P}_{\mathbf{x},\mathbf{y}} \notin \mathcal{Q}\left(\alpha,\hat{P}_Y\right)} \left|T(\hat{P}_{\mathbf{y}|\mathbf{x}})\right| \exp\left\{-n\left[D\left(\hat{P}_{\mathbf{y}|\mathbf{x}} \| P_{Y|X} \middle| \hat{P}_{\mathbf{x}}\right)\right.\right.\\
&\qquad\qquad \left.\left. + H\left(\hat{P}_{\mathbf{y}|\mathbf{x}} \middle| \hat{P}_{\mathbf{x}}\right)\right]\right\} && (3.64)\\
&\doteq \sum_{\substack{\hat{P}_{X,Y} \\ \text{subject to } \hat{P}_{X|Y} \notin \mathcal{Q}\left(\alpha,\hat{P}_Y\right)}} \exp\left\{-n\left(D\left(\hat{P}_{\mathbf{y}|\mathbf{x}} \| P_{Y|X} \middle| \hat{P}_{\mathbf{x}}\right)\right)\right\} && (3.65)\\
&\doteq \sup_{\substack{\hat{P}_{X,Y} \\ \text{subject to } \hat{P}_{X|Y} \notin \mathcal{Q}\left(\alpha,\hat{P}_Y\right)}} \exp\left\{-n\left(D\left(\hat{P}_{\mathbf{y}|\mathbf{x}} \| P_{Y|X} \middle| \hat{P}_{\mathbf{x}}\right)\right)\right\}. && (3.66)
\end{aligned}$$

We are now ready to plug (3.66) into (3.56). Recall that the position of $\mathbf{x}$ is a function of the types $\hat{P}_{\mathbf{x}|\mathbf{y}}$ and $\hat{P}_{\mathbf{y}}$. Let the set $\mathcal{A} = \{\alpha : \alpha = H(\hat{P}_{\mathbf{x}}), \text{for some sequence } \mathbf{x} \in \mathcal{X}^n\}$, be the set of empirical entropy values which can be obtained from the n-length sequences. Note that since there are only a polynomial number, in n, of valid types $\hat{P}_{\mathbf{x}}$, $\mathcal{A}$ is also of polynomial size, and thus, we have

$$\begin{aligned}
&\sum_{i=1}^{|\mathcal{X}|^n} \mathbb{P}\left\{\min_{j=1,\ldots,m} G^*\left(\mathbf{X}|\mathbf{Y}_{(j)}\right) \geq i \middle| \mathbf{X}=\mathbf{x}\right\}\\
&= \sum_{\alpha \in \mathcal{A}} e^{n\alpha} \mathbb{P}\left\{\min_{j=1,\ldots,m} G^*\left(\mathbf{X}|\mathbf{Y}_{(j)}\right) \geq \lceil |\mathcal{X}|^{n\alpha} \rceil \middle| \mathbf{X}=\mathbf{x}\right\}\\
&\doteq \sup_{\alpha \in [0,1]} \sup_{\substack{\hat{P}_{X,Y} \\ \text{subject to } \hat{P}_{X|Y} \notin \mathcal{Q}\left(\alpha,\hat{P}_Y\right)}} \exp\left\{n\left[\alpha - D\left(\hat{P}_{\mathbf{y}|\mathbf{x}} \| P_{Y|X} \middle| \hat{P}_{\mathbf{x}}\right)\right]\right\}. && (3.67)
\end{aligned}$$

Finally, plugging (3.67) into (3.55), and using once again the method of types to get that $\mathbb{P}(\mathbf{X} \in T(\hat{P}_{\mathbf{x}})) \doteq \exp\{-nD(\hat{P}_{\mathbf{x}} \| P_X)\}$, the result is deduced. ■

Using Theorem 10, we have the following corollary.

Corollary 7. *For any $\rho > 0$,*

$$\lim_{m\to\infty}\lim_{n\to\infty}\frac{1}{n}\log\mathbb{E}\left[\min_{i=1,\ldots,m}\left\{G(X^n|Y_{(i)}^n)^\rho\right\}\right] = H(X|Y). \tag{3.68}$$

Proof. Looking at (3.53), we see that as $m \to \infty$, $D(\hat{P}_{Y|X}\|P_{Y|X}|\hat{P}_X)$ must be zero, and thus $\hat{P}_{Y|X}$ must be equal to $P_{Y|X}$ for all x such $\hat{P}_X(x) > 0$. Note that, the maximizing $\hat{P}_X$ is then given by $\hat{P}_X = P_X$, and thus we get $\hat{P}_{X,Y} = P_{X,Y}$. This in turns impose a condition on α, namely that the set $\mathcal{Q}(\alpha, P_Y)$ must not contain $P_{X|Y}$. Precisely, we have

$$P_{X|Y} \notin \mathcal{Q}(\alpha|P_Y) \implies H(P_{X|Y}|P_Y) \geq H(Q^*_{X|Y}|P_Y) \geq \alpha, \tag{3.69}$$

where the second inequality follows from the definition of $Q^*_{X|Y}$. Thus, the maximal α is given by $H(P_{X|Y}|P_Y) = H(X|Y)$. ∎

To illustrate the power of the decentralized approach, we consider again the BEC and BSC side information. Note that it is possible to obtain these results by plugging in Theorem 10. However, for these two channels, it is insightful to take a direct approach. In addition, we address Remark 4, and show that under these two channels, the number of guesses does not change asymptotically even if the adversaries coordinate jointly their lists prior to observing the side-information.

Theorem 11. *For* BEC(ϵ)*,*

$$E_\rho^{(d)}(\mathrm{BEC}(\epsilon), m) = \sup_{\lambda\in[0,1]}\left(\rho\lambda - mD(\lambda||\epsilon)\right). \tag{3.70}$$

Before we proceed to the proof of Theorem 11, some remarks are in order. One can verify that the guesswork exponent for the decentralized mechanism, as the number of agents m increases, converges towards ϵ (see Fig 3-5), as expected from Corollary 7. On the other hand, Remark 1 implies that even two agents that collapse their side information are more powerful than any finite number of agents guessing X^n in a decentralized way, since the centralized scheme has a convex exponent.

Proof of Theorem 11. For simplicity of exposition, we focus on the case where $m = 2$ and $\rho = 1$, while the generalization for any ρ and m is immediate. The proof of Theorem 11 follows from two steps. First, we find an upper bound on the guesswork exponent by

considering the exponent of the shortest sequence [5]. Recall that, since $Y_{(i)}^n$ is just an erased version of X^n for any $i = 1, \ldots, m$. The adversaries must each try to guess a sequence $Z_{(i)}^{\mathcal{E}_{(i)}^n}$, where the length $\mathcal{E}_{(i)}^n$ of $Z_{(i)}^{\mathcal{E}_{(i)}^n}$ is the number of erasures in the sequence $Y_{(i)}^n$, and $Z_{(i)}^{\mathcal{E}_{(i)}^n}$ is a uniformly distributed binary sequence. We then have

$$\mathbb{E}[\min_{i=1,\ldots,m} G(Z_{(i)}^{\mathcal{E}_{(i)}^n})] \leq \mathbb{E}[G(Z_*^n)], \tag{3.71}$$

where Z_*^n is the sequence of any adversary which has $\mathcal{E}_n^* \triangleq \min_{i=1} \mathcal{E}_{(i)}^n$ erasures. Note that the probability of having $\mathcal{E}_n^* = n \cdot \lambda$ for some $\epsilon \leq \lambda \leq 1$, is given (exponentially) by $\exp\left[-n \cdot mD(\lambda \| \epsilon)\right]$. Indeed,

$$\mathbb{P}\left(\frac{1}{n}\mathcal{E}_n^* = \lambda\right) \doteq \mathbb{P}\left(\frac{1}{n}\mathcal{E}_n^* \leq \lambda\right) \tag{3.72}$$

$$= \mathbb{P}\left(\frac{1}{n}\mathcal{E}_i \leq \lambda\right)^m \tag{3.73}$$

$$\doteq \exp\left[-n \cdot mD(\lambda \| \epsilon)\right], \tag{3.74}$$

where the last step follows from Sanov's theorem. Similarly, when $0 \leq \lambda < \epsilon$, the probability $\mathbb{P}\left(\frac{1}{n}\mathcal{E}_n^* = \lambda\right)$ is exponentially equal to $\exp -nD(\lambda \| \epsilon)$. Therefore, letting $f(\lambda, m) = \mathbf{1}\left\{\lambda > \epsilon\right\} mD(\lambda \| \epsilon) + \mathbf{1}\left\{\lambda \leq \epsilon\right\} D(\lambda \| \epsilon)$, we have:

$$\mathbb{E}\left[G(\mathbf{Z}_*)\right] = \mathbb{E}\left[\mathbb{E}\left[G(\mathbf{Z}_*) | \mathcal{E}_* = n\lambda\right]\right] \tag{3.75}$$

$$= \sum_{\lambda = 0, 1/n, \ldots, 1} \mathbb{P}(\mathcal{E}_* = \lambda n) \exp(n\lambda) \tag{3.76}$$

$$\doteq \exp\left[n \sup_{\lambda \in [0,1]} \left(\lambda - f(\lambda, \epsilon)\right)\right]. \tag{3.77}$$

Noting that the maximizing λ is always greater or equal to ϵ, we have the upper-bound

$$\lim_{n \to \infty} \frac{1}{n} \log \mathbb{E}\left[\min_{i=1,\ldots,m} G(\mathbf{Z}_{(i)})\right] \leq \sup_{\lambda \in [0,1]} \left[\lambda - mD(\lambda \| \epsilon)\right]. \tag{3.78}$$

To obtain a matching lower-bound, we consider an oracle that provides additional information to both agents, strictly reducing their guesswork. The additional information from the

[5] Note that this exponent can be derived directly as a consequence of the results in [49]. The proof method in this thesis is included for completeness, and characterizes only the exponent of the guesswork, as opposed to the entire large deviation rate, as done in [49].

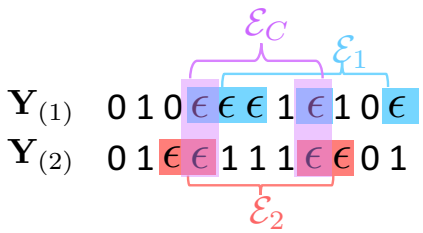

Figure 3-7: The erasures sets that the Oracle Mechanism shares. Note that $G^*(\mathbf{X}|\mathbf{Y}_{(1)})$ and $G^*(\mathbf{X}|\mathbf{Y}_{(2)})$ are not independent because of the bits in $\mathcal{E}_C$. Over this interval, the agents should query sequences which are disjoint, by for example, querying following opposite ends of a lexicographical ordering.

oracle allows to construct explicitly the optimal list of both agents. More precisely, this is achieved by transmitting the position of the common erasures for both agent. The optimal joint strategy is then to construct lists as to minimize queries that have a common subsequence in the overlapping erasures. Indeed, each incorrect query from an agent, shapes the probability distribution of the second agent because of the common sequences. We show that this probability shaping, can be again lower-bounded by a mechanism in which each agent has two guesses at each step, instead of one, therefore not affecting the guesswork exponent. This is formalized below:

Definition 8 (Oracle Mechanism)**.** Let $\mathcal{E}_1$ be the set of erased indices for agent 1, that is $\mathcal{E}_1 = \{i|Y_{(1),i} = \epsilon\}$, and define $\mathcal{E}_2$ similarly for agent 2. Also let $\mathcal{E}_C = \mathcal{E}_1 \cap \mathcal{E}_2$ be the common erasures, and denote by $n_c = |\mathcal{E}_C|$. Further, let $n_1 = |\mathcal{E}_1 \backslash \mathcal{E}_C|$ and $n_2 = |\mathcal{E}_2 \backslash \mathcal{E}_C|$, see Fig. 3-7. Suppose without loss of generality that $n_1 \geq n_2$ .We consider an helping oracle that does the following:

- Transmits to each agent the sets $\mathcal{E}_1$ and $\mathcal{E}_2$.
- Reveals $n_1 - n_2$ bits among those in $\mathcal{E}_1 \backslash \mathcal{E}_C$ to agent 1, making agent 1 as strong as agent 2.

That is, agent i has to guess a binary uniform sequence $(\tilde{X}_{(i)}^{n_1}, \tilde{X}^{n_c})$, where the subsequence $\tilde{X}^{n_c}$ is common for both agents, and the subsequences $\tilde{X}_{(1)}^{n_1}$ and $\tilde{X}_{(2)}^{n_1}$ are independent.

With the knowledge of the Oracle, the two agents will try to construct an optimal joint strategy. At step k, the agent 1 will pick its sequence assuming its previous $k-1$ were incorrect, as well as the $k-1$ sequences of the second agent. Indeed, each of the $k-1$ guesses of the second agent shapes the probability distribution over the sequences for the

first agent due to the common sequence $\tilde{X}^{n_c}$. Therefore, the optimal strategy for the agent 1 is to query a sequence for which the corresponding subsequence $\tilde{x}^{n_c}$ is as likely as possible, or in other words, has been queried the least so far by the other agent. This can be achieved simply by considering a lexicographical ordering over the subsequences x^{n_c} for one agent and an anti-lexicographical ordering for the other agent, as this guarantees that each agent queries sequences that disagree on their subsequence. Next, using the Lemma 21, we show that this process is worse, in terms of guesswork, to a process in which the agent gets one *free* query. Therefore, the guesswork is unchanged asymptotically, and we obtain the desired result. ■

We now study the BSC side-information channel.

Theorem 12. *For* BSC(δ)*,*

$$E_\rho^{(d)}(\mathrm{BSC}(\delta), m) = \rho H_{\frac{m}{\rho+m}}(\delta). \tag{3.79}$$

Proof of Theorem 12. First notice that $\mathbf{Y}_{(i)} = \mathbf{X} \oplus \mathbf{Z}_{(i)}$ where $\mathbf{Z}_{(i)}$ is the sequence of flips, and is generated i.i.d. from Bern(δ), and hence $G(\mathbf{X}|\mathbf{Y}_{(i)}) = G(\mathbf{Z}_{(i)})$. Further, all $\mathbf{Z}_{(i)}$ sequences are independent, and so are the guessworks $G(\mathbf{Z}_{(i)})$. First recall the following elementary result. Let S_i^n, for $i = 1, \ldots, m$, be the sum of n i.i.d. coin flips with parameter δ, and let $S_1^n, \ldots, S_m^n$ be independent. Then, for any $\delta < s \leq 1$:

$$\begin{aligned}
\mathbb{P}\left(\min_i S_i = sn\right) &= m \cdot \mathbb{P}(S_1 = s \cdot n) \cdot \prod_{i=2}^{m} \mathbb{P}(S_i \geq s \cdot n) \\
&\doteq \exp\{-nD(s||\delta)\} \left(\exp\{-nD(s||\delta)\}\right)^{m-1} \\
&\doteq \exp\{-nmD(s||\delta)\}.
\end{aligned}$$

Alternatively, when $0 < s \leq \delta$, we have:

$$\mathbb{P}\left(\min_{i=1,\ldots,m} S_i = sn\right) \doteq \exp\{-nD(s||\delta)\}. \tag{3.80}$$

Using the previous results, and recalling that $G(Z_{(i)}^n) \doteq 2^{S_i^n}$, where S_i^n is the number of 0's

in the sequence (the type of the binary sequence), we obtain that:

$$\mathbb{E}\left[\min_{i=1,\ldots,m} G(\mathbf{Z}_{(i)})^{\rho}\right] \doteq \exp\left\{n \cdot \sup_{\lambda\in[0,1]} \left(\rho\lambda - f(\lambda, m)\right)\right\}, \tag{3.81}$$

where $f(\lambda, m) = \mathbf{1}\{\lambda > \delta\} m D(\lambda||\delta) + \mathbf{1}\{\lambda \leq \delta\} D(\lambda||\delta)$. The desired result follows by observing that the maximization over λ always lead to a solution in the range $\lambda > \delta$, for any $\rho > 0$. ■

The main take-away from this section is that side-information, and thus targeted-attacks, are very powerful as that they reduce the asymptotic work of the adversary drastically. While we only looked at two specific types of side-information, in both cases, we saw an exponential decrease in the guesswork when there is side-information available. Next, we look at adversaries which do not have the perfect knowledge of the password distribution P_X.

3.3 Attacks with Distribution Mismatch

In this section, we study the probabilistic behavior of the so-called mismatched guesswork $G_Q(X)$, for $X \sim P$, and Q is a mismatched distribution $Q \neq P$. In many of the applications of guesswork to brute-force security, mismatch is inevitable in practice, as the source distribution is obtained via a sample estimation which is prone to imprecision. We consider the case where a sequence of length n denoted by x^n is drawn i.i.d. from P, while the mismatched distribution is the product distribution of Q, denoted Q^n. We prove that, on the one hand, $G_{Q^n}(x^n)$ is related to the entropy of the "projection" of the type of x^n on the tilted family of the mismatched distribution Q. On the other hand, the probability of a sequence x^n is related to the KL-divergence of its type with the true distribution P.

In the appendix, we also explore the application of mismatched guesswork in one-to-one source coding, i.e., source coding without the prefix constraint. Mismatched guesswork has a direct application in this setting, and is the counterpart of the usual mismatch prefix-free source coding. It is well known that, in contrast to the prefix-free source codes, the average length of the one-to-one source codes converge to the entropy rate from below at a rate $-1/2\log(n)/n$ when the distribution is matched [138]. It was also shown that the cost of universality is smaller in one-to-one codes because of one less degrees of freedom [87, 25]. To

complete the characterization, we show that one-to-one source codes are more robust to an incorrect knowledge of the source distribution. Moreover, it is possible to obtain the exact same optimal performance of an optimal one-to-one encoder with a mismatched distribution Q, under the condition that Q is on the tilted family of the true distribution P.

3.3.1 Mismatched Guesswork

In this section, we investigate the behavior of $G_{Q^n}(X^n)$, when $X^n \sim P^n$, and establish an LDP result. We also aim at characterizing the exponent of the growth of the moments of mismatched guesswork, denoted by $E_\rho(Q\|P)$, and defined as

$$E_\rho(Q\|P) = \frac{1}{\rho} \lim_{n\to\infty} \mathbb{E}_{P^n} \left[G_{Q^n}(X^n)^\rho \right]. \tag{3.82}$$

The following result, proved in [25, Theorem 1], characterizes the mismatched guesswork in the case where $P \in \mathcal{T}_Q$.

Lemma 13 (mismatched guesswork on the same tilted family)**.** *Let $P \in \mathcal{T}_Q^+$, then $G_Q(x) = G_P(x)$. Alternatively, let $P \in \mathcal{T}_Q^-$, then $G_Q(x) = |\mathcal{X}| - G_P(x)$.*

Note that the previous result is non-asymptotic. In particular, it follows readily that $E_\rho(Q\|P) = E_\rho(P)$ when $P \in \mathcal{T}_Q^+$ and $E_\rho(Q\|P) = \log(|\mathcal{X}|)$ when $P \in \mathcal{T}_Q^-$.

However, the techniques in [25] fall short on characterizing mismatch for $P \notin \mathcal{T}_Q$. Such characterization is given in the following theorem.

Theorem 13 (LDP for mismatched guesswork)**.** *For any unambiguous P and Q, such that $\Pi_{\mathcal{T}_Q}(P) \in \mathcal{T}_Q^+$, the sequence $\{\frac{1}{n} g_Q(x^n)\}_{n\in\mathbb{N}^+}$ satisfies a LDP, with rate function $J(t)$, and the rate function is implicitly given by*

$$J(t) = D(\gamma_{Q,P}(t)\|P), \tag{3.83}$$

for

$$\gamma_{Q,P}(t) = \mathcal{T}_{Q,P} \cap \mathcal{L}(Q, \alpha(t)), \tag{3.84}$$

$$\alpha(t) = \arg_{\alpha\geq 0}\{H(T(Q,\alpha)) = t\}. \tag{3.85}$$

Before we proceed to the proof, let us briefly discuss the result. Two features of this re-

sult are particularly interesting. First, note that while the value $\alpha(t)$ is determined through a similar implicit equation as the matched guesswork in Theorem 7, the rate function is controlled by $D(\gamma_{Q,P}(t)\|P)$, where $\gamma_{Q,P}(t) \in \mathcal{T}_{Q,P}$. In particular, if $P \in \mathcal{T}_Q^+$, then Theorem 13 recovers Theorem 7 by observing that $Q = T(P, \beta)$ for some $\beta > 0$, and thus $\gamma_{Q,P}(t)$ can be reparameterized in terms of P only.

The proof of Theorem 13 relies on a correspondence between guesswork, and some sets of distributions, which we will define shortly. This correspondence is implicitly used in [25, proof of Theorem 5] but it is not explicitly observed. For $\epsilon \geq 0$ and $\alpha \in \mathbb{R}$, let

$$\mathcal{D}(Q, \alpha, \epsilon) \triangleq \{\varphi \in \Delta_{\mathcal{X}} : H(\varphi\|Q) - H(T(Q, \alpha)\|Q) \leq \epsilon\} \tag{3.86}$$

$$\mathcal{E}(Q, \alpha, \epsilon) \triangleq \{\varphi \in \Delta_{\mathcal{X}} : H(\varphi\|Q) - H(T(Q, \alpha)\|Q) \geq -\epsilon\} \tag{3.87}$$

$$\mathcal{B}(Q, \alpha, \epsilon) \triangleq \{\varphi \in \Delta_{\mathcal{X}} : H(T(Q, \alpha)\|Q) - H(\varphi\|Q) \in [0, \epsilon]\}. \tag{3.88}$$

The sets above are extensions of tilted weakly typical sets of order α [25, Definition 18], and capture the set of types which are respectively, more likely, less likely, and as likely according to Q than $T(Q, \alpha)$. For these sets, we then have the following lemma.

Lemma 14. *For any $\alpha > 0$, the following inclusion relations hold, for sufficiently large n,*

$$\left|\frac{1}{n} g_Q(x^n) - H(T(Q, \alpha))\right| \leq \epsilon \Rightarrow \mathbf{q}_{x^n} \in \mathcal{D}(Q, \alpha, 2\epsilon/\alpha), \tag{3.89}$$

$$\left|\frac{1}{n} g_Q(x^n) - H(T(Q, \alpha))\right| \leq \epsilon \Rightarrow \mathbf{q}_{x^n} \in \mathcal{E}(Q, \alpha, 2\epsilon/\alpha), \tag{3.90}$$

$$\left|\frac{1}{n} g_Q(x^n) - H(T(Q, \alpha))\right| \leq \epsilon \Leftarrow \mathbf{q}_{x^n} \in \mathcal{B}(Q, \alpha, \epsilon/\alpha). \tag{3.91}$$

This was proved implicitly in the proofs of Theorems 3 and 5 in [25]. We are now equipped to provide the proof of the main theorem.

Proof of Theorem 13. Observe that by Lemma 13, for any $Q* \in \mathcal{T}_Q^+$ we have $G_{Q*}(x) = G_Q(x)$ for all $x \in \mathcal{X}$. In particular, this holds for $Q* = \Pi_{\mathcal{T}_Q}(P)$. Therefore, without loss of generality throughout the proof we assume that $Q = \Pi_{\mathcal{T}_Q}(P)$.

Next, note that as $\frac{1}{n} g_{Q^n}(X^n)$ takes values in a compact subset $[0, \log|\mathcal{X}|]$ of $\mathbb{R}$, it is sufficient to prove that the limit below exists and evaluates to the rate function (see [25,

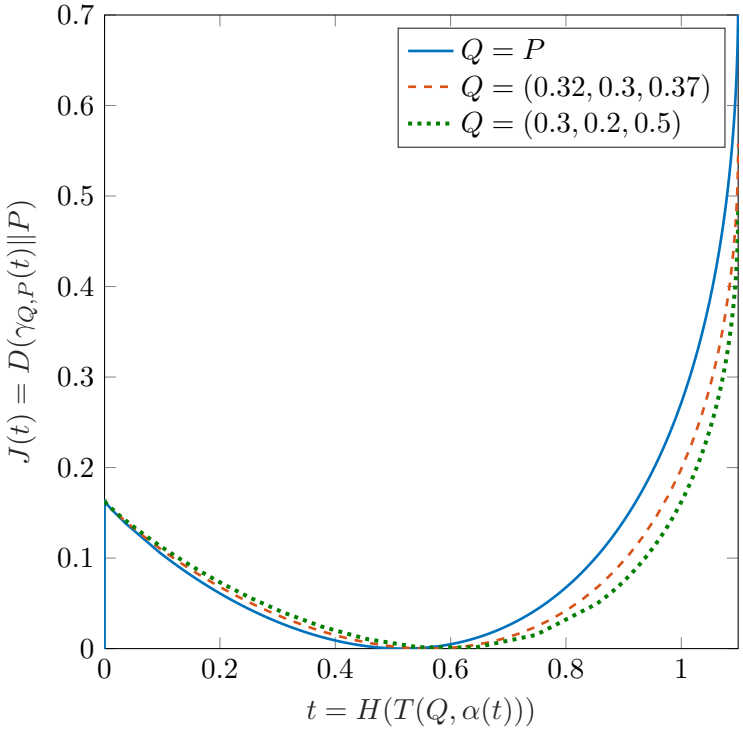

Figure 3-8: Rate function $J(t)$ of $\{\frac{1}{n} g_Q(X^n)\}$, for a distribution over three symbols $P = (0.05, 0.1, 0.85)$.

Section V] for a formal discussion), i.e.,

$$\lim_{\epsilon \downarrow 0} \lim_{n \to \infty} \frac{1}{n} \log \mathbb{P}_P^n \left(\left| \frac{1}{n} g_{Q^n}(X^n) - t \right| < \epsilon \right) = -J(t). \tag{3.92}$$

We proceed with the proof in three separate cases.

Case (a): We let $t \in (H(Q), \log |\mathcal{X}|)$, which implies $\alpha(t) \in (0, 1)$ by monotonicity of $H(T(Q, \alpha))$ for non-negative α. Note that (3.89) and (3.91) respecitvely imply

$$\begin{aligned}
&\lim_{\epsilon \downarrow 0} \limsup_{n \to \infty} \frac{1}{n} \log \mathbb{P}_{P^n} \left(\left| \frac{1}{n} g_Q(X^n) - H(T(Q, \alpha(t))) \right| \leq \epsilon \right) \\
&\leq \lim_{\epsilon \downarrow 0} \limsup_{n \to \infty} \frac{1}{n} \log \mathbb{P}_{P^n}(\mathbf{q}_{X^n} \in \mathcal{D}(Q, \alpha(t), 2\epsilon/\alpha(t))), \qquad (3.93) \\
&\lim_{\epsilon \downarrow 0} \liminf_{n \to \infty} \frac{1}{n} \log \mathbb{P}_{P^n} \left(\left| \frac{1}{n} g_Q(X^n) - H(T(Q, \alpha(t))) \right| \leq \epsilon \right) \\
&\geq \lim_{\epsilon \downarrow 0} \liminf_{n \to \infty} \frac{1}{n} \log \mathbb{P}_{P^n}(\mathbf{q}_{X^n} \in \mathcal{B}(Q, \alpha(t), \epsilon/\alpha(t))). \qquad (3.94)
\end{aligned}$$

Thus, it suffices to show that the RHS of (3.93) and (3.94) both evaluate to $-D(\gamma_{Q,P}(t)\|P)$. This is done via Sanov's Theorem. Recall that Sanov's Theorem [58, Theorem 6.2.10] states

that, for a set of distributions $\mathcal{C}$,

$$\begin{aligned} -\inf_{\gamma \in \text{int}\mathcal{C}} D(\gamma\|P) &\leq \liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}(\mathbf{q}_{x^n} \in \mathcal{C}) \\ &\leq \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}(\mathbf{q}_{x^n} \in \mathcal{C}) \\ &\leq -\inf_{\gamma \in \text{cl}\mathcal{C}} D(\gamma\|P). \end{aligned} \tag{3.95}$$

To obtain the upper bound, we apply this result to the set $\mathcal{D}(Q, \alpha(t), 2\epsilon/\alpha(t))$. Observing that this holds for any ϵ, and then letting $\epsilon \downarrow 0$, we get that the RHS of (3.93) is upper bounded

$$-\lim_{\epsilon\downarrow 0} \inf_{\gamma \in \text{cl}\mathcal{D}(Q,\alpha(t),2\epsilon/\alpha(t))} D(\gamma\|P). \tag{3.96}$$

We now make use of a basic topological fact. Observe that $D(\gamma\|P)$ is strictly convex in γ for a fixed P, and thus there is a unique minimizer $\gamma(t,\epsilon)$. Noting that the minimizer $\gamma(\epsilon, t)$ is in the set $\text{cl}\mathcal{D}(Q, \alpha(t), 2\epsilon/\alpha(t))$, by continuity of $D(\gamma\|P)$ and compactness of the set. Thus, the collection of minimizers $\gamma(t,\epsilon)$ is a collection of points such that $\gamma(t,\epsilon) \in \text{cl}\mathcal{D}(Q, \alpha(t), 2\epsilon/\alpha(t))$. It follows from compactness that the limit point $\lim_{\epsilon\downarrow 0} \gamma(\epsilon, t) \in \bigcap_{\epsilon>0} \text{cl}\mathcal{D}(Q, \alpha(t), 2\epsilon/\alpha(t)) = \text{cl}\mathcal{D}(Q, \alpha(t), 0)$, where we have used that $\alpha(t) > 0$. Therefore, we have the bound

$$\begin{aligned} \inf_{\gamma\in\Delta_{\mathcal{X}}} \quad & D(\gamma\|P) \\ \text{subject to} \quad & H(\gamma\|Q) \leq H(T(Q,\alpha(t))\|Q) \end{aligned} \tag{3.97}$$

Note that this optimization problem is convex, and thus can be solved analytically by writing the KKT conditions [36], which give a solution $\gamma_{Q,P}(t) \in \mathcal{T}_{Q,P}$, and optimal value $D(\gamma_{Q,P}(t)\|P)$.

Analogously, the RHS of (3.91) can be shown to be lower bounded by $-\inf D(\gamma\|P)$, where $\gamma \in \mathcal{B}(Q, \alpha(t), 0)$, by noting $\mathcal{B}(Q,\alpha,0) \subset \text{int}\mathcal{B}(Q,\alpha,\epsilon)$, for any $\epsilon > 0$. Again, this optimization can be solved analytically, and gives the desired output. Putting these results together, we get that

$$\lim_{\epsilon\downarrow 0} \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}_P^n \left(\left| \frac{1}{n} g_{Q^n}(X^n) - H(T(Q,\alpha(t))) \right| < \epsilon \right)$$

$$= -D(\gamma_{Q,P}(t)\|P). \tag{3.98}$$

Case (b): We now let $t \in (0, H(Q))$, which implies $\alpha(t) \in (1, \infty)$. The proof in this case follows from the same step as in Case (a), by replacing the set $\mathcal{D}(Q, \alpha(t), \epsilon)$ with the set $\mathcal{E}(Q, \alpha(t), \epsilon)$.

Case (c): Finally, let $t = H(Q)$, or equivalently, $\alpha(t) = 1$. In this case, note that $P \in \mathcal{B}(Q, 1, \epsilon)$, and thus, by the law of large numbers and (3.94), we have that

$$\lim_{\epsilon \downarrow 0} \lim_{n \to \infty} \frac{1}{n} \log \mathbb{P}_P^n \left(\left| \frac{1}{n} g_{Q^n}(X^n) - t \right| < \epsilon \right) \geq 0, \tag{3.99}$$

which implies that $J(t) = 0$ in this case. ■

As mentioned before, an attractive feature of the LDP is that it implies the asymptotic average growth rate of the ρ-th moment of the mismatched guesswork, i.e., $E_\rho(P\|Q)$. This is formalized in the following corollary, which is the second main result of this paper implied by Theorem 13.

Corollary 8. *Let $\Pi_{\mathcal{T}_Q}(P) \in \mathcal{T}_Q^+$. Then, we have*

$$E_\rho(Q\|P) = \max_{\gamma \in \mathcal{T}_{Q,P}} \quad H(\Pi_{T_Q}(\gamma)) - \frac{1}{\rho} D(\gamma\|P) \tag{3.100}$$

Proof. We use Varadhan's Lemma [58, Theorem 4.3.1], which states that if a sequence of random variables M_n satisfies a LDP with rate function $J(t)$, then we have

$$\lim_{n \to \infty} \frac{1}{n} \log \mathbb{E}_{P^n}[\exp nF(M_n)] = \sup_t F(t) - J(t), \tag{3.101}$$

for any continuous and bounded function F. Applying this results to the sequence $\{\frac{1}{n} g_Q(X^n)\}$, and letting $F(t) = \rho \cdot t$, for $\rho > 0$ and $t \in [0, \log|\mathcal{X}|]$ thus yields

$$\lim_{n \to \infty} \frac{1}{n} \log \mathbb{E}_{P^n}[G_Q^\rho(X^n)] = \sup_t \rho \cdot t - D(\gamma_{Q,P}(t)\|P). \tag{3.102}$$

Performing the optimization on γ instead of t, via the change of variables in (3.85) and (3.84) concludes the proof. ■

The following is an immediate corollary which lower bounds the mismatched guesswork.

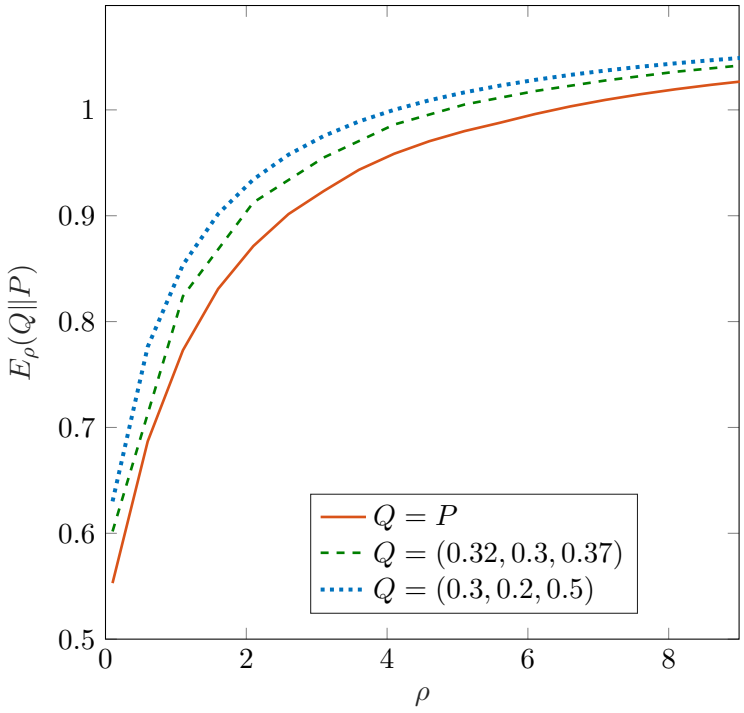

Figure 3-9: Illustration of Corollary 8. The distributions are identical to the ones in Figure 3-8. Note that, as ρ grows, the curves meet at $\log|\mathcal{X}|$.

Corollary 9 (non-negativity of mismatch penalty)**.** *Let* $\Pi_{\mathcal{T}_Q}(P) \in \mathcal{T}_Q^+$*, then the following holds:*

$$E_\rho(Q\|P) \geq E_\rho(P) = H_{\frac{1}{1+\rho}}(P), \tag{3.103}$$

with equality iff $P \in \mathcal{T}_Q^+$*.*

Proof. Consider the optimization from (3.33), and notice that it can be equivalently written as

$$\max_{\phi \in \mathcal{T}_P} \; H(\phi) - \frac{1}{\rho} D(\phi\|P) \tag{3.104}$$

Using Lemma 10, we obtain that $H(\Pi_{\mathcal{T}_Q}(\zeta)) \geq H(\zeta))$, giving the upper bound

$$\max_{\phi \in \mathcal{T}_P} \; H(\Pi_{\mathcal{T}_Q}(\phi)) - \frac{1}{\rho} D(\phi\|P). \tag{3.105}$$

Next, notice that since $H(\Pi_{\mathcal{T}_Q}(\phi)\|Q) = H(\phi\|Q)$, by definition of $\Pi_{\mathcal{T}_Q}$, it must be the case that $D(\gamma\|P) < D(\phi\|P)$ for some $\gamma \in \mathcal{T}_{Q,P}$ which satisfies $H(\gamma\|Q) = H(\Pi_{\mathcal{T}_Q}(\phi)\|Q)$. It

follows that

$$\max_{\phi \in \mathcal{T}_P} \quad H(\Pi_{\mathcal{T}_Q}(\phi)) - \frac{1}{\rho} D(\gamma \| P) \tag{3.106}$$

$$\text{such that} \quad H(\gamma \| Q) = H(\Pi_{\mathcal{T}_Q}(\phi) \| Q) \tag{3.107}$$

is an upper bound to the matched guesswork. The proof follows from performing the change of variable $H(\Pi_{\mathcal{T}_Q}(\zeta)) = T(Q, \alpha)$, and identifying the resulting optimization as being equivalent to (3.100). ■

To summarize, in this section, we revisited mismatch guesswork using geometric insights. In particular, we generalized the tilted families of [25], and showed that the LDP rate function is implicitly expressed in terms of the relative entropy between distributions on this tilted family, and the true distribution P_X. These results also find applications in one-to-one lossless coding, where one can show that, perhaps surprisingly, one-to-one coding is more robust to mismatch than prefix-free coding. We refer the interested reader to Appendix C for more details on the subject. Interestingly, similar tilted distributions have appeared in the context of error exponents, see e.g. [35]. A more in depth study of the relationship between mismatched guesswork and error exponents for random coding is of interest. The next and final section of this chapter is dedicated to unsynchronized attacks which arise, for example, in the context of attacks performed by botnets.

3.4 Randomized Attacks and Botnets

As mentioned at the start of this chapter, brute-force attacks are prevalent, despite the computational burden on the attacker. This is in part explained by the fact that attacks through huge networks of compromised computers (botnets) are now more common, giving access to significant computational resources for the attacker. More critically, these botnets help to disguise the attack by distributing it. Indeed, a main solution to the threat of online brute-force attacks is to setup a system that detects and prevents too many queries from any one user, as determined by IP addresses. As such, an attacker which using only a single IP address would be limited to a fixed number of guesses. In recent years, however, this defense was circumvented by using massive botnets, each bot querying potential passwords. In this situation, it is hard to detect legitimate users in the crowd of illegitimate attackers. These

attacks come with a cost, namely, the attack is now distributed across thousands, sometimes millions of computers, each with limited computational power and synchronization tools.

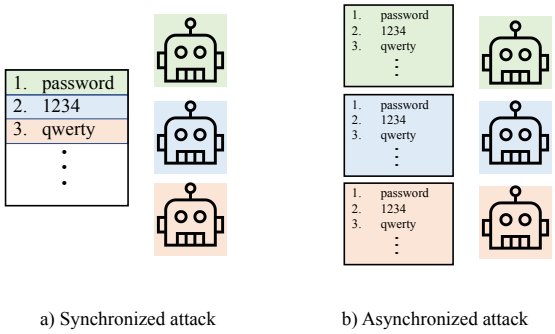

Figure 3-10: In a synchronized attack, the bots query from the password-list in a specified order. In the asynchronous attack, they do not know the order in which the queries will be sent. Our solution will consist at drawing guesses according to some distribution, instead of querying passwords one-by-one.

As a first step to understand the impact of synchronization, we put forth in this section a simplified mathematical model for passwords and brute-force attacks without synchronization. The intuition gained from this model is informative and helpful in assessing the security of systems under brute-force attacks. If multiple adversarial agents (we shall use adversary and agents interchangeably) coordinate their attack, the system will be compromised as soon as any of them succeeds. Moreover, the individual computational effort of each adversary is reduced, while the total number of queries remains the same. Indeed, an optimal strategy here would consist of having each agent query the most-likely password that has not been queried by any of the other agents. Since this strategy reduces to querying as a group from the optimal list, the average number of queries completed by each agent is thus reduced by a factor of the number of agents, with respect to the case where a single agent queries alone. This requires the agents to be able to synchronize their queries, that is, there must be a knowledge of an ordering in which the agents make guesses. However, in many practical scenarios the adversarial agents are completely distributed and have limited communication with each other. For example, in botnets agents are often oblivious to the actions taken by other adversaries, and may have limited access to shared memory or synchronization tools. Owing to constraints of the physical computers in which these bots run, the speed, latency, and reliability of these agents is heterogeneous — thus, perfect synchronization is unlikely. Note that even if a central agent distributes lists of possible guesses to the bots,

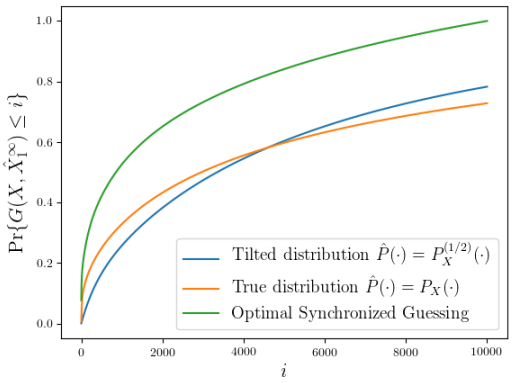

Figure 3-11: Probability of finding the password in fewer than i queries. In a synchronized attack, the passwords has to be found after at most $|\mathcal{X}| = 1e4$ queries. The blue and orange line correspond to i.i.d. guesses according to the distribution $\hat{P}$.

such that the lists form a partition of all guesses, making sure no guess is repeated, the lack of synchronization may still render the process sub-optimal. We illustrate an example of synchronized and asynchronous attack in Fig 3-10. At one extreme, a complete lack of synchronization can be modeled by a worst-case optimization, in which the guesses of each agent come in the worst possible order. The goal of this section is to study how much the lack of synchronization, as described above, might affect the overall number of queries that are made until the game ends. We discuss why deterministic strategies cannot perform well in this paradigm, while on the other hand, a simple randomized strategy in which all the guesses are drawn i.i.d. from a certain distribution asymptotically achieves the same optimal performance of a synchronous attack when guessing secrets that are long sequences drawn according to some types of distribution. This optimal guessing distribution is non-trivial, and, perhaps surprisingly, it is not the original password generating distribution P_X. It is a tilted distribution from P_X, where the tilt exponent depends on the moment of guesswork of interest. In other words, distributed and asynchronous agents can adopt a strategy for which the asymptotic number of total queries sent before a system breach is optimal, regardless of the ordering in which these queries are received, but this distribution is only optimal for a given moment of guesswork, and not optimal universally across all moments.

To illustrate the proposed scheme, we have shown our results on an extract of the Adobe Leaked password dataset (see [4] for a description of the dataset). In particular, we extracted the 10^4 most likely passwords from a subset of 10 millions passwords in the data, and

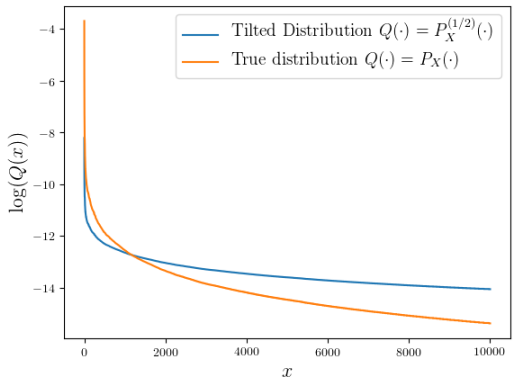

Figure 3-12: Log-probability mass function. Notice how the tilted distribution gives more weight to less likely symbols, as they correspond to the symbol which are the most costly for password guessing.

restricted our study to those passwords. We investigate the guesswork when the correct password is drawn according to the distribution P_X as computed on this restricted sample of the data. We show in Figures 3-11 and 3-12, the performance of a randomized strategy when using the optimal guessing distribution versus the naive distribution P_X, both in terms of expected number of guesses and in terms of probability of making less than a fixed number of guesses. Note that the true distribution P_X performs well if one wishes to make only a small number of guesses, but eventually takes longer to reach a high probability. This is due to less frequent passwords, which are barely ever queried if guesses are drawn according to P_X. The guessing distribution which optimizes the average number of guesses increases the probability of querying the less likely passwords, as those passwords represent the main computational burden on the adversary when they occur.

Main Contributions: We define a min-max formulation that models a worst case asynchronous attack from the attacker's perspective, and show that a randomized strategy in which each guess is drawn i.i.d. from a certain distribution achieves the same asymptotic performance (in the length of the password sequence n) as an optimal synchronized attack. This optimal distribution is non-trivial; performing guesses according to the distribution from which the password was generated yields a strategy that is exponentially worse than the optimal guessing distribution. In fact, the optimal choice is a tilted distribution, where the tilt parameter is chosen depending on the moment of guesswork to be optimized. We also discuss optimal strategies when the benchmark is to maximize the probability of success

of an attack with a fixed number of overall queries, and show that an i.i.d. guessing strategy again has optimal performance asymptotically. The optimal distribution is again a tilted distribution, where the tilt depends on the number of queries allowed. Together these results indicate that there is no loss in performance (asymptotically in n) when performing an asynchronous attack.

3.4.1 Asynchronous Brute-Force Attack

In this section, we discuss synchronization when multiple agents aim to breach a secured system. Recall that we say that distributed agents are synchronized if they know in which order every agent's queries will be received by Alice. In this case, they can query from the optimal list as a group, *i.e.*, the first query received is the most likely symbol, etc. In other words, full synchronization means they can all share a single (optimal) list, and a *pointer* to this list advancing after each new guess. As a result, the total number of queries sent is the same as the optimal single agent guesswork, namely, the optimal result from Corollary 3 is achieved, while the individual computational burden on each agent is reduced since the queries are divided among agents. Further, even if the number of adversaries grows exponentially[6] with the length of the password n, the total number of queries remains the same [7].

Instead, if agents do not know in which order the queries are delivered, they must adopt a strategy which performs well under any such ordering. In particular, we shall adopt a worst-case approach in which the goal is to minimize the number of queries in the worst ordering. Specifically, let $\mathbf{X}$, an i.i.d. sequence of length n generated from P_X, be the sequence to be guessed, and let $\{\hat{\mathbf{X}}_k^{(a)} : k \geq 1\}$ be the strategy of agent $a \in \mathcal{A}$, where $\mathcal{A}$ is a, possibly infinite, countable set. Again, we shall be interested in the regime where $|\mathcal{A}|$ grows at least exponentially fast with n, and the goal is to characterize number of queries made in total. We let the permutation $\pi : \mathbb{N}^+ \to \mathcal{A} \times \mathbb{N}^+$ denote the ordering in which the queries are received, i.e., $\pi(i) = (a_i, k_i)$ means that the i-th query received is $\hat{X}_{k_i}^{(a_i)}$. Denote by Π the set of all such possible orderings. Under an ordering π, Alice receives the sequence of

[6]Note that in practice, the number of agents usually needs to grow since most secured systems include a mechanism which blocks IP addresses after a given number of password attempts. Thus, if a single agent can only make k queries, there must be at least $\lceil |\mathcal{X}|^n/k \rceil$ agents to guarantee that a password of length n will be found.

[7]Note that in this work we use the total number of queries as the main metric for computational effort, as opposed to e.g. [38] where the average number of guesses *per agent* is characterized.

queries $\pi(\hat{\mathbf{X}}_1^\infty) \triangleq \{\hat{\mathbf{X}}_{k_i}^{(a_i)} : i \geq 1\}$. Note that this permutation allows reordering of guesses of a given agent $a \in \mathcal{A}$ which may be received in any arbitrary order. For some fixed strategies $\{\hat{\mathbf{X}}_k^{(a)} : k \geq 1\}$, the worst ordering in terms of guesswork is thus given by

$$\sup_{\pi \in \Pi} \mathbb{E}\left\{G(\mathbf{X}, \pi(\hat{\mathbf{X}}_1^\infty))^\rho\right\}. \tag{3.108}$$

The goal of the agents is to minimize the worst-case number of queries, or, in other words, solve the min-max problem

$$\inf_{\{\hat{X}_k^{(a)}, k \geq 1\} \text{ for } a \in \mathcal{A}} \sup_{\pi \in \Pi} \mathbb{E}\left\{G(\mathbf{X}, \pi(\hat{\mathbf{X}}_1^\infty))^\rho\right\}. \tag{3.109}$$

The main result of this section, presented below, characterizes the asymptotic exponent of (3.109), as $n \to \infty$. The proof of this result, along with the associated lemmas, are given after some discussion.

Theorem 14. *For $\mathbf{X}_n$ an i.i.d. sequence according to P_X, and $\{\hat{\mathbf{X}}_k^{(t)}, k \geq 1\}$ sequences of guesses which are independent over $a \in \mathcal{A}$, we have the following*

$$\begin{aligned}
&\lim_{n\to\infty} \frac{1}{n} \log \left(\inf_{\{\hat{\mathbf{X}}_k^{(t)} : k \geq 1\}} \sup_{\pi \in \Pi} \mathbb{E}\left\{G(\mathbf{X}_n, \pi(\hat{\mathbf{X}}_1^\infty))^\rho\right\} \right) \\
&\quad = \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}\left\{G^*(\mathbf{X}_n)^\rho\right\} \\
&\quad = \rho \cdot H_{\frac{1}{1+\rho}}(X).
\end{aligned} \tag{3.110}$$

Note that guesswork measures the *total number of guesses made by the agents.* Thus it is clear that with full synchronization among the agents this value will not depend on $|\mathcal{A}|$. In a sense, dependence on $|\mathcal{A}|$ for a certain scheme would indicate a *lack of synchronization*, as it would suggest that queries are repeated by the agents. Surprisingly, Theorem 14 states that even under a worst-case assumption, there exist a strategy under which the guesswork does not depend on $|\mathcal{A}|$ and is similar to the fully synchronous case. The above result show that *synchronization is not necessary to achieve the asymptotic optimal guessing performance.* This can be equivalently formulated by an achievability strategy, and a converse. The converse result is trivial, as the performance of the synchronized strategy $\mathbb{E}\{G^*(\mathbf{X})\}$ upper bounds (3.109).

Lemma 15 (Converse)**.** *For any strategy* $\hat{\mathbf{X}}^{\infty}$*,*

$$\inf_{\{\hat{X}_k^{(t)},k\geq 1\} \text{ for } a\in\mathcal{A}} \sup_{\pi\in\Pi} \mathbb{E}\left\{G(\mathbf{X},\pi(\hat{\mathbf{X}}_1^{\infty}))^{\rho}\right\} \geq \mathbb{E}\left\{G^*(\mathbf{X})\right\}. \tag{3.111}$$

We now turn to finding an appropriate strategy which would match this converse bound. Let us first examine a naive solution to this problem. Consider the strategy which consists in letting each agent construct the optimal list and query it individually, that is $X_1^{(a)}$ is the most likely symbol for all $a \in \mathcal{A}$, $X_2^{(a)}$ the second most likely symbol, etc. It is easy to see that (3.108) would evaluate to a quantity which grows with the number of agents $|T|$. Indeed, many queries are duplicated, and thus the overall number of queries grows with $|\mathcal{A}|$, without even reducing the computational burden on each adversary since they all must query the same password strings. Note that this remains true if one considers a less stringent worst-case analysis, by for example, letting the guesses of each of the agent to be consistent among themselves, i.e. the permutation does not change the relative order of the guesses of each agent.

If instead the agents agree on a partition of the guesses before the attack, in a way such that no two guesses are repeated, then the correct password is queried by one unique agent. Again, it is easy to see that the worst-case analysis yields a quantity which grows with $|\mathcal{A}|$, even though it cannot grow beyond $|\mathcal{X}|^n$, as every unique password is queried at most once. In particular, if $|\mathcal{A}| = |\mathcal{X}|^n$, then the worst-case analysis achieves its upper-bound. Note that these observations are not only an artifact of the worst-case analysis, but rather a consequence of the deterministic nature of the queries.

This motivates us to study randomized strategies. In particular, we consider guesses, which are randomly and independently drawn according to a specific distribution, independent from each other, and identically distributed. We then study this optimal distribution in terms of the expected moments of guesswork. Consider first a scalar $X \in \mathcal{X}$, generated from P_X. We let $\{\hat{X}_k^{(a)}, k \geq 1\}$ be an i.i.d. process with respect to $\hat{P}(\cdot)$, for all $a \in \mathcal{A}$. For a given $\rho > 0$, we define the quantity

$$V_{\rho}(X,\hat{X}_1^{\infty}) \triangleq \binom{G(X,\hat{X}_1^{\infty})+\rho-1}{\rho}, \tag{3.112}$$

where $\binom{x}{y}$ is the generalized binomial coefficient defined in terms of the Gamma function

$\Gamma(\cdot)$, i.e.

$$\binom{x}{y} = \frac{\Gamma(x+1)}{\Gamma(y+1)\Gamma(x-y+1)}. \tag{3.113}$$

In particular, $V_1(X, \hat{X}_1^\infty) = G(X, X_1^\infty)$. The motivation for this definition of $V_\rho(X, \hat{X}_1^\infty)$ will be made clear in the proof of Lemma 16, where it allows us to compute a particular infinite sum neatly. Note that for large $G(X, \hat{X}_1^\infty)$ and fixed integer ρ, Stirling's approximation of the binomial coefficient directly gives $V_\rho(X, \hat{X}_1^\infty) \approx G(X, \hat{X}_1^\infty)^\rho/\rho!$, therefore $V_\rho(X, \hat{X}_1^\infty)$ approximates the behavior of the guesswork moment $G(X, \hat{X}_1^\infty)^\rho$, up to some factor.

We are interested in the following optimization problem

$$\mathbb{E}\{V_\rho^*(X, \hat{X}_1^\infty)\} \triangleq \inf_{\hat{P} \in \mathcal{P}} \mathbb{E}\{V_\rho(X, \hat{X}_1^\infty)\}, \tag{3.114}$$

where $\mathcal{P}$ is the probability simplex and $\{\hat{X}_k : k \geq 1\}$ is generated i.i.d. from $\hat{P}$. We let $\hat{P}_\rho^*$ designate the minimizer. The following Lemma is the main ingredient in proving an achievability and thus Theorem 14.

Lemma 16. *For any $\rho \geq 1$,*

$$\log \mathbb{E}\{V_\rho^*(X, \hat{X}_1^\infty)\} = \rho \cdot H_{\frac{1}{1+\rho}}(X), \tag{3.115}$$

and for any $x \in \mathcal{X}$,

$$\hat{P}_\rho^*(x) = \frac{P_X(x)^{\frac{1}{1+\rho}}}{\sum_{x' \in \mathcal{X}} P_X(x')^{\frac{1}{1+\rho}}}. \tag{3.116}$$

Before providing the proof of Lemma 16 we briefly discuss our result. First, we note that contrary to Corollary 3, the above result provides an *exact* operational meaning for Rényi entropy $H_\alpha(X)$ of order $\alpha > 0$. It should be mentioned here that a similar interpretation for $H_{1/2}(X)$ was reported in [37, 78, 38]. Also, we see that the optimal guessing distribution (3.116) is simply the tilted distribution of P_X of order $1/(1+\rho)$. It should be emphasized that, since the function $f(x) = x^{1/1+\rho}$ is monotone, creating an optimal list according to $\hat{P}_X$ yields the exact same list as if done according to P_X. However, the list of guesses chosen i.i.d. according to $\hat{P}_X$ will be different from the one if guesses are made i.i.d. according to

P_X. Indeed, letting $\hat{P}(x) = P_X(x)$ gives

$$\log \mathbb{E}\{G(X, \hat{X}_1^\infty)\} = \log |\mathcal{X}| ,$$

which could be much worse than $\log \mathbb{E}\{V_1^*(X, \hat{X}_1^\infty)\} = H_{1/2}(X)$. Namely, when one is allowed only to guess passwords according to a certain distribution, independently, and without a list, then using the original distribution is strictly sub-optimal, and the tilted distribution should be used. This result is related to similar results from the source-coding literature in which a tilted distribution also appears as the solution of an optimization where longer codewords are penalized exponentially (see e.g. [42, 31]). Finally, note that the result is not asymptotic. In particular, the randomized strategy can be used over an alphabet $\mathcal{X}$ where each $x \in \mathcal{X}$ corresponds to a password. This result is thus relevant to dictionary attacks, where queries are drawn according to a dictionary of possible passwords, and suggests that distributed dictionary attacks should use a guessing distribution which is a tilted version of the true distribution.

Proof of Lemma 16. First, note that given X, $G(X, \hat{X}_1^\infty)$ is a geometric random variable, and for $k \geq 1$,

$$\Pr\{G(X, \hat{X}_1^\infty) = k\} = \sum_{x \in \mathcal{X}} P_X(x)(1 - \hat{P}(x))^{k-1} \hat{P}(x).$$

Then, for any $\rho > 0$, we have

$$\begin{aligned} \mathbb{E}\{V_\rho(X, \hat{X}_1^\infty)\} &= \sum_{m=1}^{\infty} \binom{m+\rho-1}{m-1} \Pr\{G(X, \hat{X}_1^\infty) = m\} \\ &= \sum_{x \in \mathcal{X}} P_X(x)\hat{P}(x) \sum_{m=1}^{\infty} \binom{m+\rho-1}{m-1} (1 - \hat{P}(x))^{m-1}. \end{aligned}$$

In the following, we calculate the second summation term in the r.h.s. of the last equality. This is equivalent to calculating

$$\sum_{m=1}^{\infty} \binom{m+\rho-1}{\rho} y^{m-1}.$$

Note that, using the identity $\Gamma(x+1) = x\Gamma(x)$ recursively, we get that

$$\begin{aligned}\frac{\Gamma(m+\rho)}{\Gamma(\rho+1)} &= (m+\rho-1)\cdot(m+\rho-2)\cdots(\rho+1) \\ &= (-1)^{m-1}(-\rho-1)\cdot(-\rho-2)\cdots(-\rho-m+1) \\ &= (-1)^{m-1}\frac{\Gamma(-\rho)}{\Gamma(-\rho-m+1)}, \end{aligned} \tag{3.117}$$

which yields $\binom{m+\rho-1}{\rho} = (-1)^{m-1}\binom{-\rho-1}{m-1}$, and together with the change of variable $k = m-1$ we obtain

$$\sum_{m=1}^{\infty}\binom{m+\rho-1}{m-1}y^{m-1} = \sum_{k=0}^{\infty}\binom{-\rho-1}{k}(-y)^k \tag{3.118}$$

$$= (1-y)^{-\rho-1}, \tag{3.119}$$

where the last equality follows from the binomial formula. Thus,

$$\begin{aligned}\mathbb{E}\{V_\rho(X,\hat{X}_1^\infty)\} &= \sum_{x\in\mathcal{X}} P_X(x)\hat{P}(x)\frac{1}{\hat{P}(x)^{1+\rho}} \\ &= \sum_{x\in\mathcal{X}}\frac{P_X(x)}{\hat{P}(x)^\rho}. \end{aligned} \tag{3.120}$$

Next, we minimize the last expression with respect to $\hat{P}\in\mathcal{P}$. To this end, since (3.120) is convex in $\hat{P}$, $\hat{P}^*$ is given by the solution of (for $x\in\mathcal{X}$)

$$-\rho\cdot\frac{P_X(x)}{\hat{P}^*(x)^{\rho+1}} + \lambda = 0,$$

where λ is a Lagrange multiplier, and thus,

$$\hat{P}^*(x) = \frac{P_X(x)^{\frac{1}{1+\rho}}}{\sum_{x'\in\mathcal{X}}P_X(x')^{\frac{1}{1+\rho}}}.$$

On substituting this optimal distribution in (3.120) we finally get

$$\mathbb{E}\{V_\rho^*(X,\hat{X}_1^\infty)\} = \sum_{x\in\mathcal{X}}\frac{P_X(x)}{\hat{P}^*(x)^\rho} = \left(\sum_{x\in\mathcal{X}}P_X(x)^{\frac{1}{1+\rho}}\right)^{1+\rho},$$

as claimed. ■

The previous lemma applies to a scalar RV X, but can be easily extended to sequences $\mathbf{X}_n$, as shown in the following corollary.

Corollary 10. *Let* $\mathbf{X}$ *be a sequence of length* n *generated i.i.d. from* P_X*. Then, we have,*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}\{V_\rho^*(\mathbf{X}, \hat{\mathbf{X}}_1^\infty)\} = \rho \cdot H_{\frac{1}{1+\rho}}(X). \tag{3.121}$$

Proof. Treating $\mathbf{X}$ as a random vector, a direct application of Lemma 16 yields

$$\begin{aligned} \log \mathbb{E}\{V_\rho^*(\mathbf{X}, \hat{\mathbf{X}}_1^\infty)\} &= \rho \cdot H_{\frac{1}{1+\rho}}(\mathbf{X}) \\ &= \left(\sum_{\mathbf{x}\in\mathcal{X}^n} P_{\mathbf{X}}(\mathbf{x})^{\frac{1}{1+\rho}} \right)^{1+\rho}. \end{aligned}$$

The desired result follows by the additivity of the Rényi entropy. ■

Note that when $\boldsymbol{X}$ is generated i.i.d., tilting the marginal distributions and drawing symbols i.i.d., or tilting the entire product distribution result in the same optimal distribution.

Remark 6. We note that the result above can be generalized to passwords $\boldsymbol{X}$ which are generated according to an irreducible stationary Markov Chain. More precisely, let $U = (U_{ab})$ and γ_a, for $a, b \in \mathcal{X}$, be the stochastic matrix and stationary distribution of the Markov chain, respectively, so that

$$\Pr\{\mathbf{X} = (x_1 \dots x_n)\} = \gamma_{x_1} \prod_{i=1}^{n-1} U_{x_i x_{i+1}} \tag{3.122}$$

Then, it was shown in [135] that

$$\lim_{n\to\infty} \log \mathbb{E}\{G^*(\boldsymbol{X})^\rho\} = \frac{1}{1+\rho} \log \lambda, \tag{3.123}$$

where λ is the Perron-Frobenius eigenvalue of the matrix with entries $W = (U_{ab}^{1/1+\rho})$ for $a, b \in \mathcal{X}$. Further, let $\{l_a\}$ and $\{r_a\}$ be the left and right eigenvectors of W associated with λ, that is

$$\sum_{a\in\mathcal{X}} l_a = 1, \quad \sum_{a\in\mathcal{X}} l_a W_{ab} = \lambda l_b, \quad \sum_{b\in\mathcal{X}} r_b W_{ab} = \lambda r_a. \tag{3.124}$$

Analogously to the result of Corollary 10, it can be shown that generating guesses $\hat{\mathbf{X}}$ ac-

cording to a Markov Chain with entries $W_{ab}r_b/(\lambda r_a)$ achieves the asymptotic performance in (3.122). A proof of this fact follows from steps outlined in [135] along with the proof of Lemma 16.

Remark 7. In the standard guessing problem [14] Alice tries to guess X using her knowledge of P_X. It is assumed that there are no constraints on the memory of Alice, namely, for each new guess, Alice knows her previous guesses, and thus she can adapt her new guess accordingly (i.e., she will not guess again a previous incorrect guess). The setting we consider here is equivalent to one in which Alice cannot keep track of her guesses, but still knows the distribution P_X. It should be clear that in this case all that Alice can do is to present a sequence of i.i.d. guesses $\hat{X}_1, \hat{X}_2, \ldots$, drawn from some distribution $\hat{P}(\cdot)$, which shall be optimized in some sense. Lemma 16 can be equivalently interpreted as the performance of a memoryless, (or oblivious) attacker [37, 78, 84, 38].

We are now ready to prove Theorem 14.

Proof of Theorem 14. We start by noting that letting $\{\hat{\mathbf{X}}_k^{(t)} : k \geq 1\}$ be an i.i.d. process distributed according to $\hat{P}^*$ (as defined in Lemma 16) gives an upper bound on (3.109). We prove that two bounds match asymptotically, by showing that the exponent of the upper-bound is equal to $\rho \cdot H_{1/\rho+1}(X)$. Indeed, let $\{\mathbf{X}_k^{(t)} : k \geq 1\}$ be an i.i.d. process distributed according to $\hat{P}^*$ for all $t \in T$. Then, it is evident that $\pi(\hat{\mathbf{X}}_1^\infty)$ is also an i.i.d. process distributed according to $\hat{P}^*$, for any permutation $\pi \in \Pi$. An application of Corollary 10 concludes the proof. ■

Note that the optimal distribution from Lemma 16 depends on the moment ρ. Indeed, the larger ρ, the more we are penalized for passwords which are less frequent (which increase the work significantly). Therefore, the optimal strategy gives extra weight to less frequent symbols as to make sure that they are more likely to be chosen than what their probability suggests. We do so by raising P_X to a power $1/1+\rho$. Nevertheless, the optimal distribution, and thus guessing strategy, will change as a function of the guesswork moment ρ of interest. This contrasts with the synchronous case, in which the optimal strategy consisting of querying the sequences from most likely to least likely is optimal universally for all moments ρ. This loss of universality is exploited in the following corollary, which characterizes the loss in using a distribution optimized for a moment $\rho > 0$, when measured in terms of a moment $\gamma \neq \rho$, and is illustrated for a binary source in Figure 3-13.

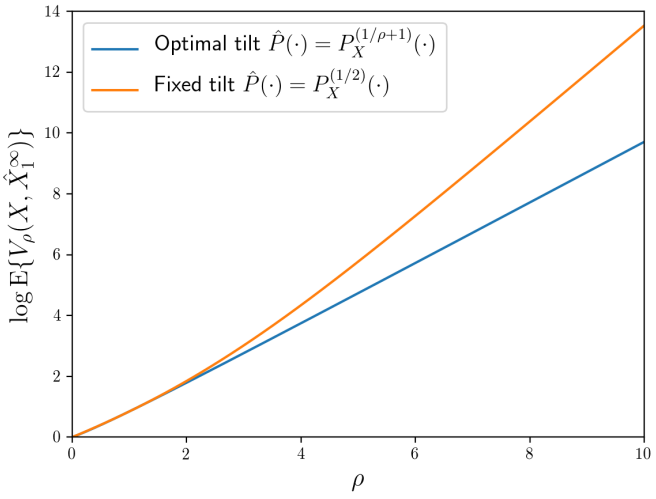

Figure 3-13: This plots compares the performance of the randomized strategy as a function of the moment ρ. We compare the optimal strategy which depends on ρ, against a fixed tilted distribution ($\gamma = 1$ in Corollary 11), when $X \sim \mathrm{Ber}(1/5)$.

Corollary 11. *Fix $\gamma > 0$, and let $\{\hat{X}_k : k \geq 1\}$ be an i.i.d. process generated according to $\hat{P}^*_\gamma(x)$. Then:*

$$\log \mathbb{E}\{V_\rho(X, \hat{X}_1^\infty)\} = \frac{\rho}{1+\gamma} H_{\frac{\gamma-\rho+1}{1+\gamma}}(X) + \frac{\gamma \cdot \rho}{1+\gamma} H_{\frac{1}{1+\gamma}}(X) \tag{3.125}$$

Proof. The proof follows by substituting $\hat{P}(\cdot) = \hat{P}^*_\gamma(\cdot)$ into (3.120). ■

Remark 8 (Zipf's distribution). We emphasize that Lemma 16 is a non-asymptotic result. As such, it can be readily used in the context of passwords generated according to a Zipf's law distribution of parameter s for some $s \geq 0$ (also known as PDF-Zipf model [142]), i.e.,

$$P_X(i) \triangleq \frac{1}{H_{m,s}} \cdot \frac{1}{i^s} \tag{3.126}$$

where $i = 1, \ldots, m$, and $H_{m,s}$ is the *generalized* harmonic number defined as $H_{m,s} = \sum_{j=1}^{m} \frac{1}{j^s}$. As pointed out in the introduction, this family of distribution has been shown in the literature to be useful in modeling password distributions, where the parameter s is dataset dependent. We refer to [142, 140] for more details about the relevance of the Zipf's law in this setting. Under this distribution, applying Lemma 16, we obtain that the optimal i.i.d. guessing strategy is to generate guesses according to a Zipf's law of parameter

$s/(\rho+1)$. Further, we get that

$$\log \mathbb{E}\left\{V_\rho^*(X, \hat{X}_1^\infty)\right\} = (1+\rho)\log H_{m,\frac{s}{1+\rho}} - \log H_{m,s}. \tag{3.127}$$

Note that this is worse than the optimal synchronized strategy which achieves $\log H_{m,(s-\rho)} - \log H_{m,s}$, for $s \geq \rho$, but can perform much better than picking the sub-optimal i.i.d. guessing distribution $P_{\hat{X}} = P_X$, which gives $\log m$. Note that a similar result would hold for the so-called CDF-Zipf's law in [142], i.e., when $P_X(i) = Ci^s - C(i-1)^s$, for some normalizing constant C and parameter $0 \geq s \leq 1$. Namely, it is easy to show that the resulting optimal i.i.d. strategy is then according to the distribution $\hat{P}_\rho^*(i) = C'(i^s - (i-1)^s)^{\frac{1}{1+\rho}}$, where C' is once again a normalizing constant.

Remark 9 (Targeted Attacks). Lemma 16 can also be generalized to the case of availability of some side information Y which is correlated with X. That is, (X, Y) is now a pair of random variables with joint distribution P_{XY}. This models targeted attacks [143] where an adversary makes use of the additional information he possess about an user (e.g. personal information, previously compromised passwords), as modeled by the side-information Y, to make guesses. Note that, as there are various kinds of side-information Y (e.g., sister password, gender), each of which has a different role in impacting password creation, how to systematically employ such side-information Y is subtle. We refer readers to [143] for a more precise treatment of targeted attacks, and the change in performance that results from them. Then, assume that the guesser generates a sequence of guesses $\hat{X}_1, \hat{X}_2, \ldots$ which are i.i.d. *given* Y, and distributed according to $\hat{P}_{X|Y}(\cdot|\cdot)$. As before, we define $G(X, \hat{X}_1^\infty|Y) \triangleq \inf\{k \geq 1 :\ \hat{X}_k(Y) = X\}$. Then, following the proof of Theorem 16 we can show that the optimal guessing distribution is

$$\hat{P}_{X|Y}^*(x|y) = \frac{P_{X|Y}(x|y)^{\frac{1}{1+\rho}}}{\sum_{x' \in \mathcal{X}} P_{X|Y}(x'|y)^{\frac{1}{1+\rho}}} \tag{3.128}$$

for any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, and

$$\log \mathbb{E}\{V_\rho^*(X, \hat{X}_1^\infty|Y)\} = \rho \cdot H_{\frac{1}{1+\rho}}(X|Y), \tag{3.129}$$

where $H_\alpha(X|Y)$ is the conditional Rényi entropy of order α, and $V_\rho^*(X, \hat{X}_1^\infty|Y)$ is defined as in (3.114) but with $G(X, \hat{X}_1^\infty)$ replaced by $G(X, \hat{X}_1^\infty|Y)$. This demonstrates that targeted

attacks can also be performed in a distributed way by employing i.i.d. guesses from the distribution $P_{X|Y}(\cdot|Y)$. Note that this assumes that all distributed agents have access to the same side-information Y. A setting in which this does not hold true, i.e. agents may use different side-information Y_i, is outside the scope of this paper, but was studied in [122]. In particular, [122] compare two mechanisms, one in which the agents do not share their side-information and attempt to breach the system independently, and one in which all the side-information is pooled.

3.4.2 Constraints on the Number of Guesses

In Section 3.4.1, we considered the case in which guesses are made until the correct sequence is found. In this section, we consider the case where adversaries can use only a fixed number of guesses denoted by J. The goal of the adversary is then to maximize her probability of success within this fixed number of queries, both in the synchronized case [14], as well as the asynchronous case. For synchronous guessers, the probability of success associated with the optimal strategy is given by

$$\mathrm{P}_{c,J}^{\mathrm{synchr}} = \sum_{x \in \mathcal{L}} P_X(x),$$

where $\mathcal{L}$ designates the set of the J most likely elements according to P_X. For asynchronous guessers, one strategy consists in generating guesses $\hat{X}$ i.i.d. from a distribution $P_{\hat{X}}$, as was done in the previous section. This setting was precisely studied in [38, Theorem 6], where the optimal guessing distribution $P_{\hat{X}}$ was characterized as a function of the password distribution P_X and of the number of guesses J. Instead, in this work, we focus on the scenario of guessing n-length i.i.d. sequences, and we assume the adversaries make $J = \lceil \mathcal{X}^{n\alpha} \rceil$ total guesses. We analyze the success probability in guessing the correct sequence and derive expressions which are exponentially tight as a function of n. We consider both the synchronized case [14] as well as the asynchronous case.

We start with synchronized guessers, and define the exponential rate of $\mathrm{P}_{c,J}^{\mathrm{synchr}}$ as

$$E_{c,\alpha}^{\mathrm{synchr}} \triangleq \liminf_{n\to\infty} -\frac{1}{n} \log \mathrm{P}_{c,J}^{\mathrm{synchr}} \tag{3.130}$$

$$= \liminf_{n\to\infty} -\frac{1}{n} \log \sum_{\mathbf{x} \in \mathcal{L}} P_{\mathbf{X}}(\mathbf{x}), \tag{3.131}$$

where again $\mathcal{L}$ represents the set of the J most likely elements distributed according this time to the product distribution $P_{\mathbf{X}}$. The following result is an immediate application of the large deviation principle of Guesswork, shown in [48].

Theorem 15 (Theorem 3 in [48]). *For any $\alpha \in [0,1]$,*

$$E_{c,\alpha}^{synchr} = \min_{Q_X \in \mathcal{Q}(\alpha)} D(Q_X \| P_X), \tag{3.132}$$

where $\mathcal{Q}(\alpha)$ is defined as:

$$\mathcal{Q}(\alpha) = \{Q_X : D(Q_X\|P_X) + H(Q_X) < D(Q_X^*\|P_X) + H(Q_X^*)\}, \tag{3.133}$$

with Q_X^* being the solution of the optimization problem:

$$\begin{aligned} \underset{Q_X}{\textit{minimize}} \quad & D(Q_X\|P_X) + H(Q_X) \\ \textit{subject to} \quad & H(Q_X) \geq \alpha \end{aligned} \tag{3.134}$$

In particular, if $\alpha > H(P_X)$, then $E_{c,\alpha}^{synchr} = 0$.

Note that the average number of guesses, roughly $2^{nH_{1/2}(X)}$, is much larger than the required list size that drives $\mathrm{P}_{c,J}^{\mathrm{synchr}}$ to one (exponentially). This great difference comes from the way atypical events are treated in each optimization. In the case of guesswork, an exponential price is payed for atypical events, since the number of queries will be exponential. For probability of error however, the scenario is closer to regular source coding in which the impact of atypical events is sub-exponential, meaning that the optimized quantity will necessarily be related to the typical events. Consider now the asynchronous case, and let $\{\hat{X}_k : k \geq 1\}$ be once again i.i.d. with distribution $P_{\hat{X}}$. In this case the probability of success is defined as

$$\mathrm{P}_{c,J}^{\mathrm{asynchr}} \triangleq \Pr\left\{G(\mathbf{X}, \hat{\mathbf{X}}_1^{\infty}) \leq J\right\}. \tag{3.135}$$

One can verify that

$$\mathrm{P}_{c,J}^{\mathrm{asynchr}} = \sum_{\mathbf{x} \in \mathcal{X}^n} P_{\mathbf{X}}(\mathbf{x}) \left[1 - (1 - P_{\hat{\mathbf{X}}}(\mathbf{x}))^J\right].$$

Finally we define

$$E_{c,\alpha}^{\text{asynchr}} \triangleq \liminf_{n\to\infty} -\frac{1}{n}\log \mathrm{P}_{c,J}^{\text{asynchr}}. \tag{3.136}$$

While, in principle, the distribution $P_{\hat{\mathbf{X}}}$ can be optimized to maximize the probability of success, we will assume that this distribution is simply given by the tilted distribution of $P_{\mathbf{X}}$, namely, for some $\beta \geq 0$, and any $\mathbf{x} \in \mathcal{X}^n$,

$$P_{\hat{X}}^{(\beta)}(x) \triangleq \frac{P_X(x)^\beta}{\sum_{x\in\mathcal{X}} P_X(x)^\beta}. \tag{3.137}$$

We motivate this choice by the results of the previous sub-section, which showed that these tilted distributions were optimal in terms of the number of guesses. We have the following result.

Theorem 16. *For any $\alpha, \beta \geq 0$,*

$$E_{c,\alpha}^{asynchr}(\beta) = \min_{Q_X\in\mathcal{Q}(\alpha)} \left\{ D(Q_X||P_X) + \left[D(Q_X||P_{\hat{X}}^{(\beta)}) + H(Q_X) - \alpha \right]_+ \right\}, \tag{3.138}$$

where $[x]_+ \triangleq \max\{x, 0\}$.

Using Theorem 16, we obtain the following immediate result.

Corollary 12.

$$\min_{\beta\geq 0} E_{c,\alpha}^{asynchr} = \min_{Q_X\in\mathcal{Q}(\alpha)} D(Q_X\|P_X) \tag{3.139}$$

$$= E_{c,\alpha}^{synchr}. \tag{3.140}$$

Corollary 12 essentially proves that the tilted family is asymptotically optimal, and that there exist a unique optimal tilt β for each size list $J = \lceil \mathcal{X}^{n\alpha} \rceil$. It follows from this that even though the optimization (3.132) is over a set of distributions $\mathcal{Q}(\alpha)$, the solution is always a tilted distribution $P_X^{(\beta)}$ for some $\beta \geq 0$ which depends on α.

Proof of Corollary 12. By definition, $\min_{\beta\geq 0} E_{c,\alpha}^{\text{asynchr}} \geq 0$. Then, for $\alpha \geq H(P_X)$, we see from Theorem 16 that by taking $Q_X = P_X$ and $\beta = 1$, we have

$$\min_{\beta\geq 0} E_{c,\alpha}^{\text{asynchr}} \leq [H(P_X) - \alpha]_+ = 0. \tag{3.141}$$

For $\alpha < H(P_X)$, we first note that by definition $\min_{\beta\geq 0} E_{c,\alpha}^{\text{asynchr}} \geq E_{c,\alpha}^{\text{synchr}}$. Hence, due to Theorem 15 and Lemma 22 in the appendix we may conclude that

$$\min_{\beta\geq 0} E_{c,\alpha}^{\text{asynchr}} \geq D(Q_X^*\|P_X), \tag{3.142}$$

where Q_X^* is the solution of the optimization

$$\begin{aligned} \underset{Q_X}{\text{minimize}} \quad & D(Q_X\|P_X) + H(Q_X) \\ \text{subject to} \quad & H(Q_X) \geq \alpha. \end{aligned} \tag{3.143}$$

On the other hand, by taking $Q_X = Q_X^*$, we have

$$\min_{\beta\geq 0} E_{c,\alpha}^{\text{asynchr}} \leq D(Q_X^*\|P_X) + \min_{\beta\geq 0} \left[D(Q_X^*||P_{\hat{X}}^{(\beta)})\right]_+.$$

It is a simple exercise to verify that Q_X^* is a tilted distribution, *i.e.* there exist a $\tilde{\beta}$ such that $Q^*(x) = \frac{Q_X(x)^{\tilde{\beta}}}{\sum_{x'} Q_X(x')^{\tilde{\beta}}}$. Letting $\beta = \tilde{\beta}$ gives

$$\min_{\beta\geq 0} E_{c,\alpha}^{\text{asynchr}} \leq D(Q_X^*||P_X). \tag{3.144}$$

The result follows from combining (3.142) and (3.144). ■

We next provide the proofs of Theorems 15 and 16.

Proof of Theorem 16. For simplicity of presentation, we prove the theorem for binary sequences, i.e. $\mathcal{X} = \{0,1\}$, and assume that $1/2 \geq p \triangleq P_X(0)$. For any given sequence $x^n \in \mathcal{X}^n$,

$$\frac{1}{n}\log \hat{P}_{X^n}(x^n) = -D(\hat{P}_{\mathbf{x}_n}||\bar{p}^{\beta}) - H(\hat{P}_{\mathbf{x}_n}) \tag{3.145}$$

where $\hat{P}_{\mathbf{x}_n}$ is the empirical measure of a given sequence x^n, and $\bar{p}^{\beta} = \frac{p^{\beta}}{p^{\beta}+(1-p)^{\beta}}$. Then,

$$\begin{aligned} \mathrm{P}_{c,J}^{\text{asynchr}} &= \sum_{x^n\in\mathcal{X}^n} P_{X^n}(x^n)\left[1-(1-\hat{P}_{\mathbf{x}_n})^J\right] \\ &= \sum_{x^n\in\mathcal{X}^n} 2^{-n\left(D(\hat{P}_{\mathbf{x}_n}||p)+H(\hat{P}_{\mathbf{x}_n})\right)} \\ &\qquad \times \left[1-(1-2^{-n(D(\hat{P}_{\mathbf{x}_n}||\bar{p}^{\beta})+H(\hat{P}_{\mathbf{x}_n}))})^J\right]. \end{aligned}$$

Letting $\mathcal{Q}_n$ denote the set of possible types, i.e. $\mathcal{Q}_n \triangleq \{0, 1/n, 2/n, \ldots, n/n\}$ we obtain,

$$\begin{aligned}
\mathrm{P}_{c,J}^{\text{asynchr}} &= \sum_{q \in \mathcal{Q}_{n,n}} |T(q)|\, 2^{-n(D(q||p)+H(q))} \\
&\qquad \times \left[1 - (1 - 2^{-n(D(q||\bar{p}^{\beta})+H(q))})^{J}\right] \\
&\doteq \sum_{q \in \mathcal{Q}_{n,n}} 2^{nH(q)} 2^{-n(D(q||p)+H(q))} 2^{-n\left[D(q||\bar{p}^{\beta})+H(q)-\alpha\right]_{+}} \\
&\doteq \max_{q \in [0,1]} 2^{-n\left[D(q||p)+\left[D(q||\bar{p}^{\beta})+H(q)-\alpha\right]_{+}\right]}
\end{aligned}$$

where the fourth equation follows from the fact that (see, e.g., [133, Lemma 1]) if $a \in [0,1]$, then $\frac{1}{2}\min\{1, aM\} \leq 1 - (1-a)^M \leq \min\{1, aM\}$. Thus, we have shown that

$$E_{c,\alpha}^{\text{asynchr}} = \min_{q \in [0,1]} \left\{ D(q||p) + \left[D(q||\bar{p}^{\beta}) + H(q) - \alpha\right]_{+} \right\}.$$

■

Together, Lemma 16 and Corollary 12 imply that i.i.d. guesses can perform optimally, both in terms of the expected number of guesses, and in terms of the probability of success. Note that, analogous to Lemma 16, the optimal distribution in Corollary 12 depends on the parameter α. As a result, asynchronous guessers can perform brute-force attacks as efficiently as synchronized guessers asymptotically, at the expense of universality. Finally, it should be emphasized that the optimality of the tilted distribution is a by-product of the asymptotic treatment. Indeed, the results of [38] show that the optimal distribution in the non-asymptotic regime is not a tilted distribution of P_X, but rather a more involved functional of the password distribution. As such, our result does not follow from [38] in a straightforward way.

Remark 10 (Probability of failure). The above results characterized the probability of success of an adversary. In particular we demonstrated that a list size J which is large enough (i.e., such that $\alpha > H(P_X)$) will have an exponent of success probability equal to 1, both under asynchronous and synchronous attacks. Note that this result can be strengthened by looking at the complementary probability of failure $P_{f,J}^{\text{synchr}}$ and $P_{f,J}^{\text{asynchr}}$. Again, in the i.i.d. setting, using essentially the same tools as for the probability of success, one can show that the exponents of the probability of failure for both synchronous and asynchronous attacks

are the same, equal to 1 when $\alpha < H(P_X)$, and decreasing as α grows. Similarly, the optimal guessing distribution for asynchronous guessers is a tilted distribution, where the tilt depends on the size of the list.

Remark 11 (J-Guesswork). We briefly mention J-Guesswork, a related notion of computational security which was introduced in [33] (denoted α-Guesswork). While the usual Guesswork captures the average number of guesses necessary for a system breach, the average J-Guesswork, denoted by $\mathbb{E}[G_J(X)]$, captures the average number of guesses for an adversary which performs no-more than J queries, where J is picked to guarantee a certain probability of success. As such, when $J = \mathcal{X}^n$, the J-Guesswork reduces to $\mathbb{E}[G(X)]$. We can rewrite the average J-Guesswork, as a sum of two terms, i.e.

$$J \times \mathbb{P}(G(X) > J) + \sum_{i=1}^{J} i \cdot P_X(i), \tag{3.146}$$

where the first term corresponds to the case where the attacker is unsuccessful and stops at J guesses, and the second terms captures his average number of guesses otherwise. In the asymptotic regime where we look at passwords generated from the product distribution $P_{\mathbf{X}}$, and letting $J = \lceil |\mathcal{X}|^{n\alpha} \rceil$, for $\alpha > H(P_X)$, it follows from the remark above that the probability $\mathbb{P}(G(X) > J)$ goes to zero with an exponent $D(P_X^{(\beta)} \| P_X)$ for some unique $\beta \geq 0$, as long as J is large enough (i.e. $\alpha > H(P_X)$). It is then easy to prove that, when $\alpha > H(P_X)$, the average J-Guesswork takes exponent

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}[G_\beta(\mathbf{X})] = \max\{\alpha - D(P_X^{(\beta)} \| P_X), H_{1/2}(P_X)\}. \tag{3.147}$$

When J is too small, i.e. when $\alpha < H(P_X)$, then with high probability $G(X) > J$, and therefore the exponent is dominated by J itself, that is $\lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}[G_\beta(\mathbf{X})] = \alpha$. Note that these result hold true in an asynchronous setting as well. Indeed, picking guesses i.i.d. from a distribution $P_{\hat{\mathbf{X}}}$ such that it is equal to the tilted distribution which achieves the maximum in (3.147) gives the same exponent of J-Guesswork. Therefore, i.i.d. guesses perform asymptotically optimally with respect to J-Guesswork as well.

Chapter 4

Conclusion

The privacy and security issues that arise when sharing data reveal how precious personal data is. In this thesis, we looked at two problems which arise when releasing such data: privacy against inference, and brute-force security. The nature of these problems does not suit itself to be solved via the traditional tools and techniques from cryptographic security. The classical computer secrecy model assumes that the sensitive data is known – but this model turns out to be unsatisfactory when the security threat associated with data that appears to be non-sensitive is unclear. While it is hard to envision a future where no personal data will be collected, it is already critical to be able to explore the trade-off between the utility of the data, and the potential threat. To explore this trade-off and handle these modern problems, it is essential to lay a strong and robust theoretical foundation. This thesis provides some outlines for this foundation, using tools from Information Theory, Statistical Learning, and Cryptography.

On the privacy end, we considered a privacy-utility trade-off encountered by users who wish to disclose some information to an analyst, that is correlated with their private data, in the hope of receiving some utility. We proposed a general framework under which data is transformed according to a probabilistic privacy-preserving mapping before it is disclosed. Applying this general framework to the setting where the adversary uses the log-loss cost function naturally led to a non-asymptotic information-theoretic formulation for characterizing the best achievable privacy subject to utility constraints. We justified the relevance and generality of the privacy metric under the log-loss by proving that the inference threat under any bounded cost function can be upper bounded by an explicit function of the mutual

information between private data and disclosed data. In addition, we showed that when the log-loss is used in this framework in both the privacy metric and the distortion metric, the average information leakage and the utility constraint can be reduced to the mutual information between private data and disclosed data, and between non-private data and disclosed data, respectively. We then showed that the privacy-utility tradeoff under the log-loss can be cast as the non-convex Privacy Funnel optimization, and we leverage its connection to the Information Bottleneck and Mrs Gerber's Lemma, to provide a greedy algorithm for solving it.

On the security end, we studied brute-force attacks in which an adversary aims at breaching a password secured system by querying tentative passwords until the correct one is found. There, we used Guesswork as a surrogate for the computational effort than an adversary has to commit before breaking such security. We looked at three brute-force attack settings, and discussed the impact of the distribution of the password (how predictable the password is), and where relevant, the amount of side-information that an adversary has (how targeted the attack is). Our studies reveal some surprising facts about such attacks. First, the impact of side-information is tremendous, even when this side-information is distributed. Next, attacks are perhaps more surprising to mismatched knowledge of adversaries than what would be expected. Finally, asynchronous attacks perform asymptotically as good as fully synchronized ones. Together, these highlight the danger of brute-force attacks. Additionally, we introduced a series of proving techniques inspired from the geometry of guesswork which shed a different light on this seminal problem.

While this is the conclusion of this thesis, there are several future directions which are worth pointing out and pursuing. In the next section, we will discuss and motivate the quest for better data representations as a main direction which could have both theoretical and practical impact in this area.

4.1 Parsimonious Data Representations: The Road Ahead

From the perspective of the users, it is in their favor to release as little data as possible, while still obtaining quality services. However, the trend is exactly the opposite, more and more personal data is being collected and processed. While data collecting services tend to benefit from this trend, they also face challenges as a consequence of this evolution. For

example, in 2004, Facebook was operated from a single server in a dorm room at Harvard University – now, the company owns massive data-centers across the globe, each housing tens of thousands of computers, connected via an intricate network of links. Each interaction with the website launches a chain of reactions over these data-centers, eventually leading to the collection, storage, and treatment of data. With 1.59B daily users, the data created is staggering, and the management of the communication overhead is complex. Interestingly, both privacy-conscious users, and the corporations collecting the data share a common desire – collect as little data as possible, while still being able to provide/obtain the desired service. This quest for parsimonious data representations spans all the steps in the data acquisition and communication pipeline, from acquisition of physical signals, to communication and processing of the data. A simple, yet important observation is that, while data is important, data is not the end-goal itself. Thus, if there is knowledge of the purpose of the data, one should aim at having a task-specific representation. Some example of questions which are worth studying in the future are presented below. In all these problems, adapting the data representation to the task is the main leverage to improve current solutions, which is in contrast with the traditional task-agnostic approaches.

- **Signal Processing for Tasks:** Consider a sensor in an autonomous car, whose goal is to identify traffic signs. When designing such sensors, an engineer might be concerned with, the frame rate at which this camera should operates, the quality at which to capture images, or the appropriate digital representation of said images, and many more parameters of concern. Each of these parameters may impact how well and how reliably traffic signs can be recognized by the vehicle. On the other hand, the engineer must take into account energy consumption, available hardware, and implementation burden. Thus, there is an inherent trade-off between the physical constraints that a system must satisfy, and how well it can perform the task. In a series of preliminary works, we have studied how one can make use of the knowledge of the task (identify traffic signs), to better optimize the number of quantization bits [125]. These results show great promise over task-agnostic solutions, and are a first step towards Signal-Processing for Tasks, where representations of signals adapt to the tasks.

- **Communication for Tasks:** Consider a healthcare monitoring application where a set of distributed sensors are capturing biomedical signals. The values from these

sensors are communicated to a central unit which pools the data to compute some health indicators, i.e. a function of the sensors value. A few questions of interest are the following: How do we effectively make use of the correlation between the sensors (e.g. one sensor capturing heart-rate and the other blood-pressure) to reduce the communication overhead? Is it better to have many low quality sensors, or few high quality ones? How robust is the computation to failures of some sensors? Importantly, how do we do this over wireless noisy communication links, or via an entire network of links? In some earlier works, we proposed to use a specific structure in the correlation between some sources to drastically reduce the communication overhead using efficient codes [124, 123]. In some more recent works [92], we explored how data-driven methods such as Neural Networks can be used to construct non-linear network codes that are robust to noise, which permit correlated sources to be communicated efficiently. The latter techniques also generalize to the case where the specific function is known, and thus reduce the communication overhead further.

- **Learning Data Representations:** So far, we have assumed that the task is known and can be expressed as a function. But many setups actually require to learn from the data itself. This is the standard Machine Learning process, where data is fed to a learning algorithm for the purpose of classification or regression. We are faced with a predicament. How can we represent the data parsimoniously, without knowing which parts of the data are useful beforehand? To address this apparent paradox, I suggest to explore how methods from unsupervised representation learning can be leveraged locally, to represent the data in a compact way before communication. One such context that we have explored is the following: consider two cameras capturing the same object, but from different angles, light conditions, and distance. In [83], we asked how to find representations of the data from each camera such that they capture what is common in both scenes (i.e. the object), and not what is superfluous (i.e. angle, light condition, etc.). For this purpose, we designed an entire framework which spans both practice and theory. We formalized the problem above in the language of Principal Inertia Components (PICs) [40], a mathematical tool which has a long history in statistical sciences and information theory. The PICs and the corresponding principal functions provide a fine-tuned decomposition of a probability distribution between two random variables in terms of maximally correlated embeddings. Then, we designed

a data-driven approach based on neural networks to find these representations from data. This approach turns out the be very versatile and provides methods to deal with various problems in machine learning, such as data visualization using correspondence analysis, comparison of black-box models, multi-modal learning, and more.

Appendix A

Proofs of Theorem 3 and 4

The following lemma [53], which bounds the difference in the entropies of two distributions, will be useful in the proof of the Theorems.

Lemma 17 ([53, Thm 17.3.3])**.** *Let P and Q be distributions with the same support $\mathcal{X}$ such that $\|P-Q\|_1 \leq \frac{1}{2}$. Then:*

$$|H(P) - H(Q)| \leq \|P-Q\|_1 \log \frac{|\mathcal{X}|}{\|P-Q\|_1}.$$

Proof of Theorem 3: The first inequality can be proved in four steps. Initially, we note that the objective function can be rewritten as

$$J(P_{S,X}, P_{Y|X}) = H(P_S) + H(P_Y) - H(P_{S,Y}). \tag{A.1}$$

Therefore, the difference between the objective functions with respect to $P_{S,X}$ and $Q_{S,X}$ is bounded as:

$$\begin{aligned}\left|J(P_{S,X}, P_{Y|X}) - J(Q_{S,X}, P_{Y|X})\right| \leq & |H(P_S) - H(Q_S)| + \\ & |H(P_Y) - H(Q_Y)| \ + \ |H(P_{S,Y}) - H(Q_{S,Y})|. \end{aligned} \tag{A.2}$$

The bound in Lemma 17 can be used to bound each of the terms in Equation (A.2). For instance:

$$\|P_{S,Y} - Q_{S,Y}\|_1 \ = \ \sum_{s,y} \left| \sum_b P(y|x)[P(s,x) - Q(s,x)] \right|$$

$$
\begin{aligned}
&\leq \sum_{s,x,y} P(y|x)\,|P(s,x) - Q(s,x)| \\
&= \sum_{s,x} \underbrace{\sum_{y} P(y|x)}_{1}\,|P(s,x) - Q(s,x)| \\
&= \|P_{S,X} - Q_{S,X}\|_1
\end{aligned} \tag{A.3}
$$

and therefore:

$$
|H(P_{S,Y}) - H(Q_{S,Y})| \leq \|P_{S,X} - Q_{S,X}\|_1 \log \frac{|\mathcal{S}|\,|\mathcal{X}|}{\|P_{S,X} - Q_{S,X}\|_1}. \tag{A.4}
$$

Similarly, it can be shown that:

$$
|H(P_S) - H(Q_S)| \leq \|P_{S,X} - Q_{S,X}\|_1 \log \frac{|\mathcal{S}|}{\|P_{S,X} - Q_{S,X}\|_1} \tag{A.5}
$$

$$
|H(P_Y) - H(Q_Y)| \leq \|P_{S,X} - Q_{S,X}\|_1 \log \frac{|\mathcal{X}|}{\|P_{S,X} - Q_{S,X}\|_1}. \tag{A.6}
$$

Finally, the three upper bounds can be substituted into Equation (A.2), which yields:

$$
\left|J(P_{S,X}, P_{Y|X}) - J(Q_{S,X}, P_{Y|X})\right| \leq 3\,\|P_{S,X} - Q_{S,X}\|_1 \log \frac{|\mathcal{S}|\,|\mathcal{X}|}{\|P_{S,X} - Q_{S,X}\|_1}. \tag{A.7}
$$

Our first claim is proved by substituting $P^*_{Y|X}$ for $P_{Y|X}$ in the above equation.

The proof of our second claim is based on the inequality:

$$
\begin{aligned}
&\left|\mathbb{E}_{P_{Y,X}}[d(Y,X)] - \mathbb{E}_{Q_{Y,X}}[d(Y,X)]\right| \\
&\quad= \left|\sum_{s,x,y} P(y|x)[P(s,x) - Q(s,x)]d(x,y)\right| \\
&\quad\leq \sum_{s,x,y} P(y|x)d(x,y)\,|P(s,x) - Q(s,x)| \\
&\quad\leq d_{\max} \sum_{s,x} \underbrace{\sum_{y} P(y|x)}_{1}\,|P(s,x) - Q(s,x)| \\
&\quad= d_{\max}\,\|P_{S,X} - Q_{S,X}\|_1\,.
\end{aligned} \tag{A.8}
$$

Based on this observation, it follows that:

$$\begin{aligned}\mathbb{E}_{P_{Y,X}}[d(Y,X)] &\leq \mathbb{E}_{Q_{Y,X}}[d(Y,X)] + \\ &\quad d_{\max} \left\| P_{S,X} - Q_{S,X} \right\|_1 \\ &\leq \Delta + d_{\max} \left\| P_{S,X} - Q_{S,X} \right\|_1 . \end{aligned} \tag{A.9}$$

The last step is due to the constraint $\mathbb{E}_{Q_{Y,X}}[d(Y,X)] \leq \Delta$ that is enforced in our problem (2.4). ■

We now move onto the next proof. First, let us introduce some useful notation. Consider the optimization problem 2.4, and denote by $R(P_{S,X}, \Delta)$ the optimal privacy leakage for input $P_{S,X}$ and distortion constraint Δ. We also denote by $\mathcal{S}(\Delta)$ the set of feasible mappings, *i.e.*, $\mathcal{S}(\Delta) = \left\{ P_{Y|X} : \mathbb{E}_{X,Y}[d(X,Y)] \leq \Delta \right\}$. The following lemma is useful in the proof of Thm. 4, and allows us to construct distributions that are close in a $\mathcal{L}_1$ sense but have specific expected distortions.

Lemma 18. *Let Q be a distribution over $\mathcal{X}$ such that $\mathbb{E}_Q[f] = \delta$, with f a non-negative function. For any $\delta > 0$, there exist a distribution P over the same support, such that $\mathbb{E}_p[f] = 0$ and $\|Q - P\|_1 \leq \frac{2\delta}{f_{min}}$, where $f_{min} = \min_{x, f(x)>0} f(x)$ is the smallest non-zero value of f.*

Proof: We do the proof by construction. Consider P such that for all $x \in \mathcal{X}$ with $f(x) > 0$, let $P(x) = 0$. For all other $x \in \mathcal{X}$, set $P(x) = Q(x) + \frac{\sum_{x \in \mathcal{X}, f(x)>0} Q(x)}{|x \in \mathcal{X} : d(x) > 0|}$, where the second term corresponds to adding uniformly the missing mass so that $\sum_x P(x) = 1$. We have:

$$\|P - Q\|_1 \leq \sum_{x \in \mathcal{X}, f(x)>0} |P(x) - Q(x)| \tag{A.10}$$

$$+ \sum_{x \in \mathcal{X}, f(x)=0} |P(x) - Q(x)| \tag{A.11}$$

$$= 2 \sum_{x \in \mathcal{X}, f(x)>0} Q(x) \tag{A.12}$$

Next, we have that:

$$\delta = \mathbb{E}_Q[f] = \sum_{x \in \mathcal{X}} f(x) Q(x) \tag{A.13}$$

$$\geq f_{\min} \sum_{x \in \mathcal{X}, f(x)>0} Q(x) \tag{A.14}$$

$$\geq \frac{f_{\min}}{2} \|P - Q\|_1 \tag{A.15}$$

where (A.15) folows from (A.12). Noticing that $\mathbb{E}_p[f] = 0$ gives the desired result. ■

Proof of Theorem 4: Recall that we denote by $R(P_{S,X}, \Delta)$ the result of the optimization problem (2.4) with input $P_{S,X}$ and distortion constraint Δ, and that we use $\mathcal{S}(\Delta)$ to denote the feasible region of this optimization problem. We use $\epsilon = \|P - Q\|_1$. Our goal is to bound $|R(P_{S,X}, \Delta) - R(Q_{S,X}, \Delta)|$. We have:

$$R(P_{S,X}, \Delta + \epsilon d_{\max}) \leq J(P_{S,X}, Q^*_{Y|X}) \tag{A.16}$$

$$\leq J(Q_{S,X}, Q^*_{Y|X}) + |J(P_{S,X}, Q^*_{Y|X}) - J(Q_{S,X}, Q^*_{Y|X})|$$

$$= R(Q_{S,X}, \Delta) + |J(P_{S,X}, Q^*_{Y|X}) - J(Q_{S,X}, Q^*_{Y|X})| \tag{A.17}$$

where (A.16) follows from the distortion inequality of Thm. 3 which means that $Q^*_{Y|X}$ is in the feasible set $\mathcal{S}(\Delta + \epsilon d_{\max})$. Adding $R(P_{S,X}, \Delta)$ on both sides of (A.17), and rearranging terms, we obtain:

$$\begin{aligned} &R(P_{S,X}, \Delta) - R(Q_{S,X}, \Delta) \\ &\leq |J(P_{S,X}, Q^*_{Y|X}) - J(Q_{S,X}, Q^*_{Y|X})| \\ &\quad + R(P_{S,X}, \Delta) - R(P_{S,X}, \Delta + \epsilon d_{\max}) \end{aligned} \tag{A.18}$$

Notice that the first term of (A.18) can be bounded using Thm. 3. The second term corresponds to the difference in the solution of the optimization problem when we have expanded the feasible set by allowing an additional distortion $\epsilon d_{\max}$. We have the following cases:

- $P^*_{Y|X}$ was not on the border of the feasible set $\mathcal{S}(\Delta)$. Then, as the problem is convex, $P^*_{Y|X}$ is also a minimizing distribution of the optimization problem with expanded feasible set $\mathcal{S}(\Delta + \epsilon d_{\max})$. Therefore, $R(P_{S,X}, \Delta) - R(P_{S,X}, \Delta + \epsilon d_{\max}) = 0$.

- $P^*_{Y|X}$ is on the border of the feasible set $\mathcal{S}(\Delta)$. First, notice that $R(P, \Delta)$ is convex in Δ. This can be seen as $\mathbb{E}_{P_{Y,X}}[d(Y, X)]$ is linear and that the mutual information $J(P_{S,X}, P_{Y|X})$ is convex in $P_{Y|X}$. Therefore, if we let Δ_1 and Δ_2 be two distortion

value, and let P_1^* and P_2^* be the respective minimizing distributions, then it is the case that for $P_\alpha = \alpha P_1^* + (1-\alpha)P_2^*$, with $0 \leq \alpha \leq 1$, we have:

$$R(P_\alpha, \Delta) \leq J(P_{S,X}, P_\alpha) \tag{A.19}$$

$$\leq \alpha J(P_{S,X}, P_1^*) + (1-\alpha) J(P_{S,X}, P_2^*) \tag{A.20}$$

$$= \alpha R(P_{S,X}, \Delta_1) + (1-\alpha) R(P_{S,X}, \Delta_2) \tag{A.21}$$

As the function $R(P, \Delta)$ is convex and non-increasing with respect to Δ, its steepest descent is at zero, that is :

$$R(P_{S,X}, \Delta) - R(P_{S,X}, \Delta + \epsilon d_{\max})$$

$$\leq R(P_{S,X}, 0) - R(P_{S,X}, \epsilon d_{\max}) \tag{A.22}$$

Then, by Lemma 18 with $f = d(Y, X)$, and $\delta = \epsilon d_{\max}$, there is a $\tilde{P}_{Y|X} \in \mathcal{S}(0)$, such that the distance between $\tilde{P}_{Y|X}$ and the minimizing distribution of the optimization problem with expanded feasible set $\mathcal{S}(\epsilon d_{\max})$ is at most $\epsilon \frac{2d_{\max}}{d_{\min}}$. If $\epsilon \leq \frac{d_{\min}}{4d_{\max}}$, we can use Lemma 17 and equations similar to those in (A.3) to obtain:

$$R(P_{S,X}, \Delta) - R(P_{S,X}, \Delta + \epsilon d_{\max})$$

$$\leq 4\epsilon \frac{d_{\max}}{d_{\min}} \log \frac{d_{\min}|\mathcal{S}||\mathcal{X}|}{\epsilon d_{\max}} \tag{A.23}$$

$$\leq 4\epsilon \frac{d_{\max}}{d_{\min}} \log \frac{|\mathcal{S}||\mathcal{X}|}{\epsilon} \tag{A.24}$$

Using (A.24) in (A.18) gives the desired bound.■

Appendix B

Additional Lemmas on Guesswork

The following lemma relate the position of a sequence $\mathbf{x}_n$ in the optimal list, with the type of that sequence.

Lemma 19. *Let $\mathbf{x}_n$ be a i.i.d. generated sequence ,and consider the position of $\mathbf{x}_n$ in the optimal list according to P_X,* i.e. *$G^*(\mathbf{x})$. For a given α, we have that $G^*(\mathbf{x}) < \lceil |\mathcal{X}|^\alpha \rceil$ if and only if the sequence $\mathbf{x}$ satisfy $\hat{P}_{\mathbf{x}} \in \mathcal{Q}(\alpha)$, where*

$$\mathcal{Q}(\alpha) = \{Q_X : D(Q_X \| P_X) + H(Q_X) < D(Q_X^* \| P_X) + H(Q_X^*)\}, \tag{B.1}$$

with Q_X^* being the solution of the optimization problem:

$$\begin{array}{ll} \underset{Q_X}{\textit{minimize}} & D(Q_X \| P_X) + H(Q_X) \\ \textit{subject to} & H(Q_X) \geq \alpha \end{array}. \tag{B.2}$$

Proof. Recall that $P_X(\mathbf{x}) = \exp\{-n\left(D(\hat{P}_{\mathbf{x}} \| P_X) + H(\hat{P}_{\mathbf{x}})\right)\}$, and that the size of the type set $T(\hat{P}_{\mathbf{x}}) \doteq 2^{nH(\hat{P}_{\mathbf{x}})}$. Let $\mathcal{Q}(\alpha)$ be the set of types of the sequences that are in the first $\mathcal{X}^{n\alpha}$ position in the list optimal list. Then, by definition of $\mathcal{Q}(\alpha)$:

$$\sum_{Q_X \in \mathcal{Q}(\alpha)} 2^{nH(Q_X)} \doteq 2^{n\alpha} \tag{B.3}$$

An application of the method of types gives that the left-hand side evaluates to $2^{n \sup_{Q_X \in \mathcal{Q}(\alpha)} H(Q_X)}$, meaning that $\sup_{Q_X \in \mathcal{Q}(\alpha)} H(Q_X) = \alpha$. Thus, the threshold probability is given by the type that solves (B.2), and any type that has lower probability must appears before in the list. ∎

The following lemma characterizes the guesswork exponent of a sequence generated by the concatenation of a uniform binary sequence, and an arbitrary *i.i.d.* sequence.

Lemma 20. *Let $U \sim \text{Bern}(1/2)$ and $V \sim \text{Bern}(p)$, with $p \leq 1/2$, and denote by U^{m_n} and V^{n-m_n} their* i.i.d. *sequences, for some sequence m_n such that $\lim_{n\to\infty} \frac{m_n}{n} = \lambda$. Then, the guesswork exponent for sequence $X^n = (U^{m_n}, V^{n-m_n})$ is:*

$$\lim_{n\to\infty} \log \mathbb{E}\left[G(\mathbf{X})^\rho\right] = \lambda\rho + (1-\lambda)\rho H_{1/1+\rho}(p). \tag{B.4}$$

Proof. We do the proof for $\rho = 1$, general case follows trivially. It is easy to verify that the optimal list is constructed by first ordering the subsequence $\mathbf{v}_{n-m_n}$ by most likely to least likely, and then concatenating to each such subsequence all the possible $\mathbf{u}_{m_n}$, in an arbitrary order. To reach a given $\mathbf{x}_n = (v^{n-m_n}, u^{m_n})$, it is necessary to reach the subsequences v^{n-m_n}, and we have:

$$\begin{aligned}
\mathbb{E}\left[G(\mathbf{X}_n)\right] &= \mathbb{E}\left[\mathbb{E}\left[G(\mathbf{X}_n)|\mathbf{V}_{n-m_n}\right]\right] \\
&\doteq \sum_{\mathbf{v}_{n-m_n}} \exp\left\{-(n-m_n)\left[D(\hat{P}_{\mathbf{v}}||P_V) + H(\hat{P}_{\mathbf{v}})\right]\right\} \\
&\qquad \times \exp\left\{(n-m_n)H(\hat{P}_{\mathbf{v}})\right\}\exp\{m_n\} \\
&\doteq \sum_{\hat{P}_V} \exp\left\{(n-m_n)\left[H(\hat{P}_V) - D(\hat{P}_V||P_V)\right] + m_n\right\} \\
&\doteq \exp\left\{n \sup_{\hat{P}_V}(1-\lambda)\left[H(\hat{P}_V) - D(\hat{P}_V||P_V)\right] + \lambda\right\}.
\end{aligned}$$

Solving the optimization yields the desired result. ■

The next lemma compares the guesswork of a random variable which takes values in a discrete alphabet uniformly at random, with a random variables for which one of the symbol has been *softly* removed. Precisely, we have

Lemma 21 (Soft Elimination)**.** *Consider a random variable U_N taking values uniformly in $[N]$, and U. For some $0 \leq s < 1$, we call a K soft-elimination, a random variable $V_{(N,K)}$ such that:*

$$Pr(V_{(N,K)} = i) = \begin{cases} \frac{1}{N-1} & \text{if } 1 \leq i \leq N-K \\ \frac{K-1}{K(N-1)} & \text{if } N-K \leq i \leq N \end{cases}. \tag{B.5}$$

Then, for any $\alpha > 0$, $\mathbb{E}[G(U_N)^\alpha] > \mathbb{E}[G(V_{(N,K)})^\alpha] \geq \mathbb{E}[G(U_{N-1})^\alpha]$.

Proof. We have :

$$\begin{aligned}
&\mathbb{E}[G(V_{(N,K)})] - \mathbb{E}[G(U_{N-1})] = \\
&\sum_{i=1}^{N-K} i^\alpha \left(\frac{1}{N-1} - \frac{1}{N-1} \right) + \\
&\sum_{i=N-K+1}^{N-1} i^\alpha \left(\frac{K-1}{K(N-1)} - \frac{1}{N-1} \right) + N^\alpha \frac{K-1}{K(N-K)}.
\end{aligned} \tag{B.6}$$

By evaluating the series and combining terms it is easy to verify that the right hand side is non-negative. ■

The following two lemmas relate the position of a sequence $\mathbf{x}$ in the optimal list, i.e. $G^*(\mathbf{x})$, with the type $\hat{P}_\mathbf{x}$ of that sequence, first without side-information, and then with side-information.

Lemma 22. *Let* $\mathbf{x}$ *be a i.i.d. generated sequence of length* n*,and consider the position of* $\mathbf{x}$ *in the optimal list according to* P_X*,* i.e. $G^*(\mathbf{x})$*. For a given* α*, we have that* $G^*(\mathbf{x}) < \lceil |\mathcal{X}|^\alpha \rceil$ *if and only if the sequence* $\mathbf{x}$ *satisfy* $\hat{P}_\mathbf{x} \in \mathcal{Q}(\alpha)$*, where*

$$\begin{aligned}
\mathcal{Q}(\alpha) = \{ Q_X : &D(Q_X \| P_X) + H(Q_X) \\
&< D(Q_X^* \| P_X) + H(Q_X^*) \},
\end{aligned} \tag{B.7}$$

with Q_X^* *being the solution of the optimization problem:*

$$\begin{aligned}
&\underset{Q_X}{\text{minimize}} \quad D(Q_X \| P_X) + H(Q_X) \\
&\text{subject to} \quad H(Q_X) \geq \alpha
\end{aligned} \tag{B.8}$$

Lemma 23. *Let* $(\mathbf{x}_n, \mathbf{y}_n)$ *be a pair of binary sequences ,and consider the position of* $\mathbf{x}$ *in the optimal list according to* $P_{X|Y}$*,* i.e. $G^*(\mathbf{x}|\mathbf{y})$*. For a given* α*, we have that* $G^*(\mathbf{x}|\mathbf{y}) < \lceil |\mathcal{X}|^\alpha \rceil$ *if and only if the sequence* $(\mathbf{x}, \mathbf{y})$ *satisfy* $\hat{P}_{\mathbf{x}|\mathbf{y}} \in \mathcal{Q}(\alpha, \hat{P}_\mathbf{y})$*, where*

$$\begin{aligned}
\mathcal{Q}(\alpha, \hat{P}_\mathbf{y}) = \Big\{ Q_{X|Y} : &D(Q_{X|Y} \| P_{X|Y} | \hat{P}_\mathbf{y}) + H(Q_{X|Y} | \hat{P}_\mathbf{y}) \\
&< D(Q_{X|Y}^* \| P_{X|Y} | \hat{P}_\mathbf{y}) + H(Q_{X|Y}^* | \hat{P}_\mathbf{y}) \Big\}.
\end{aligned} \tag{B.9}$$

with $Q^*_{X|Y}$ being the solution of the optimization problem:

$$\begin{aligned} \underset{Q_{X|Y}}{\textit{minimize}} \quad & D(Q_{X|Y}\|P_{X|Y}|\hat{P}_{\mathbf{y}}) + H(Q_{X|Y}|\hat{P}_{\mathbf{y}}) \\ \textit{subject to} \quad & H(Q_{X|Y}|\hat{P}_{\mathbf{y}}) \geq \alpha. \end{aligned} \tag{B.10}$$

Since Lemma 22 is a direct consequence of Lemma 23, we only include the proof of the latter.

Proof. Recall that $P_{Y|X}(\mathbf{x}|\mathbf{y}) = \exp\{-n\left(D(\hat{P}_{\mathbf{x}|\mathbf{y}}\|P_{X|Y}|\hat{P}_{\mathbf{y}}) + H(\hat{P}_{\mathbf{x}|\mathbf{y}}|\hat{P}_{\mathbf{y}})\right)\}$. Furthermore, note that for a given $\mathbf{y}$ the number of sequences $\mathbf{x}$ which have conditional type $Q_{\mathbf{x}|\mathbf{y}}$ is given by $|T(Q_{\mathbf{x}|\mathbf{y}})(\mathbf{y})| \doteq \exp\{nH(Q_{\mathbf{x}|\mathbf{y}}|\hat{P}_{\mathbf{y}})\}$ (see, e.g., [54, Lemma 2.5]). Let $\hat{\mathcal{Q}}(\alpha, \hat{P}_{\mathbf{y}})$ be the set of types of the sequences that are in the first $\mathcal{X}^{n\alpha}$ position in the list, that is $\hat{\mathcal{Q}}(\alpha, \hat{P}_{\mathbf{y}})$ is such that:

$$\sum_{Q_{X|Y} \in \mathcal{Q}(\alpha, \hat{P}_{\mathbf{y}})} 2^{nH(Q_{X|Y}|\hat{P}_{\mathbf{y}})} = 2^{n\alpha}. \tag{B.11}$$

An application of the method of types gives that the left-hand side evaluates (exponentially) to $2^{n \sup_{Q_{X|Y} \in \hat{\mathcal{Q}}(\alpha, \hat{P}_{\mathbf{y}})} H(Q_{X|Y}|\hat{P}_{\mathbf{y}})}$, meaning that $\sup_{Q_{X|Y} \in \mathcal{Q}} H(Q_{X|Y}|\hat{P}_{\mathbf{y}}) = \alpha$. Thus, the threshold probability is given by (B.10), and any type that has lower probability appears before in the list. ■

The list $\mathcal{Q}(\alpha, \hat{P}_{\mathbf{y}})$ is specified implicitely for any $P_{Y|X}$, but can also be made explicit for some specific channels. In particular, binary erasures channels yield to an easy characterization of $\mathcal{Q}(\alpha, \hat{P}_{\mathbf{y}})$. Indeed, in this case the only reverse channel types which need to be considered are those that are valid outputs of an erasure channel. Thus, the order of $\mathbf{x}$ in the ordered list after observation $\mathbf{y}$ solely depends on the type of $\mathbf{x}$ over the position which are erased in $\mathbf{y}$, which we shall denote by $\hat{Q}^{(\epsilon)}_{X|Y}$. Letting ϵ be the erasure symbol, $\hat{P}_{\mathbf{y}}(\epsilon)$ is thus the fraction of erasures in the received output $\mathbf{y}$, and assuming $P_X(0) > P_X(1)$, we have $\hat{P}_{X|Y} \in \mathcal{Q}(\alpha, \hat{P}_{\mathbf{y}})$ iff $\hat{Q}^{(\epsilon)}_{X|Y} < \frac{\alpha}{\hat{P}_{\mathbf{y}}(\epsilon)}$.

Appendix C

Applications to one-to-one Coding

In this section, we connect the established results on mismatched decoding to lossless source coding. We follow the notation from [52], and start by a discussion on lossless coding without mismatch. A lossless source code is an injective function $f : \mathcal{X} \to \{0,1\}^*$, and we refer to $f(x)$, for some $x \in \mathcal{X}$ as a codeword. For a codeword $c \in \{0,1\}^*$, the length of the codeword is denoted by $l(c)$. A lossless source code f^* is said to be optimal if it satisfies $\mathbb{E}[l(f^*(X))] \geq \mathbb{E}[l(f(X))]$ for all valid source codes f.

The relationship between the optimal source code f^* and the log-guesswork g_P, was discussed in [15] [79], and later in [48]. Essentially, this correspondence is due to the relation $P(x) \geq P(y) \iff l(f^*(x)) \leq l(f^*(y))$, which imposes that there is an optimal encoding with $l(f^*(x)) \geq \lfloor \log_2 G_P(x) \rfloor$ for all $x \in \mathcal{X}$. For iid sources, the asymptotic behavior of lossless codes are investigated through two quantities of interest, namely the asymptotic average length, and the reliability function

$$L(P) \triangleq \lim_{n\to\infty} \frac{1}{n} \mathbb{E}[l(f^*(X^n))], \tag{C.1}$$

$$E(R,P) \triangleq -\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}_{P^n} \left(l(f^*(X^n)) > nR \right), \tag{C.2}$$

where $H(P) < R < \log |\mathcal{X}|$. Naturally, the average length $L(P) = H(P)$, that is, the best average length for a lossless code is asymptotically converging to the entropy of the source, see [138]. By using the correspondence between $l(f^*(x^n))$ and $g_P(x^n)$, one can directly apply the results in Theorem 7 to obtain closed forms on the reliability function $E(R,P)$ (we refer to [52] for more details). In the rest of this section, we discuss analogous quantities for the case of mismatched lossless coding without prefix-free constraint.

Now, assume that an optimal lossless source code is constructed according to a mismatched source statistic Q. We let f_Q^* be the resulting optimal code for the source statistic Q, and define the asymptotic average length and reliability function similarly as in the matched case, i.e.,

$$L(Q\|P) = \lim_{n\to\infty} \frac{1}{n}\mathbb{E}[l(f_Q^*(X^n))] \tag{C.3}$$

$$E(R, Q\|P) \triangleq -\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}_{P^n}\left(l(f_Q^*(X^n)) > nR\right). \tag{C.4}$$

The following is the main result of this section, and is a direct consequence of the LDP result on the mismatched guesswork.

Theorem 17. *Let $X^n \sim P^n$, and assume $\Pi_{\mathcal{T}}(P) \in \mathcal{T}_Q^+$, then:*

$$L(Q\|P) = H(\Pi_{\mathcal{T}_Q}(P)), \tag{C.5}$$

$$E(R, Q\|P) = J(R), \tag{C.6}$$

for $H(\Pi_{\mathcal{T}_Q}(P)) < R < \log|\mathcal{X}|$.

Proof. The proof of the statement on the reliability function follows immediately by noting that there is an optimal encoding such that $g_Q(x^n) \le l(f_Q^*(x^n)) < g_Q(x^n) + 1$, and by applying Theorem 13. The result on $L(Q\|P)$ follows from:

$$L(Q\|P) = \lim_{n\to\infty} \frac{1}{n}\mathbb{E}_{P^n}\left[g_Q(X^n)\right] \tag{C.7}$$

$$= \lim_{\rho\downarrow 0} E_\rho(Q\|P), \tag{C.8}$$

where the first equality is again a consequence of the correspondence between optimal code and guesswork, while the second equality is an application of L'Hôpital's rule. Recall that, by Corollary 8, $E_\rho(Q\|P) = \max_{\gamma\in\mathcal{T}_Q^+} H(\Pi_{\mathcal{T}_Q}(\gamma))) - \frac{1}{\rho}D(\gamma\|P)$. It follows that when $\rho \downarrow 0$, it must be that $\gamma = P$, which results in $L(Q\|P) = H(\Pi_{\mathcal{T}_Q}(P))$. ∎

In prefix free coding, the average length of the coded iid sequence is governed by the cross entropy $H(P\|Q)$, where P is the true distribution, and Q is the mismatched distribution used to generate the code. In particular, since $D(P\|Q) \ge 0$, with equality only if $P = Q$, there is always a loss in performance in using a mismatched distribution. The result above

guarantees that the performance of a lossless one-to-one code always exceeds that of a prefix-free code in terms of asymptotic average length, in the presence of mismatch. Indeed, we have by Lemma 10,

$$H(\Pi_{\mathcal{T}_Q}(P)) = H(P\|\Pi_{\mathcal{T}_Q}(P)) \tag{C.9}$$

$$= H(P) + D(P\|\Pi_{\mathcal{T}_Q}(P)) \tag{C.10}$$

$$\leq H(P) + D(P\|Q), \tag{C.11}$$

where the last step follows from Lemma 9. Therefore, the penalty induced by mismatch from one-to-one coding is always upper bounded by the penalty for prefix-free codes as the asymptotic average codeword length in both cases is characterized by $H(P)$ [138]. The relative entropy $D(P\|\Pi_{\mathcal{T}_Q}(P))$ can also be 0, if $P \in \mathcal{T}_Q^+$, i.e., if P and Q are on the same tilted distribution. This implies that the cost of mismatched source coding vanishes if and only if $P \in \mathcal{T}_Q^+$ (Lemma 13), and Theorem 17 generalizes such characterization to arbitrary mismatched distributions.

Bibliography

[1] Bitcoin Wallets under siege from 'Large Collider' Attack. http://fortune.com/2017/04/15/bitcoin-collider/.

[2] 'Brute force' cyber attack on Parliament compromised up to 90 email accounts. http://www.telegraph.co.uk/news/2017/06/25/brute-force-cyber-attack-parliament-compromised-90-email-accounts/.

[3] How Companies Learn Your Secrets. https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html.

[4] Anatomy of a password disaster – Adobe's giant-sized cryptographic blunder. https://nakedsecurity.sophos.com/2013/11/04/anatomy-of-a-password-disaster-adobes-giant-sized-cryptographic-blunder/, 2013 (last accessed May 2020).

[5] McAfee Labs Threat Report. https://www.mcafee.com/ca/resources/reports/rp-quarterly-threats-sept-2017.pdf, 2017.

[6] Learning with privacy at scale. https://machinelearning.apple.com/2017/12/06/learning-with-privacy-at-scale.html, 2017 (last accessed May 2020).

[7] Cambridge analytica and facebook: The scandal and the fallout so far. https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html, 2018 (last accessed July 7, 2020).

[8] Google health-data scandal spooks researchers. https://www.nature.com/articles/d41586-019-03574-5, 2019 (last accessed July 20, 2020).

[9] Disclosure avoidance and the 2020 census. https://www.census.gov/about/policies/privacy/statistical_safeguards/disclosure-avoidance-2020-census.html, 2020 (last accessed May 2020).

[10] Lecture notes 13: 6.437 inference and information, Spring 2016.

[11] Abbas Acar, Hidayet Aksu, A Selcuk Uluagac, and Mauro Conti. A survey on homomorphic encryption schemes: Theory and implementation. *ACM Computing Surveys (CSUR)*, 51(4):1–35, 2018.

[12] Rakesh Agrawal and Ramakrishnan Srikant. Privacy-preserving data mining. *ACM Sigmod Record*, 29(2):439–450, 2000.

[13] Mário S Alvim, Miguel E Andrés, Konstantinos Chatzikokolakis, Pierpaolo Degano, and Catuscia Palamidessi. Differential privacy: on the trade-off between utility and information leakage. In *International Workshop on Formal Aspects in Security and Trust*, pages 39–54. Springer, 2011.

[14] E. Arikan. An inequality on guessing and its application to sequential decoding. *IEEE Trans. on Inf. Theory*, 42(1):99–105, Jan. 1996.

[15] E. Arikan and N. Merhav. Guessing subject to distortion. *IEEE Trans. on Inf. Theory*, 44(3):1041–1056, May 1998.

[16] Shahab Asoodeh, Fady Alajaji, and Tamás Linder. Notes on information-theoretic privacy. In *Communication, Control, and Computing (Allerton), 2014 52nd Annual Allerton Conference on*, pages 1272–1278. IEEE, 2014.

[17] Shahab Asoodeh, Mario Diaz, Fady Alajaji, and Tamás Linder. Information extraction under privacy constraints. *Information*, 7(1):15, 2016.

[18] Shahab Asoodeh, Mario Diaz, Fady Alajaji, and Tamás Linder. Privacy-aware guessing efficiency. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 754–758. IEEE, 2017.

[19] Shahab Asoodeh, Mario Diaz, Fady Alajaji, and Tamás Linder. Estimation efficiency under privacy constraints. *IEEE Transactions on Information Theory*, 65(3):1512–1534, 2018.

[20] A. Asuncion and D.J. Newman. UCI machine learning repository, 2007.

[21] Koenraad M R Audenaert. A Sharp Fannes-type Inequality for the von Neumann Entropy. *arXiv preprint quant-ph/0610146*, 2006.

[22] Siddhartha Banerjee, Nidhi Hegde, and Laurent Massoulié. The price of privacy in untrusted recommendation engines. *arXiv preprint arXiv:1207.3269*, 2012.

[23] Yuksel Ozan Basciftci, Ye Wang, and Prakash Ishwar. On privacy-utility tradeoffs for constrained data release mechanisms. In *2016 information theory and applications workshop (ITA)*, pages 1–6. IEEE.

[24] A. Beirami, R. Calderbank, M. Christiansen, K. Duffy, A. Makhdoumi, and M. Médard. A geometric perspective on guesswork. In *53rd Annual Allerton Conference (Allerton)*, Oct. 2015.

[25] A. Beirami, R. Calderbank, M. Christiansen, K. Duffy, and M. Médard. A characterization of guesswork on swiftly tilting curves. *IEEE Transactions on Information Theory*, pages 1–1, 2018.

[26] A. Beirami, R. Calderbank, K. Duffy, and M. Médard. Quantifying computational security subject to source constraints, guesswork and inscrutability. In *2015 IEEE International Symposium on Information Theory Proceedings*, Jun. 2015.

[27] Dimitris Bertsimas and John N Tsitsiklis. *Introduction to linear optimization*, volume 6. Athena Scientific Belmont, MA, 1997.

[28] Alina Beygelzimer, Sham Kakade, and John Langford. Cover trees for nearest neighbor. In *Proceedings of the 23rd international conference on Machine learning*, pages 97–104, 2006.

[29] Jeremiah Blocki, Ben Harsha, and Samson Zhou. On the economics of offline password cracking. *IEEE Security and Privacy (to appear)*, 2018.

[30] Avrim Blum, Katrina Ligett, and Aaron Roth. A learning theory approach to noninteractive database privacy. *Journal of the ACM (JACM)*, 60(2):12, 2013.

[31] Anselm C Blumer and Robert J McEliece. The rényi redundancy of generalized huffman codes. *IEEE Transactions on Information Theory*, 34(5):1242–1249, 1988.

[32] Dan Boneh, Amit Sahai, and Brent Waters. Functional encryption: Definitions and challenges. In *Theory of Cryptography Conference*, pages 253–273. Springer, 2011.

[33] Joseph Bonneau. The science of guessing: analyzing an anonymized corpus of 70 million passwords. In *Security and Privacy (SP), 2012 IEEE Symposium on*, pages 538–552. IEEE, 2012.

[34] Luca Bonomi, Liyue Fan, and Hongxia Jin. An information-theoretic approach to individual sequential data sanitization. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pages 337–346. ACM, 2016.

[35] Shashi Borade and Lizhong Zheng. I-projection and the geometry of error exponents. In *Proceedings of the Forty-Fourth Annual Allerton Conference on Communication, Control, and Computing, Sept 27-29*, 2006.

[36] Stephen Poythress Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[37] S. Boztaş. Oblivious distributed guessing. In *2012 IEEE International Symposium on Information Theory Proceedings*, pages 2162–2165, Jul. 2012.

[38] Serdar Boztas. On rényi entropies and their applications to guessing attacks in cryptography. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, 97(12):2542–2548, 2014.

[39] Annina Bracher, Eran Hof, and Amos Lapidoth. Guessing attacks on distributed-storage systems. *arXiv preprint arXiv:1701.01981*, 2017.

[40] F. P. Calmon, M. Varia, and M. Médard. An exploration of the role of principal inertia components in information theory. In *Information Theory Workshop (ITW), 2014 IEEE*, pages 252–256, 2014.

[41] Flavio P Calmon, Ali Makhdoumi, and Muriel Médard. Fundamental limits of perfect privacy. In *2015 IEEE International Symposium on Information Theory (ISIT)*, pages 1796–1800. IEEE, 2015.

[42] L Lore Campbell. A coding theorem and rényi's entropy. *Information and control*, 8(4):423–429, 1965.

[43] Moses Charikar, Chandra Chekuri, Tomás Feder, and Rajeev Motwani. Incremental clustering and dynamic information retrieval. *SIAM Journal on Computing*, 33(6):1417–1440, 2004.

[44] Konstantinos Chatzikokolakis, Tom Chothia, and Apratim Guha. Statistical measurement of information leakage. In *International Conference on Tools and Algorithms for the Construction and Analysis of Systems*, pages 390–404. Springer, 2010.

[45] Konstantinos Chatzikokolakis, Catuscia Palamidessi, and Prakash Panangaden. Anonymity protocols as noisy channels. *Information and Computation*, 206(2-4):378–401, 2008.

[46] Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *The Journal of Machine Learning Research*, 12:1069–1109, 2011.

[47] Kamalika Chaudhuri, Anand D Sarwate, and Kaushik Sinha. Near-optimal differentially private principal components. In *NIPS*, pages 998–1006, 2012.

[48] M. M. Christiansen and K. R. Duffy. Guesswork, large deviations, and Shannon entropy. *IEEE Trans. on Inf. Theory*, 59(2):796–802, Feb. 2013.

[49] Mark M Christiansen, Ken R Duffy, Flávio du Pin Calmon, and Muriel Médard. Guessing a password over a wireless channel (on the effect of noise non-uniformity). In *Signals, Systems and Computers, 2013 Asilomar Conference on*, pages 51–55. IEEE, 2013.

[50] Mark M Christiansen, Ken R Duffy, Flávio du Pin Calmon, and Muriel Médard. Multi-user guesswork and brute force security. *IEEE Transactions on Information Theory*, 61(12):6876–6886, 2015.

[51] John Horton Conway and Neil James Alexander Sloane. *Sphere packings, lattices and groups*, volume 290. Springer Science & Business Media, 2013.

[52] Thomas A Courtade and Sergio Verdú. Cumulant generating function of codeword lengths in optimal lossless compression. In *2014 IEEE International Symposium on Information Theory*, pages 2494–2498. IEEE, 2014.

[53] Thomas M Cover and Joy A Thomas. *Elements of information theory*. Wiley-interscience, 2012.

[54] Imre Csiszár and János Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press, 2 edition, August 2011.

[55] Imre Csiszár and Paul C Shields. *Information theory and statistics: A tutorial*. Now Publishers Inc, 2004.

[56] Nik Cubrilovic. Rockyou hack: From bad to worse. `https://techcrunch.com/2009/12/14/rockyou-hack-security-myspace-facebook-passwords/`, 2009 (accessed January 2020).

[57] Anupam Das, Joseph Bonneau, Matthew Caesar, Nikita Borisov, and XiaoFeng Wang. The tangled web of password reuse. In *NDSS*, volume 14, pages 23–26, 2014.

[58] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*. Springer, 1998.

[59] Wenliang Du and Mikhail J Atallah. Secure multi-party computation problems and their applications: a review and open problems. In *Proceedings of the 2001 workshop on New security paradigms*, pages 13–22, 2001.

[60] Flávio du Pin Calmon and Nadia Fawaz. Privacy against statistical inference. In *2012 50th annual Allerton conference on communication, control, and computing (Allerton)*, pages 1401–1408. IEEE, 2012.

[61] John C Duchi, Michael I Jordan, and Martin J Wainwright. Local privacy and statistical minimax rates. *arXiv preprint arXiv:1302.3203*, 2013.

[62] C. Dwork, F. Mcsherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, 2006.

[63] Cynthia Dwork. Differential privacy. In *Automata, Languages and Programming*, volume 4052, pages 1–12. Springer, 2006.

[64] Cynthia Dwork. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*, pages 1–19. Springer, 2008.

[65] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toni Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. In *Advances in Neural Information Processing Systems*, pages 2350–2358, 2015.

[66] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Leon Roth. Preserving statistical validity in adaptive data analysis. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, pages 117–126. ACM, 2015.

[67] Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *Proceedings of the 41st annual ACM symposium on Theory of computing*. ACM, 2009.

[68] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In Shai Halevi and Tal Rabin, editors, *Theory of Cryptography*, pages 265–284, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[69] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[70] Cynthia Dwork, Guy N Rothblum, and Salil Vadhan. Boosting and differential privacy. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 51–60. IEEE, 2010.

[71] H. G. Eggleston. *Convexity*. Cambridge University Press, Cambridge England, 1 edition edition, January 2009.

[72] Alexandre Evfimievski, Johannes Gehrke, and Ramakrishnan Srikant. Limiting privacy breaches in privacy preserving data mining. In *ACM PODS*, 2003.

[73] Arik Friedman and Assaf Schuster. Data mining with differential privacy. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 493–502. ACM, 2010.

[74] Allen Gersho and Robert M Gray. *Vector quantization and signal compression*, volume 159. Springer Science & Business Media, 2012.

[75] Arpita Ghosh and Aaron Roth. Selling privacy at auction. *Games and Economic Behavior*, 91:334–346, 2015.

[76] Shafi Goldwasser, Yael Tauman Kalai, Raluca Ada Popa, Vinod Vaikuntanathan, and Nickolai Zeldovich. How to run turing machines on encrypted data. In *Annual Cryptology Conference*, pages 536–553. Springer, 2013.

[77] Sergey Gorbunov, Vinod Vaikuntanathan, and Hoeteck Wee. Functional encryption with bounded collusions via multi-party computation. In *Annual Cryptology Conference*, pages 162–179. Springer, 2012.

[78] M. J. Hanawal and R. Sundaresan. Randomised attacks on passwords. In *DRDO-IISc Programme on Advanced Research in Mathematical Engineering*, Feb. 2010.

[79] M. K. Hanawal and R. Sundaresan. Guessing revisited: A large deviations approach. *IEEE Trans. on Inf. Theory*, 57(1):70–78, Jan. 2011.

[80] F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19, 2015.

[81] Kjell Jørgen Hole, Vebjørn Moen, and Thomas Tjostheim. Case study: Online banking security. *IEEE Security & Privacy*, 4(2):14–20, 2006.

[82] Hsiang Hsu, Shahab Asoodeh, and Flavio du Pin Calmon. Obfuscation via information density estimation. *arXiv preprint arXiv:1910.08109*, 2019.

[83] Hsiang Hsu, Shahab Asoodeh, Salman Salamatian, and Flavio P Calmon. Generalizing bottleneck problems. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 531–535. IEEE, 2018.

[84] Wasim Huleihel, Salman Salamatian, and Muriel Médard. Guessing with limited memory. In *Information Theory (ISIT), 2017 IEEE International Symposium on*, pages 2253–2257. IEEE, 2017.

[85] Shiva Prasad Kasiviswanathan, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

[86] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[87] Oliver Kosut and Lalitha Sankar. Asymptotics and non-asymptotics for universal fixed-to-variable source coding. *IEEE Transactions on Information Theory*, 2017.

[88] Zuxing Li, Tobias J Oechtering, and Deniz Gündüz. Privacy against a hypothesis testing adversary. *IEEE Transactions on Information Forensics and Security*, 14(6):1567–1581, 2018.

[89] Jiachun Liao, Oliver Kosut, Lalitha Sankar, and Flavio P Calmon. A tunable measure for information leakage. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 701–705. IEEE, 2018.

[90] Jiachun Liao, Lalitha Sankar, Oliver Kosut, and Flavio P Calmon. Robustness of Maximal α-Leakage to Side Information. *arXiv preprint arXiv:1901.07105*, 2019.

[91] Jiachun Liao, Lalitha Sankar, Vincent YF Tan, and Flavio du Pin Calmon. Hypothesis testing under mutual information privacy constraints in the high privacy regime. *IEEE Transactions on Information Forensics and Security*, 13(4):1058–1071, 2017.

[92] Litian Liu, Amit Solomon, Salman Salamatian, and Muriel Médard. Neural network coding. In *ICC 2020-2020 IEEE International Conference on Communications (ICC)*, pages 1–6. IEEE, 2020.

[93] Nate Lord. Uncovering Password Habits: Are Users' Password Security Habits Improving? `https://digitalguardian.com/blog/uncovering-password-habits-are-users-password-security-habits-improving-infographic`, 2018 (last accessed August 2020).

[94] Ashwin Machanavajjhala, Aleksandra Korolova, and Atish Das Sarma. Personalized social recommendations: accurate or private. *Proceedings of the VLDB Endowment*, 4(7):440–450, 2011.

[95] Ali Makhdoumi and Nadia Fawaz. Privacy-utility tradeoff under statistical uncertainty. In *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1627–1634. IEEE, 2013.

[96] Ali Makhdoumi, Salman Salamatian, Nadia Fawaz, and Muriel Médard. From the information bottleneck to the privacy funnel. In *2014 IEEE Information Theory Workshop (ITW 2014)*, pages 501–505. IEEE, 2014.

[97] James L Massey. Guessing and entropy. In *Proceedings of 1994 IEEE International Symposium on Information Theory*, page 204. IEEE, 1994.

[98] Andrew McGregor, Ilya Mironov, Toniann Pitassi, Omer Reingold, Kunal Talwar, and Salil Vadhan. The limits of two-party differential privacy. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 81–90. IEEE, 2010.

[99] Frank McSherry and Ilya Mironov. Differentially private recommender systems: building privacy into the net. In *ACM SIGKDD*, 2009.

[100] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *Foundations of Computer Science, 2007. FOCS'07. 48th Annual IEEE Symposium on*, pages 94–103. IEEE, 2007.

[101] Neri Merhav and Erdal Arikan. The shannon cipher system with a guessing wiretapper. *IEEE Transactions on Information Theory*, 45(6):1860–1866, 1999.

[102] Neri Merhav and Meir Feder. Universal prediction. *IEEE Transactions on Information Theory*, 44(6):2124–2147, 1998.

[103] Darakhshan J Mir. Information-theoretic foundations of differential privacy. In *International Symposium on Foundations and Practice of Security*, pages 374–381. Springer, 2012.

[104] Nina Mishra and Mark Sandler. Privacy via pseudorandom sketches. In *Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 143–152. ACM, 2006.

[105] Bahman Moraffah and Lalitha Sankar. Information-theoretic private interactive mechanism. In *Communication, Control, and Computing (Allerton), 2015 53rd Annual Allerton Conference on*, pages 911–918. IEEE, 2015.

[106] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 111–125. IEEE, 2008.

[107] Seyed Ali Osia, Borzoo Rassouli, Hamed Haddadi, Hamid R Rabiee, and Deniz Gündüz. Privacy against brute-force inference attacks. *arXiv preprint arXiv:1902.00329*, 2019.

[108] Jim Owens and Jeanna Matthews. A study of passwords and methods used in brute-force ssh attacks. In *USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET)*, 2008.

[109] Hari Palaiyanur and Anant Sahai. On the uniform continuity of the rate-distortion function. *ISIT 2008*, pages 1–5, jan 2008.

[110] C. E. Pfister and W. G. Sullivan. Rényi entropy, guesswork moments, and large deviations. *IEEE Trans. on Inf. Theory*, 50(11):2794–2800, Nov. 2004.

[111] Ariel Rabkin. Personal knowledge questions for fallback authentication: Security questions in the era of facebook. In *Proceedings of the 4th symposium on Usable privacy and security*, pages 13–23, 2008.

[112] Maxim Raginsky. Strong data processing inequalities and Φ-sobolev inequalities for discrete channels. *IEEE Transactions on Information Theory*, 62(6):3355–3389, 2016.

[113] Borzoo Rassouli and Deniz Gunduz. Optimal utility-privacy trade-off with total variation distance as a privacy measure. *IEEE Transactions on Information Forensics and Security*, 2019.

[114] Borzoo Rassouli, Fernando E Rosas, and Deniz Gündüz. Data disclosure under perfect sample privacy. *IEEE Transactions on Information Forensics and Security*, 15:2012–2025, 2019.

[115] Judith Rauhofer. Privacy is dead, get over it! information privacy and the dream of a risk-free society. *Information & Communications Technology Law*, 17(3):185–197, 2008.

[116] David Rebollo-Monedero, Jordi Forne, and Josep Domingo-Ferrer. From t-closeness-like privacy to postrandomization via information theory. *Knowledge and Data Engineering, IEEE Transactions on*, 22(11):1623–1636, 2010.

[117] I. S. Reed. Information Theory and Privacy in Data Banks. In *Proc. of national computer conference and exposition*. ACM, 1973.

[118] Irving S Reed. Information theory and privacy in data banks. In *Proceedings of the June 4-8, 1973, national computer conference and exposition*. ACM, 1973.

[119] Benjamin IP Rubinstein, Peter L Bartlett, Ling Huang, and Nina Taft. Learning in a large function space: Privacy-preserving mechanisms for svm learning. *arXiv preprint arXiv:0911.5708*, 2009.

[120] Ira S Rubinstein. Voter privacy in the age of big data. *Wisconsin Law Review*, page 861, 2014.

[121] Ryan Singel. Netflix Spilled Your Brokeback Mountain Secret, Lawsuit Claims, 2014, last accessed 07/2020.
`https://www.wired.com/2009/12/netflix-privacy-lawsuit/`.

[122] Salman Salamatian, Ahmad Beirami, Asaf Cohen, and Muriel Médard. Centralized vs decentralized multi-agent guesswork. In *Information Theory (ISIT), 2017 IEEE International Symposium on*, pages 2258–2262. IEEE, 2017.

[123] Salman Salamatian, Asaf Cohen, and Muriel Médard. Efficient coding for multi-source networks using gács-körner common information. In *2016 International Symposium on Information Theory and Its Applications (ISITA)*, pages 166–170. IEEE, 2016.

[124] Salman Salamatian, Muriel Médard, and Emre Telatar. A successive description property of monotone-chain polar codes for slepian-wolf coding. In *2015 IEEE International Symposium on Information Theory (ISIT)*, pages 1522–1526. IEEE, 2015.

[125] Salman Salamatian, Nir Shlezinger, Yonina C Eldar, and Muriel Médard. Task-based quantization for recovering quadratic functions using principal inertia components. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 390–394. IEEE, 2019.

[126] Lalitha Sankar, S Raj Rajagopalan, and H Vincent Poor. Utility and privacy of data sources: Can shannon help conceal and reveal information? In *Information Theory and Applications Workshop (ITA), 2010*, pages 1–7. IEEE, 2010.

[127] Lalitha Sankar, S Raj Rajagopalan, and H Vincent Poor. Utility-privacy tradeoffs in databases: An information-theoretic approach. *IEEE Transactions on Information Forensics and Security*, 8(6):838–852, 2013.

[128] Anand D Sarwate and Lalitha Sankar. A rate-disortion perspective on local differential privacy. In *Allerton*, pages 903–908, 2014.

[129] I. Sason and S. Verdú. Arimoto-rényi conditional entropy and bayesian m-ary hypothesis testing. *IEEE Transactions on Information Theory*, PP(99):1–1, 2017.

[130] Claude E Shannon. Communication theory of secrecy systems. *The Bell system technical journal*, 28(4):656–715, 1949.

[131] N. Slonim and N. Tishby. Agglomerative information bottleneck. *Proc. of Neural Information Processing Systems (NIPS-99)*, 1999.

[132] Adam Smith. Privacy-preserving statistical estimation with optimal convergence rates. In *Proceedings of the 43rd annual ACM symposium on Theory of computing*, pages 813–822. ACM, 2011.

[133] A. Somekh-Baruch and N. Merhav. Achievable error exponents for the private fingerprinting game. *IEEE Trans. on Inf. Theory*, 53(5):1827–1838, May 2007.

[134] Sreejith Sreekumar, Asaf Cohen, and Deniz Gündüz. Distributed hypothesis testing with a privacy constraint. *arXiv preprint arXiv:1807.02764*, 2018.

[135] D. Sullivan and W. G. Sullivan. Guesswork and entropy. *IEEE Trans. on Inf. Theory*, 50(3):525–526, Mar. 2004.

[136] R. Sundaresan. Guessing under source uncertainty. *IEEE Trans. on Inf. Theory*, 53(1):525–526, Jan. 2007.

[137] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

[138] Wojciech Szpankowski. A one-to-one code and its anti-redundancy. *IEEE transactions on information theory*, 54(10):4762–4766, 2008.

[139] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

[140] Ding Wang, Haibo Cheng, Ping Wang, Xinyi Huang, and Gaopeng Jian. Zipf's law in passwords. *IEEE Transactions on Information Forensics and Security*, 12(11):2776–2791, 2017.

[141] Ding Wang, Haibo Cheng, Ping Wang, Jeff Yan, and Xinyi Huang. A security analysis of honeywords. NDSS, 2018.

[142] Ding Wang and Ping Wang. On the implications of zipf's law in passwords. In *European Symposium on Research in Computer Security*, pages 111–131. Springer, 2016.

[143] Ding Wang, Zijian Zhang, Ping Wang, Jeff Yan, and Xinyi Huang. Targeted online password guessing: An underestimated threat. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 1242–1254. ACM, 2016.

[144] Hao Wang and Flavio P Calmon. An estimation-theoretic view of privacy. In *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 886–893. IEEE, 2017.

[145] Hao Wang, Lisa Vo, Flavio P Calmon, Muriel Médard, Ken R Duffy, and Mayank Varia. Privacy with estimation guarantees. *IEEE Transactions on Information Theory*, 65(12):8025–8042, 2019.

[146] Ke Wang, Philip S Yu, and Sourav Chakraborty. Bottom-up generalization: A data mining solution to privacy protection. In *Data Mining, 2004. ICDM'04. Fourth IEEE International Conference on*. IEEE, 2004.

[147] Weina Wang, Lei Ying, and Junshan Zhang. On the relation between identifiability, differential privacy, and mutual-information privacy. *IEEE Transactions on Information Theory*, 62(9):5018–5029, 2016.

[148] Stanley L Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of American Statistical Association*, 1965.

[149] H.S. Witsenhausen and A.D. Wyner. A conditional entropy bound for a pair of discrete random variables. *IEEE Transactions on Information Theory*, 21(5):493–501, September 1975.

[150] A Wyner and Jacob Ziv. A theorem on the entropy of certain binary sequences and applications–i. *IEEE Transactions on Information Theory*, 19(6):769–772, 1973.

[151] H. Yamamoto. A source coding problem for sources with additional outputs to keep secret from the receiver of wiretappers. 29(6), 1983.

[152] Hirosuke Yamamoto. A source coding problem for sources with additional outputs to keep secret from the receiver or wiretappers (corresp.). *Information Theory, IEEE Transactions on*, 29(6), 1983.

[153] Zhengmin Zhang. Estimating Mutual Information Via Kolmogorov Distance. *Information Theory, IEEE Transactions on*, 53(9).

[154] Ye Zhu and Riccardo Bettati. Anonymity vs. information leakage in anonymity systems. In *Distributed Computing Systems, 2005. ICDCS 2005. Proceedings. 25th IEEE International Conference on*, pages 514–524. IEEE, 2005.