One-Shot Coding and Applications

LIU, Yanxiao

A Thesis Submitted in Partial Fulfilment
of the Requirements for the Degree of
Doctor of Philosophy
in
Information Engineering

The Chinese University of Hong Kong
July 2025

Thesis Assessment Committee

Professor Cheuk Ting Li (Thesis Supervisor)

Professor Raymond W. Yeung (Thesis Co-supervisor)

Professor Chandra Nair (Committee Chair)

Professor Amin Gohari (Committee Member)

Professor Ying-Jun Angela Zhang (Committee Member)

Professor Vincent Y. F. Tan (External Examiner)

Abstract

One-shot information theory addresses scenarios in source coding and channel coding where the signal blocklength is assumed to be 1. In this case, each source and channel can be used only once, and the sources and channels are arbitrary and not required to be memoryless or ergodic. We study the *achievability* part of one-shot information theory, i.e., we consider explicit coding schemes in the one-shot scenario. The objective is to derive one-shot achievability results that can imply existing (first-order and second-order) asymptotic results when applied to memoryless sources and channels, or applied to systems with memory that behave ergodically.

Poisson functional representation was first proposed as a one-shot channel simulation technique by Li and El Gamal [118] for proving a strong functional representation lemma. It was later extended to the Poisson matching lemma by Li and Anantharam [117], which provided a unified one-shot coding scheme for a broad class of information-theoretic problems. The main contribution of this thesis is to extend the applicability of Poisson functional representation to various more complicated scenarios, where the original version cannot be applied directly and further extensions must be developed. Below, we highlight some of the key contributions.

1. In Chapter 3, we design a unified one-shot coding framework for the communication and compression of messages among multiple nodes across a *general*

acyclic noisy network. This framework can be viewed as a one-shot counterpart to the unified random coding bound studied by Lee and Chung [110], as well as the noisy network coding developed by Lim *et al.* [126]. Our general framework not only recovers a wide range of existing one-shot and asymptotic results but also provides novel one-shot achievability results for various network information theory problems.

2. In Chapter 4, we examine two classes of secrecy problems where the channel conditions are *unknown* to the encoder and the decoder, based on the Poisson matching lemma and a covering argument. We provide one-shot achievability results for a generalized information hiding setting [144] and the compound wiretap channel [123], each of which recovers many existing problems as special cases.

3. In Chapter 5, leveraging the Poisson functional representation, we design a novel construction called *Poisson private representation* that can compress arbitrary differential privacy mechanisms. It is the first scheme that achieves a close-to-optimal compression size (within a logarithmic gap), exactness of the output distribution (thus preserving all the desirable statistical properties of the original privacy mechanism, such as unbiasedness and Gaussianity), while ensuring local differential privacy. New trade-offs among communication, accuracy, and central and local differential privacy are established, and experimental advantages are demonstrated across different applications.

摘要

單發信息论探討的是信道编码与信源压缩在信號區塊長度 1 时的一般情境。在此情況下，每個信道與信源僅使用一次，且信道與信源都可以是任意的，並不被要求具備無記憶性或遍歷性。本論文的目標是構造出新的編碼方式，以推導出新的單發可達性結果。當這些結果應用於無記憶的信道與信源或是有遍歷性的有記憶系統時，能夠推導出既有的一階與二階的漸近結果。泊松函數表示作為一種單發信道模擬技術，最初由 Li 與 El Gamal 提出，用於證明強函數表示引理及其他結果。Li 與 Anantharam 將其擴展為泊松配對引理，並提供了一套能夠涵蓋廣泛的信息论問題的統一單發編碼方案。本論文的主要貢獻在於進一步擴展泊松函數表示與泊松配對引理的適用範圍，基於它們來設計新的工具和處理更為複雜的情境。下述分章節列舉本論文的幾個核心貢獻。

在第 3 章中，我們設計了一個針對任意無環噪聲網路中多個節點間資訊通信與壓縮的統一框架。此框架可被視為 Lee 與 Chung 所研究之漸進隨機編碼界限的單發對應版本，或者 Lim 等人所提出之噪聲網絡編碼的單發對應版本。我們的編碼框架涵蓋極廣泛的網路信息论問題，不僅能復現多種既有的單發與漸近結果，亦提出了多種新的單發可達性結果。在第 4 章中，我們基於泊松配對引理，探討了兩類與資訊安全相關的問題，其共同點和難點在於編碼器與解碼器需要在未知的信道狀態下編碼。我們對一般化的資訊隱藏問題與複合竊聽信道問題給出了新的單發可達性結果，並涵蓋了多種既有問題作為特例。在第 5 章中，我們擴展了泊松函數表示，設計出一種用於壓縮任意差分隱私方法的新型結構，名為泊松隱私表示。泊松隱私表示是首個同時達成幾乎最佳的壓

縮性能、精確無損的輸出分佈（從而保留原始隱私機制的統計性質，如無偏性與高斯性）、並確保差分隱私的構造。我們同時也於多種應用上展示了泊松隱私表示的實驗優勢。

Acknowledgement

This dissertation would not have been possible without the support of my advisors, mentors, collaborators, colleagues, friends, family, and many other individuals. I am deeply indebted to all of them for their kind assistance throughout this pleasant journey.

First and foremost, I would like to express my deepest gratitude to my advisors, Prof. Cheuk Ting Li and Prof. Raymond W. Yeung, for their support throughout my time at CUHK. I feel honored to hold the distinguished title of being one of the first two PhD students of Prof. Li, which afforded me privileges that very few PhD students enjoy: whenever I came up with vague ideas or felt confused, Prof. Li always made time to meet with me with great patience. Most of the time, these ideas would later prove to be useless, especially at the beginning of my PhD journey, but Prof. Li was always patient enough to listen, take them seriously, and think through them with me. Such support is crucial for a young PhD student. Along with his guidance on proving theorems, writing academic papers, preparing presentations, and allowing me to explore freely in various areas of information theory, I wish to express my deepest gratitude to him. If I have the opportunity to advise students in the future, I will remember the patience and generosity I received from Prof. Li, and I hope I can inspire my students as he inspired me. I also feel privileged to be co-advised by Prof. Yeung, one of the giants in information theory. As I gained more experience, I learned

more and more from Prof. Yeung about how to embrace unexpected results in research, discover the extraordinary in the ordinary, and distill complicated findings into clear insights. His two courses laid the foundation of my knowledge in information theory. I sincerely thank both of my advisors for their outstanding guidance and support.

Throughout my PhD journey, I had the privilege of visiting Stanford University for half a year. I am grateful to Prof. Li for supporting my visit and to Prof. Ayfer Özgür for hosting me. This experience was crucial for my academic development, as I attended several workshops in California, finished an interesting project with Prof. Li and Prof. Özgür, and made many friends in the Bay Area. I am also thankful to Prof. Özgür for her guidance not only on the project we worked on but also on research methodology in general. She was so kind and supportive that I felt re-energized after every meeting with her.

I would like to extend my heartfelt thanks to Prof. Chandra Nair. Though not officially my advisor, he was like one to me. From him, I learned not only research skills but also the attitude a great scholar should possess. The first time I felt capable of working something out independently was during a project in his course. No matter how naïve my questions were, he was always willing to answer them and support my ideas. I have learned a lot from his persistence, creativity, discipline, and life philosophy.

I would also like to thank Prof. Amin Gohari for his guidance on how to explain sophisticated ideas clearly and how to ask good questions to elaborate on concepts. I am grateful to Prof. Pascal O. Vontobel for his excellent course on coding theory and his inspiring papers, from which I learned how to explain complex ideas with detailed and illustrative examples. It has been a privilege to study information theory in a department with so many experts in the field, a rarity anywhere in the world.

Moreover, I would also like to thank Prof. Vincent Y. F. Tan and Prof. Angela Yingjun Zhang for serving on my thesis committee. Their comments and suggestions are valuable to this thesis.

My PhD journey would not have been possible without the help of Prof. Shenghao Yang, who guided me for more than three years when I was an undergraduate and introduced me to the field of information theory. He taught me how to conduct good research, solve difficult problems step by step, and write first-class academic papers, even before I began my PhD studies, and he continued to support me throughout the journey. This journey would also not have been possible without Prof. Ximing Fu, who collaborated with me on a project closely and taught me many things.

I would also like to extend my thanks to professors who offered me valuable advice when I was an undergraduate deciding to pursue a PhD at CUHK, including Prof. Jianwei Huang, who provided extremely helpful advice on how to plan my PhD studies wisely; Prof. Gongqiu Zhang, who hosted my final-year project and advised me on PhD opportunities; Prof. Qin Wang, who guided me through multiple courses and provided me invaluable advices; Prof. Kenneth Shum, who encouraged me to pursue information theory despite its challenges; and Prof. Christopher Kluz and Dr. Lucas Scripter.

I would like to thank my other coauthors, each of whom taught me a lot during our collaborations, including Prof. Yi Chen, Yijun Fan, Chih Wei Ling, Wei-Ning Chen, and Sepehr Heidari Advary. Special thanks to Chih Wei and Yijun, who supported me many times when I was frustrated.

Except my coauthors, my life at CUHK would have been much more tedious without my other friends from the department: Jianguo Zhao, Xiang Li, Chin Wa (Ken) Lau, Jinpei Zhao, Zijie Chen, Zhaobang Zhu, Yi Liu, Jiaxin Qing, Xiaohong Cai, Yulin Chen, Yuwen Huang, Junda Zhou, Yicheng Cui, Binghong

Wu, Zhenduo Wen, Chenyu Wang, and Chun Hei (Michael) Shiu. Many thanks also to the friends I met at Stanford: Dan Song and Andy Dong.

I am deeply appreciative of Prof. Chak Wong, who was the most important mentor to me outside of research. He taught me how to read, how to think, how to live a happy life, and how to overcome difficulties. Whenever I feel lost, I think about what Chak shared with me and the books we read together, and I regain my passion and confidence. I cannot be more grateful.

I have been fortunate to have great friends who took care of me when I was unwell: Qingyan Chen, Xiaoyu Yue, Taolin Liu, and Yue Qin. I would also like to thank my great friends with whom I can discuss not only research but also personal life: Chengchang Liu, Guodong Li, Licheng Mao, Jie Wang, and Yanyan Dong. My deepest thanks also go to Xiao Tan for her encouragement.

Last but not least, I would like to thank Dr. Raymond Tung for his treatment, along with the other doctors, nurses, and helpers from Prince of Wales Hospital whose names I do not know. Without their careful treatment and support, I would not have been able to write this thesis. There are no words that can fully express my gratitude to them.

Lastly, I dedicate this thesis to my parents, Jiayong Liu and Baorong Ma, whose unconditional love and support have been my constant motivation. I am forever indebted to them.

List of Publications

During the PhD studies, Yanxiao Liu has published the following works:

1. Liu, Yanxiao, Sepehr Heidari Advary, and Cheuk Ting Li. "Nonasymptotic Oblivious Relaying and Variable-Length Noisy Lossy Source Coding." 2025 IEEE International Symposium on Information Theory (ISIT).

2. Liu, Yanxiao, Wei-Ning Chen, Ayfer Özgür, and Cheuk Ting Li. "Universal exact compression of differentially private mechanisms." Advances in Neural Information Processing Systems 37 (2024): 91492-91531.

3. Liu, Yanxiao, and Cheuk Ting Li. "One-Shot Information Hiding." 2024 IEEE Information Theory Workshop (ITW), pp. 169-174. IEEE, 2024. (c) 2024 IEEE. Reprinted, with permission.

4. Liu, Yanxiao, and Cheuk Ting Li. "One-shot coding over general noisy networks." 2024 IEEE International Symposium on Information Theory (ISIT), pp. 3124-3129. IEEE, 2024. Full version was submitted to IEEE Transactions on Information Theory. (c) 2024 IEEE. Reprinted, with permission.

5. Ling, Chih Wei, Yanxiao Liu, and Cheuk Ting Li. "Weighted parity-check codes for channels with state and asymmetric channels." IEEE Transactions on Information Theory (2024). Short version was presented at 2023 IEEE

International Symposium on Information Theory (ISIT), pp. 3103-3108. IEEE, 2022.

6. Yang, Shenghao, Jun Ma, and Yanxiao Liu. "Wireless network scheduling with discrete propagation delays: Theorems and algorithms." IEEE Transactions on Information Theory 70, no. 3 (2023): 1852-1875. Short version was presented at IEEE INFOCOM 2021-IEEE Conference on Computer Communications, pp. 1-10.

7. Fan, Yijun, Yanxiao Liu, Yi Chen, Shenghao Yang, and Raymond W. Yeung. "Reliable throughput of generalized collision channel without synchronization." 2023 IEEE International Symposium on Information Theory (ISIT), pp. 2697-2702. IEEE, 2023.

8. Fan, Yijun, Yanxiao Liu, and Shenghao Yang. "Continuity of link scheduling rate region for wireless networks with propagation delays." 2022 IEEE International Symposium on Information Theory (ISIT), pp. 730-735. IEEE, 2022.

This thesis covers the second, third and fourth publications.

Contents

List of Figures

Chapter 1

Introduction

1.1 Background of One-Shot Information Theory

In information theory, which originated from Shannon [155], the goal is to determine optimal and reliable transmission rates over channels, or optimal compression rates for sources. Conventional information-theoretic analyses often rely on the asymptotic equipartition property, typicality-based proofs, and the law of large numbers to characterize the behavior of channels and sources in the asymptotic regime [59].

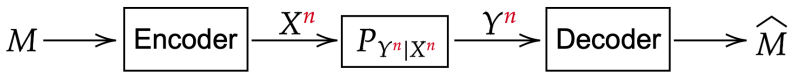

Figure 1.1: Channel coding setting in the large blocklength limit.

Take channel coding as an example. Figure 1.1 illustrates the conventional channel coding setting: a message M of length k is encoded to an input sequence $X^n = (X_1, \ldots, X_n)$; a decoder observes $Y^n = (Y_1, \ldots, Y_n)$ through a discrete

memoryless channel, and outputs $\hat{M}$. Shannon's channel coding theorem [155] states that when the blocklength n is large, i.e., $n \to \infty$, the channel capacity, which is defined as the maximum communication rate k/n in bits per channel transmissions such that $\mathbf{P}(M \neq \hat{M})$ can be made arbitrarily small [155], is given by

$$C = \max_{P_X} I(X;Y), \tag{1.1}$$

where $I(X;Y)$ is the mutual information between X and Y.

However, a critical practical issue is that packet lengths are never infinite and can, in fact, be very short in real-world applications—for example, in ultra-reliable low-latency communications [48]. Motivated by this, *finite blocklength information theory* has been extensively studied over the past decade. The goal is to provide nonasymptotic guarantees in scenarios where the number of channel uses is limited [106, 149, 165, 174]. That is, in Figure 1.1, when n is finite, what is the guarantee on the error probability $\mathbf{P}(M \neq \hat{M})$?

An even more general scenario is the *one-shot* setting [63, 88, 117, 129, 149, 156, 159, 172, 178, 189], where the blocklength is assumed to be 1. That is, each source and channel can be used only *once*. Note that "one-shot" does not mean transmitting only 1 bit. Instead, it represents the most general case, where the sources and channels can be arbitrary. This line of research is primarily motivated by the generality of the setting: no assumptions are imposed on the sources or channels (e.g., memorylessness, ergodicity, etc.). The difficulty is that well-known techniques such as joint typicality and time sharing are not applicable. This setting is more general than some finite-blocklength cases; for instance, the finite-blocklength bounds in [188] do not seem to yield one-shot results due to their use of the method of types.

We use the one-shot channel coding setting as an example in Figure 1.2. Upon observing a message $M \sim \mathrm{Unif}[1 : \mathsf{L}]$, the encoder produces X that is sent through

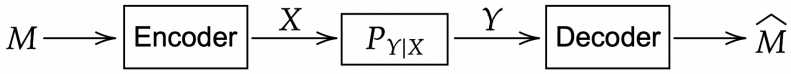

Figure 1.2: Channel coding setting in the one-shot regime.

the channel $P_{Y|X}$. The decoder observes Y and recovers $\hat{M}$ with error probability $P_e := \mathbf{P}\{M \neq \hat{M}\}$. Take the dependence testing bound by Polyanskiy, Poor and Verdú [149] as an example, we have:

$$P_e \leq \mathbf{E}\left[\min\left\{\frac{\mathsf{L}-1}{2} \cdot 2^{-\iota_{X;Y}(X;Y)}, 1\right\}\right], \tag{1.2}$$

where $\iota_{X;Y}(X;Y) := \log\left(\frac{\mathrm{d}P_{X|Y}(x|y)}{\mathrm{d}P_X(x)}\right)$ is the information density and $\frac{\mathrm{d}P_{X|Y}(x|y)}{\mathrm{d}P_X(x)} = \frac{\mathrm{d}P_{X|Y}(\cdot|y)}{\mathrm{d}P_X}(x)$ denotes the Radon-Nikodym derivative.

We show how the one-shot result (1.2) recovers the asymptotic channel capacity (1.1). Consider $\mathsf{L} = 2^{nR}$, due to the channel being discrete memoryless $P_{Y^n|X^n}(y^n|x^n) = \prod_{i=1}^n P_{Y|X}(y_i|x_i)$ in the absence of feedback, applying the one-shot result (1.2) gives

$$P_e \leq \mathbf{E}\left[\min\{(2^{nR}-1) \cdot 2^{-1-\sum_{i=1}^n \iota_{X;Y}(X_i;Y_i)}, 1\}\right] \tag{1.3}$$

where $(X_i;Y_i) \sim P_X P_{Y|X}$ i.i.d. for $i = 1, \ldots, n$.

When $n \to \infty$, by the law of large numbers we know $\sum_{i=1}^n \iota_{X;Y}(X_i;Y_i) \approx nI(X;Y)$, and therefore by (1.3) we know $P_e \to 0$ if $R < I(X;Y)$, which recovers the channel capacity C in (1.1).

One-shot settings are general, and we expect good one-shot achievability results can imply existing (first-order and second-order) asymptotic results when applied to memoryless sources and channels as above presents, or applied to systems with memory that behave ergodically [173]. For point-to-point channel coding, the achievability of the channel capacity is implied by the one-shot bounds by Feinstein [63] and Shannon [156], which are precursors of the dependence testing bound [149] in (1.2).

For settings more complex than the point-to-point channel, one-shot coding schemes have also been studied. We briefly review existing one-shot results for multi-user coding settings, and this part also appeared in [133]. In [172], one-shot versions of the covering and packing lemmas have been proposed and applied to various problems in multiuser information theory, for example, multiple access channels and broadcast channels. In [189], a proof technique based on stochastic likelihood encoders and decoders has been used to derive various one-shot achievability results in several multi-user settings, including broadcast channels, multiterminal source coding and multiple description coding. A one-shot mutual covering lemma has been proposed in [129] for broadcast channels, which recovers Marton's inner bound. In [159], the multiterminal source coding inner bound has been examined by a likelihood encoder. A finite-blocklength version of the random binning technique has been used in [188] to derive second order regions for broadcast channels. Recently, in [117], a technique called the *Poisson matching lemma* has been introduced to prove various one-shot achievability results for a range of coding settings, and the achievable one-shot bounds improve upon the best known one-shot bounds in several settings with shorter proofs. This technique has been applied to unequal message protection [100], hypothesis testing [81] and secret key generation [93]. The Poisson matching lemma is based on the *Poisson functional representation* [118], which has been applied to various fields recently, e.g., neural estimation [111], minimax learning [121] and differential privacy [132], together with other related techniques. We will provide more details on the Poisson functional representation [118] in Chapter 2.

1.2 Background of Differential Privacy

In this section we review the background of differential privacy, part of which also appeared in [132].

In modern data science and wireless communications, there is a growing dependence on large amounts of high-quality data, often generated by edge devices (e.g., photos and videos captured by smartphones, or messages hosted by social networks). However, this data inherently contains personal information, making it susceptible to privacy breaches during acquisition, collection, or utilization. For instance, despite the significant recent advancement in foundational models [20], studies have shown that these models can accidentally memorize their training data. This poses a risk where malicious users, even with just API access, can extract substantial portions of sensitive information [27, 28].

In recent years, differential privacy (DP) [51] has emerged as a powerful framework for safeguarding users' privacy by ensuring that local data is properly randomized before leaving users' devices. With the local data X, a DP mechanism $\mathcal{A}$ satisfies (ϵ, δ)-DP maps X to the output $Z = \mathcal{A}(X) \in \mathcal{Z}$, where $\mathcal{A}(\cdot)$ is a randomized function, such that for any neighboring (x, x') and $\mathcal{S} \subseteq \mathcal{Z}$,

$$\Pr(Z \in \mathcal{S} \,|\, X = x) \leq e^{\varepsilon} \Pr(Z \in \mathcal{S} \,|\, X = x') + \delta, \tag{1.4}$$

where neighboring (x, x') are neighboring if they differ in a single data point. This definition can be understood as follows: A differentially private mechanism (satisfying (1.4)) guarantees that small changes in its input lead to only insignificant changes in its output. If condition (1.4) is violated, an adversary could infer whether specific data was included in the input. At a high level, differential privacy prevents attackers from gaining significant knowledge about the input by observing changes in the output [51].

Apart from privacy concerns, communicating local data from edge devices to

the central server often becomes a bottleneck in the system pipeline, especially with high-dimensional data common in many machine learning scenarios. This leads to the following fundamental question: how can we efficiently communicate differentially privatized data?

Recent works have shown that a wide range of differential privacy mechanisms can be "simulated" and "compressed" using shared randomness, resulting in a "compressed mechanism" which has a smaller communication cost compared to the original mechanism, while retaining the (perhaps slightly weakened) privacy guarantee. This can be done via rejection sampling [65], importance sampling [153, 168], or dithered quantization [85, 91, 108, 154, 185] with each approach having its own advantages and disadvantages. For example, importance-sampling-based methods [153, 168] and the rejection-sampling-based method [65] can simulate a wide range of privacy mechanisms; however, the output distribution of the induced mechanism does not perfectly match the original mechanism. This is limiting in scenarios where the original mechanism is designed to satisfy some desired statistical properties, e.g. it is often desirable for the local randomizer to be unbiased or to be "summable" noise such as Gaussian or other infinitely divisible distributions. Since the induced mechanism is different from the original one, these statistical properties are not preserved. On the other hand, dithered-quantization-based approaches [85, 91, 92, 108, 154, 185] can ensure a correct simulated distribution, but they can only simulate additive noise mechanisms. More importantly, dithered quantization relies on shared randomness between the user and the server, and the server needs to know the dither for decoding. This annuls the local privacy guarantee on the user data, unless we are willing to assume a trusted aggregator [85], use an additional secure aggregation step [91], or restrict attention to specific privacy mechanisms (e.g., one-dimensional Laplace [154]).

1.3 One-Shot Codes Meet Differential Privacy

In this section, we discuss the interplay between information theory—specifically, one-shot coding schemes—and differential privacy, which motivates the studies presented in this thesis. Our goal is to develop a unified one-shot coding framework that is applicable to both network information theory problems and differential privacy mechanisms. We emphasize that our focus is on information-theoretic coding schemes, rather than information-theoretic measures of privacy. For discussions on the latter, we refer the reader to [170].

As discussed in Section 1.1, one-shot information theory addresses the most fundamental setting, where no assumptions are made about the signal length or the structure of the channel/source. Similar to the asymptotic analyses in network information theory [59], the goal remains to investigate the fundamental limits of signal transmission over noisy channels or source compression. To ensure reliable message reconstruction—and thus high accuracy for downstream tasks—effective one-shot coding schemes must be constructed to withstand noise. In this thesis, we present various one-shot codes based on early works on Poisson functional representation [117, 118].

However, when considering privacy, the requirement may initially appear to conflict with the goal of communication: achieving differential privacy typically involves deliberately adding noise, which increases the message entropy and reduces its compressibility. This tension is sometimes referred to as the "communication-privacy-accuracy trilemma" [33]. Nevertheless, it has been shown that through careful encoding and controlled noise injection for privacy, it is possible to simultaneously achieve communication efficiency and privacy, while still maintaining accuracy for various tasks [2, 10, 33, 65]. In this sense, noise is *utilized* as a tool for both compression and privacy.

One might initially think that controlling noise is difficult. However, the idea

of using random noise for source coding dates back to [190, 191], which proposed additive noise as a tool for universally good lossy compression schemes. These works represent early-stage studies of *channel simulation*. Channel simulation, also known as *channel synthesis* or *reverse channel coding*, aims to *simulate* a noisy channel using as few communication bits as possible [13, 40]. This approach has the potential to simultaneously achieve communication efficiency and privacy. We will provide a more detailed review of channel simulation in Chapter 5. Various channel simulation schemes—such as dithered quantization [191], rejection sampling [82], and importance sampling (or more specifically, minimal random coding) [86]—have been employed to compress differential privacy mechanisms [25, 65, 92, 153, 154, 168, 185]. Readers are referred to [119] for a comprehensive review.

Though we will first discuss one-shot codes based on the Poisson functional representation [117, 118] for various network information theory problems, it is worth noting that the Poisson functional representation is also a good channel simulation scheme [118] (with some improvements in analysis found in [113, 115, 117]). With unlimited common randomness, it can provide the smallest known bound on the expected length for one-shot channel simulation. However, it is a deterministic mapping: the input, together with the common randomness, deterministically determines the output. As a result, a small change in the input (in the sense of differential privacy) can lead to a deterministic change in the compressed output, making this method unsuitable for directly ensuring privacy. To address this issue, we propose a way to *randomize* the Poisson functional representation. This randomized variant will be shown to preserve differential privacy while achieving a compression size close to the optimal.

In summary, though at first glance the goals of privacy protection and efficient communication may appear to be in conflict, recent works have shown that

this "paradox" can be resolved, and channel simulation emerges as a promising candidate for achieving both. The various one-shot codes proposed in this thesis are based on the Poisson functional representation [118], which is also a state-of-the-art channel simulation scheme. Although its deterministic nature poses challenges for direct application to differential privacy mechanisms, we show that it can be *randomized* to provide privacy protection. From this unified perspective, the Poisson functional representation serves as a bridge between one-shot information theory and differential privacy. We note that while importance sampling has also been used in both network information theory [147] and differential privacy [153, 168], it is *not exact*; that is, the output distribution is distorted, resulting in only approximate simulation. We will elaborate on this distinction in Chapter 5.

1.4 Our Contributions

1.4.1 Contributions in Chapter 3

In Chapter 3, we present a unified one-shot coding framework designed for the communication and compression of messages among multiple nodes across a general acyclic noisy network. Our setting can be seen as a one-shot version of the acyclic discrete memoryless network studied by Lee and Chung [110], and noisy network coding studied by Lim, Kim, El Gamal and Chung [126]. We design a proof technique, called the exponential process refinement lemma, that is rooted in the Poisson matching lemma by Li and Anantharam, and can significantly simplify the analyses of one-shot coding over multi-hop networks. Our one-shot coding theorem not only recovers a wide range of existing asymptotic results, but also yields novel one-shot achievability results in different multi-hop network information theory problems, such as compress-and-forward and partial-decode-

and-forward bounds for a one-shot (primitive) relay channel, and a bound for one-shot cascade multiterminal source coding. In a broader context, our framework provides a unified one-shot bound applicable to any combination of source coding, channel coding and coding for computing problems. This chapter is based on [133].

1.4.2 Contributions in Chapter 4

In Chapter 4, we present one-shot information-theoretic analyses of two secrecy problems: a generalization of the information hiding problem [144] and the compound wiretap channel [123]. The former admits a game-theoretic formulation, where one party (the information hider and decoder) seeks to embed secret messages into a host signal for later reconstruction, while the opposing party (an attacker) attempts to remove or degrade the embedded information. The latter generalizes Wyner's wiretap channel by allowing multiple potential channel states, making it more suitable for the rapidly changing characteristics of modern wireless communications. Although these two secrecy problems seem unrelated, we study both utilizing a covering argument and similar techniques under a unified framework. We derive one-shot achievability results for both problems using techniques based on the Poisson matching lemma, which enables us to handle both discrete and continuous cases. We show that our one-shot results readily recover existing asymptotic results. Unlike previous asymptotic results, ours apply to any source distribution and any class of channels, not necessarily memoryless or ergodic. This chapter is partially based on [134].

1.4.3 Contributions in Chapter 5

In Chapter 5, we introduce a novel construction, called Poisson private representation (PPR), designed to compress and simulate any local randomizer while ensuring local differential privacy, hence reduce the communication cost of differential privacy mechanisms, Unlike previous simulation-based local differential privacy mechanisms, PPR exactly preserves the joint distribution of the data and the output of the original local randomizer. Hence, the PPR-compressed privacy mechanism retains all desirable statistical properties of the original privacy mechanism such as unbiasedness and Gaussianity. Moreover, PPR achieves a compression size within a logarithmic gap from the theoretical lower bound. Using the PPR, we give a new order-wise trade-off between communication, accuracy, central and local differential privacy for distributed mean estimation. Experiment results on distributed mean estimation show that PPR consistently gives a better trade-off between communication, accuracy and central differential privacy compared to the coordinate subsampled Gaussian mechanism, while also providing local differential privacy. This chapter is based on [132].

Chapter 2

Poisson Functional Representation

In this chapter, we review the Poisson functional representation [118] and discuss some technical background on related schemes. We begin by introducing our notations.

2.1 Notations

We assume the logarithm and entropy are to the base 2 unless otherwise stated, and logarithm to the base e is denoted as $\ln(x)$. For a statement S, we use $\mathbf{1}\{S\}$ to denote its indicator, i.e., $\mathbf{1}\{S\}$ is 1 if S holds and otherwise $\mathbf{1}\{S\} = 0$. δ_a denotes the degenerate distribution $\mathbf{P}\{X = a\} = 1$.

We use $[i..j]$ to denote $\{i, i+1, \ldots, j\}$ and $[j] := [1..j]$. For a set $\mathcal{S} \subseteq [k]$ and random sequence $U_1, \ldots, U_k$, we write $U^k := (U_1, \ldots, U_k)$, $U_{\mathcal{S}} := (U_j)_{j\in\mathcal{S}}$. For two random variables X, Y, the information density is defined as

$$\iota_{X;Y}(x;y) = \log\left(\frac{\mathrm{d}P_{X|Y}(x|y)}{\mathrm{d}P_X(x)}\right),$$

where $\frac{\mathrm{d}P_{X|Y}(x|y)}{\mathrm{d}P_X(x)}$ denotes the Radon-Nikodym derivative. For two random variables X, Y, we sometimes omit the subscript and write $\iota(X;Y)$ instead of $\iota_{X;Y}(X;Y)$ if the random variables are clear from the context. In discrete case, the conditional information density is defined to be

$$\iota_{X;Y|Z}(x;y|z) := \log\left(\frac{P_{X,Y|Z}(x,y|z)}{P_{X|Z}(x|z)P_{Y|Z}(y|z)}\right),$$

The total variation (TV) distance between two distributions P, Q over $\mathcal{X}$ is $\|P-Q\|_{\mathrm{TV}} := \sup_{A\subseteq\mathcal{X} \text{ measurable}} |P(A)-Q(A)|$.

For two distributions P and Q, we write $P \ll Q$ to denote that P is absolutely continuous with respect to Q.

2.2 Poisson Functional Representation

In this section, we introduce the Poisson Functional Representation [118] and the Poisson Matching Lemma [117], which serve as the building blocks of this thesis.

The Poisson functional representation was introduced in [118] as a channel simulation scheme, where a strong functional representation lemma is proved. Related constructions for Monte Carlo simulations can be found in [135]. Together with other techniques, the Poisson functional representation [118] has been applied to various fields, including neural estimation [111] and minimax learning [120]. As discussed in [70], the Poisson functional representation can be viewed as a certain variant of the A* sampling [135, 136], and hence an optimized version with better runtime for one-dimensional unimodal distribution has been proposed in [70]. A greedy-search version can be found in [68].

Based on the Poisson functional representation, the Poisson Matching Lemma was proposed in [117], and it has been shown to improve upon previously known one-shot bounds in various settings with simpler analyses. Recent applications of

the Poisson Matching Lemma include unequal message protection [100], hypothesis testing [81], and secret key generation [93].

We start with the discrete case, which we refer to as the exponential functional representation [118].

Definition 2.2.1 (Exponential Functional Representation [118])**.** Consider a finite set $\mathcal{U}$. Let $\mathbf{U} := (Z_u)_{u\in\mathcal{U}}$ be i.i.d. Exp(1) random variables.[1] Given a distribution P over $\mathcal{U}$,

$$\mathbf{U}_P := \operatorname{argmin}_u \frac{Z_u}{P(u)} \tag{2.1}$$

is called the exponential functional representation [118].

By [118], we have $\mathbf{U}_P \sim P$.

The exponential functional representation [118] is designed for finite alphabets, which is the case in Chapter 3. When the space is continuous, as in Chapter 4 and Chapter 5, a generalization via Poisson processes is utilized [117, 118]. Further discussions and detailed derivations of the connection between the two cases can be found in [117, 119]; we omit them here. We introduce the generalization, called the Poisson functional representation [118], as follows.

Definition 2.2.2 (Poisson Functional Representation [118])**.** Let $(T_i)_i$ be a Poisson process with rate 1 (i.e., $T_1, T_2 - T_1, T_3 - T_2, \ldots \overset{iid}{\sim} \text{Exp}(1)$), independent of $\bar{U}_i \overset{iid}{\sim} Q$ for $i = 1, 2, \ldots$, and we denote $\mathbf{U} := (\bar{U}_i)_i$. $(\bar{U}_i, T_i)_i$ is a Poisson process with intensity measure $Q \times \lambda_{[0,\infty)}$ [109], where $\lambda_{[0,\infty)}$ is the Lebesgue measure over $[0, \infty)$. Fix any distribution P over $\mathcal{U}$ that is absolutely continuous with respect to Q. Let

$$\tilde{T}_i := T_i \cdot \left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\bar{U}_i)\right)^{-1}, \tag{2.2}$$

where $\frac{\mathrm{d}P}{\mathrm{d}Q}(\cdot)$ is the Radon-Nikodym derivative. By the mapping theorem [109], $(\bar{U}_i, \tilde{T}_i)$ is a Poisson process with intensity measure $P \times \lambda_{[0,\infty)}$. Then the Poisson

[1] Exp(1) random variables follow an exponential distribution with rate parameter 1.

functional representation [118] selects

$$\mathbf{U}_P := \bar{U}_K,$$

where

$$K := \text{argmin}_i \tilde{T}_i.$$

Note that since the T_i's are continuous, with probability 1, there do not exist two equal values among $\tilde{T}_i$'s. The Poisson functional representation [118] holds for general Q which may be discrete or continuous.

The Poisson functional representation [118] selects a sample following the target distribution P among samples from another distribution Q, i.e., $\mathbf{U}_P \sim P$. It draws a random sequence $(\bar{U}_i)_i$ from Q and a sequence of times $(T_i)_i$ according to a Poisson process. If we select the sample $\bar{U}_i$ with the smallest T_i, then the selected sample will follow distribution Q. To obtain a sample from P instead, we multiply the time by the factor $(\frac{\mathrm{d}P}{\mathrm{d}Q}(\bar{U}_i))^{-1}$ in (2.2) to give $\tilde{T}_i$, so the $\bar{U}_i$ with the smallest $\tilde{T}_i$ will follow P.

The Poisson functional representation [118] was originally developed to prove the strong functional representation lemma, and possibly tighter guarantees via different analyses can be found in [113, 115].

The way this Poisson process is used in communication settings (e.g., in [117]) is that the encoder would query the process using the prior distribution of the signal to obtain the signal to be sent, and the decoder would query using the posterior distribution of the signal given the noisy observation to obtain the message. There is no error in the communication if the two queries return the same sample. The probability of error can be bounded by the Poisson matching lemma [117], which will be discussed in Section 2.3.

2.3 Poisson Matching Lemma

In this section, we introduce a technique that is based on the Poisson Functional Representation, called the Poisson matching lemma [117], which has been shown to be able to provide good one-shot achievability results on a large class of network information theory problems [117].

The Poisson matching lemma has been shown to be quite useful in proving one-shot achievability results of network information theory [117, 133]. It is rooted in the Poisson functional representation [118] that is reviewed as follows.

Lemma 2.3.1 (Poisson matching lemma [117])**.** *Consider two distributions $P_1, P_2 \ll Q$. Almost surely, we have*

$$\mathbf{P}\big(\mathbf{U}_{P_2} \neq \mathbf{U}_{P_1} \,\big|\, \mathbf{U}_{P_1}\big) \leq 1 - \Big(1 + \frac{\mathrm{d}P_1}{\mathrm{d}P_2}(\mathbf{U}_{P_1})\Big)^{-1}.$$

The Poisson matching lemma [117] provides a bound on the probability of mismatch between the Poisson functional representations applied on different distributions. Various information theory problems have been studied by using the Poisson matching lemma [117]. In chapter 3, we will extend it to a tool that can provide one-shot achievability results over arbitrary acyclic noisy networks, and hence recover many one-shot results in [117].

2.4 Discussions on Other Existing Techniques

Compared to the one-shot coding scheme in [189], the Poisson matching lemma utilizes a Poisson process to create a codebook, instead of the conventional i.i.d. random codebook [189], and each codeword is assigned a bias T_i. The scheme is thus a biased maximum likelihood decoder, rather than a stochastic decoder as in [189]. The idea of using a biased, or *soft*, coding scheme has been extended to linear codes, known as *weighted parity-check codes*; see [127, 128].

Except for the one-shot coding scheme based on Poisson functional representation [117, 118], other unified frameworks for one-shot coding include the schemes based on random binning [187, 188, 189], the likelihood encoder [159], and importance sampling [147]. Other one-shot coding schemes [63, 88, 156, 172, 178] have been reviewed in detail in Chapter 1; here, we discuss some connections between one-shot codes and channel simulation via the likelihood encoder [159] and importance sampling [147], since we will utilize channel simulation to compress differential privacy mechanisms in Chapter 5.

As mentioned above, the Poisson functional representation can be viewed as a variant of A* sampling [135, 136], and a very useful property is that the output sample follows *exactly* the input distribution. This can also be understood as a remote sampling problem; we refer readers to [119] for a comprehensive explanation. The Poisson functional representation can be used to prove the strong functional representation lemma [118]: to simulate a channel $P_{Y|X}$, the communication cost (expected number of bits) required is bounded by

$$I(X;Y) + \log\big(I(X;Y)+1\big) + 5,$$

which with some finer analyses can be improved to$I(X;Y) + \log\big(I(X;Y)+2\big) + 3$ [113, 115].

If one considers greedy rejection sampling for channel simulation, a weaker guarantee has been proved by [83], and improved by [22], as follows:

$$I(X;Y) + \log\big(I(X;Y)+1\big) + c,$$

where c is an unspecified constant. The guarantee by using greedy rejection sampling was later improved by [72] to

$$I(X;Y) + \log\big(I(X;Y) + \log(4e)\big) + \log(4e) + 1.$$

Another popular sampling scheme is importance sampling, which was considered by [44] for asymptotic channel simulation, and later dubbed the *likelihood*

encoder [159] for one-shot coding. In machine learning, a similar scheme named *minimal random coding* was also studied by [86], where it was applied to model compression and later to lossy image compression [69]. Besides the likelihood encoder, a recent work [147] used importance sampling (an *importance matching lemma* that shares some similarity with the Poisson matching lemma [117]) for coding in information theory, and it has the potential to be extended to other information theory problems. We would like to emphasize here that a major difference compared to the Poisson functional representation is that the output sample does not exactly follow the input distribution.

This difference also appears in the study of compressing differential privacy mechanisms. Compression of differential privacy mechanisms can be viewed as a channel simulation problem, where the channel is subject to an additional privacy constraint. Importance sampling (or more specifically, minimal random coding [86]) has been used for compressing differential privacy mechanisms [153, 168]. However, as mentioned in the previous paragraph, since minimal random coding is not exact, the output distribution is only approximate. On the other hand, rejection sampling can also be utilized [25, 65], although the communication cost and privacy guarantees were not close to optimal. In Chapter 5, we will utilize a variant of the Poisson functional representation [118] to compress differential privacy mechanisms.

In summary, importance sampling has been applied in both one-shot coding and the compression of differential privacy mechanisms. In contrast, the Poisson functional representation [118] offers an exact simulation framework with close-to-optimal communication cost guarantees. In the following chapter, we will demonstrate how to extend the construction of the Poisson functional representation to various problems, ensuring good performance.

Chapter 3

One-Shot Coding over General Noisy Networks

3.1 Overview

In this chapter, we study a general class of networks, which we call *acyclic discrete networks*, where there are N nodes connected by noisy channels in an acyclic manner. Each node can play the role of an encoder or a decoder (or both) in source coding or channel coding settings. This is a one-shot version of the asymptotic acyclic discrete memoryless network studied by Lee and Chung [110], and includes a wide range of settings as special cases, such as source and channel coding, primitive relay channel [55, 56, 59, 101, 142], Gelfand-Pinsker [74, 89], relay-with-unlimited-look-ahead [57, 58], Wyner-Ziv [180, 182], coding for computing [184], multiple access channels [4, 5, 125], broadcast channels [138] and cascade multiterminal source coding [42]. In a broader context, our one-shot achievability results are general enough to be applicable to any combination of source coding, channel coding and coding for computing problems. This chapter is based on [133].

In order to alleviate the difficulty of keeping track of a large number of auxiliary random variables in a general N-node network, we propose a tool called the *exponential process refinement lemma* based on the Poisson matching lemma [117],[1] which simplifies the analyses of the evolution of the posterior distribution of the sources, messages and/or auxiliary random variables at the decoder. We utilize the lemma to prove a one-shot achievability result for general acyclic discrete networks, which recovers existing one-shot results in a range of settings in [117, 172, 178, 189], and also give novel one-shot results for various multi-hop settings, namely primitive relay channels [55, 56, 59, 101, 142], relay-with-unlimited-look-ahead [57, 58], and cascade multiterminal source coding [42].

The chapter is organized as follows. We present our proof technique, called the exponential process refinement lemma, in Section 3.2. We describe our general acyclic discrete network in Section 3.3, and prove our main theorem in Section 3.4. In Section 3.5, we use a one-shot relay channel and related settings to elaborate our coding scheme in detail. We then discuss a novel one-shot cascade multiterminal source coding problem in Section 3.6. We also show our coding scheme provides one-shot bounds on various network information theory settings in Section 3.7.

3.2 Exponential Process Refinement Lemma

Recall we have introduced the Exponential functional representation [118] in Chapter 2. We briefly review it here together with the Poisson matching lemma [117]

[1] We only present the discrete case in this chapter for the sake of simplicity. Hence, instead of Poisson processes, we may use an i.i.d. exponential processes instead [118]. While we expect the results to be extended to the continuous case, this is left for future studies. For the use of Poisson functional representation in continuous case in other settings, see Chapter 4 and Chapter 5.

for the sake of completeness. We then design a tool for proving one-shot achievability results over noisy multi-hop networks based on the Poisson matching lemma, called the *exponential process refinement lemma*.

Consider a finite set $\mathcal{U}$. Let $\mathbf{U} := (Z_u)_{u\in\mathcal{U}}$ be i.i.d. Exp(1) random variables.[2] Given a distribution P over $\mathcal{U}$,

$$\mathbf{U}_P := \operatorname{argmin}_u \frac{Z_u}{P(u)} \tag{3.1}$$

is called the exponential functional representation in Chapter 2.[3]

As explained in Chapter 2 and also [118], we have $\mathbf{U}_P \sim P$.

We can generalize this by letting $\mathbf{U}_P(1), \ldots, \mathbf{U}_P(|\mathcal{U}|) \in \mathcal{U}$ be the elements of $\mathcal{U}$ sorted in ascending order of $Z_u/P(u)$:

$$\frac{Z_{\mathbf{U}_P(1)}}{P(\mathbf{U}_P(1))} \le \cdots \le \frac{Z_{\mathbf{U}_P(|\mathcal{U}|)}}{P(\mathbf{U}_P(|\mathcal{U}|))}.$$

We break ties arbitrarily and treat $1/0 = \infty$. This is similar to the mapped Poisson process in the generalized Poisson matching lemma [117], though unlike [117], $\mathbf{U}_P(1), \ldots, \mathbf{U}_P(|\mathcal{U}|)$ is not an i.i.d. sequence following P. Write $\mathbf{U}_P^{-1} : \mathcal{U} \to [|\mathcal{U}|]$ for the inverse function of $i \mapsto \mathbf{U}_P(i)$. The following is a direct corollary of the generalized Poisson matching lemma [117].

Lemma 3.2.1. *For distributions P, Q over $\mathcal{U}$, we have the following almost surely:*

$$\mathbf{E}\left[\mathbf{U}_Q^{-1}(\mathbf{U}_P) \,\middle|\, \mathbf{U}_P\right] \le \frac{P(\mathbf{U}_P)}{Q(\mathbf{U}_P)} + 1.$$

We now define a convenient tool.

[2] When the space $\mathcal{U}$ is continuous, a Poisson process is used in [117, 118].

[3] In [118], even for discrete case using exponential random variables, (3.1) was still called the *Poisson* functional representation. Here and similar to [119] we call it exponential functional representation to distinguish it with the Poisson functional representation that works for continuous cases in Chapter 4 and Chapter 5.

Definition 3.2.2 (Refining a distribution by an exponential process)**.** For a joint distribution $Q_{V,U}$ over $\mathcal{V} \times \mathcal{U}$, the refinement of $Q_{V,U}$ by $\mathbf{U}$, denoted as $Q_{V,U}^{\mathbf{U}}$, is a joint distribution

$$Q_{V,U}^{\mathbf{U}}(v,u) := \frac{Q_V(v)}{\mathbf{U}_{Q_{U|V}(\cdot|v)}^{-1}(u) \sum_{i=1}^{|\mathcal{U}|} i^{-1}}$$

for all (v,u) in the support of $Q_{V,U}$, where Q_V is the V-marginal of $Q_{V,U}$ and $Q_{U|V}$ is the conditional distribution of U given V. When $V = \emptyset$, the above definition becomes

$$Q_U^{\mathbf{U}}(u) = \frac{1}{\mathbf{U}_{Q_U}^{-1}(u) \sum_{i=1}^{|\mathcal{U}|} i^{-1}}.$$

While the exponential functional representation $\mathbf{U}_{Q_U}$ (which only gives one value of U) is used for the unique decoding of U, the refinement $Q_U^{\mathbf{U}}(u)$ is for the *soft decoding* of U, which gives a distribution over U, with $\mathbf{U}_{Q_U}$ having the largest probability. This is useful in non-unique decoding. For example, if we want to decode U_1 uniquely, while utilizing U_2 via non-unique decoding, we can first obtain the distribution $(\mathbf{U}_2)_{Q_{U_2}}$, and then compute the marginal distribution of U_1 in $(\mathbf{U}_2)_{Q_{U_2}} P_{U_1|U_2}$ and use this marginal distribution to recover U_1 via the exponential functional representation.

Loosely speaking, if the distribution $Q_{V,U}$ represents our "prior distribution" of (V,U), then the refinement $Q_{V,U}^{\mathbf{U}}$ is our updated "posterior distribution" after taking the exponential process $\mathbf{U}$ into account. In multiterminal coding settings that a node decodes multiple random variables, the prior distribution of those random variables will be refined by multiple exponential processes. To keep track of the evolution of the "posterior probability" of the correct values of those random variables through the refinement process, we use the following lemma, called the *exponential process refinement lemma*. Although its proof still relies on the Poisson matching lemma [117], it significantly simplifies our analyses.

Lemma 3.2.3 (Exponential Process Refinement Lemma)**.** *For a distribution P over $\mathcal{U}$ and a joint distribution $Q_{V,U}$ over a finite $\mathcal{V} \times \mathcal{U}$, for every $v \in \mathcal{V}$, we have, almost surely,*

$$\mathbf{E}\left[\frac{1}{Q^{\mathbf{U}}_{V,U}(v, \mathbf{U}_P)}\middle|\mathbf{U}_P\right] \leq \frac{\ln|\mathcal{U}|+1}{Q_V(v)}\left(\frac{P(\mathbf{U}_P)}{Q_{U|V}(\mathbf{U}_P|v)}+1\right).$$

Proof. We have

$$\begin{aligned}
&\mathbf{E}\left[\frac{1}{Q^{\mathbf{U}}_{V,U}(v, \mathbf{U}_P)}\,\middle|\,\mathbf{U}_P\right] \\
&\overset{(a)}{=} \mathbf{E}\left[\frac{\mathbf{U}^{-1}_{Q_{U|V}(\cdot|v)}(\mathbf{U}_P)\sum_{i=1}^{|\mathcal{U}|} i^{-1}}{Q_V(v)}\middle|\mathbf{U}_P\right] \\
&\overset{(b)}{\leq} \frac{\sum_{i=1}^{|\mathcal{U}|} i^{-1}}{Q_V(v)}\left(\frac{P(\mathbf{U}_P)}{Q_{U|V}(\mathbf{U}_P|v)}+1\right) \\
&\overset{(c)}{\leq} \frac{\ln|\mathcal{U}|+1}{Q_V(v)}\left(\frac{P(\mathbf{U}_P)}{Q_{U|V}(\mathbf{U}_P|v)}+1\right),
\end{aligned}$$

where (a) is by Definition 3.2.2, (b) is by Lemma 3.2.1 and (c) is by $\sum_{i=1}^{n} i^{-1} \leq \int_1^n x^{-1}dx + 1 = \ln n + 1$. □

3.3 Network Model

We describe a general N-node network model, which is the one-shot version of the *acyclic discrete memoryless network (ADMN)* [110]. There are N nodes labelled $1, \ldots, N$. Node i observes $Y_i \in \mathcal{Y}_i$ and produces $X_i \in \mathcal{X}_i$ (while we assume $\mathcal{X}_i, \mathcal{Y}_i$ are finite). Unlike conventional asymptotic settings (e.g. [110]), here X_i is only *one symbol*, instead of a sequence $(X_{i,1}, \ldots, X_{i,n})$. The transmission is performed sequentially, and each Y_i is allowed to depend on all previous inputs and outputs (i.e., X^{i-1}, Y^{i-1}) in a stochastic manner, as shown in Figure 3.1. Therefore, we can formally define an *N-node acyclic discrete network (ADN)* as a collection of channels $(P_{Y_i|X^{i-1},Y^{i-1}})_{i\in[N]}$, where $P_{Y_i|X^{i-1},Y^{i-1}}$ is a conditional distribution from

$(\prod_{j=1}^{i-1} \mathcal{X}_j) \times (\prod_{j=1}^{i-1} \mathcal{Y}_j)$ to $\mathcal{Y}_i$. In particular, Y_1 follows P_{Y_1} and does not depend on any other random variable. The asymptotic ADMN [110] can be considered as the n-fold ADN $(P^n_{Y_i|X^{i-1},Y^{i-1}})_{i\in[N]}$, where $P^n_{Y_i|X^{i-1},Y^{i-1}}$ denotes the n-fold product conditional distribution (i.e., n copies of a memoryless channel), and we take the blocklength $n \to \infty$.

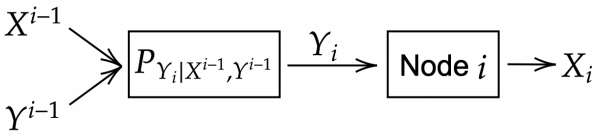

Figure 3.1: Acyclic discrete memoryless network.

We remark that, similar to the asymptotic unified random coding bound [110], the X_i's and Y_i's can represent sources, states, channel inputs, outputs and messages in source coding and channel coding settings. For example, for point-to-point channel coding, we take Y_1 to be the message, which the encoder (node 1) encodes into the channel input X_1, which in turn is sent through the channel $P_{Y_2|X_1}$. The decoder (node 2) observes Y_2 and outputs X_2, which is the decoded message. For lossless source coding, Y_1 is the source, $X_1 = Y_2$ is the description by the encoder, and X_2 is the reconstruction.

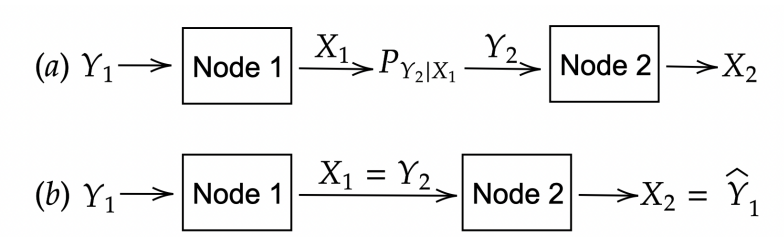

Figure 3.2: (a) Channel coding. (b) Source coding.

We give the definition of a coding scheme below.

Definition 3.3.1. A *deterministic coding scheme* consists of a sequence of en-

coding functions $(f_i)_{i\in[N]}$, where $f_i : \mathcal{Y}_i \to \mathcal{X}_i$. For $i = 1, \ldots, N$, the following operations are performed:

- **Noisy channel.** The output $\tilde{Y}_i$ is generated conditional on $\tilde{X}^{i-1}, \tilde{Y}^{i-1}$ according to $P_{Y_i|X^{i-1},Y^{i-1}}$. For $i = 1$, $\tilde{Y}_1 \sim P_{Y_1}$ can be regarded as a source or a channel state.

- **Node operation.** Node i observes $\tilde{Y}_i$ and outputs $\tilde{X}_i = f_i(\tilde{Y}_i)$.

We sometimes allow an additional unlimited public randomness available to all nodes.

Definition 3.3.2. A *public-randomness coding scheme* for the network consists of a pair $(P_W, (f_i)_{i\in[N]})$, where P_W is the distribution of the public randomness $W \in \mathcal{W}$ available to all nodes and $f_i : \mathcal{Y}_i \times \mathcal{W} \to \mathcal{X}_i$ is the encoding function of node i mapping its observation Y_i and the public randomness W to its output X_i. The operations are as follows. First, generate $W \sim P_W$. For $i = 1, \ldots, N$, generate $\tilde{Y}_i$ conditional on $\tilde{X}^{i-1}, \tilde{Y}^{i-1}$ according to $P_{Y_i|X^{i-1},Y^{i-1}}$, and take $\tilde{X}_i = f_i(\tilde{Y}_i, W)$.

We do not impose any constraint on the public randomness W. In reality, to carry out a public-randomness coding scheme, the nodes share a common random seed to initialize their pseudorandom number generators before the scheme commences.

We use $\tilde{X}_i, \tilde{Y}_i$ to denote the actual random variables from the coding scheme. In contrast, X_i, Y_i usually denote the random variables following an ideal distribution. For example, in channel coding, the ideal distribution is $Y_1 = X_2 \sim \mathrm{Unif}[\mathsf{L}]$ (i.e., the message is decoded without error), independent of $(X_1, Y_2) \sim P_{X_1} P_{Y_2|X_1}$. If we ensure that the actual $\tilde{X}^2, \tilde{Y}^2$ is "close to" the ideal X^2, Y^2, this would imply that $\tilde{Y}_1 = \tilde{X}_2$ with high probability as well, giving a small error probability.

The goal (the "achievability") is to make the actual joint distribution $P_{\tilde{X}^N,\tilde{Y}^N}$ "approximately as good as" the ideal joint distribution P_{X^N,Y^N}. If we have an "error set" $\mathcal{E} \subseteq \left(\prod_{i=1}^N \mathcal{X}_i\right) \times \left(\prod_{i=1}^N \mathcal{Y}_i\right)$ that we do not want $(\tilde{X}^N, \tilde{Y}^N)$ to fall into (e.g., for channel coding, $\mathcal{E}$ is the set where $\tilde{Y}_1 \neq \tilde{X}_2$, i.e., an error occurs; for lossy source coding, $\mathcal{E}$ is the set where $d(\tilde{Y}_1, \tilde{X}_2) > \mathsf{D}$, i.e., the distortion exceeds the limit), we want

$$\mathbf{P}\big((\tilde{X}^N, \tilde{Y}^N) \in \mathcal{E}\big) \lesssim \mathbf{P}\big((X^N, Y^N) \in \mathcal{E}\big). \tag{3.2}$$

If $P_{\tilde{X}^N,\tilde{Y}^N}$ is close to P_{X^N,Y^N} in total variation distance, i.e.,

$$\delta_{\mathrm{TV}}\big(P_{X^N,Y^N},\, P_{\tilde{X}^N,\tilde{Y}^N}\big) \approx 0, \tag{3.3}$$

then (3.2) is guaranteed. For public-randomness coding, we show that (3.3) can be achieved, which can be seen as a channel simulation [13, 40] or a coordination [41] result. For deterministic coding, since the node operations are deterministic, there might not be sufficient randomness to make $P_{\tilde{X}^N,\tilde{Y}^N}$ close to P_{X^N,Y^N}, and hence we use the error bound in (3.2).

3.4 Main Theorem for Acyclic Discrete Networks

We show a one-shot achievability result for ADN via public-randomness coding scheme.

Theorem 3.4.1. *Fix any ADN $(P_{Y_i|X^{i-1},Y^{i-1}})_{i\in[N]}$. For any collection of indices $(a_{i,j})_{i\in[N],j\in[d_i]}$ where $(a_{i,j})_{j\in[d_i]}$ is a sequence of distinct indices in $[i-1]$ for each i, any sequence $(d_i')_{i\in[N]}$ with $0 \le d_i' \le d_i$ and any collection of conditional distributions $(P_{U_i|Y_i,\overline{U}_i'}, P_{X_i|Y_i,U_i,\overline{U}_i'})_{i\in[N]}$ (where $\overline{U}_{i,\mathcal{S}} := (U_{a_{i,j}})_{j\in\mathcal{S}}$ for $\mathcal{S} \subseteq [d_i]$ and $\overline{U}_i' := \overline{U}_{i,[d_i']}$), which induces the joint distribution of X^N, Y^N, U^N (the "ideal distribution"), there exists a public-randomness coding scheme $(P_W, (f_i)_{i\in[N]})$ such*

that the joint distribution of $\tilde{X}^N, \tilde{Y}^N$ induced by the scheme (the "actual distribution") satisfies

$$\delta_{\mathrm{TV}}\big(P_{X^N,Y^N},\, P_{\tilde{X}^N,\tilde{Y}^N}\big) \le \mathbf{E}\bigg[\min\bigg\{\sum_{i=1}^{N}\sum_{j=1}^{d_i'} B_{i,j},\, 1\bigg\}\bigg],$$

where

$$B_{i,j} := \gamma_{i,j}\prod_{k=j}^{d_i}\Big(2^{-\iota(\overline{U}_{i,k};\overline{U}_{i,[d_i]\setminus[j..k]},Y_i)+\iota(\overline{U}_{i,k};\overline{U}'_{a_{i,k}},Y_{a_{i,k}})} + \mathbf{1}\{k>j\}\Big) \tag{3.4}$$

such that[4]

$$\gamma_{i,j} := \prod_{k=j+1}^{d_i}\Big(\ln|\mathcal{U}_{a_{i,k}}|+1\Big).$$

The sequences $(a_{i,j})_j$ control which auxiliaries U_j node i decodes and in which order. Node i uniquely decodes $\overline{U}'_i = (U_{a_{i,j}})_{j\in[d_i']}$ while utilizing $(U_{a_{i,j}})_{j\in[d_i'+1..d_i]}$ by non-unique decoding via the exponential process refinement (Definition 3.2.2). For brevity, we say "the *decoding order* of node i is $\overline{U}_{i,1},\dots,\overline{U}_{i,d_i'},\overline{U}_{i,d_i'+1}?,\dots,\overline{U}_{i,d_i}?$" where "?" means the random variable is only used in non-unique decoding. Node i decodes $\overline{U}'_i$, creates its own U_i by using the exponential functional representation on $P_{U_i|Y_i,\overline{U}'_i}$, and generates X_i from $P_{X_i|Y_i,U_i,\overline{U}'_i}$.

We also have the following result for deterministic coding schemes.

Theorem 3.4.2. *Fix any ADN $(P_{Y_i|X^{i-1},Y^{i-1}})_{i\in[N]}$. For any $(a_{i,j})_{i\in[N],j\in[d_i]}$, $(d_i')_{i\in[N]}$, $(P_{U_i|Y_i,\overline{U}'_i}, P_{X_i|Y_i,U_i,\overline{U}'_i})_{i\in[N]}$ as defined in Theorem 3.4.1, which induce the joint distribution of X^N, Y^N, U^N, and any set $\mathcal{E}\subseteq\big(\prod_{i=1}^N\mathcal{X}_i\big)\times\big(\prod_{i=1}^N\mathcal{Y}_i\big)$, there is a deterministic coding scheme $(f_i)_{i\in[N]}$ such that $\tilde{X}^N, \tilde{Y}^N$ induced by the scheme satisfy*

$$\mathbf{P}\big((\tilde{X}^N,\tilde{Y}^N)\in\mathcal{E}\big)$$

[4]Note that the logarithmic terms $\gamma_{i,j}$ do not affect the first and second order results.

$$\leq \mathbf{E}\left[\min\left\{\mathbf{1}\{(X^N,Y^N)\in\mathcal{E}\}+\sum_{i=1}^{N}\sum_{j=1}^{d_i'}B_{i,j},\ 1\right\}\right], \tag{3.5}$$

where $B_{i,j}$ is defined in Theorem 3.4.1.

Theorem 3.4.1 implies the following result for the asymptotic ADMN [110] by directly applying the law of large numbers.

Corollary 3.4.3. *Fix any ADN $(P_{Y_i|X^{i-1},Y^{i-1}})_{i\in[N]}$. Fix any $(a_{i,j})_{i\in[N],j\in[d_i]}$, $(d_i')_{i\in[N]}$, $(P_{U_i|Y_i,\overline{U}_i'},P_{X_i|Y_i,U_i,\overline{U}_i'})_{i\in[N]}$ as defined in Theorem 3.4.1, which induces the joint distribution of X^N,Y^N,U^N. If for every $i\in[N]$, $j\in[d_i']$,*

$$I(\overline{U}_{i,j};\overline{U}_{i,[d_i]\setminus\{j\}},Y_i)-I(\overline{U}_{i,j};\overline{U}'_{a_{i,j}},Y_{a_{i,j}})>$$
$$\sum_{k=j+1}^{d_i}\left(\max\left\{I(\overline{U}_{i,k};\overline{U}'_{a_{i,k}},Y_{a_{i,k}})-I(\overline{U}_{i,k};\overline{U}_{i,[d_i]\setminus[j..k]},Y_i),\ 0\right\}\right),$$

then there is a sequence of public-randomness coding (indexed by n) for the n-fold ADN $(P^n_{Y_i|X^{i-1},Y^{i-1}})_{i\in[N]}$ such that the induced $\tilde{X}^{N,n},\tilde{Y}^{N,n}$ (write $\tilde{X}^{N,n}=(\tilde{X}_{i,j})_{i\in[N],j\in[n]}$) satisfy

$$\lim_{n\to\infty}\delta_{\mathrm{TV}}\big(P^n_{X^N,Y^N},\ P_{\tilde{X}^{N,n},\tilde{Y}^{N,n}}\big)=0. \tag{3.6}$$

While this result is not as strong as the general asymptotic result in [110], a one-shot analogue of [110] will likely be significantly more complicated than Theorem 3.4.1. We choose to present Theorem 3.4.1 since it is simple but already general and powerful enough to give a wide range of tight one-shot results.

3.5 One-shot Relay Channel

To explain our scheme, we first discuss a *one-shot relay channel* in Figure 3.3. An encoder observes $M\sim\mathrm{Unif}[\mathsf{L}]$ and outputs X, which is passed through the channel $P_{Y_\mathrm{r}|X}$. The relay observes Y_r and outputs X_r. Then $(X,X_\mathrm{r},Y_\mathrm{r})$ is passed

through the channel $P_{Y|X,X_{\rm r},Y_{\rm r}}$. The decoder observes Y and recovers $\hat{M}$. For generality, we allow Y to depend on all of $X, X_{\rm r}, Y_{\rm r}$, and $X_{\rm r}$ may interfere with $(X, Y_{\rm r})$, which can happen if the relay outputs $X_{\rm r}$ instantaneously or the channel has a long memory, or it is a storage device. It is a one-shot version of the *relay-without-delay* and *relay-with-unlimited-look-ahead* [57, 58], and is an ADN by taking $Y_1 = M$, $X_1 = X$, $Y_2 = Y_{\rm r}$, $X_2 = X_{\rm r}$, $Y_3 = Y$, and $X_3 = M$ (in the ideal distributions).

In case if $Y = (Y', Y'')$ consists of two components and the channel $P_{Y|X,X_{\rm r},Y_{\rm r}} = P_{Y'|X,Y_{\rm r}} P_{Y''|X_{\rm r}}$ can be decomposed into two orthogonal components (so $X_{\rm r}$ does not interfere with $(X, Y_{\rm r})$), this becomes the one-shot version of the *primitive relay channel* [55, 56, 59, 101, 142] since the n-fold version of this ADN (with $n \to \infty$) is precisely the asymptotic primitive relay channel. However, the n-fold version of the ADN in Figure 3.3 in general is not the conventional relay channel [36, 59, 171] (it is the relay-with-unlimited-look-ahead instead). The conventional relay channel, due to its causal assumption that the relay can only look at past $Y_{{\rm r},t}$'s, has no one-shot counterpart.

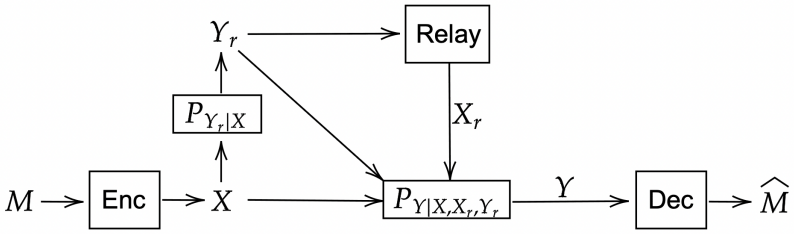

Figure 3.3: One-shot relay channel setting.

We use the following corollary of Theorem 3.4.2 to demonstrate the use of the exponential process refinement lemma (Lemma 3.2.3).

Corollary 3.5.1. *For any P_X, $P_{U|Y_{\rm r}}$, function $x_{\rm r}(y_{\rm r}, u)$, there is a deterministic coding scheme for the one-shot relay channel such that the error probability*

satisfies

$$P_e \le \mathbf{E}\Big[\min\left\{\gamma \mathsf{L} 2^{-\iota(X;U,Y)}\left(2^{-\iota(U;Y)+\iota(U;Y_{\mathrm{r}})}+1\right),1\right\}\Big], \tag{3.7}$$

where $(X, Y_{\mathrm{r}}, U, X_{\mathrm{r}}, Y) \sim P_X P_{Y_{\mathrm{r}}|X} P_{U|Y_{\mathrm{r}}} \delta_{x_{\mathrm{r}}(Y_{\mathrm{r}},U)} P_{Y|X,Y_{\mathrm{r}},X_{\mathrm{r}}}$, *and* $\gamma := \ln|\mathcal{U}| + 1$.

Proof. For the sake of demonstration, we first give a detailed proof via the exponential process refinement lemma without invoking Theorem 3.4.2. Let $U_1 := (X, M)$, $U_2 := U$. Let $\mathbf{U}_1$, $\mathbf{U}_2$ be two independent exponential processes, which serve as the "random codebooks". The encoder (node 1) uses the exponential functional representation (3.1) to compute $U_1 = (\mathbf{U}_1)_{P_{U_1}\times\delta_M}$ and outputs X-component of U_1. The relay (node 2) computes $U_2 = (\mathbf{U}_2)_{P_{U_2|Y_{\mathrm{r}}}(\cdot|Y_{\mathrm{r}})}$ and outputs $X_{\mathrm{r}} = x_{\mathrm{r}}(Y_{\mathrm{r}}, U_2)$. Note that $X, Y_{\mathrm{r}}, U_2, X_{\mathrm{r}}, Y$ follow the ideal distribution in the corollary due to the property of exponential functional representation, and hence we write X_{r} instead of $\tilde{X}_{\mathrm{r}}$. The decoder (node 3) observes Y, and performs the following steps.

1. Refine $P_{U_2|Y}(\cdot|Y)$ (written as $P_{U_2|Y}$ for brevity) to $Q_{U_2} := P_{U_2|Y}^{\mathbf{U}_2}$ using Definition 3.2.2. By the exponential process refinement lemma (Lemma 3.2.3, with $V = \emptyset$),

$$\mathbf{E}\left[\frac{1}{Q_{U_2}(U_2)}\middle| U_2, Y, Y_{\mathrm{r}}\right] \le (\ln|\mathcal{U}_2|+1)\left(\frac{P_{U_2|Y_{\mathrm{r}}}(U_2)}{P_{U_2|Y}(U_2)}+1\right).$$

2. Compute the joint distribution $Q_{U_2}P_{U_1|U_2,Y}$ over $\mathcal{U}_1 \times \mathcal{U}_2$, the semidirect product between Q_{U_2} and $P_{U_1|U_2,Y}(\cdot|\cdot, Y)$. Let its U_1-marginal be $\tilde{Q}_{U_1}$.

3. Let $\tilde{U}_1 = (\mathbf{U}_1)_{\tilde{Q}_{U_1}\times P_M}$, and output its M-component.

Let $A := (X, Y_{\mathrm{r}}, U_2, X_{\mathrm{r}}, Y, M)$ and $\gamma := \ln|\mathcal{U}_2| + 1$,

$$\mathbf{P}(\tilde{U}_1 \neq U_1 \,|\, A)$$

$$
\begin{aligned}
&\overset{(a)}{\leq} \mathbf{E}\left[\min\left\{\frac{P_{U_1}(U_1)\delta_M(M)}{P_{U_1|U_2,Y}(U_1|U_2,Y)Q_{U_2}(U_2)P_M(M)},1\right\}\,\middle|\,A\right]\\
&\overset{(b)}{=} \mathbf{E}\left[\min\left\{\mathsf{L}\frac{P_{U_1}(U_1)}{P_{U_1|U_2,Y}(U_1|U_2,Y)Q_{U_2}(U_2)},1\right\}\,\middle|\,A\right]\\
&\overset{(c)}{\leq} \min\left\{\mathsf{L}\frac{P_{U_1}(U_1)}{P_{U_1|U_2,Y}(U_1|U_2,Y)}\gamma\left(\frac{P_{U_2|Y_\mathrm{r}}(U_2)}{P_{U_2|Y}(U_2)}+1\right),1\right\}\\
&= \min\left\{\gamma\mathsf{L}2^{-\iota(X;U_2,Y)}\left(2^{-\iota(U_2;Y)+\iota(U_2;Y_\mathrm{r})}+1\right),1\right\},
\end{aligned}
$$

where (a) is by the generalized Poisson matching lemma [117] (Lemma 3.2.1), (b) is by $\delta_M(M) = 1$ and $P_M(M) = 1/\mathsf{L}$, and (c) is by step 1) and Jensen's inequality. Taking expectation over A gives the desired error bound. Although the codebooks $\mathbf{U}_1$, $\mathbf{U}_2$ are random (so this is a public-randomness scheme), we can convert it to a deterministic scheme by fixing one particular choice $(\mathbf{u}_1, \mathbf{u}_2)$ that satisfies the error bound. Alternatively, Theorem 3.4.2 allows us to derive bounds for general acyclic discrete networks in a systematic manner, without going through the above arguments for every specific ADN. To prove Corollary 3.5.1, we can invoke Theorem 3.4.2 on the ADN with nodes $1, 2, 3$, with inputs Y_i's, outputs X_i's, auxiliaries U_i's and the terms $B_{i,j}$'s as follows:

1. Node 1 has input $Y_1 = M$, output $X_1 = X$ and auxiliary $U_1 = (X, M)$.

2. Node 2 has input $Y_2 = Y_r$, output $X_2 = X_r$ and auxiliary $U_2 = U$.

3. Node 3 has input $Y_3 = Y$, output $X_3 = M$ and decodes with the order "U_1, U_2?". Applying Theorem 3.4.2 (note that $d_3 = 2$ and $d_3' = 1$), we have

$$
B_{3,1} = (\ln|\mathcal{U}_2| + 1)\mathsf{L}2^{-\iota(X;U_2,Y)}\left(2^{-\iota(U_2;Y)+\iota(U_2;Y_\mathrm{r})}+1\right),
$$

and hence we obtain the bound (3.7) by invoking Theorem 3.4.2.

□

Corollary 3.5.1 yields the following asymptotic achievable rate:

$$R \le I(X;U,Y) - \max\left\{I(U;Y_{\mathrm{r}}) - I(U;Y),\, 0\right\}$$

for some $P_{U|Y_{\mathrm{r}}}$ and function $x_{\mathrm{r}}(y_{\mathrm{r}}, u_2)$.

We also consider a one-shot primitive relay channel (as shown in Figure 3.4), where $P_{Y|X,X_{\mathrm{r}},Y_{\mathrm{r}}} = P_{Y'|X,Y_{\mathrm{r}}} P_{Y''|X_{\mathrm{r}}}$ can be decomposed into two orthogonal components. Consider (X, Y_{r}, Y') independent of (X_{r}, Y'') in the ideal distribution and take $U = (U', X_{\mathrm{r}})$ where U' follows $P_{U'|Y_{\mathrm{r}}}$, Corollary 3.5.1 specializes to the following Corollary 3.5.2.

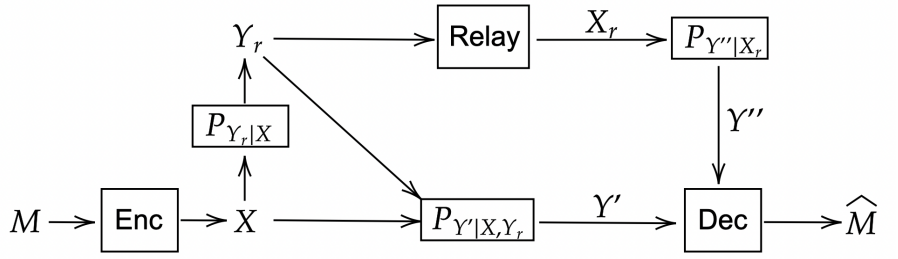

Figure 3.4: One-shot primitive relay channel setting.

Corollary 3.5.2. *For any P_X, $P_{X_{\mathrm{r}}}$, $P_{U'|Y_{\mathrm{r}}}$, there is a deterministic coding scheme for the one-shot primitive relay channel with $M \sim \mathrm{Unif}[\mathsf{L}]$ such that the error probability satisfies*

$$P_e \le \mathbf{E}\left[\min\left\{\gamma \mathsf{L} 2^{-\iota(X;U',Y')}\left(2^{-\iota(X_{\mathrm{r}};Y'')+\iota(U';Y_{\mathrm{r}}|Y')}+1\right), 1\right\}\right],$$

where $(X, Y_{\mathrm{r}}, U', Y') \sim P_X P_{Y_{\mathrm{r}}|X} P_{U'|Y_{\mathrm{r}}} P_{Y'|X,Y_{\mathrm{r}}}$ is independent of $(X_{\mathrm{r}}, Y'') \sim P_{X_{\mathrm{r}}} P_{Y''|X_{\mathrm{r}}}$, and $\gamma := \ln(|\mathcal{U}'||\mathcal{X}_{\mathrm{r}}|) + 1$.

This gives the asymptotic achievable rate $R \le I(X;U',Y') - \max\{I(U';Y_{\mathrm{r}}|Y') - C_{\mathrm{r}},\, 0\}$ where $C_{\mathrm{r}} = \max_{P_{X_{\mathrm{r}}}} I(X_{\mathrm{r}};Y'')$ is the capacity of the channel $P_{Y''|X_{\mathrm{r}}}$. It implies the compress-and-forward bound [101], which is the maximum of $I(X;U',Y')$

subject to the constraint $C_{\mathrm{r}} \geq I(U'; Y_{\mathrm{r}}|Y')$ (where the random variables are distributed as in Corollary 3.5.2). Hence, Corollary 3.5.2 can be treated as a one-shot compress-and-forward bound.

3.5.1 Partial-Decode-and-Forward Bound

We extend Corollary 3.5.1 to allow partial decoding of the message [36, 58, 101]. To this end, we split the message and encoder into two. The message $M \sim \mathrm{Unif}[\mathsf{L}]$ is split into $M_1 \sim \mathrm{Unif}[\mathsf{J}]$ and $M_2 \sim \mathrm{Unif}[\mathsf{L}/\mathsf{J}]$ (assume J is a factor of L). The encoder controls two nodes (node 1 and 2), where node 1 observes $Y_1 = M_1$, outputs $X_1 = V$, and has an auxiliary $U_1 = (M_1, V)$; node 2 observes $Y_2 = (M_1, M_2, V)$, outputs $X_2 = X$, and has an auxiliary $U_2 = (M_1, M_2, X)$. The relay (node 3) observes $Y_3 = Y_r$, decodes U_1, outputs $X_3 = X_r$, and has an auxiliary $U_3 = (M_1, U)$. The decoder (node 4) observes $Y_4 = Y$ and uses the decoding order "$U_2, U_3?, U_1?$".

Corollary 3.5.3. *Fix any $P_{X,V}$, $P_{U|Y_{\mathrm{r}},V}$, function $x_{\mathrm{r}}(y_{\mathrm{r}}, u, v)$, and J which is a factor of L. There exists a deterministic coding scheme for the one-shot relay channel with*

$$P_e \leq \mathbf{E}\Big[\min\Big\{\mathsf{J}2^{-\iota(V;Y_{\mathrm{r}})} + \gamma\mathsf{L}\mathsf{J}^{-1}2^{-\iota(X;U,Y|V)} \cdot \big(2^{-\iota(U;V,Y)+\iota(U;V,Y_{\mathrm{r}})} + 1\big)\big(\mathsf{J}2^{-\iota(V;Y)} + 1\big), 1\Big\}\Big],$$

where $(X, V, Y_{\mathrm{r}}, U, X_{\mathrm{r}}, Y) \sim P_{X,V} P_{Y_{\mathrm{r}}|X,V} P_{U|Y_{\mathrm{r}},V} \delta_{x_{\mathrm{r}}(Y_{\mathrm{r}},U,V)} P_{Y|X,Y_{\mathrm{r}},X_{\mathrm{r}}}$ and $\gamma := (\ln(\mathsf{J}|\mathcal{U}|)+1)(\ln(\mathsf{J}|\mathcal{V}|)+1)$.

Applying the law of large numbers to Corollary 3.5.3, and Fourier-Motzkin elimination (using the PSITIP software [112]), we obtain the following asymptotic

achievable rate:

$$\min\begin{Bmatrix} I(V;Y)+I(U,Y;X|V), \\ I(V;Y_{\mathrm{r}})+I(U,Y;X|V), \\ I(V,U;Y)+I(U,Y;X|V)-I(U;Y_{\mathrm{r}}|V), \\ I(V;Y_{\mathrm{r}})+I(U;Y|V)+I(U,Y;X|V)-I(U;Y_{\mathrm{r}}|V) \end{Bmatrix},$$

where $(X,V,Y_{\mathrm{r}},U,X_{\mathrm{r}},Y)\sim P_{X,V}P_{Y_{\mathrm{r}}|X,V}P_{U|Y_{\mathrm{r}},V}\delta_{x_{\mathrm{r}}(Y_{\mathrm{r}},U,V)}P_{Y|X,Y_{\mathrm{r}},X_{\mathrm{r}}}$, subject to the constraint $I(U;Y_{\mathrm{r}}|V)\le I(U;Y|V)+I(U,Y;X|V)$.

Taking $x_{\mathrm{r}}(y_{\mathrm{r}},(v',x'_{\mathrm{r}}))=x'_{\mathrm{r}}$, $U=\emptyset$, $V=(V',X'_{\mathrm{r}})$, it gives an achievable rate $\min\{I(X,X_{\mathrm{r}};Y),\ I(V';Y_{\mathrm{r}})+I(X;Y|X_{\mathrm{r}},V')\}$, recovering the partial noncausal decode-forward bound for relay-with-unlimited-look-ahead [58, Prop. 3].

Specializing to the primitive relay channel, and again substituting $U=\emptyset$, $V=(V',X'_{\mathrm{r}})$, $x_{\mathrm{r}}(y_{\mathrm{r}},(v',x'_{\mathrm{r}}))=x'_{\mathrm{r}}$, we have

$$P_e\le \mathbf{E}\Big[\min\Big\{\mathsf{J}2^{-\iota(V';Y_{\mathrm{r}})}+2\gamma\mathsf{L}\mathsf{J}^{-1}2^{-\iota(X;Y'|V')} \cdot\big(\mathsf{J}2^{-\iota(V';Y')-\iota(X_{\mathrm{r}};Y'')}+1\big),1\Big\}\Big],$$

where $\gamma:=(\ln\mathsf{J}+1)(\ln(\mathsf{J}|\mathcal{V}'||\mathcal{X}_{\mathrm{r}}|)+1)$ and $(X,V',Y_{\mathrm{r}},Y')\sim P_{X,V'}P_{Y_{\mathrm{r}}|X}P_{Y'|X,Y_{\mathrm{r}}}$ is independent of $(X_{\mathrm{r}},Y'')\sim P_{X_{\mathrm{r}}}P_{Y''|X_{\mathrm{r}}}$. It gives the asymptotic rate $\min\{I(V';Y_{\mathrm{r}})+I(X;Y|V'),\ I(X;Y)+C_{\mathrm{r}}\}$ and recovers the partial decode-forward lower bound for primitive relay channels [36, 101]. One-shot versions of other asymptotic bounds for primitive relay channels (e.g., [55, 142]) are left for future studies.

3.6 Cascade multiterminal source coding with computing

We consider the cascade multiterminal source coding problem [42] (which is also called the *cascade coding for computing* in [59, Section 21.4]). It is similar to

the traditional multiterminal source coding problem introduced by Berger and Tung [14, 169], where two information sources are encoded in a distributed fashion with loss, though the communication between encoders replaces one of the direct channels to the decoder in the cascade case. It can include different variations, e.g., the decoder desires to estimate both X and Y, X only, Y only and some functions of both. It is tightly related to real problems where it is required to pass messages to neighbors in order to compute functions of data, e.g., distributed data collection, aggregating measurements in sensor networks, interactive coding for computing and distributed lossy averaging (see [42] and references therein).

The asymptotic rate-distortion region for the general cascade multiterminal source coding problem is unknown, even for the case where X and Y are independent. We study the one-shot setting of this problem, which has not been discussed in literature to the best of our knowledge. We provide a novel one-shot bound on the cascade multiterminal source coding problem, and show that our one-shot achievability result recovers the best known asymptotic inner bound, i.e., the *local-computing-and-forwarding inner bound* [42] (which in turn recovers various other existing bounds as special cases, also see [59, Section 21.4] for a detailed discussion).

The one-shot cascade multiterminal source coding problem is described as follows (see Figure 3.5). Consider two sources X and Y that are jointly distributed according to $P_{X,Y}$. They will be described by separate encoders and passed to a single decoder in a cascade fashion. Upon observing X, encoder a sends a message $M \in [\mathsf{L}_1]$ about X to encoder b. Encoder b then creates a final message $M' \in [\mathsf{L}_2]$ summarizing both sources X and Y and sends it to the decoder. We investigate the error probability P_e, which is the probability of the decoder recovers $\tilde{Z} \in \mathcal{Z}$ with excess distortion $P_e := \mathbf{P}\{d(X, Y, \tilde{Z}) > \mathsf{D}\}$, where $d : \mathcal{X} \times \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}_{\geq 0}$ is a distortion measure. Due to the flexibility of the distortion function d, in

general one can estimate any function of X and Y to tackle various objectives in practice [42].

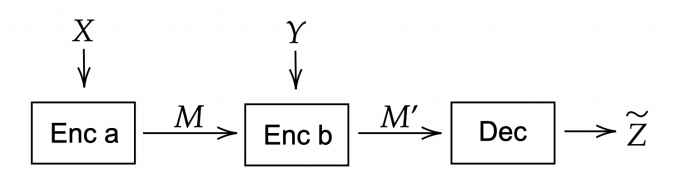

Figure 3.5: One-shot cascade multiterminal source coding setting.

By Theorem 3.4.2, we bound P_e by the following corollary.

Corollary 3.6.1. *Fix $P_{X,Y}$, $P_{U,V|X}$, function $z : \mathcal{U}\times\mathcal{V}\times\mathcal{Y} \to \mathcal{Z}$ and $\tilde{\mathsf{L}}_i, i = 1,2,3$ with $\tilde{\mathsf{L}}_1\tilde{\mathsf{L}}_2 \le \mathsf{L}_1$ and $\tilde{\mathsf{L}}_2\tilde{\mathsf{L}}_3 \le \mathsf{L}_2$, there exists a deterministic coding scheme for the one-shot cascade multiterminal source coding problem such that the probability of excess distortion is bounded by*

$$\begin{aligned} P_e \le \mathbf{E}\Bigg[\min\Big\{ & \mathbf{1}\{d(X,Y,Z) > \mathsf{D}\} + \gamma\tilde{\mathsf{L}}_1^{-1}\tilde{\mathsf{L}}_2^{-1}2^{\iota(U,V;X|Y)} \\ & + \gamma\tilde{\mathsf{L}}_1^{-1}2^{-\iota(V;U,Y)+\iota(V;U,X)} + \tilde{\mathsf{L}}_2^{-1}2^{-\iota(U;V,Y)+\iota(U;X)} \\ & + \gamma\tilde{\mathsf{L}}_3^{-1}2^{\iota(Z;V,Y|U)}\left(\tilde{\mathsf{L}}_2^{-1}2^{\iota(U;X)}+1\right),\ 1\Big\}\Bigg], \end{aligned}$$

where $\gamma = \ln(|\mathcal{U}|\tilde{\mathsf{L}}_2) + 1$ and $X, Y, Z, U, V \sim P_X P_{Y|X} P_{U,V|X} P_{Z|Y,U,V}$.

Proof. We adapt the problem into our ADN framework by splitting the encoder a, as shown in Figure 3.6 (see next page). The encoder a1, encoder a2, encoder b and decoder are referred to as nodes $1, 2, 3, 4$, respectively.

Let $M_i \in [\mathsf{L}_i]$ for $i = 1, 2, 3$. Encoder a1 (node 1) observes X, outputs U, and has an auxiliary $U_1 = (U, M_2)$. Encoder a2 (node 2) observes (U, X), outputs (M_1, M_2), and has an auxiliary $U_2 = (V, M_1)$. Encoder b (node 3) observes M_1, M_2 and Y, outputs (M_2, M_3), and has an auxiliary $U_3 = (Z, M_3)$. The decoder observes M_2, M_3 and recovers Z by using the function z and our coding scheme.

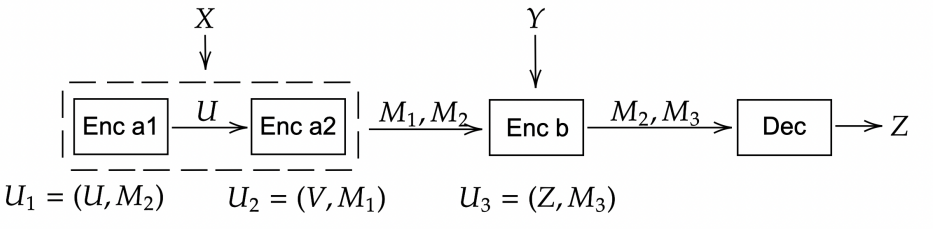

Figure 3.6: One-shot cascade multiterminal source coding in AND framework by splitting the first encoder.

For each node $i = 1, \ldots, 4$ in the ADN, we describe its input Y_i, output X_i, auxiliary U_i and the terms $B_{i,j}$ in Theorem 3.4.2 as follows:

1. Node 1 has input $Y_1 = X$, output $X_1 = U$ and auxiliary $U_1 = (U, M_2)$.

2. Node 2 has input $Y_2 = (U, X)$, output $X_2 = M_1$ and auxiliary $U_2 = (V, M_1)$.

3. Node 3 has input $Y_3 = (Y, M_1, M_2)$, output $X_3 = M_3$, auxiliary $U_3 = (Z, M_3)$, and decodes with the order "U_2, U_1" (i.e., "$U_{\text{Enc 1}b}, U_{\text{Enc 1a}}$"). We have $d_3' = d_3 = 2$, and

$$B_{3,1} = \gamma \tilde{\mathsf{L}}_1^{-1} \tilde{\mathsf{L}}_2^{-1} 2^{\iota(U,V;X|Y)} + \gamma \tilde{\mathsf{L}}_1^{-1} 2^{-\iota(V;U,Y)+\iota(V;U,X)},$$
$$B_{3,2} = \tilde{\mathsf{L}}_2^{-1} 2^{-\iota(U;V,Y)+\iota(U;X)},$$

   where $\gamma = \ln(|\mathcal{U}|\tilde{\mathsf{L}}_2) + 1$.

4. Node 4 has input $Y_4 = (M_2, M_3)$, output $X_4 = Z$, and decodes with the order "U_3, U_1?" (i.e., "$U_{\text{Enc 2}}, U_{\text{Enc 1a}}$?"). By applying Theorem 3.4.2 (note that $d_4' = 1$ and $d_4 = 2$), it gives

$$B_{4,1} = \gamma \tilde{\mathsf{L}}_3^{-1} 2^{\iota(Z;V,Y|U)} \left(\tilde{\mathsf{L}}_2^{-1} 2^{\iota(U;X)} + 1 \right),$$

   where $\gamma = \ln(|\mathcal{U}|\tilde{\mathsf{L}}_2) + 1$.

Therefore, by applying Theorem 3.4.2, the probability of distortion exceeds the limit $P_e := \mathbf{P}\{d(X,Y,\tilde{Z}) > \mathsf{D}\}$ can be bounded as the result stated in this corollary:

$$\begin{aligned} P_e &\leq \mathbf{E}\left[\min\left\{\mathbf{1}\{d(X,Y,Z)>\mathsf{D}\} + B_{3,1} + B_{3,2} + B_{4,1}, 1\right\}\right] \\ &= \mathbf{E}\Bigg[\min\Big\{\mathbf{1}\{d(X,Y,Z)>\mathsf{D}\} + \gamma\tilde{\mathsf{L}}_1^{-1}\tilde{\mathsf{L}}_2^{-1}2^{\iota(U,V;X|Y)} \\ &\qquad + \gamma\tilde{\mathsf{L}}_1^{-1}2^{-\iota(V;U,Y)+\iota(V;U,X)} + \tilde{\mathsf{L}}_2^{-1}2^{-\iota(U;V,Y)+\iota(U;X)} \\ &\qquad + \gamma\tilde{\mathsf{L}}_3^{-1}2^{\iota(Z;V,Y|U)}\left(\tilde{\mathsf{L}}_2^{-1}2^{\iota(U;X)}+1\right),\ 1\Big\}\Bigg], \end{aligned}$$

where $\gamma = \ln(|\mathcal{U}|\tilde{\mathsf{L}}_2) + 1$. □

Following the one-shot bound as shown above, we let $\tilde{\mathsf{L}}_i = 2^{n\tilde{R}_i}$ for $i = 1, 2, 3$ and apply the law of large numbers. We obtain the asymptotic achievable region

$$\begin{aligned} \tilde{R}_1 + \tilde{R}_2 &> I(X;U,V|Y), \\ \tilde{R}_1 &> I(V;U,X) - I(V;U,Y), \\ \tilde{R}_2 &> I(X;U) - I(V,Y;U), \\ \tilde{R}_2 + \tilde{R}_3 &> I(X;U) + I(Z;V,Y|U), \\ \tilde{R}_3 &> I(Z;V,Y|U), \end{aligned}$$

and $D > \mathbf{E}[d(X,Y,Z)]$. By $\tilde{\mathsf{L}}_1\tilde{\mathsf{L}}_2 \leq \mathsf{L}_1$ and $\tilde{\mathsf{L}}_2\tilde{\mathsf{L}}_3 \leq \mathsf{L}_2$ and consider $R_1 = \tilde{R}_1 + \tilde{R}_2$ and $R_2 = \tilde{R}_2 + \tilde{R}_3$, by applying the Fourier-Motzkin elimination (using the PSI-TIP software [112]), we recover an asymptotic achievable region for the cascade multiterminal source coding problem:

$$\begin{aligned} R_1 &> I(X;U,V|Y), \\ R_2 &> I(X;U) + I(Z;V,Y|U), \end{aligned}$$

and $D > \mathbf{E}[d(X,Y,Z)]$, where $X,Y,Z,U,V \sim P_X P_{Y|X} P_{U,V|X} P_{Z|Y,U,V}$. This is the local-computing-and-forwarding inner bound in asymptotic case, as discussed in [42] and also in [59, Section 21.4].

3.7 Examples of Acyclic Discrete Networks

In this section, to demonstrate the use of our main results, we apply Theorem 3.4.1 and Theorem 3.4.2 on several settings in network information theory: Gelfand-Pinsker [74, 89], Wyner-Ziv [180, 182], coding for computing [184], multiple access channels [4, 5, 125] and broadcast channels [138], recovering similar results as the results in [117] and other works.

3.7.1 Gelfand-Pinsker Problem

The one-shot version of the Gelfand-Pinsker problem [74] is described as follows.

Upon observing $M \sim \mathrm{Unif}[\mathsf{L}]$ and $S \sim P_S$, the encoder generates X and sends X through a channel $P_{Y|X,S}$. The decoder receives Y and recovers $\hat{M}$. This can be considered as an ADN as follows (see Figure 3.7 for an illustration): in the ideal situation, let $Y_1 := (M,S)$ represent all the information coming into node 1, $Y_2 := Y$, $P_{Y_2|Y_1,X_1}$ be $P_{Y|S,X}$, and $X_2 := M$. The auxiliary of node 1 is $U_1 = (U,M)$ for some U following $P_{U|S}$ given S. The decoding order of node 2 is "U_1" (i.e., it only wants U_1). Since node 2 has decoded U_1, X_2 is allowed to depend on $U_1 = (U,M)$, and hence the choice $X_2 := M$ is valid in the ideal situation. Nevertheless, in the actual situation where we have $\tilde{X}, \tilde{Y}$ instead of X, Y, the actual output $\tilde{X}_2$ will not be exactly M, though the error probability $P_e := \mathbf{P}(\tilde{X}_2 \neq M)$ can still be bounded. Applying Theorem 3.4.2, we obtain the following Corollary 3.7.1.

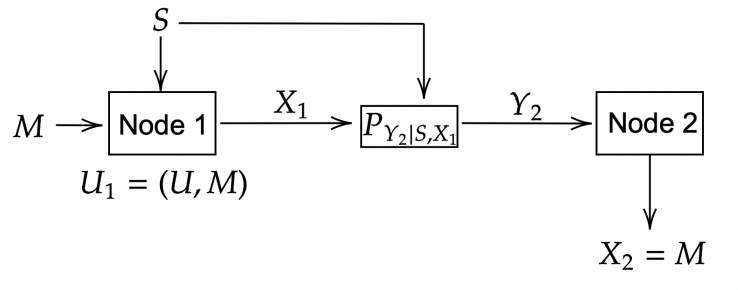

Figure 3.7: Gelfand-Pinsker problem in ADN framework.

Corollary 3.7.1. *Fix $P_{U|S}$ and function $x : \mathcal{U} \times \mathcal{S} \to \mathcal{X}$. There exists a deterministic coding scheme for the channel $P_{Y|X,S}$ with state $S \sim P_S$ and message $M \sim \mathrm{Unif}[\mathsf{L}]$ such that*

$$P_e \le \mathbf{E}\big[\min\big\{\mathsf{L}2^{-\imath(U;Y)+\imath(U;S)}, 1\big\}\big],$$

where $S, U, X, Y \sim P_S P_{U|S} \delta_{x(U,S)} P_{Y|X,S}$.

This bound is similar to the one given in [117] (which is stronger than the one-shot bounds in [172, 178, 189] in the second order). Both of them attain the second-order result in [151].

3.7.2 Wyner-Ziv Problem and Coding for Computing

The Wyner-Ziv problem [180, 182] in a one-shot setting is described as follows (see Figure 3.9 for an illustration).

Upon observing $X \sim P_X$, the encoder outputs $M \in [\mathsf{L}]$. The decoder receives M and the side information $T \sim P_{T|X}$, and recovers $\tilde{Z} \in \mathcal{Z}$ with probability of excess distortion $P_e := \mathbf{P}\{d(X, \tilde{Z}) > \mathsf{D}\}$, where $d : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}_{\ge 0}$ is a distortion measure. This can be considered as an ADN: in the ideal situation, $Y_1 := X$, $X_1 := M$, $Y_2 := (M, T)$, $X_2 := Z$. The auxiliary of node 1 is $U_1 = (U, M)$ for some U following $P_{U|X}$ given X. By Theorem 3.4.2, we bound P_e as follows.

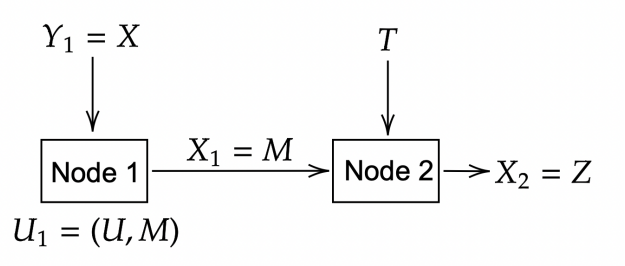

Figure 3.8: Wyner-Ziv problem in ADN framework.

Corollary 3.7.2. *Fix $P_{U|X}$ and function $z : \mathcal{U} \times \mathcal{Y} \to \mathcal{Z}$. There exists a deterministic coding scheme for lossy source coding with source $X \sim P_X$, side information at the decoder $T \sim P_{T|X}$ and description $M \in [\mathsf{L}]$ such that*

$$P_e \leq \mathbf{E}\Big[\min\Big\{\mathbf{1}\{d(X,Z) > \mathsf{D}\} + \mathsf{L}^{-1}2^{-\imath(U;T)+\imath(U;X)}, 1\Big\}\Big], \tag{3.8}$$

where $X, Y, U, Z \sim P_X P_{Y|X} P_{U|X} \delta_{z(U,Y)}$.

This bound is similar to, though slightly weaker than, the bound given in [117] (which improves upon the one-shot bounds in [172, 178] in second-order performance). Our main contribution lies in the generality of our one-shot coding framework, and we do not always derive bounds identical to those in [117], despite both methods employing Poisson functional representations.

This reduces to lossy source coding with $T = \emptyset$. Let $U = Z$, we have $P_e \leq \mathbf{P}(d(X,Z) > \mathsf{D}) + \mathbf{E}\left[\min\left\{\mathsf{L}^{-1}2^{\imath(Z;X)}, 1\right\}\right]$.

We also consider coding for computing [184], where node 2 recovers a function $f(X,T)$ of X and T with respect to distortion level D with a distortion measure $d(\cdot,\cdot)$. The probability of excess distortion is $P_e := \mathbf{P}\{d(f(X,T), \tilde{Z}) > \mathsf{D}\}$. We obtain a result similar to Corollary 3.7.2, where (3.8) is changed to

$$P_e \leq \mathbf{E}\Big[\min\Big\{\mathbf{1}\{d(f(X,T),Z) > \mathsf{D}\} + \mathsf{L}^{-1}2^{-\imath(U;T)+\imath(U;X)},\ 1\Big\}\Big].$$

3.7.3 Multiple Access Channel

The multiple access channel [4, 5, 125] in a one-shot setting is described as follows.

There are two encoders, one decoder, and two independent messages $M_j \sim \text{Unif}[\mathsf{L}_j]$ for $j = 1, 2$. Encoder j observes M_j and creates X_j for $j = 1, 2$. The decoder observes the output Y of the channel $P_{Y|X_1,X_2}$ and produces the reconstructions $(\hat{M}_1, \hat{M}_2)$ of the messages. The error probability is defined as $P_e := \mathbf{P}\{(M_1, M_2) \neq (\hat{M}_1, \hat{M}_2)\}$. To consider this as an ADN, in the ideal situation, we let $Y_1 := M_1$, $Y_2 := M_2$, $Y_3 := Y$ and $X_3 := (M_1, M_2)$. We let $U_1 := (X_1, M_1)$ and $U_2 := (X_2, M_2)$. The decoding order of node 3 is "U_2, U_1" (i.e., decode U_1 (soft), and then U_2 (unique), and then U_1 (unique)). By Theorem 3.4.2, we have the following result.

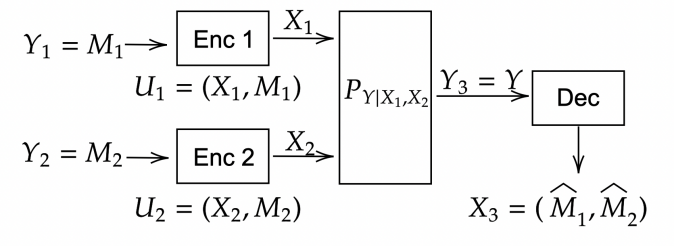

Figure 3.9: Multiple access channel in ADN framework.

Corollary 3.7.3. *Fix P_{X_1}, P_{X_2}. There exists a deterministic coding scheme for the multiple access channel $P_{Y|X_1,X_2}$ for messages $M_j \sim \text{Unif}[1 : \mathrm{L}_j]$ for $j = 1, 2$ such that*

$$P_e \leq \mathbf{E}\Big[\min\Big\{\gamma \mathsf{L}_1 \mathsf{L}_2 2^{-\iota(X_1,X_2;Y)} + \gamma \mathsf{L}_2 2^{-\iota(X_2;Y|X_1)} + \mathsf{L}_1 2^{-\iota(X_1;Y|X_2)}, 1\Big\}\Big],$$

where $\gamma := \ln(\mathsf{L}_1|\mathcal{X}_1|) + 1$, $(X_1, X_2, Y) \sim P_{X_1} P_{X_2} P_{Y|X_1,X_2}$.

This bound is similar to the one-shot bounds in [117, 172]. In the asymptotic setting, this will give the region $R_1 < I(X_1; Y|X_2)$, $R_2 < I(X_2; Y|X_1)$, $R_1 + R_2 < I(X_1, X_2; Y)$.

3.7.4 Broadcast Channel with Private Messages

The broadcast channel with private messages [138] in a one-shot setting is described as follows.

Upon observing independent messages $M_j \sim \mathrm{Unif}[\mathsf{L}_j]$ for $j = 1, 2$, the encoder produces X and sends it through a channel $P_{Y_1,Y_2|X}$. Decoder j observes Y_j and reconstructs $\hat{M}_j$ for $j = 1, 2$. By Theorem 3.4.2, we have the following result.

Corollary 3.7.4. *Fix any P_{U_1,U_2} and function $x : \mathcal{U}_1 \times \mathcal{U}_2 \to \mathcal{X}$. There exists a deterministic coding scheme for the broadcast channel $P_{Y_1,Y_2|X}$ for independent messages $M_k \sim \mathrm{Unif}[\mathsf{L}_j]$ for $j = 1, 2$, with the error probability bounded by*

$$P_e \le \mathbf{E}\Big[\min\Big\{\mathsf{L}_1 2^{-\iota(U_1;Y_1)} + \mathsf{L}_2 2^{-\iota(U_2;Y_2)+\iota(U_1;U_2)}, 1\Big\}\Big],$$

where $(U_1, U_2, X, Y_1, Y_2) \sim P_{U_1,U_2}\delta_{x(U_1,U_2)}P_{Y_1,Y_2|X}$.

In the asymptotic case, this gives a corner point in Marton's region [138]: $R_1 < I(U_1; Y_1), R_2 < I(U_2; Y_2) - I(U_1; U_2)$. Another corner point can be obtained by swapping the decoders.

Chapter 4

One-Shot Coding on Secrecy Problems with Channel Uncertainties

4.1 Overview

In this chapter, we study two fundamental information theory problems in the one-shot regime, namely the information hiding problem [144] and the compound wiretap channel [123]. The former concerns active attacks during information transmission, while the latter addresses passive eavesdropping and information leakage. These two problems have become crucial in the era of data science, where secrecy and privacy are increasingly important due to the growing dependence on and reliance upon large amounts of communicated, analyzed, and utilized data, which inherently contain sensitive and personal information. This chapter is partially based on [134].

For the information hiding problem, [144] formulated it as a communication system from a game-theoretic perspective, where an encoder-decoder team

seeks to transmit a confidential message *embedded* in a host data source, while the opposing side is an attacker, modeled as a noisy channel, attempts to destroy or degrade the message. The information-theoretic limits of different variations of information hiding have been extensively investigated over the past two decades [35, 157, 158], due to its wide range of applications, including watermarking, fingerprinting, steganography, and copyright protection. Existing analyses of information hiding problems borrow techniques from various fields, including wireless communication, signal processing, cryptography, and game theory.

For the compound wiretap channel, [123] modeled the problem as a generalization of Wyner's wiretap channel setting [181], where the communication channel can take multiple potential states. The objective is to ensure reliable transmission and minimize information leakage regardless of which state occurs. This model is more general and better suited to practical scenarios where the transmitter may not have knowledge of the channel conditions or where channel characteristics change rapidly, yet communication performance must still be guaranteed.

In all existing studies on these two problems, the information-theoretic limits have been analyzed in the *asymptotic* regime, assuming that the signal has a blocklength approaching infinity. However, this assumption does not hold in practice, as packets have bounded lengths, which can be quite short in many applications [95]. Similar to Chapter 3, we study the one-shot achievability results of a generalized information hiding setting and the compound wiretap channel, where the channel or source is arbitrary and used only *once*, the law of large numbers does not apply and conventional typicality-based tools are inapplicable.

Most of the existing asymptotic analyses on the two problems [123, 144] assume the decoder know the channel condition (in information hiding, the attacker's strategy), since when the blocklength is large, one can utilize training symbols at the beginning of transmission, whose size becomes negligible compared

to the blocklength. However, this assumption is questionable in the one-shot case (see Section 4.3.3 for discussions).[1] We only assume the attack channel belongs to a set (which may have infinite cardinality). Our goal is to provide *distributionally robust* coding strategies (also see [137]) for the two problems, by utilizing a classical covering argument [18] to handle the uncertainties of channels. To the best of the authors' knowledge, our results are novel and have not been studied in the literature.

This chapter is organized as follows. We begin with a literature review on information hiding, watermarking, compound wiretap channels, and one-shot information theory in Section 4.2. Next, we present the one-shot generalized information hiding problem in Section 4.3 and recover existing asymptotic results in Section 4.4. Within the same framework, we study the one-shot compound wiretap channel in Section 4.5.

4.2 Related Work

We review related literature on the information hiding and the compound wiretap channel in this section.

4.2.1 Information Hiding

The information hiding problem has been studied since [35, 144, 157, 158], due to its wide range of applications, including watermarking, fingerprinting, audio/image/video processing, copyright protection, and steganography. The goal

[1] Note that this assumption was also removed in [158], although that work assumes the side information is an independent shared key of unlimited size and is chosen as part of the coding scheme, whereas in our information hiding setting, we allow it to be correlated with the host and fixed, as in [144].

is to *hide* a message into some host signal (by introducing a certain level of distortion), so that the message can be correctly reconstructed after suffering *attacks* (which introduce another level of distortion). This problem was modeled as a communication problem, and asymptotic information-theoretic capacity was derived in [144]. In general, information hiding is closely related to the Gelfand–Pinsker problem [74, 90], and various extensions have been studied, e.g., the case where side information is available to the encoder, decoder, and adversary [146], and the case where the decoder has rate-limited side information [161]. See [9] for its duality with the Wyner–Ziv problem, and [99] for a comprehensive survey. We discuss its applications and related settings with different objectives as follows.

Watermarking, Fingerprinting, and Steganography

The setting in [144] can be viewed as *public* watermarking [158], where the host signal is available only at the encoder. In contrast, when it is also available at the decoder, *private* watermarking has been studied in [37, 157]. In the Gaussian case, public and private watermarking have the same capacity [35], but this is not true in general. Watermarking problems consider messages containing personal identification information to be protected from attacks, but secrecy is not always required. In comparison, *digital fingerprinting* [21, 144] embeds fingerprints into the host data to uniquely identify users for tracing illegal data usage, which can be more challenging due to potential collusion. A provably good data embedding strategy was introduced by [31]. Random coding error exponents have been investigated in these problems [140, 143, 146]. Although [144, Sec. VII.C] indicated the applicability of information hiding to steganography, the discussion was later extended by [145, 175] to the capacity of perfectly secure steganographic systems. Other steganographic code designs include using trellis codes [79] and polar codes [122].

Host, Stegotext, and Reversibility

In the conventional information hiding setting [144], the message is embedded into host data by producing an encoded signal ("stegotext"), with the goal of recovering the message only. Other objectives have been considered later, such as conveying the host [102] or reconstructing the stegotext [78, 183]. *Reversible* information embedding has also been investigated [97, 160, 162], where the host signal needs to be decoded. However, this can incur a high cost when the host has high entropy [97], making perfect reversibility even impossible for continuous host signals [162]. Nevertheless, in practice, the goal is often to enable retransmission of the stegotext, and codes for stegotext recovery have been studied [78, 183]. For this setting, single-letter capacity-distortion tradeoffs are known only for logarithmic distortion [102] and quadratic distortion in the Gaussian case [164].

4.2.2 Compound Wiretap Channels

Compound wiretap channels [123] generalize the wiretap channel model by Wyner [181] by allowing both the legitimate channel and the eavesdropper's channel to have multiple possible states. The objective is to guarantee reliable and secure signal transmission regardless of which state occurs. This is a practical model for channel uncertainty, where the transmitter may have no knowledge of the channel (due to the dynamic nature of the wireless medium or unavoidable implementation/estimation inaccuracies), but zero performance outage is still required (e.g., for ultra-reliable communications [95]). [123] proposed achievable and converse results, with the converse bounds shown to be tight in certain cases by [17]. They also studied the achievable secrecy degrees of freedom (s.d.o.f.) region for a multi-input multi-output (MIMO) model, which was later extended to the case of two confidential messages in [103]. The s.d.o.f. of compound wiretap parallel

channels was also studied in [131]. See [53, 54] for discussions on Gaussian MIMO compound wiretap channels.

In [17, 123], the results focused on discrete memoryless channels with a countably finite uncertainty set (i.e., the set from which the exact channel realization is drawn). This was later extended to *arbitrary uncertainty sets*, including continuous alphabets, by [152], which is also one of our contributions. Moreover, [19] showed that the secrecy capacity is a continuous function of the uncertainty set.

4.3 One-shot Generalized Information Hiding

In this section, we formulate the generalized one-shot information hiding problem, which is more general than the one-shot information hiding setting in [134] and can be seen as a natural generalization of [134], [183], and [146]. More specifically, as an extension of our conference version [134], we adopt the idea from [146] that considers the side information available at both the encoder and the decoder.[2] Moreover, similar to [183], we require the decoder to not only reconstruct the message M, but also the stegotext X.

Our generalized information hiding problem recovers many existing settings as special cases, including the conventional information hiding problem [134, 144, 157], the information embedding with stegotext reconstruction problem [78, 183], the conventional Gelfand-Pinsker coding [74, 90], the generalized Gelfand–Pinsker family [146] and the compound channel [18, 45, 148, 179], and also the special cases recovered therein.

[2]In [146], it was assumed that there were three sources of side information, available at the encoder, the attacker, and the decoder, respectively. We model this scenario by considering sources of side information available at the encoder and the decoder, together with the attack channel being in an unspecified set, since the decoder has no knowledge on the attacker. We can recover the Gelfand-Pinsker coding [74, 90] by letting $\mathcal{A} = \{A_{Y|X}\}$ be a singleton set.

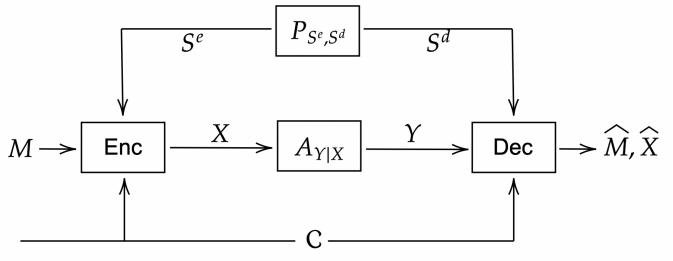

Figure 4.1: One-shot generalized information hiding setting.

4.3.1 Problem Formulation

The one-shot generalized information hiding problem is shown in Figure 4.1. The goal is to hide a message M into a host signal S^e, so that even though there is an attacker during the signal transmission which aims at removing the hidden message, the decoder can still reconstruct the original message and also the stegotext X, within a range defined by the distortion functions. We further elaborate their roles and assumptions in detail as follows.

- **Encoder**: The encoder observes a message M that is uniformly chosen from the set $[1 : \mathsf{L}]$, and the goal of encoding is to *hide* M into a host data source $S^e \in \mathcal{S}^e$ by introducing some tolerable level of distortions. Given S^e and M, the encoding function $f : \mathcal{S}^e \times [1 : \mathsf{L}] \to \mathcal{X}$ outputs $X = f(S^e, M)$. It is expected that X is close to S^e, in the sense that $d_1(S, X)$ is small, where $d_1 : \mathcal{S}^e \times \mathcal{X} \to [0, \infty)$ is a distortion measure. We want $d_1(S, X) \leq \mathsf{D}_1$ with high probability. This will be elaborated later. The encoded signal X is then transmitted through the a channel $A_{Y|X} \in \mathcal{A}$.

- **Attacker**: The attacker is formulated as a noisy channel $P_{Y|X}$. With input X, it performs data processing attacks on X by introducing another level of distortion and produces Y, a corrupted version of X. Its objective is to (partially) remove or degrade the message and/or the stegotext X, so

that the decoder cannot have a correct reconstruction with a high fidelity. Unlike the conventional asymptotic information hiding [144], we do not assume the attacker's strategy is known by the encoder and the decoder. Instead, the attacker is free to choose from a class of channels $\mathcal{A}$ (e.g., the class of channels satisfying some distortion constraint between X and Y, or the class of memoryless channels in case X and Y are sequences). Both deterministic attacks or randomized attacks could be performed. We assume the attacker has knowledge of the distributions (but not the values) of M, S^e, S^d, and also the code $\mathcal{C}$ that is used by the encoder and the decoder.

- **Decoder**: The decoder observes Y, the output of the attacker, together with another source of side information S^d, and computes $(\hat{M}, \hat{X}) = \phi(S^d, Y)$, the distorted versions of M and X, where $\phi : \mathcal{S}^d \times \mathcal{Y} \to [1 : \mathsf{L}] \times \mathcal{X}$. When the stogetext X is expected to be reconstructed, we expect $d_2(X, \hat{X})$ is small, where $d_2 : \mathcal{X} \times \mathcal{X} \to [0, \infty)$ is another distortion measure. Since we assume that the decoder is *uninformed* of the attacker's strategy (different from [78, 144, 183]), we intend to bound the encoder-decoder team's worst case failure probability

$$P_e := \sup_{A_{Y|X} \in \mathcal{A}} \mathbf{P}\big(d_1(S, X) > \mathsf{D}_1 \ \text{ OR } \ d_2(X, \hat{X}) > \mathsf{D}_2 \ \text{ OR } \ M \neq \hat{M}\big). \quad (4.1)$$

to be small, where we assume $(S^e, S^d, M) \sim P_{S^e,S^d} \times \mathrm{Unif}[\mathsf{L}]$, $X = f(S^e, M)$, $Y|X \sim A_{Y|X}$ and $(\hat{M}, \hat{X}) = \phi(S^d, Y)$ in the probability.[3]

[3]Note that [144] imposes a constraint on the expected distortion $\mathbf{E}[d_1(S, X)]$, which is reasonable in the context of [144] because the memoryless assumption and the law of large numbers ensure that the actual distortion is close to the expected distortion. Since we are considering a one-shot setting where we only assume the attack channel is chosen from a set $\mathcal{A}$, if constraint need to be specified, it might be more reasonable to consider $d_1(S, X) > \mathsf{D}_1$ as a failure event

- **Side information**: The side information S_e, S_a, S_d can be viewed as certain *common randomness* (or some resource) available at the encoder, the attacker and the decoder, respectively. The joint distribution P_{S^e,S^d} reveals information about the host data source S^e to the decoder.

Remark 4.3.1. Our formulation can be viewed as a one-shot version of the generalized Gelfand-Pinsker problem [146] (which only considered discrete case though), or a one-shot compound channel with side information at the encoder and/or the decoder.

Remark 4.3.2. As noted in [144, 158] (also see [35, 157]), these settings can be viewed as a *game* between two parties: the first party consists of the encoder (information hider) and the decoder, who are cooperatively transmitting the message M; the second party is an attacker, who is trying to destroy or degrade the hidden message M in S^e so that the decoder cannot correctly decode M or reconstruct a good $\hat{X}$. More discussions on such game-theoretic perspective can be found in [144].

4.3.2 One-shot Achievability Results

We then provide one-shot achievability results of the generalized information hiding problem.

Note that in one-shot settings, as we discussed above, the techniques in [144, 158] (e.g., the tools based on the *typical sets*), which have resemblances to the Gelfand-Pinsker coding [74, 90] as discussed in [144]) are not suitable. Similar to Chapter 3, we utilize Poisson Matching Lemma has been shown to perform well in various one-shot settings [117] and was been introduced in 2.3 as one part of

and bound the probability of failure, i.e., the excess distortion probability instead, compared to expected distortion.

our proof technique. Our one-shot results apply to both discrete and continuous cases.

Briefly recall Chapter 2. Fix a distribution Q over $\mathcal{U}$. Let $(T_i)_{i=1,2,\dots}$ be a Poisson process with rate 1. Let $\mathbf{U} := (\bar{U}_i)_i$ be an independent i.i.d. sequence with distribution Q. The "marked" Poisson process $(\bar{U}_i, T_i)_i$ supports a "query operation" given by the Poisson functional representation, where one can input a distribution P over $\mathcal{U}$, and obtain one sample $\tilde{U}_P$ with distribution P. The Poisson functional representation is given by

$$\mathbf{U}_P := \bar{U}_K, \quad \text{where } K := \arg\min_i T_i \cdot \Big(\frac{\mathrm{d}P}{\mathrm{d}Q}(\bar{U}_i)\Big)^{-1}.$$

Since we let the encoder-decoder team account for all possible attack channels in a set $\mathcal{A}$, the achievability results have to suffer a penalty depending on the "size" of $\mathcal{A}$. Though the cardinality of $\mathcal{A}$ could be infinite, we can often find a finite subset $\tilde{\mathcal{A}}$ such that every attack channel $A \in \mathcal{A}$ is close enough to some $\tilde{A} \in \tilde{\mathcal{A}}$. We capture this notion of size by the ϵ-covering number defined below (see similar covering arguments in [18, 144]).

Definition 4.3.3. Given a set of channels $\mathcal{A}$ from $\mathcal{X}$ to $\mathcal{Y}$, its *ϵ-covering number* is defined as

$$N_\epsilon(\mathcal{A}) := \min\Big\{|\tilde{\mathcal{A}}| : \tilde{\mathcal{A}} \subseteq \mathcal{A}, \sup_{A\in\mathcal{A}} \min_{\tilde{A}\in\tilde{\mathcal{A}}} \sup_{x\in\mathcal{X}} \Big\|A_{Y|X}(\cdot|x) - \tilde{A}_{Y|X}(\cdot|x)\Big\|_{\mathrm{TV}} \le \epsilon\Big\},$$

where $\|A_{Y|X}(\cdot|x) - \tilde{A}_{Y|X}(\cdot|x)\|_{\mathrm{TV}} \in [0,1]$ denotes the total variation distance between $A_{Y|X}(\cdot|x)$ (the distribution of Y if $X = x$, and Y follows $A_{Y|X}$) and $\tilde{A}_{Y|X}(\cdot|x)$.

We now present the main result, which is a one-shot achievability result with a bound on the error probability in terms of $N_\epsilon(\mathcal{A})$ and information density terms.

Theorem 4.3.4. *Fix any $P_{U,X|S^e,S^d}$ and channel $\hat{A}_{Y|X}$. For any $\epsilon \geq 0$, there exists a scheme for the generalized information hiding problem satisfying*

$$P_e \leq N_\epsilon(\mathcal{A}) \sup_{A \in \mathcal{A}} \mathbf{E}_{Y|X\sim A}\Bigg[1 - \mathbf{1}\{d_1(S^e, X) \leq \mathsf{D}_1\} \\ \cdot \mathbf{1}\{d_2(X, \hat{X}) \leq \mathsf{D}_2\} \cdot \left(1 + \mathsf{L} \cdot 2^{-\hat{\imath}(U;Y,S^d)+\imath(U;S^e)}\right)^{-1}\Bigg] + \epsilon,$$

where we assume $(S^e, S^d, U, X, Y) \sim P_{S^e,S^d} P_{U,X|S^e} A_{Y|X}$ in the expectation, and $\hat{\imath}(U; Y, S^d)$ is the information density computed by the joint distribution $P_{S,K} P_{U,X|S,K} \hat{A}_{Y|X}$ (instead of $A_{Y|X}$), assuming that $\imath(U; S^e), \hat{\imath}(U; Y, S^d)$ are almost surely finite for every $A_{Y|X} \in \mathcal{A}$.

Proof. The idea is that we design the decoder assuming that the attack channel is fixed to $A_{Y|X}$, and hope that this decoder works for every attack channel $A_{Y|X} \in \mathcal{A}$. Let $\mathcal{C} := ((\bar{U}_i, \bar{M}_i), T_i)_i$ where $(T_i)_i$ is a Poisson process, $\bar{U}_i \overset{\text{iid}}{\sim} P_U$, and $\bar{M}_i \overset{\text{iid}}{\sim} P_M$ (where $P_M = \text{Unif}[\mathsf{L}]$). This will act as a random codebook shared between the encoder and the decoder (and this codebook will be fixed later).

The encoder observes the message $M \sim P_M$ and the encoder-side host signal S^e, by the Poisson functional representation [117, 118] on the distribution $P_{U|S^e}(\cdot|S^e) \times \delta_M$ over $\mathcal{U} \times [1 : \mathsf{L}]$ it produces $U = \mathbf{U}_{P_{U|S^e}(\cdot|S^e)\times\delta_M}$,[4] and sends the generated $X|(S^e, U) \sim P_{X|S^e,U}$. The decoder observes Y, S^d, outputs $\hat{M} = \mathbf{M}_{\hat{P}_{U|Y,S^d}(\cdot|Y,K)\times P_M}$ by the Poisson functional representation, and computes the reconstruction sequence $\hat{X}$ by $\hat{X} = \hat{\mathbf{x}}(\hat{M}, Y)$, where $\hat{P}_{U|Y,S^d}$ is the conditional distribution computed by the joint distribution $P_{S^e,S^d} P_{U,X|S^e} \hat{A}_{Y|X}$. When the attack channel is $A_{Y|X} \in \mathcal{A}$, the error probability is bounded as follows:

$$P_e(A) := 1 - \mathbf{P}_{Y|X\sim A_{Y|X}}\big(d_1(S^e, X) \leq \mathsf{D}_1 \ \text{ AND } \ M = \hat{M}\big)$$

[4] The Poisson functional representation produces a pair $(\mathbf{U}, \mathbf{M})$, and U is set to the first component of the pair.

$$
\begin{aligned}
&= \mathbf{E}\Big[1 - \mathbf{1}\{d_1(S^e, X) \le \mathsf{D}_1\} \\
&\qquad\qquad \cdot \mathbf{1}\{d_2(X, \hat{X}) \le \mathsf{D}_2\} \cdot \mathbf{1}\{M = \hat{M}\}\Big] \\
&\le \mathbf{E}\Big[1 - \mathbf{1}\{d_1(S^e, X) \le \mathsf{D}_1\} \cdot \mathbf{1}\{d_2(X, \hat{X}) \le \mathsf{D}_2\} \\
&\qquad\qquad \cdot \mathbf{P}\big(M = \hat{M} | M, S^e, S^d, U, Y\big)\Big] \\
&\le \mathbf{E}\Big[1 - \mathbf{1}\{d_1(S^e, X) \le \mathsf{D}_1\} \cdot \mathbf{1}\{d_2(X, \hat{X}) \le \mathsf{D}_2\} \\
&\cdot \mathbf{P}\big((U, M) = (\mathbf{U}, \mathbf{M})_{\hat{P}_{U|Y,S^d}(\cdot|Y,S^d) \times P_M} | M, S^e, S^d, U, Y\big)\Big] \\
&\overset{(a)}{\le} \mathbf{E}\Big[1 - \mathbf{1}\{d_1(S^e, X) \le \mathsf{D}_1\} \cdot \mathbf{1}\{d_2(X, \hat{X}) \le \mathsf{D}_2\} \\
&\qquad\qquad \cdot \Big(1 + \frac{\mathrm{d}P_{U|S^e}(\cdot|S^e) \times \delta_M}{\mathrm{d}\hat{P}_{U|Y,S^d}(\cdot|Y, S^d) \times P_M}(U, M)\Big)^{-1}\Big] \\
&= \mathbf{E}\Big[1 - \mathbf{1}\{d_1(S^e, X) \le \mathsf{D}_1\} \cdot \mathbf{1}\{d_2(X, \hat{X}) \le \mathsf{D}_2\} \\
&\qquad\qquad \cdot \Big(1 + \mathsf{L} \cdot \frac{P_{U|S^e}(\cdot|S^e)}{\hat{P}_{U|Y,S^d}(\cdot|Y, S^d)}(U)\Big)^{-1}\Big] \\
&= \mathbf{E}\Big[1 - \mathbf{1}\{d_1(S^e, X) \le \mathsf{D}_1\} \cdot \mathbf{1}\{d_2(X, \hat{X}) \le \mathsf{D}_2\} \\
&\qquad\qquad \cdot \Big(1 + \mathsf{L} \cdot 2^{-\hat{\imath}(U;Y,S^d) + \imath(U;S^e)}\Big)^{-1}\Big] \\
&\le \sup_{A_{Y|X} \in \mathcal{A}} \mathbf{E}_{Y|X \sim A_{Y|X}}\Big[1 - \mathbf{1}\{d_1(S^e, X) \le \mathsf{D}_1\} \\
&\qquad \cdot \mathbf{1}\{d_2(X, \hat{X}) \le \mathsf{D}_2\} \cdot \Big(1 + \mathsf{L} \cdot 2^{-\hat{\imath}(U;Y,S^d) + \imath(U;S^e)}\Big)^{-1}\Big]
\end{aligned}
$$

where (a) is by the Poisson matching lemma.[5] If we allow the encoder and the decoder to share unlimited additional common randomness, we can assume the codebook $\mathcal{C} = ((\bar{U}_i, \bar{M}_i), T_i)_i$ is actually shared, and conclude that $P_e =$

[5]The Poisson matching lemma is applied on the conditional distributions given M, S^e, S^d, U, Y. Also see the conditional Poisson matching lemma [117].

$\sup_{A\in\mathcal{A}} P_e(A) \le \overline{P_e}$. Nevertheless, the only actual common randomness between the encoder and the decoder is K, which we cannot control. Therefore, we have to fix the codebook.

Let $P_e(A,c)$ be the probability of error when the attack channel is A and the codebook is $\mathcal{C}=c$. We have $P_e(A)=\mathbf{E}_{\mathcal{C}}[P_e(A,\mathcal{C})]$. Let $\tilde{\mathcal{A}}\subseteq\mathcal{A}$ attain the minimum in $N_\epsilon(\mathcal{A})$.

Consider any $A\in\mathcal{A}$, and let $\tilde{A}\in\tilde{\mathcal{A}}$ satisfy

$$\sup_{x\in\mathcal{X}} \left\| A_{Y|X}(\cdot|x) - \tilde{A}_{Y|X}(\cdot|x) \right\|_{\mathrm{TV}} \le \epsilon.$$

The total variation distance between the joint distribution of M,S,K,U,X,Y under the attack channel A conditional on $\mathcal{C}=c$ and the joint distribution under the attack channel $\tilde{A}$ conditional on $\mathcal{C}=c$ is also bounded by ϵ. Hence $|P_e(A,c)-P_e(\tilde{A},c)|\le\epsilon$ and

$$\begin{aligned} P_e(A,c) &\le P_e(\tilde{A},c)+\epsilon \\ &\le \sum_{\tilde{A}\in\tilde{\mathcal{A}}} P_e(\tilde{A},c)+\epsilon. \end{aligned}$$

Therefore,

$$\begin{aligned} \mathbf{E}_{\mathcal{C}}\Big[\sup_{A\in\mathcal{A}} P_e(A,\mathcal{C})\Big] &\le \mathbf{E}_{\mathcal{C}}\Big[\sum_{\tilde{A}\in\tilde{\mathcal{A}}} P_e(\tilde{A},\mathcal{C})+\epsilon\Big] \\ &= \sum_{\tilde{A}\in\tilde{\mathcal{A}}} P_e(\tilde{A})+\epsilon \\ &\le |\tilde{\mathcal{A}}|\cdot\overline{P_e}+\epsilon. \end{aligned}$$

The proof is completed by the existence of a codebook c such that

$$\sup_{A\in\mathcal{A}} P_e(A,c) \le |\tilde{\mathcal{A}}|\cdot\overline{P_e}+\epsilon.$$

□

Remark 4.3.5. It is straightforward to convert this to a finite blocklength result where n is a fixed number using the Berry-Esseen theorem [15, 62].

Remark 4.3.6. In Theorem 4.3.4, we use a penalty term $N_\epsilon(\mathcal{A})$ to measure the effect of the "size" of $\mathcal{A}$, which introduces a degradation in the error probability. The choice of ϵ can be viewed as a way to balance the two terms in Theorem 4.3.4: increasing ϵ will result in a larger ϵ but a smaller $N_\epsilon(\mathcal{A})$ (see also Proposition 4.4.1). Although directly investigating $N_\epsilon(\mathcal{A})$ following Definition 4.3.3 may not be straightforward, it is possible to optimize the one-shot bound in Theorem 4.3.4 with respect to ϵ and the bound on $N_\epsilon(\mathcal{A})$ in Proposition 4.4.1. We leave more detailed analysis of these manipulations and potential second-order results as future work. For now, we only require that our one-shot bound suffices to recover the asymptotic hiding capacity [144] when applied to discrete memoryless channels, as demonstrated in Section 4.4 where we take $\epsilon = 1/n$ in the asymptotic analysis.

Remark 4.3.7. Note that when $S^d = S^e = \emptyset$, $d_1(s, x) = 0$, and $\mathcal{A} = \{A_{Y|X}\}$ is a singleton set, taking $\hat{A}_{Y|X} = A_{Y|X}$, Theorem 4.3.4 reduces to the one-shot Gelfand-Pinsker coding result in [117].

4.3.3 Discussions

In [144], it is assumed that the attack channel must be memoryless, and hence the decoder can obtain full knowledge about the attack channel, justified by the large blocklength of signals. In this paper, similar to [146, 158] (which focus on different targets or are under settings different to us), we drop this assumption, and consider a one-shot setting where the set of possible attack channels $\mathcal{A}$ can be *any* set of channels. Also, we do not assume that the decoder knows the attack channel, which is unrealistic in the one-shot setting where the attacker can be arbitrary. In [158] (which is a specialized information hiding setting that is similar

to Section 4.4) the memoryless assumption is also dropped, and an asymptotic hiding capacity expressed as the limit of a sequence of single-letter expressions has been derived using constant composition codes. The key difference between [158] and our setting in Section 4.4 (and also [144]) is that the side information in [158] is a shared key of unlimited size independent of M, S^e that can be chosen as a part of the coding scheme, whereas in our paper and [144] the side information is given and may be correlated with the host signal (where the dependence is from the joint distribution), and cannot be changed. In some watermarking problems [37, 84] certain components can be further constrained, e.g., there may exist a mapping from the message M to a codeword $V(M)$ which is independent of the host, and then composite data are obtained by a mapping from $V(M)$ and the side information.

The information hiding can be regarded as a variant of Gelfand-Pinsker coding for channels with side information at the encoder [74, 90], where the channel is fixed and not chosen by the attacker, and there is no shared side information between the encoder and the decoder. Since the encoder and the decoder have to account for all possible attack channels, this can be regarded as a combination of Gelfand-Pinsker coding and compound channel [18, 45, 179]. The analyses in [144, 158] utilize techniques such as random binning, joint typicality decoding and constant composition codes, which are also commonly utilized in the asymptotic analyses of Gelfand-Pinsker coding [74, 151]. These techniques may not be suitable for our one-shot setting. Strong typicality and constant composition codes are inapplicable when the blocklength is 1. While random binning can be applied to one-shot Gelfand-Pinsker coding [172, 177, 189], it produces weaker results compared to the Poisson matching lemma [117]. To obtain tight one-shot bounds for information hiding, we utilize the Poisson matching lemma instead. The main tool used to prove the coding theorems of generalized Gelfand-

Pinsker problems [146] is also the method of types [38], which does not work in the one-shot analysis in general.

4.4 Recovery of the Asymptotic Information Hiding

In this section, we discuss a special case of our generalized information hiding setting, which is the information hiding setting that was investigated in [144]. We show that one-shot achievability results readily recover their asymptotic results on this setting, when we apply our results on discrete and memoryless channels.

We first provide a simple bound on the ϵ-covering number in the case that X and Y are discrete and finite.

Proposition 4.4.1. *If $\mathcal{X}$ and $\mathcal{Y}$ are finite, then*

$$N_\epsilon(\mathcal{A}) \leq \left(\frac{1}{2\epsilon} + \frac{|\mathcal{Y}|+1}{2}\right)^{|\mathcal{X}|\cdot|\mathcal{Y}|}.$$

The proof can be found in Appendix B.1.

We show that we recover the information hiding capacity that was discovered by [144]. We can employ similar procedure to recover either the achievable bound of information hiding with stegotext reconstruction [183, Theorem 1] (which in turn is an extension of [78] and [162]), or the similar bounds in [146]. For the simplicity, we only show the details of recovering the hiding capacity in [144] here.

The setting is shown in Figure 4.2, where there exist a host signal S available to the encoder, in which the encoder hides the message, and another source of side information K that is available to the encoder and the decoder. By letting $S_e := (S, K)$ and $S_d := K$, we now show that Theorem 4.3.4 recovers the asymptotic result in [144] when S, K, X, Y are finite and discrete, and the attack

channel must be memoryless and is subject to a distortion constraint, and hence giving a simple alternative proof to [144].

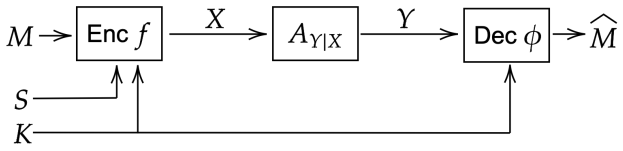

Figure 4.2: Information hiding setting [144, 158].

Consider sequences $S^n = (S_1, \ldots, S_n)$, K^n, X^n, Y^n where $(S_i, K_i) \overset{\text{iid}}{\sim} P_{S,K}$. Consider a channel input distribution P_X. The class of attack channels $\mathcal{A}_n = \mathcal{A}_n(P_X)$ (which depends on P_X) is taken to be

$$\mathcal{A}_n(P_X) := \big\{A^n_{Y|X} : A_{Y|X} \in \mathcal{A}(P_X)\big\},$$

and we let

$$\mathcal{A}(P_X) := \big\{A_{Y|X} : \mathbf{E}_{(X,Y)\sim P_X A_{Y|X}}[d_3(X,Y)] \le \mathsf{D}_3\big\},$$

where $d_3 : \mathcal{X}\times\mathcal{Y} \to [0,\infty)$ is a distortion measure, and D_3 is the allowed distortion level. In other words, the attacker can only use memoryless channels $A^n_{Y|X}$ that satisfy the expected distortion constraint $\mathbf{E}[d_3(X,Y)] \le \mathsf{D}_3$. The asymptotic hiding capacity given in [144] is

$$C = \max_{P_{U,X|S,K}} \min_{A_{Y|X}:\mathbf{E}[d_2(X,Y)]\le \mathsf{D}_3} \big(I(U;Y|K) - I(U;S|K)\big).$$

where the maximum is over $P_{U,X|S,K}$ with $\mathbf{E}[d_1(S,X)] \le \mathsf{D}_1$.

We now show the achievability of the above asymptotic rate as a direct corollary of Theorem 4.3.4. Fix $P_{U,X|S,K}$ which achieves the above maximum subject to $\mathbf{E}[d_1(S,X)] \le \mathsf{D}'_1$ where $\mathsf{D}'_1 < \mathsf{D}_1$. Take $\hat{A}_{Y|X}$ to be the minimizer of the rate-distortion function $\min_{A_{Y|X}:\mathbf{E}[d_2(X,Y)]\le \mathsf{D}_2} I(U;Y|K)$, and assume $(S,K,U,X,Y) \sim$

$P_{S,K}P_{U,X|S,K}\hat{A}_{Y|X}$. Write the information density and mutual information obtained from this distribution as $\hat{\imath}_{U;Y|K}$ and $\hat{I}(U;Y|K)$, respectively. Fix a coding rate $R < \hat{I}(U;Y|K) - I(U;S|K)$. We want to show that this rate is achievable.

Consider any attack channel $A_{Y|X}$ with $\mathbf{E}[d_3(X,Y)] \leq \mathsf{D}_3$. Let $A^{\lambda}_{Y|X} := (1-\lambda)\hat{A}_{Y|X} + \lambda A_{Y|X}$ for $0 \leq \lambda \leq 1$. Write $I_\lambda(U;Y|K)$ for the mutual information computed assuming $Y|X \sim A^{\lambda}_{Y|X}$. It is straightforward to check that

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} I_\lambda(U;Y|K)\Big|_{\lambda=0} = \mathbf{E}_{Y|X\sim A_{Y|X}}[\hat{\imath}(U;Y|K)] - \hat{I}(U;Y|K).$$

By the optimality of $\hat{A}$, the above derivative is nonnegative, and hence

$$\mathbf{E}_{Y|X\sim A_{Y|X}}[\hat{\imath}(U;Y|K)] \geq \hat{I}(U;Y|K).$$

Therefore, when we have i.i.d. sequences $(S^n, K^n, U^n, X^n, Y^n) \sim P^n_{S,K} P^n_{U,X|S,K} A^n_{Y|X}$ and $\mathsf{L} = \lfloor 2^{nR} \rfloor$, by the law of large numbers,

$$\begin{aligned}
&\mathsf{L} 2^{-\hat{\imath}(U^n;Y^n|K^n)+\imath(U^n;S^n|K^n)} \\
&\leq 2^{nR - \sum_{i=1}^n (\hat{\imath}(U_i;Y_i|K_i) - \imath(U_i;S_i|K_i))} \\
&\to\ 0
\end{aligned}$$

exponentially as $n \to \infty$ since

$$\begin{aligned}
&\mathbf{E}[\hat{\imath}(U_i;Y_i|K_i) - \imath(U_i;S_i|K_i))] \\
&\geq \hat{I}(U;Y|K) - I(U;S|K) \\
&> R
\end{aligned}$$

We also have

$$d_1(S^n, X^n) = \sum_{i=1}^n d_1(S_i, X_i) > n\mathsf{D}_1$$

with probability approaching 0 exponentially since $\mathsf{D}'_1 < \mathsf{D}_1$. These convergences are uniform over all such attack channels $A_{Y|X}$ since the random variables are discrete and finite.

Therefore, to bound P_e using Theorem 4.3.4, it is left to bound the ϵ-covering number $N_\epsilon(\mathcal{A}_n(P_X))$. Note that

$$\begin{aligned}&\left\|A^n_{Y|X}(\cdot|x^n) - \tilde{A}^n_{Y|X}(\cdot|x^n)\right\|_{\mathrm{TV}}\\ &\leq \sum_{i=1}^n \left\|A_{Y|X}(\cdot|x_i) - \tilde{A}_{Y|X}(\cdot|x_i)\right\|_{\mathrm{TV}},\end{aligned}$$

and hence we can construct an ϵ-cover of $\mathcal{A}_n(P_X)$ using an (ϵ/n)-cover of $\mathcal{A}(P_X)$. Therefore,

$$\begin{aligned}N_\epsilon(\mathcal{A}_n(P_X)) &\leq N_{\epsilon/n}(\mathcal{A}(P_X))\\ &= O((n/\epsilon)^{|\mathcal{X}|\cdot|\mathcal{Y}|})\end{aligned}$$

by Proposition 4.4.1, which grows much slower than the exponential decrease of the expectation in Theorem 4.3.4. Therefore, taking $\epsilon = 1/n$, we have $P_e \to 0$ as $n \to \infty$. Taking $\mathsf{D}_1' \to \mathsf{D}_1$ completes the proof.

4.5 One-shot Compound Wiretap Channels

In this section, we consider the compound wiretap channel [123] in the one-shot setting. We utilize the Poisson matching lemma [117], under a framework similar to the one-shot codes of the information hiding problem. We provide novel one-shot achievablity results for the compound wiretap channel [123]. To the best of our knowledge, the one-shot results of this problem has not been discussed in literature, though finite-blocklength bounds on single (without channel uncertainties) wiretap channels can be found in [87, 130, 186, 187].

Unlike the asymptotic analysis of the compound wiretap channel [123], our results also apply to continuous cases. Note that [152] also studied the continuous case of compound wiretap channels, but the focus in [152] was mainly on the compound Gaussian MIMO wiretap channels, and the analysis was not one-shot.

In modern wireless communication, handling continuous cases can be essential in various applications for capturing the inherent variability and nuances of real-world signal propagation, which has a dynamic nature.

4.5.1 Problem Formulation

The one-shot compound wiretap channel setting is described as follows. A message M is uniformly chosen from Unif[L]. Upon observing $M \sim$ Unif[L], the encoder produces $X = f(M)$, where $f : [\mathsf{L}] \to \mathcal{X}$ is a randomized encoding function. Then X is sent through a channel $P_{Y,Z|X}$ that is unknown to the encoder and the decoder but known to the eavesdropper. A legitimate decoder observes Y and recovers $\hat{M} = g(Y)$, where $g : \mathcal{Y} \to [\mathsf{L}]$ is a decoding function. The eavesdropper observes $Z \in \mathcal{Z}$. Justified by [123] and [124, Lemma 1], we can assume the transition probability distribution is $P_{Y|X}P_{Z|X}$ by decomposing $P_{Y,Z|X}$ without loss of optimality.

We assume $P_{Y|X}$ is from a set $\mathcal{D}$ for **d**ecoding, while $P_{Z|X}$ is from a set $\mathcal{E}$ for **e**avesdropping. Unlike [123], we assume $\mathcal{D}, \mathcal{E}$ can be infinite, which captures the infinite variability of real-world signals and their propagation characteristics in practical applications. Even though their cardinalities can be infinite, we can often find a finite subset $\tilde{\mathcal{D}}$ (or $\tilde{\mathcal{E}}$) such that every receiver (or eavesdropper) in $\tilde{\mathcal{D}}$ (or $\tilde{\mathcal{E}}$) would be close enough to some $\tilde{D} \in \tilde{\mathcal{D}}$ (or $\tilde{E} \in \tilde{\mathcal{E}}$). This idea has appeared in Section 4.3.2 and also in [152].

The objective is to bound the worst case error probability

$$P_e := \sup_{P_{Y|X} \in \mathcal{D}} \mathbf{P}\left(M \neq \hat{M}\right), \tag{4.2}$$

while also ensure the secrecy is guaranteed, which is measured by the total variation distance

$$\gamma := \sup_{P_{Z|X} \in \mathcal{E}} \|P_{M,Z} - P_M \times P_Z\|_{\mathrm{TV}} \tag{4.3}$$

being small.

4.5.2 One-Shot Achievability Results

We then provide the one-shot achievability results of the compound wiretap channel. Note the result can be viewed as a combination of the covering argument appeared in Section 4.3 and the one-shot soft covering lemma in [117, Proposition 3]. Other existing one-shot wiretap channel results [87, 130, 187] might also be utilized in a similar framework as well.

Theorem 4.5.1. *Fix any $P_{U,X}$ and any wiretap channel $\hat{P}_{Y|X}\hat{P}_{Z|X}$. For any $\nu \geq 0$, any $\epsilon_1, \epsilon_2 \geq 0$ and $\mathsf{A}, \mathsf{B} \in \mathbb{N}$, there exists a code for the compound wiretap channel setting, with message $M \sim \mathrm{Unif}[\mathsf{L}]$, satisfying*

$$\begin{aligned}
&P_e + \nu \cdot \gamma \\
&\leq N_{\epsilon_1}(\mathcal{D}) \sup_{P_{Y|X} \in \mathcal{D}} \mathbf{E}_{Y|X \sim P_{Y|X}} \left[\min \left\{ \mathsf{L}\mathsf{A} 2^{-\hat{\imath}(U;Y)}, 1 \right\} \right] + \epsilon_1 \\
&+ \nu \cdot N_{\epsilon_2}(\mathcal{E}) \left(\sup_{P_{Z|X} \in \mathcal{E}} 2 \cdot \mathbf{E}_{Z|X \sim P_{Z|X}} \left[\left(1 + 2^{-\hat{\imath}(U;Z)}\right)^{-\mathsf{B}} \right] + \sqrt{\mathsf{B}\mathsf{A}^{-1}} \right) + \nu \cdot \epsilon_2,
\end{aligned}$$

where we assume $(U, X, Y, Z) \sim P_{U,X} P_{Y|X} P_{Z|X}$ in the expectation, and $\hat{\imath}(U;Y)$, $\hat{\imath}(U;Z)$ are the information densities computed for compound channels by the joint distribution $P_{U,X}\hat{P}_{Y|X}\hat{P}_{Z|X}$, assuming that $\hat{\imath}(U;Y), \hat{\imath}(U;Z)$ are almost surely finite for every $P_{Y|X} \in \mathcal{D}$, $P_{Z|X} \in \mathcal{E}$.

Proof. We first design our coding strategy assuming that the transmission channel is fixed to $\hat{P}_{X|Y} \in \mathcal{D}$ and the eavesdropping channel is fixed to $\hat{P}_{X|Z} \in \mathcal{E}$.

Let $\mathcal{C} := ((\bar{U}_i, \bar{M}_i), T_i)_i$ where $(T_i)_i$ is a Poisson process, $\bar{U}_i \overset{\text{iid}}{\sim} P_U$, and $\bar{M}_i \overset{\text{iid}}{\sim} P_M$ (where $P_M = \mathrm{Unif}[\mathsf{L}]$). This is a random codebook that is known to both the encoder and the decoder, and it will be fixed later.

Let $A \sim \mathrm{Unif}[\mathsf{A}]$ be independent of $(M, \mathcal{C})$. The encoder observes the message $M \sim P_M$, computes $U = \mathbf{U}_{P_U \times \delta_M}(A)$, and sends the generated $X|U \sim$

$P_{X|U}$. The decoder observes Y and recovers $\hat{M} = \mathbf{M}_{\hat{P}_{U|Y}(\cdot|Y)\times P_M}$. We have $(M, A, U, X, Y, Z) \sim P_M \times P_A \times P_{U,X}\hat{P}_{Y|X}\hat{P}_{Z|X}$. For the fixed $\hat{P}_{Y|X} \in \mathcal{D}$, we have:

$$
\begin{aligned}
&\mathbf{P}\left\{M \neq \hat{M}\right\} \\
&\leq \mathbf{E}\left[\mathbf{P}\big((U,M) \neq (\mathbf{U},\mathbf{M})_{\hat{P}_{U|Y}(\cdot|Y)\times P_M}|M,A,U,Y\big)\right] \\
&\overset{(a)}{\leq} \mathbf{E}\left[\min\left\{\mathsf{A}\frac{\mathrm{d}P_U \times \delta_M}{\mathrm{d}\hat{P}_{U|Y}(\cdot|Y)\times P_M}(U,M), 1\right\}\right] \\
&= \mathbf{E}\left[\min\left\{\mathsf{LA}2^{-\hat{\imath}_{U;Y}(U;Y)}, 1\right\}\right] \\
&\leq \sup_{P_{Y|X}\in\mathcal{D}} \mathbf{E}_{Y|X\sim P_{Y|X}}\left[\min\left\{\mathsf{LA}2^{-\hat{\imath}_{U;Y}(U;Y)}, 1\right\}\right] \qquad (4.4) \\
&=: \overline{P_e}
\end{aligned}
$$

where (a) is by the conditional generalized Poisson matching lemma [117] applied on $(M, A, (U, M), Y, \hat{P}_{U|Y} \times P_M)$, and we define (4.4) to be $\overline{P_e}$.

For the secrecy measure, for the fixed $\hat{P}_{Z|X} \in \mathcal{E}$, we have:

$$
\begin{aligned}
&\mathbf{E}\left[\left\|\hat{P}_{M,Z|\mathcal{C}}(\cdot,\cdot|\mathcal{C}) - P_M \times \hat{P}_{Z|\mathcal{C}}(\cdot|\mathcal{C})\right\|_{\mathrm{TV}}\right] \\
&= \mathbf{E}\left[\left\|\hat{P}_{Z|M,\mathcal{C}}(\cdot,\cdot|\mathcal{C}) - \hat{P}_{Z|\mathcal{C}}(\cdot|\mathcal{C})\right\|_{\mathrm{TV}}\right] \\
&\leq \mathbf{E}\left[\left\|\hat{P}_{Z|M,\mathcal{C}}(\cdot,\cdot|\mathcal{C}) - \hat{P}_{Z}(\cdot)\right\|_{\mathrm{TV}}\right] + \mathbf{E}\left[\left\|\hat{P}_{Z|\mathcal{C}}(\cdot|\mathcal{C}) - \hat{P}_{Z}(\cdot)\right\|_{\mathrm{TV}}\right] \\
&\overset{(a)}{\leq} 2\cdot\mathbf{E}\left[\left\|\hat{P}_{Z|M,\mathcal{C}}(\cdot,\cdot|\mathcal{C}) - \hat{P}_{Z}(\cdot)\right\|_{\mathrm{TV}}\right] \\
&= 2\cdot\mathbf{E}\left[\left\|\mathsf{A}^{-1}\sum_{a=1}^{\mathsf{A}}\hat{P}_{Z|U}(\cdot|\mathbf{U}_{P_U\times\delta_M}(a)) - \hat{P}_{Z}(\cdot)\right\|_{\mathrm{TV}}\right] \\
&\overset{(b)}{\leq} 2\cdot\mathbf{E}\left[\left(1+2^{-\hat{\imath}(U;Z)}\right)^{-\mathsf{B}}\right] + \sqrt{\mathsf{BA}^{-1}} \\
&\leq \sup_{P_{Z|X}\in\mathcal{E}} 2\cdot\mathbf{E}_{Z|X\sim P_{Z|X}}\left[\left(1+2^{-\hat{\imath}(U;Z)}\right)^{-\mathsf{B}}\right] + \sqrt{\mathsf{BA}^{-1}} \qquad (4.5) \\
&=: \overline{\gamma}
\end{aligned}
$$

where (a) is by the convexity of total variation distance, (b) is by [117, Proposition 3] since $\left\{\mathbf{U}_{P_U\times\delta_m(a)}\right\}_{a\in[\mathsf{A}]} \overset{\text{iid}}{\sim} P_U$ for any m, and we define (4.5) to be $\overline{\gamma}$.

Let $P_e\left(P_{Y|X}, c\right)$ denote be the probability of error when the legitimate channel is $P_{Y|X}$ and the codebook is $\mathcal{C} = c$ and also let $\gamma\left(P_{Z|X}, c\right)$ denote the total variation distance γ when the wiretap channel is $P_{Z|X}$ and the codebook is $\mathcal{C} = c$. Let $P_e(P_{Y|X}, P_{Z|X}, c) = P_e(P_{Y|X}, c) + \nu \cdot \gamma(P_{Z|X}, c)$. Let $\tilde{\mathcal{D}} \subseteq \mathcal{D}$ attain the minimum in $N_{\epsilon_1}(\mathcal{D})$ and $\tilde{\mathcal{E}} \subseteq \mathcal{E}$ attain the minimum in $N_{\epsilon_2}(\mathcal{E})$.

Consider any $P_{Y|X} \in \mathcal{D}$ and any $P_{Z|X} \in \mathcal{E}$, and let $\tilde{P}_{Y|X} \in \tilde{\mathcal{D}}, \tilde{P}_{Z|X} \in \tilde{\mathcal{E}}$ satisfy

$$\sup_{x\in\mathcal{X}} \left\| P_{Y|X}(\cdot|x) - \tilde{P}_{Y|X}(\cdot|x) \right\|_{\text{TV}} \leq \epsilon_1,$$
$$\sup_{x\in\mathcal{X}} \left\| P_{Z|X}(\cdot|x) - \tilde{P}_{Z|X}(\cdot|x) \right\|_{\text{TV}} \leq \epsilon_2.$$

The total variation distance between the joint distribution of M, A, U, X, Y, Z under the channel $P_{Y|X}$ (or $P_{Z|X}$) conditional on $\mathcal{C} = c$ and the joint distribution under the channel $\tilde{P}_{Y|X}$ (or $\tilde{P}_{Z|X}$) conditional on $\mathcal{C} = c$ is also bounded by ϵ_1 (or ϵ_2). Therefore, we have $\left|P_e(P_{Y|X}, c) - P_e(\tilde{P}_{Y|X}, c)\right| \leq \epsilon_1$ and $\left|\gamma(P_{Z|X}, c) - \gamma(\tilde{P}_{Z|X}, c)\right| \leq \epsilon_2$. Hence,

$$\begin{aligned}
&P_e\left(P_{Y|X}, P_{Z|X}, c\right) \\
&\leq P_e\left(\tilde{P}_{Y|X}, c\right) + \epsilon_1 + \nu \cdot \gamma\left(\tilde{P}_{Z|X}, c\right) + \nu \cdot \epsilon_2 \\
&\leq \sum_{\tilde{P}_{Y|X}\in\tilde{\mathcal{D}}} P_e\left(\tilde{P}_{Y|X}, c\right) + \epsilon_1 + \nu \cdot \sum_{\tilde{P}_{Z|X}\in\tilde{\mathcal{E}}} \gamma\left(\tilde{P}_{Z|X}, c\right) + \nu \cdot \epsilon_2.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbf{E}_{\mathcal{C}}\Big[\sup_{P_{Y|X}\in\mathcal{D}, P_{Z|X}\in\mathcal{E}} P_e(P_{Y|X}, P_{Z|X}, \mathcal{C})\Big] \\
&\leq \mathbf{E}_{\mathcal{C}}\left[\sum_{\tilde{P}_{Y|X}\in\tilde{\mathcal{D}}} P_e\left(\tilde{P}_{Y|X}, \mathcal{C}\right) + \epsilon_1 + \nu \cdot \sum_{\tilde{P}_{Z|X}\in\tilde{\mathcal{E}}} \gamma\left(\tilde{P}_{Z|X}, \mathcal{C}\right) + \nu \cdot \epsilon_2\right] \\
&= \sum_{\tilde{P}_{Y|X}\in\tilde{\mathcal{D}}} P_e(\tilde{P}_{Y|X}) + \epsilon_1 + \nu \cdot \sum_{\tilde{P}_{Z|X}\in\tilde{\mathcal{E}}} \gamma\left(\tilde{P}_{Z|X}\right) + \nu \cdot \epsilon_2
\end{aligned}$$

$$\leq |\tilde{\mathcal{D}}| \cdot \overline{P_e} + \nu \cdot |\tilde{\mathcal{E}}| \cdot \overline{\gamma} + \epsilon_1 + \nu \cdot \epsilon_2.$$

Hence the proof is completed by the existence of a codebook c such that

$$\sup_{P_{Y|X} \in \mathcal{D}, P_{Z|X} \in \mathcal{E}} P_e(P_{Y|X}, P_{Z|X}, c)$$
$$\leq |\tilde{\mathcal{D}}| \cdot \overline{P_e} + \nu \cdot |\tilde{\mathcal{E}}| \cdot \overline{\gamma} + \epsilon_1 + \nu \cdot \epsilon_2.$$

□

Remark 4.5.2. This scheme can be viewed as a combination of the covering argument that has been discussed in Section 4.3 and the one-shot soft covering lemma [117, Proposition 3]. One can possibly provide different one-shot achievability results utilizing other existing one-shot results on single wiretap channels [87, 130, 187].

4.5.3 Recovery of the Asymptotic Results

We recover the existing asymptotic results [123] as follows. In [123], they assume all the random variables are discrete and the channels are memoryless, and $\mathcal{D} := \{P_{Y_1|X}, \ldots, P_{Y_\mathsf{J}|X}\}$ and $\mathcal{E} := \{P_{Z_1|X}, \ldots, P_{Z_\mathsf{K}|X}\}$ for some finite J, K. The setting can be understood as Figure 4.3.

By [123], for the discrete memoryless channels, the achievable secrecy rate is

$$R = \max \left[\min_j I(U;Y_j) - \max_k I(U;Z_k) \right] \tag{4.6}$$
$$= \max \min_{j,k} \left[I(U;Y_j) - I(U;Z_k) \right], \tag{4.7}$$

where the maximum is taken over all distributions $P_{U,X}$ such that the auxiliary random variable U satisfies the Markov chain $U \leftrightarrow X \leftrightarrow (Y_j, Z_k)$ for $j = 1, \ldots, \mathsf{J}$ and $k = 1, \ldots, \mathsf{K}$. This can be understood as the *worst-case* result, i.e., one considers the worst receiver in $\mathcal{D}$ and the most-powerful eavesdropper in $\mathcal{E}$.

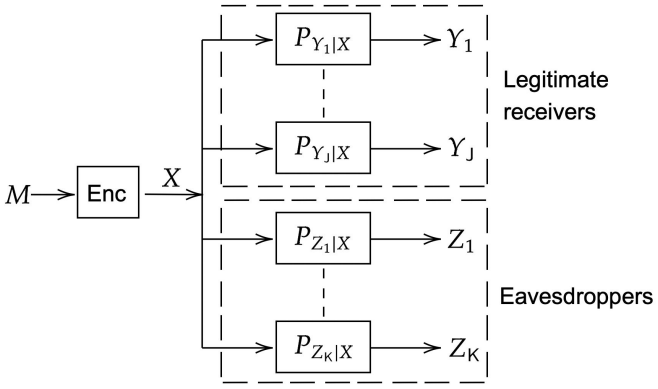

Figure 4.3: Discrete memoryless compound wiretap channel setting in [123].

To recover the above asymptotic results from our Theorem 4.5.1, fix $P_{U,X}$, take $\hat{P}_{Y|X}\hat{P}_{Z|X}$ which minimizes $I(U;Y) - I(U;Z)$, and assume $(M, A, U, X, Y, Z) \sim P_M \times P_A \times P_{U,X}\hat{P}_{Y|X}\hat{P}_{Z|X}$. Write the information density and mutual information obtained terms from this distribution as $\hat{\imath}(U;Y), \hat{\imath}(U;Z), \hat{I}(U;Y), \hat{I}(U;Z)$. Fix a coding rate $R = \hat{I}(U;Y) - \hat{I}(U;Z) - \epsilon$, we are left to show that this rate is achievable.

When we have i.i.d. sequences $(A^n, U^n, X^n, Y^n, Z^n) \sim P_A^n P_{U,X}^n P_{Y|X}^n P_{Z|X}^n$, take $\mathsf{L} = \lfloor 2^{nR} \rfloor$, $\mathsf{A} = 2^{n(I(U;Z)+\epsilon/2)}$ and $\mathsf{B} = 2^{n(I(U;Z)+\epsilon/3)}$, by the law of large numbers, the following terms in Theorem 4.5.1:

$$\begin{aligned}
\mathsf{L}\mathsf{A}2^{-\hat{\imath}(U;Y)} &\leq 2^{nR+n(I(U;Z)+\epsilon/2)-\sum_{i=1}^n \hat{\imath}(U_i;Y_i)} \to 0, \\
\left(1+2^{-\hat{\imath}(U;Z)}\right)^{-\mathsf{B}} &\overset{(a)}{\leq} 2^{\sum_{i=1}^n \hat{\imath}(U_i;Z_i)-n(I(U;Z)+\epsilon/3)} \to 0, \\
\sqrt{\mathsf{B}\mathsf{A}^{-1}} &= \sqrt{2^{n(I(U;Z)+\epsilon/3)-n(I(U;Z)+\epsilon/2)}} \to 0
\end{aligned}$$

exponentially as $n \to \infty$, where (a) used $(1+2^{-x})^{-2^y} \leq 2^{x-y}$. It is left to bound the ϵ-covering numbers $N_{\epsilon_1}(\mathcal{D}), N_{\epsilon_2}(\mathcal{E})$ in Theorem 4.5.1. Similar to Section 4.3, by constructing an ϵ-covers of them and by Proposition 4.4.1, we know $N_{\epsilon_1}(\mathcal{D}_n) \leq N_{\epsilon_1/n}(\mathcal{D}) = O((n/\epsilon)^{|\mathcal{X}|\cdot|\mathcal{Y}|})$ and similarly $N_{\epsilon_2}(\mathcal{E}_n) \leq O((n/\epsilon)^{|\mathcal{X}|\cdot|\mathcal{Z}|})$. Hence they grow much slower than the exponential decrease of above terms in Theorem 4.5.1.

Therefore we have $P_e + \nu \cdot \gamma \to 0$ as $n \to \infty$.

Chapter 5

One-Shot Channel Simulation with Differential Privacy

5.1 Overview

In this chapter, we introduce a novel "DP mechanism compressor", called *Poisson private representation*, designed to compress and *exactly* simulate *any* local randomizer while ensuring local DP, through the use of shared randomness. The Poisson private representation (PPR) can be viewed as a "meta-mechanism", in the sense that it compresses arbitrary differential privacy mechanisms.[1] This chapter is based on [132].

We elaborate on three main advantages of PPR, namely universality, exactness and communication efficiency.

1. **Universality**: Unlike dithered-quantization-based approaches which can

[1]Here "meta-mechanism" means a method that takes a privacy mechanism $\mathcal{A}$, and produces a new compressed mechanism $\mathcal{A}'$. While it is intuitively similar to a higher-order function in functional programming, we allow a meta-mechanism to look at the output distribution induced by $\mathcal{A}$, instead of only treating $\mathcal{A}$ as a black box.

only simulate additive noise mechanisms, PPR can simulate any local or central DP mechanism with discrete or continuous input and output. Moreover, PPR is *universal* in the sense that the user and the server only need to agree on the output space and a proposal distribution, and the user can simulate any DP mechanism with the same output space. The user can choose a suitable DP mechanism and privacy budget according to their communication bandwidth and privacy requirement, without divulging their choice to the server.

2. **Exactness**: Unlike previous DP mechanism compressors such as [65, 153, 168], PPR enables *exact* simulation, ensuring that the reproduced distribution perfectly matches the original one. Exact distribution recovery offers several advantages. Firstly, the compressed sample maintains the same statistical properties as the uncompressed one. If the local randomizer is unbiased (a crucial requirement for many machine learning tasks like DP-SGD), the outcome of PPR remains unbiased. In contrast, reconstruction distributions in prior simulation-based compression methods [65, 153] are often biased unless specific debiasing steps are performed (only possible for certain DP mechanisms [153]). Secondly, when the goal is to compute the mean (e.g., for private mean or frequency estimation problems) and the local noise is "summable" (e.g., Gaussian noise or other infinitely divisible distributions [77, 107]), exact distribution recovery of the local noise enables precise privacy accounting for the final *central* DP guarantee, without relying on generic privacy amplification techniques like shuffling [61, 64]. PPR can compress a central DP mechanism (e.g., the Gaussian mechanism [50]) and simultaneously achieve weaker local DP (i.e., with a larger $\varepsilon_{\mathsf{local}}$) and stronger central DP (i.e., with a smaller $\varepsilon_{\mathsf{central}}$), while maintaining exactly the same privacy-utility trade-offs as the uncompressed Gaussian mecha-

nism.

3. **Communication efficiency**: PPR compresses the output of any DP mechanism to a size close to the theoretical lower bound. For a mechanism on the data X with output Z, the compression size of PPR is $I(X;Z) + \log(I(X;Z)+1) + O(1)$, with only a logarithmic gap from the mutual information lower bound $I(X;Z)$.[2] The "$O(1)$" constant can be given explicitly in terms of a tunable parameter $\alpha > 1$ which controls the trade-off between compression size, computational time and privacy. An α close to 1 provides a better local DP guarantee, but requires a larger compression size and longer computational time.

The main technical tool we utilize for PPR is the Poisson functional representation [117, 118], which provides precise control over the reconstructed joint distribution in channel simulation problems [12, 13, 22, 40, 68, 76, 83, 118]. Channel simulation aims to achieve the minimum communication for simulating a channel (i.e., a specific conditional distribution). Typically, these methods rely on shared randomness between the user and server, and privacy is only preserved *when the shared randomness is hidden from the adversary*. This setup conflicts with local DP, where the server (which requires access to shared randomness for decoding) is considered adversarial. To ensure local DP, we introduce a randomized encoder based on the Poisson functional representation, which stochastically maps a private local message to its representation. Hence, PPR achieves order-wise trade-offs between privacy, communication, and accuracy, while preserving the original distribution of local randomizers.

[2]This is similar to channel simulation [83] and the strong functional representation lemma [118], though [83, 118] do not concern privacy.

5.2 Related Work

5.2.1 Generic compression of local DP mechanisms

In this work, we consider both central DP [51] and local DP [98, 176]. Recent research has explored methods for compressing local DP randomizers when shared randomness is involved. For instance, when $\varepsilon \leq 1$, [11] demonstrated that a single bit can simulate any local DP randomizer with a small degradation of utility, as long as the output can be computed using only a subset of the users' data. [24] proposed another generic compression technique based on rejection sampling, which compresses a ε-DP mechanism into a 10ε-DP mechanism. [65] proposed a distributed simulation approach using rejection sampling with shared randomness, while [153, 168] utilized importance sampling (or more specifically, minimum random coding [39, 86, 159]). However, all these methods only *approximate* the original local DP mechanism, unlike our scheme, which achieves an *exact* distribution recovery.

5.2.2 Distributed mean estimation under DP

Mean estimation is the canonical problems in distributed learning and analytics. They have been widely studied under privacy [8, 16, 46, 47], communication [23, 73, 163], or both constraints [30, 32, 34, 65, 80, 153]. Among them, [8] has demonstrated that the optimal unbiased mean estimation scheme under local differential privacy is privUnit [16]. Subsequently, communication-efficient mechanisms introduced by [65, 94, 153] aimed to construct communication-efficient versions of privUnit, either through distributed simulation or discretization. However, these approaches only approximate the privUnit distribution, while our proposed method ensures exact distribution recovery.

5.2.3 Distributed channel simulation

Our approach relies on the notion of channel simulation [12, 13, 22, 40, 68, 76, 83, 118]. One-shot channel simulation is a lossy compression task, which aims to find the minimum amount of communications over a noiseless channel that is in need to "simulate" some channel $P_{Z|X}$ (a specific conditional distribution). By [83, 118], the average communication cost is $I(X;Z) + O(\log(I(X;Z)))$. In [83], algorithms based on rejection sampling are proposed, and it is further generalized in [71] by introducing the greedy rejection coding. Dithered quantization [191] has also been used to simulate an additive noise channel in [3] for neural compression. As also shown in [3], the time complexity of channel simulation protocols (e.g., in [118]) is usually high, and [68, 76, 166] try to improve the runtime under certain assumptions. Moreover, channel simulation tools have also been used in neural network compression [86], image compression via variational autoencoders [69], diffusion models with perfect realism [167] and differentially private federated learning [153].

5.3 Preliminaries

We begin by reviewing the formal definitions of differential privacy (DP). We consider two models of DP data analysis. In the central model, introduced in [51], the data of the individuals is stored in a database $X \in \mathcal{X}$ by the server. The server is then trusted to perform data analysis whose output $Z = \mathcal{A}(X) \in \mathcal{Z}$ (where $\mathcal{A}$ is a randomized algorithm), which is sent to an untrusted data analyst, does not reveal too much information about any particular individual' s data. While this model requires a higher level of trust than the local model, it is possible to design significantly more accurate algorithms. We say that two databases $X, X' \in \mathcal{X}$ are neighboring if they differ in a single data point. More generally,

we can consider a symmetric neighbor relation $\mathcal{N} \subseteq \mathcal{X}^2$, and regard X, X' as neighbors if $(X, X') \in \mathcal{N}$.

On the other hand, in the local model, each individual (or client) randomizes their data before sending it to the server, meaning that individuals are not required to trust the server. A local DP mechanism [98] is a local randomizer $\mathcal{A}$ that maps the local data $X \in \mathcal{X}$ to the output $Z = \mathcal{A}(X) \in \mathcal{Z}$. Note that here X is the data at one user, unlike central-DP where X is the database with the data of all users. We now review the notion of (ε, δ)-central and local DP.

Definition 5.3.1 (Differential privacy [51, 98])**.** Given a mechanism $\mathcal{A}$ which induces the conditional distribution $P_{Z|X}$ of $Z = \mathcal{A}(X)$, we say that it satisfies (ε, δ)-DP if for any neighboring $(x, x') \in \mathcal{N}$ and $\mathcal{S} \subseteq \mathcal{Z}$, it holds that

$$\Pr(Z \in \mathcal{S} \,|\, X = x) \leq e^{\varepsilon} \Pr(Z \in \mathcal{S} \,|\, X = x') + \delta.$$

In particular, if $\mathcal{N} = \mathcal{X}^2$, we say that the mechanism satisfies (ε, δ)-local DP [98].[3]

When a mechanism satisfies $(\varepsilon, 0)$-central/local DP, we will refer to it simply as ε-central/local DP. ε-DP can be generalized to *metric privacy* by considering a metric $d_{\mathcal{X}}(x, x')$ over $\mathcal{X}$ [6, 29].

Definition 5.3.2 ($\varepsilon \cdot d_{\mathcal{X}}$-privacy [6, 29])**.** Given a mechanism $\mathcal{A}$ with conditional distribution $P_{Z|X}$, and a metric $d_{\mathcal{X}}$ over $\mathcal{X}$, we say that $\mathcal{A}$ satisfies $\varepsilon \cdot d_{\mathcal{X}}$-privacy if for any $x, x' \in \mathcal{X}$, $\mathcal{S} \subseteq \mathcal{Z}$, we have

$$\Pr(Z \in \mathcal{S} \,|\, X = x) \leq e^{\varepsilon \cdot d_{\mathcal{X}}(x, x')} \Pr(Z \in \mathcal{S} \,|\, X = x').$$

This recovers the original ε-central DP by considering $d_{\mathcal{X}}$ to be the Hamming distance among databases, and recovers the original ε-local DP by considering $d_{\mathcal{X}}$ to be the discrete metric [29].

[3] Equivalently, local DP can be viewed as a special case of central DP with dataset size $n = 1$.

The reason we use X to refer to both the database in central DP and the user's data in local DP is that our proposed method can compress both central and local DP mechanisms in exactly the same manner. In the following sections, the mechanism $\mathcal{A}$ to be compressed (often written as a conditional distribution $P_{Z|X}$) can be either a central or local DP mechanism, and the neighbor relation $\mathcal{N}$ can be any symmetric relation. The "encoder" refers to the server in central DP, or the user in local DP. The "decoder" refers to the data analyst in central DP, or the server in local DP.

5.4 Poisson Private Representation

As introduced in Chapter 2, given $(T_i)_i$, a Poisson process with rate 1 (i.e., $T_1, T_2 - T_1, T_3 - T_2, \ldots \overset{iid}{\sim} \mathrm{Exp}(1)$) that is independent of $Z_i \overset{iid}{\sim} Q$ for $i = 1, 2, \ldots$, the Poisson functional representation [117, 118] selects the point $Z := Z_K$ where

$$K = \mathrm{argmin}_i T_i \cdot \left(\frac{\mathrm{d}P}{\mathrm{d}Q}(Z_i)\right)^{-1}.$$

The calculation of K can be viewed as a search problem over a Poisson process. The "marked" Poisson process $(Z_i, T_i)_i$ supports a "query operation" provided by the Poisson functional representation, where one can input a distribution P over $\mathcal{Z}$, and obtain one sample with distribution P, i.e., the Poisson functional representation guarantees that $Z \sim P$ [118].

To simulate a DP mechanism with a conditional distribution $P_{Z|X}$ using the Poisson functional representation, we can use $(Z_i)_i$ as the shared randomness between the encoder and the decoder. [4] Upon observing X, the encoder generates the Poisson process $(T_i)_i$, computes $\tilde{T}_i$ and K using $P = P_{Z|X}$, and transmits K

[4] The original Poisson functional representation [117, 118] uses the whole $(Z_i, T_i)_i$ as the shared randomness. It is clear that $(T_i)_i$ is not needed by the decoder, and hence we can use only $(Z_i)_i$ as the shared randomness.

to the decoder. The decoder simply outputs Z_K, which follows the conditional distribution $P_{Z|X}$. The issue is that K is a function of X and the shared randomness $(Z_i, T_i)_i$, and a change of X may affect K in a deterministic manner, and hence this method cannot be directly used to protect the privacy of X.

Poisson private representation. To ensure privacy, we introduce randomness in the encoder by a generalization of the Poisson functional representation, which we call *Poisson private representation (PPR)* with parameter $\alpha \in (1, \infty]$, proposal distribution Q and the simulated mechanism $P_{Z|X}$. Both X and Z can be discrete or continuous, though as a regularity condition, we require $P_{Z|X}(\cdot|X)$ to be absolutely continuous with respect to Q almost surely. The PPR-compressed mechanism is given as:

1. We use $(Z_i)_{i=1,2,\ldots}$, $Z_i \overset{iid}{\sim} Q$ as the shared randomness between the encoder and the decoder. Practically, the encoder and the decoder can share a random seed and generate $Z_i \overset{iid}{\sim} Q$ from it using a pseudorandom number generator.[5]

2. The encoder knows $(Z_i)_i, X, P_{Z|X}$ and performs the following steps:

   (a) Generates the Poisson process $(T_i)_i$ with rate 1.

   (b) Computes $\tilde{T}_i := T_i \cdot (\frac{\mathrm{d}P}{\mathrm{d}Q}(Z_i))^{-1}$, where $P := P_{Z|X}(\cdot|X)$. Take $\tilde{T}_i = \infty$ if $\frac{\mathrm{d}P}{\mathrm{d}Q}(Z_i) = 0$.

   (c) Generates $K \in \mathbb{Z}_+$ using local randomness with

$$\Pr(K = k) = \frac{\tilde{T}_k^{-\alpha}}{\sum_{i=1}^{\infty} \tilde{T}_i^{-\alpha}}.$$

[5]We note that our analyses assume that the adversary knows both the index K and the shared randomness $(Z_i)_i$, and we prove that the mechanism is still private despite the shared randomness between the encoder and the decoder, since the privacy is provided by locally randomizing K in Step 2c.

(d) Compress K (e.g., using Elias delta coding [60]) and sends K.

3. The decoder, which knows $(Z_i)_i, K$, outputs $Z = Z_K$.

We will provide an algorithm with implementation details later.

Note that when $\alpha = \infty$, we have $K = \operatorname{argmin}_i \tilde{T}_i$, and PPR reduces to the original Poisson functional representation [117, 118]. PPR can simulate the privacy mechanism $P_{Z|X}$ precisely, as shown in the following proposition. The proof is in Appendix C.1.

Proposition 5.4.1. *The output Z of PPR follows the conditional distribution $P_{Z|X}$ exactly.*

Due to the *exactness* of PPR, it guarantees unbiasedness for tasks such as DME. If the goal is only to design a stand-alone privacy mechanism, we can focus on the privacy and utility of the mechanism without studying the output distribution. However, if the output of the mechanism is used for downstream tasks (e.g., for DME, after receiving information from clients, the server sends information about the aggregated mean to data analysts, where central DP is crucial), having an exact characterization of the conditional distribution of the output given the input allows us to obtain precise (central) privacy and utility guarantees.

Notably, PPR is *universal* in the sense that only the encoder needs to know the simulated mechanism $P_{Z|X}$. The decoder can decode the index K as long as it has access to the shared randomness $(Z_i)_i$. This allows the encoder to choose an arbitrary mechanism $P_{Z|X}$ with the same $\mathcal{Z}$, and adapt the choice of $P_{Z|X}$ to the communication and privacy constraints without explicitly informing the decoder which mechanism is chosen.

Practically, the algorithm cannot compute the whole infinite sequence $(\tilde{T}_i)_i$. We can truncate the method and only compute $\tilde{T}_i, \ldots, \tilde{T}_N$ for a large N and

select $K \in \{1, \ldots, N\}$, which incurs a small distortion in the distribution of Z.[6] While this method is practically acceptable, it might defeat the purpose of having an exact algorithm that ensures the correct conditional distribution $P_{Z|X}$. In Appendix C.2, we will present an exact algorithm for PPR that terminates in a finite amount of time, using a reparametrization that allows the encoder to know when the optimal point Z_i has already been encountered (see Algorithm 1 in Appendix C.2).

By the lower bound for channel simulation [13, 118], we must have $H(K) \geq I(X; Z)$, i.e., the compression size is at least the mutual information between the data X and the output Z. The following result shows that the compression provided by PPR is "almost optimal", i.e., close to the theoretical lower bound $I(X; Z)$. The proof is given in Appendix C.6.

Theorem 5.4.2 (Compression size of PPR)**.** *For PPR with parameter $\alpha > 1$, when the encoder is given the input x, the message K given by PPR satisfies*

$$\mathbb{E}[\log K] \leq D(P\|Q) + (\log(3.56))/\min\{(\alpha - 1)/2, 1\},$$

where $P := P_{Z|X}(\cdot|x)$. As a result, when the input $X \sim P_X$ is random, taking $Q = P_Z$, we have

$$\mathbb{E}[\log K] \leq I(X; Z) + (\log(3.56))/\min\{(\alpha - 1)/2, 1\}.$$

Note that the running time complexity (which depends on the number of samples Z_i the algorithm must examine before outputting the index K) can be

[6] To compare to the minimal random coding (MRC) [39, 86, 159] scheme in [153], which also utilizes a finite number N of samples $(Z_i)_{i=1,\ldots,N}$, while truncating the number of samples to N in both PPR and MRC results in a distortion in the distribution of Z that tends to 0 as $N \to \infty$, the difference is that $\log K$ (which is approximately the compression size) in MRC grows like $\log N$, whereas $\log K$ does not grow as $N \to \infty$ in PPR. The size N in truncated PPR merely controls the trade-off between accuracy of the distribution of Z and the running time of the algorithm.

quite high. Since $\mathbb{E}[\log K] \approx I(X;Z)$, K (and hence the running time) is at least exponential in $I(X;Z)$. See more discussions in Section 5.9.

If a prefix-free encoding of K is required, then the number of bits needed is slightly larger than $\log_2 K$. For example, if Elias delta code [60] is used, the expected compression size is $\leq \mathbb{E}[\log_2 K] + 2\log_2(\mathbb{E}[\log_2 K] + 1) + 1$ bits. If the Shannon code [155] (an almost-optimal prefix-free code) for the Zipf distribution $p(k) \propto k^{-\lambda}$ with $\lambda = 1 + 1/\mathbb{E}[\log_2 K]$ is used, the expected compression size is $\leq \mathbb{E}[\log_2 K] + \log_2(\mathbb{E}[\log_2 K] + 1) + 2$ bits (see [118]). Both codes yield an $I(X;Z) + O(\log I(X;Z))$ size, within a logarithmic gap from the lower bound $I(X;Z)$. This is similar to some other channel simulation schemes such as [22, 83, 118], though these schemes do not provide privacy

Remark 5.4.3. Note that PPR requires a variable-length code to encode the index $K \in \mathbb{Z}_+$, which is common in channel simulation [83, 118] and distributed mean estimation [163]. If we impose a fixed limit of b bits on the encoding, since Theorem 5.4.2 and Markov's inequality yields $\Pr(\log_2 K > b) \leq P_e := I(X;Z)/b + \log_2(3.56)/(b \min\{(\alpha - 1)/2, 1\})$

Note that if $P_{Z|X}$ is ε-DP, then by definition, for any $z \in \mathcal{Z}$ and $x, x_0 \in \mathcal{X}$, it holds that

$$D\left(P_{Z|X=x} \middle\| P_{Z|X=x_0}\right) = \mathbb{E}_{Z \sim P_{Z|X=x}}\left[\log\left(\frac{\mathrm{d}P_{Z|X=x}}{\mathrm{d}P_{Z|X=x_0}}(Z)\right)\right] \leq \varepsilon \log e.$$

Setting the proposal distribution $Q = P_{Z|X=x_0}$ for an arbitrary $x_0 \in \mathcal{X}$ gives the following bound.

Corollary 5.4.4 (Compression size under ε-LDP)**.** *Let $P_{Z|X}$ satisfy ε-differential privacy. Let $x_0 \in \mathcal{X}$ and $Q = P_{Z|X=x_0}$. Then for PPR with parameter $\alpha > 1$, the expected compression size is at most $\ell + \log_2(\ell + 1) + 2$ bits, where $\ell := \varepsilon \log_2 e + (\log_2(3.56))/\min\{(\alpha - 1)/2, 1\}$.*

Next, we analyze the privacy guarantee of PPR. The PPR method induces a conditional distribution $P_{(Z_i)_i,K|X}$ of the knowledge of the decoder $((Z_i)_i, K)$, given the data X. To analyze the privacy guarantee, we study whether the randomized mapping $P_{(Z_i)_i,K|X}$ from X to $((Z_i)_i, K)$ satisfies ε-DP or (ε, δ)-DP. [7] This is similar to the privacy condition in [153], and is referred as *decoder privacy* in [154], which is stronger than *database privacy* which concerns the privacy of the randomized mapping from X to the final output Z [154] (which is simply the privacy of the original mechanism $P_{Z|X}$ to be compressed since PPR simulates $P_{Z|X}$ precisely). Since the decoder knows $((Z_i)_i, K)$, more than just the final output Z, we expect that the PPR-compressed mechanism $P_{(Z_i)_i,K|X}$ to have a worse privacy guarantee than the original mechanism $P_{Z|X}$, which is the price of having a smaller communication cost. The following result shows that, if the original mechanism $P_{Z|X}$ is ε-DP, then the PPR-compressed mechanism is guaranteed to be $2\alpha\varepsilon$-DP.

Theorem 5.4.5 (ε-DP of PPR)**.** *If the mechanism $P_{Z|X}$ is ε-differentially private, then PPR $P_{(Z_i)_i,K|X}$ with parameter $\alpha > 1$ is $2\alpha\varepsilon$-differentially private.*

Similar results also apply to (ε, δ)-DP and metric DP.

Theorem 5.4.6 ((ε, δ)-DP of PPR)**.** *If the mechanism $P_{Z|X}$ is (ε, δ)-differentially private, then PPR $P_{(Z_i)_i,K|X}$ with parameter $\alpha > 1$ is $(2\alpha\varepsilon, 2\delta)$-differentially private.*

[7] Note that the encoder does not actually send $((Z_i)_i, K)$; it only sends K. The common randomness $(Z_i)_i$ is independent of the data X, and can be pre-generated using a common random seed in practice. While this seed must be communicated between the client and the server as a small overhead, the client and the server only ever need to communicate *one* seed to initialize a pseudorandom number generator, that can be used in *all* subsequent privacy mechanisms and communication tasks (to transmit high-dimensional data or use DP mechanisms for many times). The conditional distribution $P_{(Z_i)_i,K|X}$ is only relevant for privacy analysis.

Theorem 5.4.7 (Metric privacy of PPR)**.** *If the mechanism $P_{Z|X}$ satisfies $\varepsilon \cdot d_{\mathcal{X}}$-privacy, then PPR $P_{(Z_i)_i,K|X}$ with parameter $\alpha > 1$ satisfies $2\alpha\varepsilon \cdot d_{\mathcal{X}}$-privacy.*

Refer to Appendices C.3 and C.4 for the proofs. In Theorem 5.4.5 and Theorem 5.4.6, PPR imposes a multiplicative penalty 2α on the privacy parameter ε. This penalty can be made arbitrarily close to 2 by taking α close to 1, which increases the communication cost (see Theorem 5.4.2). Compared to minimal random coding which has a factor 2 penalty in the DP guarantee [86, 153], the 2α factor in PPR is slightly larger, though PPR ensures exact simulation (unlike [86, 153] which are approximate). The method in [65] does not have a penalty on ε, but the utility and compression size depends on computational hardness assumptions on the pseudorandom number generator, and there is no guarantee that the compression size is close to the optimum. In comparison, the compression and privacy guarantees of PPR are *unconditional* and does not rely on computational assumptions. In order to make the penalty of PPR close to 1, we have to consider (ε, δ)-differential privacy, and allow a small failure probability, i.e., a small increase in δ. The following result shows that PPR can compress any ε-DP mechanism into a $(\approx \varepsilon, \approx 0)$-DP mechanism as long as α is close enough to 1 (i.e., almost no inflation). More generally, PPR can compress an (ε, δ)-DP mechanism into an $(\approx \varepsilon, \approx 2\delta)$-DP mechanism for α close to 1. The proof is in Appendix C.5.

Theorem 5.4.8 (Tighter (ε, δ)-DP of PPR)**.** *If the mechanism $P_{Z|X}$ is (ε, δ)-differentially private, then PPR $P_{(Z_i)_i,K|X}$ with parameter $\alpha > 1$ is $(\alpha\varepsilon + \tilde{\varepsilon}, 2(\delta + \tilde{\delta}))$-differentially private, for every $\tilde{\varepsilon} \in (0, 1]$ and $\tilde{\delta} \in (0, 1/3]$ that satisfy $\alpha \leq e^{-4.2\tilde{\delta}\tilde{\varepsilon}^2/(-\ln \tilde{\delta})} + 1$.*

Remark 5.4.9. For the computation-privacy trade-off, in general, a larger α results in a smaller compression size (i.e., smaller K and hence shorter running time)

but larger privacy leakage, while a smaller α leads to worse compression but better privacy guarantee. Regarding the randomness requirement, although in the theory of this paper we assume "unlimited common randomness", it is of interest to study the trade-off between the amount of common randomness used and the required communication cost, similar to the study in [40]. We leave the investigation of the randomness-communication-privacy trade-off for future work.

5.5 PPR on Distributed Mean Estimation

We demonstrate the efficacy of PPR by applying it to distributed mean estimation (DME) [163]. Note that this problem is closely related to the federated learning problems [96, 105], or similar stochastic optimization problems, e.g., [139]. In the DME problem, for each iteration, the server sends a message to update the global model by a noisy mean of the local model updates. The noisy estimation is usually from some DME framework, hence a distributed differentially-private SGD (or a differentially-private federated learning) can be constructed based on a differentially-private DME framework. For such problems, as discussed in [75], if the estimate of the gradient is unbiased in each round, the convergence rates of SGD are dependent on the ℓ_2 estimation error. In short, private DME is the core sub-routine in various private and federated optimization algorithms, such as DP-SGD [1] or DP-FedAvg [139].

In such distributed settings, each local client communicates a length-limited message to the central server, and the privacy (explicit differential privacy guarantee [51]) of the data can be guaranteed by adding noise to the estimated mean at the central server before releasing it to downstream components. For example, after estimating the average model update, the central server corrupts it with the addition of Gaussian noise). This is usually called the trusted server or *central*

DP guarantee (see Section 5.3 for definitions), since the central server is trusted in privatizing the computed mean, and it is one of the most common methods in practice for federated learning and analytics. However, as we have discussed above, our scheme not only achieves the same level of central DP guarantee, but also ensures local DP guarantee.

Consider the following general distributed setting: each of n clients holds a local data point $X_i \in \mathcal{X}$, and a central server aims to estimate a function of all local data $\mu(X^n)$, subject to privacy and local communication constraints. To this end, each client i compresses X_i into a message $Z_i \in \mathcal{Z}_n$ via a local encoder, and we require that each Z_i can be encoded into a bit string with an expected length of at most b bits. Upon receiving $Z^n := (Z_1, \ldots, Z_n)$, the central server decodes it and outputs a DP estimate $\hat{\mu}$. Two DP criteria can be considered: the (ε, δ)-central DP of the randomized mapping from X^n to $\hat{\mu}$, and the (ε, δ)-local DP of the randomized mapping from X_i to Z_i for each client i.

In the distributed L_2 mean estimation problem,

$$\mathcal{X} = \mathcal{B}_d(C) := \left\{ v \in \mathbb{R}^d \,\middle|\, \|v\|_2 \le C \right\},$$

and the central server aims to estimate the sample mean $\mu(X^n) := \frac{1}{n}\sum_{i=1}^n X_i$ by minimizing the mean squared error (MSE) $\mathbb{E}[\|\mu - \hat{\mu}\|_2^2]$. It is recently proved that under ε-local DP, privUnit [8, 16] is the optimal mechanism. By simulating privUnit with PPR and applying Corollary 5.4.4 and Theorem 5.4.6, we immediately obtain the following corollary:

Corollary 5.5.1 (PPR simulating privUnit)**.** *Let P be the density defined by ε-privUnit$_2$ [16, Algorithm 1]. Let Q be the uniform density over the sphere $\mathbb{S}^{d-1}(1/m)$ where the radius $1/m$ is defined in [16, (15)]. Let $r^* := e^{\varepsilon}$. Then the outcome of PPR (see Algorithm 1) satisfies (1) $2\alpha\varepsilon$-local DP; and (2) $(\alpha\varepsilon + \tilde{\varepsilon}, 2\tilde{\delta})$-DP for any $\alpha \le e^{-4.2}\tilde{\delta}\tilde{\varepsilon}^2/\log(1/\tilde{\delta}) + 1$. In addition, the average compression size*

is at most $\ell + \log_2(\ell+1) + 2$ *bits where* $\ell := \varepsilon + (\log_2(3.56))/\min\{(\alpha-1)/2, 1\}$. *Moreover, PPR achieves the same MSE as* ε-$\mathsf{privUnit}_2$*, which is* $O\left(d/\min\left(\varepsilon, \varepsilon^2\right)\right)$.

Note that PPR can simulate arbitrary local DP mechanisms. However, we present only the result of $\mathsf{privUnit}_2$ because it achieves the optimal privacy-accuracy trade-off. Besides simulating local DP mechanisms, PPR can also compress central DP mechanisms while still preserving some (albeit weaker) local guarantees. We give a corollary of Theorems 5.4.2 and 5.4.6. The proof is in Appendix C.8.

Corollary 5.5.2 (PPR-compressed Gaussian mechanism)**.** *Let* $\varepsilon, \delta \in (0,1)$. *Consider the Gaussian mechanism* $P_{Z|X}(\cdot|x) = \mathcal{N}(x, \frac{\sigma^2}{n}\mathbb{I}_d)$, *and the proposal distribution* $Q = \mathcal{N}(0, (\frac{C^2}{d} + \frac{\sigma^2}{n})\mathbb{I}_d)$, *where* $\sigma \geq \frac{C\sqrt{2\ln(1.25/\delta)}}{\varepsilon}$. *For each client* i, *let* Z_i *be the output of PPR applied on* $P_{Z|X}(\cdot|X_i)$. *We have:*

- $\hat{\mu}(Z^n) := \frac{1}{n}\sum_i Z_i$ *yields an unbiased estimator of* $\mu(X^n) = \frac{1}{n}\sum_{i=1}^n X_i$ *satisfying* (ε, δ)*-central DP and has MSE* $\mathbb{E}[\|\mu - \hat{\mu}\|_2^2] = \sigma^2 d/n^2$.
- *As long as* $\varepsilon < 1/\sqrt{n}$, *PPR satisfies* $(2\alpha\sqrt{n}\varepsilon, 2\delta)$*-local DP.*[8]
- *The average per-client communication cost is at most* $\ell + \log_2(\ell+1) + 2$ *bits where*

$$\ell := \frac{d}{2}\log_2\left(\frac{C^2 n}{d\sigma^2} + 1\right) + \eta_\alpha \;\leq\; \frac{d}{2}\log_2\left(\frac{n\varepsilon^2}{2d\ln(1.25/\delta)} + 1\right) + \eta_\alpha,$$

  where $\eta_\alpha := (\log_2(3.56))/\min\{(\alpha-1)/2,\, 1\}$.

[8]The restricted range on $\varepsilon < 1/\sqrt{n}$ is due to the simpler privacy accountant [49]. By using the Rényi DP accountant instead, one can achieve a tighter result that applies to any n. We present the Rényi DP version of the corollary in Appendix C.7. Moreover, in the context of federated learning, n refers to the number of clients in *each round*, which is typically much smaller than the total number of clients. For example, as observed in [96], the per-round cohort size in Google's FL application typically ranges from 10^3 to 10^5, significantly smaller than the number of trainable parameters $d \in [10^6, 10^9]$ or the number of available users $N \in [10^6, 10^8]$.

A few remarks are in order. First, notice that when α is fixed, for an $O(\frac{C^2 d}{n^2\varepsilon^2}\log(1/\delta))$ MSE, the per-client communication cost is

$$O\Big(d\log\Big(\frac{n\varepsilon^2}{d\log(1/\delta)}+1\Big)+1\Big),$$

which is at least as good as the $O(n\varepsilon^2/\log(1/\delta)+1)$ bound in [34, 163], and can be better than $O(n\varepsilon^2/\log(1/\delta)+1)$ when $n \gg d$. Hence, the PPR-compressed Gaussian mechanism is order-wise optimal. Second, compared to other works that also compress the Gaussian mechanism, PPR is the only lossless compressor; schemes based on random sparsification, projection, or minimum random coding (e.g., [34, 168]) are *lossy*, i.e., they introduce additional distortion on top of the DP noise. Finally, other DP mechanism compressors tailored to local randomizers [65, 153] do not provide the same level of central DP guarantees when applied to local Gaussian noise since the reconstructed noise is no longer Gaussian. Refer to Section 5.6 for experiments.

5.6 Empirical Results on Distributed Mean Estimation

We empirically evaluate our scheme on the Distributed Mean Estimation (DME) problem (which is formally introduced in Section 5.5).

We examine the privacy-accuracy-communication trade-off, and compare it with the Coordinate Subsampled Gaussian Mechanism (CSGM) [34, Algorithm 1], an order-optimal scheme for DME under central DP. In [34], each client only communicates partial information (via sampling a subset of the coordinates of the data vector) about its samples to amplify the privacy, and the compression is mainly from subsampling. Moreover, CSGM only guarantees central DP.

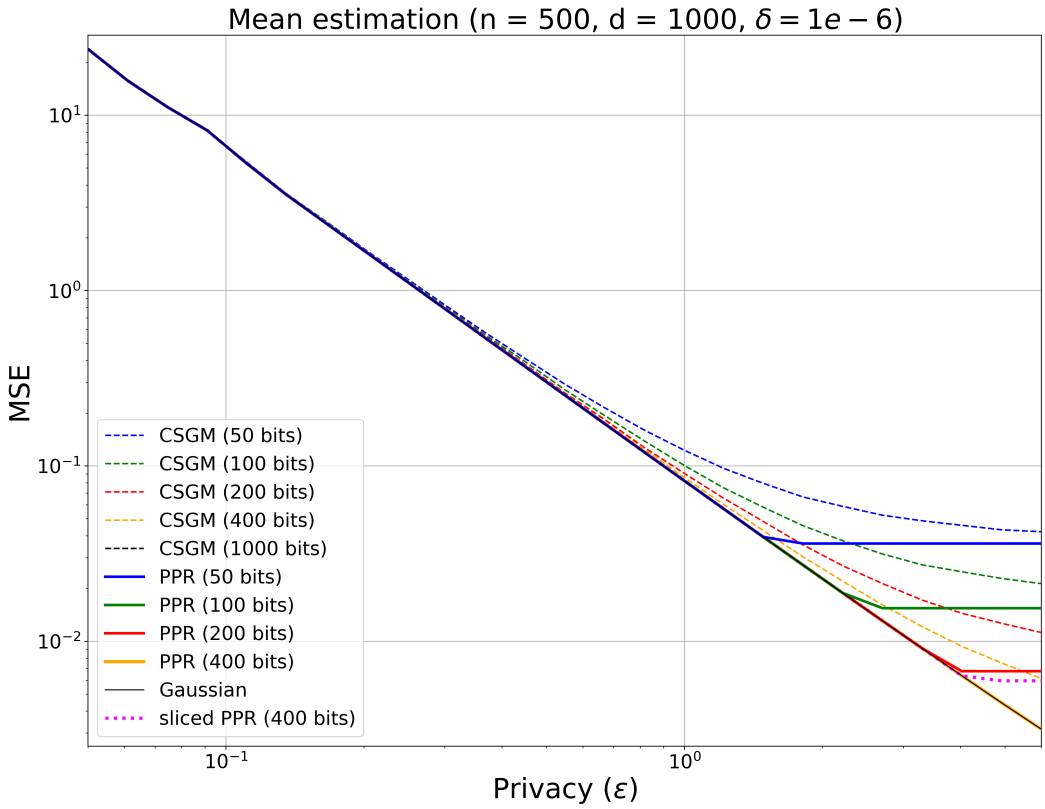

Figure 5.1: MSE of distributed mean estimation for PPR and CSGM [34] for different ε's.

5.6.1 Experiment

We use the same setup that has been used in [34]: consider $n = 500$ clients, and the dimension of local vectors is $d = 1000$, each of which is generated according to $X_i(j) \overset{\text{i.i.d.}}{\sim} (2 \cdot \text{Ber}(0.8) - 1)$, where $\text{Ber}(0.8)$ is a Bernoulli random variable with parameter $p = 0.8$. We require (ε, δ)-central DP with $\delta = 10^{-6}$ and $\varepsilon \in [0.05, 6]$ and apply the PPR with $\alpha = 2$ to simulate the Gaussian mechanism, where the privacy budgets are accounted via Rényi DP.

We compare the MSE of PPR ($\alpha = 2$, using Theorem 5.4.2) and CSGM un-

der various compression sizes in Figure 5.1 (the y-axis is in logarithmic scale).[9] Note that the MSE of the (uncompressed) Gaussian mechanism coincides with the CSGM with 1000 bits, and the PPR with only 400 bits. We see that PPR consistently achieves a smaller MSE compared to CSGM for all ε's and compression sizes considered. For $\epsilon = 1$ and we compress $d = 1000$ to 50 bits, CSGM has an MSE 0.1231 , while PPR has an MSE 0.08173, giving a 33.61% reduction. For $\epsilon = 0.5$ and we compress $d = 1000$ to 25 bits (the case of high compression and conservative privacy), CSGM has an MSE 0.3877, while PPR has an MSE 0.3011, giving a 22.33% reduction. These reductions are significant, since all considered mechanisms are asymptotically close to optimal and a large improvement compared to an (almost optimal) mechanism is unexpected. See Section C.10 for more about MSE against the compression sizes.

We also emphasize that PPR provides *both* central and local DP guarantees according to Theorem 5.4.5, 5.4.6 and 5.4.8, benefiting from the fact that PPR *exactly* compresses the privacy mechanism and hence control the distributions exactly. In contrast, CSGM only provides central DP guarantees. Another advantage of PPR under conservative privacy (small ϵ) is that the trade-off between ϵ and MSE of PPR exactly coincides with the trade-off of the Gaussian mechanism for small ϵ (see Figure 5.1), and CSGM is only close to (but strictly worse than) the Gaussian mechanism. This means that for small ϵ, PPR provides compression without any drawback in terms of ϵ-MSE trade-off compared to the Gaussian

[9]Source code: https://github.com/cheuktingli/PoissonPrivateRepr. Experiments were executed on M1 Pro Macbook, 8-core CPU ($\approx$ 3.2 GHz) with 16GB memory. For PPR under a privacy budget ε and communication budget b, we find the largest $\varepsilon' \leq \varepsilon$ such that the communication cost bound in Theorem 5.4.2 (with Shannon code [155]) for simulating the Gaussian mechanism with (ε', δ)-central DP is at most b, and use PPR to simulate this Gaussian mechanism. Thus, MSE of PPR in Figure 5.1 becomes flat for large ε, as PPR falls back to using a smaller ε' instead of ε due to the communication budget.

mechanism (which requires an infinite size communication to exactly realize).

Moreover, although directly applying PPR on the d-dimensional vectors is impractical for a large d, one can ensure an efficient $O(d)$ running time (see Section 5.9 for details) by breaking the vector with $d = 1000$ dimensions into small chunks of fixed lengths (we use $d_{\text{chunk}} = 50$ dimensions for each chunk), and apply the PPR to each chunk. We call it the *sliced PPR* in Figure 5.1. Though the sliced PPR has a small penalty on the MSE (as shown in Figure 5.1), it still outperforms the CSGM (400 bits) for the range of ε in the plot. For the sliced PPR for one $d = 1000$ vector, when $\epsilon = 0.05$, the running time is 1.3348 seconds on average.[10] For larger ϵ's, we can choose smaller d_{chunk}'s to have reasonable running time: For $\epsilon = 6$ and $d_{\text{chunk}} = 2$ we have an average running time 0.0127 seconds and with $d_{\text{chunk}} = 4$ we have an average running time 0.6343 seconds; for $\epsilon = 10$ and $d_{\text{chunk}} = 2$ we have an average running time 0.0128 seconds and with $d_{\text{chunk}} = 4$ we have an average running time 0.7301 seconds. See Section 5.9 for more experiments on the running time of the sliced PPR.

5.7 Applications to Metric Privacy

Metric privacy [6, 29] (see Definition 5.3.2) allows users to send privatized version $Z \in \mathbb{R}^d$ of their data vectors $X \in \mathbb{R}^d$ to an untrusted server, so that the server can know X approximately but not exactly. A popular mechanism is the *Laplace mechanism* [6, 29, 66, 67], where a d-dimensional Laplace noise is added to X. The conditional density function of Z given X is $f_{Z|X}(z|x) \propto e^{-\varepsilon d_{\mathcal{X}}(x,z)}$, where ε is the privacy parameter, and the metric $d_{\mathcal{X}}(x, z) = \|x - z\|_2$ is the Euclidean distance. The Laplace mechanism achieves $\varepsilon \cdot d_{\mathcal{X}}$-privacy, and has been used, for

[10] The running time is calculated by $\frac{1000}{50} \times T_{\text{chunk}}$, where each chunk's running time T_{chunk} is averaged over 1000 trials. The estimate of the mean of T_{chunk} is 0.0667, whereas the standard deviation is 0.2038.

example, to privatize high-dimensional word embedding vectors [66, 67], or for geo-indistinguishability [6] to privatize the users' locations, where the purpose is to allow users to send privatized version of their location information to an untrusted server, so that the server can approximate the locations (to provide some remote services) without knowing the exact locations.

A problem is that the real vector Z cannot be encoded into finitely many bits. To this end, [6] studies a *discrete Laplace mechanism* where each coordinate of Z is quantized to a finite number of levels, introducing additional distortion to Z. PPR provides an alternative compression method that preserves the statistical behavior of Z (e.g., unbiasedness) exactly. We give a corollary of Theorems 5.4.2 and 5.4.7. The proof is in Appendix C.9. Refer to Section 5.8 for an experiment on metric privacy.

Corollary 5.7.1 (PPR-compressed Laplace mechanism)**.** *Consider PPR applied to the Laplace mechanism $P_{Z|X}$ where $X \in \mathcal{B}_d(C) = \{x \in \mathbb{R}^d \,|\, \|x\|_2 \le C\}$, with a proposal distribution $Q = \mathcal{N}(0, (\frac{C^2}{d} + \frac{d+1}{\varepsilon^2})\mathbb{I}_d)$. It achieves an MSE $\frac{d(d+1)}{\varepsilon^2}$, a $2\alpha\epsilon \cdot d_{\mathcal{X}}$-privacy, and a compression size at most $\ell + \log_2(\ell+1) + 2$ bits, where*

$$\ell := \frac{d}{2}\log_2\left(\frac{2}{e}\left(\frac{C^2\varepsilon^2}{d} + d + 1\right)\right) - \log_2\frac{\Gamma(d+1)}{\Gamma(\frac{d}{2}+1)} + \eta_\alpha,$$

where $\eta_\alpha := (\log_2(3.56))/\min\{(\alpha-1)/2,\, 1\}$.

5.8 Empirical Results on Metric Privacy

In [6], to privatize the users' location information for some remote services provided by an untrusted server, 2-dimensional Laplace noises have been used [6] to obtain metric privacy, where the continuous planar Laplace mechanism [6] is given by the following conditional density function $f_{Z|X}(z|x) = \frac{\varepsilon^2}{2\pi}e^{-\varepsilon d_{\mathcal{X}}(x,z)}$.

We use PPR to simulate the Laplace mechanism [6, 66, 67] $f_{Z|X}(z|x) \propto e^{-\varepsilon d_{\mathcal{X}}(x,z)}$ discussed in Section 5.7. We consider $X \in \mathcal{B}_d(C)$ where $C = 10000$

and $d = 500$. A large number of dimensions d is common, for example, in privatizing word embedding vectors [66, 67]. We compare the performance of PPR-compressed Laplace mechanism (Corollary 5.7.1) with the discrete Laplace mechanism [6]. The discrete Laplace mechanism is described as follows (slightly modified from [6] to work for the d-ball $\mathcal{B}_d(C)$): 1) generate a Laplace noise Y with probability density function $f_Y(y) \propto e^{-\varepsilon\|y\|_2}$; 2) compute $\hat{Z} = X + Y$; 3) truncate $\hat{Z}$ to the closest point Z in $\mathcal{B}_d(C)$; and 4) quantize each coordinate of Z by a quantizer with step size $u > 0$. The number of bits required by the discrete Laplace mechanism is $\lceil \log_2(\mathrm{Vol}(\mathcal{B}_d(C))/u^d) \rceil$, where $\mathrm{Vol}(\mathcal{B}_d(C))/u^d$ is the number of quantization cells (hypercube of side length u) inside $\mathcal{B}_d(C)$. The parameter u is selected to fit the number of bits allowed.

Figure 5.2 shows the mean squared error of PPR-compressed Laplace mechanism ($\alpha = 2$) and the discrete Laplace mechanism for different ε's, when the number of bits is limited to 500, 1000 and 1500.[11] We can see that PPR performs better for larger ϵ or smaller MSE, whereas the discrete Laplace mechanism performs better for smaller ϵ or larger MSE. The performance of discrete Laplace mechanism for smaller ϵ is due to the truncation step which limits Z to $\mathcal{B}_d(C)$, which reduces the error at the expense of introducing distortion to the distribution of Z, and making Z a biased estimate of X. In comparison, PPR preserves the Laplace conditional distribution $f_{Z|X}$ exactly, and hence produces an unbiased Z.

[11] The MSE of PPR is computed using the closed-form formula in Corollary 5.7.1, which is tractable since Z follows the Laplace conditional distribution $f_{Z|X}$ exactly. The number of bits used by PPR is given by the bound in Corollary 5.7.1. The MSE of the discrete Laplace mechanism is estimated using 5000 trials per data point. Although we do not plot the error bars, the largest coefficient of variation of the sample mean (i.e., standard error of the mean divided by the sample mean) is only 0.00117, which would be unnoticeable even if the error bars were plotted.

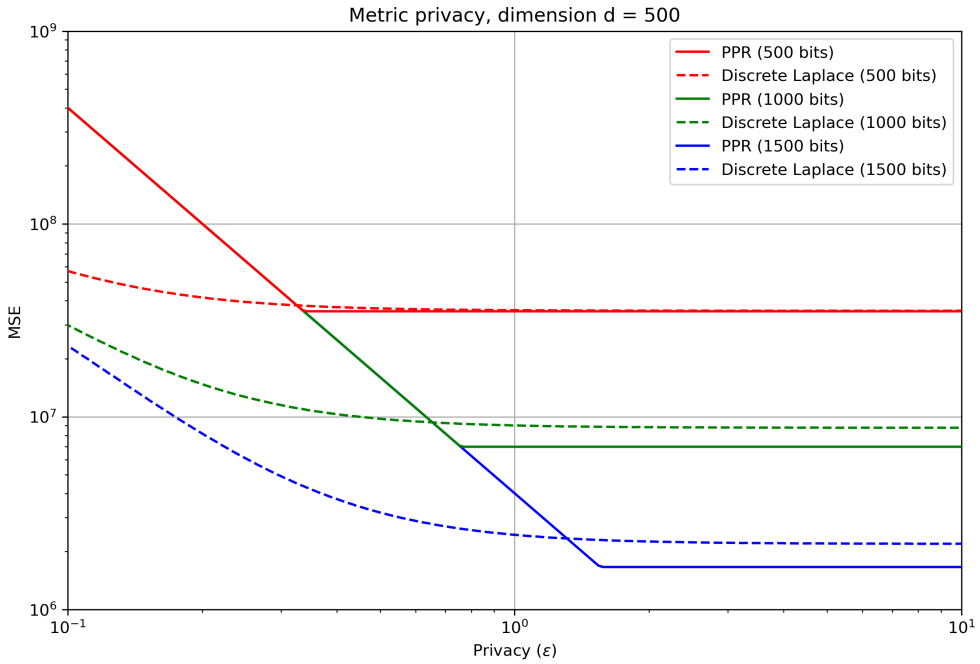

Figure 5.2: MSE of PPR-compressed Laplace mechanism and discrete Laplace mechanism [6] for different ε's.

5.9 Running Time of PPR

5.9.1 Discussions

While PPR is communication-efficient, having only a logarithmic gap from the theoretical lower bound on the compression size as shown in Theorem 5.4.2, the running time complexity can be high. However, we note that an exponential complexity is also needed in sampling methods that do not ensure privacy, such as [86, 135]. It has been proved in [3] that no polynomial time general sampling-based method exists (even without privacy constraint), if $RP \neq NP$. All existing polynomial time exact channel simulation methods can only simulate specific noisy channels.[12] Hence, a polynomial time algorithm for exactly compressing a general DP mechanism is likely nonexistent.

Nevertheless, this is not an obstacle for simulating local DP mechanisms, since the mutual information $I(X;Z)$ for a reasonable local DP mechanism must be small, or else the leakage of the data X in Z would be too large. For an ε-local DP mechanism, we have $I(X;Z) \leq \min\{\varepsilon, \varepsilon^2\}$ (in nats) [43]. Hence, the PPR algorithm can terminate quickly even if has a running time exponential in $I(X;Z)$.

Another way to ensure a polynomial running time is to divide the data into small chunks and apply the mechanism to each chunk separately. For example, to apply the Gaussian mechanism to a high-dimensional vector, we break it into several shorter vectors and apply the mechanism to each vector. Experiments in Section 5.6 show that this greatly reduces the running time while having only a small penalty on the compression size.

[12] For example, [72] and dithered-quantization-based schemes [92, 154] can only simulate additive noise mechanisms. Among these existing works, only [154] ensures local DP.

5.9.2 Empirical Results

We show empirical results on the running time of PPR on distributed mean estimation task, as discussed in Section 5.5 and Section 5.6.

5.9.3 Running Time of Sliced PPR against chunk size

As discussed in Section 5.6, we can ensure an $O(d)$ running time for the Gaussian mechanism by using the sliced PPR, where the d-dimensional vector X is divided into $\lceil d/d_{\text{chunk}} \rceil$ chunks, each with a fixed dimension d_{chunk} (possibly except the last chunk if d_{chunk} is not a factor of d). The average total running time is $\lceil d/d_{\text{chunk}} \rceil T_{\text{chunk}}$, where T_{chunk} is the average running time of PPR applied on one chunk.[13] Therefore, to study the running time of the sliced PPR, we study how T_{chunk} depend on d_{chunk}.

In Figure 5.3 we show the running time T_{chunk} of PPR applied on one chunk with dimension d_{chunk}, where d_{chunk} ranges from 40 to 110.[14] With $d = 1000$, $n = 500$, $\varepsilon = 0.05$ and $\delta = 10^{-6}$, we require a Gaussian mechanism with noise $\mathcal{N}(0, n\tilde{\sigma}^2 \mathbb{I}_{d_{\text{chunk}}})$ where $\tilde{\sigma} = 1.0917$ at each user in order to ensure (ε, δ)-central DP. We record the mean T_{chunk} and the standard error of the mean[15] of the running time of PPR applied to simulate this Gaussian mechanism (averaged over 20000 trials).

[13] Note that the chunks may be processed in parallel for improved efficiency.

[14] Experiments were executed on M1 Pro Macbook, 8-core CPU ($\approx$ 3.2 GHz) with 16GB memory.

[15] The standard error of the mean is given by $\sigma_{\text{mean}} = \sigma_{\text{time}}/\sqrt{n_{\text{trials}}}$, where σ_{time} is the standard deviation of the running time among the $n_{\text{trials}} = 20000$ trials.

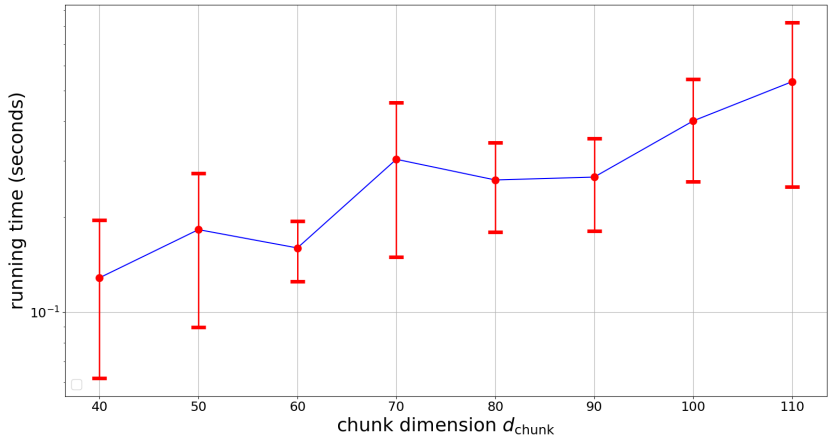

Figure 5.3: Average running time of PPR applied to a chunk of dimension d_{chunk}, with error bars indicating the interval $T_{\text{chunk}} \pm 2\sigma_{\text{mean}}$, where T_{chunk} is the sample mean of the running time, and σ_{mean} is the standard error of the mean (see Footnote 15).

5.9.4 Running Time of PPR against privacy budget ϵ

We plot the average running time (over 20000 trials for each data point) against the values of $\epsilon \in [0.06, 10]$, with d_{chunk} always chosen to be 4. The average running time is denoted as T_{chunk}, and the standard error of the mean is given by $\sigma_{\text{mean}} = \sigma_{\text{time}}/\sqrt{n_{\text{trials}}}$, where σ_{time} is the standard deviation of the running time among the $\sigma_{\text{time}} = 20000$ trials.

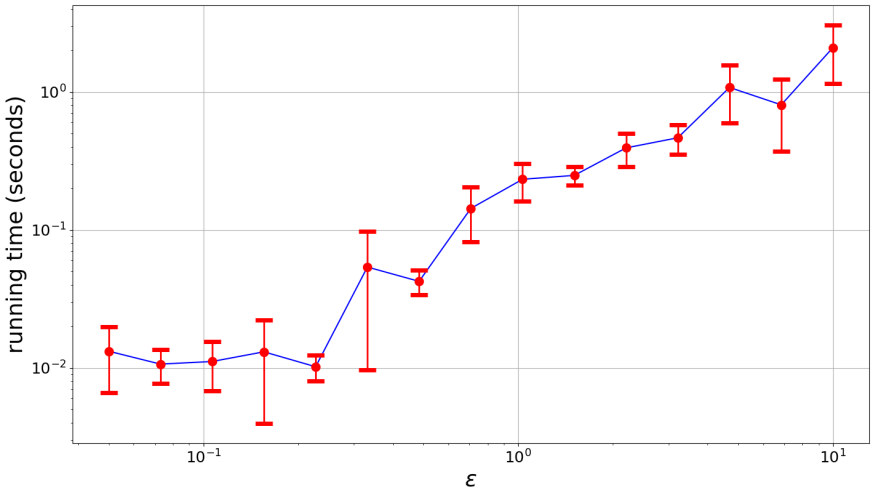

Figure 5.4: Average running time (over 20000 trials), $d_{\text{chunk}} = 4$ and $\varepsilon \in [0.06, 10]$, with error bars indicating the interval $T_{\text{chunk}} \pm 2\sigma_{\text{mean}}$, where T_{chunk} is the sample mean of the running time, and σ_{mean} is the standard error of the mean.

Chapter 6

Discussion and Conclusion

In conventional network information theory, the information-theoretic limits are investigated *asymptotically* in the large blocklength regime, based on the law of large numbers. However, this assumption is impractical, as packets have bounded lengths. Over the past decade, *finite blocklength* [106, 149, 165] and *one-shot* [88, 117, 159, 172, 177, 189] information theory have been widely studied.

In this thesis, we studied one-shot information theory, code constructions, and applications to differential privacy. In the one-shot setting, we assumed the channel or source was used only *once* (i.e., it need not be memoryless or ergodic), and the blocklength was 1. Therefore, the blocklength did not approach infinity, and hence existing tools (e.g., *typical sets* [59]) and the law of large numbers could not be used. We provided several novel constructions for one-shot settings and proved achievability results, based on the Poisson functional representation [118]. These results were expected to recover existing (first-order and second-order) asymptotic bounds when applied to memoryless channels or sources.

In Chapter 3, we provide a unified one-shot coding framework over a general noisy network, applicable to any combination of source coding, channel coding, and coding for computing problems. Compared to the original Pois-

son matching lemma [117], our scheme works for arbitrary discrete acyclic noisy networks and can be viewed as a one-shot counterpart of the asymptotic acyclic discrete memoryless network studied by Lee and Chung [110]. Various one-shot results [117] and asymptotic results [59] have been recovered, and novel one-shot results have been derived, including source coding, channel coding, primitive relay channel [55, 56, 59, 101, 142], Gelfand-Pinsker [74, 89], relay-with-unlimited-look-ahead [57, 58], Wyner-Ziv [180, 182], coding for computing [184], multiple access channels [4, 5, 125], broadcast channels [138], and cascade multiterminal source coding [42].

In Chapter 4, we derived novel one-shot achievability results for two classical secrecy problems in information theory: the information hiding problem [144] and the compound wiretap channels [123]. Our bounds are based on the Poisson matching lemma together with other techniques, and are applicable to both continuous and discrete cases. Our one-shot achievability results apply to any distribution of the source data, and any class of the channels (not necessarily memoryless or ergodic), and can readily recover the existing asymptotic results on both problems when apply to discrete memoryless channels subject to potential distortion constraints, thus providing alternative proofs that are potentially simpler. Moreover, we generalized the information hiding setting [144] and extended its reconstruction objective. For both the generalized information hiding problem and the compound wiretap channels, unlike most existing studies, we do not assume that the decoder knows the channel state in the one-shot setting.

In Chapter 5, we proposed a novel scheme for compressing differential privacy mechanisms, called Poisson Private Representation (PPR), to reduce the communication cost of differential privacy mechanisms. Unlike previous schemes, which were either constrained to special classes of DP mechanisms or introduced additional distortions to the output, our scheme could compress and exactly sim-

ulate arbitrary mechanisms while locally protecting differential privacy, thereby preserving the joint distribution of the data and the output of the original local randomizer. PPR achieved a compression size within a logarithmic gap from the theoretical lower bound, and a new order-wise trade-off between communication, accuracy, and central and local differential privacy for distributed mean estimation was derived. One possible issue was the running time of PPR, which we discussed and tackled by using sliced PPR, a method that divided long data vectors into small chunks. Moreover, we presented experimental results on distributed mean estimation to show that, while providing local differential privacy, PPR and sliced PPR consistently offered a better trade-off between communication, accuracy, and central differential privacy compared to the coordinate-subsampled Gaussian mechanism [34].

6.1 Future Directions

While we have discussed the analysis and applications of one-shot codes based on the Poisson functional representation [118], several avenues for future research remain, which we discuss below.

Based on the unified one-shot coding scheme over arbitrary noisy networks described in Chapter 3, automated theorem provers can potentially be designed. For example, an existing automated theorem proving tool [112] provides an algorithm for deriving asymptotic inner and outer bounds for general acyclic discrete memoryless networks [110]. Considering our coding scheme in Chapter 3 as a one-shot counterpart of [110], designing an automated theorem prover for one-shot inner bounds appears promising. Developing different schemes for one-shot outer bounds could be another potential direction for future work. Moreover, we should note that for the sake of universality and simplicity, our one-shot cod-

ing scheme sacrifices some performance. For example, for broadcast channels, we showed in Section 3.7 that two corner points can be recovered, but not the entire region of Marton's inner bound, which could be derived using the original Poisson matching lemma with a more complicated analysis [117]. A possible extension is to generalize our scheme to improve its capability without significantly compromising its universality and simplicity.

In Chapter 5, we discussed the application of channel simulation schemes (in particular, a variant of the Poisson functional representation [118]) for compressing differential privacy mechanisms. As detailed in Section 5.9, the running time of general channel simulation schemes poses challenges for practical implementation, although certain strategies, such as the sliced PPR method shown in Section 5.9, can improve efficiency. Although it has been proven in [3] that no general polynomial-time exact sampling-based method exists (even without privacy constraints) unless $RP = NP$, fast exact channel simulation schemes can still be designed for specific noisy channels. For the channel simulation task itself (without privacy constraints), recent works have proposed using linear error correction codes, such as polar codes [7], for fast channel simulation; techniques from [68, 70] are also useful when $P_{Z|X}$ is unimodal. On the one hand, developing fast channel simulation schemes for specific channels is an interesting research direction on its own (for example, the additive Gaussian noise channel, which plays an important role in neural compression [86] and differential privacy [85, 185]; special schemes based on vector quantization have been considered in [104]). On the other hand, applying these schemes to compress differential privacy mechanisms is another promising future direction. Moreover, as discussed in Chapter 5, our scheme suffers a two-fold increase in the privacy budget, similar to the results based on importance sampling [153] (which is an approximate but not exact scheme). It is worth investigating whether this two-fold increase is a fundamental

limit for privacy mechanism simulation.

Appendix A

Proofs for Chapter 3

A.1 Proof of Theorem 3.4.1 and Theorem 3.4.2

We first generate N independent exponential processes $\mathbf{U}_i$ for $i \in [N]$ according to Section 3.2, which serve as the random codebooks. Each node i will perform two steps: the *decoding step* and the *encoding step*.

We describe the decoding step at node i. The node observes Y_i and wants to decode $\overline{U}_i' = (U_{a_{i,j}})_{j\in[d_i']}$, while utilizing $(U_{a_{i,j}})_{j\in[d_i'+1..d_i]}$ by non-unique decoding. For the sake of notational simplicity, we omit the subscript i and write $d = d_i$, $d' = d_i'$, $a_k = a_{i,k}$, $\overline{U}_{\mathcal{S}} = \overline{U}_{i,\mathcal{S}} = (U_{a_{i,j}})_{j\in\mathcal{S}}$, $\overline{\mathbf{U}}_k := \mathbf{U}_{a_{i,k}}$. For each $j = 1, \ldots, d'$, the node will perform soft decoding via the exponential process refinement (see Section 3.2) on $\overline{U}_d$, and then on $\overline{U}_{d-1}$, and so on up to $\overline{U}_{j+1}$, and then use all the distributions obtained to decode $\overline{U}_j$ uniquely using the Poisson functional representation. For example, when $d = 3$, $d' = 2$, the decoding process will be: $\overline{U}_3$ (soft), $\overline{U}_2$ (soft), $\overline{U}_1$ (unique), $\overline{U}_3$ (soft), $\overline{U}_2$ (unique). The choice of the sequence $a_{i,k}$ controls the decoding ordering of the random variables. The goal is to obtain the decoded variables $\hat{\overline{U}}_1, \ldots, \hat{\overline{U}}_{d'}$ that equal $\overline{U}_1, \ldots, \overline{U}_{d'}$ with high probability.

More precisely, for $j = 1, \ldots, d'$, the node computes the decoded variable $\hat{\overline{U}}_j \in \overline{\mathcal{U}}_j$ by first computing the joint distributions $Q^{(j)}_{\overline{U}_{[k..d]}}$ over $\overline{\mathcal{U}}_k \times \ldots \times \overline{\mathcal{U}}_d$ for $k = d, d-1, \ldots, j+1$ recursively using the exponential process refinement as

$$Q^{(j)}_{\overline{U}_{[k..d]}} := \left(Q^{(j)}_{\overline{U}_{[k+1..d]}} P_{\overline{U}_k|\overline{U}_{[k+1..d]},\overline{U}_{[j-1]},Y_i}(\cdot \,|\, \cdot, \hat{\overline{U}}_{[j-1]}, Y_i)\right)^{\overline{\mathbf{U}}_k},$$

i.e., first compute the semidirect product between $Q^{(j)}_{\overline{U}_{[k+1..d]}}$ and the conditional distribution $P_{\overline{U}_k|\overline{U}_{[k+1..d]},\overline{U}_{[j-1]},Y_i}(\cdot \,|\, \cdot, \hat{\overline{U}}_{[j-1]}, Y_i)$ (computed using the ideal joint distribution of X^N, Y^N, U^N) to obtain a distribution over $\overline{\mathcal{U}}_k \times \ldots \times \overline{\mathcal{U}}_d$, and then refine it by $\overline{\mathbf{U}}_k$ using Definition 3.2.2. For the base case, we assume $Q^{(j)}_{\overline{U}_{[d+1..d]}}$ is the degenerate distribution. After we have computed $Q^{(j)}_{\overline{U}_{[j+1..d]}}$, we can obtain $\hat{\overline{U}}_j$ using the Poisson functional representation (3.1) as $\hat{\overline{U}}_j = (\overline{\mathbf{U}}_j)_{\tilde{Q}^{(j)}_{\overline{U}_j}}$, where $\tilde{Q}^{(j)}_{\overline{U}_j}$ is the $\overline{U}_j$-marginal of

$$Q^{(j)}_{\overline{U}_{[j+1..d]}} P_{\overline{U}_j|\overline{U}_{[j+1..d]},\overline{U}_{[j-1]},Y_i}(\cdot \,|\, \cdot, \hat{\overline{U}}_{[j-1]}, Y_i). \tag{A.1}$$

The node repeats this process for $j = 1, \ldots, d'$ to obtain $\hat{\overline{U}}'_i = (\hat{\overline{U}}_1, \ldots, \hat{\overline{U}}_{d'})$.

We then describe the encoding step at node i. It uses the Poisson functional representation (see Section 3.2) to obtain

$$U_i = (\mathbf{U}_i)_{P_{U_i|Y_i,\overline{U}'_i}(\cdot|Y_i,\hat{\overline{U}}'_i)}. \tag{A.2}$$

Finally, it generates X_i from the conditional distribution $P_{X_i|Y_i,U_i,\overline{U}'_i}(\cdot|Y_i, U_i, \hat{\overline{U}}'_i)$.

For the error analysis, we create a fictitious "ideal network" (with N "ideal nodes") that is almost identical to the actual network. The only difference is that the ideal node i uses the true $\overline{U}'_i$ (supplied by a genie) instead of the decoded $\hat{\overline{U}}'_i$ for the encoding step. The random variables induced by the ideal network will have the same distribution as the ideal distribution of X^N, Y^N, U^N in Theorem 3.4.1. Hence, we assume X^N, Y^N, U^N are induced by the ideal network. We couple the channels in the ideal network and the channels in the actual network, such

that $Y_i = \tilde{Y}_i$ if $(X^{i-1}, Y^{i-1}) = (\tilde{X}^{i-1}, \tilde{Y}^{i-1})$ (i.e., the "channel noises" in the two networks are the same). If none of the actual nodes makes an error (i.e., $\hat{\overline{U}}'_i = \overline{U}'_i$ for all i), the actual network would coincide with the ideal network, and $(\tilde{X}^N, \tilde{Y}^N) = (X^N, Y^N)$. We consider the error probability conditional on $A := (X^N, Y^N, U^N)$:

$$F := \mathbf{P}\big(\exists\, i:\ \hat{\overline{U}}'_i = \overline{U}'_i \,\big|\, A\big).$$

Note that F is a random variable and is a function of $A = (X^N, Y^N, U^N)$. We have

$$\begin{aligned} F &= \mathbf{P}\big(\exists\, i \in [N], j \in [d'_i] : \hat{\overline{U}}_{i,j} \neq \overline{U}_{i,j} \,\big|\, A\big) \\ &= \sum_{i=1}^{N} \sum_{j=1}^{d'_i} \mathbf{P}\Big(\hat{\overline{U}}'_{[i-1]} = \overline{U}'_{[i-1]},\, \hat{\overline{U}}_{i,[j-1]} = \overline{U}_{i,[j-1]},\, \hat{\overline{U}}_{i,j} \neq \overline{U}_{i,j} \,\Big|\, A\Big). \end{aligned}$$

For the term inside the summation (which is the probability that the first error we make is at $\overline{U}_{i,j}$), by (A.1), (A.2) and the Poisson matching lemma [117] (we again omit the subscripts i as in the description of the decoding step, e.g., we write $\overline{U}_j = \overline{U}_{i,j} = U_{a_j} = U_{a_{i,j}}$; we also simply write $P(\overline{U}_j|Y_{a_j}, \overline{U}'_{a_j}) = P_{\overline{U}_j|Y_{a_j},\overline{U}'_{a_j}}(\overline{U}_j|Y_{a_j}, \overline{U}'_{a_j})$), we have

$$\begin{aligned} &\mathbf{P}\big(\hat{\overline{U}}'_{[i-1]} = \overline{U}'_{[i-1]},\, \hat{\overline{U}}_{i,[j-1]} = \overline{U}_{i,[j-1]},\, \hat{\overline{U}}_{i,j} \neq \overline{U}_{i,j} \,\big|\, A\big) \\ &\le \mathbf{E}\left[\frac{P(\overline{U}_j|Y_{a_j}, \overline{U}'_{a_j})}{Q^{(j)}(\overline{U}_{[j+1..d]})P(\overline{U}_j \,|\, \overline{U}_{[j+1..d]},\, \overline{U}_{[j-1]}, Y_i)} \,\middle|\, A\right] \\ &\overset{(a)}{\le} \mathbf{E}\left[\frac{P(\overline{U}_j|Y_{a_j}, \overline{U}'_{a_j})}{P(\overline{U}_j \,|\, \overline{U}_{[j+1..d]},\, \overline{U}_{[j-1]}, Y_i)} \right. \\ &\qquad \left. \cdot\, \mathbf{E}\left[\frac{1}{Q^{(j)}(\overline{U}_{[j+1..d]})} \,\middle|\, \overline{U}_{[d]}, Y_i, Y_{a_j}, \overline{U}'_{a_j}, \overline{\mathbf{U}}_{[j+1..d]}\right] \,\middle|\, A\right] \\ &\overset{(b)}{\le} \mathbf{E}\left[\frac{P(\overline{U}_j|Y_{a_j}, \overline{U}'_{a_j})}{P(\overline{U}_j \,|\, \overline{U}_{[j+1..d]},\, \overline{U}_{[j-1]}, Y_i)} (\ln|\overline{\mathcal{U}}_{j+1}| + 1) \right. \\ &\qquad \left. \cdot\, \frac{1}{Q^{(j)}(\overline{U}_{[j+2..d]})} \left(\frac{P(\overline{U}_{j+1}|Y_{a_{j+1}}, \overline{U}'_{a_{j+1}})}{P(\overline{U}_{j+1}|\overline{U}_{[j+2..d]}, \overline{U}_{[j-1]}, Y_i)} + 1\right) \,\middle|\, A\right] \end{aligned}$$

$$
\begin{aligned}
&\overset{(c)}{\leq} \mathbf{E}\Bigg[\frac{P(\overline{U}_j|Y_{a_j},\overline{U}'_{a_j})}{P(\overline{U}_j \,|\, \overline{U}_{[j+1..d]},\, \overline{U}_{[j-1]}, Y_i)} \\
&\cdot \prod_{k=j+1}^{d'} (\ln|\overline{\mathcal{U}}_k|+1)\left(\frac{P(\overline{U}_k|Y_{a_k},\overline{U}'_{a_k})}{P(\overline{U}_k|\overline{U}_{[k+1..d]},\overline{U}_{[j-1]},Y_i)}+1\right)\Bigg|\, A\Bigg] \\
&= B_{i,j},
\end{aligned}
$$

where (a) is by Jensen's inequality, (b) is due to Lemma 3.2.3, (c) is by applying the same steps as (a) and (b) $d'-j-1$ times, and $\beta_{i,j}$ is given in (3.4). The proof of Theorem 3.4.1 is completed by noting that $\delta_{\mathrm{TV}}(P_{X^N,Y^N}, P_{\tilde{X}^N,\tilde{Y}^N}) \leq \mathbf{P}((X^N,Y^N)\neq(\tilde{X}^N,\tilde{Y}^N)) \leq \mathbf{E}[F] = \mathbf{E}[\min\{F,1\}]$.

We now prove Theorem 3.4.2. Recall that the scheme we have constructed requires the public randomness W, which we have to fix in order to construct a deterministic coding scheme for Theorem 3.4.2. We have

$$
\begin{aligned}
&\mathbf{E}\big[\mathbf{P}\big((\tilde{X}^N,\tilde{Y}^N)\in\mathcal{E}\,\big|\,W\big)\big] \\
&= \mathbf{P}((\tilde{X}^N,\tilde{Y}^N)\in\mathcal{E}) \\
&\leq \mathbf{P}\big((X^N,Y^N)\in\mathcal{E} \text{ or } (X^N,Y^N)\neq(\tilde{X}^N,\tilde{Y}^N)\big) \\
&= \mathbf{E}\big[\mathbf{P}\big((X^N,Y^N)\in\mathcal{E} \text{ or } (X^N,Y^N)\neq(\tilde{X}^N,\tilde{Y}^N)\,\big|\,A\big)\big] \\
&\leq \mathbf{E}\big[\min\big\{\mathbf{1}\{(X^N,Y^N)\in\mathcal{E}\}+\mathbf{P}\big((X^N,Y^N)\neq(\tilde{X}^N,\tilde{Y}^N)\,\big|\,A\big),\,1\big\}\big] \\
&\leq \mathbf{E}\big[\min\big\{\mathbf{1}\{(X^N,Y^N)\in\mathcal{E}\}+F,\,1\big\}\big].
\end{aligned}
$$

Therefore, there exists a value w such that $\mathbf{P}((\tilde{X}^N,\tilde{Y}^N)\in\mathcal{E}\,|\,W=w)$ satisfies the upper bound. Fixing the value of W to w gives a deterministic coding scheme.

Appendix B

Proofs for Chapter 4

B.1 Proof of Proposition 4.4.1

Proof. Write $d(A, \tilde{A}) := \sup_{x \in \mathcal{X}} \|A_{Y|X}(\cdot|x) - \tilde{A}_{Y|X}(\cdot|x)\|_{\mathrm{TV}}$. We use the standard method to bound the covering number, where we start with $\tilde{\mathcal{A}} = \emptyset$, and add $A \in \mathcal{A}$ not currently covered by $\tilde{\mathcal{A}}$ (i.e., $\min_{\tilde{A} \in \tilde{\mathcal{A}}} d(A, \tilde{A}) > \epsilon$) to $\tilde{\mathcal{A}}$ one by one until all of $\mathcal{A}$ is covered. Note that every two different $\tilde{A}, \tilde{A}' \in \tilde{\mathcal{A}}$ produced this way must satisfy $d(\tilde{A}, \tilde{A}') > \epsilon$, and hence the $(\epsilon/2)$-balls $\{A : d(A, \tilde{A}) \leq \epsilon/2\}$ must be disjoint for $\tilde{A} \in \tilde{\mathcal{A}}$.

We now treat $A_{Y|X}$ as a transition probability matrix $A \in \mathbb{R}^{|\mathcal{Y}| \times |\mathcal{X}|}$. We have

$$\begin{aligned} d(A, \tilde{A}) &= \frac{1}{2}\|A - \tilde{A}\|_1 \\ &= \frac{1}{2} \max_x \sum_y |A_{y,x} - \tilde{A}_{y,x}|. \end{aligned}$$

The volume of the ball $\{A \in \mathbb{R}^{|\mathcal{Y}| \times |\mathcal{X}|} : d(A, \tilde{A}) \leq \epsilon/2\}$ (i.e., its Lebesgue measure in the space $\mathbb{R}^{|\mathcal{Y}| \cdot |\mathcal{X}|}$) is $((2\epsilon)^{|\mathcal{Y}|}/(|\mathcal{Y}|!))^{|\mathcal{X}|}$, and all these balls are subsets of $\{A \in \mathbb{R}^{|\mathcal{Y}| \times |\mathcal{X}|} : \min_{x,y} A_{y,x} \geq -\epsilon, \max_x \sum_y A_{y,x} \leq 1 + \epsilon\}$, which has a volume $((1 +$

$(|\mathcal{Y}|+1)\epsilon)^{|\mathcal{Y}|}/(|\mathcal{Y}|!))^{|\mathcal{X}|}$. Hence, the size of $\tilde{\mathcal{A}}$ is upper-bounded by

$$\frac{\Big(\big(1+(|\mathcal{Y}|+1)\epsilon\big)^{|\mathcal{Y}|}/(|\mathcal{Y}|!)\Big)^{|\mathcal{X}|}}{\big((2\epsilon)^{|\mathcal{Y}|}/(|\mathcal{Y}|!)\big)^{|\mathcal{X}|}} = \Big(\frac{1}{2\epsilon}+\frac{|\mathcal{Y}|+1}{2}\Big)^{|\mathcal{X}|\cdot|\mathcal{Y}|}.$$

□

Appendix C

Proofs for Chapter 5

C.1 Proof of Proposition 5.4.1

Write $(X_i)_i \sim \mathrm{PP}(\mu)$ if the points $(X_i)_i$ (as a multiset, ignoring the ordering) form a Poisson point process with intensity measure μ. Similarly, for $f : [0,\infty)^n \to [0,\infty)$, we write $\mathrm{PP}(f)$ for the Poisson point process with intensity function f (i.e., the intensity measure has a Radon-Nikodym derivative f against the Lebesgue measure).

Let $(T_i)_i \sim \mathrm{PP}(1)$ be a Poisson process with rate 1, independent of $Z_1, Z_2, \ldots \overset{iid}{\sim} Q$. By the marking theorem [109], $(Z_i, T_i)_i \sim \mathrm{PP}(Q \times \lambda_{[0,\infty)})$, where $Q \times \lambda_{[0,\infty)}$ is the product measure between Q and the Lebesgues measure over $[0,\infty)$. Let $P = P_{Z|X}(\cdot|x)$, and $\tilde{T}_i = T_i \cdot (\frac{\mathrm{d}P}{\mathrm{d}Q}(Z_i))^{-1}$. By the mapping theorem [109] (also see [117, 118]), $(Z_i, \tilde{T}_i)_i \sim \mathrm{PP}(P \times \lambda_{[0,\infty)})$. Note that the points $(Z_i, \tilde{T}_i)_i$ may not be sorted in ascending order of $\tilde{T}_i$. Therefore, we will sort them as follows. Let j_1 be the j such that $\tilde{T}_j$ is the smallest, j_2 be the j other than j_1 such that $\tilde{T}_j$ is the smallest, and so on. Break ties arbitrarily. Then $(\tilde{T}_{j_i})_i$ is an ascending sequence, and we still have $(Z_{j_i}, \tilde{T}_{j_i})_i \sim \mathrm{PP}(P \times \lambda_{[0,\infty)})$ since we are merely re-arranging the points. Comparing $(Z_{j_i}, \tilde{T}_{j_i})_i \sim \mathrm{PP}(P \times \lambda_{[0,\infty)})$ with the definition

of $(Z_i, T_i)_i \sim \mathrm{PP}(Q \times \lambda_{[0,\infty)})$, we can see that $(\tilde{T}_{j_i})_i \sim \mathrm{PP}(1)$ is independent of $Z_{j_1}, Z_{j_2}, \ldots \overset{iid}{\sim} P$.

Recall that in PPR, we generate $K \in \mathbb{Z}_+$ with

$$\Pr(K = k) = \frac{\tilde{T}_k^{-\alpha}}{\sum_{i=1}^{\infty} \tilde{T}_i^{-\alpha}},$$

and the final output is Z_K. Rearranging the points according to $(j_i)_i$, the distribution of the final output remains the same if we instead generate $K' \in \mathbb{Z}_+$ with

$$\Pr(K' = k) = \frac{\tilde{T}_{j_k}^{-\alpha}}{\sum_{i=1}^{\infty} \tilde{T}_{j_i}^{-\alpha}},$$

and the final output is $Z_{j_{K'}}$. Since $(\tilde{T}_{j_i})_i \sim \mathrm{PP}(1)$ is independent of $Z_{j_i} \overset{iid}{\sim} P$, we know that K' is independent of $(Z_{j_i})_i$, and hence $Z_{j_{K'}} \sim P$ follows the desired distribution.

C.2 Reparametrization and Detailed Algorithm of PPR

We now discuss the implementation of the Poisson private representation in Section 5.4. Practically, the algorithm cannot compute the whole infinite sequence $(\tilde{T}_i)_i$. We can truncate the method and only compute $\tilde{T}_i, \ldots, \tilde{T}_N$ for a large N and select $K \in \{1, \ldots, N\}$, which incurs a small distortion in the distribution of Z.[1] While this method is practically acceptable, it might defeat the purpose of

[1] To compare to the minimal random coding (MRC) [39, 86, 159] scheme in [153], which also utilizes a finite number N of samples $(Z_i)_{i=1,\ldots,N}$, while truncating the number of samples to N in both PPR and MRC results in a distortion in the distribution of Z that tends to 0 as $N \to \infty$, the difference is that $\log K$ (which is approximately the compression size) in MRC grows like $\log N$, whereas $\log K$ does not grow as $N \to \infty$ in PPR. The size N in truncated PPR merely controls the tradeoff between accuracy of the distribution of Z and the running time of the algorithm.

having an exact algorithm that ensures the correct conditional distribution $P_{Z|X}$. We now present an exact algorithm for PPR that terminates in a finite amount of time using a reparametrization.

In the proof of Theorem C.6.1, we showed that, letting $(T_i)_i \sim \text{PP}(1)$, $Z_1, Z_2, \ldots \overset{iid}{\sim} Q$, $R_i := (\mathrm{d}P/\mathrm{d}Q)(Z_i)$, $V_1, V_2, \ldots \overset{iid}{\sim} \text{Exp}(1)$, PPR can be equivalently expressed as

$$K = \operatorname*{argmin}_k T_k^\alpha R_k^{-\alpha} V_k.$$

The problem of finding K is that there is no stopping criteria for the argmin. For example, if we scan the points $(T_i, R_i, V_i)_i$ in increasing order of T_i, it is always possible that there is a future point with V_i so small that it makes $T_i^\alpha R_i^{-\alpha} V_i$ smaller than the current minimum. If we scan the points in increasing order of V_i instead, it is likewise possible that there is a future point with a very small T_i. We can scan the points in increasing order of $U_i := T_i^\alpha V_i$, but we would not know the indices of the points in the original process where $T_1 \leq T_2 \leq \cdots$ is in increasing order, which is necessary to find out the Z_i corresponding to each point (recall that in PPR, the point with the smallest T_i corresponds to Z_1, the second smallest T_i corresponds to Z_2, etc.).

Therefore, we will scan the points in increasing order of $B_i := T_i^\alpha \min\{V_i, 1\}$ instead. By the mapping theorem [109], $(T_i^\alpha)_i \sim \text{PP}(\alpha^{-1} t^{1/\alpha - 1})$. By the marking theorem [109],

$$(T_i^\alpha, V_i)_i \sim \text{PP}(\alpha^{-1} t^{1/\alpha - 1} e^{-v}).$$

By the mapping theorem,

$$(T_i^\alpha, T_i^\alpha V_i)_i \sim \text{PP}(\alpha^{-1} t^{1/\alpha - 2} e^{-vt^{-1}}).$$

Since $B_i = \min\{T_i^\alpha, T_i^\alpha V_i\}$, again by the mapping theorem,

$$(B_i)_i \sim \text{PP}\left(\int_b^\infty \alpha^{-1} b^{1/\alpha - 2} e^{-vb^{-1}} \mathrm{d}v + \int_b^\infty \alpha^{-1} t^{1/\alpha - 2} e^{-bt^{-1}} \mathrm{d}t \right)$$

$$= \mathrm{PP}\left(\alpha^{-1}b^{1/\alpha-1}e^{-1} + \alpha^{-1}b^{1/\alpha-1}\gamma(1-\alpha^{-1},1)\right)$$
$$= \mathrm{PP}\left(\alpha^{-1}\left(e^{-1}+\gamma_1\right)b^{1/\alpha-1}\right),$$

where $\gamma_1 := \gamma(1-\alpha^{-1},1)$ and $\gamma(\beta,x) = \int_0^x e^{-\tau}\tau^{\beta-1}\mathrm{d}\tau$ is the lower incomplete gamma function. Comparing the distribution of $(B_i)_i$ and $(T_i^\alpha)_i$, we can generate $(B_i)_i$ by first generating $(U_i)_i \sim \mathrm{PP}(1)$, and then taking $B_i = (U_i\alpha/(e^{-1}+\gamma_1))^\alpha$. The conditional distribution of (T_i, V_i) given $B_i = b$ is described as follows:

- With probability $e^{-1}/(e^{-1}+\gamma_1)$, we have $T_i^\alpha = b$ and $T_i^\alpha V_i \sim b(\mathrm{Exp}(1)+1)$, and hence $T_i = b^{1/\alpha}$ and $V_i \sim \mathrm{Exp}(1)+1$.

- With probability $\gamma_1/(e^{-1}+\gamma_1)$, we have $T_i^\alpha V_i = b$ and

$$T_i^\alpha \sim \frac{\alpha^{-1}t^{1/\alpha-2}e^{-bt^{-1}}}{\alpha^{-1}\gamma(1-\alpha^{-1},1)b^{1/\alpha-1}}.$$

  Hence, for $0 < \tau \le 1$,

$$\Pr(V_i \le \tau) = \Pr(T_i^\alpha \ge b/\tau) = \frac{\gamma(1-\alpha^{-1},\tau)}{\gamma(1-\alpha^{-1},1)},$$

  and V_i follows the truncated gamma distribution with shape $1-\alpha^{-1}$ and scale 1, truncated within the interval $[0,1]$. We then have $T_i = (b/V_i)^{1/\alpha}$.

The algorithm is given in Algorithm 1. The encoder and decoder require a shared random seed s. One way to generate s is to have the encoder and decoder maintain two synchronized pseudorandom number generators (PRNGs) that are always at the same state, and invoke the PRNGs to generate s, guaranteeing that the s at the encoder is the same as the s at the decoder. The encoder maintains a collection of points (T_i, V_i, Θ_i), stored in a heap to allow fast query and removal of the point with the smallest T_i. The value $\Theta_i \in \{0,1\}$ indicates whether it is possible that the point (T_i, V_i) attains the minimum of $T_k^\alpha R_k^{-\alpha} V_k$. The encoding algorithm repeats until there is no possible points left in the heap, and it is impossible for

any future point to be better than the current minimum of $T_k^\alpha R_k^{-\alpha} V_k$. The encoding time complexity is $O(\sup_z(\mathrm{d}P/\mathrm{d}Q)(z)\log(\sup_z(\mathrm{d}P/\mathrm{d}Q)(z)))$, which is close to other sampling-based channel simulation schemes [72, 83].[2] The decoding algorithm simply outputs the k-th sample generated using the random seed s, which can be performed in $O(1)$ time.[3]

The PPR is implemented by Algorithm 1. We write $x \leftarrow \mathrm{Exp}_{\mathscr{G}}(1)$ to mean that we generate an exponential random variate x with rate 1 using the pseudorandom number generator $\mathscr{G}$. Write $x \leftarrow \mathrm{Exp}_{\mathrm{local}}(1)$ to mean that x is generated using a local pseudorandom number generator (not $\mathscr{G}$).

Algorithm 1: Poisson private representation

Procedure PPRENCODE(α, Q, r, r^*, s) :

Input: parameter $\alpha > 1$, distribution Q, density $r(z) := (\mathrm{d}P/\mathrm{d}Q)(z)$, bound $r^* \geq \sup_z r(z)$, random seed s

Output: index $k \in \mathbb{Z}_{>0}$

Initialize PRNG $\mathscr{G}$ using the seed s

$u \leftarrow 0$, $w^* \leftarrow \infty$, $k \leftarrow 0$, $k^* \leftarrow 0$, $n \leftarrow 0$

$\gamma_1 \leftarrow \gamma(1-\alpha^{-1}, 1) = \int_0^1 e^{-\tau}\tau^{-\alpha^{-1}}\mathrm{d}\tau$

$h \leftarrow \emptyset$ (empty heap)

while true **do**

  $u \leftarrow u + \mathrm{Exp}_{\mathrm{local}}(1)$ ▷ *Generated using local randomness (not $\mathscr{G}$)*

  $b \leftarrow (u\alpha/(e^{-1}+\gamma_1))^\alpha$

[2] It was shown in [72] that greedy rejection sampling [83] runs in $O(\sup_z(\mathrm{d}P/\mathrm{d}Q)(z))$ time. The PPR algorithm has an additional log term due to the use of heap.

[3] A counter-based PRNG [150] allows us to directly jump to the state after k uses of the PRNG, without the need of generating all k samples, greatly improving the decoding efficiency. This technique is applicable to greedy rejection sampling [83] and the original Poisson functional representation [117, 118] as well.

if $n = 0$ and $b(r^*)^{-\alpha} \geq w^*$ **then** ▷ *No possible points left and future points impossible*

return k^*

end if

if $\text{Unif}_{\text{local}}(0,1) < e^{-1}/(e^{-1}+\gamma_1)$ **then** ▷ *Run with prob.* $e^{-1}/(e^{-1}+\gamma_1)$

$t \leftarrow b^{1/\alpha}$, $v \leftarrow \text{Exp}_{\text{local}}(1) + 1$

else

repeat

$v \leftarrow \text{Gamma}_{\text{local}}(1-\alpha^{-1}, 1)$ ▷ *Gamma distribution*

until $v \leq 1$

$t \leftarrow (b/v)^{1/\alpha}$

end if

$\theta \leftarrow \mathbf{1}\{(t/r^*)^\alpha v \leq w^*\}$ ▷ *Is it possible for this point to be optimal*

Push (t, v, θ) to h

$n \leftarrow n + \theta$ ▷ *Number of possible points in heap*

while $h \neq \emptyset$ and $\min_{(t',v',\theta') \in h} t' \leq b^{1/\alpha}$ **do** ▷ *Assign* Z_i*'s to points in heap with small* T_i

$(t, v, \theta) \leftarrow \arg\min_{(t',v',\theta') \in h} t'$, and pop (t, v, θ) from h

$n \leftarrow n - \theta$

$k \leftarrow k + 1$

Generate $z \sim Q$ using $\mathscr{G}$

$w \leftarrow (t/r(z))^\alpha v$

if $w < w^*$ **then**

$w^* \leftarrow w$

$k^* \leftarrow k$

end if

end while

33: **end while**

Procedure PPRDecode(Q, k, s) :

Input: Q, index $k \in \mathbb{Z}_{>0}$, seed s

Output: sample z

1: Initialize PRNG $\mathscr{G}$ using the seed s
2: **for** $i = 1, 2, \ldots, k$ **do**
3: Generate $z \sim Q$ using $\mathscr{G}$ ▷ *See footnote 3*
4: **end for**
5: **return** z

Algorithm 1: Poisson private representation

C.3 Proofs of Theorem 5.4.5 and Theorem 5.4.7

First prove Theorem 5.4.5. Consider a ε-DP mechanism $P_{Z|X}$. Consider neighbors x_1, x_2, and let $P_j := P_{Z|X}(\cdot|x_j)$, $\tilde{T}_{j,i} := T_i/(\frac{\mathrm{d}P_j}{\mathrm{d}Q}(Z_i))$, and K_j be the output of PPR applied on P_j, for $j = 1, 2$. Since $P_{Z|X}$ is ε-DP,

$$e^{-\varepsilon}\frac{\mathrm{d}P_2}{\mathrm{d}Q}(z) \le \frac{\mathrm{d}P_1}{\mathrm{d}Q}(z) \le e^{\varepsilon}\frac{\mathrm{d}P_2}{\mathrm{d}Q}(z) \tag{C.1}$$

for Q-almost every z,[4] and hence $e^{-\varepsilon}\tilde{T}_{2,i} \le \tilde{T}_{1,i} \le e^{\varepsilon}\tilde{T}_{2,i}$. For $k \in \mathbb{Z}_+$, we have, almost surely,

$$\begin{aligned}\Pr(K_1 = k \,|\, (Z_i, T_i)_i) &= \frac{\tilde{T}_{1,k}^{-\alpha}}{\sum_{i=1}^{\infty} \tilde{T}_{1,i}^{-\alpha}} \\ &\le \frac{e^{\alpha\varepsilon}\tilde{T}_{2,k}^{-\alpha}}{\sum_{i=1}^{\infty} e^{-\alpha\varepsilon}\tilde{T}_{2,i}^{-\alpha}} \\ &= e^{2\alpha\varepsilon} \Pr(K_2 = k \,|\, (Z_i, T_i)_i).\end{aligned}$$

For any measurable $\mathcal{S} \subseteq \mathcal{Z}^\infty \times \mathbb{Z}_{>0}$,

$$\begin{aligned}&\Pr\left(((Z_i)_i, K_1) \in \mathcal{S}\right) \\ &= \mathbb{E}\left[\Pr\left(((Z_i)_i, K_1) \in \mathcal{S} \,\middle|\, (Z_i, T_i)_i\right)\right] \\ &= \mathbb{E}\left[\sum_{k:\,((Z_i)_i,k)\in\mathcal{S}} \Pr\left(K_1 = k \,\middle|\, (Z_i, T_i)_i\right)\right] \\ &\le e^{2\alpha\varepsilon} \cdot \mathbb{E}\left[\sum_{k:\,((Z_i)_i,k)\in\mathcal{S}} \Pr\left(K_2 = k \,\middle|\, (Z_i, T_i)_i\right)\right] \\ &= e^{2\alpha\varepsilon} \Pr\left(((Z_i)_i, K_2) \in \mathcal{S}\right). \qquad \text{(C.2)}\end{aligned}$$

Hence, $P_{(Z_i)_i,K|X}$ is $2\alpha\varepsilon$-DP.

For Theorem 5.4.7, consider a $\varepsilon \cdot d_{\mathcal{X}}$-private mechanism $P_{Z|X}$, and consider $x_1, x_2 \in \mathcal{X}$. We have

$$e^{-\varepsilon \cdot d_{\mathcal{X}}(x_1,x_2)} \frac{\mathrm{d}P_2}{\mathrm{d}Q}(z) \le \frac{\mathrm{d}P_1}{\mathrm{d}Q}(z) \le e^{\varepsilon \cdot d_{\mathcal{X}}(x_1,x_2)} \frac{\mathrm{d}P_2}{\mathrm{d}Q}(z) \qquad \text{(C.3)}$$

[4] ε-DP only implies that (C.1) holds for P_1-almost every z (or equivalently P_2-almost every z since P_1, P_2 are absolutely continuous with respect to each other). We now show that (C.1) holds for Q-almost every z. Apply Lebesgue's decomposition theorem to find measures $\tilde{Q}, \hat{Q}$ such that $Q = \tilde{Q} + \hat{Q}$, $\tilde{Q} \ll P_1$ and $\hat{Q} \perp P_1$. There exists $\mathcal{Z}' \subseteq \mathcal{Z}$ such that $P_1(\mathcal{Z}') = 1$ and $\hat{Q}(\mathcal{Z}') = 0$. Since $P_1 \ll Q$, we have $P_1 \ll \tilde{Q}$. We have (C.1) for $\tilde{Q}$-almost every z. Also, we have (C.1) for $\hat{Q}$-almost every z since $z \notin \mathcal{Z}'$ gives $\frac{\mathrm{d}P_1}{\mathrm{d}Q}(z) = \frac{\mathrm{d}P_1}{\mathrm{d}\hat{Q}}(z) = 0$ for $\hat{Q}$-almost every z, and also $\frac{\mathrm{d}P_2}{\mathrm{d}Q}(z) = 0$ for $\hat{Q}$-almost every z since $P_2 \ll P_1$.

for Q-almost every z. By exactly the same arguments as in the proof of Theorem 5.4.5, $\Pr\left(((Z_i)_i, K_1) \in \mathcal{S}\right) \leq e^{2\alpha\varepsilon \cdot d_{\mathcal{X}}(x_1,x_2)} \Pr\left(((Z_i)_i, K_2) \in \mathcal{S}\right)$, and hence $P_{(Z_i)_i,K|X}$ is $2\alpha\varepsilon \cdot d_{\mathcal{X}}$-private.

C.4 Proof of Theorem 5.4.6

Consider a (ε, δ)-DP mechanism $P_{Z|X}$. Consider neighbors x_1, x_2, and let $P_j := P_{Z|X}(\cdot|x_j)$, and K_j be the output of PPR applied on P_j, for $j = 1, 2$. By the definition of (ε, δ)-differential privacy, we have

$$\int \max\left\{\rho_1(z) - e^{\varepsilon}\rho_2(z),\, 0\right\} Q(\mathrm{d}z) \leq \delta, \tag{C.4}$$

$$\int \max\left\{\rho_2(z) - e^{\varepsilon}\rho_1(z),\, 0\right\} Q(\mathrm{d}z) \leq \delta. \tag{C.5}$$

Let

$$\overline{\rho}(z) := \min\left\{\max\left\{\rho_1(z),\, e^{-\varepsilon}\rho_2(z)\right\},\, e^{\varepsilon}\rho_2(z)\right\}.$$

Note that $e^{-\varepsilon}\rho_2(z) \leq \overline{\rho}(z) \leq e^{\varepsilon}\rho_2(z)$. We then consider two cases:
Case 1: $\int \overline{\rho}(z)Q(\mathrm{d}z) \leq 1$. Let $\rho_3(z)$ be such that $\int \rho_3(z)Q(\mathrm{d}z) = 1$ and

$$\overline{\rho}(z) \leq \rho_3(z) \leq e^{\varepsilon}\rho_2(z).$$

We can always find such ρ_3 by taking an appropriate convex combination of the lower bound above (which integrates to ≤ 1) and the upper obund above (which integrates to ≥ 1). We then have

$$e^{-\varepsilon}\rho_2(z) \leq \rho_3(z) \leq e^{\varepsilon}\rho_2(z). \tag{C.6}$$

If $\rho_1(z) - e^{\varepsilon}\rho_2(z) \leq 0$, then $\rho_1(z) - \rho_3(z) \leq \rho_1(z) - \overline{\rho}(z) \leq 0$. If $\rho_1(z) - e^{\varepsilon}\rho_2(z) > 0$, then $\rho_3(z) = \overline{\rho}(z) = e^{\varepsilon}\rho_2(z)$. Either way, we have $\max\left\{\rho_1(z) - \rho_3(z),\, 0\right\} = \max\left\{\rho_1(z) - e^{\varepsilon}\rho_2(z),\, 0\right\}$. By (C.4), we have

$$\int \max\left\{\rho_1(z) - \rho_3(z),\, 0\right\} Q(\mathrm{d}z) \leq \delta.$$

Let $P_3 = \rho_3 Q$ be the probability measure with $\mathrm{d}P_3/\mathrm{d}Q = \rho_3$. Then the total variation distance $d_{\mathrm{TV}}(P_1, P_3)$ between P_1 and P_3 is at most δ, and by (C.6),

$$e^{-\varepsilon}\frac{\mathrm{d}P_2}{\mathrm{d}Q}(z) \le \frac{\mathrm{d}P_3}{\mathrm{d}Q}(z) \le e^{\varepsilon}\frac{\mathrm{d}P_2}{\mathrm{d}Q}(z). \tag{C.7}$$

Case 2: $\int \overline{\rho}(z)Q(\mathrm{d}z) > 1$. Let $\rho_3(z)$ be such that $\int \rho_3(z)Q(\mathrm{d}z) = 1$ and

$$e^{-\varepsilon}\rho_2(z) \le \rho_3(z) \le \overline{\rho}(z).$$

We can always find such ρ_3 by taking an appropriate convex combination of the lower bound above (which integrates to ≤ 1) and the upper obund above (which integrates to > 1). We again have $e^{-\varepsilon}\rho_2(z) \le \rho_3(z) \le e^{\varepsilon}\rho_2(z)$. If $e^{-\varepsilon}\rho_2(z) - \rho_1(z) \le 0$, then $\rho_3(z) - \rho_1(z) \le \overline{\rho}(z) - \rho_1(z) \le 0$. If $e^{-\varepsilon}\rho_2(z) - \rho_1(z) > 0$, then $\rho_3(z) = \overline{\rho}(z) = e^{-\varepsilon}\rho_2(z)$. Either way, we have $\max\{\rho_3(z) - \rho_1(z), 0\} = \max\{e^{-\varepsilon}\rho_2(z) - \rho_1(z), 0\}$. By (C.5), we have

$$\int \max\{\rho_3(z) - \rho_1(z), 0\}\, Q(\mathrm{d}z) \le e^{-\varepsilon}\delta \le \delta.$$

Let $P_3 = \rho_3 Q$ be the probability measure with $\mathrm{d}P_3/\mathrm{d}Q = \rho_3$. Again, we have $d_{\mathrm{TV}}(P_1, P_3) \le \delta$ and (C.7). Therefore, regardless of whether Case 1 or Case 2 holds, we can construct P_3 satisfying $d_{\mathrm{TV}}(P_1, P_3) \le \delta$ and (C.7). Let K_3 be the output of PPR applied on P_3.

In the proof of Theorem C.6.1, we see that PPR has the following equivalent formulation. Let $(T_i)_i \sim \mathrm{PP}(1)$ be a Poisson process with rate 1, independent of $Z_1, Z_2, \ldots \overset{iid}{\sim} Q$. Let $R_i := (\mathrm{d}P/\mathrm{d}Q)(Z_i)$, and let its probability measure be P_R. Let $V_1, V_2, \ldots \overset{iid}{\sim} \mathrm{Exp}(1)$. PPR can be equivalently expressed as

$$K = \operatorname*{argmin}_k T_k^{\alpha} R_k^{-\alpha} V_k = \operatorname*{argmin}_k \frac{T_k V_k^{1/\alpha}}{R_k}.$$

Note that $(T_i V_i^{1/\alpha})_i \sim \mathrm{PP}(\int_0^{\infty} v^{-1/\alpha} e^{-v}\mathrm{d}v) = \mathrm{PP}(\Gamma(1-\alpha^{-1}))$ is a uniform Poisson process. Therefore PPR is the same as the Poisson functional representation [117,

118] applied on $(T_i V_i^{1/\alpha})_i$. By the grand coupling property of Poisson functional representation [116, 117] (see [116, Theorem 3]), if we apply the Poisson functional representation on P_1 and P_3 to get K_1 and K_3 respectively, then

$$\Pr(K_1 \neq K_3) \leq 2d_{\mathrm{TV}}(P_1, P_3) \leq 2\delta.$$

Therefore, for any measurable $\mathcal{S} \subseteq \mathcal{Z}^\infty \times \mathbb{Z}_{>0}$,

$$\begin{aligned}\Pr\left(((Z_i)_i, K_1) \in \mathcal{S}\right) &\leq \Pr\left(((Z_i)_i, K_3) \in \mathcal{S}\right) + 2\delta \\ &\leq e^{2\alpha\varepsilon} \Pr\left(((Z_i)_i, K_2) \in \mathcal{S}\right) + 2\delta,\end{aligned}$$

where the last inequality is by applying (C.2) on P_3, P_2 instead of P_1, P_2. Hence, $P_{(Z_i)_i, K|X}$ is $(2\alpha\varepsilon, 2\delta)$-DP.

C.5 Proof of Theorem 5.4.8

We present the proof of (ε, δ)-DP of PPR (i.e., Theorem 5.4.8).

Proof. We assume

$$\alpha - 1 \leq \frac{\beta\tilde{\delta}\tilde{\varepsilon}^2}{-\ln\tilde{\delta}}, \tag{C.8}$$

where $\beta := e^{-4.2}$. Using the Laplace functional of the Poisson process $(\tilde{T}_i)_i$ [109, Theorem 3.9], for $w > 0$,

$$\begin{aligned}\mathbb{E}\left[\exp\left(-w\sum_i \tilde{T}_i^{-\alpha}\right)\right] &= \exp\left(-\int_0^\infty (1 - \exp(-wt^{-\alpha}))\mathrm{d}t\right) \\ &= \exp\left(-w^{1/\alpha}\Gamma(1-\alpha^{-1})\right).\end{aligned} \tag{C.9}$$

We first bound the left tail of $\sum_i \tilde{T}_i^{-\alpha}$. By Chernoff bound, for $d \geq 0$,

$$\Pr\left(\sum_i \tilde{T}_i^{-\alpha} \leq d\right)$$

$$
\begin{aligned}
&\leq \inf_{w>0} e^{wd}\mathbb{E}\left[\exp\left(-w\sum_i \tilde{T}_i^{-\alpha}\right)\right] \\
&= \inf_{w>0} \exp\left(wd - w^{1/\alpha}\Gamma(1-\alpha^{-1})\right) \\
&\leq \exp\left(\left(\frac{\Gamma(1-\alpha^{-1})}{\alpha d}\right)^{\frac{\alpha}{\alpha-1}} d - \left(\frac{\Gamma(1-\alpha^{-1})}{\alpha d}\right)^{\frac{1}{\alpha-1}} \Gamma(1-\alpha^{-1})\right) \\
&= \exp\left(\left(\Gamma(1-\alpha^{-1})\right)^{\frac{\alpha}{\alpha-1}} d^{-\frac{1}{\alpha-1}} \left(\alpha^{-\frac{\alpha}{\alpha-1}} - \alpha^{-\frac{1}{\alpha-1}}\right)\right) \\
&= \exp\left(-\left(\frac{\alpha d}{(\Gamma(1-\alpha^{-1}))^{\alpha}}\right)^{-\frac{1}{\alpha-1}} \left(1-\alpha^{-1}\right)\right) \\
&= \exp\left(-\left(\frac{\alpha d(1-\alpha^{-1})^{\alpha}}{(\Gamma(2-\alpha^{-1}))^{\alpha}}\right)^{-\frac{1}{\alpha-1}} \left(1-\alpha^{-1}\right)\right) \\
&= \exp\left(-\left(\frac{(\alpha-1)d}{(\Gamma(2-\alpha^{-1}))^{\alpha}}\right)^{-\frac{1}{\alpha-1}}\right).
\end{aligned}
$$

Therefore, to guarantee $\Pr(\sum_i \tilde{T}_i^{-\alpha} \leq d) \leq \tilde{\delta}/3$, we require

$$
d \leq \frac{\Gamma(2-\alpha^{-1})^{\alpha}\left(-\ln(\tilde{\delta}/3)\right)^{-(\alpha-1)}}{\alpha-1},
$$

where

$$
\begin{aligned}
&\Gamma(2-\alpha^{-1})^{\alpha}\left(-\ln(\tilde{\delta}/3)\right)^{-(\alpha-1)} \\
&\geq \left(\exp\left(-\gamma(\alpha-1)\right)\right)^{\alpha}\left(-\ln(\tilde{\delta}^2)\right)^{-(\alpha-1)} \\
&\geq \exp\left(-\gamma\alpha\frac{\beta\tilde{\delta}\tilde{\varepsilon}^2}{-\ln\tilde{\delta}}\right)\left(-2\ln\tilde{\delta}\right)^{-\frac{\beta\tilde{\delta}\tilde{\varepsilon}^2}{-\ln\tilde{\delta}}} \\
&\geq \exp\left(-2\gamma\frac{\beta\tilde{\delta}\tilde{\varepsilon}^2}{-\ln\tilde{\delta}} - 2e^{-1}\beta\tilde{\delta}\tilde{\varepsilon}^2\right) \\
&\geq \exp\left(-\left(\frac{2\gamma}{3\ln 2} + \frac{2}{3e}\right)\beta\tilde{\varepsilon}^2\right) \\
&\geq \exp\left(-0.81\cdot\beta\tilde{\varepsilon}^2\right) \\
&\geq e^{-\tilde{\varepsilon}/2},
\end{aligned}
$$

since $1 < \alpha \leq 2$, $0 < \tilde{\delta} \leq 1/3$, $\beta = e^{-4.2}$ and $0 < \tilde{\varepsilon} \leq 1$, where γ is the Euler-Mascheroni constant. Hence, we have

$$\Pr\left(\sum_i \tilde{T}_i^{-\alpha} \leq \frac{e^{-\tilde{\varepsilon}/2}}{\alpha - 1}\right) \leq \frac{\tilde{\delta}}{3}. \tag{C.10}$$

We then bound the right tail of $\sum_i \tilde{T}_i^{-\alpha}$. Unfortunately, the Laplace functional (C.9) does not work since the integral diverges for small t. Therefore, we have to bound t away from 0. Note that $\min_i \tilde{T}_i \sim \text{Exp}(1)$, and hence

$$\Pr(\min_i \tilde{T}_i \leq \tilde{\delta}/3) \leq \tilde{\delta}/3. \tag{C.11}$$

Write $\tau = \tilde{\delta}/3$. Using the Laplace functional again, for $w > 0$,

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(w \sum_{i:\tilde{T}_i > \tau} \tilde{T}_i^{-\alpha}\right)\right] \\
&= \exp\left(-\int_\tau^\infty (1 - \exp(wt^{-\alpha}))\mathrm{d}t\right) \\
&= \exp\left(\int_\tau^\infty (\exp(wt^{-\alpha}) - 1)\mathrm{d}t\right) \\
&\leq \exp\left(\int_\tau^\infty (\exp(w\tau^{-\alpha}) - 1)\frac{t^{-\alpha}}{\tau^{-\alpha}}\mathrm{d}t\right) \\
&= \exp\left(\frac{\exp(w\tau^{-\alpha}) - 1}{\tau^{-\alpha}} \cdot \frac{\tau^{-(\alpha-1)}}{\alpha - 1}\right) \\
&= \exp\left(\frac{\tau(\exp(w\tau^{-\alpha}) - 1)}{\alpha - 1}\right).
\end{aligned}$$

Therefore, by Chernoff bound, for $d \geq 0$,

$$\begin{aligned}
&\Pr\Big(\sum_{i:\tilde{T}_i > \tau} \tilde{T}_i^{-\alpha} \geq d\Big) \\
&\leq \inf_{w>0} \exp\left(-wd + \frac{\tau(\exp(w\tau^{-\alpha}) - 1)}{\alpha - 1}\right) \\
&\leq \exp\left(-d\tau^\alpha \ln(d(\alpha - 1)\tau^{\alpha-1}) + \frac{\tau(\exp(\ln(d(\alpha - 1)\tau^{\alpha-1})) - 1)}{\alpha - 1}\right)
\end{aligned}$$

$$
\begin{aligned}
&= \exp\left(-d\tau^{\alpha}\ln(d(\alpha-1)\tau^{\alpha-1}) + \tau\frac{d(\alpha-1)\tau^{\alpha-1}-1}{\alpha-1}\right)\\
&= \exp\left(-\frac{c\tau}{\alpha-1}\ln c + \tau\frac{c-1}{\alpha-1}\right)\\
&= \exp\left(-\frac{\tau}{\alpha-1}\left(c\ln c - c + 1\right)\right)\\
&\le \exp\left(-\frac{\tau(2\ln 2-1)(c-1)^2}{\alpha-1}\right), \qquad \text{(C.12)}
\end{aligned}
$$

where

$$c := d(\alpha-1)\tau^{\alpha-1},$$

and the last inequality holds whenever $c \in [1,2]$ since in this range,

$$c\ln c - c + 1 \ge (2\ln 2-1)(c-1)^2.$$

Substituting

$$d = \frac{e^{\tilde{\varepsilon}/2}}{\alpha-1},$$

we have $c = e^{\tilde{\varepsilon}/2}\tau^{\alpha-1}$. By (C.12), to guarantee $\Pr(\sum_{i:\tilde{T}_i>\tau}\tilde{T}_i^{-\alpha} \ge d) \le \tilde{\delta}/3 = \tau$, we require

$$
\begin{aligned}
\frac{\tau(2\ln 2-1)(e^{\tilde{\varepsilon}/2}\tau^{\alpha-1}-1)^2}{\alpha-1} &\ge -\ln\tau,\\
e^{\tilde{\varepsilon}/2}\tau^{\alpha-1} &\ge \sqrt{\frac{(\alpha-1)(-\ln\tau)}{\tau(2\ln 2-1)}} + 1. \qquad \text{(C.13)}
\end{aligned}
$$

Substituting (C.8), we have

$$
\begin{aligned}
e^{\tilde{\varepsilon}/2}\tau^{\alpha-1} &\ge e^{\tilde{\varepsilon}/2}\tau^{\frac{\beta\tilde{\delta}\tilde{\varepsilon}^2}{-\ln\tilde{\delta}}}\\
&= \exp\left(\frac{\tilde{\varepsilon}}{2} + \left(\ln\frac{\tilde{\delta}}{3}\right)\frac{\beta\tilde{\delta}\tilde{\varepsilon}^2}{-\ln\tilde{\delta}}\right)\\
&\ge \exp\left(\frac{\tilde{\varepsilon}}{2} + \left(2\ln\tilde{\delta}\right)\frac{\beta\tilde{\varepsilon}}{-3\ln\tilde{\delta}}\right)\\
&= \exp\left(\tilde{\varepsilon}\left(\frac{1}{2} - \frac{2\beta}{3}\right)\right),
\end{aligned}
$$

since $0 < \tilde{\delta} \le 1/3$. Note that this also guarantees $c = e^{\tilde{\varepsilon}/2}\tau^{\alpha-1} \in [1,2]$ since $\beta = e^{-4.2}$ and $0 < \tilde{\varepsilon} \le 1$. We also have

$$\begin{aligned}
\frac{(\alpha-1)(-\ln\tau)}{\tau(2\ln 2-1)} &\le \frac{\frac{\beta\tilde{\delta}\tilde{\varepsilon}^2}{-\ln\tilde{\delta}}(-\ln\tau)}{\tau(2\ln 2-1)} \\
&\le \frac{\frac{\beta\tilde{\delta}\tilde{\varepsilon}^2}{-\ln\tilde{\delta}}(-2\ln\tilde{\delta})}{(\tilde{\delta}/3)(2\ln 2-1)} \\
&= \frac{6\beta\tilde{\varepsilon}^2}{2\ln 2-1} \\
&\le 16\beta\tilde{\varepsilon}^2.
\end{aligned}$$

Hence,

$$\begin{aligned}
\sqrt{\frac{(\alpha-1)(-\ln\tau)}{\tau(2\ln 2-1)}} + 1 &\le 4\tilde{\varepsilon}\sqrt{\beta} + 1 \\
&\le \exp\left(4\tilde{\varepsilon}\sqrt{\beta}\right) \\
&\overset{(a)}{\le} \exp\left(\tilde{\varepsilon}\left(\frac{1}{2} - \frac{2\beta}{3}\right)\right) \\
&\le e^{\tilde{\varepsilon}/2}\tau^{\alpha-1},
\end{aligned}$$

where (a) is by $\beta = e^{-4.2}$. Hence (C.13) is satisfied, and

$$\Pr\Big(\sum_{i:\tilde{T}_i>\tau} \tilde{T}_i^{-\alpha} \ge \frac{e^{\tilde{\varepsilon}/2}}{\alpha-1}\Big) \le \frac{\tilde{\delta}}{3}.$$

Combining this with (C.10) and (C.11),

$$\begin{aligned}
&\Pr\Big(\sum_i \tilde{T}_i^{-\alpha} \notin \Big[\frac{e^{-\tilde{\varepsilon}/2}}{\alpha-1}, \frac{e^{\tilde{\varepsilon}/2}}{\alpha-1}\Big]\Big) \\
&\le \Pr\Big(\sum_i \tilde{T}_i^{-\alpha} \le \frac{e^{-\tilde{\varepsilon}/2}}{\alpha-1}\Big) + \Pr(\min_i \tilde{T}_i \le \tilde{\delta}/3) \\
&\quad + \Pr\Big(\sum_{i:\tilde{T}_i>\tilde{\delta}/3} \tilde{T}_i^{-\alpha} \ge \frac{e^{\tilde{\varepsilon}/2}}{\alpha-1}\Big) \\
&\le \tilde{\delta}.
\end{aligned}$$

Consider an (ε, δ)-differentially private mechanism $P_{Z|X}$. Consider neighbors x_1, x_2, and let $P_j := P_{Z|X}(\cdot|x_j)$, $\tilde{T}_{j,i} := T_i/(\frac{\mathrm{d}P_j}{\mathrm{d}Q}(Z_i))$, and K_j be the output of PPR applied on P_j, for $j = 1, 2$. We first consider the case $\delta = 0$, which gives $\frac{\mathrm{d}P_1}{\mathrm{d}Q}(z) \le e^{\varepsilon}\frac{\mathrm{d}P_2}{\mathrm{d}Q}(z)$ for every z. For any measurable $\mathcal{S} \subseteq \mathcal{Z}^\infty \times \mathbb{Z}_{>0}$,

$$
\begin{aligned}
&\Pr\left(((Z_i)_i, K_1) \in \mathcal{S}\right) \\
&= \mathbb{E}\left[\Pr\left(((Z_i)_i, K_1) \in \mathcal{S} \,\middle|\, (Z_i, T_i)_i\right)\right] \\
&= \mathbb{E}\left[\sum_{k:\,((Z_i)_i,k)\in\mathcal{S}} \Pr\left(K_1 = k \,\middle|\, (Z_i, T_i)_i\right)\right] \\
&= \mathbb{E}\left[\sum_{k:\,((Z_i)_i,k)\in\mathcal{S}} \frac{\tilde{T}_{1,k}^{-\alpha}}{\sum_i \tilde{T}_{1,i}^{-\alpha}}\right] \\
&\le \mathbb{E}\left[\mathbf{1}\left\{\sum_i \tilde{T}_{1,i}^{-\alpha} \in \left[\frac{e^{-\tilde{\varepsilon}/2}}{\alpha-1}, \frac{e^{\tilde{\varepsilon}/2}}{\alpha-1}\right]\right\} \min\left\{\sum_{k:\,((Z_i)_i,k)\in\mathcal{S}} \frac{\tilde{T}_{1,k}^{-\alpha}}{\sum_i \tilde{T}_{1,i}^{-\alpha}}, 1\right\}\right] + \tilde{\delta} \\
&\le \mathbb{E}\left[\min\left\{\sum_{k:\,((Z_i)_i,k)\in\mathcal{S}} \frac{\tilde{T}_{1,k}^{-\alpha}}{e^{-\tilde{\varepsilon}/2}/(\alpha-1)}, 1\right\}\right] + \tilde{\delta} \\
&= \mathbb{E}\left[\min\left\{\sum_{k:\,((Z_i)_i,k)\in\mathcal{S}} \frac{(\frac{\mathrm{d}P_1}{\mathrm{d}Q}(Z_k))^\alpha T_k^{-\alpha}}{e^{-\tilde{\varepsilon}/2}/(\alpha-1)}, 1\right\}\right] + \tilde{\delta} \\
&\le \mathbb{E}\left[\min\left\{\sum_{k:\,((Z_i)_i,k)\in\mathcal{S}} \frac{(e^{\varepsilon}\frac{\mathrm{d}P_2}{\mathrm{d}Q}(Z_k))^\alpha T_k^{-\alpha}}{e^{-\tilde{\varepsilon}/2}/(\alpha-1)}, 1\right\}\right] + \tilde{\delta} \\
&\le \mathbb{E}\left[\mathbf{1}\left\{\sum_i \tilde{T}_{2,i}^{-\alpha} \in \left[\frac{e^{-\tilde{\varepsilon}/2}}{\alpha-1}, \frac{e^{\tilde{\varepsilon}/2}}{\alpha-1}\right]\right\} \min\left\{\sum_{k:\,((Z_i)_i,k)\in\mathcal{S}} \frac{(e^{\varepsilon}\frac{\mathrm{d}P_2}{\mathrm{d}Q}(Z_k))^\alpha T_k^{-\alpha}}{e^{-\tilde{\varepsilon}/2}/(\alpha-1)}, 1\right\}\right] + 2\tilde{\delta} \\
&\le \mathbb{E}\left[\min\left\{e^{\alpha\varepsilon} \sum_{k:\,((Z_i)_i,k)\in\mathcal{S}} \frac{(\frac{\mathrm{d}P_2}{\mathrm{d}Q}(Z_k))^\alpha T_k^{-\alpha}}{e^{-\tilde{\varepsilon}}\sum_i \tilde{T}_{2,i}^{-\alpha}}, 1\right\}\right] + 2\tilde{\delta} \\
&\le \mathbb{E}\left[e^{\alpha\varepsilon+\tilde{\varepsilon}} \sum_{k:\,((Z_i)_i,k)\in\mathcal{S}} \frac{\tilde{T}_{2,k}^{-\alpha}}{\sum_i \tilde{T}_{2,i}^{-\alpha}}\right] + 2\tilde{\delta} \\
&= e^{\alpha\varepsilon+\tilde{\varepsilon}} \Pr\left(((Z_i)_i, K_2) \in \mathcal{S}\right) + 2\tilde{\delta}. \qquad \text{(C.14)}
\end{aligned}
$$

Hence PPR is $(\alpha\varepsilon + \tilde{\varepsilon},\, 2\tilde{\delta})$-differentially private.

For the case $\delta > 0$, by the definition of (ε, δ)-differential privacy, we have

$$\int \max\left\{\frac{\mathrm{d}P_1}{\mathrm{d}Q}(z) - e^{\varepsilon}\frac{\mathrm{d}P_2}{\mathrm{d}Q}(z),\, 0\right\} Q(\mathrm{d}z) \le \delta.$$

Let P_3 be a probability measure that satisfies

$$\min\left\{\frac{\mathrm{d}P_1}{\mathrm{d}Q}(z),\, e^{\varepsilon}\frac{\mathrm{d}P_2}{\mathrm{d}Q}(z)\right\} \le \frac{\mathrm{d}P_3}{\mathrm{d}Q}(z) \le e^{\varepsilon}\frac{\mathrm{d}P_2}{\mathrm{d}Q}(z),$$

for every z. Such P_3 can be constructed by taking an appropriate convex combination of the lower bound above (which integrates to ≤ 1) and the upper bound above (which integrates to ≥ 1) such that P_3 integrates to 1. We have

$$\int \max\left\{\frac{\mathrm{d}P_1}{\mathrm{d}Q}(z) - \frac{\mathrm{d}P_3}{\mathrm{d}Q}(z),\, 0\right\} Q(\mathrm{d}z) \le \delta,$$

and hence the total variation distance $d_{\mathrm{TV}}(P_1, P_3)$ between P_1 and P_3 is at most δ. Let K_3 be the output of PPR applied on P_3.

In the proof of Theorem C.6.1, we see that PPR has the following equivalent formulation. Let $(T_i)_i \sim \mathrm{PP}(1)$ be a Poisson process with rate 1, independent of $Z_1, Z_2, \ldots \overset{iid}{\sim} Q$. Let $R_i := (\mathrm{d}P/\mathrm{d}Q)(Z_i)$, and let its probability measure be P_R. Let $V_1, V_2, \ldots \overset{iid}{\sim} \mathrm{Exp}(1)$. PPR can be equivalently expressed as

$$K = \operatorname*{argmin}_k T_k^{\alpha} R_k^{-\alpha} V_k = \operatorname*{argmin}_k \frac{T_k V_k^{1/\alpha}}{R_k}.$$

Note that $(T_i V_i^{1/\alpha})_i \sim \mathrm{PP}(\int_0^{\infty} v^{-1/\alpha} e^{-v}\mathrm{d}v) = \mathrm{PP}(\Gamma(1-\alpha^{-1}))$ is a uniform Poisson process. Therefore PPR is the same as the Poisson functional representation [117, 118] applied on $(T_i V_i^{1/\alpha})_i$. By the grand coupling property of Poisson functional representation [116, 117] (see [116, Theorem 3]), if we apply the Poisson functional representation on P_1 and P_3 to get K_1 and K_3 respectively, then

$$\Pr(K_1 \ne K_3) \le 2 d_{\mathrm{TV}}(P_1, P_3) \le 2\delta.$$

Therefore, for any measurable $\mathcal{S} \subseteq \mathcal{Z}^\infty \times \mathbb{Z}_{>0}$,

$$\begin{aligned}
&\Pr\left(((Z_i)_i, K_1) \in \mathcal{S}\right) \\
&\leq \Pr\left(((Z_i)_i, K_3) \in \mathcal{S}\right) + 2\delta \\
&\leq e^{\alpha\varepsilon+\tilde{\varepsilon}} \Pr\left(((Z_i)_i, K_2) \in \mathcal{S}\right) + 2\tilde{\delta} + 2\delta,
\end{aligned}$$

where the last inequality is by applying (C.14) on P_3, P_2 instead of P_1, P_2. This completes the proof.

□

C.6 Proof of Theorem 5.4.2

We now bound the size of the index output by the Poisson private representation. The following is a refined version of Theorem 5.4.2.

Theorem C.6.1. *For PPR with parameter $\alpha > 1$, when the encoder is given the input x, the message K given by PPR satisfies*

$$\begin{aligned}
\mathbb{E}[\log K] &\leq D(P\|Q) \\
&\quad + \inf_{\eta \in (0,1] \cap (0, \alpha-1)} \frac{1}{\eta} \log\left(\frac{\Gamma(1-\frac{\eta+1}{\alpha})\Gamma(\eta+1)}{(\Gamma(1-\frac{1}{\alpha}))^{\eta+1}} + 1\right) \qquad \text{(C.15)} \\
&\leq D(P\|Q) + \frac{\log(3.56)}{\min\{(\alpha-1)/2, 1\}}, \qquad \text{(C.16)}
\end{aligned}$$

where $P := P_{Z|X}(\cdot|x)$.

Note that for $\alpha = \infty$, (C.15) with $\eta = 1$ gives $\mathbb{E}[\log K] \leq D(P\|Q) + \log 2$, recovering the bound in [114] (which strengthened [118]).

Proof. Write $(X_i)_i \sim \mathrm{PP}(\mu)$ if the points $(X_i)_i$ (as a multiset, ignoring the ordering) form a Poisson point process with intensity measure μ. Similarly, for $f : [0,\infty)^n \to [0,\infty)$, we write $\mathrm{PP}(f)$ for the Poisson point process with intensity

function f (i.e., the intensity measure has a Radon-Nikodym derivative f against the Lebesgue measure). Let $(T_i)_i \sim \mathrm{PP}(1)$ be a Poisson process with rate 1, independent of $Z_1, Z_2, \ldots \overset{iid}{\sim} Q$. Let $R_i := (\mathrm{d}P/\mathrm{d}Q)(Z_i)$, and let its probability measure be P_R. We have $\tilde{T}_i = T_i/R_i$. Let $V_1, V_2, \ldots \overset{iid}{\sim} \mathrm{Exp}(1)$. By the property of exponential random variables, for any $p_1, p_2, \ldots \geq 0$ with $\sum_i p_i < \infty$, we have $\Pr(\mathrm{argmin}_k V_k/p_k = k) = p_k/\sum_i p_i$. Therefore, PPRF can be equivalently expressed as

$$K = \operatorname*{argmin}_k T_k^\alpha R_k^{-\alpha} V_k.$$

By the marking theorem [109], $(T_i, R_i, V_i)_i$ is a Poisson process over $[0,\infty)^3$ with intensity measure

$$(T_i, R_i, V_i)_i \sim \mathrm{PP}\left(e^{-v} P_R(r)\right).$$

By the mapping theorem [109], letting $W_i := T_i^\alpha R_i^{-\alpha} V_i$, we have

$$(T_i, R_i, W_i)_i \sim \mathrm{PP}\left(r^\alpha t^{-\alpha} e^{-w r^\alpha t^{-\alpha}} P_R(r)\right). \tag{C.17}$$

Again by the mapping theorem,

$$\begin{aligned}
(W_i)_i &\sim \mathrm{PP}\left(\mathbb{E}_{R\sim P_R}\left[\int_0^\infty R^\alpha t^{-\alpha} e^{-wR^\alpha t^{-\alpha}} \mathrm{d}t\right]\right) \\
&= \mathrm{PP}\left(\mathbb{E}\left[\alpha^{-1}(wR^\alpha)^{1/\alpha-1}\Gamma(1-\alpha^{-1})R^\alpha\right]\right) \\
&= \mathrm{PP}\left(\mathbb{E}\left[\alpha^{-1}w^{1/\alpha-1}\Gamma(1-\alpha^{-1})R\right]\right) \\
&= \mathrm{PP}\left(\alpha^{-1}w^{1/\alpha-1}\Gamma(1-\alpha^{-1})\right)
\end{aligned}$$

since $\mathbb{E}[R] = \int (\mathrm{d}P/\mathrm{d}Q)(z) Q(\mathrm{d}z) = 1$. Recall that $W_K = \min_i W_i$ by the definition of K. We have

$$\begin{aligned}
\Pr(W_K > w) &= \exp\left(-\int_0^w \alpha^{-1} v^{1/\alpha-1}\Gamma(1-\alpha^{-1})\mathrm{d}v\right) \\
&= \exp\left(-w^{1/\alpha}\Gamma(1-\alpha^{-1})\right).
\end{aligned}$$

Hence the probability density function of W_K is

$$\begin{aligned}&-\frac{\mathrm{d}}{\mathrm{d}w}\exp\left(-w^{1/\alpha}\Gamma(1-\alpha^{-1})\right)\\&=\alpha^{-1}w^{1/\alpha-1}\Gamma(1-\alpha^{-1})\exp\left(-w^{1/\alpha}\Gamma(1-\alpha^{-1})\right).\end{aligned}\tag{C.18}$$

By (C.17), the Radon-Nikodym derivative between the conditional distribution of R_K given $W_K = w$ and P_R is

$$\begin{aligned}&\Pr(R_K\in[r,r+\mathrm{d}r)\,|\,W_K=w)/P_R(\mathrm{d}r)\\&=\frac{\int_0^\infty r^\alpha t^{-\alpha}e^{-wr^\alpha t^{-\alpha}}\mathrm{d}t}{\mathbb{E}_{R\sim P_R}\left[\int_0^\infty R^\alpha t^{-\alpha}e^{-wR^\alpha t^{-\alpha}}\mathrm{d}t\right]}\\&=\frac{\alpha^{-1}w^{1/\alpha-1}\Gamma(1-\alpha^{-1})r}{\alpha^{-1}w^{1/\alpha-1}\Gamma(1-\alpha^{-1})}\\&=r\end{aligned}$$

does not depend on w. Hence R_K is independent of W_K. By (C.17), for $0\le\eta<\alpha-1$,

$$\begin{aligned}&\mathbb{E}[T_K^\eta\,|\,R_K=r,\,W_K=w]\\&=\frac{\int_0^\infty t^\eta r^\alpha t^{-\alpha}e^{-wr^\alpha t^{-\alpha}}\mathrm{d}t}{\int_0^\infty r^\alpha t^{-\alpha}e^{-wr^\alpha t^{-\alpha}}\mathrm{d}t}\\&=\frac{\alpha^{-1}w^{(\eta+1)/\alpha-1}\Gamma(1-(\eta+1)\alpha^{-1})r^{\eta+1}}{\alpha^{-1}w^{1/\alpha-1}\Gamma(1-\alpha^{-1})r}.\end{aligned}\tag{C.19}$$

Since R_K is independent of W_K, using (C.19) and (C.18), for $\eta\in(0,1]\cap(0,\alpha-1)$,

$$\begin{aligned}&\mathbb{E}[T_K^\eta\,|\,R_K=r]\\&=\int_0^\infty\alpha^{-1}w^{(\eta+1)/\alpha-1}\Gamma(1-(\eta+1)\alpha^{-1})r^\eta\exp\left(-w^{1/\alpha}\Gamma(1-\alpha^{-1})\right)\mathrm{d}w\\&=r^\eta\Gamma(1-(\eta+1)\alpha^{-1})\int_0^\infty\alpha^{-1}w^{(\eta+1)/\alpha-1}\exp\left(-w^{1/\alpha}\Gamma(1-\alpha^{-1})\right)\mathrm{d}w\\&=r^\eta\Gamma(1-(\eta+1)\alpha^{-1})(\Gamma(1-\alpha^{-1}))^{-(\eta+1)}\Gamma(\eta+1)\end{aligned}$$

$$=: c_{\alpha,\eta} r^{\eta}, \tag{C.20}$$

where $c_{\alpha,\eta} := \Gamma(1-(\eta+1)\alpha^{-1})(\Gamma(1-\alpha^{-1}))^{-(\eta+1)}\Gamma(\eta+1)$. Hence,

$$\begin{aligned}
&\mathbb{E}[\log(T_K+1) \,|\, R_K = r] \\
&\le \mathbb{E}[\log((T_K^{\eta}+1)^{1/\eta}) \,|\, R_K = r] \\
&= \mathbb{E}[\eta^{-1}\log(T_K^{\eta}+1) \,|\, R_K = r] \\
&\le \eta^{-1}\log(c_{\alpha,\eta} r^{\eta}+1).
\end{aligned} \tag{C.21}$$

Note that

$$K - 1 = \left|\{i : T_i < T_K\}\right|,$$

and hence the expecation of $K-1$ given T_K should be around T_K. This is not exact since conditioning on T_K changes the distribution of the process $(T_i, R_i, V_i)_i$. To resolve this problem, we define a new process $(T_i', R_i', V_i')_i$ which includes all points in $(T_i, R_i, V_i)_i$ excluding the point (T_K, R_K, V_K), together with newly generated points according to

$$\mathrm{PP}\left(e^{-v} P_R(r) \mathbf{1}\{t^{\alpha} r^{-\alpha} v < T_K^{\alpha} R_K^{-\alpha} V_K\}\right).$$

Basically, $\{t^{\alpha} r^{-\alpha} v < T_K^{\alpha} R_K^{-\alpha} V_K\}$ is the "impossible region" where the points in $(T_i, R_i, V_i)_i$ cannot be located in, since K attains the minimum of $T_K^{\alpha} R_K^{-\alpha} V_K$. The new process $(T_i', R_i', V_i')_i$ removes the point (T_K, R_K, V_K), and then fills in the impossible region. It is straightforward to check that $(T_i', R_i', V_i')_i \sim \mathrm{PP}(e^{-v} P_R(r))$, independent of (T_K, R_K, V_K). We have

$$\begin{aligned}
&\mathbb{E}[K \,|\, T_K] \\
&= \mathbb{E}\left[\left|\{i : T_i < T_K\}\right| \,\middle|\, T_K\right] + 1 \\
&\le \mathbb{E}\left[\left|\{i : T_i' < T_K\}\right| \,\middle|\, T_K\right] + 1 \\
&= T_K + 1.
\end{aligned}$$

Therefore, by (C.21) and Jensen's inequality,

$$
\begin{aligned}
&\mathbb{E}[\log K] \\
&= \mathbb{E}\left[\mathbb{E}[\log K \,|\, T_K]\right] \\
&\le \mathbb{E}\left[\log(T_K+1)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\log(T_K+1) \,\middle|\, R_K\right]\right] \\
&\le \mathbb{E}\left[\eta^{-1}\log(c_{\alpha,\eta}R_K^{\eta}+1)\right] \\
&= \eta^{-1}\mathbb{E}_{Z\sim P}\left[\log\left(c_{\alpha,\eta}\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(Z)\right)^{\eta}+1\right)\right] \\
&= \eta^{-1}\mathbb{E}\left[\log\left(\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(Z)\right)^{\eta}\right)\right] + \eta^{-1}\mathbb{E}\left[\log\left(c_{\alpha,\eta}+\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(Z)\right)^{-\eta}\right)\right] \\
&\le D(P\|Q) + \eta^{-1}\log\left(c_{\alpha,\eta}+\left(\mathbb{E}\left[\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(Z)\right)^{-1}\right]\right)^{\eta}\right) \\
&\le D(P\|Q) + \eta^{-1}\log(c_{\alpha,\eta}+1),
\end{aligned}
$$

where the last line is due to $\mathbb{E}[((\mathrm{d}P/\mathrm{d}Q)(Z))^{-1}] = \int((\mathrm{d}P/\mathrm{d}Q)(Z))^{-1}P(\mathrm{d}z) \le 1$ (this step appeared in [114]). The bound (C.15) follows from minimizing over $\eta \in (0,1] \cap (0, \alpha-1)$.

To show (C.16), substituting $\eta = \min\{(\alpha-1)/2, 1\}$,

$$
\begin{aligned}
c_{\alpha,\eta} &= \frac{\Gamma(1-(\eta+1)\alpha^{-1})\Gamma(\eta+1)}{(\Gamma(1-\alpha^{-1}))^{\eta+1}} \\
&\overset{(a)}{\le} \frac{(1-\alpha^{-1})^{\eta+1}}{0.885^{\eta+1}\cdot(1-(\eta+1)\alpha^{-1})} \\
&\le \frac{(1-\alpha^{-1})^{\eta+1}}{0.885^{2}\cdot(1-((\alpha-1)/2+1)\alpha^{-1})} \\
&= \frac{2}{0.885^2}(1-\alpha^{-1})^{\eta} \\
&\le 2.56,
\end{aligned}
$$

where (a) is because $0.885 \le x\Gamma(x) = \Gamma(x+1) \le 1$ for $0 < x \le 1$. Hence,

$$
\mathbb{E}[\log K] \le D(P\|Q) + \eta^{-1}\log(c_{\alpha,\eta}+1),
$$

$$\leq D(P\|Q) + \frac{\log(3.56)}{\min\{(\alpha-1)/2, 1\}}.$$

□

C.7 Distributed Mean Estimation with Rényi DP

In many machine learning applications, privacy budgets are often accounted in the moment space, and one popular moment accountant is the Rényi DP accountant. For completeness, we provide a Rényi DP version of Corollary 5.5.2 in this section. We begin with the following definition of Rényi DP:

Definition C.7.1 (Rényi Differential privacy [1, 141])**.** Given a mechanism $\mathcal{A}$ which induces the conditional distribution $P_{Z|X}$ of $Z = \mathcal{A}(X)$, we say that it satisfies (γ, ε)- Rényi DP if for any neighboring $(x, x') \in \mathcal{N}$ and $\mathcal{S} \subseteq \mathcal{Z}$, it holds that

$$D_\gamma\left(P_{Z|X=x} \middle\| P_{Z|X=x'}\right) \leq \varepsilon,$$

where

$$D_\gamma\left(P\|Q\right) := \frac{1}{\gamma-1}\log\left(\mathbb{E}_Q\left[\left(\frac{P}{Q}\right)^\gamma\right]\right)$$

is the Rényi divergence between P and Q. If $\mathcal{N} = \mathcal{X}^2$, we say that the mechanism satisfies (γ, ε)-local DP.

The following conversion lemma from [26] relates Rényi DP to $(\varepsilon_{\mathsf{DP}}(\delta), \delta)$-DP.

Lemma C.7.2. *If $\mathcal{A}$ satisfies (γ, ε)-Rényi DP for some $\gamma \geq 1$, then, for any $\delta > 0$, $\mathcal{A}$ satisfies $(\varepsilon_{\mathsf{DP}}(\delta), \delta)$-DP, where*

$$\varepsilon_{\mathsf{DP}}(\delta) = \varepsilon + \frac{\log\left(1/\gamma\delta\right)}{\gamma-1} + \log(1-1/\gamma). \tag{C.22}$$

The following theorem states that, when simulating the Gaussian mechanism, PPR satisfies the following both central and local DP guarantee:

Corollary C.7.3 (PPR-compressed Gaussian mechanism)**.** *Let $\varepsilon \geq 0$ and $\gamma \geq 1$. Consider the Gaussian mechanism $P_{Z|X}(\cdot|x) = \mathcal{N}(x, \frac{\sigma^2}{n}\mathbb{I}_d)$, and the proposal distribution $Q = \mathcal{N}(0, (\frac{C^2}{d} + \frac{\sigma^2}{n})\mathbb{I}_d)$, where $\sigma \geq \sqrt{\frac{C\gamma}{2\varepsilon}}$. For each client i, let Z_i be the output of PPR applied on $P_{Z|X}(\cdot|X_i)$. We have:*

- *$\hat{\mu}(Z^n) := \frac{1}{n}\sum_i Z_i$ yields an unbiased estimator of $\mu(X^n) = \frac{1}{n}\sum_{i=1}^n X_i$ satisfying (γ, ε)-(central) Rényi DP and $(\varepsilon_{\mathsf{DP}}(\delta), \delta)$-DP, where $\varepsilon_{\mathsf{DP}}(\delta)$ is defined in (C.22).*
- *$P_{Z|X_i}$ satisfies $(2\alpha\tilde{\varepsilon}_{\mathsf{DP}}(\delta), 2\delta)$-local DP, where*

$$\tilde{\varepsilon}_{\mathsf{DP}}(\delta) := \sqrt{n}\varepsilon + \frac{\log(1/\gamma\delta)}{\gamma - 1} + \log(1 - 1/\gamma).$$

- *$\hat{\mu}(Z^n)$ has MSE $\mathbb{E}[\|\mu - \hat{\mu}\|_2^2] = \sigma^2 d/n^2$.*
- *The average per-client communication cost is at most $\ell + \log_2(\ell + 1) + 2$ bits where*

$$\ell := \frac{d}{2}\log_2\left(\frac{C^2 n}{d\sigma^2} + 1\right) + \eta_\alpha \;\leq\; \frac{d}{2}\log_2\left(\frac{n\varepsilon^2}{2d\ln(1.25/\delta)} + 1\right) + \eta_\alpha,$$

where $\eta_\alpha := (\log_2(3.56))/\min\{(\alpha - 1)/2, 1\}$.

Proof. The central DP guarantee follows from [141] and Lemma C.7.2. The local DP guarantee follows from Lemma C.7.2 and Theorem 5.4.8. Finally, the communication bound can be obtained from the same analysis as in Corollary 5.5.2. □

C.8 Proof of Corollary 5.5.2

Consider the PPR applied on the Gaussian mechanism $P_{Z|X}(\cdot|x) = \mathcal{N}(x, \frac{\sigma^2}{n}\mathbb{I}_d)$, with the proposal distribution $Q = \mathcal{N}(0, (\frac{C^2}{d} + \frac{\sigma^2}{n})\mathbb{I}_d)$. PPR ensures that Z_i

follows the distribution $\mathcal{N}(X_i, \frac{\sigma^2}{n}\mathbb{I}_d)$. Therefore the MSE is

$$\begin{aligned}\mathbb{E}\left[\|\mu - \hat{\mu}\|_2^2\right] &= \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}(X_i - Z_i)\right\|_2^2\right] \\ &= \frac{1}{n}\cdot d \cdot \frac{\sigma^2}{n} \\ &= \frac{\sigma^2 d}{n^2}.\end{aligned}$$

For the compression size, for $x \in \mathbb{R}^d$ with $\|x\|_2 \le C$, we have

$$\begin{aligned}&D(P_{Z|X}(\cdot|x)\|Q) \\ &= \mathbb{E}_{Z\sim P_{Z|X}(\cdot|x)}\left[\log\frac{\mathrm{d}P_{Z|X}(\cdot|x)}{\mathrm{d}Q}(Z)\right] \\ &= \mathbb{E}_{Z\sim P_{Z|X}(\cdot|x)}\left[\log\frac{(2\pi\sigma^2/n)^{-d/2}\exp(-\frac{1}{2}\|Z-x\|_2^2/(\sigma^2/n))}{(2\pi(\frac{C^2}{d}+\frac{\sigma^2}{n}))^{-d/2}\exp(-\frac{1}{2}\|Z\|_2^2/(\frac{C^2}{d}+\frac{\sigma^2}{n}))}\right] \\ &= \mathbb{E}_{Z\sim P_{Z|X}(\cdot|x)}\left[\frac{d}{2}\log\frac{\frac{C^2}{d}+\frac{\sigma^2}{n}}{\sigma^2/n} + \frac{1}{2}\left(\frac{\|Z\|_2^2}{\frac{C^2}{d}+\frac{\sigma^2}{n}} - \frac{\|Z-x\|_2^2}{\sigma^2/n}\right)\right] \\ &\le \frac{d}{2}\log\frac{\frac{C^2}{d}+\frac{\sigma^2}{n}}{\sigma^2/n} + \frac{1}{2}\left(\frac{C^2+\sigma^2 d/n}{\frac{C^2}{d}+\frac{\sigma^2}{n}} - \frac{\sigma^2 d/n}{\sigma^2/n}\right) \\ &= \frac{d}{2}\log\left(\frac{C^2 n}{d\sigma^2}+1\right).\end{aligned}$$

Hence, by Theorem 5.4.2, the compression size is at most $\ell + \log_2(\ell+1) + 2$ bits, where

$$\begin{aligned}\ell &:= \frac{d}{2}\log_2\left(\frac{C^2 n}{d\sigma^2}+1\right) + \eta_\alpha \\ &\le \frac{d}{2}\log_2\left(\frac{n\epsilon^2}{2d\ln(1.25/\delta)}+1\right) + \eta_\alpha \\ &\le \frac{n\epsilon^2\log_2(e)}{4\ln(1.25/\delta)} + \eta_\alpha,\end{aligned}$$

where $\eta_\alpha := (\log_2(3.56))/\min\{(\alpha-1)/2,\, 1\}$.

The central-DP guarantee follows from (ε, δ)-DP of Gaussian mechanism [52, Appendix A] since the output distribution of PPR is exactly the same as the

Gaussian mechanism, whereas the local-DP guarantee follows from Theorem 5.4.6 and [52, Appendix A].

C.9 Proof of Corollary 5.7.1

Let $\|X - Z\|_2 = RS$ where $R \in [0, \infty)$ is the magnitude of $X - Z$, and $\|S\|_2 = 1$. As shown in [66], R follows the Gamma distribution with shape d and scale $1/\varepsilon$. Hence the MSE is

$$\mathbb{E}\left[\|X - Z\|_2^2\right] = \mathbb{E}\left[R^2\right] = \left(\frac{d}{\varepsilon}\right)^2 + \frac{d}{\varepsilon^2} = \frac{d(d+1)}{\varepsilon^2}.$$

The conditional differential entropy (in nats) of Z given X is

$$\begin{aligned}
h(Z|X) &= h(R) + h(S|R) \\
&= d + \ln\Gamma(d) - (d-1)\psi(d) - \ln\varepsilon + \mathbb{E}\left[\ln(nR^{d-1}\mathrm{Vol}(\mathcal{B}_d(1)))\right] \\
&= d + \ln\Gamma(d) - (d-1)\psi(d) - \ln\varepsilon + \ln d + \ln(\mathrm{Vol}(\mathcal{B}_d(1))) + (d-1)\mathbb{E}\left[\ln R\right] \\
&= d + \ln\Gamma(d) - (d-1)\psi(d) - \ln\varepsilon + \ln d + \frac{d}{2}\ln\pi - \ln\Gamma\left(\frac{d}{2}+1\right) \\
&\qquad - (d-1)\ln\epsilon + (d-1)\psi(d) \\
&= d\ln\frac{e\sqrt{\pi}}{\varepsilon} + \ln\frac{d\Gamma(d)}{\Gamma(\frac{d}{2}+1)},
\end{aligned}$$

where ψ is the digamma function. Therefore, the KL divergence between $P_{Z|X}(\cdot|x)$ and Q (in nats) is

$$\begin{aligned}
&D(P_{Z|X}(\cdot|x)\|Q) \\
&= -\mathbb{E}_{Z\sim P_{Z|X}(\cdot|x)}\left[\ln\left(\left(2\pi\left(\frac{C^2}{d} + \frac{d+1}{\varepsilon^2}\right)\right)^{-d/2}\exp\left(-\frac{\|Z\|_2^2}{2(\frac{C^2}{d}+\frac{d+1}{\varepsilon^2})}\right)\right)\right] - h(Z|X) \\
&= \frac{d}{2}\ln\left(2\pi\left(\frac{C^2}{d} + \frac{d+1}{\varepsilon^2}\right)\right) + \frac{\mathbb{E}_{Z\sim P_{Z|X}(\cdot|x)}\left[\|Z\|_2^2\right]}{2(\frac{C^2}{d}+\frac{d+1}{\varepsilon^2})} - d\ln\frac{e\sqrt{\pi}}{\varepsilon} - \ln\frac{d\Gamma(d)}{\Gamma(\frac{d}{2}+1)}
\end{aligned}$$

$$\leq \frac{d}{2}\ln\left(2\pi\left(\frac{C^2}{d}+\frac{d+1}{\varepsilon^2}\right)\right)+\frac{C^2+\frac{d(d+1)}{\varepsilon^2}}{2(\frac{C^2}{d}+\frac{d+1}{\varepsilon^2})}-d\ln\frac{e\sqrt{\pi}}{\varepsilon}-\ln\frac{d\Gamma(d)}{\Gamma(\frac{d}{2}+1)}$$
$$=\frac{d}{2}\ln\left(\frac{2}{e}\left(\frac{C^2\varepsilon^2}{d}+d+1\right)\right)-\ln\frac{\Gamma(d+1)}{\Gamma(\frac{d}{2}+1)}.$$

Hence, by Theorem 5.4.2, the compression size is at most $\ell+\log_2(\ell+1)+2$ bits. The metric privacy guarantee follows from Theorem 5.4.7.

C.10 MSE against Compression Size

We plot the MSE against the compression size (ranging from 25 to 1000 bits) for $\epsilon \in \{0.25, 0.5, 1.0, 2.0\}$ in Figure C.1 as follows.

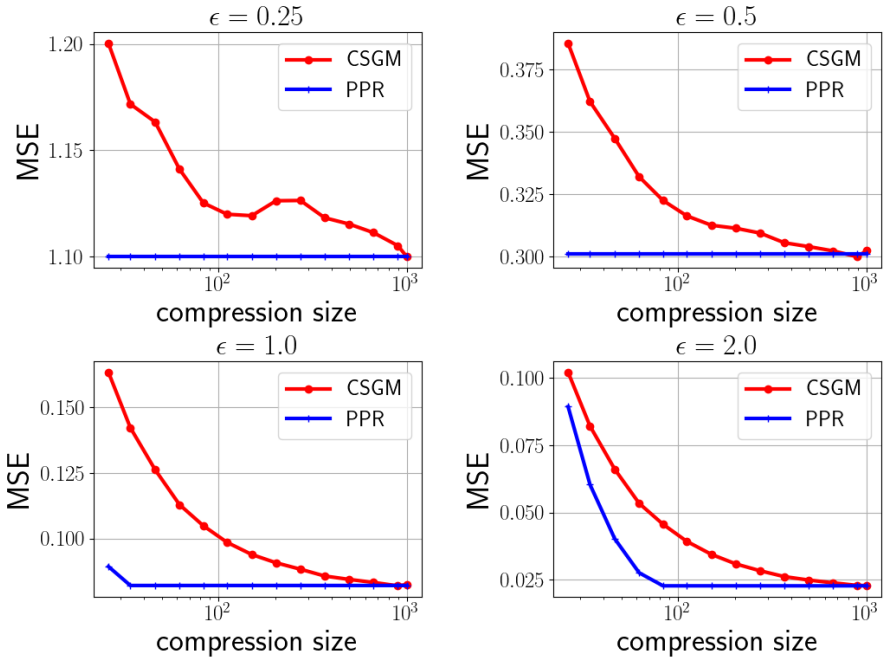

Figure C.1: The MSE of PPR and CSGM against the compression size in bits, where ε is chosen from $\{0.25, 0.5, 1.0, 2.0\}$ and compression sizes vary from 25 to 1000 bits. Note that parts of the curves for PPR are flat, because a lower compression size is already sufficient for PPR to exactly simulate the best Gaussian mechanism for that value of ε, so a higher compression size than necessary will not affect the result.

Bibliography

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016. 86, 135

[2] Jayadev Acharya and Ziteng Sun. Communication complexity in locally private distribution estimation and heavy hitters. In *International Conference on Machine Learning*, pages 51–60. PMLR, 2019. 7

[3] Eirikur Agustsson and Lucas Theis. Universally quantized neural compression. *Advances in neural information processing systems*, 33:12367–12376, 2020. 77, 96, 104

[4] Rudolf Ahlswede. Multi-way communication channels. In *2nd Int. Symp. Inform. Theory, Tsahkadsor, Armenian SSR*, pages 23–52, 1971. 21, 41, 44, 102

[5] Rudolf Ahlswede. The capacity region of a channel with two senders and two receivers. *The annals of probability*, 2(5):805–814, 1974. 21, 41, 44, 102

[6] Miguel E Andrés, Nicolás E Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. Geo-indistinguishability: Differential privacy for

location-based systems. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, pages 901–914, 2013. xv, 78, 92, 93, 94, 95

[7] Erdal Arikan. Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels. *IEEE Transactions on information Theory*, 55(7):3051–3073, 2009. 104

[8] Hilal Asi, Vitaly Feldman, and Kunal Talwar. Optimal algorithms for mean estimation under local differential privacy. In *International Conference on Machine Learning*, pages 1046–1056. PMLR, 2022. 76, 87

[9] Richard J Barron, Brian Chen, and Gregory W Wornell. The duality between information embedding and source coding with side information and some applications. *IEEE Transactions on Information Theory*, 49(5):1159–1180, 2003. 50

[10] Raef Bassily, Kobbi Nissim, Uri Stemmer, and Abhradeep Guha Thakurta. Practical locally private heavy hitters. *Advances in Neural Information Processing Systems*, 30, 2017. 7

[11] Raef Bassily and Adam Smith. Local, private, efficient protocols for succinct histograms. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 127–135, 2015. 76

[12] Charles H Bennett, Igor Devetak, Aram W Harrow, Peter W Shor, and Andreas Winter. The quantum reverse shannon theorem and resource tradeoffs for simulating quantum channels. *IEEE Transactions on Information Theory*, 60(5):2926–2959, 2014. 75, 77

[13] Charles H Bennett, Peter W Shor, John Smolin, and Ashish V Thapliyal. Entanglement-assisted capacity of a quantum channel and the reverse

Shannon theorem. *IEEE Trans. Inf. Theory*, 48(10):2637–2655, 2002. 8, 28, 75, 77, 82

[14] Toby Berger. Multiterminal source coding. In G. Longo, editor, *The Information Theory Approach to Communications*, pages 171–231. Springer-Verlag, New York, 1978. 37

[15] Andrew C Berry. The accuracy of the Gaussian approximation to the sum of independent variates. *Transactions of the American Mathematical Society*, 49(1):122–136, 1941. 60

[16] Abhishek Bhowmick, John Duchi, Julien Freudiger, Gaurav Kapoor, and Ryan Rogers. Protection against reconstruction and its applications in private federated learning. *arXiv preprint arXiv:1812.00984*, 2018. 76, 87

[17] Igor Bjelaković, Holger Boche, and Jochen Sommerfeld. Secrecy results for compound wiretap channels. *Problems of Information Transmission*, 49(1):73–98, 2013. 51, 52

[18] David Blackwell, Leo Breiman, AJ Thomasian, et al. The capacity of a class of channels. *The Annals of Mathematical Statistics*, 30(4):1229–1241, 1959. 49, 52, 56, 61

[19] Holger Boche, Rafael F Schaefer, and H Vincent Poor. On the continuity of the secrecy capacity of compound and arbitrarily varying wiretap channels. *IEEE Transactions on Information Forensics and Security*, 10(12):2531–2546, 2015. 52

[20] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 5

[21] Dan Boneh and James Shaw. Collusion-secure fingerprinting for digital data. *IEEE Transactions on Information Theory*, 44(5):1897–1905, 1998. 50

[22] Mark Braverman and Ankit Garg. Public vs private coin in bounded-round information. In *International Colloquium on Automata, Languages, and Programming*, pages 502–513. Springer, 2014. 18, 75, 77, 83

[23] Mark Braverman, Ankit Garg, Tengyu Ma, Huy L Nguyen, and David P Woodruff. Communication lower bounds for statistical estimation problems via a distributed data processing inequality. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 1011–1020, 2016. 76

[24] Mark Bun, Jelani Nelson, and Uri Stemmer. Heavy hitters and the structure of local privacy. In *Proceedings of the 37th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, SIGMOD/PODS ' 18, page 435–447, New York, NY, USA, 2018. Association for Computing Machinery. 76

[25] Mark Bun, Jelani Nelson, and Uri Stemmer. Heavy hitters and the structure of local privacy. *ACM Transactions on Algorithms (TALG)*, 15(4):1–40, 2019. 8, 19

[26] Clément L Canonne, Gautam Kamath, and Thomas Steinke. The discrete Gaussian for differential privacy. *Advances in Neural Information Processing Systems*, 33:15676–15688, 2020. 135

[27] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models.

In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650, 2021. 5

[28] Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 5253–5270, 2023. 5

[29] Konstantinos Chatzikokolakis, Miguel E Andrés, Nicolás Emilio Bordenabe, and Catuscia Palamidessi. Broadening the scope of differential privacy using metrics. In *Privacy Enhancing Technologies: 13th International Symposium, PETS 2013, Bloomington, IN, USA, July 10-12, 2013. Proceedings 13*, pages 82–102. Springer, 2013. 78, 92

[30] Kamalika Chaudhuri, Chuan Guo, and Mike Rabbat. Privacy-aware compression for federated data analysis. In *Uncertainty in Artificial Intelligence*, pages 296–306. PMLR, 2022. 76

[31] Brian Chen and Gregory W Wornell. Quantization index modulation: A class of provably good methods for digital watermarking and information embedding. *IEEE Transactions on Information theory*, 47(4):1423–1443, 2001. 50

[32] Wei-Ning Chen, Peter Kairouz, and Ayfer Özgür. Breaking the communication-privacy-accuracy trilemma. *Advances in Neural Information Processing Systems*, 33:3312–3324, 2020. 76

[33] Wei-Ning Chen, Peter Kairouz, and Ayfer Özgür. Breaking the communication-privacy-accuracy trilemma. *IEEE Transactions on Information Theory*, 69(2):1261–1281, 2022. 7

[34] Wei-Ning Chen, Dan Song, Ayfer Özgür, and Peter Kairouz. Privacy amplification via compression: Achieving the optimal privacy-accuracy-communication trade-off in distributed mean estimation. *Advances in Neural Information Processing Systems*, 36, 2024. xv, 76, 89, 90, 103

[35] Aaron S Cohen and Amos Lapidoth. The gaussian watermarking game. *IEEE transactions on Information Theory*, 48(6):1639–1667, 2002. 48, 49, 50, 55

[36] Thomas Cover and Abbas El Gamal. Capacity theorems for the relay channel. *IEEE Transactions on information theory*, 25(5):572–584, 1979. 31, 35, 36

[37] Ingemar J Cox, Joe Kilian, F Thomson Leighton, and Talal Shamoon. Secure spread spectrum watermarking for multimedia. *IEEE transactions on image processing*, 6(12):1673–1687, 1997. 50, 61

[38] Imre Csiszár. The method of types [information theory]. *IEEE Transactions on Information Theory*, 44(6):2505–2523, 1998. 62

[39] Paul Cuff. Communication requirements for generating correlated random variables. In *2008 IEEE International Symposium on Information Theory*, pages 1393–1397. IEEE, 2008. 76, 82, 114

[40] Paul Cuff. Distributed channel synthesis. *IEEE Trans. Inf. Theory*, 59(11):7071–7096, Nov 2013. 8, 28, 75, 77, 86

[41] Paul Cuff, Haim Permuter, and Thomas M Cover. Coordination capacity. *IEEE Trans. Inf. Theory*, 56(9):4181–4206, Sept 2010. 28

[42] Paul Cuff, Han-I Su, and Abbas El Gamal. Cascade multiterminal source

coding. In *2009 IEEE International Symposium on Information Theory*, pages 1199–1203. IEEE, 2009. 21, 22, 36, 37, 38, 41, 102

[43] Paul Cuff and Lanqing Yu. Differential privacy as a mutual information constraint. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 43–54, 2016. 96

[44] Paul W Cuff. *Communication in networks for coordinating behavior*. Stanford University, 2009. 18

[45] RL Dobrushin. Optimum information transmission through a channel with unknown parameters. *Radio Eng. Electron*, 4(12):1–8, 1959. 52, 61

[46] John Duchi and Ryan Rogers. Lower bounds for locally private estimation via communication complexity. In *Conference on Learning Theory*, pages 1161–1191. PMLR, 2019. 76

[47] John C Duchi, Michael I Jordan, and Martin J Wainwright. Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 429–438. IEEE, 2013. 76

[48] Giuseppe Durisi, Tobias Koch, and Petar Popovski. Toward massive, ultra-reliable, and low-latency wireless communication with short packets. *Proceedings of the IEEE*, 104(9):1711–1726, 2016. 2

[49] Cynthia Dwork. Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer, 2006. 88

[50] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Advances in Cryptology-EUROCRYPT 2006: 24th Annual International Conference on the Theory and Applications of Cryptographic*

Techniques, St. Petersburg, Russia, May 28-June 1, 2006. Proceedings 25, pages 486–503. Springer, 2006. 74

[51] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006. 5, 76, 77, 78, 86

[52] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014. 137, 138

[53] Ersen Ekrem and Sennur Ulukus. On gaussian mimo compound wiretap channels. In *2010 44th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE, 2010. 52

[54] Ersen Ekrem and Sennur Ulukus. Degraded compound multi-receiver wiretap channels. *IEEE Transactions on Information Theory*, 58(9):5681–5698, 2012. 52

[55] Abbas El Gamal, Amin Gohari, and Chandra Nair. Achievable rates for the relay channel with orthogonal receiver components. In *2021 IEEE Information Theory Workshop (ITW)*, pages 1–6. IEEE, 2021. 21, 22, 31, 36, 102

[56] Abbas El Gamal, Amin Gohari, and Chandra Nair. A strengthened cutset upper bound on the capacity of the relay channel and applications. *IEEE Transactions on Information Theory*, 68(8):5013–5043, 2022. 21, 22, 31, 102

[57] Abbas El Gamal and Navid Hassanpour. Relay-without-delay. In *Proceedings. International Symposium on Information Theory, 2005. ISIT 2005.*, pages 1078–1080. IEEE, 2005. 21, 22, 31, 102

[58] Abbas El Gamal, Navid Hassanpour, and James Mammen. Relay networks with delays. *IEEE Transactions on Information Theory*, 53(10):3413–3431, 2007. 21, 22, 31, 35, 36, 102

[59] Abbas El Gamal and Young-Han Kim. *Network information theory*. Cambridge university press, 2011. 1, 7, 21, 22, 31, 36, 37, 41, 101, 102

[60] Peter Elias. Universal codeword sets and representations of the integers. *IEEE transactions on information theory*, 21(2):194–203, 1975. 81, 83

[61] Úlfar Erlingsson, Vitaly Feldman, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Abhradeep Thakurta. Amplification by shuffling: From local to central differential privacy via anonymity. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2468–2479. SIAM, 2019. 74

[62] Carl-Gustaf Esseen. On the Liapunov limit error in the theory of probability. *Ark. Mat. Astr. Fys.*, 28:1–19, 1942. 60

[63] Amiel Feinstein. A new basic theorem of information theory. *IRE Trans. Inf. Theory*, (4):2–22, 1954. 2, 3, 18

[64] Vitaly Feldman, Audra McMillan, and Kunal Talwar. Hiding among the clones: A simple and nearly optimal analysis of privacy amplification by shuffling. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 954–964. IEEE, 2022. 74

[65] Vitaly Feldman and Kunal Talwar. Lossless compression of efficient private local randomizers. In *International Conference on Machine Learning*, pages 3208–3219. PMLR, 2021. 6, 7, 8, 19, 74, 76, 85, 89

[66] Natasha Fernandes, Mark Dras, and Annabelle McIver. Generalised differential privacy for text document processing. In *Principles of Security and Trust: 8th International Conference, POST 2019*, pages 123–148. Springer International Publishing, 2019. 92, 93, 94, 138

[67] Oluwaseyi Feyisetan, Borja Balle, Thomas Drake, and Tom Diethe. Privacy- and utility-preserving textual analysis via calibrated multivariate perturbations. In *Proceedings of the 13th international conference on web search and data mining*, pages 178–186, 2020. 92, 93, 94

[68] Gergely Flamich. Greedy poisson rejection sampling. *Advances in Neural Information Processing Systems*, 36, 2024. 14, 75, 77, 104

[69] Gergely Flamich, Marton Havasi, and José Miguel Hernández-Lobato. Compressing images by encoding their latent representations with relative entropy coding. *Advances in Neural Information Processing Systems*, 33:16131–16141, 2020. 19, 77

[70] Gergely Flamich, Stratis Markou, and José Miguel Hernández-Lobato. Fast relative entropy coding with A* coding. In *International Conference on Machine Learning*, pages 6548–6577. PMLR, 2022. 14, 104

[71] Gergely Flamich, Stratis Markou, and José Miguel Hernández Lobato. Faster relative entropy coding with greedy rejection coding. *arXiv preprint arXiv:2309.15746*, 2023. 77

[72] Gergely Flamich and Lucas Theis. Adaptive greedy rejection sampling. In *2023 IEEE International Symposium on Information Theory (ISIT)*, pages 454–459. IEEE, 2023. 18, 96, 117

[73] Ankit Garg, Tengyu Ma, and Huy Nguyen. On communication cost of

distributed statistical estimation and dimensionality. In *Advances in Neural Information Processing Systems*, pages 2726–2734, 2014. 76

[74] S. I. Gel'fand and M. S. Pinsker. Coding for channel with random parameters. *Probl. Contr. and Inf. Theory*, 9(1):19–31, 1980. 21, 41, 50, 52, 55, 61, 102

[75] Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM journal on optimization*, 23(4):2341–2368, 2013. 86

[76] Daniel Goc and Gergely Flamich. On channel simulation with causal rejection samplers. *arXiv preprint arXiv:2401.16579*, 2024. 75, 77

[77] Slawomir Goryczka and Li Xiong. A comprehensive comparison of multiparty secure additions with differential privacy. *IEEE transactions on dependable and secure computing*, 14(5):463–477, 2015. 74

[78] Pulkit Grover, Aaron B Wagner, and Anant Sahai. Information embedding and the triple role of control. *IEEE Transactions on Information Theory*, 61(4):1539–1549, 2015. 51, 52, 54, 62

[79] Qingxiao Guan, Peng Liu, Weiming Zhang, Wei Lu, and Xinpeng Zhang. Double-layered dual-syndrome trellis codes utilizing channel knowledge for robust steganography. *IEEE Transactions on Information Forensics and Security*, 18:501–516, 2022. 50

[80] Chuan Guo, Kamalika Chaudhuri, Pierre Stock, and Michael Rabbat. Privacy-aware compression for federated learning through numerical mechanism design. In *International Conference on Machine Learning*, pages 11888–11904. PMLR, 2023. 76

[81] Yuanxin Guo, Sadaf Salehkalaibar, Stark C Draper, and Wei Yu. One-shot achievability region for hypothesis testing with communication constraint. In *2024 IEEE Information Theory Workshop (ITW)*, pages 55–60. IEEE, 2024. 4, 15

[82] Prahladh Harsha, Rahul Jain, David McAllester, and Jaikumar Radhakrishnan. The communication complexity of correlation. In *Twenty-Second Annual IEEE Conference on Computational Complexity (CCC'07)*, pages 10–23. IEEE, 2007. 8

[83] Prahladh Harsha, Rahul Jain, David McAllester, and Jaikumar Radhakrishnan. The communication complexity of correlation. *IEEE Trans. Inf. Theory*, 56(1):438–449, Jan 2010. 18, 75, 77, 83, 117

[84] Frank Hartung and Martin Kutter. Multimedia watermarking techniques. *Proceedings of the IEEE*, 87(7):1079–1107, 1999. 61

[85] Burak Hasırcıoğlu and Deniz Gündüz. Communication efficient private federated learning using dithering. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7575–7579. IEEE, 2024. 6, 104

[86] Marton Havasi, Robert Peharz, and José Miguel Hernández-Lobato. Minimal random code learning: Getting bits back from compressed model parameters. In *7th International Conference on Learning Representations, ICLR 2019*, 2019. 8, 19, 76, 77, 82, 85, 96, 104, 114

[87] Masahito Hayashi. General nonasymptotic and asymptotic formulas in channel resolvability and identification capacity and their application to the wiretap channel. *IEEE Transactions on Information Theory*, 52(4):1562–1575, 2006. 65, 67, 70

[88] Masahito Hayashi. Information spectrum approach to second-order coding rate in channel coding. *IEEE Transactions on Information Theory*, 55(11):4947–4966, 2009. 2, 18, 101

[89] C. Heegard and A. El Gamal. On the capacity of computer memory with defects. *IEEE Transactions on Information Theory*, 29(5):731–739, 1983. 21, 41, 102

[90] Chris Heegard and A El Gamal. On the capacity of computer memory with defects. *IEEE transactions on Information theory*, 29(5):731–739, 1983. 50, 52, 55, 61

[91] Mahmoud Hegazy, Rémi Leluc, Cheuk Ting Li, and Aymeric Dieuleveut. Compression with exact error distribution for federated learning. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 613–621. PMLR, 02–04 May 2024. 6

[92] Mahmoud Hegazy and Cheuk Ting Li. Randomized quantization with exact error distribution. In *2022 IEEE Information Theory Workshop (ITW)*, pages 350–355. IEEE, 2022. 6, 8, 96

[93] Henri Hentilä, Yanina Y Shkel, and Visa Koivunen. Communication-constrained secret key generation: Second-order bounds. *IEEE Transactions on Information Theory*, 2024. 4, 15

[94] Berivan Isik, Wei-Ning Chen, Ayfer Özgür, Tsachy Weissman, and Albert No. Exact optimality of communication-privacy-utility tradeoffs in distributed mean estimation. *Advances in Neural Information Processing Systems*, 36, 2024. 76

[95] Hyoungju Ji, Sunho Park, Jeongho Yeo, Younsun Kim, Juho Lee, and Byonghyo Shim. Ultra-reliable and low-latency communications in 5g downlink: Physical layer aspects. *IEEE Wireless Communications*, 25(3):124–130, 2018. 48, 51

[96] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2):1–210, 2021. 86, 88

[97] TON Kalker and Frans MJ Willems. Capacity bounds and constructions for reversible data-hiding. In *2002 14th International Conference on Digital Signal Processing Proceedings. DSP 2002 (Cat. No. 02TH8628)*, volume 1, pages 71–76. IEEE, 2002. 51

[98] 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. 76, 78

[99] Guy Keshet, Yossef Steinberg, Neri Merhav, et al. Channel coding in the presence of side information. *Foundations and Trends® in Communications and Information Theory*, 4(6):445–586, 2008. 50

[100] Ashish Khisti, Arash Behboodi, Gabriele Cesa, and Pratik Kumar. Unequal message protection: One-shot analysis via poisson matching lemma. In *2024 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2024. 4, 15

[101] Young-Han Kim. Coding techniques for primitive relay channels. In *Proc.*

Forty-Fifth Annual Allerton Conf. Commun., Contr. Comput, page 2007, 2007. 21, 22, 31, 34, 35, 36, 102

[102] Young-Han Kim, Arak Sutivong, and Thomas M Cover. State amplification. *IEEE Transactions on Information Theory*, 54(5):1850–1859, 2008. 51

[103] Mari Kobayashi, Yingbin Liang, Shlomo Shamai, and Mérouane Debbah. On the compound mimo broadcast channels with confidential messages. In *2009 IEEE International Symposium on Information Theory*, pages 1283–1287. IEEE, 2009. 51

[104] Szymon Kobus, Lucas Theis, and Deniz Gündüz. Gaussian channel simulation with rotated dithered quantization. In *2024 IEEE International Symposium on Information Theory (ISIT)*, pages 1907–1912. IEEE, 2024. 104

[105] Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016. 86

[106] Victoria Kostina and Sergio Verdú. Lossy joint source-channel coding in the finite blocklength regime. *IEEE Transactions on Information Theory*, 59(5):2545–2575, 2013. 2, 101

[107] Samuel Kotz, Tomasz Kozubowski, and Krzystof Podgorski. *The Laplace distribution and generalizations: a revisit with applications to communications, economics, engineering, and finance*. Springer Science & Business Media, 2012. 74

[108] Natalie Lang, Elad Sofer, Tomer Shaked, and Nir Shlezinger. Joint privacy

enhancement and quantization in federated learning. *IEEE Transactions on Signal Processing*, 71:295–310, 2023. 6

[109] Günter Last and Mathew Penrose. *Lectures on the Poisson process*, volume 7. Cambridge University Press, 2017. 15, 113, 115, 123, 131

[110] Si-Hyeon Lee and Sae-Young Chung. A unified random coding bound. *IEEE Transactions on Information Theory*, 64(10):6779–6802, 2018. ii, 9, 21, 25, 26, 30, 102, 103

[111] Eric Lei, Hamed Hassani, and Shirin Saeedi Bidokhti. Neural estimation of the rate-distortion function with applications to operational source coding. *IEEE Journal on Selected Areas in Information Theory*, 3(4):674–686, 2022. 4, 14

[112] Cheuk Ting Li. An automated theorem proving framework for information-theoretic results. *IEEE Transactions on Information Theory*, 69(11):6857–6877, 2023. 35, 40, 103

[113] Cheuk Ting Li. Pointwise redundancy in one-shot lossy compression via Poisson functional representation. In *International Zurich Seminar on Information and Communication (IZS 2024)*, 2024. 8, 16, 18

[114] Cheuk Ting Li. Pointwise redundancy in one-shot lossy compression via Poisson functional representation. *arXiv preprint*, 2024. 130, 134

[115] Cheuk Ting Li. Discrete layered entropy, conditional compression and a tighter strong functional representation lemma. *arXiv preprint arXiv:2501.13736*, 2025. 8, 16, 18

[116] Cheuk Ting Li and Venkat Anantharam. Pairwise multi-marginal opti-

mal transport and embedding for earth mover's distance. *arXiv preprint arXiv:1908.01388*, 2019. 123, 129

[117] Cheuk Ting Li and Venkat Anantharam. A unified framework for one-shot achievability via the Poisson matching lemma. *IEEE Transactions on Information Theory*, 67(5):2624–2651, 2021. i, 2, 4, 7, 8, 14, 15, 16, 17, 18, 19, 22, 23, 24, 33, 41, 42, 43, 44, 55, 57, 58, 60, 61, 65, 67, 68, 69, 70, 75, 79, 81, 101, 102, 104, 109, 113, 117, 122, 123, 129

[118] Cheuk Ting Li and Abbas El Gamal. Strong functional representation lemma and applications to coding theorems. *IEEE Transactions on Information Theory*, 64(11):6967–6978, 2018. i, 4, 7, 8, 9, 13, 14, 15, 16, 17, 18, 19, 22, 23, 57, 75, 77, 79, 81, 82, 83, 101, 103, 104, 113, 117, 123, 129, 130

[119] Cheuk Ting Li et al. Channel simulation: Theory and applications to lossy compression and differential privacy. *Foundations and Trends® in Communications and Information Theory*, 21(6):847–1106, 2024. 8, 15, 18, 23

[120] Cheuk Ting Li, Xiugang Wu, Ayfer Özgür, and Abbas El Gamal. Minimax learning for remote prediction. In *2018 IEEE ISIT*, pages 541–545, June 2018. 14

[121] Cheuk Ting Li, Xiugang Wu, Ayfer Ozgur, and Abbas El Gamal. Minimax learning for distributed inference. *IEEE Transactions on Information Theory*, 66(12):7929–7938, 2020. 4

[122] Weixiang Li, Weiming Zhang, Li Li, Hang Zhou, and Nenghai Yu. Designing near-optimal steganographic codes in practice based on polar codes. *IEEE Transactions on Communications*, 68(7):3948–3962, 2020. 50

[123] Yingbin Liang, Gerhard Kramer, and H Vincent Poor. Compound wiretap channels. *EURASIP Journal on Wireless Communications and Networking*, 2009:1–12, 2009. ii, xv, 10, 47, 48, 51, 52, 65, 66, 70, 71, 102

[124] Yingbin Liang and H Vincent Poor. Multiple-access channels with confidential messages. *IEEE Transactions on Information Theory*, 54(3):976–1002, 2008. 66

[125] H Liao. *Multiple Access Channels*. PhD thesis, Department of Electrical Engineering, University of Hawaii, Honolulu, HI, 1972. 21, 41, 44, 102

[126] Sung Hoon Lim, Young-Han Kim, Abbas El Gamal, and Sae-Young Chung. Noisy network coding. *IEEE Transactions on Information Theory*, 57(5):3132–3152, 2011. ii, 9

[127] Chih Wei Ling, Yanxiao Liu, and Cheuk Ting Li. Weighted parity-check codes for channels with state and asymmetric channels. In *2022 IEEE International Symposium on Information Theory (ISIT)*, pages 3103–3108. IEEE, 2022. 17

[128] Chih Wei Ling, Yanxiao Liu, and Cheuk Ting Li. Weighted parity-check codes for channels with state and asymmetric channels. *IEEE Transactions on Information Theory*, 2024. 17

[129] Jingbo Liu, Paul Cuff, and Sergio Verdú. One-shot mutual covering lemma and Marton's inner bound with a common message. In *2015 IEEE International Symposium on Information Theory (ISIT)*, pages 1457–1461. IEEE, 2015. 2, 4

[130] Jingbo Liu, Paul Cuff, and Sergio Verdú. e_γ-resolvability. *IEEE Transactions on Information Theory*, 63(5):2629–2658, 2016. 65, 67, 70

[131] Tie Liu, Vinod Prabhakaran, and Sriram Vishwanath. The secrecy capacity of a class of parallel gaussian compound wiretap channels. In *2008 IEEE International Symposium on Information Theory*, pages 116–120. IEEE, 2008. 52

[132] Yanxiao Liu, Wei-Ning Chen, Ayfer Özgür, and Cheuk Ting Li. Universal exact compression of differentially private mechanisms. *Advances in Neural Information Processing Systems*, 37:91492–91531, 2024. 4, 5, 11, 73

[133] Yanxiao Liu and Cheuk Ting Li. One-shot coding over general noisy networks. In *2024 IEEE International Symposium on Information Theory (ISIT)*, pages 3124–3129. IEEE, 2024. (c) 2024 IEEE. Reprinted, with permission. 4, 10, 17, 21

[134] Yanxiao Liu and Cheuk Ting Li. One-shot information hiding. In *2024 IEEE Information Theory Workshop (ITW)*, pages 169–174. IEEE, 2024. (c) 2024 IEEE. Reprinted, with permission. 10, 47, 52

[135] Chris J Maddison. A Poisson process model for Monte Carlo. *Perturbation, Optimization, and Statistics*, pages 193–232, 2016. 14, 18, 96

[136] Chris J Maddison, Daniel Tarlow, and Tom Minka. A* sampling. *Advances in neural information processing systems*, 27, 2014. 14, 18

[137] Vikrant Malik, Taylan Kargin, Victoria Kostina, and Babak Hassibi. A distributionally robust approach to shannon limits using the wasserstein distance. In *2024 IEEE International Symposium on Information Theory (ISIT)*, pages 861–866. IEEE, 2024. 49

[138] Katalin Marton. A coding theorem for the discrete memoryless broadcast channel. *IEEE Transactions on Information Theory*, 25(3):306–311, 1979. 21, 41, 45, 102

[139] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017. 86

[140] Neri Merhav. On random coding error exponents of watermarking systems. *IEEE Transactions on Information Theory*, 46(2):420–430, 2000. 50

[141] Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pages 263–275. IEEE, 2017. 135, 136

[142] Marco Mondelli, S Hamed Hassani, and Rüdiger Urbanke. A new coding paradigm for the primitive relay channel. *Algorithms*, 12(10):218, 2019. 21, 22, 31, 36, 102

[143] Pierre Moulin. Universal fingerprinting: Capacity and random-coding exponents. In *2008 IEEE International Symposium on Information Theory*, pages 220–224. IEEE, 2008. 50

[144] Pierre Moulin and Joseph A O'Sullivan. Information-theoretic analysis of information hiding. *IEEE Transactions on information theory*, 49(3):563–593, 2003. ii, xv, 10, 47, 48, 49, 50, 51, 52, 54, 55, 56, 60, 61, 62, 63, 102

[145] Pierre Moulin and Ying Wang. New results on steganographic capacity. In *Proc. CISS Conference*. Citeseer, 2004. 50

[146] Pierre Moulin and Ying Wang. Capacity and random-coding exponents for channel coding with side information. *IEEE Transactions on Information Theory*, 53(4):1326–1347, 2007. 50, 52, 55, 60, 62

[147] Buu Phan, Ashish Khisti, and Christos Louizos. Importance matching lemma for lossy compression with side information. In *International Conference on Artificial Intelligence and Statistics*, pages 1387–1395. PMLR, 2024. 9, 18, 19

[148] Yury Polyanskiy. On dispersion of compound dmcs. In *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 26–32. IEEE, 2013. 52

[149] Yury Polyanskiy, H Vincent Poor, and Sergio Verdú. Channel coding rate in the finite blocklength regime. *IEEE Trans. Inf. Theory*, 56(5):2307–2359, 2010. 2, 3, 101

[150] John K Salmon, Mark A Moraes, Ron O Dror, and David E Shaw. Parallel random numbers: as easy as 1, 2, 3. In *Proceedings of 2011 international conference for high performance computing, networking, storage and analysis*, pages 1–12, 2011. 117

[151] Jonathan Scarlett. On the dispersions of the Gel' fand–Pinsker channel and dirty paper coding. *IEEE Transactions on Information Theory*, 61(9):4569–4586, 2015. 42, 61

[152] Rafael F. Schaefer and Sergey Loyka. The secrecy capacity of compound gaussian mimo wiretap channels. *IEEE Transactions on Information Theory*, 61(10):5535–5552, 2015. 52, 65, 66

[153] Abhin Shah, Wei-Ning Chen, Johannes Balle, Peter Kairouz, and Lucas Theis. Optimal compression of locally differentially private mechanisms. In *International Conference on Artificial Intelligence and Statistics*, pages 7680–7723. PMLR, 2022. 6, 8, 9, 19, 74, 76, 77, 82, 84, 85, 89, 104, 114

[154] Ali Moradi Shahmiri, Chih Wei Ling, and Cheuk Ting Li. Communication-efficient laplace mechanism for differential privacy via random quantization. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4550–4554. IEEE, 2024. 6, 8, 84, 96

[155] Claude E Shannon. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423, 1948. 1, 2, 83, 91

[156] Claude E Shannon. Certain results in coding theory for noisy channels. *Information and control*, 1(1):6–25, 1957. 2, 3, 18

[157] Anelia Somekh-Baruch and Neri Merhav. On the error exponent and capacity games of private watermarking systems. *IEEE Transactions on Information Theory*, 49(3):537–562, 2003. 48, 49, 50, 52, 55

[158] Anelia Somekh-Baruch and Neri Merhav. On the capacity game of public watermarking systems. *IEEE Transactions on Information Theory*, 50(3):511–524, 2004. xv, 48, 49, 50, 55, 60, 61, 63

[159] Eva C Song, Paul Cuff, and H Vincent Poor. The likelihood encoder for lossy compression. *IEEE Trans. Inf. Theory*, 62(4):1836–1849, 2016. 2, 4, 18, 19, 76, 82, 101, 114

[160] Yossef Steinberg. Reversible information embedding with compressed host at the decoder. In *2006 IEEE International Symposium on Information Theory*, pages 188–191. IEEE, 2006. 51

[161] Yossef Steinberg. Coding for channels with rate-limited side information at the decoder, with applications. *IEEE transactions on information theory*, 54(9):4283–4295, 2008. 50

[162] Orna Sumszyk and Yossef Steinberg. Information embedding with reversible stegotext. In *2009 IEEE International Symposium on Information Theory*, pages 2728–2732. IEEE, 2009. 51, 62

[163] Ananda Theertha Suresh, X Yu Felix, Sanjiv Kumar, and H Brendan McMahan. Distributed mean estimation with limited communication. In *International conference on machine learning*, pages 3329–3337. PMLR, 2017. 76, 83, 86, 89

[164] Arak Sutivong, Mung Chiang, Thomas M Cover, and Young-Han Kim. Channel capacity and state estimation for state-dependent gaussian channels. *IEEE Transactions on Information Theory*, 51(4):1486–1495, 2005. 51

[165] Vincent YF Tan and Oliver Kosut. On the dispersions of three network information theory problems. *IEEE Transactions on Information Theory*, 60(2):881–903, 2013. 2, 101

[166] Lucas Theis and Noureldin Y Ahmed. Algorithms for the communication of samples. In *International Conference on Machine Learning*, pages 21308–21328. PMLR, 2022. 77

[167] Lucas Theis, Tim Salimans, Matthew D Hoffman, and Fabian Mentzer. Lossy compression with Gaussian diffusion. *arXiv preprint arXiv:2206.08889*, 2022. 77

[168] Aleksei Triastcyn, Matthias Reisser, and Christos Louizos. DP-REC: Private & communication-efficient federated learning. *arXiv preprint arXiv:2111.05454*, 2021. 6, 8, 9, 19, 74, 76, 89

[169] S.-Y. Tung. *Multiterminal source coding*. PhD thesis, School of Electrical Engineering, Cornell University, Ithaca, NY, 1978. 37

[170] Ayşe Ünsal and Melek Önen. Information-theoretic approaches to differential privacy. *ACM Computing Surveys*, 56(3):1–18, 2023. 7

[171] Edward C Van Der Meulen. Three-terminal communication channels. *Advances in applied Probability*, 3(1):120–154, 1971. 31

[172] Sergio Verdú. Non-asymptotic achievability bounds in multiuser information theory. In *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1–8. IEEE, 2012. 2, 4, 18, 22, 42, 43, 44, 61, 101

[173] Sergio Verdú and Te Sun Han. A general formula for channel capacity. *IEEE Trans. Inf. Theory*, 40(4):1147–1157, 1994. 3

[174] D. Wang, A. Ingber, and Y. Kochman. The dispersion of joint source-channel coding. In *2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 180–187, Sep. 2011. 2

[175] Ying Wang and Pierre Moulin. Perfectly secure steganography: Capacity, error exponents, and code constructions. *IEEE Transactions on Information Theory*, 54(6):2706–2722, 2008. 50

[176] Stanley L Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965. 76

[177] S. Watanabe, S. Kuzuoka, and V. Y. F. Tan. Nonasymptotic and second-order achievability bounds for coding with side-information. *IEEE Trans. Inf. Theory*, 61(4):1574–1605, April 2015. 61, 101

[178] Shun Watanabe, Shigeaki Kuzuoka, and Vincent YF Tan. Nonasymptotic and second-order achievability bounds for coding with side-information.

IEEE Transactions on Information Theory, 61(4):1574–1605, 2015. 2, 18, 22, 42, 43

[179] J Wolfowitz. *Simultaneous Channels*. New York: Springer-Verlag, 1980. 52, 61

[180] Aaron Wyner and Jacob Ziv. The rate-distortion function for source coding with side information at the decoder. *IEEE Transactions on information Theory*, 22(1):1–10, 1976. 21, 41, 42, 102

[181] Aaron D Wyner. The wire-tap channel. *Bell system technical journal*, 54(8):1355–1387, 1975. 48, 51

[182] Aaron D Wyner. The rate-distortion function for source coding with side information at the decoder-ii. general sources. *Information and control*, 38(1):60–80, 1978. 21, 41, 42, 102

[183] Yinfei Xu, Jian Lu, Xuan Guang, and Wei Xu. Information embedding with stegotext reconstruction. *IEEE Transactions on Information Forensics and Security*, 19:1415–1428, 2023. 51, 52, 54, 62

[184] Hirosuke Yamamoto. Wyner-ziv theory for a general function of the correlated sources (corresp.). *IEEE Transactions on Information Theory*, 28(5):803–807, 1982. 21, 41, 43, 102

[185] Guangfeng Yan, Tan Li, Tian Lan, Kui Wu, and Linqi Song. Layered randomized quantization for communication-efficient and privacy-preserving distributed learning. *arXiv preprint arXiv:2312.07060*, 2023. 6, 8, 104

[186] Wei Yang, Rafael F Schaefer, and H Vincent Poor. Wiretap channels: Nonasymptotic fundamental limits. *IEEE Transactions on Information Theory*, 65(7):4069–4093, 2019. 65

[187] Mohammad Hossein Yassaee. One-shot achievability via fidelity. In *Proc. IEEE Int. Symp. Inf. Theory*, pages 301–305. IEEE, 2015. 18, 65, 67, 70

[188] Mohammad Hossein Yassaee, Mohammad Reza Aref, and Amin Gohari. Non-asymptotic output statistics of random binning and its applications. In *2013 IEEE International Symposium on Information Theory*, pages 1849–1853. IEEE, 2013. 2, 4, 18

[189] Mohammad Hossein Yassaee, Mohammad Reza Aref, and Amin Gohari. A technique for deriving one-shot achievability results in network information theory. In *2013 IEEE International Symposium on Information Theory*, pages 1287–1291. IEEE, 2013. 2, 4, 17, 18, 22, 42, 61, 101

[190] Ram Zamir and Meir Feder. On universal quantization by randomized uniform/lattice quantizers. *IEEE Transactions on Information Theory*, 38(2):428–436, 2002. 8

[191] Jacob Ziv. On universal quantization. *IEEE Transactions on Information Theory*, 31(3):344–347, 1985. 8, 77