EPFL

Thèse n° 9869

On Speed and Advantage : Results in Information Velocity and Monitoring Problems

Présentée le 21 mai 2024

Faculté informatique et communications
Laboratoire de théorie de l'information
Programme doctoral en informatique et communications

pour l'obtention du grade de Docteur ès Sciences

par

Reka INOVAN

Acceptée sur proposition du jury

Prof. N. Macris, président du jury
Prof. E. Telatar, directeur de thèse
Prof. A. Khina, rapporteur
Prof. A. Lapidoth, rapporteur
Prof. M. Gastpar, rapporteur

École polytechnique fédérale de Lausanne

2024

Abstract

Information theory has allowed us to determine the fundamental limit of various communication and algorithmic problems, e.g., the channel coding problem, the compression problem, and the hypothesis testing problem. In this work, we revisit the assumptions underlying two of the classical information theoretic problems: the channel coding problem and the hypothesis testing problem.

In the first part, we study the information velocity problem. If the channel coding problem answers the question of how much information we can send per time unit, the information velocity problem tackles the question of the latency of communicating said information on a communication network composed of relays. In the literature, this problem is commonly studied in the regime of finite message size but with a growing number of relays. In this work, we consider an asymptotic regime where we let the message size to grow to infinity. We present a converse result and two achievability results: one for Binary Erasure Channels (BEC) and one for Additive White Gaussian Noise (AWGN) channels with feedback. The converse result is obtained by extending the argument given in (Rajagopalan and Schulman, 1994) using the tools of F-divergences. The achievability results that we obtain are based on two different ideas. In the achievability result for BEC, we exploited the property of tree codes which ensure that all message bits can eventually be correctly decoded after a certain time delay. We use this property to build a tape abstraction which allows for the streaming of message bits through the relay chain. For AWGN channels, we modify the Schalkwijk-Kailath scheme to allow each relay to focus on locally transmitting its estimate of the message bits to its neighboring relay. We analyze the local behavior of this scheme, and show that we can prove results about the information velocity of the whole network based on these local results.

In the second part we study the monitoring problem. This problem capture a scenario where several regular data-generating processes maximize their own reward, with one adversarial data-generating process hiding among these regular processes and privy to certain private information. This model introduces an interesting trade-off where the adversarial data-generating process aims to exploit its private information without deviating too much from the regular data-generating processes. As by increasing its deviation, it also becomes more distinguishable from the regular data generating processes. We will analyze this problem using tools from information theory and characterize the extent of the advantage that can be obtained by the adversarial data generation process. In doing so, we showed that classification problems, which are commonly modeled as hypothesis testing problems, become more

complex when an adversarial data-generating process can adapt to the tester's protocol.

Keywords : Information Velocity, Tree Codes, Feedback, Schalkwijk-Kailath Scheme, Channel Codes, Monitoring Problem, Hypothesis Testing, Type Method

Résumé

La théorie de l'information nous a permis de déterminer la limite fondamentale de diverses tâches de communication et d'algorithmiques, telles que le problème de codage de canal, le problème de codage de sources et le problème de tests d'hypothèse. Dans cette thèse, nous revisitons les hypothèses sous-jacentes à deux problèmes classiques de la théorie de l'information : le problème de codage de canal et le problème de tests d'hypothèse. Ce faisant, nous explorons des modèles qui remettent en question ces hypothèses fondamentales.

Dans la première partie, nous étudions le problème de la vitesse d'information (*information velocity* en anglais). Si le problème de codage de canal répond à la question de la quantité d'information que nous pouvons envoyer par unité de temps, le problème de la vitesse d'information aborde la question de la latence de communication de cette information sur un réseau de communication composé de relais. Dans la littérature, ce problème est couramment étudié dans le régime de taille de message finie mais avec un nombre croissant de relais. Dans ce travail, nous considérons un régime asymptotique où nous laissons la taille du message croître à l'infini. Nous présentons un résultat contradictoire et deux résultats de réalisabilité : l'un pour les canaux binaire d'effacement (BEC en anglais) et l'autre pour les canaux de bruit additif blanc gaussien (AWGN en anglais) avec retour d'informations. Le résultat contradictoire est obtenu en étendant l'argument donné dans (Rajagopalan and Schulman, 1994) en utilisant les outils des F-divergences. Les résultats de réalisabilité que nous obtenons sont basés sur deux idées différentes. Dans le résultat de réalisabilité pour le BEC, nous avons exploité la propriété des codes de l'arbre (*tree codes* en anglais) qui garantissent que tous les bits du message peuvent finalement être décodés après un certain délai. Nous utilisons cette propriété pour construire une abstraction de bande qui permet le transfert en continu des bits de message à travers la chaîne de relais. Pour les canaux AWGN avec retour d'informations, nous modifions le schéma de Schalkwijk-Kailath qui permet à chaque relais de se concentrer sur la transmission locale de son estimation des bits de message à son relais voisin. Nous analysons le comportement local de ce schéma et montrons que nous pouvons prouver des résultats sur la vitesse d'information de la totalité du réseau basés sur ces résultats locaux.

Dans la deuxième partie, nous étudions le problème de surveillance. Ce problème modélise un scénario où plusieurs processus de génération de données réguliers opèrent pour maximiser leur propre récompense, avec un processus de génération de données adverses se cachant parmi ces processus réguliers et ayant accès à certaines informations privées. Ce modèle introduit un compromis intéressant où le processus de génération de données adverses cherche

à exploiter ses informations privées sans s'écarter trop des processus règuliers, risquant ainsi d'être facilement différenciable des processus réguliers. Nous analyserons ce scénario à l'aide d'outils de la théorie de l'information et caractériserons l'étendue de l'avantage pouvant être obtenu par le processus de génération de données adverses. Ce faisant, nous avons montré que les problèmes de classification, souvent modélisés comme des problèmes de tests d'hypothèse, deviennent plus complexes lorsque le processus de génération de données adverses peut s'adapter au protocole du testeur.

Mots-clés : Vitesse d'Information, Codes de l'Arbre, Retour d'Informations, Schéma de Schalkwijk-Kailath, Codage de Canaux, Problème de Surveillance, Problèmes de Tests d'Hypothèse

Acknowledgements

I consider myself incredibly fortunate to have Emre as my advisor. When I first come to EPFL, I heard many rumors about his kindness, breadth of knowledge, and generosity. Now, after half a decade being his student, I have enough evidence to assure everyone that not only those rumors are true, but there is an argument to be made that the rumors are severely underselling it.

As his PhD student, I benefited immensely from Emre's encyclopaedic knowledge. I fondly remember the hours that we spent at the whiteboard in his office, sometimes until very late, attempting to prove some conjectures. In those moments I learned from his examples the value of being creative and tenacious when faced with seemingly impenetrable problems. Emre also provides me with encouragement to explore and pursue my research interest, and to be comfortable in indulging my intellectual curiosities.

Not to mention that during the later years of my PhD, where things become a bit tumultuous for me and my family, he provided generous supports and a source of constancy. I always know that I can talk and confide to him.

They say that imitation is the biggest form of admiration. Perhaps there is no better way to express my admiration than to say that he is the model of a researcher and a person that I aspire to be; maybe with much less fascination with wines though.

This thesis also benefited from the questions and suggestions from my thesis committee. One of my regret is not knowing Anatoly and Amos earlier, but even our brief discussions greatly improves this thesis and equips me with a lot of interesting directions to pursue later. Nicolas and Michael also plays a pivotal roles in my academic career. It is Nicolas that first told me that I should consider doing a PhD during the oral exam of his Quantum Information Theory class. Whilst one of my earliest experience of doing research is done under Michael's supervision when I was a project student and a master thesis student in his lab.

IPG has been my home for the last few years. My stay has been greatly enriched by the opportunities to discuss and to attend lectures from IPG's professors, scientist and post-doc. Also, a thanks to Muriel which somehow always been able to find a solution to any administrative matters, and to Damir, the resident computer expert.

I would like to express my heartfelt gratitude to all my wonderful friends at IPG and EDIC, whose camaraderie has enriched my Ph.D. journey with unforgettable memories. I fondly recall the Christmas dinner and the delightful trip to Montreux Christmas Market with Eric

and Ying;When Erixhen and Arda took me on a surprise trip in search of the best kebab in Lausanne; Or when I was walking from Flon to Malley with Sephand “to help us digest our dinner”; When I crashed on Praneeth’s and Clement’s couch after a new year party; When Amaedeo taught me his tricks to break probability dependence using Cauchy-Schwarz; When I was borrowing a bladeRF from Nicolae; Or when it is the middle of the pandemic and I found Pierre standing with a whiteboard in front of my appartment; Or the bebefoot match, when Kirill, Dasha, and I faced off against the INR concierge guy in a 3 vs. 1 showdown (and still losing badly!). Also, thank you to the participants of LTHI lab meetings, Aditya, Bora, 2nd Emre (Serhat) and Millen for always bringing the weekly basket of interesting questions; And also to the other members of IPG – Aditya, Anand, Anastasia, Andreas, Antoine, Dina, Thomas, Marco, Bora, Anuj, Cemre, Yunzhen – for our daily interactions and the IPG board game nights. Special thanks to my office mate Yunus. Few years ago, during a lunch in the middle of January, he said that we should take a course in stochastic calculus from the school of financial engineering together. This offhand idea seems to be the flap of the butterfly wings that leads to my next step after my PhD.

A quick shout out to my friends in Indonesia – Dzuhri, Umam, Tiko, Andreas, and Sudiro – for being dependable friends and sources of TV show recommendations.

It’s only fair to admit that this thesis wouldn’t exist without the effort of my incredible parents, Kusnoto and Sunarmi, in giving me an opportunity to pursue my education. After all, writing a thesis becomes a bit tricky if you haven’t quite mastered the art of reading and writing, or if you’re not born in the first place. Also my gratitude to my brothers Cendekia and Digit, for helping me spending the long winter nights with Dungeon and Dragon sessions.

In my house, it is customary to give the last piece of a dish to the one we love the most. So here we are at the last part of my acknowledgement, I want to thank my beloved wife Nia for her unending support and encouragement. Thank you for simply existing and being here with me.

Lausanne, March 22, 2024 Reka Inovan

Contents

1 Introduction and Preliminaries

It is not an overstatement to say that information theory is one of the most important mathematical theories in the 21st century. What began as an attempt to understand the limitations and quantify the performance of the newly developed telephony systems in 1948 (Shannon, 1948) has since found applications in diverse fields such as machine learning, cryptography, statistics, etc.

Among these applications, digital communications can be considered as the poster child of this theory. In fact, the story of digital communications can be told as an extended journey to achieve the performance that Shannon promised to be possible in 1948. Of course, as we know, Shannon actually provides a constructive method on how to design communication systems which achieve the fundamental limit he derived. If not for the very tiny problem that a) the maximum-likelihood decoding process he proposed is exactly as hard[1] as solving the traveling salesman problem or determining the 3-SAT problem[2] (Berlekamp et al., 1978), and b) his method only exhibits optimality in the regime where the size of the data grows to infinity.

Since then, generations of researchers have attempted to find the holy grail for problem a), namely, a cleverly constructed mathematical object that achieves the same performance as the random codes that Shannon proposed while requiring much less computation to decode. From these continuous effort, we have discovered many interesting mathematical objects such as Hamming codes (Hamming, 1950), Golay codes (Golay, 1949), Reed-Solomon codes (Reed and Solomon, 1960), Reed-Muller codes (Reed, 1954), BCH codes (Bose and Ray-Chaudhuri, 1960), convolutional codes (Massey et al., 1973), etc., which nowadays can be classified as "classical" channel codes. Particularly interesting developments are the so-called modern channel codes, with two of the most representative examples being the Spatially-Coupled Low-Density Parity Check codes (Jimenez Felstrom and Zigangirov, 1999) (Gallager, 1962) and the Polar codes (Arikan, 2009); both of which have recently been shown to achieve the fundamental limit of performance that Shannon has derived.

[1] More accurately, it has been shown that the problem of decoding general linear codes is NP-complete

[2] Or finding the best way to flip a stack of pancakes, if one prefers a more down-to-earth comparison.

Another interesting line of research is based on developing methods to directly tackle problem b) by refining the fundamental limits in the regime where the number of data is more "reasonable", with most of these results in this area found under the heading of "finite-blocklength" information theory (Polyanskiy et al., 2010). A rather surprising result shows that the existing channel codes are actually already very close to optimal.

Given these results, it is natural to have a sense that the problem of channel coding as a mathematical problem is largely "solved". However, channel coding as a practical problem is still rich in unsolved problems.

One can argue that one of the main reasons why the channel coding problem is practically important is because it allows us to characterize the possible performances of real-world communication systems. The blocklength in the channel coding problem is directly proportional to the number of "independence" that is consumed by a practical communication system (whether this "independence" is achieved by transmitting through different frequency sub-bands or different time slices). In this work, we will focus on using the blocklength as a proxy of time. Hence, the current results in channel coding problems essentially give us a very accurate characterization of the time needed to communicate a certain number of bits.

But seen from this light, the channel coding problem only characterizes one aspect of time in communication problems, namely, the long-time ratio between the number of bits and the required time to communicate the said bits. Another aspect that has recently received more attention is the notion of delay[3]. There are several approaches that have been proposed to address this notion of delay, either through designing channel codes with finite delay (Guo and Kostina, 2023), or through defining a model to account for the cost of the delay of information update (Kaul et al., 2012).

In this thesis, we will focus on one of the approaches to study and quantify the delays in a network setting, namely the approach of "information velocity" where we study the delay of propagating information through a network in terms of the ratio between the length of transmission time vs the number of relays which intermediated the transmission. We will present several results on this approach in the first part of this thesis.

In the second part of this thesis, we will consider a particular model of hypothesis testing problem. Information-theoretic methods have a long history of being associated with the analysis of hypothesis testing problems. There has been a very fruitful line of research using the hypothesis testing problem to prove converse results for communication problems, e.g., (Blahut, 1974). More recently, the hypothesis testing framework is also being used to model classification problems in machine learning, especially in the context of distributed learning,

[3] In fact, the author was exposed to this delay vs rate distinction while taking the Distributed System Engineering class. It turns out that most practical communication protocols seem to choose a combination of a simple forward-error correction and an ARQ scheme to achieve reliable communication. Simple forward error correction schemes generally require much shorter blocklength to achieve reasonable performance compared to the more complex capacity achieving schemes. Combined with ARQ scheme to account for hard-to-correct error events, this combination performs surprisingly well in practices.

for example (Kayaalp et al., 2023), (Lalitha et al., 2018).

An important assumption in the classical hypothesis testing problem is that the data-generating process which we tested is not influenced by the hypothesis testing scheme itself. This assumption is certainly reasonable if the hypothesis that we want to study is induced by a natural phenomenon. However, this assumption might not hold if the data-generating process is incentivized to increase the error rate, for example, if we assume that the hypothesis testing problem is conducted to monitor for suspicious communication in an attempt to detect for intrusion in a network. There has been several research in this direction, for example to determine the optimal error exponent of classification error in (Yasodharan and Loiseau, 2019), or to design an optimal classifier and adversary, for example in (Jin and Lai, 2021). In the second part of the thesis, we propose a model that can capture this dynamic while being amenable to information-theoretic analysis.

But before we present these results, we will review several classical information theory results in channel coding and hypothesis testing, for the purpose of understanding the context and motivation for our result, while introducing mathematical notations that we will use in the later chapters.

1.1 Hypothesis Testing

The simplest form of hypothesis testing is binary hypothesis testing. In this problem, a tester is given an observation Y and needs to determine whether this observation is a realization of a data-generating process in $\mathscr{D}_0$ or $\mathscr{D}_1$, i.e., $P_Y \in \mathscr{D}_0$ or $P_Y \in \mathscr{D}_1$.

For example, consider an environmental scientist who wants to determine whether Lausanne's autumn temperature in 2023 is warmer than the temperature in 2022. In this case, the statistician has two hypotheses,

- H_0: Lausanne's autumn temperature in 2023 is not higher than the autumn temperature in 2022.
- H_1: Lausanne's autumn temperature in 2023 is higher than the autumn temperature in 2022.

The scientist will base his conclusion on the average autumn 2023 temperature Y_{2023}. To determine the test function, the scientist needs to specify the data-generating process that corresponds to their hypotheses. As this thesis is not a treatise in environmental science, let us simply assume that the data-generating processes are Gaussian distributions with different means, i.e., $\mathscr{D}_0 = \{N(\mu,\sigma^2) : \mu \leq y_{2022}\}$ and $\mathscr{D}_1 = \{N(\mu,\sigma^2) : \mu > y_{2022}\}$, where $N(\mu,\sigma)$ is the Gaussian distribution with mean μ and variance σ^2.

Let us model the decision process of the scientist as the test function $T(\cdot)$. We will consider a test function where the scientist simply compares the observed average temperature in 2023

against a threshold t, i.e., $T(x) = \mathbb{1}\{x > t\}$. Based on this test function, we can determine the set where the scientist decides 1 as $D_1 = \{x : x > t\}$ and where they decide 0 as $D_0 = D_1^c$. Now the question is how to quantify the performance of this simplistic decision rule that the scientist proposes. Common performance metrics in this framework are the probability of two error events[4], namely,

1. Type-I error, the probability that the scientist decides H_1 even though the data-generating process is actually in $\mathscr{D}_0$, i.e., $\max_{P_Y \in \mathscr{D}_0} P_Y(T(Y) = 1)$.

2. Type-II error, the probability that the scientist decides H_0 if the data-generating process is actually in $\mathscr{D}_1$, i.e., $\max_{P_Y \in \mathscr{D}_1} P_Y(T(Y) = 0)$.

In general, each choice of test function will generate a different trade-off in terms of Type-I and Type-II errors. For example, in this case, we can see that for this example problem, we have

$$\text{Type-I Error } + \text{ Type-II Error } = 1. \tag{1.1}$$

Although this relation does not hold for general hypothesis testing problems.

Given a family of test functions, it is reasonable to ask whether this family is "optimal", in the sense that given any test function with Type-I error and Type-II error $(\epsilon_I, \epsilon_{II})$, there exists a test function in this family with Type-I error and Type-II error $(\epsilon'_I, \epsilon'_{II})$ such that $\epsilon'_I \leq \epsilon_I$ or $\epsilon'_{II} \leq \epsilon_{II}$. The family of test functions which fulfills this criterion is commonly referred as the universally most powerful (UMP) tests.

One of the earliest results in this area is the existence of UMP tests for simple binary hypothesis testing, i.e., if $\mathscr{D}_0$ and $\mathscr{D}_1$ are singleton sets(Neyman et al., 1933)[5].

Theorem 1.1 (Neyman-Pearson)**.** *Consider a binary hypothesis testing problem with observation Y, and simple hypotheses such that, $\mathscr{D}_0 = \{P_Y\}$ and $\mathscr{D}_1 = \{Q_Y\}$. We define the likelihood-ratio test as,*

$$T_t(y) = \mathbb{1}\left\{\log \frac{dQ_Y}{dP}(y) > t\right\}, \tag{1.2}$$

and the randomized likelihood-ratio test as the

$$T(y) = T_{t_R}(y) \tag{1.3}$$

where (t_1, t_2) is a pair of threshold values and R is an integer random variable in $\{1,2\}$. The family of randomized likelihood-ratio tests is UMP tests for binary hypothesis testing problem with simple hypotheses.

[4] Indeed, this is the Frequentist approach to hypothesis testing. But the distinction between Bayesian and Frequentist approach is irrelevant for this thesis.

[5] Although back then they referred to the test family which fulfills our optimality condition as "the most efficient" test as opposed to UMP tests terminology that we use in this work.

Interestingly, a modification to the randomized likelihood-ratio tests can also be shown to be UMP for a certain subclass of binary hypothesis testing problem(Karlin and Rubin, 1956).

Theorem 1.2 (Karlin-Rubin)**.** *Consider a binary hypothesis testing problem where the data-generating process can be parameterized by a real parameter θ i.e., P_Y^θ such that $\mathscr{D}_0 = \{P_Y^{(\theta)} : \theta \leq t\}$ and $\mathscr{D}_1 = \{P_Y^{(\theta)} : \theta > t\}$. Furthermore, let us assume that there exists a statistics of Y, i.e., $g(y)$, such that we have*

$$f(y) = \log \frac{dP_Y^{(\theta_2)}}{dP_Y^{(\theta_1)}}\big(g(y)\big) \tag{1.4}$$

is non-decreasing on $g(y)$ for every $\theta_1 \leq \theta_2$. Then the family of test functions $T(y) = \mathbb{1}\{g(y) > t\}$ is UMP tests for this binary hypothesis testing problem.

This theorem lends validity to the practice of taking one representative distribution from $\mathscr{D}_0$ and $\mathscr{D}_1$, and then performing hypothesis testing using likelihood-ratio tests as if $\mathscr{D}_0$ and $\mathscr{D}_1$ are singletons each containing their representative distribution. If we assume that the data-generating process fulfills the condition of Karlin-Rubin theorem then these seemingly ad-hoc tests are also UMP. Given this theorem, we can see that the scientist's decision in the running example is already optimal.

In practice, there might be a reason to prefer sub-optimal hypothesis testing rules if the likelihood ratio tests are computationally intractable, for example in (Newey and West, 1987). Nevertheless, the likelihood-ratio tests play an important role in the theoretical treatment of hypothesis testing problems, simply by virtue of its theoretical amenability.

The setting in which likelihood-ratio tests particularly shine is if the observation is multivariate, i.e., $Y[1:n]$,[6] and the data-generating process in $\mathscr{D}_0$ and $\mathscr{D}_1$ are independent and identical distributions, i.e., for every $P_{Y[1:n]} \in \mathscr{D}_0 \bigcup \mathscr{D}_1$, $P_{Y[1:n]} = \prod_{i=1}^n P_{Y[i]}$ and for all i we have $P_{Y[i]} = P_Y[1]$.

In the case of a simple binary hypothesis testing problem with $\mathscr{D}_0 = \{P_Y\}$ and $\mathscr{D}_1 = \{Q_Y\}$, we have the log likelihood-ratio becomes,

$$\mathrm{LLR}(Y[1:n]) = \sum_{k=1}^{n} \log \frac{dQ_Y}{dP_Y}(Y[k]). \tag{1.5}$$

This log likelihood-ratio is interesting as it is a sum of i.i.d. random variables.Thus, provided that the required moments exist, it is subject to the Law of Large Numbers and Central Limit Theorem. This observation motivates the definition of KL divergence(Kullback, 1959)[7].

Definition 1.1. *Consider two distributions P and Q, we define KL divergence as,*

$$D(Q;P) = \int \log \frac{dQ}{dP} dQ. \tag{1.6}$$

[6] In this work, we will use $Y[1:n]$ to denote a vector of $(Y[1],\ldots,Y[n])$.

[7] Curiously enough Kullback refers to this quantity as discrimination-information in his book.

Furthermore, we have that

$$\mathbb{E}_Q\left[\mathrm{LLR}(Y[1:n])\right] = nD(Q_Y;P_Y) \qquad \mathbb{E}_P\left[\mathrm{LLR}(Y[1:n])\right] = -nD(P_Y;Q_Y). \tag{1.7}$$

As the log likelihood-ratio factorizes neatly into a sum, astute readers might have noticed that likelihood-ratio tests do not require precise information about the observations. It is sufficient to record the number of occurrences of each value in the alphabet. For example, let observation $Y[1:5]$ takes values in $\{0,1\}^5$, then the log likelihood-ratio of $y[1:5]=[0,0,0,1,1]$ is exactly equivalent as the log likelihood-ratio of $y[1:5]=[0,1,0,1,0]$. This insight motivates the definition of the type of a sequence, namely the empirical distribution of that sequence. This definition is the foundation of the type method, readers can find a comprehensive treatment of this method in Csiszár and Körner (2011).

Definition 1.2. *Given a sequence of realizations from a discrete random variable, $y[1:n]$, we define the type of this sequence as a function,*

$$P_{y[1:n]}(y') = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{y[i]=y'\}. \tag{1.8}$$

We denote the set of all possible types of length n over an alphabet $\mathscr{X}$ can take as $\mathscr{P}_n(\mathscr{X})$.

We also denote the set of all sequences of length n with type P_X as $\mathscr{T}^n_{P_Y}$.

Given a type Q, we define the neighborhood of this type as

$$[Q]_{\delta,n} = \left\{P \in \mathscr{P}_n(\mathscr{X}) : |Q-P|_1 \le \delta\right\} \tag{1.9}$$

and correspondingly,

$$\mathscr{T}_{[Q]_{\delta,n}} = \bigcup_{P\in[Q]_{\delta,n}} \mathscr{T}^n_P. \tag{1.10}$$

Let us recall several basic properties of the types of sequences.

Proposition 1.1. *We have,*

1. *The size of possible types is polynomial in n, i.e., $|\mathscr{P}_n(\mathscr{X})| \le (n+1)^{|\mathscr{X}|}$,*
2. *Given a type $P_X \in \mathscr{P}_n(\mathscr{X})$, then we have for every distribution Q_X,*

$$Q_X(\mathscr{T}^n_{P_X}) \le e^{-nD(P_X;Q_X)}, \tag{1.11}$$

3. *For any discrete probability distribution P_X, we have,*

$$\lim_{n\to\infty} P_X(\mathscr{T}^n_{[P_X]_{\delta_n,n}}) = 1 \tag{1.12}$$

if $\lim_{n\to\infty} n\delta_n^2 = \infty$.

Proof. 1. This is a simple application of counting with stars and bars (Feller, 1968).

2. We have for all $x[1:n] \in \mathcal{T}_{P_X}$,

$$P_{X^n}(x[1:n]) = \exp\left(n\mathbb{E}_{P_X}\left[\log P_X(X)\right]\right) \tag{1.13}$$

simply from the definition.

This gives us

$$|\mathcal{T}^n_{P_X}| \exp\left(n\mathbb{E}_{P_X}\left[\log P_X(X)\right]\right) \leq \sum_{P'_X \in \mathcal{P}_n(\mathcal{X})} |\mathcal{T}^n_{P'_X}| \exp\left(n\mathbb{E}_{P'_X}\left[\log P'_X(X)\right]\right) = 1 \tag{1.14}$$

which implies that,

$$|\mathcal{T}^n_{P_X}| \leq \exp\left(-n\mathbb{E}_{P_X}\left[\log P_X(X)\right]\right). \tag{1.15}$$

This leads to,

$$\begin{aligned} Q_X(\mathcal{T}^n_{P_X}) &= \sum_{x[1:n] \in \mathcal{T}^n_{P_X}} Q_X(x[1:n]) \\ &= |\mathcal{T}^n_{P_X}| e^{n\mathbb{E}_{P_X}[\log Q_X(X)]} \\ &\leq e^{-nD(P_X;Q_X)}. \end{aligned} \tag{1.16}$$

3. Note that,

$$\forall_{x' \in \mathcal{X}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{x' = x[i]\} - P_X(x') \right| \leq \frac{\delta_n}{|\mathcal{X}|} \tag{1.17}$$

implies,

$$x[1:n] \in \mathcal{T}_{[P_X]_{\delta_n,n}}. \tag{1.18}$$

Hence we have,

$$\Pr\left(\bigcap_{x' \in \mathcal{X}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{x' = x[i]\} - P_X(x') \right| \leq \frac{\delta_n}{|\mathcal{X}|} \right) \leq \Pr\left(\mathcal{T}_{[P_X]_{\delta_n,n}}\right) \tag{1.19}$$

Let us consider the complement of the left-hand side; by the union bound, we have

$$\Pr\left(\bigcup_{x' \in \mathcal{X}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{x' = x[i]\} - P_X(x') \right| > \frac{\delta_n}{|\mathcal{X}|} \right) \leq \sum_{x' \in \mathcal{X}} \Pr\left(\left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{x' = x[i]\} - P_X(x') \right| > \frac{\delta_n}{|\mathcal{X}|} \right). \tag{1.20}$$

Note that the indicators are Bernoulli random variables. We use Chebyshev's inequality (see appendix) and the fact that the variance of Bernoulli random variables is upper bounded by $1/4$ to obtain

$$\Pr\left(\bigcup_{x' \in \mathcal{X}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{x' = x[i]\} - P_X(x') \right| > \frac{\delta_n}{|\mathcal{X}|} \right) \leq \frac{|\mathcal{X}|^3}{4n\delta_n^2}. \tag{1.21}$$

Hence we have,

$$\Pr\left(\mathscr{T}_{[P_X]_{\delta_n,n}}\right) \geq 1 - \Pr\left(\bigcup_{x' \in \mathscr{X}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{x' = x[i]\} - P_X(x') \right| > \frac{\delta_n}{|\mathscr{X}|} \right) \geq 1 - \frac{|\mathscr{X}|^3}{4n\delta_n^2} \tag{1.22}$$

which converges to 1 under the condition on δ_n. □

We can directly apply this result to study the error probability of a hypothesis testing problem. The following theorem gives us the asymptotic rate of decay for the Type-II error of any hypothesis testing scheme if we require that the Type-I error is smaller than a certain value.

Theorem 1.3. *Let us define the best Type-II error for a given binary hypothesis testing problem between P_{X^n} and Q_{X^n} while ensuring that the Type-I error is less than ϵ as,*

$$\beta(P_{X^n}, Q_{X^n}, \epsilon) = \min_{D_0 : P_{X^n}(D_0) \geq 1-\epsilon} Q_{X^n}(D_0). \tag{1.23}$$

We have for $\epsilon \in (0,1)$

$$\lim_{n\to\infty} -\frac{1}{n} \log \beta(P_{X^n}, Q_{X^n}, \epsilon) \geq D(P_X; Q_X). \tag{1.24}$$

Proof. Let us take $D_0 = \mathscr{T}_{[P_X]_{\delta_n,n}}$. From point 3 in Proposition 1.1, this D_0 is a feasible decision set for a large enough n and δ_n which fulfills the condition of the proposition. From point 2 in Proposition 1.1, we can also show that this D_0 gives the correct exponent in eq. 1.24. □

This theorem is originally proved in (Sanov, 1958), with a different proof than what we presented here. The original proof actually gives equality instead of the upper bound that we showed here. We will delay the discussion about the corresponding lower bound to the next section where we use it to illustrate the use of KL divergence and data processing inequality.

In this section, we have reviewed several results in hypothesis testing and introduced the basic information measures and type method. One can directly see that the techniques introduced in this section is designed specifically for the i.i.d. cases and under the assumption that the data-generating process does not change with the given test function.

1.2 Information Measures

In the previous section, we gave an example of the operational importance of KL divergence by showing a lower bound on the exponent of a binary hypothesis testing problem. In this section, we will delve deeper into the properties of KL divergence and introduce other forms of information measures.

Before we discuss these information measures, we recall the definition of convex functions and Jensen's inequality (Jensen, 1906).

Definition 1.3. *A function $f : \mathbb{R} \to \mathbb{R}$ is said to be convex if for all x, y in the domain of the function, then,*

$$f\left(\lambda x + (1-\lambda) y\right) \leq \lambda f(x) + (1-\lambda) f\left(y\right) \tag{1.25}$$

for all $\lambda \in [0,1]$. Function $f(\cdot)$ is said to be concave if $x \to -f(x)$ is convex.

Proposition 1.2 (Jensen's Inequality)**.** *For any random variable X and convex function $f(\cdot)$, we have,*

$$f\left(\mathbb{E}[X]\right) \leq \mathbb{E}\left[f(X)\right]. \tag{1.26}$$

We will introduce the notion of conditional divergence and several properties of the divergence,

Definition 1.4. *We define conditional KL divergence as,*

$$D\left(W_{Y|X}; Q_{Y|X} \middle| P_X\right) = \mathbb{E}_{P_X}\left[D\left(W_{Y|X=X}; Q_{Y|X=X}\right)\right]. \tag{1.27}$$

Proposition 1.3. *We have,*

1. *KL Divergence is non-negative, i.e., $D(P;Q) \geq 0$,*
2. *Conditional KL divergence can be expressed as unconditional KL divergence,*

$$D(P_{Y|X}; Q_{Y|X}|P_X) = D(P_{Y|X}P_X; Q_{Y|X}P_X), \tag{1.28}$$

3. *Adding a random variable can only increase KL divergence,*

$$D(P_{X,Y}; Q_{X,Y}) \geq D(P_X; Q_X), \tag{1.29}$$

4. *Data Processing Inequality, for any P_X, Q_X and $W_{Y|X}$*

$$D(P_Y; Q_Y) \leq D(P_X; Q_X), \tag{1.30}$$

   where $P_Y = W_{Y|X} \circ P_X$ and $Q_Y = W_{Y|X} \circ Q_X$,

5. *KL divergence is subject to the chain rule,*

$$D(P_{X,Y}; Q_{X,Y}) = D(P_{Y|X}; Q_{Y|X}|P_X) + D(P_X; Q_X), \tag{1.31}$$

6. *KL divergence can be tensorized,*

$$D\left(\prod_{j=1}^{n} P_{X[j]}; \prod_{j=1}^{n} Q_{X[j]}\right) = \sum_{j=1}^{n} D(P_{X[j]}; Q_{X[j]}), \tag{1.32}$$

7. *KL divergence can only increases by conditioning,*

$$D\left(P_{Y|X}; Q_{Y|X}|P_X\right) \geq D\left(P_Y; Q_Y\right) \tag{1.33}$$

where $P_Y = P_{Y|X} \circ P_X$ *and* $Q_Y = Q_{Y|X} \circ P_X$.

Proof. Points 1 to 4 can be proven in a more general setting (F-divergences). To prevent duplication, readers are advised to flip to chapter 3 to find the proof. We will only prove points 5-7 which are peculiar to KL divergence.

Point 5 can be shown as,

$$\begin{aligned} D\left(P_{Y,X}; Q_{Y,X}\right) &= \mathbb{E}_{P_{Y,X}}\left[\log \frac{P_{Y|X}P_X}{Q_{Y|X}Q_X}\right] \\ &= \mathbb{E}_{P_{Y,X}}\left[\log \frac{P_{Y|X}}{Q_{Y|X}}\right] + \mathbb{E}_{P_{Y,X}}\left[\log \frac{P_X}{Q_X}\right] \\ &= D\left(P_{Y|X}; Q_{Y|X}|P_X\right) + D\left(P_X; Q_X\right). \end{aligned} \tag{1.34}$$

Point 6 can be shown using the chain rule,

$$D\left(P_{X_1}P_{X_2}; Q_{X_1}Q_{X_2}\right) = D\left(P_{X_1}; Q_{X_1}\right) + D\left(P_{X_2}; Q_{X_2}\right). \tag{1.35}$$

Point 7 is also a consequence of the chain rule,

$$\begin{aligned} D\left(P_{Y|X}P_X; Q_{Y|X}P_X\right) &= D\left(P_{Y|X}; Q_{Y|X}|P_X\right) + \underbrace{D(P_X; P_X)}_{=0} \\ &= \underbrace{D\left(P_{X|Y}; Q_{X|Y}|P_Y\right)}_{\geq 0} + D(P_Y; Q_Y). \end{aligned} \tag{1.36}$$

Hence we have,

$$D\left(P_{Y|X}; Q_{Y|X}|P_X\right) = D\left(P_{X|Y}; Q_{X|Y}|P_Y\right) + D(P_Y; Q_Y) \geq D(P_Y; Q_Y). \tag{1.37}$$

□

One of the applications of KL divergence and data processing inequality is to give an upper bound to the exponent of Type-II error.

Theorem 1.4. *For* $\epsilon \in (0,1)$

$$\lim_{n\to\infty} -\frac{1}{n}\log \beta(P_{X^n}, Q_{X^n}, \epsilon) \leq \frac{1}{1-\epsilon} D(P_X; Q_X) \tag{1.38}$$

Proof. Consider the data processing inequality, where we consider $X[1:n] \to Y = \mathbb{1}\{X[1:n] \in D_0\}$, we have

$$D\left(P_Y; Q_Y\right) \leq D\left(P_{X[1:n]}; Q_{X[1:n]}\right) = nD\left(P_X; Q_X\right). \tag{1.39}$$

Hence we have,

$$(1-\epsilon)\log\frac{1-\epsilon}{\beta(P_{X^n},Q_{X^n},\epsilon)}+\epsilon\log\frac{\epsilon}{1-\beta(P_{X^n},Q_{X^n},\epsilon)}\leq nD(P_X;Q_X) \tag{1.40}$$

Rearranging gives us,

$$-\log\beta(P_{X^n},Q_{X^n},\epsilon)\leq n\frac{1}{1-\epsilon}D(P_X;Q_X)-\frac{\epsilon}{1-\epsilon}\log\frac{\epsilon}{1-\beta(P_{X^n},Q_{X^n},\epsilon)}. \tag{1.41}$$

As we have $\beta(P_{X^n},Q_{X^n},\epsilon)\to 0$, dividing both sides by n and taking $n\to\infty$ gives us the desired inequality. □

Intuitively, KL divergence measures the ease of distinguishing two different hypotheses. On the other hand, we have the notion of "entropy" which captures the "dispersion" a random variable[8]. Entropy is defined as follows.

Definition 1.5. *Given a distribution P_X, we define*

$$H(P_X)=-\mathbb{E}_{P_X}[\log P_X(X)]. \tag{1.42}$$

We also define the notion of conditional entropy for distribution P_X and $P_{Y|X}$.

$$H(P_{Y|X}|P_X)=-\mathbb{E}_{P_{Y|X}P_X}\left[\log P_{Y|X=X}(Y)\right]. \tag{1.43}$$

We have the following properties of entropy.

Proposition 1.4. *We have*

1. *Entropy is non-negative: $H(P_X)\geq 0$ and $H(P_{Y|X}|P_X)\geq 0$.*
2. *Entropy is upper-bounded by the alphabet size:*
$$H(P_X)\leq\log|\mathscr{X}|. \tag{1.44}$$
3. *Chain rule:*
$$H\left(P_{X[1:n]}\right)=\sum_{j=2}^{n}H\left(P_{X[j]|X[1:j-1]}\middle|P_{X[1:j-1]}\right)+H(P_{X[1]}). \tag{1.45}$$
4. *Monotonicity:*
$$H\left(P_{X,Y}\right)\geq H\left(P_X\right). \tag{1.46}$$

[8]Although we will not discuss it further in this thesis, this intuitive notion can be backed by formal results. In the sense that we can prove a lower bound to the difficulty of "guessing" a random variable which increases as its entropy increases (Massey, 1994). In fact, in his seminal work on cryptography (Shannon, 1949), Shannon argues about the security of a cryptosystem in terms of the "work" needed to guess the secret key, which can be lower bounded by a function of conditional entropy.

5. *Conditioning reduces entropy:*

$$H\left(P_{Y|X}|P_X\right) \leq H\left(P_Y\right). \tag{1.47}$$

Proof. 1. Note that $x \to -x\log x$ is a non-negative function for $x \in [0,1]$.

2. We note that $x \to -x\log x$ is a concave function; hence, the optimum is achieved by a uniform distribution on its support, which gives us $H(P_X) = \log|\text{support}(X)|$. Hence, the maximum is achieved at $H(P_X) = \log|\mathscr{X}|$.

3. We have

$$\begin{aligned} H\left(P_{X,Y}\right) &= -\mathbb{E}\left[\log P_{X,Y}(X,Y)\right] \\ &= -\mathbb{E}\left[\log P_{Y|X}(Y,X)\right] - \mathbb{E}\left[\log P_X(X)\right] \\ &= H\left(P_{Y|X}|P_X\right) + H\left(P_X\right). \end{aligned} \tag{1.48}$$

   Applying this equality repeatedly gives us the result.

4. We have

$$\begin{aligned} H\left(P_{X,Y}\right) &= H\left(P_{Y|X}|P_X\right) + H\left(P_X\right) \\ &\geq H\left(P_X\right). \end{aligned} \tag{1.49}$$

5. As $x \to -x\log x$ is a concave function, and $\mathbb{E}_{P_X}\left[P_{Y|X=X}\right] = P_Y$, using Jensen's inequality, we have

$$H(P_{Y|X}|P_X) = -\mathbb{E}_{P_X}\left[\mathbb{E}_{P_{Y|X=X}}\left[\log P_{Y|X=X}\right]\right] \leq \mathbb{E}_{P_{X,Y}}\left[\log P_Y\right] = H(P_Y). \tag{1.50}$$

□

Entropy also plays an important role in giving a bound on the size of a type. Let us also use this opportunity to introduce the notion of conditional type. We also define $\mathscr{X}$ and $\mathscr{Y}$ as the input and the output alphabet respectively.

Definition 1.6. *For a given realization of $x[1:n]$ and $y[1:n]$, let us define the conditional type of Y given $x[1:n]$ as*

$$P_{y[1:n]|x[1:n]}(y',x') = \frac{\sum_{i=1}^{n} \mathbb{1}\{x[i]=x', y[i]=y'\}}{\sum_{i=1}^{n} \mathbb{1}\{x[i]=x'\}}. \tag{1.51}$$

Given a realization of $x[1:n]$, let us denote the set of $y[1:n]$ with conditional type $P_{Y|X}$ as $\mathscr{T}^n_{P_{Y|X}}(x[1:n])$.

Proposition 1.5. *We have*

$$(n+1)^{-|\mathscr{X}|}\exp\left(nH(P_X)\right) \leq \left|\mathscr{T}^n_{P_X}\right| \leq \exp\left(nH(P_X)\right) \tag{1.52}$$

and for all $x[1:n]$ *with type* P_X

$$(n+1)^{-|\mathscr{Y}||\mathscr{X}|}\exp\left(nH(P_{Y|X}|P_X)\right) \leq \left|\mathscr{T}^n_{P_{Y|X}}(x[1:n])\right| \leq \exp\left(nH(P_{Y|X}|P_X)\right). \tag{1.53}$$

Proof. The proof for conditional type mirrors the proof for the unconditional case, with the main difference being that we consider each conditioning value in $\mathscr{X}$ separately. Hence, we will only discuss the unconditional case.

We have proof of the upper bounds in 1.15. Hence, we will only show the lower bound. Let us consider a specific P_X,

$$1 = P_X\left(\bigcup_{P'_X \in \mathscr{P}_n(\mathscr{X})} \mathscr{T}^n_{P'_X}\right) = \sum_{P'_X \in \mathscr{P}_n(\mathscr{X})} P_X\left(\mathscr{T}^n_{P'_X}\right). \tag{1.54}$$

Note that the largest terms in the summation are given by $P'_X = P_X$. We have,

$$1 \leq (n+1)^{|\mathscr{X}|} P_X\left(\mathscr{T}^n_{P_X}\right) = (n+1)^{|\mathscr{X}|}|\mathscr{T}^n_{P_X}|\exp\left(-nH(P_X)\right). \tag{1.55}$$

This implies that,

$$(n+1)^{-|\mathscr{X}|}\exp\left(nH(P_X)\right) \leq |\mathscr{T}^n_{P_X}|. \tag{1.56}$$

□

More importantly, for a given P_X, the set $\mathscr{T}^n_{P_X}$ forms the "core" of any high probability sets, which can be expressed more formally by the following proposition.

Proposition 1.6. *For any set* $\mathscr{A}$ *with high probability in set* $P_{X[1:n]}$*, i.e.,*

$$P_{X[1:n]}(\mathscr{A}) \geq \gamma, \tag{1.57}$$

for all $\gamma > 0$*. We have*

$$\liminf_{n\to\infty} \frac{1}{n}\log|\mathscr{A}| \leq H(P_X) - O\left(\frac{\log n + \log\delta_n}{n}\right). \tag{1.58}$$

Proof. We have,

$$P_{X[1:n]}\left(\mathscr{A}\bigcup \mathscr{T}^n_{[P_X]_{\delta_n,n}}\right) = P_{X[1:n]}\left(\mathscr{A}\right) + P_{X[1:n]}\left(\mathscr{T}^n_{[P_X]_{\delta_n,n}}\right) - P_{X[1:n]}\left(\mathscr{A}\bigcap \mathscr{T}^n_{[P_X]_{\delta_n,n}}\right) \tag{1.59}$$

By upper bounding the LHS by 1 and using the result of Proposition 1.1, we have,

$$P_{X[1:n]}\left(\mathscr{A}\bigcap\mathscr{T}^n_{[P_X]_{\delta_n,n}}\right)\geq\gamma-O\left(\frac{1}{n\delta_n^2}\right). \tag{1.60}$$

Note that for all $x[1:n]\in\mathscr{A}\bigcap\mathscr{T}^n_{[P_X]_{\delta_n,n}}$, we have,

$$P_{X[1:n]}(x[1:n])\leq\exp\left(-n(H(P_X)-O(\delta_n))\right). \tag{1.61}$$

Therefore,

$$P_{X[1:n]}(\mathscr{A}\bigcap\mathscr{T}^n_{[P_X]_{\delta_n,n}})\leq|\mathscr{A}|\exp\left(-n(H(P_X)-O(\delta_n))\right). \tag{1.62}$$

This gives us

$$|\mathscr{A}|\geq\left(\gamma-O\left(\frac{1}{n\delta_n^2}\right)\right)\exp\left(n(H(P_X)-O(\delta_n))\right) \tag{1.63}$$

which implies the proposition. □

Now, we can define the mutual information. This quantity can be understood intuitively through two different viewpoints: i) As the conditional KL divergence of the conditional distribution $P_{Y|X}$ against the marginal distribution P_Y, ii) As the reduction of a random variable entropy given another random variable.

Definition 1.7. *Given a distribution $P_{X,Y}$, we define mutual information as*

$$I(P_X;P_{Y|X})=D(P_{Y|X};P_Y|P_X), \tag{1.64}$$

with $P_Y=P_{Y|X}\circ P_X$ and the conditional mutual information as,

$$I(P_{X|Z};P_{Y|XZ}|P_Z)=\mathbb{E}_{P_Z}\left[D(P_{Y|X,Z=Z};P_{Y|Z=Z}|P_{X|Z=Z})\right]. \tag{1.65}$$

We have the following properties of mutual information.

Proposition 1.7. *We have,*

1. *Positivity: $I(P_X;P_{Y|X})\geq 0$,*
2. *Data processing inequality: consider $X\to Y\to Z$,*

$$I\left(P_X;P_{Z|X}\right)\leq I\left(P_X;P_{Y|X}\right), \tag{1.66}$$

3. *Mutual information in terms of entropy:*

$$I(P_X;P_{Y|X})=H(P_Y)-H(P_{Y|X}|P_X), \tag{1.67}$$

4. *Chain rule:*

$$I\left(P_X; P_{Y[1:n]|X}\right) = \sum_{i=2}^{n} I\left(P_{X|Y[1:i-1]}; P_{Y[i:n]|X,Y[1:i-1]} \middle| P_{Y[1:i-1]}\right) + I\left(P_X; P_{Y[1]|X}\right), \quad (1.68)$$

5. *Mutual information as the center of conditional distribution:*

$$I(P_X; P_{Y|X}) = D(P_{Y|X}; Q_Y|P_X) - D(P_Y; Q_Y). \quad (1.69)$$

Proof. 1. This is a consequence of mutual information being a divergence.

2. Data processing inequality will be shown in a more general form in chapter 3.

3. We have

$$I(P_X; P_{Y|X}) = \mathbb{E}_{P_{X,Y}}\left[\log \frac{P_{Y|X}}{P_Y}\right] = H(P_Y) - H(P_{Y|X}|P_X). \quad (1.70)$$

4. We have

$$\begin{aligned} I(P_X; P_{Y[1:2]|X}) &= H(P_{Y[1:2]}) - H(P_{Y[1:2]|X}|P_X) \\ &= H(P_{Y[2]|Y[1]}|P_{Y[1]}) + H(P_{Y[1]}) - H(P_{Y[2]|X,Y[1]}|P_{X,Y[1]}) - H(P_{Y[1]|X}|P_X) \\ &= I(P_X; P_{Y[1]|X}) + I(P_{X|Y[1]}; P_{Y[2]|X,Y[1]}|P_{Y[1]}) \end{aligned} \quad (1.71)$$

   Repeating this equality gives us the chain rule.

5. We have

$$\begin{aligned} I(P_X; P_{Y|X}) &= \mathbb{E}_{P_{X,Y}}\left[\log \frac{P_{Y|X}}{P_Y}\right] \\ &= \mathbb{E}_{P_{X,Y}}\left[\log \frac{P_{Y|X}}{Q_Y} + \log \frac{Q_Y}{P_Y}\right] \\ &= D(P_{Y|X}; Q_Y|P_X) - D(P_Y; Q_Y) \end{aligned} \quad (1.72)$$

□

In this section, we have discussed commonly used information measures. In the next section, we will discuss how these information measures are used to give a converse to practical channel coding problems.

1.3 Channel Coding

In Shannon's 1948 paper, he proposed a model of a digital communication system and derived a result that gives us the fundamental performance that can be achieved by any communication schemes. The result is derived by considering the abstracted model of point-to-point communication shown in Figure 1.1.

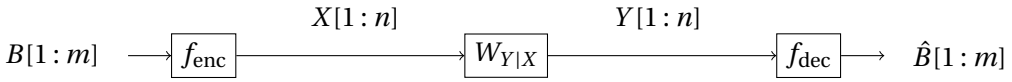


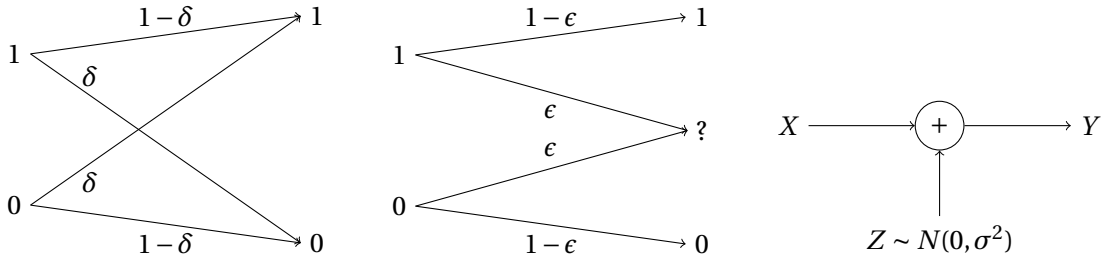
Figure 1.1: Point-to-point communication block diagram

In Figure 1.1, we assume that $B[1:m]$ is i.i.d., and $B[i] \sim \text{Bernoulli}(1/2)$. The result presented in Shannon's paper is more general as it does not require the source to be a bitstream. In fact, there is a vibrant theory on how to represent a message in terms of a bitstream.

In this model, the transmitter wants to communicate m bits of information by using the channel n times. In practice, each channel use "consumes" a degree of freedom[9]. For example, in Orthogonal Frequency Division Multiplexing (Wu and Zou, 1995), each channel use corresponds to a different sub-frequency bands; or in Time Division Multiplexing, each channel use corresponds to a different time slices. Hence, there is a practical importance in trying to understand the trade-off between m and n.

Let us refer to the mapping between $B[1:m]$ and $X[1:n]$ as the *encoding* scheme, respectively, the mapping between $Y[1:n]$ and $\hat{B}[1:m]$ will be referred to as the *decoding* scheme. We will refer to this pair as a channel coding scheme.

In this model, the main challenge hindering reliable communication between the transmitter and the receiver is mathematically modeled in terms of the channel. During Shannon's time, these channels mainly model the thermal noise which distorts long-range wired communication [10]. In this work, we will mainly deals with channel models in Figure 1.1.



(a) Binary Symmetric Channels, $BSC(\delta)$ (b) Binary Erasure Channels, $BEC(\epsilon)$ (c) Additive White Gaussian Noise Channels, $AWGN(\sigma^2)$

Figure 1.2: Examples of channel models that we will use in this work.

Moreover, throughout this work, we will only consider *memoryless* channels, i.e., channels

[9] Degree of freedom in the sense of number of "independent" dimensions of the signal, not in the sense of "independent" paths in wireless communication.

[10] In fact, it is a miracle that a theory built on this simplistic mathematical model manages to provide guidance for designing real communication systems operating under the complex physical realities. But then again, the existence of this theory surely spurs engineers to design their systems to make these channel models a reasonable approximation of the physical dynamics of their systems.

that fulfill,

$$W_{Y[1:n]|X[1:n]}(y[1:n]|x[1:n]) = \prod_{i=1}^{n} W_{Y[1]|X[1]}(y[i]|x[i]). \tag{1.73}$$

Channel coding schemes are evaluated on their capability to send messages reliably. More formally, for a given channel coding scheme, we define the (average block) error probability as $P_e = \Pr\left(B[1:m] \neq \hat{B}[1:m]\right)$. Consider a sequence of tuples (m_k, n_k) with the corresponding error probability $P_e^{(k)}$. Classical channel coding result is concerned with finding the largest $\liminf_{k\to\infty} m_k/n_k$ such that $\limsup_{k\to\infty} P_e^{(k)} = 0$. As an application of the information measures that we discussed in the previous section, we can give an upper bound on these m_k/n_k.

Before that, let us consider the inequality between error probability and its entropy; this result is commonly known as Fano's inequality(Fano, 1961).

Lemma 1.1. *Consider discrete random variables $(X, \hat{X})$ both defined in the same alphabet $\mathcal{X}$, we have,*

$$H(P_{\hat{X}|X}|P_X) \leq \log(2) + \Pr(\hat{X} \neq X)\log|\mathcal{X}|. \tag{1.74}$$

Theorem 1.5. *Given a channel $W_{Y|X}$, for every channel coding scheme, we have,*

$$P_e \geq 1 - \frac{nC}{m\log(2)} - \frac{1}{m}, \tag{1.75}$$

with $C = \max_{P_X} I(P_X; W_{Y|X})$.

Proof. We have,

$$\begin{aligned}
m\log(2) &= H(P_{B[1:m]}) \\
&= I(P_{B[1:m]}; P_{\hat{B}[1:m]|B[1:m]}) + H(P_{\hat{B}[1:m]|B[1:m]}|P_{B[1:m]}) \\
&\leq I(P_{B[1:m]}; P_{\hat{B}[1:m]|B[1:m]}) + \log(2) + P_e m\log(2) \\
&\leq I(P_{X[1:n]}; W_{Y[1:n]|X[1:n]}) + \log(2) + P_e m\log(2) \\
&\leq \sum_{i=1}^{n} I(P_{X[i]}; W_{Y|X}) + \log(2) + P_e m\log(2) \\
&\leq nC + \log(2) + P_e m\log(2)
\end{aligned} \tag{1.76}$$

with

$$1 - \frac{nC}{m\log(2)} - \frac{1}{m} \leq P_e. \tag{1.77}$$

□

Surprisingly, proving the existence of a channel coding scheme that achieves this upper bound is not difficult. There are several constructions. We will present the construction method presented in Csiszár and Körner's textbook(Csiszár and Körner, 2011) to illustrate the use of the concept of types.

Proposition 1.8. *Given a channel $W_{Y|X}$, for any $p_e > 0$, there exists a channel coding scheme with an error probability p_e if,*

$$\frac{m\log(2)}{n} \leq \sup_{P_X:\forall_{x\in\mathscr{X}} P_X(x)>0} I(P_X, W_{Y|X}) - O(n^{-1/2+\epsilon}) \tag{1.78}$$

for any $\epsilon > 0$.

Proof. We will design the channel coding schemes sequentially. Namely, by assigning the set of $y[1:n]$'s that will be decoded to message i, one by one. We will use the variable B_i to hold the set of $y[1:n]$ that has been occupied by a codeword in the i-th step. We start with $B_0 = \emptyset$.

At the $i+1$-th step, we check if we can find a codeword in $\mathscr{T}^n_{[P_X]_{\delta_n,n}}$ with high decoding probability among the unoccupied $y[1:n]$, i.e., we are trying to find $x[1:n]$ such that

$$W_{Y[1:n]|X[1:n]=x[1:n]}\left(\mathscr{T}_{[W_{Y|X}]_{\delta'_n,n}}(x[1:n]) - B_i | x[1:n]\right) \geq 1 - p_e. \tag{1.79}$$

If we can find such codeword, we will augment our current encoding function and decoding function such that $f_{\text{enc}}(i+1) = x[1:n]$ and $f_{\text{dec}}(y[1:n]) = i+1$ if $y[1:n] \in \mathscr{T}_{[W_{Y|X}]_{\delta_n,n}}(x[1:n]) - B_i$. Otherwise, we stop the design process.

Note that by construction, this channel coding scheme will have an error probability of p_e. We only need to characterize the number of steps that we can perform before being stopped.

Let us assume that we stopped at time i^*, we have that for all $x[1:n] \in \mathscr{T}_{[P_X]_{\delta_n,n}}$,

$$W_{Y[1:n]|X[1:n]=x[1:n]}\left(\mathscr{T}_{[W_{Y|X}]_{\delta'_n,n}}(x[1:n]) - B_i | x[1:n]\right) < 1 - p_e \tag{1.80}$$

which implies that for all $x[1:n] \in \mathscr{T}_{[P_X]_{\delta_n,n}}$

$$W_{Y[1:n]|X[1:n]=x[1:n]}\left(B_{i^*} | x[1:n]\right) \geq p_e - O\left(\frac{1}{n{\delta'_n}^2}\right). \tag{1.81}$$

This implies that

$$P_{Y[1:n]}\left(B_{i^*}\right) \geq p_e - O\left(\frac{1}{n{\delta_n}^2}\right) - O\left(\frac{1}{n^2\delta_n^2{\delta'_n}^2}\right). \tag{1.82}$$

Hence this B_{i^*} must be a high probability set in P_Y, which given our previous results shows that

$$\frac{1}{n}\log|B_{i^*}| \geq H(P_Y) - O\left(\frac{\log_n + \log\delta_n + \log\delta'_n}{n}\right) \tag{1.83}$$

Given the construction process, we also have that

$$|B_{i^*}| \leq \sum_{j=1}^{i^*}\left|\mathscr{T}_{[W_{Y|X}]_{\delta'_n,n}}\left(f_{\text{enc}}(j)\right)\right| \leq i^* \exp\left(n(H(W_{Y|X}|P_X) + O(\delta' + \delta))\right). \tag{1.84}$$

Combining these two bounds gives us

$$\frac{1}{n}\log i^* \geq H(P_Y) - H(W_{Y|X}|P_X) - O\left(\frac{\log_n + \log\delta_n + \log\delta'_n}{n}\right) - O(\delta_n + \delta'_n). \tag{1.85}$$

We only need to make sure that $i^* \geq m$. Taking δ_n and δ'_n as $O(n^{-0.5+\epsilon})$ gives us the statement of the proposition. □

Combining the achievability and converse, we can conclude that the best trade-off of n and m that one can hope for is given by $I(P_X; W_{Y|X})/\log 2$. Note that in the achievability part, we have control over P_X, hence we can optimize the performance of our scheme by choosing P_X that maximizes $I(P_X; W_{Y|X})$. This allows us to define the notion of channel capacity $C(W_{Y|X}) = \max_{P_X} I(P_X; W_{Y|X})$.

1.4 Thesis Outline

This work is divided into two parts covering our results in information velocity and monitoring problems. The first composed of 4 chapters discussing our results in information velocity, and the second part is composed of 1 chapter discussing the monitoring problems.

1. **Chapter 1 (Blockwise Information Velocity):** The subfield of information velocity has only recently attracted interest; hence, it might not be familiar to most information theory researchers. In this chapter, we will introduce the mathematical formulation of information velocity and review several known results in this area. We will also introduce blockwise information velocity, which studies information velocity in the regime where the message size grows to infinity.

2. **Chapter 2 (A Converse for Blockwise Information Velocity):** In this chapter, we will show how information-theoretic arguments can be used to give a converse for blockwise information velocity. Intuitively, the idea behind the converse is to model the flow of information through the relay chain in a similar manner to a transport differential equation, where the "amount of information" flowing at each channel use is proportional to the difference between the information that the sender possess and the information that the receiver already possess. To quantify this information, we will use tools from F-divergences(Ali and Silvey, 1966)(Csiszár, 1967) and strong data processing inequality(Ahlswede and Gacs, 1976) (Polyanskiy and Wu, 2015).

3. **Chapter 3 (Blockwise Information Velocity in Binary Erasure Channels):** In this chapter, we will present our results on the achievability scheme for binary erasure channels. In single-bit relaying problems for BEC, the intermediate relays only have two states, namely whether it has received the bit or not. But in multiple bit settings, the transmission progress at each intermediate relay can be more fine-grained, making the analysis

difficult. Here, we present a relaying scheme that achieves positive blockwise information velocity by building a tape abstraction on top of tree codes. This scheme builds on the previous results of using tree codes for distributed computation (Schulman, 1993).

4. **Chapter 4 (Blockwise Information Velocity in AWGN Channels with Feedback):** A well-known result in information theory states that the fundamental limit of the rate of communication cannot be improved by adding feedback (Shannon, 1961). However, it is also known that feedback allows us to achieve this fundamental limit with a simple scheme. One of the most striking results is presented by Schalkwijk and Kailath, which showed that a simple linear scheme can achieve reliable communication with doubly-exponentially decaying error probability in AWGN channels with feedback(Schalkwijk and Kailath, 1966). In this chapter, we show that a modification to the Schalkwijk-Kailath scheme achieves positive blockwise information velocity of AWGN with feedback, both in the case where the relays' noise levels are homogeneous or heterogeneous.

5. **Chapter 5 (Monitoring Problem as Adversarial Hypothesis Testing Problems):** As we discussed in this chapter, the classical hypothesis testing problem relies on the fact that the data-generating process does not react to the hypothesis testing process. In this chapter, we will present a monitoring model where there is an adversarial relationship between the tester and the data-generating process. In this model, an adversarial data-generating process obtains rewards depending on the realization of its output, but the reward will be revoked if the tester manages to differentiate the adversarial data-generating process from a group of regular data-generating processes. We will show that in this case, surprisingly, the "optimal" statistics in the asymptotic case are equal to the empirical KL divergence.

Finally, at the end of this work, we will present concluding remarks on our results, and more importantly, we will discuss several future steps that we have not managed to explore in this work.

1.5 Appendix: Results on Random Variables

Concentration Inequalities

In this thesis, we will use several concentration results. A common foundation for a large class of concentration inequalities is the following observation for non-negative random variable X:

Proposition 1.9 (Markov's Inequality)**.** *For a non-negative random variable X where $\mathbb{E}[X]$ exists,*

$$\Pr(X \geq \alpha) \leq \frac{\mathbb{E}[X]}{\alpha}. \tag{1.86}$$

Proof. We have,

$$\begin{aligned}\Pr(X \geq \alpha) &= \int_0^\infty \mathbb{1}\{X \geq \alpha\} dP_X \\ &\leq \int_0^\infty \frac{X}{\alpha} dP_X \\ &= \frac{\mathbb{E}[X]}{\alpha}.\end{aligned} \tag{1.87}$$

□

A straightforward extension to Markov's inequality can be obtained by noting that the random variable X can be a function of other random variables. A common extension where we bound the deviation from the mean of any random variable (not only non-negative random variables) is Chebyshev's inequality.

Proposition 1.10 (Chebyshev's Inequality)**.** *For a random variable X with finite first and second moments, we have,*

$$\Pr(X - \mathbb{E}[X] \geq \alpha) \leq \frac{Var(X)}{\alpha^2}. \tag{1.88}$$

Proof. We consider the following function $f(x) = (x - \mathbb{E}[X])^2$. This function is non-negative, and plugging this function into Markov's inequality gives us the required inequality. □

This inequality gives us a way to show that the empirical average converges to the mean, even if the distributions are not exactly similar.

Proposition 1.11. *Given independent random variables $X[1:n]$ such that $\mathbb{E}[X[i]] = \mu$ and $Var(X[i]) = \sigma_i^2$, we have*

$$\limsup_{n\to\infty} \Pr\left(\sum_{i=1}^n \frac{X[i]}{n} - \mu \geq \alpha_n\right) = 0$$

if $\lim_{n\to\infty} n\alpha_n^2 = \infty$.

Proof. By Chebyshev's inequality, we have that,

$$\begin{aligned}\Pr\left(\sum_{i=1}^n \frac{X[i]}{n} - \mu \geq \alpha_n\right) &\leq \sum_{i=1}^n \frac{\sigma_i^2}{n^2\alpha_n^2} \\ &\leq \frac{\max_i \sigma_i^2}{n\alpha_n^2}.\end{aligned} \tag{1.89}$$

We can see that the upper bound vanishes as $n \to \infty$ under our hypothesis. □

If the moment generating function of X exists, we can use this information to assert a faster decay of the distribution's tail. The following inequality is an application of Markov's inequality

using the function $f(x) = e^{\lambda x}$. This inequality allows us to establish an exponential upper bound to the tail probability of the deviation.

Proposition 1.12 (Chernoff-Cramer Inequality)**.** *Given a random variable X such that $\mathbb{E}[e^{\lambda X}]$ is finite for $\lambda(0, \lambda_+)$ then,*

$$\Pr(X \geq \alpha) \leq \frac{\mathbb{E}[e^{\lambda X}]}{e^{\lambda \alpha}} \tag{1.90}$$

for all $\lambda \in (0, \lambda_+)$. Consider i.i.d. random variables $X[1:n]$ then we have,

$$\Pr\left(\sum_{i=1}^{n} \frac{X[i]}{n} \geq \alpha\right) \leq \exp\left(-n\left(\lambda\alpha - \log\mathbb{E}[e^{\lambda X}]\right)\right). \tag{1.91}$$

Interestingly enough, we can find a lower bound of the same form by a similar argument. The argument for this fact is based on the notion of the tilted distribution. Given a distribution P, we define a tilted distribution P_γ as,

$$P_\gamma(X) = \frac{e^{\gamma X} P(X)}{\mathbb{E}[e^{\gamma X}]}. \tag{1.92}$$

Proposition 1.13. *Consider i.i.d. random variables $X[1:n]$ such that there exists moment generating functions $\mathbb{E}[e^{\lambda X}]$ for $\lambda \in (0, \lambda_+)$. Then the following lower bound holds for $\lambda \in (0, \lambda_+)$,*

$$\lim_{n\to\infty} \frac{1}{n} \log \Pr\left(\sum_{i=1}^{n} \frac{X[i]}{n} \geq \alpha\right) \geq -\left(\lambda(\alpha)\alpha - \log\mathbb{E}\left[e^{\lambda(\alpha)X}\right]\right) \tag{1.93}$$

where $\lambda(\alpha)$ is such that $\mathbb{E}_{P_{\lambda(\alpha)}}[X] = \alpha$ if $\alpha > \mathbb{E}[X]$ or 0 otherwise.

Proof. Let us denote $\sum_{i=1}^{n} \frac{X[i]}{n} = Z[n]$. We have the following,

$$\Pr(Z[n] \geq \alpha n) \geq \Pr(Z[n] \geq (\alpha+\epsilon)n) = \int_{(\alpha+\epsilon)n}^{\infty} dP_X^{\otimes n}. \tag{1.94}$$

Let us introduce $\tilde{Z}[n]$ to be $Z[n]$ under the tilted distribution $P_{X,\lambda(\alpha+2\epsilon)}^{\otimes n}$. By multiplying with the change of measure variable and then truncating the integration region, we have

$$\begin{aligned}
\Pr(Z[n] \geq \alpha n) &\geq \int_{(\alpha+\epsilon)n}^{(\alpha+3\epsilon)n} \mathbb{E}_{P_X^{\otimes n}}[e^{\lambda(\alpha+2\epsilon)Z[n]}] e^{-\lambda(\alpha+2\epsilon)z} dP_{\tilde{Z}}(z) \\
&\geq \mathbb{E}_{P_X^{\otimes n}}[e^{\lambda(\alpha+2\epsilon)Z[n]}] e^{-\lambda(\alpha+2\epsilon)(\alpha+3\epsilon)} \int_{(\alpha+\epsilon)n}^{(\alpha+3\epsilon)n} dP_{\tilde{Z}}(z) \\
&= \exp\left(-\left(\lambda(\alpha+2\epsilon)(\alpha+3\epsilon) - \log\mathbb{E}_{P_X^{\otimes n}}[e^{\lambda(\alpha+2\epsilon)Z[n]}]\right)\right) P_{\tilde{Z}}(z)\left(\left|\tilde{Z}[n] - (\alpha+2\epsilon)\right| \leq \epsilon\right).
\end{aligned} \tag{1.95}$$

Note that because the moment generating function for X exists, it implies that the second

moment of the tilted distribution also exists; hence, we have, by Chebyshev's inequality,

$$\lim_{n\to\infty} P_{\tilde{Z}}(z)\left(\left|\tilde{Z}[n]-(\alpha+2\epsilon)\right|\leq\epsilon\right)=1. \tag{1.96}$$

Note that the lower bound is valid for all ϵ. By taking $\epsilon\to 0$, we have the statement of the proposition. □

Finding the tightest exponent of the upper bound using the Chernoff-Cramer's inequality is equal to finding λ which maximizes $\lambda\alpha-\log\mathbb{E}[e^{\lambda X}]$.

The first-order condition for optimality gives us,

$$\alpha=\frac{\mathbb{E}\left[Xe^{\lambda X}\right]}{\mathbb{E}\left[e^{\lambda X}\right]}=\mathbb{E}_{P_\lambda}[X], \tag{1.97}$$

and taking the second derivatives gives us $-\left(\mathbb{E}_{P_\lambda}[X^2]-\mathbb{E}_{P_\lambda}[X]^2\right)$. Hence, for a given α, the tightest exponent is achieved at $\lambda(\alpha)$. This shows that the upper bound given by the Chernoff-Cramer's inequality is exponentially tight.

A particularly useful results is the specialization of Chernoff-Cramer's inequality for Bernoulli random variables.

Corollary 1.1. *Consider i.i.d. random variables $X[1:n]$ with $X[i]$ distributed according to Bernoulli(β) then we have,*

$$\Pr\left(\sum_{i=1}^{n}\frac{X[i]}{n}\geq\alpha\right)\leq\exp\left(-nd_2(\alpha;\beta)\right), \tag{1.98}$$

where $d_2(\cdot;\cdot)$ is the binary KL divergence.

Proof. By using Chernoff-Cramer's bound and minimizing the upper bound, we obtain the optimal λ

$$\lambda=\frac{(1-\beta)\alpha}{(1-\alpha)\beta}. \tag{1.99}$$

Plugging the optimal value to Chernoff-Cramer's bound gives us the corollary. □

Expected Value of Maximum of Random Variables

Another result that will be useful in the later chapters is on the rate of growth of the expected value of the maximum of a finite number of random variables.

Proposition 1.14. *Let us consider m i.i.d. random variables $X[1:m]$, such that $\mathbb{E}[e^X]$ exists. We have*

$$\mathbb{E}[\max_i X[i]]\leq\log m+\log\mathbb{E}\left[e^X\right]. \tag{1.100}$$

Proof. We have, by Jensen's,

$$\begin{aligned}\mathbb{E}[\max_i X[i]] &\leq \log \mathbb{E}\left[e^{\max_i X[i]}\right] &(1.101)\\ &\leq \log \mathbb{E}\left[\sum_i e^{X[i]}\right]\\ &= \log m \mathbb{E}\left[e^X\right]\\ &= \log m + \log \mathbb{E}\left[e^X\right]. &(1.102)\end{aligned}$$

□

Stochastic Dominance of Sum of Random Variables

We also need the concept stochastic dominance between random variables for our result in chapter 4.

Definition 1.8. *A random variable A stochastically dominates B if*

$$\Pr(A \geq x) \leq \Pr(B \geq x) \tag{1.103}$$

for every x.

For our purpose, the stochastic domination relation is useful, since if B stochastically dominates A, then any tail bounds for B can also be used as a tail bound for A. This is especially useful if we can easily use concentration results on B. Moreover, because summation preserves stochastic dominance.

Proposition 1.15. *Given sequences of independent random variables $A[1:n]$ and $B[1:n]$ such that $A[i]$ is stochastically dominated by $B[i]$ for all i, then we have $\sum_{i=1}^n A[i]$ is stochastically dominated by $\sum_{i=1}^n B[i]$.*

Proof. We can prove this statement by characterizing the stochastic dominance relation in terms of the existence of coupling fulfilling certain criteria. Then we show the statement by explicitly providing this coupling.

We can express the stochastic dominance as

$$1 - F_A(x) \leq 1 - F_B(x) \tag{1.104}$$

for all x, where F_A and F_B are the cumulative distribution functions of A and B respectively. Let us define the function which is the inverse of $1 - F_A(x)$,

$$G_A(z) = \inf\{x : 1 - F_A(x) \leq z\}, \tag{1.105}$$

for $z \in [0,1]$. By a simple change of variable, we can see that A is stochastically dominated by B if and only if,

$$G_A(z) \leq G_B(z). \tag{1.106}$$

Now, we can consider a coupling between A and B. Let us consider a random variable Z distributed uniformly between 0 and 1. We have $G_A(Z) = A$ and $G_B(Z) = B$ gives us the correct marginal. We can see that under this coupling $B - A \geq 0$ almost surely. Hence we show that A stochastically dominated by B implies that we can find a coupling for A and B such that $B - A \geq 0$ almost surely.

We show that the other direction also holds. Consider that we are given a coupling of $\tilde{A}$ and $\tilde{B}$ such that $\tilde{B} - \tilde{A} \geq 0$ almost surely. From this coupling, we have the following sequence of inequalities,

$$\begin{aligned} \Pr(B \geq x) &= \Pr(\tilde{B} \geq x) \\ &\geq \Pr(\tilde{A} \geq x) \\ &= \Pr(A \geq x) \end{aligned} \tag{1.107}$$

where the inequality is due to the fact that $\tilde{A} \geq x$ implies that $\tilde{B} \geq x$ due to the condition that $\tilde{B} - \tilde{A} \geq 0$ almost surely. Hence the existence of a coupling with these properties is equivalent to stochastic domination.

Consider the coupled version of $A[n]$ and $B[n]$, i.e., we have $(\tilde{A}[i], \tilde{B}[i])$ such that $\tilde{B}[i] - \tilde{A}[i] \geq 0$. Let us denote this difference as $\delta[i]$. We can construct a joint coupling such that the coupling $(\tilde{A}[i], \tilde{B}[i])$ is independent of $(\tilde{A}[j], \tilde{B}[j])$ where $i \neq j$. Under this joint coupling, we have

$$\sum_{i=1}^{n} \tilde{B}[i] - \sum_{i=1}^{n} \tilde{A}[i] = \sum_{i=1}^{n} \delta_i \tag{1.108}$$

where the right-hand side is almost surely non-negative. Hence this joint coupling implies that $\sum_{i=1}^{n} \tilde{A}[i]$ is stochastically dominated by $\sum_{i=1}^{n} \tilde{B}[i]$. □

Commonly Used Inequalities

This simple result is used surprisingly often in many proofs.

Proposition 1.16. *We have,* $\log(1+x) \leq x$.

Proof. Consider $f(x) = x - \log(1+x)$. We have $f(0) = 0$. We also have $f'(x) < 0$ for $x < 0$ and $f'(x) > 0$ for $x > 0$. Hence the inequality. □

In chapter 5, there are many cases where we need to bound the exponent of binomial coefficients. We will use the following bound.

Proposition 1.17. *We have for $q \in (0,1)$,*

$$\binom{n}{k} q^k (1-q)^{n-k} \leq e^{-n d_2(k/n;q)}, \tag{1.109}$$

where $d_2(\cdot;\cdot)$ is the binary KL divergence.

Proof. Note that $\binom{n}{k}$ is exactly $|\mathscr{T}^n_{P_X}|$ for $P_X(1) = k/n$ and $P_X(0) = 1-k/n$. If we define $Q_X(1) = q$ and $Q_X(0) = 1-q$, then we have the desired inequality by Proposition 1.1. □

Part I
Results on Information Velocity

2 Blockwise Information Velocity

> An error does not become truth by reason of multiplied propagation, nor does truth become error because nobody sees it.
>
> Mahatma Gandhi

When Shannon proved the channel coding theorem, it is reasonable to assume that the point-to-point model, where the transmitter directly communicates with the receiver, is the dominating mode of communication. However, since then, the world has been subsumed into a giant network spanning the globe. In our global network, communication between the sender and the receiver is intermediated by several routers, repeaters, relays and gateways, which, for the purpose of this work will be grouped together, under the term "relays". We argue that the following block diagram in Figure 2.1 is a more accurate model of modern communication networks.

One complicating aspect of this model is the assumption that the communication process is conducted "continuously", in the sense that that for any i, the transmission of $X_1[i], X_2[i], \dots X_\ell[i]$ happen at the same time.

Due to this assumption, we must be careful when defining the relaying function $f^{(i)}_{\text{relay}}$, such that its output at time n' should not depend on its inputs after time n'. The constraint that we will impose hinges on our definition of "causality" between a sequence of random variable.

Definition 2.1. *A sequence of random variables $T[1:n]$ is causal with respect to another*

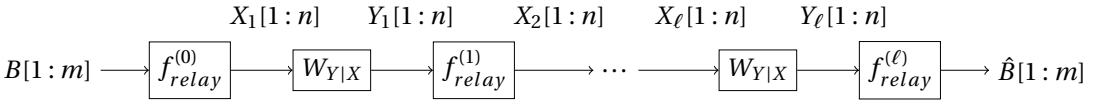


Figure 2.1: Relay communication block diagram

sequence $S[1:n]$ *if for all* $1 \le i \le n$ *we have* $T[i] \perp\!\!\!\perp S[i:n] | S[1:i-1]$.

Definition 2.2. *A relaying scheme composed of a sequence of mappings* $\left(f_{relay}^{(i)}\right)_{i=1}^{\ell}$ *is said to be admissible if for all* $i \in \{1,\dots,\ell-1\}$, $X_{i+1}[1:n]$ *is causal with respect to* $Y_i[1:n]$.

This model can be understood as adding another dimension to channel coding problems. In channel coding problems, we characterized the achievable rate $R \approx m/n$, i.e., how the transmission time grows as the number of bits that we have to send increases. But now, we add another dimension ℓ, which is the number of intermediating relays. To make an analogue with transmission through a physical medium, we will refer to the number of relays as the *length* of the relaying scheme. Given the success of information theory in characterizing the achievable rate, one might ask whether a similar success can be achieved on determining the optimal trade-off between the number of channel uses (time) and the number of relays (space) that allows reliable communication.

This part of this thesis is dedicated to answering this question: what do we know about v in the relation,

$$n \approx \frac{\ell}{v} + \frac{m}{R}. \tag{2.1}$$

By analogy to physics, the problem of determining v is also known as the problem of information velocity.[1]

As in the channel coding problem, the reliability of communication in the relaying problem is measured in terms of its (average block) error probability, namely $\Pr(B[1:m] \neq \hat{B}[1:m])$.

In previous works, e.g., (Huleihel et al., 2019) (Ling and Scarlett, 2022), the problem of information velocity is commonly posed as the problem of determining the time needed to send a single bit of information such that the receiver can determine the value of the bit with probability larger than a certain fixed threshold. In this work, we will refer to this quantity as the single-bit information velocity. More generally, we can define m-bit information velocity as follows.

Definition 2.3. *For a given channel,* $W_{Y|X}$, *m-bit information velocity* v'_m *is said to be achievable, if there exists a function* $\ell(n) : \mathbb{Z} \to \mathbb{Z}$ *such that,*

i. There exists a sequence of admissible relaying schemes with length $\ell(n)$ *with the probability of error* $p_e^{(n)}$ *such that,*

$$\limsup_{n\to\infty} p_e^{(n)} = 0, \tag{2.2}$$

ii. The ratio of channel uses and the relaying scheme's length approaches v'_m, *i.e.,*

$$\limsup_{n\to\infty} \frac{\ell(n)}{n} = v'_m. \tag{2.3}$$

[1] I do wonder why it is called information velocity and not information speed, as I do not see a vectorial direction. But this might be because, as currently stated, we only consider this problem on a chain of relay instead on a more complex network of relays.

We define the m-bit information velocity of the channel, $v_m(W_{Y|X})$, as the supremum of achievable v'_m.

Currently, single-bit information velocity is the most studied form of information velocity, with some results that we will discuss in the next section. The focus of this work is on a different version of information velocity which we will refer to as the blockwise information velocity.

Definition 2.4. *A blockwise information velocity v' is said to be achievable for channel $W_{Y|X}$, if there exists a function $\ell(m,n) : \mathbb{Z} \times \mathbb{Z} \to \mathbb{Z}$ such that,*

i. There exists a sequence of admissible relaying schemes of length $\ell(m,n)$ with decaying error probability $p_e^{(n,m)}$ for all m, i.e.,

$$\limsup_{n\to\infty} p_e^{(n,m)} = 0, \tag{2.4}$$

ii. The ratio of channel uses and the relaying scheme's length approaches v, i.e.,

$$\limsup_{m\to\infty} \limsup_{n\to\infty} \frac{\ell(m,n)}{n} = v'. \tag{2.5}$$

We define the blockwise information velocity of the channel $v(W_{Y|X})$ as the supremum of achievable v'.

Note that the order of limit on condition (ii) is important, such that $1/v$ captures the growth of channel uses as the number of relays increases regardless of how many bits are being sent.

But on the other hand, m-bit information velocity allows us to bound the blockwise information velocity.

Proposition 2.1. *We have for every channel $W_{Y|X}$,*

1. *Information velocity (blockwise or m-bit) is upper bounded by,*

$$0 \le v(W_{Y|X}) \le 1, \tag{2.6}$$

2. *Information velocity is non-increasing in m, i.e. for all $m' > m$*

$$v(W_{Y|X}) \le v_{m'}(W_{Y|X}) \le v_m(W_{Y|X}) \le v_1(W_{Y|X}), \tag{2.7}$$

3. *m-bit information velocity can be lower bounded by single-bit information velocity, i.e.,*

$$\frac{v_1(W_{Y|X})}{m} \le v_m(W_{Y|X}). \tag{2.8}$$

Proof. The non-negativity of information velocity trivially follows from the definition. The upper bound is due to the causality constraints, where we need at least k channel uses such that $X_k[k]$ can start to depends on $B[1:m]$.

The second statement follows from the fact that we can use any relaying scheme which send m'-bits of information to send m-bits of information.

The third statement is based on the idea of interleaving the relaying schemes. Given an achievability scheme for single-bit information velocity, we can construct a reliability scheme for m-bit information velocity by using time indices $\{i : i \mod m = r\}$ to send the r-th bit. Due to the memorylessness of the channel, this is equivalent to having m independent relaying schemes. Let us denote the error probability of each relaying schemes as $p_{e,r}^{(n/m)}$ we have the error probability of the whole scheme as $1-\prod_{r=1}^{m}(1-p_{e,r}^{(n/m)})$. This error probability goes to 0 if $p_{e,r}^{(n/m)}$ goes to 0 as n grows to infinity. From the definition of v_1, for each bit we have that there exists $\ell(\cdot)$ which fulfills this requirement with,

$$\limsup_{n\to\infty}\frac{\ell(n/m)}{n/m} = v_1. \tag{2.9}$$

The same $\ell(\cdot)$ implies that information velocity of v_1/m is achievable. □

In light of this proposition, one might guess that the blockwise information velocity for the majority of channels is equal to 0. However, the achievability results that we will present in the subsequent chapters show that the lower bound given by this proposition is loose.

2.1 Naive Forwarding Scheme in Binary Memoryless Symmetric Channels

One common question is how this blockwise information velocity differs from the notion of channel capacity or single-bit information velocity. To illustrate this difference, in this section we will analyze the information velocity of a naive relaying scheme. In this scheme, we would simply consider the relaying problem as a series of channel coding problems, i.e., each relay waits until it receives the complete message before starting to transmit. The idea is to reuse the results that we have developed in channel coding settings for the relaying setting.

Here, we will focus on the class of Binary Memoryless Symmetric (BMS) channels.

Definition 2.5. *A channel $W_{Y[1:n]|X[1:n]}$ is a Binary Memoryless Symmetric channel if i) its input is binary, i.e., $X[1:n]$ is defined on $\{0,1\}^n$, ii) it is a memoryless channel, i.e., $W_{Y[1:n]|X[1:n]} = W_{Y|X}^{\otimes n}$, iii) it is symmetric, i.e., there exists a permutation π such that $W_{Y|X}(y|0) = W_{Y|X}(\pi(y)|1)$.*

Consider a relaying scheme of length ℓ with n channel uses. We can divide these n channel uses into several blocks of n_i channel uses such that $\sum_{i=1}^{\ell(n)} n_i$. The basic idea of this scheme is to complete the transmission between the i-th and $(i+1)$-th relay during the i-th block. Let us

start on the first relay. The first relay will read the message and use a channel coding scheme to encode the message. It will then transmit the encoded message using $X_1[1:n_1]$. The second relay will try to decode the message from $Y_1[1:n_1]$, it will then uses a coding scheme (possibly the same coding scheme as the one used in the previous relay) to transmit the message that it decodes at $X_2[n_1+1:n_2+n_1]$. Depending of the channel coding scheme that we use, there are many possible variations of this scheme, let us refer to these schemes as the class of NAIVE relaying schemes.

We can see that this relaying scheme is admissible, as for any $k<\ell$ we have a) $X_k[1:\sum_{l=1}^{k-1} n_l]$ is independent of the actual message, b) $X_k[1+\sum_{l=1}^{k-1} n_l : n]$ only depends on $Y_{k-1}[1:\sum_{l=1}^{k-1} n_l]$.

The main observation of this section is that the single-bit information velocity of this class of NAIVE relaying scheme is equal to 0. But before that, let us review the lower bound on error probability of sending 1 bit of information through BMS channels after n channel uses.

Proposition 2.2. *Consider a BMS channel $W_{Y|X}$. Let us define a Markov chain $B[1]\to X[1:n]\to Y[1:n]\to \hat{B}[1]$, with $M\sim$ Bernoulli$(1/2)$ such that,*

1. *We have $x_0[1:n]$ and $x_1[1:n]$, such that*

$$f(x)=\begin{cases} x_0[1:n] & x=0\\ x_1[1:n] & x=1\end{cases}\qquad X[1:n]=f(B[1]) \tag{2.10}$$

2. *There exists $\mathscr{D}_1$ such that,*

$$\hat{B}[1]=\mathbb{1}\{Y[1:n]\in\mathscr{D}_1\}. \tag{2.11}$$

Then we have,

$$\frac{1}{4}e^{2n\log Z(W_{Y|X})}\le \Pr(\hat{B}[1]\neq B[1]) \tag{2.12}$$

with

$$Z(W_{Y|X})=\int\sqrt{W_{Y|X}(y|0)W_{Y|X}(y|1)}\,dy. \tag{2.13}$$

Proof. This proposition is a consequence of Cauchy-Schwartz. Let us consider,

$$\begin{aligned}\frac{1}{4}Z(W_{Y|X})^2&=\frac{1}{4}\left(\int\sqrt{W_{Y|X}(y|\mathbb{1}\{y\in\mathscr{D}_1\})W_{Y|X}(y|\mathbb{1}\{y\notin\mathscr{D}_1\})}dy\right)^2\\&\le\frac{1}{4}\int W_{Y|X}(y|\mathbb{1}\{y\in\mathscr{D}_1\})dy\int W_{Y|X}(y'|\mathbb{1}\{y\notin\mathscr{D}_1\})dy'\\&=\big(1-\Pr(B[1]\neq\hat{B}[1])\big)\Pr(B[1]\neq\hat{B}[1])\\&\le\Pr(B[1]\neq\hat{B}[1]).\end{aligned} \tag{2.14}$$

We can obtain the proposition by noting that $Z(W_{Y|X}^{\otimes n})=Z(W_{Y|X})^n$. □

Proposition 2.3. *Consider a BMS channel $W_{Y|X}$. The single-bit information velocity of scheme in the class of* NAIVE *relaying schemes, $v_1^{\text{NAIVE}}(W_{Y|X})$, is equal to* 0.

Proof. We will prove by contradiction. Let us assume that there exists a length function $\ell(\cdot)$ such that $0 < v_1^{\text{NAIVE}}$. Let us consider a fixed n (and therefore a fixed $\ell(n)$).

In NAIVE relaying schemes, each relay will need to decode the message explicitly. Hence, we can refer to the message decoded by the i-th relay as M_i. To be consistent, let us denote the original message as M_0.

Furthermore, let us define the one-hop error probability of the i-th relay as the probability that the message decoded by the i-th relay differs from the message decoded by the $(i-1)$-th relay. Due to the BMS channel properties the error event does not depend on the actual value of the message. Note that this is not necessarily means that the message decoded by the i-th relay differs from the original message. Let us denote this one-hop error probability for the i-th relay as ϵ_i. We can also observe that the one-hop error events are independent.

It is easy to see that we have the following recurrence relation,

$$\begin{bmatrix} P_{M_{i+1}}(0) \\ P_{M_{i+1}}(1) \end{bmatrix} = \begin{bmatrix} 1-\epsilon_i & \epsilon_i \\ \epsilon_i & 1-\epsilon_i \end{bmatrix} \begin{bmatrix} P_{M_i}(0) \\ P_{M_i}(1) \end{bmatrix} \tag{2.15}$$

The eigenvector of this stochastic matrix is given by $(1, 1-2\epsilon_i)$ with the eigenvector is given by $\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \end{bmatrix}^T$ and $\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & -1 \end{bmatrix}^T$ respectively. Without loss of generality, let us assume that $M_0 = 0$. Given the starting probability $P_{M_0}(0) = 1$ and $P_{M_0}(1) = 0$, the error probability of the whole scheme is given by,

$$P_{M_{\ell(n)}}(1) = \frac{1 - \prod_{i=1}^{\ell(n)}(1-2\epsilon_i)}{2} \tag{2.16}$$

by using the standard inequality $(1-x) \leq e^{-x}$, we have

$$P_{M_{\ell(n)}}(1) \geq \frac{1}{2} - \frac{\exp\left(-2\sum_{i=1}^{\ell(n)} \epsilon_i\right)}{2}. \tag{2.17}$$

Let us consider the exponent term, given the lower bound to error probability that we have proved, we have that,

$$\sum_{i=1}^{\ell(n)} \epsilon_i \geq \frac{1}{4} \sum_{i=1}^{\ell(n)} e^{2n_i \log Z(W_{Y|X})} \geq n e^{\frac{2n}{\ell(n)} \log Z(W_{Y|X})} = n e^{\frac{2 \log Z(W_{Y|X})}{v_1^{\text{NAIVE}}}} \tag{2.18}$$

where the second inequality is due to AM-GM inequality. We can also see that as $n \to \infty$ then the lower bound on the exponent will also grows to infinity. This implies that

$$\lim_{n\to\infty} P_{M_{(\ell(n))}}(1) \geq \frac{1}{2}. \tag{2.19}$$

This contradicts the fact that $\ell(n)$ is a valid function to establish a strictly positive information velocity v_1^{NAIVE}. Hence, the proposition. □

This proposition combined with the fact that single-bit information velocity is an upper bound to the blockwise information velocity gives us the following corollary.

Corollary 2.1. *The blockwise information velocity of the naive scheme, v^{NAIVE}, fulfills*

$$v^{\text{NAIVE}} = 0. \tag{2.20}$$

This example gives us a reason to believe that a relaying scheme with positive information velocity will need to either be "adaptive", in the sense that the time taken by each hop will need to depend on the actual realization of the channel, or be able to start transmitting before deciding on its final value of the message. This insight also provides the rationale why we study the tree codes in chapter 4.

Furthermore, this class of schemes will not be able to operate at positive blockwise information velocity, even if we disregard the problem with its error probability. Let us say that we want to transmit m-bits using this scheme; at the very least we require that each block length n_i is $\approx mC(W_{Y|X})$. If we want to transmit through a relay chain of length k, then we would have $n \approx kmC(W_{Y|X})$. This implies that the inner limit in point ii in definition 2.4 is equal to $1/mC(W_{Y|X})$. Taking the outer limit implies that the blockwise information velocity is equal to 0, even though the m-bit information velocity is strictly positive.

2.2 Single-Bit Information Velocity for BEC

Considering the previous example, doubts may arise about whether there exists a relaying scheme in which we can analytically prove a positive information velocity. We will present an example in which a positive single-bit information velocity can be easily established.

In this scheme, the relay mapping functions behave as follows,

- At the beginning of the chain, we have $X_1[k] = B[1]$ for all $1 \le k \le n$.
- For the subsequent relays, the relay will default to guessing that the message is 0. However, once the relay observes 1 on its input, it will change its guess to 1. At any time, the relay will transmit its guess. More formally, we have

  $$X_j[k] = \mathbb{1}\{\exists_{i<k} Y_{j-1}[i] = 1\} \tag{2.21}$$

  for all $0 \le k \le n$ and $0 \le j \le \ell$.
- The decoder at the end of the chain will emit its guess.

It is easy to see that this scheme is admissible since each relay's guess at time i depends only on its input before time i.

An interesting feature of this scheme is that the error probability depends on the actual message. In the sense that the error probability conditioned on $B[1] = 0$ is exactly 0, while the error probability is not zero if we condition on $B[1] = 1$.

We can show that this scheme has a positive information velocity if the BEC erasure probability $\epsilon < 1$.

Proposition 2.4. *The single-bit information velocity of BEC relay is lower bounded as follows,*

$$v_1\left(BEC(\epsilon)\right) \geq 1 - \epsilon. \tag{2.22}$$

Proof. Here we can observe an asymmetry in the error probability. As all errors in BEC are detected errors, the final relay can be certain that $M = 1$ if it has received an unerased 1. However, if it has only received an unerased 0, it is either because the correct message is indeed 0, or it is due to the unerased 1 not being propagated to the final relay.

To study the error probability of this scheme, it is sufficient to characterize the propagation times of the unerased message. Let us define a sequence of times, $T_1, T_2, T_3, \ldots$, such that

$$T_i = \min\{t : t > T_{i-1}, Y_i[t] \neq \epsilon\} \tag{2.23}$$

where we define $T_0 = 0$. In other words, T_i is the time where the i-th relay can be said to be sure of its reception, i.e., for all $t > T_i$, we have $X_{i+1}[t] = M$. We must emphasize that if $M = 0$, the i-th relay cannot determine T_i with its local information.

Due to the memorylessness of the channel, we have that $T_i - T_{i-1} \sim \text{Geometric}(1-\epsilon)$, and $T_i - T_{i-1} \perp T_{i-1}$. Hence, by considering the telescopic sum

$$T_i = \sum_{k=1}^{i} T_k - T_{k-1}, \tag{2.24}$$

we have that

1. By linearity of expectation, we have $E[T_k] = k\frac{1}{1-\epsilon}$,

2. As the terms in the telescopic sum are i.i.d. random variables, the empirical average T_i/i is subject to concentration of measure, for example Proposition 1.12. Consequently, we have,
$$\Pr\left(T_k - \frac{k}{1-\epsilon} > k^\gamma\right) \to 0 \tag{2.25}$$
where $\gamma > 1/2$.

Hence, in this case, let us consider $\ell(n) = (1-\epsilon)n - O(n^{1/2+\delta})$ for $\delta > 0$. It fulfills the requirements to establish $(1-\epsilon)$ as an achievable single-bit information velocity, namely

1. We will have an error if $T_{\ell(n)} > n$. We have that

$$\begin{aligned} p_e^{(n)} &= \Pr\left(T_{\ell(n)} - n > 0\right) \\ &= \Pr\left(T_{\ell(n)} - \frac{\ell(n)}{1-\epsilon} > O(n^{1/2+\delta})\right) \end{aligned} \tag{2.26}$$

   which vanishes due to the concentration.

2. We have the ratio between channel uses and the length of the relaying scheme, i.e.,

$$\limsup_{n\to\infty} \frac{\ell(n)}{n} = 1-\epsilon. \tag{2.27}$$

Hence this scheme provides us with a lower bound on the information velocity. □

On the next chapter, we will show that this lower bound is in fact an equality for single-bit information velocity by using the converse result that we will develop.

Unfortunately, this scheme cannot be used to establish a positive blockwise information velocity for BEC. As we can see, this scheme fundamentally exploits the fact that a) all error in BEC is detected errors (as opposed to undetected errors, which is more common to other channel models), and b) the message set only contains two possible messages.

Using the argument in proposition 2.1, this scheme also establishes the positivity of information velocity for any m-bit information velocity. However, this method does not imply the positivity of the blockwise information velocity.

As we can see in the proof of the previous proposition, the time that is taken for the message to traverse a single hop is not fixed. The transmission time taken by each hop will depends on the actual realization of the erasures. In a sense, this scheme overcomes the result that we showed in the previous section as the scheme adaptively lengthen the transmission time in the face of more erasures. As our analysis shows, we rely on the fact that the length of these transmission times is independent, and therefore the empirical average of these transmission times exhibit concentration to the expected value. In this regards, the fact that the deviation from the long-time behavior is in the order of $o(k)$ is essential to our argument. We will see the same pattern in the later chapter where we developed a scheme with positive blockwise information velocity for BEC relaying problems.

2.3 Related Works

To our knowledge, the first work that attempts to quantify the relation between the “spatial” dimension in a network and the number of channel uses is (Rajagopalan and Schulman, 1994). In this work, the authors attempted to study the consequences of noisy communication channels in the context of distributed computation. The inherent unreliability of communication channels contradicts the commonly used abstract model of computation (i.e., Turing machine, distributed state machine, etc.). To bridge this gap, the authors introduced a communication scheme based on tree codes, allowing diverse computation units to agree on the system's state, assuming feedback exists between each computation unit.

In their work, they also introduced an ad-hoc scheme which they refer to as "Broadcasting on a chain", we will discuss this scheme in the appendix as it represents another important line of thought in designing a relaying scheme.

But their main concern is mainly in solving the problem of distributed computing, which explains the simplicity of their discussion of trade-off between time and spatial dimension. In contrast to their ad-hoc scheme for relaying, the scheme that we will develop in chapter 4, can be understood as an extension of their idea of using tree code to simulate a tape abstraction (in our case we use this tape to track the transmission of message bits), which surprisingly achieves positive information velocity in BEC relaying problems[2].

The work that can be said to revive the interest in the question of information velocity is (Huleihel et al., 2019). The authors studied the problem of tandem relaying for BSC with finite number of relays. The setting in their paper, where the number of relays is fixed, is characteristically different than our asymptotic formulation of the problem. In fact, this the first paper which presents a conjecture on the nature of (single-bit) information velocity, namely that information velocity is equal to the channel capacity. Although this relation holds for BEC, it turns out that the conjecture is incorrect for other channels.

The subsequent progress is on the asymptotic formulation of this problem, which is given in the work of (Domanovitz et al., 2022). The author studied the problem of relaying on packet erasure channels with feedback, which the analysis is closely similar to BEC with feedback. In this paper, we first see the treatment of information velocity problems in terms of the ratio between the number of relays and the number of channel uses. Another innovation in this paper is the use of the notion of streaming sources, where the packets arrive with geometric interarrival, as opposed to previous study which considers a transmission of a single bit.

Another notable work is the line of research given in (Ling and Scarlett, 2022). In this paper, the authors presented a simple relaying scheme which achieves positive information velocity by modifying the "Broadcasting on a chain" scheme presented in (Rajagopalan and Schulman,

[2] In fact, I wonder why they choose to define an ad-hoc scheme instead of conducting a more careful error analysis of their tree codes scheme. But then again, maybe their focus on distributed computing precludes this consideration.

1994). In our view, their main insight is on the use of more powerful error correction codes, i.e., Maximum Distance Separation codes (namely Hamming codes), as opposed to the abstracted channel coding scheme used in the original scheme. They also optimize the parameter of the original scheme to achieve a probability of error upper bounded by a constant (although not decaying, hence does not fulfill our definition of information velocity), with the number of channel uses scaling linearly with the number of relays.

We would be remiss if we did not mention a recent result by (Domanovitz et al., 2023) which independently achieves the same result as ours regarding the information velocity for AWGN channels with feedback. We will elaborate on our presentation of this scheme later on chapter 5.

2.4 Appendix : Rajagopalan's Relaying Scheme

The scheme proposed by Rajagopalan is defined recursively. The first iteration of this scheme, which we will refer to level 1 scheme, is designed for two relays (a single hop) to send a single bit by repeating the message bit R times. We will keep track of three parameters of the i-th iteration of the scheme (as a shorthand we will refer this as the i-th scheme), namely

- The number of hop traversed by the scheme k_i,
- The number of bits that it sends m_i,
- The time elapsed between the transmission by the first relay to the reception by the final relay n_i.

Hence, we have $k_1 = 1$, $m_1 = 1$, and $n_i = R$.

Given the i-th scheme with parameter (k_i, m_i, n_i), the $i+1$-th scheme is designed to operate at parameter $(4k_i, 2^{m_i}, (3+\alpha 2^{m_i}/m_i)n_i)$, namely,

1. Given 2^{m_i} bits that we want to send, we divide it into $2^{m_i}/m_i$ blocks.
2. We apply channel coding to correct for $1/4$ error. Let us denote the rate of this code by $1/\alpha$ hence we have $\alpha 2^{m_i}/m_i$ coded blocks.
3. We sent these blocks by using the i-th scheme 4 times. Hence, we can send these blocks through $4k_i$ hops. Note that we can send blocks in streaming manner; i.e., after receiving the first i-th scheme's block, the endpoint of the i-th scheme can start running the i-th scheme for the next hop while simultaneously receiving the subsequent i-th scheme's block from the previous hop. Hence, the time spent by this relaying scheme is equal to

$$3n_i + n_i\alpha \times 2^{m_i}/m_i. \tag{2.28}$$

First, let us analyze the error probability of this scheme. Let us denote the error probability of the i-th scheme as q_i. By the design of our channel code, the error event of the $(i+1)$-th scheme can be upper bounded by the event where a quarter of i-th scheme blocks are in error. The error probability of each block can be upper bounded by union bound to be equal to $q_i 2^{m_i}/{m_i}^3$. Hence we have by Chernoff bounds for Bernoulli random variables,

$$\begin{aligned} q_{i+1} &\leq e^{-\alpha d_2(1/4;4q_i)2^{m_i}/m_i} \\ &\leq e^{-2\alpha d_2(1/4;4q_i)} \end{aligned} \tag{2.29}$$

Regardless of q_i, we can increase α, i.e., reduce the channel code's rate, such that $q_{i+1} \leq q_i$. Hence, we can achieve reliable communication using this scheme.

We can also study the scaling of number of hops and the number of bits transmitted, where we can see that to design a relaying scheme for ℓ hops and m bits, then we need to repeat the iteration schemes i^* times with $i^* = \max(\log_2 \ell, \log_2^*(m))$. Finally, we can see that the number of channel uses for the i-th scheme can be upper bounded by $(2\alpha)^i \exp_2^*(i)$ where $\exp_2^*(i)$ is defined as $\exp_2^*(0) = 1$ and $\exp_2^*(i+1) = 2^{\exp_2^*(i+1)}$. Assuming that $\log_2^*(m) > \log_2 \ell$, we have the number of channel uses is equal $m(2\alpha)^{\log_2^*(m)}$. We can see that this scheme also fails to achieve positive information velocity, as the information velocity of this scheme is equal to $4^{i^*}/2\alpha^{i^*} \exp^*(i^*)$ which vanishes as $i^* \to \infty$.

[3]The original analysis in (Rajagopalan and Schulman, 1994) contains an error whereas the error probability of the block at the $i+1$-th scheme is mistakenly replaced with the error probability of a single application of the i-th scheme, instead of being properly upper bounded using union bound. This also explains why our parameter here deviates from the original analysis to make the whole scheme work.

3 A Converse for Blockwise Information Velocity

In the time-honored tradition of information theory, we will start our study of information velocity by gaining insights into the fundamental limit that can be achieved by arbitrary relaying scheme. The argument for our converse for information velocity is driven by the intuition that the amount of information conveyed in a single channel use is limited. Moreover, we can upper bound the amount of information that a relay can send as a fixed proportion of the difference of information between the relay's observations and the subsequent relay's observations. This restriction introduces a "transport equation"-like structure, dictating the speed of information propagation through the relay chain.

In this chapter, we will begin by discussing F-divergences, with the main focus on defining the objects used for our converse result and on establishing connections between F-divergences and the decoding error of communication schemes. Afterwards, we will delve into the notion of information transfer curves, that characterizes the amount of information that can be propagated at each channel use as a function of the difference in information between the transmitter and the receiver in a single hop. We will also review the concept of channel contraction coefficient which can be understood as the maximum slope of the information transfer curve.

We will then use the local bound that we have developed for a single channel use between two relays in a single hop to develop a bound on the information velocity of the entire relay chain. Originally, an early form of this method is presented in (Rajagopalan and Schulman, 1994)[1]. Our main contribution lies in connecting the method with the notion of F-divergences, hence allowing the method to be used with greater class of information measures. Through this connection, we demonstrate that we can analytically derive an upper bound on information velocity for BMS channels. For more general channels, the upper bound on information velocity will depend on our ability to evaluate the channel contraction coefficient.

[1] Actually, we developed this argument as an offshoot of the achievability argument for AWGN channels with feedback, which we will discuss in chapter 5. Later on, we realized that the same technique can be used to prove a converse result.

3.1 F-Divergences

The notion of F-divergences generalizes the notion of KL divergence while maintaining the desirable property of being subject to a data processing inequality.

Definition 3.1. *Given a convex function $f : \mathbb{R}_+ \to \mathbb{R}$ with $f(1) = 0$, we define*

$$D_f(Q;P) = \mathbb{E}_P\left[f\left(\frac{dQ}{dP}\right)\right] \tag{3.1}$$

for any Q and P are probability distributions satisfying $P \ll Q$.

We will refer to $D_f(\cdot;\cdot)$ as the F-divergence induced by the convex function f.

Three particular F-divergences of interest are:

- KL divergence, where $f(x) = x\log x$,
- TV divergence, where $f(x) = \frac{1}{2}|x-1|$,
- χ^2 divergence, where $f(x) = (x-1)^2$.

We can prove several properties of F-divergences.

Proposition 3.1. *We have,*

1. $D_f(Q;P) \geq 0$ *with equality if* $Q = P$,
2. $D_f(Q_X P_{Y|X}; P_X P_{Y|X}) = D_f(Q_X; P_X)$,
3. $D_f(Q_{X,Y}; P_{X,Y}) \geq D_f(Q_X; P_X)$.
4. $D_f(Q_X; P_X) \geq D_f(Q_Y; P_Y)$ *if* $Q_Y = W_{Y|X} \circ Q_X$ *and* $P_Y = W_{Y|X} \circ P_X$ *(Data Processing Inequality).*

Proof. The second property is obvious from the definition. The first and third properties are consequence of Jensen's inequality. We have

$$\begin{aligned} D_f(Q;P) &= \mathbb{E}_P\left[f\left(\frac{dQ}{dP}\right)\right] \\ &\geq f\left(\mathbb{E}_P\left[\frac{dQ}{dP}\right]\right) \\ &= f(1) \\ &= 0 \end{aligned} \tag{3.2}$$

and,

$$\begin{aligned} D_f\left(Q_{X,Y};P_{X,Y}\right) &= \mathbb{E}_{P_X}\left[\mathbb{E}_{P_{Y|X}}\left[f\left(\frac{Q_{X,Y}}{P_{X,Y}}\right)\right]\right] \\ &\geq \mathbb{E}_{P_X}\left[f\left(\mathbb{E}_{P_{Y|X}}\left[\frac{Q_{X,Y}}{P_{X,Y}}\right]\right)\right] \\ &= \mathbb{E}_{P_X}\left[f\left(\mathbb{E}_{Q_{Y|X}}\left[\frac{Q_X}{P_X}\right]\right)\right] \\ &= \mathbb{E}_{P_X}\left[f\left(\frac{Q_X}{P_X}\right)\right] \\ &= D_f(Q_X;P_X). \end{aligned} \tag{3.3}$$

The last property can be shown by combining the second and the third property, namely,

$$\begin{aligned} D_f(Q_X;P_X) &= D_f(Q_X W_{Y|X};P_X W_{Y|X}) \\ &\geq D_f(Q_Y;P_Y). \end{aligned} \tag{3.4}$$

□

Although all F-divergences roughly measure the difficulty of distinguishing two distributions, it is not obvious how to transfer an estimate on the divergence induced by convex function f to an estimate on the divergence induced by another convex function g, for example, how to find an estimate of total variation distance (which is operationally important) from an estimate on KL divergence. The reader is suggested to consult (Harremoës and Vajda, 2010) for a general method of solving this problem, and (Sason and Verdú, 2016) for a compendium of results on F-divergences inequalities.

However, for our purpose, we are more interested in a qualitative result which asserts that F-divergences tending to 0 implies that the distribution is hard to distinguish in a binary hypothesis testing setting.

Proposition 3.2. *Let f be a convex function which is thrice differentiable on 1 and $f''(1) > 0$. Consider a family of binary distributions $\left\{Q_X^{(n)}\right\}$ and a family of binary distributions $\left\{P_X^{(n)}\right\}$, then we have*

$$\lim_{n\to\infty} D_f\left(Q_X^{(n)};P_X^{(n)}\right) = 0 \tag{3.5}$$

if and only if

$$\lim_{n\to\infty} TV(Q_X^{(n)},P_X^{(n)}) = 0. \tag{3.6}$$

Proof. Note that for any convex function f we can consider a modified function

$$\tilde{f}(x) = f(x) - f'(1)(x-1), \tag{3.7}$$

such that $\tilde{f}(x) \geq 0$ for all x, and $\tilde{f}'(1) = 0$. We also have that

$$\begin{aligned} D_{\tilde{f}}(Q_X; P_X) &= \mathbb{E}_{P_X}\left[f\left(\frac{dQ_X}{dP_X}\right)\right] - f'(1)\mathbb{E}_{P_X}\left[\frac{dQ_X}{dP_X} - 1\right] \\ &= \mathbb{E}_{P_X}\left[f\left(\frac{dQ_X}{dP_X}\right)\right] \\ &= D_f(Q_X; P_X). \end{aligned} \tag{3.8}$$

As we are considering binary distribution, let us consider the edge case where $P_X(0)$ or $P_X(1)$ is equal to 0. Without loss of generality, let us assume that $P_X(0) = 0$. For this edge case, we have that $D_{\tilde{f}}(Q_X; P_X) = \tilde{f}(Q_X(1))$, and $TV(Q_X, P_X) = 2(1 - Q_X(1))$. In other words, $D_{\tilde{f}}(Q_X, P_X) = \tilde{f}\,(1 - TV(Q_X, P_X)/2)$. We also know that due to the hypothesis, we have that $\tilde{f}(x)$ is continuous at $x = 1$. This implies the statement of the proposition.

Given the assumption on f, the set of binary probability such that $D_{\tilde{f}}(Q_X; P_X) \leq \epsilon$ is compact and due to Taylor's theorem there exists a constant K_ϵ with $\frac{1}{2}f''(1) < K_\epsilon < f''(1)$ for small enough ϵ such that,

$$\begin{aligned} D_{\tilde{f}}(Q_X; P_X) \geq& \tilde{f}(1) + \tilde{f}'(1)\mathbb{E}_{P_X}\left[\frac{Q_X(X)}{P_X(X)} - 1\right] + \frac{K_\epsilon}{2}\mathbb{E}_{P_X}\left[\frac{(P_X(X) - Q_X(X))^2}{P_X(X)^2}\right] \\ \geq& \frac{K_\epsilon}{2}\mathbb{E}_{P_X}\left[\frac{|P_X(X) - Q_X(X)|}{P_X(X)}\right]^2 \\ \geq& \frac{K_\epsilon}{2}\left(\sum_{x\in\mathcal{X}} |P_X(x) - Q_X(x)|\right)^2 \\ =& 2K_\epsilon TV(P_X, Q_X)^2 \end{aligned} \tag{3.9}$$

which implies

$$TV(Q_X, P_X) \leq \sqrt{\frac{1}{f''(1)} D_{\tilde{f}}(Q_X; P_X)}. \tag{3.10}$$

This implies that if

$$\lim_{n\to\infty} D_f\left(Q_X^{(n)}; P_X^{(n)}\right) = 0 \tag{3.11}$$

then

$$\lim_{n\to\infty} TV(Q_X^{(n)}, P_X^{(n)}) = 0. \tag{3.12}$$

On the other hand, we have that,

$$\begin{aligned} D_{\tilde{f}}(Q_X;P_X) &\leq \mathbb{E}_P\left[\tilde{f}\left(1-\frac{P_X-Q_X}{P_X}\right)\right] \\ &\leq \mathbb{E}_P\left[\max\left\{\tilde{f}\left(1-\frac{|P_X-Q_X|}{P_X}\right),\tilde{f}\left(1+\frac{|P_X-Q_X|}{P_X}\right)\right\}\right] \\ &\leq \mathbb{E}_P\left[\max\left\{\tilde{f}\left(1-\frac{TV(P_X,Q_X)}{P_X}\right),\tilde{f}\left(1+\frac{TV(P_X,Q_X)}{P_X}\right)\right\}\right] \\ &\leq \max\left\{\tilde{f}\left(1-\frac{TV(P_X,Q_X)}{\min_x P_X(x)}\right),\tilde{f}\left(1+\frac{TV(P_X,Q_X)}{\min_x P_X(x)}\right)\right\} \end{aligned} \tag{3.13}$$

This implies the other direction by continuity, i.e.,

$$\lim_{n\to\infty} TV(Q_X^{(n)},P_X^{(n)}) = 0 \tag{3.14}$$

then

$$\lim_{n\to\infty} D_f\left(Q_X^{(n)};P_X^{(n)}\right) = 0. \tag{3.15}$$

□

Fortunately, most of the interesting F-divergences fulfill the condition of this proposition . For example, we have for KL divergence, $f''(1)=1$, and for χ^2 divergence, $f''(1)=2$.

We will also define the analogue of mutual information for F-divergences as follows.

Definition 3.2. *Given a convex function f, we define the F-information between random variable X and observation Y as,*

$$I_f(Q_X;W_{Y|X}) = \mathbb{E}_{Q_X}\left[D_f\left(W_{Y|X=X};Q_Y\right)\right]. \tag{3.16}$$

We note that there are many alternatives for extending mutual information in the F-divergences framework; see (Csiszár, 1974). The definition used here is the definition proposed in (Csiszár, 1967)[2]. This specific definition is suitable for our purpose, as we can show results with similar flavor to i) data processing inequality, ii) hypothesis testing converse.

Proposition 3.3. *We have,*

$$I_f(Q_X;W_{Y|X}) \geq 0. \tag{3.17}$$

Proof. This obvious due to the non-negativity of F-divergences. □

[2]Please be cautious as the notation used in that paper differs from the notation of this thesis. They used I_f to denote F-divergences, instead of D_f that we used in this work. However, we agree with the sentiment expressed in the paper that emphasizes the operational characteristic of the proposed generalization of mutual information (in our case: its ability to produce results similar to the data processing inequality).

Proposition 3.4. *Given $X \to Y \to Z$ we have,*

$$I_f(Q_X; W_{Y|X}) \geq I_f(Q_X; W_{Z|X}). \tag{3.18}$$

Proof. This proposition follows from the data processing inequality for F-divergences. We have,

$$\begin{aligned} I_f(Q_X; W_{Y|X}) &= \mathbb{E}_{Q_X}\left[D_f\left(W_{Y|X=X}; Q_Y\right)\right] \\ &\geq \mathbb{E}_{Q_X}\left[D_f\left(W_{Z|X=X}; Q_Z\right)\right] \\ &= I_f(Q_X; W_{Z|X}). \end{aligned} \tag{3.19}$$

□

The following proposition justifies the use of F-information to study the qualitative behavior of a hypothesis testing scheme (and consequently, the feasibility of reliable communication).

Proposition 3.5. *Let f be a convex function which fulfills the conditions of Proposition 3.2. Consider a discrete distribution Q_X with the size of support is at least 2. Let us consider a sequence of channels $W^{(n)}_{Y|X}$ and an optimal decoding function with conditional distribution sequence $Q^{(n)}_{\tilde{X}|Y}$ such that $\mathbb{E}_{Q^{(n)}_{\tilde{X}|Y} W^{(n)}_{Y|X} Q_X}\left[\mathbb{1}\{\tilde{X} \neq X\}\right]$ is minimized. If*

$$\lim_{n\to\infty} I_f\left(Q_X; W^{(n)}_{Y|X}\right) \to 0 \tag{3.20}$$

then

$$\lim_{n\to\infty} \mathbb{E}_{Q^{(n)}_{\tilde{X}|Y} W^{(n)}_{Y|X} Q_X}\left[\mathbb{1}\{\tilde{X} \neq X\}\right] \geq \frac{1 - \max_x Q_X(x)}{2}. \tag{3.21}$$

Proof. The condition of the proposition implies that there exists $n(\epsilon)$ such that

$$I_f(Q_X; W^{(n)}_{Y|X}) \leq \epsilon \tag{3.22}$$

for all $n \geq n(\epsilon)$. Consider the following set

$$\mathscr{E} = \left\{x : D_f(W^{(n)}_{Y|X=x}; Q^{(n)}_Y) \leq \sqrt{\epsilon}\right\}. \tag{3.23}$$

Then the definition of $I_f(\cdot,\cdot)$ implies that we have

$$Q_X(\mathscr{E}) \geq 1 - \sqrt{\epsilon} \text{ and } Q_X(\mathscr{E}^c) < \sqrt{\epsilon}. \tag{3.24}$$

Hence the probability of set $\mathscr{E}$ is close to 1 under the hypothesis as n goes infinity. Let us define the event $J_{x'}$ as the event where the decoder output x', i.e., $J_{x'} = \mathbb{1}\{\tilde{X} = x'\}$. We have

$$D_f(W^{(n)}_{Y|X=x}; Q^{(n)}_Y) = D_f(Q^{(n)}_{\tilde{X}|Y} W^{(n)}_{Y|X=x}; Q^{(n)}_{\tilde{X}Y}) \geq D_f(Q^{(n)}_{J_{\tilde{x}}|X=x}; Q^{(n)}_{J_{\tilde{x}}}) \tag{3.25}$$

which imply that, for all $x \in \mathscr{E}$ and any $\tilde{x}$, we have

$$D_f(Q^{(n)}_{J_{\tilde{x}}|X=x}; Q^{(n)}_{J_{\tilde{x}}}) \leq \sqrt{\epsilon}. \tag{3.26}$$

Let us denote the error probability, $\mathbb{E}_{Q^{(n)}_{\tilde{X}|Y} W^{(n)}_{Y|X} Q_X}[\mathbb{1}\{\tilde{X} \neq X\}]$ as P_e, we have

$$\begin{aligned}
1 - P_e &= \sum_x Q_X(x) Q^{(n)}_{J_X|X=}(1) \\
&= \sum_{x \in \mathscr{E}} Q_X(x) Q^{(n)}_{J_X|X=}(1) + \sum_{x \notin \mathscr{E}} Q_X(x) Q^{(n)}_{J_x|X=x}(1) \\
&\leq \sum_{x \in \mathscr{E}} Q_X(x) Q^{(n)}_{J_x|X=x}(1) + \sqrt{\epsilon} \\
&\leq Q_X(x') Q^{(n)}_{J_{x'}|X=x'}(1) + \sum_{x \in \mathscr{E}, x \neq x'} Q_X(x) Q^{(n)}_{J_x|X=x}(1) + \sqrt{\epsilon}
\end{aligned} \tag{3.27}$$

for any $x' \in \mathscr{E}$. We then have,

$$\begin{aligned}
1 - P_e &\leq Q_X(x') Q^{(n)}_{J_{x'}|X=x'}(1) + \sum_{x \in \mathscr{E}, x \neq x'} Q_X(x) \left(Q^{(n)}_{J_{x'}|X=x}(1) + \max_{x, x' \in \mathscr{E}} TV\left(Q^{(n)}_{J_{x'}|X=x}, Q^{(n)}_{J_{x'}|X=x} \right) \right) + \sqrt{\epsilon} \\
&\leq Q_X(x') Q^{(n)}_{J_{x'}|X=x'}(1) + \sum_{x \in \mathscr{E}, x \neq x'} Q_X(x) Q^{(n)}_{J_{x'}|X=x}(1) + \max_{x, x' \in \mathscr{E}} TV\left(Q^{(n)}_{J_{x'}|X=x}, Q^{(n)}_{J_{x'}|X=x} \right) + \sqrt{\epsilon} \\
&\leq Q_X(x') Q^{(n)}_{J_{x'}|X=x'}(1) + P_E + \max_{x, x' \in \mathscr{E}} TV\left(Q^{(n)}_{J_{x'}|X=x}, Q^{(n)}_{J_{x'}|X=x} \right) + \sqrt{\epsilon} \\
&\leq Q_X(x') Q^{(n)}_{J_{x'}|X=x'}(1) + P_E + 2 \max_{x' \in \mathscr{E}} TV\left(Q^{(n)}_{J_{x'}|X=x}, Q^{(n)}_{J_{x'}} \right) + \sqrt{\epsilon} \\
&\leq \max_{x'} Q_X(x') + P_E + 2 \max_{x' \in \mathscr{E}} TV\left(Q^{(n)}_{J_{x'}|X=x}, Q^{(n)}_{J_{x'}} \right) + \sqrt{\epsilon}
\end{aligned} \tag{3.28}$$

which implies that,

$$\frac{1 - \max_x Q_X(x) - 2 \max_{x' \in \mathscr{E}} TV\left(Q^{(n)}_{J_{x'}|X=x}, Q^{(n)}_{J_{x'}} \right) - \sqrt{\epsilon}}{2} \leq P_e. \tag{3.29}$$

Taking the limit, we have $\epsilon \to 0$ and the maximization terms goes to 0 due to the Proposition 3.2. □

The previous proposition allows us to assert that the error probability of any communication scheme is strictly bounded away from 0 if the F-information between the message and decoder's observations vanishes.

Finally, we will introduce the concept of α-tilted F-divergence.

Definition 3.3. *Let f be a convex function. We define the α-tilted F-divergence as,*

$$D^{(\alpha)}_f(Q_X; P_X) = \frac{\mathbb{E}_P\left[f\left(\alpha \frac{dQ_X}{dP_X} \right) - f(\alpha) \right]}{\alpha} \tag{3.30}$$

for $\alpha > 0$.

Note that α-tilted F-divergence is a valid F-divergence as it is a perspective transformation on the convex function f. It is also obvious that if $\alpha = 1$, the F-tilted divergence reduces to the original F-divergence.

For example, we have the tilted version of the KL divergence and the χ^2 divergence as,

- KL Divergence,

$$\begin{aligned} D_{KL}^{\alpha}(Q_X;P_X) &= \frac{\mathbb{E}_P\left[\alpha\frac{dQ_X}{dP_X}\log\left(\alpha\frac{dQ_X}{dP_X}\right)\right]-\alpha\log\alpha}{\alpha} \\ &= \mathbb{E}_Q\left[\log\left(\alpha\frac{dQ_X}{dP_X}\right)\right]-\log\alpha \\ &= \mathbb{E}_Q\left[\log\left(\frac{dQ_X}{dP_X}\right)\right] \\ &= D_{KL}(Q_X,P_X) \end{aligned} \tag{3.31}$$

- χ^2 Divergence,

$$\begin{aligned} \chi^{2(\alpha)}(Q_X;P_X) &= \frac{\mathbb{E}_Q\left[\left(\alpha\frac{dQ_X}{dP_X}-1\right)^2\right]-(\alpha-1)^2}{\alpha} \\ &= \mathbb{E}_Q\left[\alpha\left(\frac{dQ_X}{dP_X}\right)^2-2\left(\frac{dQ_X}{dP_X}\right)\right]-(\alpha-2) \\ &= \alpha\mathbb{E}_Q\left[\left(\frac{dQ_X}{dP_X}\right)^2-1\right] \\ &= \alpha\mathbb{E}_Q\left[\left(\frac{dQ_X}{dP_X}\right)^2-2\frac{dQ_X}{dP_X}+1\right] \\ &= \alpha\mathbb{E}_Q\left[\left(\frac{dQ_X}{dP_X}-1\right)^2\right] \\ &= \alpha\chi^2(Q_X,P_X) \end{aligned} \tag{3.32}$$

In essence, the tilting operation changes the "center" of the convex function that induced the F-divergence from 1 to α.

3.2 Channel Contraction Coefficients and Information Transfer Curves

For the purpose of determining the information velocity, the most crucial characteristic of a channel is the amount of information that can be transferred to the receiver given the amount of information available at the sender. More formally, we define an information transfer curve

of the channel[3] as follows,

Definition 3.4. *For a channel $W_{Y|X}$, the information transfer curve for F-divergence induced by a convex function f is defined as,*

$$F_{I_f}(\gamma; W_{Y|X}) = \sup_{Q_{M,X}: I_f(Q_M, Q_{X|M}) \le \gamma} I_f(Q_M, Q_{Y|M}) \tag{3.33}$$

where $Q_{Y|M} = W_{Y|X} \circ Q_{X|M}$.

Information transfer curves fulfill the following properties.

Proposition 3.6. *For any channel $W_{Y|X}$,*

1. *We have,*
$$0 \le F_{I_f}(\gamma; W_{Y|X}) \le \gamma \tag{3.34}$$
2. *For $\gamma > 0$ we have $F_{I_f}(\gamma; W_{Y|X})/\gamma$ is a decreasing function in x. In other words, $\sup_x F_{I_f}(x; W_{Y|X})/x = \lim_{x\to 0^+} \frac{dF_{I_f}(x; W_{Y|X})}{dx}$.*

Proof. 1. The lower bound is due to the non-negativity of F-information. The upper bound is a consequence of data processing inequality.

2. We only need to show that for $0 < \gamma < x$, we have
$$\frac{\gamma}{x} F_{I_f}(x; W_{Y|X}) \le F_{I_f}(\gamma; W_{Y|X}). \tag{3.35}$$
Consider the sequence $Q^{(n)}_{X,M}$ which achieves $F_{I_f}(x; W_{Y|X})$. We consider a modified sequence
$$\tilde{Q}^{(n)}_{X,A,M} = \begin{cases} \lambda Q^{(n)}_{X,M} & \text{if } A = 1 \\ (1-\lambda)\, Q^{(n)}_X Q^{(n)}_M, & \text{if } A = 0. \end{cases} \tag{3.36}$$
Note that we have augmented M with A, such that M is decoupled from X if $A = 0$, otherwise, it is distributed according to $Q^{(n)}_{X,M}$. As we augment M, the Markov chain that we should study is $(M, A) \to X \to Y$. We have,
$$I_f(\tilde{Q}^{(n)}_{A,M}; \tilde{Q}^{(n)}_{X|A,M}) = \mathbb{E}_{A,M}\left[D_f(\tilde{Q}_{X|A,M=A,M}; \tilde{Q}_X)\right] = \lambda I_f(Q^{(n)}_M; Q^{(n)}_{X|M}) \tag{3.37}$$
and
$$I_f(\tilde{Q}^{(n)}_{A,M}; \tilde{Q}^{(n)}_{Y|A,M}) = \mathbb{E}_{A,M}\left[D_f(\tilde{Q}_{Y|A,M=A,M}; \tilde{Q}_Y)\right] = \lambda I_f(Q^{(n)}_M; Q^{(n)}_{Y|M}) \tag{3.38}$$
We have shown that $F_{I_f}(\lambda x; W_{Y|X})$ is lower bounded by $\lambda I_f(Q^{(n)}_M; Q^{(n)}_{Y|M})$. Taking $\lambda = \gamma/x$ gives us our desired inequality.

[3] Some literature refer to this curve as the F_I curve.

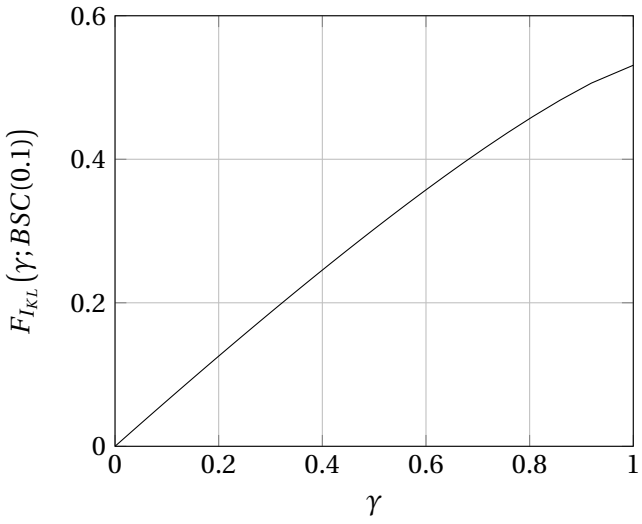


Figure 3.1: KL Information Curve for Binary Symmetric Channel with 0.1 flipping probability. The axes are denoted in bits as opposed to nats.

□

An example of numerical evaluation of the information transfer curve for the BSC is given in Figure 3.1. In this case, we can see that the curve is concave and strictly increasing, although this is not necessarily true for general channels.

Later on, we will see that the important quantity is the multiplicative reduction factor on information. We will refer to this quantity as the contraction coefficient of the channel. In the literature, the contraction coefficient is more commonly measured for F-divergence, and is defined as follows.

Definition 3.5. *Given a channel $W_{Y|X}$, the channel contraction coefficient for the F-divergence induced by a convex function f is defined as,*

$$\eta_f(W_{Y|X}) = \sup_{Q_X, P_X : 0 < D_f(P;Q) < \infty} \frac{D_f(W_{Y|X} \circ Q_X; W_{Y|X} \circ P_X)}{D_f(Q_X; P_X)}. \tag{3.39}$$

Even though the channel contraction coefficient is defined in terms of divergences, one can show that for certain choices of F-divergences, the channel contraction coefficient in terms of divergences is equivalent to the corresponding ratio in terms of F-information. This proof is a specific case of results presented in (Polyanskiy and Wu, 2015).

Proposition 3.7. *Let f be a convex function which fulfills the conditions of Proposition 3.2. Then we have,*

$$\eta_f(W_{Y|X}) = \sup_{Q_{M,X}} \frac{I_f(Q_M; Q_{Y|M})}{I_f(Q_M; Q_{X|M})} \tag{3.40}$$

where $Q_{Y|M} = W_{Y|X} \circ Q_{X|M}$.

Proof. First, let us prove the easy direction. We have,

$$\begin{aligned}\sup_{Q_{M,X}} \frac{I_f(Q_M, Q_{Y|M})}{I_f(Q_M, Q_{X|M})} &= \sup_{Q_{M,X}} \frac{E[D_f(Q_{Y|M=M}; Q_Y)]}{E[D_f(Q_{X|M=M}; Q_X)]} \\ &\leq \sup_{Q_{M,X}} \max_m \frac{D_f(Q_{Y|M=m}; Q_Y)}{D_f(Q_{X|M=m}; Q_X)} \\ &\leq \sup_{Q_X, P_X : 0 < D_f(P;Q) \leq \infty} \frac{D_f(W_{Y|X} \circ Q_X; W_{Y|X} \circ P_X)}{D_f(Q_X; P_X)}. \end{aligned} \tag{3.41}$$

For the other direction, consider any pair Q_X, P_X such that $\frac{dQ_X}{dP_X} < \infty$. We will define a distribution,

$$\tilde{P}_{M,X}(m,x) = \begin{cases} \frac{\epsilon}{1+\epsilon} Q_X & m = 1 \\ \frac{1}{1+\epsilon}(P_X + \epsilon(P_X - Q_X)) & m = 0. \end{cases} \tag{3.42}$$

We note that this family of distribution is a valid probability distribution with $\tilde{P}_X = P_X$ if

$$\epsilon < \frac{1}{\sup_x \frac{dQ_X}{dP_X}(x) - 1}. \tag{3.43}$$

Let us define $P_{Y|M} = W_{Y|X} \circ P_{X|M}$, then we have,

$$\begin{aligned}\frac{I_f(\tilde{P}_M, W_{Y|X} \circ \tilde{P}_{X|M})}{I_f(\tilde{P}_M, \tilde{P}_{X|M})} &= \frac{\epsilon^{-1} D_f(Q_{Y|X} \circ \tilde{P}_{X|M=0}; P_Y) + D_f(Q_{Y|M=1}; P_Y)}{\epsilon^{-1} D_f(\tilde{P}_{X|M=0}; P_X) + D_f(Q_{X|M=1}; P_X)} \\ &= \frac{\epsilon^{-1} D_f(Q_{Y|X} \circ \tilde{P}_{X|M=0}; P_Y) + D_f(Q_Y; P_Y)}{\epsilon^{-1} D_f(\tilde{P}_{X|M=0}; P_X) + D_f(Q_X; P_X)}. \end{aligned} \tag{3.44}$$

Hence, we can prove our assertion if we can show that $\epsilon^{-1} D_f(Q_{Y|X} \circ \tilde{P}_{X|M=0}; P_Y) \to 0$ and $\epsilon^{-1} D_f(\tilde{P}_{X|M=0}; P_X) \to 0$. We will only show this fact for $\tilde{P}_{X|M=0}$ and P_X but the argument also holds for the output divergence. We have,

$$\begin{aligned}\frac{1}{\epsilon} D_f(\tilde{P}_{X|M=0}; P_X) &= \frac{1}{\epsilon} E\left[f\left(1 + \epsilon\left(1 - \frac{dQ_X}{dP_x}\right)\right)\right] \\ &= \frac{1}{\epsilon} E\left[f(1)\right] + O(1) E\left[1 - \frac{dQ_X}{dP_X}\right] + O(\epsilon) \\ &= O(\epsilon). \end{aligned} \tag{3.45}$$

This shows that for any pair of distributions Q_X, P_X which achieves the supremum for the channel contraction coefficient, we can design another sequence $\tilde{P}_{M,X}$ which also achieves the same value for F-information, hence proving the proposition. □

In principle, the channel contraction coefficient of a channel might depend on the F-divergence that we use to measure the contraction. A natural question would be to ask: what is the f which minimizes the channel contraction coefficient? The following proposition is interesting, as we can prove that the worst measure among F-divergences which fulfills the condition of Proposition 3.2 is precisely the χ^2 divergence.

Proposition 3.8. *Let f be a convex function which fulfills the conditions of Proposition 3.2. Then we have,*

$$\eta_f(W_{Y|X}) \geq \eta_{\chi^2}(W_{Y|X}) \tag{3.46}$$

Proof. Consider any pair Q_X, P_X. We define another distribution $\tilde{Q}_X = (1-\epsilon)P_X + \epsilon Q_X$. Under our choice of f we have,

$$\begin{aligned} D_f(\tilde{Q}_X, P_X) &= \mathbb{E}[f(1)] + O(1)\epsilon\mathbb{E}\left[\left(\frac{dQ_X}{dP_x} - 1\right)\right] + \left.\frac{\partial^2 f(x)}{\partial x^2}\right|_{x=1} \epsilon^2 \mathbb{E}\left[\left(\frac{dQ_X}{dP_x} - 1\right)^2\right] + O(\epsilon^3) \\ &= \left.\frac{\partial^2 f(x)}{\partial x^2}\right|_{x=1} \epsilon^2 D_{\chi^2}(Q_X, P_X) + O(\epsilon^3). \end{aligned} \tag{3.47}$$

Therefore, consider the sequence $Q_X^{(n)}, P_X^{(n)}$ which achieves the supremum for χ^2 divergence. Then, we can construct another sequence of distributions $\tilde{Q}_X^{(n)} = (1-\epsilon_n)P_X^{(n)} + \epsilon_n Q_X^{(n)}$ with $\epsilon_n \to 0$ such that,

$$\frac{D_f(\tilde{Q}_Y^{(n)}; P_Y^{(n)})}{D_f(\tilde{Q}_X^{(n)}; P_X^{(n)})} \to \eta_{\chi^2}(Q_{Y|X}). \tag{3.48}$$

Hence proving the proposition. □

3.3 Upper Bound on Information Velocity

In relaying problems, the F-information between the $i+1$-th relay's observations and the message cannot be larger than the information that the i-th relay has, as the relays themselves form a Markov chain. Thus, at each time increment, the gain in information on the $(i+1)$-th relay in the n-th channel use cannot be larger than their difference in information after the $(n-1)$-th channel use[4]. So we have,

$$\begin{aligned} &I_f(Q_M; Q_{Y_{i+1}[1:n]|M}) - I_f(Q_M; Q_{Y_{i+1}[1:n-1]|M}) \\ &\leq \left(I_f(Q_M; Q_{Y_i[1:n-1]|M}) - I_f(Q_M; Q_{Y_{i+1}[1:n-1]|M})\right). \end{aligned} \tag{3.49}$$

As currently stated, this inequality is just a convoluted restatement of the data processing inequality. At this point there is no reason to believe that the information velocity is finite, as given this inequality, there is no limit of information that the i-th relay can convey as long as

[4] Due to our assumption on the causality structure, the output of the i-th relay at time n can only depend on its information at time $n-1$.

there is an information difference between the relays. In fact, the characteristic of the channel does not come into consideration for this inequality.

The information velocity only comes into play if we can show that at each time step, the i-th relay cannot convey its whole information. In fact, for the majority of the channels, it can only convey at most a fixed fraction of its information due to the channel noise. Here, the contraction coefficient of the channel comes into play.

Proposition 3.9. *We have,*

$$\begin{aligned} &I_f(Q_M; Q_{Y_{i+1}[1:n]|M}) - I_f(Q_M; Q_{Y_{i+1}[1:n-1]|M}) \\ &\leq \gamma \left(I_f(Q_M; Q_{Y_i[1:n-1]|M}) - I_f(Q_M; Q_{Y_{i+1}[1:n-1]|M}) \right). \end{aligned} \tag{3.50}$$

where

$$\gamma = \sup_{\alpha} \eta_f^{(\alpha)}(Q_{Y|X}) \tag{3.51}$$

with $\eta_f^{(\alpha)}(Q_{Y|X}) = \sup_{C,M} \frac{D_f^{(\alpha)}(Q_{Z_{out}|C,M}, Q_{Z_{out}|C})}{D_f^{(\alpha)}(Q_{Z_{in}|C,M}, Q_{Z_{in}|C})}$ *for any convex function* f *which fulfills the conditions of Proposition 3.2.*

Proof. In this case, we will have,

$$\begin{aligned} &\frac{I_f(Q_M; Q_{Y_{i+1}[1:n]|M}) - I_f(Q_M; Q_{Y_{i+1}[1:n-1]|M})}{I_f(Q_M; Q_{Y_i[1:n-1]|M}) - I_f(Q_M; Q_{Y_{i+1}[1:n-1]|M})} \\ &\qquad \leq \frac{I_f(Q_M; Q_{Z_{out},C|M}) - I_f(Q_M; Q_{C|M})}{I_f(Q_M; Q_{Z_{in},C|M}) - I_f(Q_M; Q_{C|M})}. \end{aligned} \tag{3.52}$$

where for the ease of notation we define,

$$Z_{in} = X_{i+1}[n], \quad Z_{out} = Y_{i+1}[n], \quad C = Y_{i+1}[1:n-1] \tag{3.53}$$

and the inequality is due to the fact that

$$M \to Y_i[1:n-1] \to (Z_{in}, X_{i+1}[1:n-1]) \to (Z_{out}, C), \tag{3.54}$$

which implies that $I_f(Q_M; Q_{Y_i[1:n-1]|M}) \geq I_f(Q_M; Q_{Z_{in},C|M})$.

Focusing on the numerator, we have,

$$\begin{aligned} I_f(Q_M; Q_{Z_{out},C|M}) - I_f(Q_M; Q_{C|M}) &= \mathbb{E}_{Q_M Q_C} \left[\mathbb{E}_{Q_{Z_{out}|C}} \left[f\left(\frac{dQ_{Z_{out}|C,M}}{dQ_{Z_{out}|C}} \frac{dQ_{C|M}}{dQ_C} \right) \right] - f\left(\frac{dQ_{C|M}}{dQ_C} \right) \right] \\ &= \mathbb{E}_{Q_{MC}} \left[\frac{\mathbb{E}_{Q_{Z_{out}|C}} \left[f\left(\frac{dQ_{Z_{out}|C,M}}{dQ_{Z_{out}|C}} \frac{dQ_{C|M}}{dQ_C} \right) \right] - f\left(\frac{dQ_{C|M}}{dQ_C} \right)}{\frac{dQ_{C|M}}{dQ_C}} \right]. \end{aligned} \tag{3.55}$$

Using the definition of the tilted F-divergences, we have,

$$I_f(Q_M; Q_{Z_{out},C|M}) - I_f(Q_M; Q_{C|M}) = \mathbb{E}_{Q_{MC}}\left[D_f^{\left(\frac{dQ_{C|M}}{dQ_C}\right)}\left(Q_{Z_{out}|C,M}, Q_{Z_{out}|C}\right)\right]. \tag{3.56}$$

We can perform similar steps for the denumerator, hence we have,

$$\begin{aligned}\frac{I_f(Q_M; Q_{Y_{i+1}[1:n]|M}) - I_f(Q_M; Q_{Y_{i+1}[1:n-1]|M})}{I_f(Q_M; Q_{Y_i[1:n-1]|M}) - I_f(Q_M; Q_{Y_{i+1}[1:n-1]|M})} &\leq \frac{\mathbb{E}_{Q_{MC}}\left[D_f^{\left(\frac{dQ_{C|M}}{dQ_C}\right)}\left(Q_{Z_{out}|C,M}, Q_{Z_o|C}\right)\right]}{\mathbb{E}_{Q_{MC}}\left[D_f^{\left(\frac{dQ_{C|M}}{dQ_C}\right)}\left(Q_{Z_{in}|C,M}, Q_{Z_i|C}\right)\right]} \\ &\leq \sup_{\alpha}\sup_{C,M}\frac{D_f^{(\alpha)}\left(Q_{Z_{out}|C,M}, Q_{Z_{out}|C}\right)}{D_f^{(\alpha)}\left(Q_{Z_{in}|C,M}, Q_{Z_{in}|C}\right)} \\ &\leq \sup_{\alpha}\eta_f^{(\alpha)}(W_{Y|X}). \end{aligned} \tag{3.57}$$

Thus proving the preposition. □

The main idea of the next proposition can be traced back to (Rajagopalan and Schulman, 1994). Our main contribution here is extending the proof to any F-divergences which fulfills the condition of Proposition 3.2.

Proposition 3.10. *For any convex function f that fulfills the conditions of proposition 3.2, the following holds,*

$$v_1 \leq \gamma \tag{3.58}$$

where,

$$\gamma = \sup_{\alpha}\eta_f^{(\alpha)}(W_{Y|X}). \tag{3.59}$$

Proof. The general idea of the proof is to form an upper bounding function $g(i,n)$ such that

$$I_f(Q_M; Q_{Y_i[1:n+i]|M}) \leq g(i,n) \tag{3.60}$$

for all i and n. Then we show that this upper bound goes to 0 if i is growing faster than n. By Proposition 3.5, this implies that the average error probability of a decoder is strictly away from 0.

We will define $g(i,n)$ recursively as,

$$g(i,n) = \begin{cases} 0 & n = 0 \\ f(|M|) & i = 0 \\ \gamma g(i-1,n) + (1-\gamma)g(i,n-1) & \text{otherwise} \end{cases} \tag{3.61}$$

The exact initial value for $g(0,n)$ is inconsequential for the proof, as the argument still holds for any finite positive value of $g(0,n)$. It is simply chosen to ensure that it is a valid upper bound for $I_f(Q_M; Q_{\{Y_0[j]\}_{j\leq n}|M})$.

From Proposition 3.9, we have,[5]

$$\begin{aligned} I_f(Q_M; Q_{Y_i[1:n+i]|M}) &\leq \gamma I_f(Q_M; Q_{Y_{i-1}[1:n+i-1]|M}) + (1-\gamma) I_f(Q_M; Q_{Y_i[1:n+i-1]|M}) \\ &\leq \gamma g(i-1,n) + (1-\gamma) g(i,n-1) \\ &= g(i,n). \end{aligned} \tag{3.62}$$

Let us note that the boundaries, i.e., the $g(i,0)$'s and $g(0,n)$'s is true by the definition of the F-divergences. Note that in inequality eq. 3.62 the $g(i,n)$ with $i+n=k$ only depends on $g(i-1,n)$ and $g(i,n-1)$, namely the terms of g where the sum of its first and second argument is equal to $k-1$. This observations allows us to structure an induction argument, where we first prove the inequality for $n+i=0$ by appealing to the boundary conditions. Thereafter, we apply the inequality for $g(n,i)$'s with $n+i=k$ assuming that it holds for $g(n',i')$'s with $n'+i'=k-1$. Hence by induction we have

$$I_f(Q_M; Q_{Y_i[1:n+i]|M}) \leq g(i,n) \tag{3.63}$$

for all i and n.

Finally, we want to show that for any $\upsilon > \frac{\gamma}{1-\gamma}$,

$$\lim_{n\to\infty} g(\upsilon n, n) = 0. \tag{3.64}$$

The crucial observation here is that the recursion for $g(\cdot,\cdot)$ forms a linear system. Hence by the superposition principle, we can decompose the value of $g(k,n)$ into a sum of $g_i(k,n)$ for $i<n$, where $g_i(k,n)$ is the value of $g(k,n)$ with boundary condition $g(k,0)=0$ for $k>0$, $g(0,n)=0$ for $n\neq i$ and $g(0,i)=f(|M|)$. We have,

$$g_i(k,n) = \binom{k+n-i}{k} \gamma^k (1-\gamma)^{n-i} \tag{3.65}$$

and therefore

$$g(k,n) = \sum_{i=1}^{n} g_i(k,n). \tag{3.66}$$

To prove eq. 3.64 we will use the estimate of the exponent of $g(k,n)$, here we will use prop. 1.17

[5] Under our model, teh

in the appendix of chapter 1. We have,

$$\lim_{n\to\infty}\frac{1}{n}\log g(\lfloor \upsilon \rfloor n, n) = \max_{0\le i\le n} -D_{KL}\left(\frac{\upsilon}{\upsilon+1-\frac{i}{n}},\gamma\right)$$
$$\le \max_{0\le x\le 1} -D_{KL}\left(\frac{\upsilon}{\upsilon+1-x},\gamma\right). \tag{3.67}$$

The upper bound can be explicitly computed, which gives the optimal x as,

$$x = \left[1-\upsilon\frac{1-\gamma}{\gamma}\right]_+ \tag{3.68}$$

and implies that,

$$\lim_{n\to\infty}\frac{1}{n}\log g(\lfloor \upsilon \rfloor n, n) \le \begin{cases} 0 & \upsilon \le \frac{\gamma}{1-\gamma} \\ -D_{KL}\left(\frac{\upsilon}{1+\upsilon},\gamma\right) & \upsilon > \frac{\gamma}{1-\gamma}. \end{cases} \tag{3.69}$$

Due to this bound on the exponent, we have eq. 3.64. Which also implies that

$$\lim_{n\to\infty} I_f(Q_M; Q_{Y_{\lfloor \upsilon \rfloor n}[1:n]|M}) = 0 \tag{3.70}$$

for $\upsilon > \gamma/(1-\gamma)$. Converting this value in terms of the ratio between the number of relays and the number of channel uses gives us the following upper bound,

$$\upsilon_1 \le \frac{\frac{\gamma}{1-\gamma}}{1-\frac{\gamma}{1-\gamma}} = \gamma. \tag{3.71}$$

Hence proving the proposition. □

A natural question arises regarding whether the upper bound that we developed can be easily evaluated. In general, the answer is negative. The channel contraction coefficient for general channel and general F-divergences is generally hard to compute, since the resulting optimization problem is generally not convex, which leads to a great interest in developing bounds for this problem.

However, we will show that for BMS channels, we have a simple characterization of the channel contraction coefficient under the χ^2 divergence. The χ^2 divergence is an important F-divergence as we have shown in the previous section that the channel contraction coefficient under χ^2 divergence is a lower bound on any channel contraction coefficient. Hence, the following proposition essentially gives the tightest bound for BMS channels that we can prove using this proof technique.

Proposition 3.11. *For BMS channels, we have,*

$$\eta_{\chi^2}(W_{Y|X}) = \frac{1}{2}\sum_{y\in\mathscr{Y}}\frac{(W_{Y|X=1}(y)-W_{Y|X=0}(y))^2}{(W_{Y|X=1}(y)+W_{Y|X=0}(y))}. \tag{3.72}$$

Proof. From the result in Proposition 3.6, we have that $F_{\chi^2}(x;W_{Y|X})/x$ is a non-increasing function. Hence, we know that the optimal coefficient can be achieved by the sequence where $D_{\chi^2}(Q_X;P_X)\to 0$. We would only consider Q_X and P_X as perturbed version of each other.

In this case, we can parameterize $Q_X(1) = P_X(1)+\epsilon$ and $Q_X(0) = P_X(0)-\epsilon$, let us denote $P_X(1) = p$. We have,

$$D_{\chi^2}(Q_X;P_X) = \sum_{x\in\{0,1\}}\frac{(Q_X(x)-P_X(x))^2}{P_X(x)} = \frac{\epsilon^2}{p}+\frac{\epsilon^2}{\bar{p}} = \frac{\epsilon^2}{p\bar{p}}. \tag{3.73}$$

We have

$$D_{\chi^2}(W_{Y|X}\circ P_X;W_{Y|X}\circ Q_X) = \sum_{y\in\mathscr{Y}}\frac{(W_{Y|X=1}(y)\epsilon - W_{Y|X=0}(y)\epsilon)^2}{W_{Y|X=1}(y)p+W_{Y|X=0}(y)\bar{p}}. \tag{3.74}$$

Hence we have,

$$\frac{D_{\chi^2}(W_{Y|X}\circ P_X;W_{Y|X}\circ Q_X)}{D_{\chi^2}(P_X;Q_X)} = p\bar{p}\sum_{y\in\mathscr{Y}}\frac{(W_{Y|X=1}(y)-W_{Y|X=0}(y))^2}{W_{Y|X=1}(y)p+W_{Y|X=0}(y)\bar{p}}. \tag{3.75}$$

As we are having BMS channel, we can find π as the BMS permutation such that,

$$\begin{aligned}\frac{D_{\chi^2}(W_{Y|X}\circ P_X;W_{Y|X}\circ Q_X)}{D_{\chi^2}(P_X;Q_X)} &= \frac{p\bar{p}}{2}\sum_{y\in\mathscr{Y}}\frac{(W_{Y|X=1}(y)-W_{Y|X=0}(y))^2}{W_{Y|X=1}(y)p+W_{Y|X=0}(y)\bar{p}}+\frac{(W_{Y|X=1}(\pi(y))-W_{Y|X=0}(\pi(y)))^2}{W_{Y|X=1}(\pi(y))p+W_{Y|X=0}(\pi(y))\bar{p}}\\ &= \frac{p\bar{p}}{2}\sum_{y\in\mathscr{Y}}\frac{(W_{Y|X=1}(y)-W_{Y|X=0}(y))^2}{W_{Y|X=1}(y)p+W_{Y|X=0}(y)\bar{p}}+\frac{(W_{Y|X=0}(y)-W_{Y|X=1}(y))^2}{W_{Y|X=0}(y)p+W_{Y|X=1}(y)\bar{p}}\\ &= \frac{1}{2}\sum_{y\in\mathscr{Y}}\frac{(W_{Y|X=1}(y)-W_{Y|X=0}(y))^2(W_{Y|X=1}(y)+W_{Y|X=0}(y))}{W_{Y|X=1}(y)^2+W_{Y|X=0}(y)^2+W_{Y|X=1}(y)W_{Y|X=0}(y)\left(\frac{\bar{p}}{p}+\frac{p}{\bar{p}}\right)}.\end{aligned} \tag{3.76}$$

We can minimize the denumerator by taking $p=\bar{p}$. Taking $p = 1/2$ gives us the proposition.

□

3.4 Discussions

3.4.1 Single-Bit Information Velocity of BEC

In chapter 2, we have shown a lower bound for single-bit information velocity of the BEC. With the converse that we have developed, we can show that the lower bound is in fact tight.

Proposition 3.12. *The single-bit information velocity of BEC with erasure probability ϵ, v_1, is given by*

$$v_1 = 1 - \epsilon. \tag{3.77}$$

Proof. The lower bound has been established in the previous chapter, we will only focus on the upper bound. We will use Proposition 3.10, by taking the contraction coefficient under the KL divergence.

$$v_1 \leq \sup_{Q_{X,M}} \frac{I(M,Y)}{I(M,X)} = \sup_{Q_{X,M}} \frac{(1-\epsilon)I(M,X)}{I(M,X)} = 1 - \epsilon. \tag{3.78}$$

□

3.4.2 Information Velocity through Information Transfer Curves?

In this chapter, we introduced the notion of information transfer curves to motivates the definition of the channel contraction coefficient. We also used the property that the channel contraction coefficient is achieved at the point where the input information is approaching 0 to characterize the channel contraction coefficient for BMS channels.

However, there might be a suspicion that we can enhance the converse by directly using the information transfer curve. After all, by taking a point on the curve where the input information is away from 0, the ratio between the input and the output information can be decreased. In the argument of the converse, the following inequality plays an important role

$$\begin{aligned} &I_f(Q_M; Q_{Y_{i+1}[1:n]|M}) - I_f(Q_M; Q_{Y_{i+1}[1:n-1]|M}) \\ &\leq \gamma \left(I_f(Q_M; Q_{Y_i[1:n-1]|M}) - I_f(Q_M; Q_{Y_{i+1}[1:n-1]|M}) \right). \end{aligned} \tag{3.79}$$

Instead of taking a static γ, could we improve the inequality by varying γ to take into account the input information on the information transfer curve? For example, could we map

$$I_f(Q_M; Q_{Y_{i+1}[1:n]|M}) - I_f(Q_M; Q_{Y_{i+1}[1:n-1]|M}). \tag{3.80}$$

to a point on the x-axis in the information transfer curve to find the corresponding upper bound on the output information, á la the EXIT curve argument?[6]

Let us assume that we have the F_I curve, i.e., the information transfer curve of the KL diver-

[6] I hope the reader is familiar with LDPC analysis to get the reference, one of the most authoritative source is (Richardson and Urbanke, 2008). But this analogue is not crucial to our discussion.

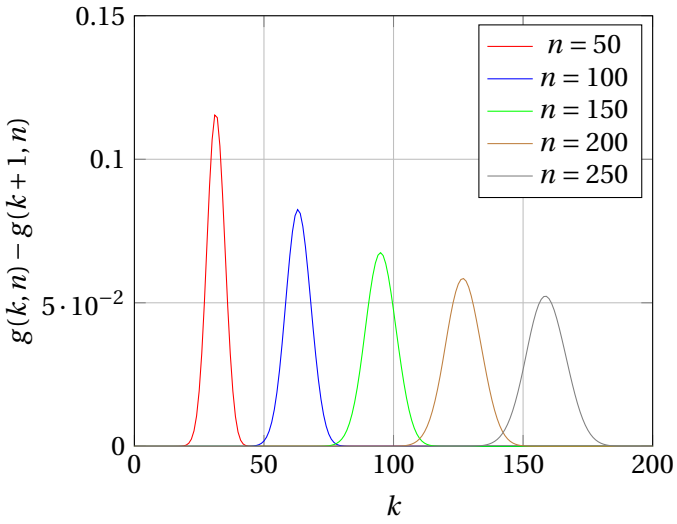


Figure 3.2: Information gap between relays as a function of time for BSC(0.1). We directly use information transfer curve evaluated in Figure 3.1 to calculate the upper bound $g(\cdot,\cdot)$.

gence, of the channel (or any upper bound on this curve, such that the upper bound is concave and non-decreasing). We can construct an upper bounds on the information between the relays' observations and the message given the following recursive relation,

$$F_I\left(\min\left\{g(k-1,n)-g(k,n), H(X)\right\}\right)+g(k,n)=g(k+1,n). \tag{3.81}$$

As promising as this might sound, unfortunately, a closer investigation will reveal that this approach will not lead to a better bound. The reason is that most of the time the information gap between the relays is close to 0, so the system will spend most of its time operating on the region where the ratio of information that it can transmit in a single channel use is dictated by the channel contraction coefficient. To give an empirical support to this claim, let us consider the numerical result given in Figure 3.2. As we can see, there are basically two forces operating on the curve. It is true that the center of the mass of the curve is moving with respect to time. However, we can also see that the curve itself is dispersed over an increasingly larger region as the time increases.

We can formalize this intuition as follows.

Proposition 3.13. *Given a F_I curve of a channel, we have,*

$$I(Q_M; Q_{Y_k[1:n]|M}) \leq g(k,n) \tag{3.82}$$

such that,

$$g(k,0) = \begin{cases} 0 & k > 0 \\ H(Q_M) & n > 0 \\ \tilde{F}_I\big(g(k-1,n) - g(k,n)\big) + g(k,n) & \textit{otherwise} \end{cases}$$

where $\tilde{F}_I$ is any concave and non-decreasing upper bound on F_I. Let us assume that $x^* = \inf\{x : \tilde{F}_I(x)/x < 1\}$. Then we have,

$$\lim_{n\to\infty} \max_k \big(g(k,n) - g(k+1,n)\big) \le x^*. \tag{3.83}$$

Proof. We will show that $g(k,n)$ is a valid upper bound through induction. The base case is simply from the fact that the sender has complete information of the message at the beginning. We have the following inequalities,

$$\begin{aligned} I(Q_M; Q_{Y_{k+1}[1:n+1]}) &= I(Q_M; Q_{Y_{k+1}[1:n]}) + I(Q_M; Q_{Y_{k+1}[n+1]} | Q_{Y_{k+1}[1:n]}) \\ &\le I(Q_M; Q_{Y_{k+1}[1:n]}) + \mathbb{E}_{Y_{k+1}[1:n]}\left[F_I\left(I(Q_M; Q_{X_{k+1}[n+1]} | Q_{Y_{k+1}[1:n]})\right)\right] \\ &\le I(Q_M; Q_{Y_{k+1}[1:n]}) + \tilde{F}_I\left(I(Q_M; Q_{X_{k+1}[n+1]} | Q_{Y_{k+1}[1:n]})\right) \end{aligned} \tag{3.84}$$

where the first inequality is due to the definition of F_I, the second inequality due to concavity of F_I. Due to data processing inequality and the fact that F_I is non-decreasing, we have,

$$\begin{aligned} I(Q_M; Q_{Y_{k+1}[1:n+1]}) &\le I(Q_M; Q_{Y_{k+1}[1:n]}) + \tilde{F}_I\left(I(Q_M; Q_{Y_k[1:n]} | Q_{Y_{k+1}[1:n]})\right) \\ &= I(Q_M; Q_{Y_{k+1}[1:n]}) + \tilde{F}_I\left(I(Q_M; Q_{Y_k[1:n]}, Q_{Y_{k+1}[1:n]}) - I(Q_M; Q_{Y_{k+1}[1:n]})\right) \\ &= I(Q_M; Q_{Y_{k+1}[1:n]}) + \tilde{F}_I\left(I(Q_M; Q_{Y_k[1:n]}) - I(Q_M; Q_{Y_{k+1}[1:n]})\right). \end{aligned} \tag{3.85}$$

The last equality due to the fact that $M - Y_k[1:n] - Y_{k+1}[1:n]$ forms a Markov chain.

As concavity implies subadditivity, we also have,

$$\begin{aligned} \tilde{F}_I\left(I(Q_M; Q_{Y_k[1:n]}) - I(Q_M; Q_{Y_{k+1}[1:n]})\right) &\le \tilde{F}_I\left(I_f(Q_M; Q_{Y_k[1:n]}) - g(k+1,n)\right) + \tilde{F}_I\left(g(k+1,n) - I(Q_M; Q_{Y_{k+1}[1:n]})\right) \\ &\le \tilde{F}_I\left(I(Q_M; Q_{Y_k[1:n]}) - g(k+1,n)\right) + g(k+1,n) - I(Q_M; Q_{Y_{k+1}[1:n]}) \end{aligned} \tag{3.86}$$

where the last inequality is due to the fact that $F_I(x) \le x$. Hence we have,

$$\begin{aligned} I(Q_M; Q_{Y_{k+1}[1:n+1]}) &\le g(k+1,n) + \tilde{F}_I\left(I(Q_M; Q_{Y_k[1:n]}) - g(k+1,n))\right) \\ &\le g(k+1,n) + \tilde{F}_I\left(g(k,n) - g(k+1,n))\right) \end{aligned} \tag{3.87}$$

as $\tilde{F}_{I_f}$ is non-decreasing, which implies the induction step.

Let us denote $g(k,n) - g(k+1,n)$ as $\Delta(k,n)$. Therefore we have,

$$\Delta(k+1,n+1) = \Delta(k+1,n) + \tilde{F}_I(\Delta(k,n)) - \tilde{F}_I(\Delta(k+1,n)). \tag{3.88}$$

Consider a sequence $x_{(1)}, x_{(2)}, \ldots$ which converges to x^* from above, from this sequence we can define a sequence $\gamma_{(1)}, \gamma_{(2)}, \ldots$ as $\tilde{F}_I(x_{(i)})/x = \gamma_{(i)}$. Consider $T_{(i)}$ to be the smallest n such that $\max_k \Delta(k,n) \leq x_{(i)}$. We have the following inequality to hold for $n \leq T_{(i)}$,

$$\Delta(k, n+1) \leq \Delta(k,n) + \gamma_{(i)} \left(\Delta(k-1,n) - \Delta(k,n)\right). \tag{3.89}$$

Hence, we have that,

$$\max_k \Delta(k,n) \leq \max_k \binom{k+n}{k} \left(1-\gamma_{(i)}\right)^n \gamma_{(i)}^k. \tag{3.90}$$

By Stirling's inequality, we have that,

$$\max_k \Delta(k,n) \leq \frac{1}{\sqrt{2en}} \frac{n^n k^k}{(n+k)^{n+k}} \left(1-\gamma_{(i)}\right)^n \gamma_{(i)}^k. \tag{3.91}$$

The optimal is achieved such that $k/(n+k) \approx 1-\gamma_{(i)}$. Hence, we have that

$$\max_k \Delta(k,n) \sim O\left(\frac{1}{\sqrt{n}}\right), \tag{3.92}$$

and there must exists a finite $T_{(i)}$ such that this maximum is less than $x_{(i)}$. This proves the proposition. □

Hence, the only region of F_I curve which determines the information velocity is in the neighborhood 0, which exactly corresponds to the region where the slope of the curve is equal to the channel contraction coefficient.

4 Blockwise Information Velocity in Binary Erasure Channels

> Wovon man nicht sprechen kann,
> darüber muss man schweigen.
>
> Ludwig Wittgenstein

In chapter 2, we discussed a relaying scheme which achieves positive single-bit information velocity for BEC without feedback. We have also discussed the challenges of extending this scheme to achieve positive blockwise information velocity.

In this chapter, we present an achievability scheme based on tree codes which achieves positive blockwise information velocity. This method is inspired by (Rajagopalan and Schulman, 1994), which uses tree codes to synchronize a state in distributed system. In the first section, we will introduce the ensemble of tree codes that we will use. In the second section, we specify the relaying scheme built on top of the tree codes and prove that a positive blockwise information velocity is achieved by this scheme. Additionally, we support our findings with numerical results.

4.1 Tree Code Ensemble

Before we describe the relaying scheme, we will discuss how we use a tree code for single hop communication. For ease of exposition, let us consider a case where the sender already has the whole data to be sent $D[1:m]$. For reasons that will become obvious in the next section, instead of assuming that $D[1:m]$ is only composed of bits, let us assume that $D[1:m]$ uniformly takes value from $\{0,1,?\}^M$.

Let us define a sequence of times $T[1], T[2]\ldots$ which we will refer as "speaking" times. This sequence is such that $T[1] \sim \text{Geometric}(p_s)$ and $T[i+1] - T[i] \sim \text{Geometric}(p_s)$, with the interspeaking times are i.i.d. and independent of other quantities. These speaking times denote the time indices at which the relay will attempt to send a message symbol.

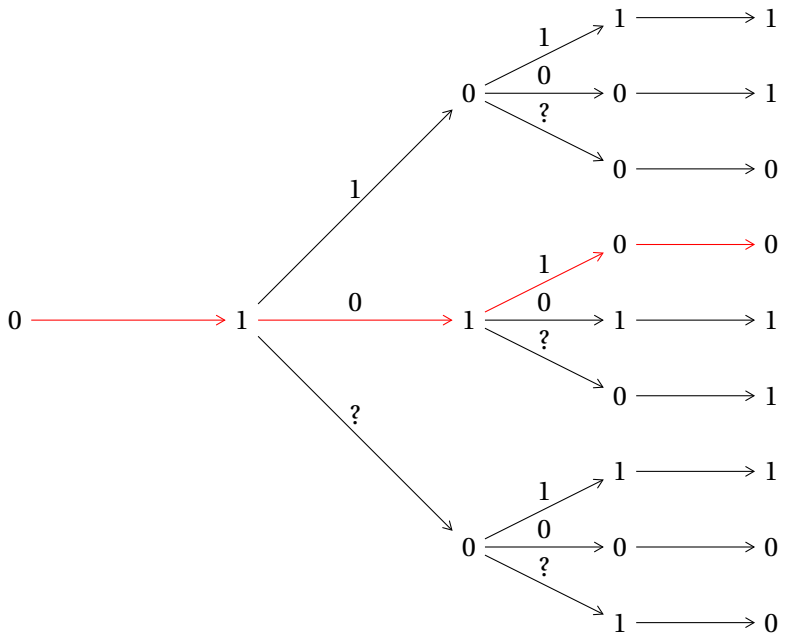


Figure 4.1: A possible realization from the tree code ensemble conditioned on $T[1] = 2, T[2] = 3$. Each label on the nodes represent the symbol being sent through the channel, which is sampled independently and uniformly from $\{0,1\}$. Each label on the edge represents possible values of input sequence $D[1:2]$. The red edges represents the path that the encoding process takes assuming $D[1:2] = (0,1)$ in this case we have $\mathscr{T}(1:4; D[1:2]) = (0,1,1,0,0)$.

We build a tree, which will function as our codebook, based on these speaking times. Starting at the root of the tree, there are two possibilities: a) $T[1] = 1$, in this case the root will split into three different branches (each branch representing a different message symbol) with each branch is given a unique label from $\{0,1,?\}$, or b) $T[1] \neq 1$, in this case the relay will not attempt to sent any message symbol, in other words the root will only grow to a single branch. At the second level of the tree, the process is repeated, if 2 (the current height of the tree) is a speaking time then each node will grow into three different branches, otherwise the node will only grow into a single branch.

Once we define the structure of the tree, we will assign a label to each node by uniformly sampling from $\{0,1\}$. The sampling process of the speaking times sequence and the labels of each node defines an ensemble of tree codes. An example realization of this ensemble is given in Figure 4.1.

Given a sampled tree and the sequence of symbols $D[1:m]$, we can find the corresponding relay's emissions by tracing a path in the sampled tree. Imagine that we have a marker that tracks our position in the tree. At the beginning, the marker is placed at the root of the tree. If the marker visits a node which corresponds to a speaking time, for example the current time index is at T_i, the marker will take the branch which corresponds to the message symbol $D[i]$, otherwise, there is only a single outgoing branch in that node and the marker will simply

take the branch. Once the marker arrives at a node, the relay will emit the label of the said node. Given a tree $\mathscr{T}$, let us denote the message that the relay emits at time n given a message sequence $D[1:m]$ as $\mathscr{T}(n;D[1:m])$. [1]

As we are working with BEC, given a received sequence $Y[1:n]$, we can define a set of message sequences which are compatible with the received sequence. More formally, given a tree $\mathscr{T}$ let us define,

$$\mathscr{D}(Y[1:n];\mathscr{T}) = \left\{D[1:n] : \forall_{n:Y[n]\neq ?} Y[n] = \mathscr{T}(n;D[1:m])\right\} \tag{4.1}$$

namely the set of message sequences where all unerased parts of the received sequence agree with the output of the tree given the message sequence.

The receiving relay is able to measure the length of the correct message that it received by taking the longest common prefix among all compatible sequences. Due to the property of the BEC, this longest common prefix is also necessarily a prefix of the correct message sequence. Let us define a sequence $T^{(r)}[i]$ as the first time where all the compatible sequences in the tree agree on the i-th bit, i.e.,

$$T^{(r)}[i] = \min\{n : LLCP(\mathscr{D}(Y[1:n];\mathscr{T})) \geq i\} \tag{4.2}$$

where $LLCP(\cdot)$ is the function which returns the length of the longest common prefix of its argument. Note that $T^{(r)}[i]$ is the time that the i-th bit is "received".

An important notion for this tree code ensemble is the difference between the speaking time and the receiving time, i.e.,

$$\Delta[i] = T^{(r)}[i] - T[i]. \tag{4.3}$$

To make the discussion more concrete, let us return to the tree code presented in Figure 4.1. Here we have the true sequence sent by the receiver $\mathscr{T}(1:4,(0,1)) = (0,1,1,0,0)$. Let us assume that due to the erasures, the receiver observes $Y[1:5] = (0,1,1,0,?)$. Hence we have $\mathscr{D}(Y[1:5];\mathscr{T}) = \{(0,1),(0,?)\}$. This implies that at this point in time, the receiver has received the first message symbol while still in doubt about the second message symbol. Furthermore, we can also see that $T^{(r)}[1] = 2$ and $\Delta[1] = 0$.

Given this example, there might be doubt whether $\Delta[i]$ is a properly defined random variables, in the sense that we want to ensure $\Pr(\Delta[i] = \infty) = 0$. The following proposition shows that this worry is unfounded if the speaking probability is sufficiently low, i.e., if the rate at which

[1] Most works in the literature seem to credit Schulman's work (Schulman, 1993) as the seminal work on tree codes. But we can also see the idea of encoding trees and decoding trees in the work of Wozencraft (Wozencraft, 1957). The ensemble that we discussed here is more closely related to the tree codes construction presented in (Sahai and Mitter, 2006), and the propositions that we presented here are in the same spirit as the results that they developed for anytime capacity. Although, as far as our knowledge, the notion of using "speaking times" to control the branching of the tree is new and, to be fair, is unnecessary for standard channel coding or control applications. However, as we can see later, this notion is important to decouple the communication process between each hop in our scheme.

we send the message symbols is sufficiently low.

Proposition 4.1. *The ensemble average of* $\Delta[i]$ *satisfies,*

$$\mathbb{E}[\Delta[i]] \leq 1 + \frac{6p_s(1+2p_s)(1+\epsilon)^2}{(2-(1+2p_s)(1+\epsilon))(1-\epsilon)} \tag{4.4}$$

if $p_s \leq \frac{1}{1+\epsilon} - \frac{1}{2}$. *We refer to a tree code ensemble which fulfills this condition as a stable tree code ensemble.*

Proof. Consider the path that is taken by the correct message sequence in the tree. At each speaking time, we have two other trees branching from the siblings of the correct branch. Let us define $S_{i\to j}$ as the number of paths in these branching trees that starts at time index i and are compatible with the received output sequence up to time j. If i is a speaking time then we have $S_{i\to i} = 2$, otherwise $S_{i\to i} = 0$. We have,

$$\mathbb{E}[\Delta[i]] = 1 + \sum_{j=1}^{\infty} \Pr\left(\bigcup_{k=1}^{i} \{S_{k\to i+j} > 0\} \,\middle|\, S_{i\to i} = 2 \right). \tag{4.5}$$

By the union bound we have,

$$\mathbb{E}[\Delta[i]] \leq 1 + \sum_{j=1}^{\infty} \sum_{k=1}^{i} \Pr\left(S_{k\to i+j} > 0 \,\middle|\, S_{i\to i} = 2\right). \tag{4.6}$$

The probability term can be further upper bounded by considering each leaf in the branching tree separately and applying the union bound on the probability that each such leaf is compatible with the received output sequence, i.e.,

$$\begin{aligned}\Pr\left(S_{k\to i+j} > 0 \,\middle|\, S_{i\to i} = 2\right) &\leq 6p \sum_{l=0}^{i-k} \binom{i-k}{l} 3^l (1-p_s)^{i-k-l} p_s^l \sum_{r=0}^{i+j-k} \binom{i+j-k}{r} 2^{-r} (1-\epsilon)^r \epsilon^{i+j-k-r} \\ &= 6p_s \left((1+2p_s)\frac{1+\epsilon}{2} \right)^{i-k} \left(\frac{1+\epsilon}{2} \right)^{j}. \end{aligned} \tag{4.7}$$

By plugging this upper bound to the sum, and extending the sum indexed by k to infinity gives us the desired upper bound. □

In the next proposition, we will show that the expected delay is non-decreasing. Intuitively, the decoder has an easier time decoding the nodes closer to the root of the tree as they have to contend with fewer competing branches than the later nodes.

Proposition 4.2. *For stable tree code ensembles, we have,*

$$\mathbb{E}[\Delta[i]] \leq \mathbb{E}[\Delta[i']] \tag{4.8}$$

for $i \leq i'$.

Proof. Consider the decomposition of the expectation in eq. 4.5, we can see by the memorylessness of the channel and the speaking time distribution that the probability term is shift invariant, i.e.,

$$\Pr\left(\bigcup_{k=1}^{i}\left\{S_{k\to i+j}>0\right\}\,\middle|\,S_{i\to i}=2\right)=\Pr\left(\bigcup_{k=1}^{i}\left\{S_{k+l\to i+j+l}>0\right\}\,\middle|\,S_{i+l\to i+l}=2\right) \tag{4.9}$$

for $l\geq 0$. Hence, we have,

$$\begin{aligned}\Pr\left(\bigcup_{k=1}^{i}\left\{S_{k\to i+j}>0\right\}\,\middle|\,S_{i\to i}=2\right)&=\Pr\left(\bigcup_{k=1}^{i}\left\{S_{k+(i'-i)\to i+j+(i'-i)}>0\right\}\,\middle|\,S_{i+(i'-i)\to i+(i'-i)}=2\right)\\&=\Pr\left(\bigcup_{k=1+(i'-i)}^{i'}\left\{S_{k\to i'+j}>0\right\}\,\middle|\,S_{i'\to i'}=2\right)\\&\leq\Pr\left(\bigcup_{k=1}^{i'}\left\{S_{k\to i'+j}>0\right\}\,\middle|\,S_{i'\to i'}=2\right).\end{aligned} \tag{4.10}$$

This implies that all probability terms in the decomposition in eq. 4.2 is larger for $\mathbb{E}[\Delta[i']]$ than the decomposition for $\mathbb{E}[\Delta[i]]$. Hence, the desired inequality. □

The main utility of these propositions is to show that the expected delay converges to a certain value if the rate of message sequence transmission is sufficiently slow (p_s is low enough).

4.2 A Relaying Scheme for BEC

In the relaying scheme that we will develop, the transmitting and the receiving relay at each hop shares a single realization of a tree code that they will use to communicate.

The chain relaying scheme that we will develop is based on the concept of "tape" abstraction. All relays in the chain maintain an internal tape, we will denote the tape of k relay as $D_k[1:n]$, where each $D_k[i]$ takes a value in $\{0,1,?\}$, hence the reason why we use this alphabet in the previous section. We will refer to ? as the erasure symbol, and the other symbols as the non-erased symbols. The k-th relay will communicates the content of its tape to the k-th using tree codes, i.e., the channel symbol that it emits corresponds to the edge label traced by its tape content $D_k[1:n]$

Imagine that the relay maintains a pointer to its tape. At the beginning, the pointer points to $D_k[0]$. After each speaking time $T_k[i]$, the relay moves its pointer to $D_k[i]$ and is allowed to write the value to $D_k[i]$. We would also like to emphasize that we now have two notions of time. The first notion is of the "physical clock", i.e., the index of the current channel use which is synchronized across all relays. The second notion is of the "logical clock", i.e., the number of speaking times that have passed which dictates how many symbols have been written to a relay's tape. These logical clocks are not synchronized across all relays and only shared by two relays in a single hop due to the virtue of them sharing a single tree code.

Let us start at the first hop. In this hop, the transmitting relay is also acting as the source and has a complete information of $B[1:m]$. At each speaking time, the transmitting relay has the opportunity to write a single value to its tape, which it will use according to the following rule

$$D_1[i] = \begin{cases} B[i] & i \le m \\ ? & \text{otherwise} \end{cases} \tag{4.11}$$

i.e., the transmitting relay copies the content of $B[1:m]$ to its tape. The transmitting relay will emit channel symbol based on its tape and its realization of tree code $\mathscr{T}_1$,

$$X_1[n] = \mathscr{T}_1(n, D_1[1:\infty]). \tag{4.12}$$

Although the expression in the LHS depends on $D_1[1:\infty]$, note that this operation still maintains causality as $\mathscr{T}_1(n, D_1[1:\infty])$ for $n < T_1[k]$ only depends on $D_1[1:k-1]$. At time n, the receiving relay in this hop has access to $Y_1[1:n]$. Based on this information, it also has access to a prefix of $D_1[1:\infty]$, which we define as,

$$D_1^{(n)}[1:] = LCP\left(\mathscr{D}(Y[1:n]; \mathscr{T}_1)\right),$$

where LCP is a function which returns the longest common prefix function given a set. We also omit the explicit length of the sequence in $D_1^{(n)}[1:]$ as this length is also a random variable depending on the realization of erasures in the channel. Note that due to its construction, we have that $D_1^{(n)}[1:]$ is a prefix of $D_1^{(n')}[1:]$ for all $n < n'$. [2]

Now, let us define the relaying process for subsequent hops, i.e., between the k-th relay and $(k+1)$-th relay for $k > 1$. At time $T_k[i]$, the transmitting relay has access to $D_{k-1}^{(T_k[i])}[1:]$ and has to decide what to write to its tape at $D_k[i]$. The main principle is for the relay to act as a queue, with the non-erased symbols in $D_{k-1}^{(T_k[i])}[1:]$ act as its input. It will compute the number of unerased symbols that it retrieves, i.e., $\sum_{j=1}^{\text{len}\left(D_{k-1}^{(T_k[i])}[1:]\right)} \mathbb{1}\left\{D_{k-1}^{(T_k[j])}[1:] = ?\right\}$, and the number of unerased symbols that it has sent, i.e., $\sum_{j=1}^{i-1} \mathbb{1}\left\{D_k[j] = ?\right\}$.

If the number of unerased input symbols is larger, it takes the oldest unerased input symbol that it has not sent, let's say $D_{k-1}^{(T_k[i])}[s]$ and it writes the symbol to its tape, i.e, $D_k[i] = D_{k-1}^{(T_k[i])}[s]$. Otherwise, the relay will write ? on its tape to signify that it currently has no new information.

We will denote the time when the i-th bit is written on the tape by the k-th relay as

$$R_k[i] = \min\left\{ T_k[j] : j \in \mathbb{Z}_+, \sum_{l=1}^{j} \mathbb{1}\{D_k[l] \neq ?\} = i \right\}. \tag{4.13}$$

Let us define the delay component $\Delta_k[i]$ as the difference between the time when the i-th bit

[2] In other words, the relay can explicitly measure its communication's progress. The inability of doing so is the main obstacle in extending this scheme to other channels which allow for undetected error.

is written to the k-th relay tape and when it is received by $k+1$-th relay, i.e., it is equivalent to

$$\Delta_k[i] = \Delta\left[\min\left\{j : j \in \mathbb{Z}_+, \sum_{l=1}^{j} \mathbb{1}\{D_k[l] \neq ?\} = i\right\}\right] \tag{4.14}$$

if we use the definition of $\Delta[i]$ in eq. 4.3 under the tree code of the k-th relay.

The main difficulty in analyzing $R_k[i]$ lies in the fact that it is not independent of $R_k[i']$ for $i' < i$. But despite this dependence, we can show that there exists a "self-regulating" behavior, namely that the delays tend to increasingly behave like independent variables as the separations between $R_k[m]$'s increase. Consequently, we can state the following fact about the expected delays.

Proposition 4.3. *For stable tree code ensembles, we have,*

$$\mathbb{E}[\Delta_k[i+1] - \Delta_k[i] \,|\, R_k[i+1] = n_2, R_k[i] = n_1] \leq e^{-O_{\epsilon,p_s}(n_2-n_1)} + e^{-O_{\epsilon,p_s}(n_1)} \tag{4.15}$$

for $m \geq 1$. We use $f(n) \in O_x(g(n))$ to denote that $\lim_{n\to\infty} g(n)/f(n) = c(x)$, i.e., the multiplicative factor is a function of x.

Proof. Care must be taken as the indexing for $R_k[\cdot]$'s is different from the indexing for $T_k[\cdot]$. Let us define i_1 and i_2 such that $T_k[i_1] = R_k[i]$ and $T_k[i_2] = R_k[i+1]$. We note that the sum in eq. 4.5 absolutely converges, hence we can decompose the expectation such that,

$$\begin{aligned}
&\mathbb{E}[\Delta_k[i+1] - \Delta_k[i] \,|\, R_k[i+1] = n_2, R_k[i] = n_1] \\
&\quad = \sum_{j=1}^{\infty} \Pr\left(\bigcup_{l=1}^{i_2} \left\{S_{T_k[l]\to T_k[i_2]+j} > 0\right\} \,\middle|\, \mathscr{C}\right) - \Pr\left(\bigcup_{l=1}^{i_1} \left\{S_{T_k[l]\to T_k[i_1]+j} > 0\right\} \,\middle|\, \mathscr{C}\right) \\
&\quad \leq \sum_{j=1}^{\infty} \Pr\left(\bigcup_{\substack{l: \\ T_k[l]\leq n_2-n_1}} \left\{S_{T_k[l]\to T_k[i_2]+j} > 0\right\} \,\middle|\, \mathscr{C}\right) + \Pr\left(\left\{S_{T_k[i_1]\to T_k[i_2]+j} > 0\right\} \,\middle|\, \mathscr{C}\right) \\
&\qquad + \Pr\left(\bigcup_{\substack{l: \\ i\neq i_1 \\ l\leq i_2 \\ T_k[i]>n_2-n_1}} \left\{S_{T_k[l]\to T_k[i_2]+j} > 0\right\} \,\middle|\, \mathscr{C}\right) - \Pr\left(\bigcup_{l=1}^{i_1} \left\{S_{T_k[l]\to T_k[i_1]+j} > 0\right\} \,\middle|\, \mathscr{C}\right)
\end{aligned} \tag{4.16}$$

with the conditioning event

$$\mathscr{C} = \{R_k[i+1] = n_2, R_k[i] = n_1\}. \tag{4.17}$$

Let us focus on the last two terms in the expression. First of all, we note that by shift invariance

property, we have,

$$\Pr\left(\bigcup_{\substack{l:\\ l\le i_2\\ T_k[l]>n_2-n_1}}\left\{S_{T_k[l]\to T_k[i_2]+j}>0\right\}\,\middle|\,R_k[i+1]=n_2\right)=\Pr\left(\bigcup_{l=1}^{i_1}\left\{S_{T_k[l]\to T_k[i_1]+j}>0\right\}\,\middle|\,R_k[i]=n_1\right). \tag{4.18}$$

Using this equality, we can then expand the last two terms in eq. 4.16 into,

$$\Pr\left(\bigcup_{\substack{l:\\ l\neq i_1\\ i\le i_2\\ T_k[l]>n_2-n_1}}\left\{S_{T_k[l]\to T_k[i_2]+j}>0\right\}\,\middle|\,\mathscr{C}\right)-\Pr\left(\bigcup_{\substack{l:\\ l\le i_2\\ T_k[l]>n_2-n_1}}\left\{S_{T_k[l]\to T_k[i_2]+j}>0\right\}\,\middle|\,R_k[i+1]=n_2\right) \tag{4.19}$$

and

$$\Pr\left(\bigcup_{l=1}^{i_1}\left\{S_{T_k[l]\to T_k[i_1]+j}>0\right\}\,\middle|\,R_k[i]=n_1\right)-\Pr\left(\bigcup_{l=1}^{i_1}\left\{S_{T_k[l]\to T_k[i_1]+j}>0\right\}\,\middle|\,\mathscr{C}\right). \tag{4.20}$$

We can show that eq. 4.19 is non-positive, as the event in both probability terms covers exactly the same time indices, with the difference that in the positive term, n_2 is guaranteed to not contain a speaking time. Hence one can couple the tree realization (except at n_2) and the erasure events in both probability terms. We can see that in this coupling, the event in the negative term is a superset of the event in the positive term. Hence this quantity is always non-positive.

A similar argument can be used for showing that eq. 4.20 is non-positive. But now the argument is based on the fact that n_1 is guaranteed to be a speaking time in the negative term.

As eq. 4.19 and eq. 4.20 are both non-positive, we can drop these terms from eq. 4.16. Hence we have,

$$\begin{aligned}&\mathbb{E}[\Delta_k[i+1]-\Delta_k[i]\,|\,R_k[i+1]=n_2,R_k[i]=n_1]\\ &\le\underbrace{\sum_{j=1}^{\infty}\Pr\left(\bigcup_{\substack{l:\\ T_k[l]\le n_2-n_1}}\left\{S_{T_k[l]\to T_k[i_2]+j}>0\right\}\,\middle|\,\mathscr{C}\right)}_{\le e^{-O_{\epsilon,p_s}(n_1)}}+\underbrace{\sum_{j=1}^{\infty}\Pr\left(\left\{S_{T_k[i_1]\to T_k[i_2]+j}>0\right\}\,\middle|\,\mathscr{C}\right)}_{\le e^{-O_{\epsilon,p_s}(n_2-n_1)}},\end{aligned} \tag{4.21}$$

where the estimate is obtained in similar manner as in eq. 4.7. □

Now, we can state the main result of this chapter, which gives the asymptotic trade-off between $R_k[m]$ and k.

Theorem 4.1. *We have,*

$$\mathbb{E}[R_k[1]] \leq k\left(\mu + \frac{1}{p_s}\right) \tag{4.22}$$

and

$$\mathbb{E}[R_k[m]] - \mathbb{E}[R_k[1]] \sim O_m(\sqrt{k}), \tag{4.23}$$

with $\mu = \mathbb{E}[\Delta_\infty]$.

Proof. We will start by the case of $R_k[1]$, where we do not have to worry about the dependence of the delays. Due to the behavior of the first relay, we have $R_1[i] = T_1[i]$. For $k \geq 1$, we can write $R_k[1]$ as,

$$R_{k+1}[1] = \min\left\{T_{k+1}[j] : j \in \mathbb{Z}_+, T_{k+1}[j] > R_k[1] + \Delta_k[1]\right\}. \tag{4.24}$$

We can also express $R_{k+1}[1]$ as,

$$R_{k+1}[1] = R_k[1] + \Delta_k[1] + N_k\left(R_k[1] + \Delta_k[1]\right), \tag{4.25}$$

for $k \geq 1$, where

$$N_k(l) = \min_j \left\{T_{k+1}[j] - l : T_{k+1}[j] > l\right\}. \tag{4.26}$$

Due to the memorylessness of the speaking times and the result in Proposition 4.2, we can upper bound the expectation as,

$$\mathbb{E}[R_{k+1}[1]] \leq \mathbb{E}[R_k[1]] + \mu + \frac{1}{p_s} \tag{4.27}$$

as $\mathbb{E}[N_k(\cdot)] = 1/p_s$. Repeatedly applying this inequality gives us,

$$\mathbb{E}[R_k[1]] \leq k\left(\mu + \frac{1}{p_s}\right). \tag{4.28}$$

Our goal now is to show that,

$$\mathbb{E}[R_k[m] - R_k[1]] \sim O_m(\sqrt{k}). \tag{4.29}$$

To do so, we will show that

$$\mathbb{E}[(R_k[i+1] - R_k[i])^2] \sim O_m(k), \tag{4.30}$$

which implies our goal as,

$$\mathbb{E}[R_k[m] - R_k[1]] = \sum_{i=1}^{M-1} E\left[R_k[i+1] - R_k[i]\right] \leq \sum_{i=1}^{M-1} \sqrt{\mathbb{E}\left[(R_k[i+1] - R_k[i])^2\right]}. \tag{4.31}$$

Hence, we have

$$\mathbb{E}[R_k[m] - R_k[1]] \sim O_m(\sqrt{k}). \tag{4.32}$$

To show eq. 4.30, note that we have

$$R_{k+1}[m] = \max_{0<i\le m} R_k[i] + \Delta_k[i] + N_k^{\circ M+1-i}(R_k[i] + \Delta_k[i]) \tag{4.33}$$

where $N_k^{\circ l}(x) = \underbrace{N_k \circ \cdots \circ N_k}_{l \text{ times}}(x)$. Hence we have

$$\begin{aligned} R_{k+1}[m] - R_{k+1}[m-1] = \max \bigcup_{i=1}^{m} \left\{ R_k[i] + \Delta_k[i] + N_k^{\circ m+1-i}(R_k[i] + \Delta_k[i]) \right\} \\ - \max \bigcup_{i=1}^{m-1} \left\{ R_k[i] + \Delta_k[i] + N_k^{\circ m-i}(R_k[i] + \Delta_k[i]) \right\} \end{aligned} \tag{4.34}$$

Let us recall the following fact about the difference of max functions,

$$\begin{aligned} \max\{x_1,\ldots,x_{n_x}\} - \max\{y_1,\ldots,y_{n_y}\} &= \max\{x_1 - \max\{y_1,\ldots,y_{n_y}\},\ldots,x_{n_x} - \max\{y_1,\ldots,y_{n_y}\}\} \\ &\le \max\{x_1 - y_{\pi(1)},\ldots,x_{n_x} - y_{\pi(n_x)}\} \end{aligned} \tag{4.35}$$

where π is any mapping from $\{1,\ldots,n_x\}$ to $\{1,\ldots,n_y\}$.

By choosing the following mapping

$$\pi(i) = \begin{cases} i & 1 \le i \le m-1 \\ m-1 & i = m \end{cases} \tag{4.36}$$

and squaring both sides, we have

$$\begin{aligned} (R_{k+1}[m] - R_{k+1}[m-1])^2 &\le \max G_e \cup \{(G_1+G_2+G_3)^2\} \\ &\le (\max G_e) + \left(G_1^2 + G_2^2 + G_3^3 + 2G_1G_2 + 2G_3(G_1+G_2)\right) \end{aligned} \tag{4.37}$$

where

$$\begin{aligned} G_e &= \bigcup_{i=1}^{m-1} \left\{ \left(N_k^{\circ m+1-i}(R_k[i] + \Delta_k[i]) - N_k^{\circ m-i}(R_k[i] + \Delta_k[i]) \right)^2 \right\} \\ G_1 &= R_k[m] - R_k[m-1] \\ G_2 &= \Delta_k[m] - \Delta_k[m-1] \\ G_3 &= N_k(R_k[m] + \Delta_k[m]) - N_k(R_k[m-1] + \Delta_k[m-1]). \end{aligned} \tag{4.38}$$

Note that squaring operation preserves the direction of the inequality, as $R_{k+1}[m] - R_{k+1}[m-1]$

is guaranteed to be non-negative.

As geometric distributions are memoryless and have a finite moment generating function, we can show that, $\mathbb{E}[\max G_e] \sim O(\log m)$, see appendix in chapter 1. Furthermore, it is easy to show that, $\mathbb{E}[G_2^2] \sim O(1)$ and $\mathbb{E}[G_3^2] \sim O(1)$.

Let us consider the $G_3(G_1+G_2)$ term. Observe that regardless of the value of G_1 and G_2, due to the memorylessness of geometric distribution and the fact that N_k only depends on the tree structure of the $(k+1)$-th relay, G_3 is independent of G_1 and G_2. Furthermore we have,

$$\mathbb{E}\left[N_k(x_1) - N_k(x_2)\middle| x_1 - x_2 = y\right] = \mathbb{E}\left[N_k(y) - N_k(0)\right] = 0. \tag{4.39}$$

This implies that $\mathbb{E}[G_3(G_1+G_2)] = 0$.

Finally, we consider the G_1G_2 term. We have

$$\begin{aligned}\mathbb{E}[G_1G_2] &= \mathbb{E}[G_1\mathbb{E}[G_2|G_1]] \\ &\leq \mathbb{E}[G_1 e^{-O_{\epsilon,p_s}(R_k[m-1])} + G_1 e^{-O_{\epsilon,p_s}(G_1)}] \\ &\leq \mathbb{E}[G_1 e^{-O_{\epsilon,p_s}(R_k[1])}] + O(1) \\ &\leq \mathbb{E}[G_1^2] e^{-O_{\epsilon,p_s}(R_k[1])} + O(1)\end{aligned} \tag{4.40}$$

where p_{G_1} is the PMF of G_1 and we used Proposition 4.3 for the inequality.

Combining all of these terms gives us,

$$\mathbb{E}\left[(R_{k+1}[m] - R_{k+1}[m-1])^2\right] \leq \mathbb{E}\left[(R_k[m] - R_k[m-1])^2\right]\left(1 + e^{-O_{\epsilon,p_s}(R_k[1])}\right) + O(\log m). \tag{4.41}$$

Note that we have $R_k[1] \sim O(k)$, hence we have,

$$\begin{aligned}\mathbb{E}\left[(R_k[m] - R_k[m-1])^2\right] &\leq \sum_{j=2}^{k}\prod_{l=2}^{j}(1+e^{-O(l)})O(\log m) \\ &\sim O(k\log m)\end{aligned} \tag{4.42}$$

which proved eq. 4.30 and completed the proof. □

But it is not sufficient to show that the expectation has the correct scaling; we also need to show that $R_k[M]$ concentrates to this expectation.

Proposition 4.4. *We have,*

$$\Pr\left(R_K[1] > \left(E[\Delta_\infty] + \frac{1}{p_s} + \delta_1\right)K\right) \to 0 \tag{4.43}$$

for any $\delta_1 > 0$.

Proof. We can expand $R_k[1]$ as

$$R_k[1] = \sum_{j=1}^{k-1} \Delta_j[1] + N_k\left(\sum_{j=1}^{k-1} \Delta_j[1]\right). \tag{4.44}$$

Now consider $k-1$ random variables $\{\Delta_{\infty,i}\}_{i=1}^{k-1}$ such that these random variables are i.i.d. copies of Δ_∞. One could note that regardless of the values of $\{\Delta_{j'}[1]\}_{j'<j}$ the random variable $\Delta_j[1]$ is stochastically dominated by $\Delta_{\infty,j}$, see appendix in chapter 1. Hence, we can consider a surrogate version of $R_k[1]$, namely,

$$\tilde{R}_k[1] = \sum_{j=1}^{k-1} \Delta_{\infty,j} + N_k\left(\sum_{j=1}^{k-1} \Delta_j[1]\right) \tag{4.45}$$

such that $\tilde{R}_k[1]$ stochastically dominates $R_k[1]$. We can see that $\tilde{R}_k[1]$ is composed of i.i.d. random variables, hence it is subject to the law of large number, such that we have

$$\Pr\left(\tilde{R}_k[1] > \left(\mathbb{E}[\Delta_\infty] + \frac{1}{p_s} + \delta_1\right)k\right) \to 0, \tag{4.46}$$

for any positive δ_1. Due to the stochastic domination, this tail bound on $\tilde{R}_k[1]$ also applies to $R_k[1]$. □

Hence by combining these results, we can show that the blockwise information velocity of this scheme is strictly bounded away from 0.[3]

Corollary 4.1. *We have,*

$$v\left(BEC(\epsilon)\right) \geq \sup_{0\leq p_s<\frac{1}{1+\epsilon}-\frac{1}{2}} v_{tree}\left(BEC(\epsilon); p_s\right) \tag{4.47}$$

with $v_{tree}\left(BEC(\epsilon); p_s\right) = \frac{1}{\mu+\frac{1}{p_s}}$ *where* $\mu = \mathbb{E}[\Delta_\infty]$.

Proof. With $N = K(\mu + 1/p_s + \delta_1 + \delta_2)$, $\delta_1, \delta_2 > 0$, the decoding scheme succeeds if (i) $R_K[1] \leq K(\mu + 1/p_s + \delta_1)$ and (ii) $R_K[M] - R_K[1] \leq \delta_2 K$, as these imply that $R_K[M] \geq N$. Using Proposition 4.4, we show that the complement of (i) has vanishing probability as K gets large. By Theorem 4.1 and Markov's inequality, the complement of (ii) also vanishes. As δ_1 and δ_2 are arbitrary positive numbers, this proves the lower bound. □

[3] During author's discussion with his colleague in physics, the colleague has brought to our attention the similarity of the process analyzed here and the problem of determining the speed of asymmetric simple exclusion process, for example results in (Aggarwal et al., 2023). Unfortunately, this discussion happens late enough into the writing process and we do not have the time to properly assess the connection or whether there are techniques that can be adapted to our problem.

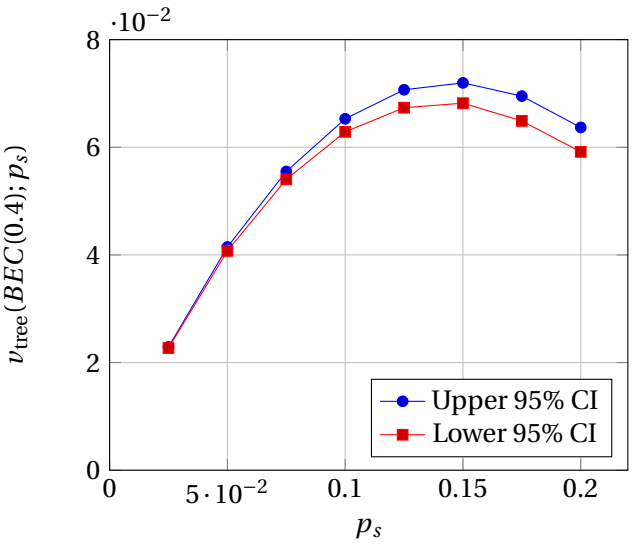


Figure 4.2: Trade-off between the ensemble speaking probability vs the achievability result in eq. 4.47. This result is obtained through monte-carlo simulations, hence we plot the 95% confidence interval range.

4.3 Discussion

4.3.1 Optimizing the Tree Code Ensemble

The results of this chapter show that the information velocity term is dominated by the transmission delay of the first bit.

In Proposition 4.1, we present an upper bound to the expectation of this delay. As we later remarked, this upper bound is pretty loose. As far as we know, there is no closed-form formula for the expected delay. Hence the maximization must be done numerically. To gain some insights to this quantity, we used Monte Carlo simulations to estimate this expectation, the result can be seen in Figure 4.2 for BEC(0.4).

The concave shaped of the curve can be explained by the trade-off between two phenomena: i) as p_s decreases the difference between the time when an unerased bit is received at the relay and the closest future speaking time increases, ii) as p_s increases, the effective rate of the tree codes increased and therefore the number of competing branches during the decoding process also increased, leading to longer delay.

4.3.2 Numerical Results

To corroborate our theoretical results, we also conducted Monte Carlo simulations of our scheme for BEC(0.4).

Our main results gives an upper bound which scales as $O(\sqrt{k})$ for the expected time difference

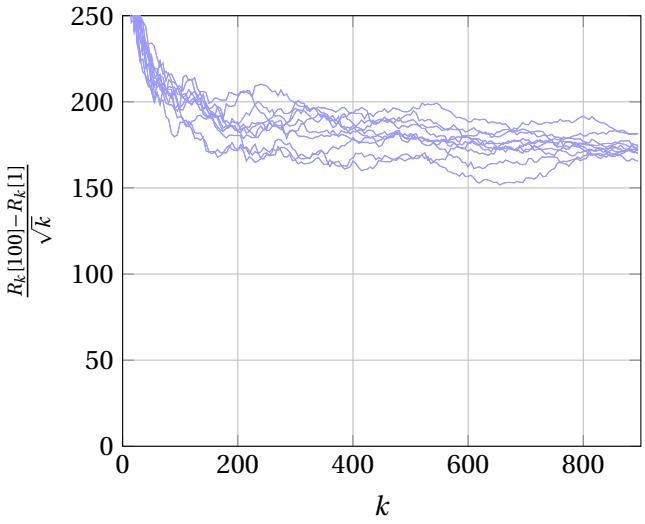


Figure 4.3: Time difference between the first and last message bit reception normalized by $\sqrt{k}$ for $p_s = 0.15$.

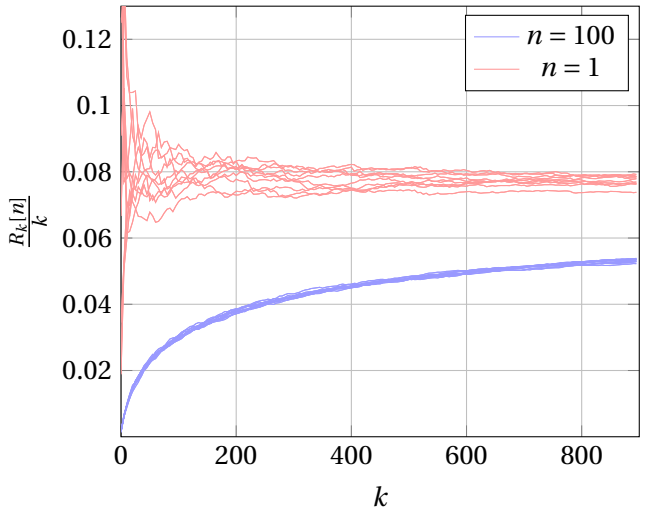


Figure 4.4: Reception time for the first and last message bit normalized by k for $p_s = 0.15$.

between the receipt of the first and the last bit. However, this leaves the question on whether the order of this upper bound is tight. Surprisingly, we can see in Figure 4.3 that it seems that the order of our upper bound is most likely tight. This could motivate further research in finding the matching lower bound.

In Figure 4.4 we show the scaling of the receiption time of the first and the last bit also tends to agree with our result in eq. 4.47. In our previous section, we showed that for $p_s = 0.15$ we would expect that the designed information velocity i.e., $1/(1/p_s + \mathbb{E}[\Delta_\infty])$ is roughly around 0.07.

4.3.3 Feasibility of Applying Tree Codes in Practical Setting

One of the main concerns in applying tree codes to a practical setting is the amount of memory needed to store the codebook for the encoding process and to store the candidate paths for the decoding process. For the decoding process, this problem is not a significant concern as it can be shown that the expected number of candidate paths is bounded for BEC, but this might be a concern if we want to generalize the scheme to other channels. There have been several attempts to circumvent this problem. For example, in regular tree codes, there has been an attempt to assign the label of the edges based on LTI codes with a sequential decoding algorithm, e.g., (Khina et al., 2016), to reduce complexity. But we cannot adopt this technique for the tree codes in our scheme as the branching process is randomized in our code.

Another commonly used approximation is to approximate tree codes by using convolutional codes with finite memory. However, this approach is also not suitable for our scheme. In our scheme, each bit is transmitted sequentially. Hence any error induced in the tape can cause catastrophic errors as all bit positions are now shifted. Tree codes ensembles have an advantage in this regard, as any decoding error will cause the decoded branch to differ from the transmitted branch, hence eventually the receiver can correct the decoding error. This behavior cannot be replicated by any channel codes with finite memory.

Hence the question of the feasibility of applying this scheme in a practical setting is still open.

5 Blockwise Information Velocity in AWGN Channels with Feedback

> When the facts change, I change my mind. What do you do, sir?
>
> John Maynard Keynes

In channel coding problems, it is well known that feedback does not increase the capacity of memoryless channels. However, it significantly reduces the complexity of achievability scheme. For example, Horstein's method (Horstein, 1963), or Schalkwijk-Kailath scheme (Schalkwijk and Kailath, 1966) achieve the channel capacity with a relatively simple scheme. In fact, an even simpler scheme, where the feedback is only used to send a "terminating" message, is also shown to achieve the full benefit of having a noiseless feedback channel (Polyanskiy et al., 2011).

The apparent ease of finding a good scheme for channels with feedback motivates us to study the problem of information velocity for channels with feedback, more specifically for the Additive White Gaussian Noise channels.

In this chapter, we will define the relaying problems with feedback. We will then define and analyze our modification to Schalkwijk-Kailath (SK) scheme. We will show that this scheme achieves an information velocity of $1/(1+\sigma^2)$. A similar result is also derived independently in (Domanovitz et al., 2023). We will then review the information velocity that we obtained against what we know about the contraction coefficient of AWGN channels. Finally, we will discuss what we will refer to as heterogeneous AWGN relaying problems with feedback, namely the relaying problems where the channel between each hop is an AWGN channel with variance that might be different for each hop.

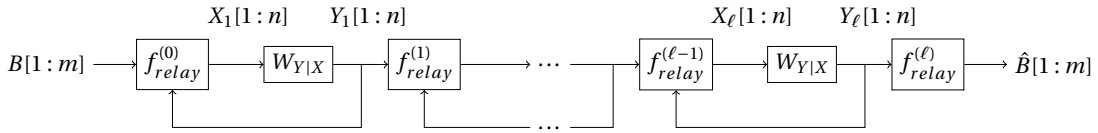


Figure 5.1: AWGN relaying problem with feedback. Here, $W_{Y|X}$ is $AWGN(\sigma^2)$.

5.1 Schalkwijk-Kailath Scheme on AWGN Relaying Problems with Feedback

Let us consider a relaying problem where the channel in each hop is an AWGN(σ^2). We introduce noiseless feedback at each hop, such that the k-th relay at time n has access to $Y_k[1:n-1]$. The illustration of the relaying problem that we will consider is given in Figure 5.1.

Our method to show achievability is based on the Schalkwijk-Kailath scheme for point-to-point communication with feedback.

In the SK scheme, the sender maps the message to equidistant points in the real line, while still respecting the power constraint. The transmitter sends this point at its first channel use, and the receiver recovers the Mean Mean Square Error (MMSE) estimate of the the transmitter's transmission. This estimate is of course noisy. At the subsequent transmissions, the transmitter sends the difference between its actual transmission at the previous time instant and receiver's estimate of its transmission at the previous time instant (which it can recreate due to the noiseless feedback). One can show that the variance of this difference decays doubly exponential with respect to the number of channel uses.

Let us define a random variable $U_i[n]$ to represent the value that the i-th relay wants to transmit at time t. For the first relay (i.e., the source), this random variable is completely equivalent to the SK scheme on point-to-point channel, i.e.,

$$\begin{aligned} U_1[1] &= f_{enc}(B[1:m]) \\ U_1[n] &= U_1[n-1] - \tilde{U}_1[n-1], \qquad \text{for } i > 1. \end{aligned} \tag{5.1}$$

The transmitter scales these values and transmit them through the channel, and the receiver forms the following estimates of the transmitted values,

$$\begin{aligned} X_1[n] &= \frac{U_1[n]}{\sqrt{\mathrm{Var}(U_1[n])}} \\ \tilde{U}_1[n] &= \sqrt{\mathrm{Var}(U_1[n])}\frac{Y_1[n]}{1+\sigma^2}. \end{aligned} \tag{5.2}$$

The function $f_{enc} : \{0,1\}^m \to \mathbb{R}$ is chosen such that the message point is mapped in an equidistant manner, while fulfilling $\mathbb{E}[U_1[1]] = 0$ and $\mathrm{Var}(U_1[1]) = 1$.

However, for the second relay onward, we introduce a small modification to the SK scheme, namely we change the value that the relay transmits as,

$$U_k[i] = \tilde{U}_{k-1}[i-1] + U_k[i-1] - \tilde{U}_k[i-1]. \tag{5.3}$$

Finally, for our decoding function, we decode by finding the message point that is closest to $\sum_{i=1}^{n-1} \tilde{U}_\ell[i]$.

Proposition 5.1. *Given a sequence $\ell(1), \ldots, \ell(n)$ such that*

$$\lim_{n\to\infty} \frac{\ell(n)}{n} = 1 - \alpha \tag{5.4}$$

for $(1+1/\sigma^2)^{-1} < \alpha < 1$, then,

$$\lim_{n\to\infty} -\frac{1}{n} \log Var\left(f_{enc}(B[1:m]) - \sum_{i=1}^{n-1} \tilde{U}_{\ell(n)}[i] \right) \geq d_2(\alpha; (1+1/\sigma^2)^{-1}). \tag{5.5}$$

Before we show the proof, we need the following fact about the independence structure of the error terms in the SK scheme.

Lemma 5.1. *For all n, we have $(U_i[n] - \tilde{U}_i[n]) \perp (U_j[n] - \tilde{U}_j[n])$ where $i \neq j$.*

Proof. We will prove by induction. First, we want to prove the relation $U_i[n] \perp U_j[n]$ if $i \neq j$. By construction, the property holds for $n = 1$.

Let us assume that the statement is true for $\{U_i[n-1]\}_{i=1}^{\ell}$. To show that it also holds for $\{U_i[n]\}_{i=1}^{\ell}$, we will show that the components of $U_i[n]$ and $U_j[n]$ for $i \neq j$, i.e., $(\tilde{U}_{i-1}[n-1], U_i[n-1] - \tilde{U}_i[n-1])$ and $(\tilde{U}_{j-1}[n-1], U_j[n-1] - \tilde{U}_j[n-1])$, are also independent.

We have $\tilde{U}_{i-1}[n-1] \perp \tilde{U}_{j-1}[n-1]$ as the induction hypothesis implies that $U_{i-1}[n-1] \perp U_{j-1}[n-1]$. This also implies that $U_{i-1}[n-1] - \tilde{U}_{i-1}[n-1] \perp U_j[n-1]$ if $i-1 \neq j$. In the case that $i-1 = j$, we also have $U_{i-1}[n-1] - \tilde{U}_{i-1}[n-1] \perp \tilde{U}_j[n-1]$ due to the property of minimum mean square error estimate for values transmitted through AWGN. Finally, we also have $U_{i-1}[n-1] - \tilde{U}_{i-1}[n-1] \perp U_{j-1}[n-1] - \tilde{U}_{j-1}[n-1]$ due to the induction assumption, as $\tilde{U}_i[n]$'s is linear combination of $U_i[n]$'s with independent noises. Hence this completes the induction, while also showing that this induction implies the proposition. □

Now, we are ready to give a proof of Proposition 5.1.

Proof. First of all, let us analyze the scheme for a fixed ℓ. From the definition of $U_k[i]$, we have,

$$\tilde{U}_k[i-1] = \tilde{U}_{k-1}[i-1] - (U_k[i] - U_k[i-1]). \tag{5.6}$$

Summing both sides from $i = 1$ to $i = n$ yields,

$$\begin{aligned}\sum_{i=1}^{n-1} \tilde{U}_k[i] &= \sum_{i=1}^{n-1} \tilde{U}_{k-1}[i] - U_k[n] \\ &= \sum_{i=1}^{n-2} \tilde{U}_{k-1}[i] - (U_k[n] - \tilde{U}_{k-1}[n-1]).\end{aligned} \tag{5.7}$$

We also have from the definition of $U_k[i]$,

$$U_k[i] - \tilde{U}_{k-1}[i-1] = U_k[i-1] - \tilde{U}_k[i-1]. \tag{5.8}$$

Hence, we have

$$\begin{aligned}\sum_{i=1}^{n-1} \tilde{U}_k[i] &= \sum_{i=1}^{n-2} \tilde{U}_{k-1}[i] - (U_k[n] - \tilde{U}_{k-1}[n-1]) \\ &= \sum_{i=1}^{n-2} \tilde{U}_{k-1}[i] - (U_k[n-1] - \tilde{U}_k[n-1]).\end{aligned} \tag{5.9}$$

Repeatedly substituting $\sum_{i=1}^{n-2} \tilde{U}_{k-1}[i]$ in this recurrence relation gives us,

$$\sum_{i=1}^{n-1} \tilde{U}_\ell[i] = \sum_{i=1}^{n-\ell} \tilde{U}_1[i] - \sum_{k=2}^{\ell} (U_k[n-\ell+k-1] - \tilde{U}_k[n-\ell+k-1]). \tag{5.10}$$

Substituting $f_{enc}(B[1:m]) - U_1[n-\ell+1]$ for $\sum_{i=1}^{n-\ell} \tilde{U}_1[i]$ gives us,

$$\begin{aligned}\sum_{i=1}^{n-1} \tilde{U}_\ell[i] &= f_{enc}(B[1:m]) - U_1[n-\ell+1] - \sum_{k=2}^{\ell} (U_k[n-\ell+k-1] - \tilde{U}_k[n-\ell+k-1]) \\ &= f_{enc}(B[1:m]) - \sum_{k=1}^{\ell} (U_k[n-\ell+k-1] - \tilde{U}_k[n-\ell+k-1]).\end{aligned} \tag{5.11}$$

Therefore, we have,

$$f_{enc}(B[1:m]) - \sum_{i=1}^{n-1} \tilde{U}_\ell[i] = \sum_{k=1}^{\ell} (U_k[n-\ell+k-1] - \tilde{U}_k[n-\ell+k-1]). \tag{5.12}$$

As all $U_k[i]$ are linear combinations of Gaussians, it is easy to see that the error term, i.e., $f_{enc}(B[1:M]) - \sum_{i=1}^{n-1} \tilde{U}_\ell[i]$, is also distributed as a Gaussian. Furthermore, taking expectation on both sides of the equation implies that the error term is a zero-centered Gaussian. Thus, we only need to understand the variance of this quantity.

Due to this independence structure in Lemma 5.1, we have,

$$f_{enc}(B[1:M]) - \sum_{i=1}^{n-1} \tilde{U}_\ell[i] = \sum_{k=1}^{\ell} \mathrm{Var}\left((U_k[n-\ell+k-1] - \tilde{U}_k[n-\ell+k-1])\right) \tag{5.13}$$

and the following relations between the variance terms,

$$\mathrm{Var}(U_k[t]) = \begin{cases} 1 & k=1, t=1 \\ 0 & t<k \\ \mathrm{Var}(\tilde{U}_{k-1}[t-1]) + \mathrm{Var}(U_k[t-1] - \tilde{U}_k[t-1]) & \text{otherwise} \end{cases} \tag{5.14}$$

where,

$$\begin{aligned} \mathrm{Var}(U_k[t] - \tilde{U}_k[t]) &= \frac{\sigma^2}{1+\sigma^2}\mathrm{Var}(U_k[t]) \\ \mathrm{Var}(\tilde{U}_k[t]) &= \frac{1}{1+\sigma^2}\mathrm{Var}(U_k[t]). \end{aligned} \tag{5.15}$$

To ease notation and simplify the boundary condition, let us denote $S_k[t] = \mathrm{Var}(U_k[t+k])$. So we have,

$$S_k[t] = \begin{cases} 1 & k=1, t=0 \\ 0 & k>1, t<0 \\ \frac{1}{1+\sigma^2} S_{k-1}[t] + \frac{\sigma^2}{1+\sigma^2} S_k[t-1] & \text{otherwise.} \end{cases} \tag{5.16}$$

Playing around with small values of k and t, one quickly notices that this summation is a similar summation that defines the Pascal's Triangle. Hence the solution to this recurrence relation is given by,

$$S_k[n] = \binom{k+n}{n} \sigma^{2n} (1+\sigma^2)^{-(k+n)}. \tag{5.17}$$

A more formal argument can be obtained by using induction. Therefore we have that

$$\begin{aligned} \mathrm{Var}\left(f_{enc}(B[1:m]) - \sum_{i=1}^{n-1} \tilde{U}_\ell[i] \right) &= \frac{\sigma^2}{1+\sigma^2} \sum_{k=1}^{\ell} \mathrm{Var}(U_k[n-\ell-1+k]) \\ &= \frac{\sigma^2}{1+\sigma^2} \sum_{k=1}^{\ell} S_k[n-\ell-1] \end{aligned} \tag{5.18}$$

and correspondingly,

$$\log \mathrm{Var}\left(f_{enc}(B[1:m]) - \sum_{i=1}^{n-1} \tilde{U}_\ell[i] \right) = -(n-\ell)\log(1+1/\sigma^2) + \log\sigma^2 + \log \sum_{k=1}^{\ell} \binom{k+n-\ell-1}{n-\ell-1} \left(1+\sigma^2\right)^{-k}. \tag{5.19}$$

Now let us assume that n and $\ell(n)$ also grows with the following proportion,

$$\lim_{n\to\infty} \frac{\ell(n)}{n} = 1-\alpha \tag{5.20}$$

with $0 < \alpha < 1$.

Then we have,

$$\lim_{n\to\infty}\frac{1}{n}\log\mathrm{Var}\left(f_{enc}(B[1:m])-\sum_{i=1}^{n-1}\tilde{U}_{\ell_n}[i]\right)$$
$$\leq -\alpha\log(1+1/\sigma^2)+\sup_{0<\beta<1-\alpha}(\alpha+\beta)h_2\left(\frac{\alpha}{\alpha+\beta}\right)-\beta\log\left(1+\sigma^2\right). \tag{5.21}$$

The variational problem is concave, so we can find the maximum by taking the derivative. This gives us the value of optimal $\beta=\min\{\sigma^2\alpha,1-\alpha\}$. Plugging $\beta=\sigma^2\alpha$, we can see that the upper bound is equal to 0. On the other hand, if we plug $\beta=1-\alpha$, the upper bound is equal to $-d_2(\alpha;1/(1+1/\sigma^2))$. □

In essence, Proposition 5.1 gives a threshold on the growth rate of the relaying channel length if we want to ensure that the error of the receiver estimates decays exponentially. This implies a lower bound on the information velocity of the Gaussian channel.

Corollary 5.1. *We have a lower bound on the information velocity of AWGN channels with feedback, i.e.,*

$$v\left(AWGN(\sigma^2)\right)\geq\frac{1}{1+\sigma^2}. \tag{5.22}$$

Proof. If the distance of each message point is $\Delta=O(2^{-|m|})$, then the error probability is given by,

$$\Pr(B[1:m]\neq\hat{B}[1:m])=\Pr\left(\left|f_{enc}(B[1:m-\sum_{i=1}^{n-1}\tilde{U}_{\ell_n}[i]\right|\geq\frac{\Delta}{2}\right)$$
$$\leq 2\exp\left(-\frac{\Delta^2}{8\sigma_e^2}\right) \tag{5.23}$$

where σ_e^2 is the variance of $\left|f_{enc}(B[1:M])-\sum_{i=1}^{n-1}\tilde{U}_{\ell_n}[i]\right|$. From the previous proposition, we know that if $(1+\sigma^2)^{-1}<\alpha$ and

$$\lim_{n\to\infty}\frac{\ell(m,n)}{n}=1-\alpha \tag{5.24}$$

then the upper bound can be expressed as,

$$\Pr(B[1:M]\neq\hat{B}[1:M])\leq 2\exp\left(-\frac{\Delta^2 e^{d_2(\alpha;(1+1/\sigma^2)^{-1})n}}{8}\right) \tag{5.25}$$

We can see that this implies that that the upper bound to the error probability vanishes as $n\to\infty$ provided that,

$$\lim_{n\to\infty}\frac{\ell(m,n)}{n}<\frac{1/\sigma^2}{1+\frac{1}{\sigma^2}}. \tag{5.26}$$

This implies that this scheme is a valid achievability scheme. Thus proving the proposition. □

Astute readers might feel a sense of déjà vu here. Indeed, the structure of the argument for Proposition 5.1 is closesly related to the argument that we used to prove the converse in chapter 3. In fact, structurally, the main difference between these two lies only on the boundary conditions. For proving the converse, we held the boundary of $k = 0$ to be equal to a positive value (representing the condition where the source has complete information at the beginning of the transmission) and the boundary of $n = 0, k > 0$ is equal to 0 (representing the condition where all other relays begin with no information). However, in this case, the boundary $k = 0$ (representing the condition that the message point at the source has zero variance at the beginning of the transmission), and the other relays have non-zero variance at the beginning of the transmission.

5.2 Comparison of Achievability Result with Converse Result

Having established an achievability result on the information velocity of a Gaussian channel with feedback, it is natural to ask about the relationship between this result and the converse result developed in chapter 3.[1]

At first glance, the result from chapter 3 merely provides a trivial bound on the information velocity. As we know that the channel contraction coefficient of AWGN is equal to 1 for KL divergence (Polyanskiy and Wu, 2016).

However, we can argue that the information velocity of our scheme is optimal if we consider the set of relaying schemes where the distribution of the relay output is a Gaussian random variable. More formally, we have the following proposition.

Proposition 5.2. *Consider a AWGN relaying problem with feedbacks, for all k and $n > 1$, $Q_{X_i[n]|Y_{i-1}[1:n-1]} \sim N(0,1)$, then*

$$v\left(AWGN(\sigma^2)\right) \leq \frac{1}{1+\sigma^2}. \tag{5.27}$$

In other words, if the hypothesis of the proposition holds, namely if the distribution of the relay's emission conditioned on its observations at the previous times is Gaussian, then our achievability scheme is optimal. In general, we would expect that this condition holds for every schemes which communicates at channel capacity; as Gaussian is the capacity-achieving distribution for AWGN[2].

Proposition 5.2 can be thought as a corollary of Proposition 3.10, as long as we can prove that the channel contraction coefficient under the condition of the proposition is equal to

[1] Readers might have doubts regarding applicability of the converse result to relaying problems with feedback. But the converse is indeed valid. After all, the motivation behind the tilting operation is to allow us to upper bound the k-th relay's information over the conditioning by the $k+1$-th relay's observations on the previous time indices.

[2] For schemes with discrete message alphabet, this statement is not entirely accurate, as proven by the ongoing research on constellation design.

$1/(1+\sigma^2)$. We will show this fact on the following proposition. This proposition is originally proved in (Makur and Zheng, 2015) which we modify for our purpose. The original result is more general and applies to any marginal distribution in the exponential family with moment constraints.

Proposition 5.3. *Consider a Markov chain $M \to X \to Y$ where the channel $W_{Y|X}$ is an AWGN channel, $Y = X + Z$, $Z \perp\!\!\!\perp X$, $Z \sim N(0,\sigma_Z^2)$. Let $\mathscr{A}$ bet the set of admissible distributions $Q_{X,M}$ such that $Q_X \sim N(0,\sigma_X^2)$. Then we have,*

$$\sup_{Q_{X,M}\in\mathscr{A}} \frac{I_f(Q_M, Q_{Y|M})}{I_f(Q_M, Q_{X|M})} \le \frac{\sigma_X^2}{\sigma_X^2+\sigma_Z^2}. \tag{5.28}$$

Proof. We have,

$$\begin{aligned}\sup_{Q_{X,M}\in\mathscr{A}} \frac{I_f(Q_M; Q_{Y|M})}{I_f(Q_M; Q_{X|M})} &= \sup_{Q_{X,M}\in\mathscr{A}} \frac{\mathbb{E}_{Q_M}[D_f(Q_{Y|M}; Q_Y)]}{\mathbb{E}_{Q_M}[D_f(Q_{X|M}; Q_X)]} \\ &\le \sup_{P_X, \mathbb{E}_{P_X}[X^2]<\infty} \frac{D_f(W_{Y|X}\circ P_X; Q_Y)}{D_f(P_X; Q_X)}\end{aligned} \tag{5.29}$$

Let $P_Y = W_{Y|X}\circ P_X$. Expanding the divergence terms, we have for all P_X,

$$\begin{aligned}D(P_X;Q_X) &= \frac{1}{2}\log(2\pi\sigma_X^2) + \frac{1}{2}\frac{\mathbb{E}_{P_X}[X^2]}{\sigma_X^2} - h(P_X) \\ &= h(N(0,\sigma_X^2)) - h(P_X) + \frac{1}{2}\left(\frac{\mathbb{E}_{P_X}[X^2]}{\sigma_X^2} - 1\right), \\ D(P_Y;Q_Y) &= \frac{1}{2}\log(2\pi(\sigma_X^2+\sigma_Z^2)) + \frac{1}{2}\frac{\mathbb{E}_{P_X}[X^2]+\sigma_Z^2}{\sigma_X^2+\sigma_Z^2} - h(P_Y) \\ &= h(N(0,\sigma_X^2+\sigma_Z^2)) - h(P_Y) + \frac{1}{2}\left(\frac{\mathbb{E}_{P_X}[X^2]+\sigma_Z^2}{\sigma_X^2+\sigma_Z^2} - 1\right).\end{aligned} \tag{5.30}$$

We want to show that

$$(\sigma_X^2+\sigma_Z^2)D(P_Y;Q_Y) \overset{?}{\le} \sigma_X^2 D(P_X;Q_X). \tag{5.31}$$

If we expand the divergence terms explicitly, we obtain

$$\begin{aligned}&(\sigma_X^2+\sigma_Z^2)\left[h(N(0,\sigma_X^2+\sigma_Z^2)) - h(P_Y)\right] + \frac{\mathbb{E}_P[X^2]-\sigma_X^2}{2} \\ &\overset{?}{\le} \sigma_X^2\left[h(N(0,\sigma_X^2)) - h(P_X)\right] + \frac{\mathbb{E}_P[X^2]-\sigma_X^2}{2}.\end{aligned} \tag{5.32}$$

By removing identical terms on both sides gives us,

$$(\sigma_X^2+\sigma_Z^2)\left[h(N(0,\sigma_X^2+\sigma_Z^2)) - h(P_Y)\right] \overset{?}{\le} \sigma_X^2\left[h(N(0,\sigma_X^2) - h(P_X)\right]. \tag{5.33}$$

The next step is to transform this inequality such that we can use the entropy-power inequality. We have,

$$\left(\frac{e^{h(N(0,\sigma_X^2+\sigma_Z^2))}}{e^{h(P_Y)}}\right)^{(\sigma_X^2+\sigma_Z^2)} \overset{?}{\leq} \left(\frac{e^{h(N(0,\sigma_X^2))}}{e^{h(P_X)}}\right)^{(\sigma_X^2)} \tag{5.34}$$

By plugging the differential entropy of the Gaussian, we have,

$$\left(\frac{2e\pi(\sigma_X^2+\sigma_Z^2)}{e^{h(P_Y)}}\right)^{\sigma_X^2+\sigma_Z^2} \overset{?}{\leq} \left(\frac{2e\pi\sigma_X^2}{e^{h(P_X)}}\right)^{\sigma_X^2}. \tag{5.35}$$

Note that this equation is equivalent with eq. 5.31, in the sense that we have not loosen the inequality so far. To prove this inequality, we will divide this inequality into two inequalities,

$$\left(\frac{2e\pi(\sigma_X^2+\sigma_Z^2)}{e^{h(P_Y)}}\right)^{\sigma_X^2+\sigma_Z^2} \overset{(i)}{\leq} \left(\frac{2e\pi(\sigma_X^2+\sigma_Z^2)}{e^{h(P_X)}+2e\pi\sigma_Z^2}\right)^{\sigma_X^2+\sigma_Z^2} \overset{(ii)}{\leq} \left(\frac{2e\pi\sigma_X^2}{e^{h(P_X)}}\right)^{\sigma_X^2}. \tag{5.36}$$

Note that we have inequality (i), i.e.,

$$\left(\frac{2e\pi(\sigma_X^2+\sigma_Z^2)}{e^{h(P_Y)}}\right)^{\sigma_X^2+\sigma_Z^2} \leq \left(\frac{2e\pi(\sigma_X^2+\sigma_Z^2)}{e^{h(P_X)}+2e\pi\sigma_Z^2}\right)^{\sigma_X^2+\sigma_Z^2} \tag{5.37}$$

by the entropy-power inequality, i.e.,

$$e^{h(X)}+e^{h(Z)} \leq e^{h(X+Z)}. \tag{5.38}$$

Namely that we can ensure that the denominator at the RHS is smaller than the LHS.

Therefore, we only need to show inequality (ii), i.e.,

$$\left(\frac{2e\pi(\sigma_X^2+\sigma_Z^2)}{e^{h(P_X)}+2e\pi\sigma_Z^2}\right)^{\sigma_X^2+\sigma_Z^2} \leq \left(\frac{2e\pi\sigma_X^2}{e^{h(P_X)}}\right)^{\sigma_X^2}, \tag{5.39}$$

then we proved the proposition for all P_X. This inequality is equivalent to showing that

$$\left(\frac{a+\gamma}{b+\gamma}\right)^{a+\gamma} \leq \left(\frac{a}{b}\right)^{a} \tag{5.40}$$

where a,b, and γ are positive values. Let us define a function

$$f(\gamma) = \left(\frac{a+\gamma}{b+\gamma}\right)^{a+\gamma}. \tag{5.41}$$

Our goal is to show,

$$\frac{\partial f}{\partial \gamma} \leq 0 \tag{5.42}$$

for $\gamma \geq 0$. We have

$$\frac{\partial f}{\partial \gamma} = f(\gamma)\left[\log\frac{a+\gamma}{b+\gamma} + \left(1 - \frac{a+\gamma}{b+\gamma}\right)\right] \tag{5.43}$$

which is always non-positive as $f(\gamma) > 0$ due to $\log x \leq x - 1$. □

5.3 Information Velocity on Heterogeneous AWGN Relaying Problems with Feedback

We can extend the model in the previous section such that the noise power of the AWGN channel at each hop is different. To adapt this model to our analysis in the previous section, we also need to assume that each relay has complete knowledge of the variance of all channels. Due to this assumption, each relay can track the variance of $U_i[k]$'s and therefore it can compute the optimal combining weight. Hence, Lemma 5.1 holds for this model.

More formally, we will consider a heterogeneous AWGN relaying model, where we have

$$Y_i[t] = X_i[t] + Z_i[t] \tag{5.44}$$

where $Z_i[t]$ is independent of all other random variables and distributed according to the distribution $N(0,\sigma_i^2)$. However, the $\{\sigma_i^2\}_{i=1}^{\ell}$ are i.i.d. discrete random variables on a finite alphabet. Let us denote the distribution of this random variable as P_{σ^2}. We still use the SK scheme where we have

$$\begin{aligned} U_1[1] &= f_{enc}(B[1:m]) \\ U_1[i] &= U_1[i-1] - \tilde{U}_1[i-1], \qquad \text{for } i > 1 \\ X_1[i] &= \frac{U_1[i]}{\sqrt{\mathrm{Var}(U_1[i])}}, \end{aligned} \tag{5.45}$$

and for $k > 1$,

$$\begin{aligned} U_k[i] &= \tilde{U}_{k-1}[i-1] + U_k[i-1] - \tilde{U}_k[i-1] \\ X_k[i] &= \frac{U_k[i]}{\sqrt{\mathrm{Var}(U_k[i])}}. \end{aligned} \tag{5.46}$$

The only difference being that $\tilde{U}_k[i]$ denotes $\mathbb{E}[U_k[i] \mid Y_k[i], \{\sigma_j^2\}_{j=1}^{k}]$.

As in the previous section, we will also use $S_k[t] = \mathrm{Var}(U_k[t+k])$. So we have,

$$S_k[t] = \begin{cases} 1 & k = 1, t = 1 \\ 0 & k > 1, t <= 1 \\ \frac{1}{1+\sigma_{k-1}^2} S_{k-1}[t] + \frac{\sigma_k^2}{1+\sigma_k^2} S_k[t-1] & \text{otherwise.} \end{cases} \tag{5.47}$$

The structure of this system of equations is given in Figure 5.2. We can also express the closed

form of $S_k[n]$'s in a more concise manner using the notion of composition from combinatorics.

Definition 5.1. *A k-composition of an integer n is defined as*

$$Comp(k,n) = \left\{ (x_1,\ldots,x_k) \in \mathbb{Z}_{\geq 0}^k : \sum_{i=1}^{k} x_i = n \right\}. \tag{5.48}$$

Proposition 5.4. *We have*

$$S_k[t] = (1+\sigma_k^2) \sum_{(x_i)_{i=1}^k \in Comp(k,t)} \prod_{i=1}^{k} \left(\frac{\sigma_i^2}{1+\sigma_i^2} \right)^{x_i} \left(\frac{1}{1+\sigma_i^2} \right). \tag{5.49}$$

Proof. We will prove by induction. The base case is trivial, hence we only need to prove the recurrence relation.

Let us introduce the following notation,

$$\mathrm{Comp}^{(0)}(k-1,n) = \left\{ (x_1,\ldots,x_{k-1},0) \in \mathbb{Z}_{\geq 0}^k : \sum_{i=1}^{k-1} x_i = n \right\} \tag{5.50}$$

and

$$\mathrm{Comp}^{(1)}(k,n-1) = \left\{ (x_1,\ldots,x_k+1) \in \mathbb{Z}_{\geq 0}^k : \sum_{i=1}^{k} x_i = n-1 \right\}. \tag{5.51}$$

We then have,

$$\mathrm{Comp}(k,n) = \mathrm{Comp}^{(0)}(k-1,n) + \mathrm{Comp}^{(1)}(k,n-1). \tag{5.52}$$

Assuming that the relation holds for $S_k[t-1]$ and $S_{k-1}[t]$, then we have,

$$\frac{\sigma_k^2}{1+\sigma_k^2} S_k[t-1] = (1+\sigma_k^2) \sum_{(x_i)_{i=1}^k \in \mathrm{Comp}^{(1)}(k,t-1)} \prod_{i=1}^{k} \left(\frac{\sigma_i^2}{1+\sigma_i^2} \right)^{x_i} \left(\frac{1}{1+\sigma_i^2} \right) \tag{5.53}$$

and

$$\frac{1}{1+\sigma_k^2} S_{k-1}[t] = (1+\sigma_k^2) \sum_{(x_i)_{i=1}^k \in \mathrm{Comp}^{(0)}(k-1,t)} \prod_{i=1}^{k} \left(\frac{\sigma_i^2}{1+\sigma_i^2} \right)^{x_i} \left(\frac{1}{1+\sigma_i^2} \right). \tag{5.54}$$

Adding these two terms and using eq. 5.3 gives us the recurrence relation. □

Our analysis is based on the analytical combinatorics approach, especially the following saddle-point bound lemma. Our proof approach follows the exposition in (Flajolet and Sedgewick, 2009), albeit highly simplified as we are only interested in using the result for proving bounds on generating functions.

Definition 5.2. *For a sequence of non-negative numbers $\{c_k\}_{k=0}^{\infty}$, the generating function of this*

sequence is given by

$$G(z) = \sum_{k=0}^{\infty} c_k z^k \tag{5.55}$$

for complex z. The region of convergence of $G(z)$ is the subset of complex number where the RHS exists. It is said to be polynomial if there exists k such that $c_j = 0$ for all $k \leq j$.

Furthermore, we will define an operator $[z^n]$ such that $[z^n]G(z)$ for $G(z) = \sum_{n=0}^{\infty} c_n z^n$ is equal to c_n. The following result is a classical result in complex analysis.

Theorem 5.1 (Cauchy's Coefficient Formula)**.** *For a generating function $G(z)$, then we have*

$$[z^k]G(z) = \frac{1}{2\pi \mathrm{i}} \oint \frac{G(z)}{z^{k+1}} dz \tag{5.56}$$

where the integral is taken over any counterclockwise contour in the region of convergence of $G(z)$ which encircles the origin.

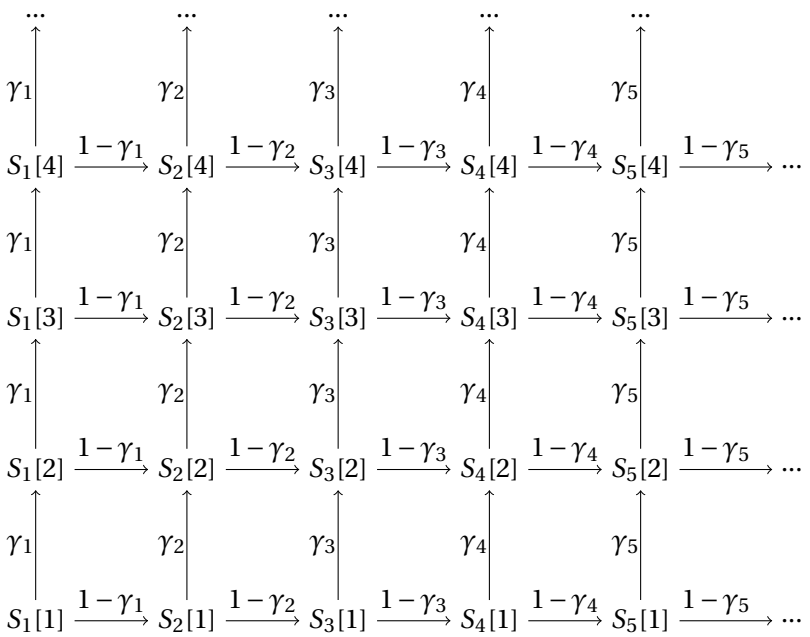


Figure 5.2: Lattice structure induced by eq. 5.47 with $\gamma_k = \frac{1}{1+\sigma_k^2}$. Note that the value of node $S_k[n]$ in the interior of this latice can be calculated as the sum of the weights of all paths taken from $S_1[1]$ to $S_k[n]$, where the weight of a path is defined as the product of the weight of all edges in the path.

Lemma 5.2 (Saddle-point Bound)**.** *Given a generating function $G(z)$ of a sequence $c_k{}_{k=0}^{\infty}$ such that $c_k \geq 0$ for all k, then we have,*

$$[z^n]G(z) \leq \frac{G(\xi)}{2\pi \xi^n} \tag{5.57}$$

for $0 < \xi < R$, where R is the radius of convergence of $G(z)$.

Proof. It is easy to see the following,

$$[z^n]G(z) = \frac{1}{2\pi \mathrm{i}} \oint \frac{G(z)}{z^{n+1}} dz \leq \frac{1}{2\pi} \oint \left| \frac{G(z)}{z^{n+1}} \right| dz \tag{5.58}$$

As this holds for any contour, let us consider a contour C where C is a counterclockwise circle of radius ξ centered on the origin. From the definition of radius of convergence, this circle is a valid contour. We have the following,

$$[z^k]G(z) \leq \frac{1}{2\pi} \oint \left| \frac{G(z)}{z^{n+1}} \right| dz \leq \frac{|C|}{2\pi \xi^{n+1}} \sup_{z \in C} |G(z)| \tag{5.59}$$

Finally, as $G(z)$ is a generating function, it can be expressed as

$$G(\xi e^{\mathrm{i}\theta}) = \sum_{k=0}^{\infty} c_n \xi^n e^{n\mathrm{i}\theta} \tag{5.60}$$

for non-negative c_n, hence its supremum is attained at $\theta = 0$. Hence the upper bound. □

We understand that there are many version of saddle point bounds, and this version is might not be the most commonly used by information theorist, but it seems that this form is what commonly referred to as the saddle-point bound in analytical combinatorics.

Proposition 5.5. *Let us assume that the variance of the heterogeneous AWGN relaying problems with feedback is given by random variable* $\{\sigma_1^2, \ldots, \sigma_{\ell(n)}^2\}$ *which is upper bounded by a constant* A*. Then we have,*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}\left[Var\left(f_{enc}(B[1:m]) - \sum_{i=1}^{n-1} \tilde{U}_{\ell(n)}[i] \right) \right] \leq \sup_{0<\alpha<\beta} \inf_{0<z<1+\frac{1}{A}} \alpha \log G(z) - (1-\beta)\log z \tag{5.61}$$

where,

$$\lim_{n\to\infty} \frac{\ell(n)}{n} = \beta. \tag{5.62}$$

Proof. Consider the expectation of $S_k[t]$ where the expectation is taken with respect to the randomness due to σ_i^2's.

$$S_k[n] = (1+\sigma_k^2) \sum_{(x_i)_{i=1}^k \in \mathrm{Comp}(k,n)} \prod_{i=1}^{k} \left(\frac{\sigma_i^2}{1+\sigma_i^2} \right)^{x_i} \frac{1}{1+\sigma_i^2}. \tag{5.63}$$

For each i, let us consider the generating function, which is valid for $0 < z < 1 + 1/A$,

$$\sum_{i=0}^{\infty} \left(\frac{\sigma_i^2}{1+\sigma_i^2} \right)^i \frac{1}{1+\sigma_i^2} z^i = \frac{1}{1+\sigma_i^2(1-z)}. \tag{5.64}$$

It is easy to see that this function does not depend on i except through σ_i^2. Hence we have the following bound,

$$\begin{aligned} S_k[n] &\le (1+\sigma_k^2)[z^n]\prod_{i=1}^{k}\sum_{i=0}^{\infty}\left(\frac{\sigma_i^2}{1+\sigma_i^2}\right)^i \frac{1}{1+\sigma_i^2}z^i \\ &\le (1+\sigma_k^2)[z^n]\prod_{i=1}^{k}\left(\frac{1}{1+\sigma_i^2(1-z)}\right) \\ &\le \frac{(1+\sigma_k^2)}{2\pi}\frac{\prod_{i=1}^{k}\left(\frac{1}{1+\sigma_i^2(1-z)}\right)}{z^n}. \end{aligned} \tag{5.65}$$

With similar argument as in the proof of Proposition 5.1, we have

$$\mathrm{Var}\left(f_{enc}(B[1:m]) - \sum_{i=1}^{n-1}\tilde{U}_\ell[i]\right) = \sum_{k=1}^{\ell}\frac{\sigma_k^2}{1+\sigma_k^2}S_k[n-\ell-1]. \tag{5.66}$$

Taking the expectation and using upper bound in 5.3, we have,

$$\begin{aligned} \mathbb{E}\left[\mathrm{Var}\left(f_{enc}(B[1:m]) - \sum_{i=1}^{n-1}\tilde{U}_\ell[i]\right)\right] &\le \mathbb{E}\left[\sum_{k=1}^{\ell}A\frac{\prod_{i=1}^{k}\left(\frac{1}{1+\sigma_i^2(1-z)}\right)}{z^{n-\ell-1}}\right] \\ &= A\sum_{k=1}^{\ell}\frac{\prod_{i=1}^{k}\mathbb{E}\left[\left(\frac{1}{1+\sigma_i^2(1-z)}\right)\right]}{z^{n-\ell-1}} \end{aligned} \tag{5.67}$$

Let us define,

$$G(z) = \mathbb{E}\left[\frac{1}{1+\sigma^2(1-z)}\right]. \tag{5.68}$$

We then have,

$$\log\mathbb{E}\left[\mathrm{Var}\left(f_{enc}(B[1:m]) - \sum_{i=1}^{n-1}\tilde{U}_\ell[i]\right)\right] \le \max_{1\le k\le \ell} k\log G(z) - (n-\ell-1)\log z + O(1). \tag{5.69}$$

Hence the proposition. □

An important consequence of this preposition is a lower bound to the information velocity.

Corollary 5.2. *We have a lower bound to information velocity of heterogeneous AWGN channels with feedback, i.e.,*

$$v(AWGN(\sigma^2)) \ge \frac{1}{1+\mathbb{E}[\sigma^2]}. \tag{5.70}$$

Proof. Let us denote the optimal point in the inner infimization in eq. 5.61 at α as $z^*(\alpha)$.

Consider the first-order condition,

$$z\frac{\partial \log G(z)}{\partial z} = \frac{1-\beta}{\alpha}. \tag{5.71}$$

Hence, we have,

$$z\frac{\mathbb{E}\left[\frac{\sigma^2}{(1+\sigma^2(1-z))^2}\right]}{\mathbb{E}\left[\frac{1}{1+\sigma^2(1-z)}\right]} = \frac{1-\beta}{\alpha}. \tag{5.72}$$

We note that $z^*(\alpha) = 1$ if $\alpha = \frac{1-\beta}{\mathbb{E}[\sigma^2]}$, $z^*(\alpha) = 1$ also implies that the exponent is equal to 0. We also note that $z^*(0) = 1 + \frac{1}{A}$. By eq. 5.72, we also have that $\frac{\partial z^*(\alpha)}{\partial \alpha} < 0$. Hence we know that $z^*(\alpha) \in (1, 1+\frac{1}{A})$ for $\alpha \in (0, \frac{1-\beta}{\mathbb{E}[\sigma^2]})$. Let us take the derivative of exponent with respect to α,

$$\frac{dE(\alpha)}{\alpha} = \frac{\partial E(\alpha)}{\partial \alpha} + \frac{\partial E(\alpha)}{\partial z^*(\alpha)}\frac{dz^*(\alpha)}{d\alpha} = \log G(z^*(\alpha)) \tag{5.73}$$

where the second term is equal to zero due to $z^*(\alpha)$ is the minimizer of the exponent. Furthermore, note that $z^*(\alpha) \in (1, 1+\frac{1}{A})$ implies that $G(z^*(\alpha)) > 1$, which implies that $0 \leq \frac{dE(\alpha)}{\alpha}$ if $\alpha \in (0, \frac{1-\beta}{\mathbb{E}[\sigma^2]})$. Hence, we have $E(\alpha) < 0$ if $\alpha \in (0, \frac{1-\beta}{\mathbb{E}[\sigma^2]})$.

Note that if $\beta < \frac{1-\beta}{\mathbb{E}[\sigma^2]}$, then $E(\alpha) < 0$ in the supremum domain of the exponent. This implies that,

$$\lim_{n\to\infty}\frac{1}{n}\log\mathbb{E}\left[\mathrm{Var}\left(f_{enc}(B[1:m]) - \sum_{i=1}^{n-1}\tilde{U}_{\ell(n)}[i]\right)\right] < 0 \tag{5.74}$$

if $\beta < \frac{1}{1+\mathbb{E}[\sigma^2]}$. Hence proving the proposition by using the same argument as in the proof of Corollary 5.1. □

5.4 Numerical Results

To corroborate the theoretical results obtained in this chapter, we conduct Monte Carlo simulations of the scheme. We simulate a heterogeneous AWGN relaying problem where σ_i^2 is uniformly sampled from $\{0.1, 0.2\}$ and the initial value $U_1[n]$ is sampled from $N(0,1)$. The result of this simulation is given in Figure 5.3.

An interesting feature of this scheme is the fact that the relays are actually only effectively transmitting on a small time window. This phenomenon can be seen in Figure 5.3, where the transition only happened between 50 to 70 channel uses. The reason for this phenomenon is due to the fact that the variance of $U_k[n]$ tends to 0 most of the time, as we can see in Figure 5.4.

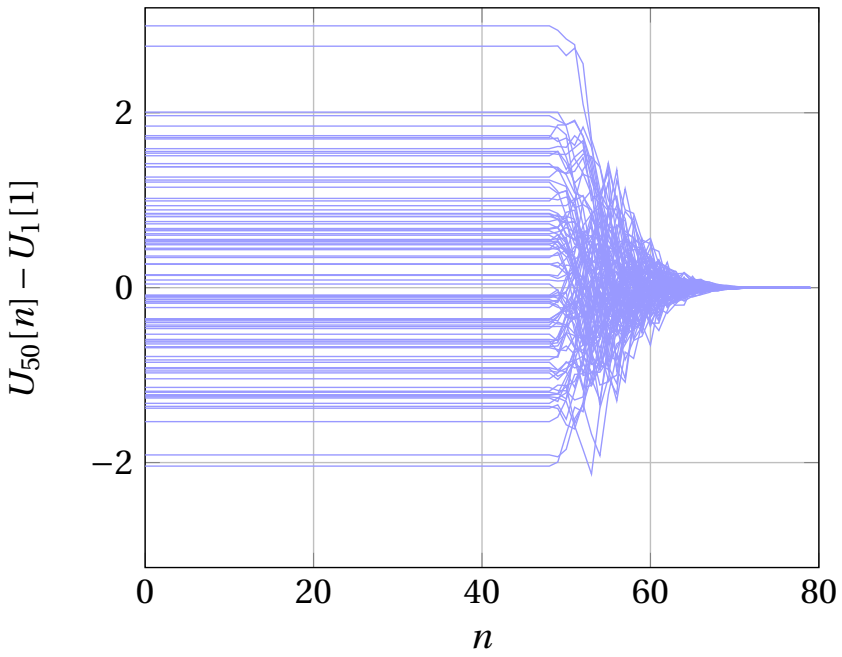


Figure 5.3: Paths generated by the Monte Carlo simulations. We can see that the value estimated by the relay converges to the correct value, with the transition happening between 50 to 70 channel uses, which is close to the value predicted by Corollary 5.2.

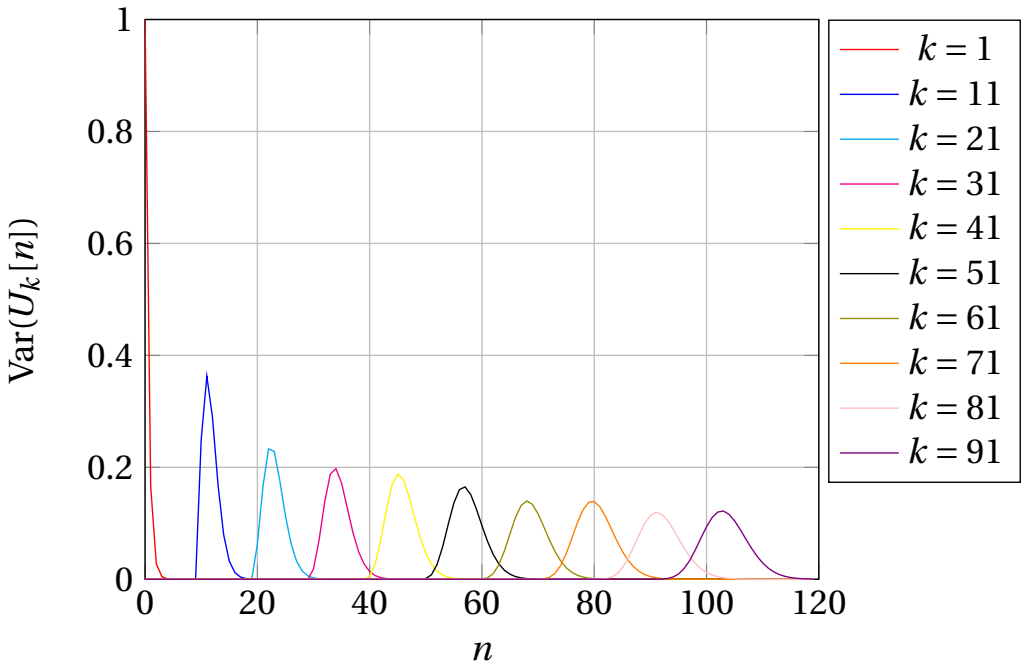


Figure 5.4: The variance of $U_k[t]$. We can see that for each relay, the variance is only significant for a small part of the transmission times.

Results on Monitoring Problems Part II

6 Monitoring Problems as Adversarial Hypothesis Testing Problems

As discussed in the first chapter, one of the main assumptions of the hypothesis testing problem is that the data-generating process is assumed to be oblivious to the testing process. This assumption might not hold if the data-generating process has some incentives to adapt its behavior to the testing process.

In this chapter, we will examine a specific form of hypothesis testing. We introduce a setting with multiple players and a hidden random variable. In this case, the tester aims to determine whether a player possesses certain side-information about the hidden random variable, essentially testing whether the player's behavior is independent of the hidden random variable. But here, we add a twist where the player with the side-information can alter its behavior to maximize the error probability.

We will present a model which captures the dynamics between the player with side-information exploiting its advantage, and the tester reducing its error probability. Several settings where this might arise include:

- In financial markets, where a regulator aims to prevent financial actors from benefiting from non-public information. In this case, these actors might want to deliberately avoid the regulator suspicions while exploiting the non-public information.

- In cybersecurity, where an intruder wants to blend in with regular users. In this case, the intruder is incentivized to imitate the behavior of regular users.

Furthermore, we will show that there is a concise representation of the strategies employed by the player with side-information and the tester under particular assumptions on the distributions of the random variables in the model.

6.1 Model Formulation

Consider a scenario where the state of an environment X is observed by $M+1$ players. Among these players, there are M regular players and a cheating player. The index of the cheating player will be denoted by H which is distributed uniformly in $\{0,\dots,M\}$.[1]

Let us denote the players' observations as $Z_0,\dots,Z_M$. For the i-th player, conditioned on X and H, the observation Z_i is independent of $\{Z_j\}_{j\neq i}$. These observations are noisy. For the regular players, i.e., $i \neq H$, these observations are obtained through identical channels characterized by $V_{Z^{(r)}|X}$. We assume that the cheating player benefits from a certain "information advantage", hence its observation is obtained through a different channel $V_{Z^{(c)}|X}$. We denote the distribution of X as V_X. We use superscripts (c) and (r) to differentiate the random variables of the cheating player and the regular player.

Each player uses its observation Z_i as a basis to take an action Y_i. In this model, the action of the i-th player can only depend on Z_i. This allows us to characterize the players' policies as (possibly probabilistic) mappings f_i for each player, so that $Y_i = f_i(Z_i)$.

A player obtains a reward from its action depending on the state of the environment. The reward that the players obtain is determined by a bounded reward function $R(X,Y)$. All players, regular and cheating, have the same reward function. We assume that $\sup_{f^{(r)}} \mathbb{E}\left[R(X,Y^{(r)})\right] \leq \sup_{f^{(c)}} \mathbb{E}\left[R(X,Y^{(c)})\right]$, i.e., the cheating player can obtain a greater expected reward if it is allowed to act without any constraint.

The cheating player is constrained by a monitoring entity who gets to observe X and noisy observations of the players' actions. Let us denote the monitor observations of the i-th player action as $\tilde{Y}_i$. We also assume that conditioned on Y_i, the observation $\tilde{Y}_i$ is independent of $\{Z_j\}_{j=0}^M, \{Y_j\}_{j\neq i}^M$ and $\{\tilde{Y}_j\}_{j\neq i}^M$. We assume that this observation is distributed as $V_{\tilde{Y}_i|Y_i}$. Based on $\left(X, \{\tilde{Y}_i\}_{i=0}^M\right)$, the monitor gives a guess of the cheating player's index. Let us denote this guess as $\hat{H}$.

We assume that the regular players disregard the possibility of being falsely accused. Hence the reward of a regular player is independent of $\hat{H}$ and is given by $\mathbb{E}\left[R(X,Y^{(r)})\right]$. For the cheating player, it can only reap its reward if it is not detected. So its reward is $\mathbb{E}[\mathbb{1}\{\hat{H} \neq H\}R(X,Y_H)]$. The cheating player wants to maximize its reward; this typically requires a small probability of being caught. On the other hand, the monitoring entity aims to minimize the reward of the cheating player. The probabilistic model can be illustrated as in Figure 6.1. Let us assume that all variables have finite alphabets.

Assuming that the monitor and the cheater are rational and that they both know about the reward structure, then the expected reward will be a minimax value.

[1] The author apologizes for reusing M with different meaning than in the previous part.

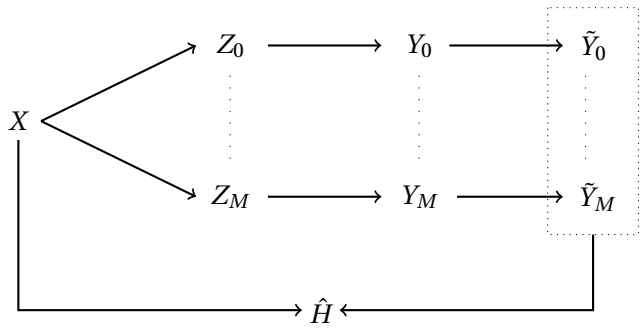


Figure 6.1: Illustration of the monitoring problem.

Definition 6.1. *We define the minimax reward of the cheating player, $R^{(c)}$, as,*

$$\min_{\hat{H}\in\hat{\mathscr{H}}} \max_{f^{(c)}} \mathbb{E}\big[\mathbb{1}\{\hat{H}\neq H\}R\big(X,Y_H\big)\big]. \tag{6.1}$$

At first glance, one would think that the formulation of reward is specific to the case where the decision rule is taken before the cheating player decides on its policy. But this is not the case. This is a consequence of Von Neumann's minimax theorem (section 17.6 in von Neumann et al. (1944)).

Proposition 6.1. *We have*

$$\min_{\hat{H}\in\hat{\mathscr{H}}} \max_{f^{(c)}} \mathbb{E}\big[\mathbb{1}\{\hat{H}\neq H\}R\big(X,Y_H\big)\big] = \max_{f^{(c)}} \min_{\hat{H}\in\hat{\mathscr{H}}} \mathbb{E}[\mathbb{1}\{\hat{H}\neq H\}R(X,Y_H)] \tag{6.2}$$

Proof. This is a consequence of the Von Neumann's minimax theorem (see section 17.6 in von Neumann et al. (1944)). Due to our assumption that the alphabet of the random variables are finite, then the set of deterministic policies of the cheater and the set of deterministic decision rules of the monitor are finite. By the linearity of expectation, one can express the reward function as a bilinear form. Hence this problem fulfils the hypothesis of Von Neumann's minimax theorem. The fact that we allow the policy $f^{(c)}$ to be probabilistic is essential to the proof of proposition 1. □

The minimax characterization of hypothesis testing problem has been studied since the time of Hoeffding (Hoeffding, 1965). There has been several information-theoretic studies of hypothesis testing problems with an adversary (Barni and Tondi, 2013; Feder and Merhav, 2002; Tondi et al., 2015). The main difference between the setting of our work and the literature on universal hypothesis testing lies on our assumption that there is a reward function which the cheating player wants to maximize.

There is also a recent interest on a related problem in the context of adversarial classification, e.g., (Yasodharan and Loiseau, 2019; Dritsoula et al., 2017), which shares a similar concern of an adversary behaving strategically to maximize a reward function. Our work is different

in that we are mainly interested in the asymptotic behavior of the sequences of decision strategies instead of characterizing the equilibrium for a one-shot instance.

6.2 One Shot Case

Before we consider the asymptotic setting, it is useful to examine the one-shot version of the problem and establish several properties of the set of decision rules and the set of cheating policies. It is easy to see that we can characterize the behaviour of a regular player as follows.

Proposition 6.2. *For the optimal $f^{(r)}$, we have, $f^{(r)}(z) \in \arg\max_y \mathbb{E}\big[R(X,y) \mid Z^{(r)} = z\big]$. where the expectation is taken with respect to the regular player probability model.*

Furthermore, even if the set $\hat{\mathscr{H}}$ is uncountable due to randomization, one can characterize a subset which attains the minimum in (6.1). We will refer to this subset as the metric-based decision rules.

Definition 6.2. *We say that $\hat{H}$ is a* metric-based *decision rule if there exists a $g : (\mathscr{X} \times \tilde{\mathscr{Y}}) \to \mathbb{R}_{\geq 0}$ such that $\hat{H}\big((x,(\tilde{y}_i)_{i=0}^M)\big) \in \arg\max_i g(x,\tilde{y}_i)$. In this case, we say that $\hat{H}$ is induced by g. Let $\hat{\mathscr{H}}_m$ be the set of all* metric-based *decision rules.*

Proposition 6.3. *Given a deterministic policy $f^{(c)}$, there exists a $\hat{H} \in \hat{\mathscr{H}}_m$ which attains the minimum of the optimization problem,* $\min_{\hat{H}\in\hat{\mathscr{H}}} \mathbb{E}\big[\mathbb{1}\{\hat{H} \neq H\} R\big(X, Y_H\big)\big]$.

Proof. If for a certain i, $(x,\tilde{y}_i)$ has probability 0 under the regular player probabilistic model, then the monitor can choose that player and incur no error probability. Hence, such $(x,\tilde{y}_i)$ pairs do not contribute to the reward calculation.

Hence, let us only consider the case where the probabilities of all pairs $(x,\tilde{y}_i)$ are non-zero under the regular player probabilistic model. Note that we can write

$$\mathbb{E}[\mathbb{1}\{\hat{H} \neq H\} R(X, Y_H)] = \mathbb{E}[R(X, Y_H)] - \mathbb{E}[\mathbb{1}\{\hat{H} = H\} R(X, Y_H)]. \tag{6.3}$$

The goal of the monitor decision rule is equivalent to maximizing the second term in the RHS.

$$\mathbb{E}[\mathbb{1}\{\hat{H} = H\} R(X, Y_H)] = \mathbb{E}\left[\mathbb{E}[R(X, Y_H)|\mathbb{1}\{\hat{H} = H\}]\right] \tag{6.4}$$

$$= \sum_{x,(\tilde{y}[k])_{k=0}^M} \frac{V_X(x)}{M+1} \prod_{j=0}^{M} V_{\tilde{Y}^{(r)}|X}(\tilde{y}_j|x) \mathbb{E}\left[\frac{V_{\tilde{Y}^{(c)}|X}(\tilde{y}_{\hat{H}}|x)}{V_{\tilde{Y}^{(r)}|X}(\tilde{y}_{\hat{H}}|x)} \mathbb{E}[R(x, Y^{(c)})|X = x, \tilde{Y}^{(c)} = \tilde{y}_{\hat{H}}]\right]. \tag{6.5}$$

This quantity is maximized if $\hat{H}$ is chosen to maximize the following metric,

$$g(x,\tilde{y}) = \frac{V_{\tilde{Y}^{(c)}|X}(\tilde{y}|x)}{V_{\tilde{Y}^{(r)}|X}(\tilde{y}|x)} \mathbb{E}[R(x, Y^{(c)})|X = x, \tilde{Y}^{(c)} = \tilde{y}] \tag{6.6}$$

for $(x,\tilde{y})$ with non-zero probability in $V_{\tilde{Y}^{(r)}|X}$, and ∞ otherwise. □

Due to Proposition 6.3, we can define an equivalence class on the set of all possible mappings g from $\mathscr{X}, \mathscr{Y}$ to $\mathbb{R}_{\geq 0}$, where we say that g_1 and g_2 are equivalent if they induce the same $\hat{H}$. We can also see that although the space of g is uncountable, but the set of $\mathscr{H}_m$ is of finite size.

Proposition 6.4. *Given a decision rule $\hat{H}$, consider a deterministic policy, $f^{(c)}$ such that*

$$f^{(c)}(z) \in \arg\max_{y} \mathbb{E}[\mathbb{1}\{\hat{H} \neq H\} R(X, y) | Z^{(c)} = z] \tag{6.7}$$

where the expectation is taken under the cheater's probabilistic model. This policy attains the maximum of optimization problem,

$$\max_{f^{(c)}} \mathbb{E}[\mathbb{1}\{\hat{H} \neq H\} R(X, Y_H)]. \tag{6.8}$$

I.e., one can restrict the feasible set of cheating player's policies to a set of deterministic cheating policies.

As a consequence of the previous propositions, the class of randomized metric-based decision rules and the class of deterministic cheating policy is sufficient to characterize the minimax value. We state this formally below by using $P(\hat{\mathscr{H}}_m)$ and $\mathscr{F}_d$ to denote the class of randomized metric-based decision rule and the set of deterministic mapping from Z to Y.

Proposition 6.5. *We have*

$$\max_{f^{(c)}} \min_{\hat{H} \in \hat{\mathscr{H}}} \mathbb{E}[\mathbb{1}\{\hat{H} \neq H\} R(X, Y_H)] = \min_{\hat{H} \in P(\hat{\mathscr{H}}_m)} \max_{f^{(c)} \in \mathscr{F}_d} \mathbb{E}[\mathbb{1}\{\hat{H} \neq H\} R(X, Y_H)]. \tag{6.9}$$

Proof. Consider,

$$\begin{aligned}
\max_{f^{(c)}} \min_{\hat{H}} \mathbb{E}[\mathbb{1}\{\hat{H} \neq H\} R(X, Y_H)] &= \max_{f^{(c)}} \min_{\hat{H} \in \hat{\mathscr{H}}_m} \mathbb{E}[\mathbb{1}\{\hat{H} \neq H\} R(X, Y_H)] \\
&= \max_{f^{(c)}} \min_{\hat{H} \in P(\hat{\mathscr{H}}_m)} \mathbb{E}[\mathbb{1}\{\hat{H} \neq H\} R(X, Y_H)] \\
&= \min_{\hat{H} \in P(\hat{\mathscr{H}}_m)} \max_{f^{(c)}} \mathbb{E}[\mathbb{1}\{\hat{H} \neq H\} R(X, Y_H)] \\
&= \min_{\hat{H} \in P(\hat{\mathscr{H}}_m)} \max_{f^{(c)} \in \mathscr{F}_d} \mathbb{E}[\mathbb{1}\{\hat{H} \neq H\} R(X, Y_H)]
\end{aligned} \tag{6.10}$$

where the extension to randomized strategies on the second line is necessary so that we can use the minimax theorem. □

However, note that we cannot exchange the minimum and the maximum after we constrain the admissible sets. This is due to the fact that the set of all possible deterministic policies is not a convex set. One must be careful to note that constraining the inner maximization only on the deterministic policy is merely a calculation device (à la theorem 17:A in (von Neumann et al., 1944)). In general, the optimal cheater strategy will not correspond to a deterministic

policy, but it is the case that the reward of the optimal (randomized) policy is equal to the reward of a certain deterministic policy.

By only considering metric-based decision rules, we transform the problem of optimizing a decision rule $\hat{H}$ over all possible realizations of $(X, \{\tilde{Y}_i\}_{i=0}^M)$ into an easier problem of whether the cheating player metric is larger than the maximum of the regular players' metric. We present a restatement of Proposition 6.4 which will be useful in the next section.

Corollary 6.1. *Given a decision rule in $P(\mathscr{H}_m)$, which is composed of $(\hat{H}_1, \ldots, \hat{H}_k)$ with a probability distribution $(p_1, \ldots, p_k)$, the deterministic policy for $f^{(c)}$ is given by,*

$$f^{(c)}(z) \in \arg\max_{y \in \mathscr{Y}} \sum_x V_{X|Z^{(c)}}(x|z) R(x,y) \sum_{i=1}^{k} p_i \Pr(\tilde{g}_i(x) \leq g_i(x, \tilde{Y}^{(c)}) \mid Y^{(c)} = y) \tag{6.11}$$

where g_i is a metric which induces $\hat{H}_i$, and $\tilde{g}_i(x)$ is the maximum value of the metric g_i among the regular players given the realization of X.

6.3 Memoryless Channels and Product Distributions

Consider a version of the monitoring problem where $(X, Z, Y, \tilde{Y})$ are replaced by n-vectors $X[1:n], Z[1:n], Y[1:n], \tilde{Y}[1:n]$. We study a special case where,

- $(X[1:n], Z[1:n])$ is distributed according to $V_{Z|X}^{\otimes n}$ and $V_X^{\otimes n}$.
- Probabilistic mapping between $Y[1:n]$ and $\tilde{Y}[1:n]$ is given by multiple uses of memoryless channel $V_{\tilde{Y}|Y}$, i.e., it is described by $V_{\tilde{Y}|Y}^{\otimes n}$.
- Both the regular players and the cheating player get to observe the whole realization of their respective $Z[1:n]$ before forming their action.

Furthermore, we will take the number of regular players M to scale as $M = \lfloor e^{nK} \rfloor$. We also assume that the reward function is additive, i.e., $R(X[1:n], Y[1:n]) = \sum_{i=1}^n R(X[i], Y[i])$.

We are interested in studying the normalized reward of the cheating player,

$$\limsup_{n\to\infty} \frac{1}{n} \max_{f^{(c)}} \min_{\hat{H} \in \hat{\mathscr{H}}} \mathbb{E}\big[\mathbb{1}\{\hat{H} \neq H\} R\big(X[1:n], Y_H[1:n]\big)\big]. \tag{6.12}$$

We need to emphasize that this setting only captures the linear growth rate of an increasingly larger single-stage game, as opposed to a repeated game in the game theory setting.

We will refer to the setting which fulfills these requirements as the *product*-regime. Under this regime, the regular players' random variables fulfill a permutation symmetry. Hence, it is not very surprising that we can simplify the problem by only considering a class of "permutation-invariant" $\hat{H}$ and $f^{(c)}$.

In the subsequent discussion, we will heavily use the notion of type that we discussed in chapter 1. In this chapter, we will use P to denote an arbitrary type or distribution, while V is used to denote the distribution in our model or the distribution that is induced by the policies.

Definition 6.3. *We say that $\hat{H}$ is a* permutation-invariant *metric decision rule if there exists a metric $g(X[1:n], \tilde{Y}[1:n])$ which induces $\hat{H}$ and this metric depends on $X[1:n], \tilde{Y}[1:n]$ only through its joint type i.e., $g(X[1:n], \tilde{Y}[1:n]) = g(P_{X,\tilde{Y}})$ where $P_{X,\tilde{Y}}$ is the empirical probability of realization $(X[1:n], \tilde{Y}[1:n])$. Let us denote the set of* permutation-invariant *metric decision rules as $\hat{\mathscr{H}}_T$.*

Definition 6.4. *For a given n, we say that $f^{(c)}$, is a* permutation-invariant *policy if for every $\underline{z}$ and type $P_{Y,Z} \in \mathscr{P}(Y \times Z)$ we have $\Pr_{Y[1:n]|Z[1:n]}(.|z[1:n])$ is uniform on $T_{P_{Y|Z}}(z[1:n])$. Let us denote the set of* permutation-invariant *policies as $\mathscr{F}_T$.*

The analogue of Proposition 6.5 for the product-regime is:

Proposition 6.6. *Under the* product*-regime, we have*

$$\max_{f^{(c)}} \min_{\hat{H} \in \hat{\mathscr{H}}} \mathbb{E}\big[\mathbb{1}\{\hat{H} \neq H\} R\big(X[1:n], Y_H[1:n]\big)\big] = \max_{f^{(c)} \in \mathscr{F}_T} \min_{\hat{H} \in \hat{\mathscr{H}}_T} \mathbb{E}[\mathbb{1}\{\hat{H} \neq H\} R(X, Y_H[1:n])] \quad (6.13)$$

Proof. First we show that,

$$\max_{f^{(c)}} \min_{\hat{H}} \mathbb{E}[\mathbb{1}\{\hat{H} \neq H\} R(X[1:n], Y_H[1:n])] = \max_{f^{(c)} \in \mathscr{F}_T} \min_{\hat{H}} \mathbb{E}[\mathbb{1}\{\hat{H} \neq H\} R(X[1:n], Y_H[1:n])]. \quad (6.14)$$

Given a $f^{(c)}$, let us consider the *permutation-invariant* version which takes a permutation π from the set of permutation Π uniformly at random, such that $\tilde{f}^{(c)}(z[1:n]) = \pi^{-1}(f^{(c)}(\pi(z[1:n]))$. Let us denote the distribution induced by this *permutation-invariant* version as $\tilde{P}$.

We have

$$\begin{aligned}
\min_{\hat{H}} \mathbb{E}_{\tilde{P}}[\mathbb{1}\{\hat{H} \neq H\} R(X, Y_H[1:n])] &= \min_{\hat{H}} \sum_{\pi} \frac{1}{|\Pi|} \mathbb{E}_{\tilde{P}}[\mathbb{1}\{\hat{H} \neq H\} R(X[1:n], Y_H[1:n]) | \pi] \\
&\geq \sum_{\pi} \frac{1}{|\Pi|} \min_{\hat{H}} \mathbb{E}_{\tilde{P}}[\mathbb{1}\{\hat{H} \neq H\} R(X[1:n], Y_H[1:n]) | \pi] \\
&\overset{(1)}{=} \sum_{\pi} \frac{1}{|\Pi|} \min_{\hat{H}} \mathbb{E}[\mathbb{1}\{\hat{H} \neq H\} R(\pi^{-1}(X[1:n]), \pi^{-1}(Y_H[1:n]))] \\
&\overset{(2)}{=} \min_{\hat{H}} \mathbb{E}[\mathbb{1}\{\hat{H} \neq H\} R(X[1:n], Y_H[1:n])]. \quad (6.15)
\end{aligned}$$

We require the facts that the channel is memoryless in (1), and that the reward is additive in (2). This inequality implies that there exists $\tilde{f}^{(c)} \in \mathscr{F}_t$ which is as good as any $f^{(c)} \notin \mathscr{F}_t$.

Finally, we show that,

$$\max_{f^{(c)}\in\mathscr{F}_T}\min_{\hat{H}}\mathbb{E}[\mathbb{1}\{\hat{H}\neq H\}R(X,Y_H[1:n])]=\max_{f^{(c)}\in\mathscr{F}_T}\min_{\hat{H}\in\hat{\mathscr{H}}_T}\mathbb{E}[\mathbb{1}\{\hat{H}\neq H\}R(X,Y_H[1:n])]. \tag{6.16}$$

We can show this by showing that the optimal metric that we established in eq. (6.6) is a *permutation-invariant* decision rule under the assumption that $f^{(c)}$ is *permutation-invariant*. More formally, consider two tuples $(x[1:n],\tilde{y}[1:n])$ and $(x'[1:n],\tilde{y}'[1:n])$ both of the same joint type. This implies that there exists a permutation π such that $(x'[1:n],\tilde{y}'[1:n])=(\pi(x[1:n]),\pi(\tilde{y}[1:n]))$. Notice that for any $\pi\in\Pi$, then $\Pi=\cup_{\pi'\in\Pi}\{\pi\circ\pi'\}$. We have,

$$\begin{aligned}
&V_{\tilde{Y}^{(c)}[1:n]|X[1:n]}(\tilde{y}'[1:n],x'[1:n])\\
&=\sum_{y'[1:n]}V_{Y[1:n],\tilde{Y}^{(c)}[1:n]|X[1:n]}(y,\tilde{y}'[1:n],x'[1:n])\\
&=\sum_{y'[1:n]}V_{\tilde{Y}^{(c)}[1:n]|Y^{(c)}[1:n]}(\tilde{y}'[1:n],y'[1:n])V_{Y^{(c)}[1:n]|X[1:n]}(y'[1:n],x'[1:n])\\
&\overset{(1)}{=}\sum_{y'[1:n]}V_{\tilde{Y}^{(c)}[1:n]|Y^{(c)}}[1:n](\tilde{y}'[1:n],y'[1:n])\sum_{\pi'}\frac{V_{Y^{(c)}[1:n]|X[1:n]}(\pi'(y'[1:n]),\pi'(x'[1:n]))}{|\Pi|}\\
&\overset{(2)}{=}\sum_{y'[1:n]}V_{\tilde{Y}^{(c)}[1:n]|Y^{(c)}}[1:n](\pi(\tilde{y}[1:n]),\pi(y[1:n]))\sum_{\pi'}\frac{V_{Y[1:n]^{(c)}|X[1:n]}(\pi'\circ\pi(y[1:n]),\pi'\circ\pi(x[1:n]))}{|\Pi|}\\
&=\sum_{y'[1:n]}V_{\tilde{Y}^{(c)}[1:n]|Y^{(c)}}[1:n](\tilde{y}[1:n],y[1:n])\sum_{\pi'}\frac{V_{Y^{(c)}[1:n]|X}[1:n](\pi'(y[1:n]),\pi'(x[1:n]))}{|\Pi|}\\
&=V_{\tilde{Y}^{(c)}[1:n]|X[1:n]}(\tilde{y}[1:n],x[1:n])
\end{aligned} \tag{6.17}$$

We used the fact that $f^{(c)}$ is *permutation-invariant* in (1), and the fact that we are on the *product*-regime in (2). We can use the same argument to show that $\mathbb{E}[R(x[1:n],Y^{(c)}[1:n])|X[1:n]=x[1:n],\tilde{Y}^{(c)}[1:n]=\tilde{y}[1:n]]$ also has the same property. As every term of the optimal metric only depends on the joint type, the optimal metric also only depends only on the joint-type. □

Given an $x[1:n]$ each permutation-invariant metric decision rule essentially ranks each possible type $P_{X,\tilde{Y}}\in\mathscr{P}(X\times\tilde{Y})$ which is compatible with $P_{x[1:n]}$. Therefore, the probability that the monitor makes an error is equivalent to the probability that one of the realizations of $\{\tilde{Y}_i[1:n]\}_{i\neq H}$ is such that the joint type of $(X[1:n],\tilde{Y}_i[1:n])$ is ranked higher than the joint type of $(X,\tilde{Y}_H[1:n])$.

For asymptotic analysis, there is a technical subtlety that the decision metrics that the monitor chooses can depend on n. It is reasonable to worry whether the ordering of types induced by the metrics will be preserved across the values of n. Hence, we need to introduce a formulation that take this concern into account.

Consider a sequence of metrics $g_1,g_2,\ldots$ such that the monitoring entity uses g_n to form its decision rule for length n.

Let us define the sets,

$$\mathscr{P}_\infty(X) = \bigcup_{i=1}^{\infty} \mathscr{P}_i(X) \qquad \mathscr{P}_\infty(X \times \tilde{Y}) = \bigcup_{i=1}^{\infty} \mathscr{P}_i(X \times \tilde{Y}). \tag{6.18}$$

Given any joint type $P_{X,\tilde{Y}} \in \mathscr{P}_\infty(X \times \tilde{Y})$, we can define a set

$$\mathscr{U}_n(P_{X,\tilde{Y}}) = \left\{ P'_{X,\tilde{Y}} \in \mathscr{P}_n(X \times \tilde{Y}) \,|\, P_X = P'_X, g_n(P'_{X,\tilde{Y}}) > g_n(P_{X,\tilde{Y}}) \right\} \tag{6.19}$$

if $P_{X,\tilde{Y}} \in \mathscr{P}_n(X \times \tilde{Y})$, otherwise $\mathscr{U}_n(P_X, P_{X,\tilde{Y}}) = \varnothing$. Let us define

$$\mathscr{U}\left(P_{X,\tilde{Y}}\right) = \limsup_{n\to\infty} \mathscr{U}_n\left(P_{X,\tilde{Y}}\right). \tag{6.20}$$

An intuitive explanation of $\mathscr{U}(P_{X,\tilde{Y}})$ is that it corresponds to "level sets" on the probability simplex in which infinitely many g_n's agrees that the distribution on this set is ranked higher than $P_{X,\tilde{Y}}$.

We also define

$$G\left(P_{X,\tilde{Y}}\right) = \inf_{P'_{X,\tilde{Y}} \in \mathscr{U}(P_{X,\tilde{Y}})} D\left(P'_{X,\tilde{Y}}; V_{X,\tilde{Y}^{(r)}}\right) \tag{6.21}$$

i.e., $G(P_{X,\tilde{Y}})$ is the distance between the regular player joint distribution $V_{X,\tilde{Y}^{(r)}}$ and the set $\mathscr{U}(P_{X,\tilde{Y}})$ in KL divergence.

Proposition 6.7. *For a sequence of metrics $g_1, g_2, \ldots$ and any distribution $P_{X,\tilde{Y}} \in \mathscr{P}_\infty(X \times \tilde{Y})$, we have*

- *if $K > G(P_{X,\tilde{Y}})$ then*

$$\limsup_{n\to\infty} \Pr\left(\hat{H} \neq H \,\middle|\, (X[1:n], \tilde{Y}_H[1:n]) \in T_{P_{X,\tilde{Y}^{(c)}}}\right) = 1, \tag{6.22}$$

- *if $K < G(P_{X,\tilde{Y}})$ then*

$$\limsup_{n\to\infty} -\frac{1}{n} \log \Pr\left(\hat{H} \neq H \,\middle|\, (X[1:n], \tilde{Y}_H[1:n]) \in T_{P_{X,\tilde{Y}^{(c)}}}\right) > 0. \tag{6.23}$$

Proof. For $P'_{\tilde{Y},X}$ such that $P'_X \neq P_X$, the conditioning event is a null event, hence we can ignore this event as we are working with discrete variables. So we will only focus on n such that $P'_X = P_X$. We have,

$$\Pr\left(\hat{H} = H \,\middle|\, (X[1:n], \tilde{Y}_H[1:n]) \in T_{P_{X,\tilde{Y}}}\right) = \left(1 - \Pr(g_n(X, \tilde{Y}^{(r)}) > g_n(P_{X,\tilde{Y}}) | (X[1:n], \tilde{Y}_H[1:n]) \in T_{P_{X,\tilde{Y}}}\right)^{e^{nK}} \tag{6.24}$$

By using type estimates, we have,

$$\begin{aligned}\Pr\left(\hat{H}=H\,\middle|\,(X[1:n],\tilde{Y}_H[1:n])\in T_{P_{X,\tilde{Y}}}\right) &\overset{(a)}{\leq} (1-e^{-n\min_{P'_{X,\tilde{Y}}\in U_n(P_{X,\tilde{Y}})} D(P'_{X,\tilde{Y}};V_{X,\tilde{Y}(r)})+o(n)})e^{nK}\\ &\overset{(b)}{\leq} \exp\left(-e^{n(K-\min_{P'_{X,\tilde{Y}}\in U_n(P_{X,\tilde{Y}})} D(P'_{X,\tilde{Y}};V_{X,\tilde{Y}(r)})+o(n)}\right)\\ &\overset{(c)}{\leq} \exp\left(-e^{n(K-D(P'_{X,\tilde{Y}};V_{X,\tilde{Y}(r)})+o(n)}\right) \end{aligned} \tag{6.25}$$

where the last inequality holds for any $P'_{X,\tilde{Y}}\in\mathscr{U}_n(P_{X,\tilde{Y}})$. For inequality (a), we use the method of type estimate of the probability. For inequality (b), we use $(1-x)\leq\exp(-x)$. For inequality (c), we possibly choose a non-minimizer of the upper bound. Note that if a type $P'_{X,\tilde{Y}}$ is included in $\mathscr{U}(P_{X,\tilde{Y}})$, it is included in $\mathscr{U}_n(P_{X,\tilde{Y}})$ for infinitely many n. Hence, for any K and $P'_{X,\tilde{Y}}\in\mathscr{U}(P_{X,\tilde{Y}})$, we have

$$\liminf_{n\to\infty}\Pr\left(\hat{H}=H\,\middle|\,(X[1:n],\tilde{Y}_H[1:n])\in T_{P_{X,\tilde{Y}^{(c)}}}\right)\leq\liminf_{n\to\infty}\exp\left(-e^{n(K-D(P'_{X,\tilde{Y}};V_{X,\tilde{Y}(r)})+o(n)}\right), \tag{6.26}$$

and thus,

$$\liminf_{n\to\infty}\Pr\left(\hat{H}=H\,\middle|\,(X[1:n],\tilde{Y}_H[1:n])\in T_{P_{X,\tilde{Y}^{(c)}}}\right)\leq\liminf_{n\to\infty}\exp\left(-e^{n(K-G(P_{X,\tilde{Y}^{(c)}}))+o(n)}\right), \tag{6.27}$$

by the definition of $G(P_{X,\tilde{Y}^{(c)}})$. Where $K>G(P_{X,\tilde{Y}})$, eq. (6.27) proves the first part of the proposition.

For the second part, it is sufficient to use the union bound. We have,

$$\begin{aligned}&\Pr\left(\hat{H}\neq H\,\middle|\,(X[1:n],\tilde{Y}_H[1:n])\in T_{P_{X,\tilde{Y}}}\right)\\ &\leq e^{nK}\Pr(g_n(X[1:n],\tilde{Y}^{(r)}[1:n])>g_n(P_{X,\tilde{Y}})|(X[1:n],\tilde{Y}_H[1:n])\in T_{P_{X,\tilde{Y}}})\\ &\leq e^{n(K-\min_{P'_{X,\tilde{Y}}\in U_n(P_{X,\tilde{Y}})} D(P'_{X,\tilde{Y}};V_{X,\tilde{Y}(r)})+o(n)}\\ &\leq e^{n(K-\min_{P'_{X,\tilde{Y}}\in \cup_{i\geq n}U_i(P_{X,\tilde{Y}})} D(P'_{X,\tilde{Y}};V_{X,\tilde{Y}(r)})+o(n)}\end{aligned} \tag{6.28}$$

This upper bound is valid for all n. Since $\cup_{i\geq n}U_i(P_{X,\tilde{Y}})$ is a decreasing sequence of sets in n with limit $\mathscr{U}(P_{X,\tilde{Y}})$, for all $\epsilon>0$ there exists n^* such that for all $n>n^*$

$$\min_{P'_{X,\tilde{Y}}\in\cup_{i\geq n}U_i(P_{X,\tilde{Y}})} D(P'_{X,\tilde{Y}};V_{X,\tilde{Y}})\geq G(P_{X,\tilde{Y}})-\epsilon \tag{6.29}$$

due to the definition of $G(P_{X,\tilde{Y}})$ and the continuity of the KL divergence. Hence,

$$\limsup_{n\to\infty}-\frac{1}{n}\log\Pr\left(\hat{H}\neq H\,\middle|\,(X[1:n],\tilde{Y}_H[1:n])\in T_{P_{X,\tilde{Y}^{(c)}}}\right)>0 \tag{6.30}$$

if $K<G(P_{X,\tilde{Y}})$. □

Proposition 6.7 is the main technical result of this work. This proposition asserts that there is a qualitative change depending on whether $G(P_{X,\tilde{Y}})$ is smaller or larger than K. Hence, given K we can determine the set of joint types of cheater's realization for which the monitoring entity will make an error with high probability. Let us define for $P_X \in \mathscr{P}_\infty(X)$

$$G^{-1}_{(g_n)_{i=1}^\infty}(P_X, K) = \limsup_{i\to\infty} G_n^{-1} \tag{6.31}$$

where

$$G_n^{-1} = \{P_{X,\tilde{Y}} \in \mathscr{P}_n(X,\tilde{Y}) : P_X = P'_X \qquad g_n(P'_{X,\tilde{Y}}) \leq \max_{\substack{P'_{X,\tilde{Y}}: \\ D(P'_{X,\tilde{Y}}; V_{X,\tilde{Y}(r)})<K}} g_n(P'_{X,\tilde{Y}})\}, \tag{6.32}$$

and

$$G^{-1}_{(g_n)_{n=1}^\infty}(K) = \bigcap_{\epsilon>0} \bigcup_{\substack{P_X \in T_X^\infty \\ |V_X - P_X|\leq\epsilon}} G^{-1}_{(g_n)_{n=1}^\infty}(P_X, K). \tag{6.33}$$

Using Proposition 6.7, we can give upper and lower bound on the expected linear growth of cheating player's reward.

Proposition 6.8. *Given a sequence of type-invariant metrics $(g_n)_{n=1}^\infty$, we have,*

$$\limsup_{n\to\infty} \frac{1}{n} \max_{f^{(c)}\in\mathscr{F}_T} \mathbb{E}\big[\mathbb{1}\{\hat{H}\neq H\} R\big(X[1:n], Y_H[1:n]\big)\big] \geq \sup_{\substack{P_{Y^{(c)}|Z^{(c)}}: \\ V_{X,\tilde{Y}^{(c)}} \in IntCl(G^{-1}_{(g_n)_{n=1}^\infty}(K))}} \mathbb{E}[R(X, Y^{(c)})], \tag{6.34}$$

where $IntCl(G^{-1}_{(g_n)_{n=1}^\infty}(K))$ is the interior of the closure of $G^{-1}_{(g_n)_{n=1}^\infty}(K)$ under the total variation metric. We also have,

$$\limsup_{n\to\infty} \frac{1}{n} \max_{f^{(c)}\in\mathscr{F}_T} \mathbb{E}\big[\mathbb{1}\{\hat{H}\neq H\} R\big(X[1:n], Y_H[1:n]\big)\big] \leq \max_{\substack{P_{Y^{(c)}|Z^{(c)}}: \\ V_{X,\tilde{Y}^{(c)}} \in Cl(G^{-1}_{(g_n)_{n=1}^\infty}(K))}} \mathbb{E}[R(X, Y^{(c)})], \tag{6.35}$$

where $Cl(G^{-1}_{(g_n)_{n=1}^\infty}(K))$ is the closure of $G^{-1}_{(g_n)_{n=1}^\infty}(K)$ and $V_{X,\tilde{Y}^{(c)}}$ is the induced distributed on the monitoring problem given $P_{Y^{(c)}|Z^{(c)}}$.

Proof. To show the first part namely eq. (6.34), fix $\epsilon > 0$ and consider a $P_{Y^{(c)}|Z^{(c)}}$ such that $V_{X,\tilde{Y}^{(c)}} \in \mathrm{IntCl}(G^{-1}_{(g_n)_{n=1}^\infty}(K))$ and $\mathbb{E}_{P_{X,Y^{(c)}}}[R(X, Y^{(c)})]$ is ϵ-close to the supremum on the RHS of eq. (6.34).

Now fix n. Let us construct our $f^{(c)}$ by sampling from the distribution $P_{Y^{(c)}|Z}$ given the realization of $Z[1:n]$. From results in chapter 1 we know that for every $\delta > 0$, $P_X(X[1:n] \in T_{[P_X]_\delta}) \geq 1 - Ae^{-n\delta^2}$ and $V_{\tilde{Y}|X}(\tilde{Y}[1:n] \in T_{[V_{\tilde{Y}|X}]_{2\delta}}(X[1:n]) | X[1:n] \in T_{[P_X]_\delta}) \geq 1 - A'e^{-n\delta^2}$ (A and A' only depend on the sizes of the alphabets of X and $\tilde{Y}$), where $T_{[V_{X,\tilde{Y}^{(c)}}]_\delta}$ is the set of all types

with total variation distance at most δ from $V_{X,\tilde{Y}^{(c)}}$. As the $V_{X,\tilde{Y}^{(c)}}$ is in the $\mathrm{IntCl}(G^{-1}_{(g_n)_{n=1}^{\infty}}(K))$, then for small enough δ we have,

$$T_{[V_{X,\tilde{Y}^{(c)}}]_{2\delta}} \subseteq \mathrm{IntCl}(G^{-1}_{(g_n)_{n=1}^{\infty}}(K)) \tag{6.36}$$

Let us define the event Q_n as $\{X \in T_{[P_X]_\delta}, \tilde{Y}_H[1:n] \in T_{[V_{\tilde{Y}^{(c)}|X}]_{2\delta}}(\underline{X})\}$. We note that $\lim_{n\to\infty}\Pr(Q_n) = 1$. So we have under $V_{X,\tilde{Y}^{(c)}}$,

$$\begin{aligned}
&\mathbb{E}[R(X[1:n], Y_H[1:n])|Q_n]\Pr(Q_n)\\
&= \mathbb{E}[R(X[1:n], Y_H[1:n])] - \mathbb{E}[R(X[1:n], Y_H[1:n])|Q_n^C](1-\Pr(Q_n))\\
&\geq \mathbb{E}[R(X[1:n], Y_H[1:n])] - O(ne^{-n\delta^2}).
\end{aligned} \tag{6.37}$$

We also have,

$$\frac{1}{n}\mathbb{E}[\mathbb{1}\{\hat{H} \neq H\}R(X[1:n], Y_H[1:n])] \geq \frac{1}{n}\mathbb{E}[\mathbb{1}\{\hat{H} \neq H\}R(X[1:n], Y_H[1:n])|Q_n]\Pr(Q_n) \tag{6.38}$$

Conditioned on Q_n, we have,

$$\frac{1}{n}R(X[1:n], Y_H[1:n]) \geq \mathbb{E}_V[R(X[1:n], Y^{(c)}[1:n])](1-O(\delta)). \tag{6.39}$$

Therefore

$$\frac{1}{n}\mathbb{E}[\mathbb{1}\{\hat{H} \neq H\}R(X[1:n], Y_H[1:n])] \geq \mathbb{E}_V[R(X[1:n], Y^{(c)}[1:n])](1-O(\delta))\Pr(\hat{H} \neq H|Q_n)\Pr(Q_n). \tag{6.40}$$

Note that

$$\Pr(\hat{H} \neq H|Q_n) \geq b_n \tag{6.41}$$

where

$$b_n = \min_{P'_{\tilde{Y}|X} \in T_{[V_{\tilde{Y}|X}]_{2\delta}}} \Pr(\hat{H} \neq H|X[1:n] \in T_{[P_X]_\delta}, \tilde{Y}[1:n] \in T_{P'_{\tilde{Y}|X}}(\underline{X})).$$

Combining these terms give us,

$$\begin{aligned}
&\limsup_{n\to\infty}\frac{1}{n}\mathbb{E}[\mathbb{1}\{\hat{H} \neq H\}R(X[1:n], Y_H[1:n])]\\
&\geq (\mathbb{E}_V[R(X[1:n], Y^{(c)}[1:n])](1-O(\delta))\limsup_{n\to\infty}\Pr(Q_n)b_n\\
&= (\mathbb{E}_V[R(X[1:n], Y^{(c)}[1:n])](1-O(\delta))\limsup_{n\to\infty} b_n\\
&\overset{(*)}{=} \mathbb{E}_V[R(X[1:n], Y^{(c)}[1:n])](1-O(\delta))
\end{aligned} \tag{6.42}$$

for equality $(*)$ we used the fact that $T_{[V_{\tilde{Y}^{(c)}|X}]_{2\delta}} \subseteq \mathrm{IntCl}(G^{-1}_{(g_n)_{n=1}^{\infty}}(K))$, therefore $\limsup_{n\to\infty} b_n = 1$ due to the first part of Proposition 6.7. Since $\mathbb{E}_V[R(X[1:n], Y^{(c)}[1:n])]$ is ϵ-close to the

supremum and δ is arbitrary, eq. (6.34) follows.

Now we will show eq. (6.35). As in the previous part, for any $\delta > 0$, let us define, $Q_n = \{(X[1:n], Z[1:n]) \in T_{[V_{X,Z^{(c)}}]_\delta}\}$. So we have

$$\begin{aligned}&\frac{1}{n}\mathbb{E}[\mathbb{1}\{\hat{H} \neq H\}R(X[1:n], Y_H[1:n])] \\ &\leq \frac{1}{n}\mathbb{E}[\mathbb{1}\{\hat{H} \neq H\}R(X[1:n], Y_H[1:n])|Q_n]\Pr(Q_n) + O(\exp(-n\delta^2))\end{aligned} \tag{6.43}$$

where we used the fact that $\Pr\left((X[1:n], Z[1:n]) \notin T_{[V_{X,Z^{(c)}}]_\delta}\right) \leq \exp(-n\delta^2)$. Note that $f^{(c)} \in \mathscr{F}_T$ can be parametrized by set of tuples $\{(P_Z, P_{Y|Z}) : P_Z \in \mathscr{P}(Z)\}$ i.e., the policy chooses a single conditional type $P_{Y|Z}$ for each P_Z that it observes. One can use this parametrization to further upper bound the non-vanishing part of eq. (6.43) as

$$\begin{aligned}&\frac{1}{n}\mathbb{E}[\mathbb{1}\{\hat{H} \neq H\}R(X, Y_H[1:n])|Q_n]\Pr(Q_n) \\ &\leq \sum_{P_{X,Z} \in T_{[V_{X,Z^{(c)}}]_\delta}} \max_{P_{Y|Z}} \frac{1}{n}\mathbb{E}[\mathbb{1}\{\hat{H} \neq H\}R(X, Y_H[1:n])|(X[1:n], Z[1:n]) \in T_{P_{X,Z}}] \\ &\qquad \Pr((X[1:n], Z[1:n]) \in T_{P_{X,Z}}).\end{aligned} \tag{6.44}$$

We will use the fact that $P_{X,Z} \in T_{[V_{X,Z^{(c)}}]_\delta}$ to note that $(X[1:n], Z[1:n], Y[1:n]) \in T_{[V_{X,Z^{(c)}} \circ P_{Y|Z}]_{2\delta}}$ with probability of 1, as given $Z[1:n]$ the policy takes $Y[1:n] \in T_{P_{Y|Z}}(Z[1:n])$. Finally, $\tilde{Y}[1:n]$ fulfils $(X[1:n], Z[1:n], Y[1:n], \tilde{Y}[1:n]) \in T_{[V_{X,Z^{(c)}} \circ P_{Y|Z} \circ V_{\tilde{Y}|Y}]_{A\delta}}$ (A is a constant that depends on the size of the alphabet of Y) with high probability since the channel $V_{\tilde{Y}|Y}$ is memoryless. Note that $V_{X,Z^{(c)}} \circ P_{Y|Z} \circ V_{\tilde{Y}|Y}$ is the definition of $V_{X,\tilde{Y}^{(c)}}$, i.e., the distribution of $X, \tilde{Y}^{(c)}$ under the cheating player probabilistic model. Hence we have,

$$\begin{aligned}&\frac{1}{n}\mathbb{E}[\mathbb{1}\{\hat{H} \neq H\}R(X[1:n], Y_H[1:n])|Q_n]\Pr(Q_n) \\ &\leq \max_{P_{Y|Z}}(1 + O(\delta))\mathbb{E}_V[R(X, Y^{(c)})]c_{n,\delta}(V_{X,\tilde{Y}^{(c)}}) + O(-nA^2\delta^2)\end{aligned} \tag{6.45}$$

where,

$$c_{n,\delta}(V_{X,\tilde{Y}^{(c)}}) = \sum_{P_{X,\tilde{Y}} \in T_{[V_{X,\tilde{Y}^{(c)}}]_{A'\delta}}} \Pr(\hat{H} \neq H, (X[1:n], \tilde{Y}[1:n]) \in T_{P_{X,\tilde{Y}}}) \tag{6.46}$$

with A' is a constant which only depends on the size of the alphabets.

This allows us to give an upper bound,

$$\begin{aligned}&\limsup_{n\to\infty} \frac{1}{n} \max_{f^{(c)} \in \mathscr{F}_T} \mathbb{E}[\mathbb{1}\{\hat{H} \neq H\}R(X[1:n], Y_H[1:n])] \\ &\leq \limsup_{n\to\infty} \max_{P_{Y|Z}}(1 + O(\delta))E_V[R(X, Y^{(c)})]c_{n,\delta}(V_{X,\tilde{Y}^{(c)}}).\end{aligned}$$

We can decompose the region of the maximization into,

$$\limsup_{n\to\infty}\frac{1}{n}\max_{f^{(c)}\in\mathscr{F}_T}\mathbb{E}[\mathbb{1}\{\hat{H}\neq H\}R(X[1:n],Y_H[1:n])]\leq\limsup_{n\to\infty}\max\{u_n,v_n\}$$
$$=\max\{\limsup_{n\to\infty}u_n,\limsup_{n\to\infty}v_n\} \tag{6.47}$$

with,

$$u_n=\max_{\substack{P_{Y|Z}:\\ V_{X,Y^{(c)}}\in D_\delta}}(1+O(\delta))\mathbb{E}_V[R(X,Y^{(c)})]c_{n,\delta}(V_{X,\tilde{Y}^{(c)}})$$
$$v_n=\max_{\substack{P_{Y|Z}:\\ V_{X,Y^{(c)}}\notin D_\delta}}(1+O(\delta))\mathbb{E}_V[R(X,Y^{(c)})]c_{n,\delta}(V_{X,\tilde{Y}^{(c)}}) \tag{6.48}$$

where

$$D_\delta=\{V_{X,Y^{(c)}}:|V_{X,Y^{(c)}}-G^{-1}_{(g_n)_{n=1}^\infty}(K)|\leq 2A'\delta\}, \tag{6.49}$$

i.e., D_δ is the $2A'\delta$ neighborhood of $G^{-1}_{(g_n)_{n=1}^\infty}(K)$ in total variation distance.

By the second part of preposition 6.7, we have that

$$\lim_{n\to\infty}c_{n,\delta}(V_{X,\tilde{Y}^{(c)}})=0 \tag{6.50}$$

if $V_{X,\tilde{Y}^{(c)}}\notin\mathrm{Int}(D_\delta)$. Hence, v_n vanishes as $n\to\infty$. For the u_n, we can upper bound $c_{n,\delta}(\cdot)$ by 1. This gives us,

$$\limsup_{n\to\infty}\frac{1}{n}\max_{f^{(c)}\in\mathscr{F}_T}\mathbb{E}[\mathbb{1}\{\hat{H}\neq H\}R(X,Y_H[1:n])]\leq\max_{\substack{P_{Y|Z}:\\ V_{X,Y^{(c)}}\in D_\delta}}(1+O(\delta))E_V[R(X,Y^{(c)})]. \tag{6.51}$$

Finally, note that this upper bound holds for arbitrary δ and we have

$$\mathrm{Cl}(G^{-1}_{(g_n)_{n=1}^\infty}(K))=\bigcap_{\delta>0}D_\delta. \tag{6.52}$$

Combined with the fact that expectation is continuous under total variation distance, this upper bound establishes the second part of Proposition 6.8. □

We have to split the proposition into two cases to take into account the effect of isolated points in $\mathrm{Cl}\big(G^{-1}_{(g_n)_{n=1}^\infty}(K)\big)$. Our achievability method requires the existence of a non-empty neighborhood around the sequence that approximates the optimal point. Note that this kind of separation also exists in several large deviation principle theorems given in (Dembo and Zeitouni, 1998), hence it might be non-trivial to remove the influence of isolated points. However, as we discuss later on, the sequence that induces the optimal strategy will not have isolated points, hence obviating the need for this consideration.

The interpretation of $\mathscr{U}$ as level sets is useful when we try to compare the asymptotic behaviour of two sequences of decision rules. In fact this notion allows us to argue that the empirical

divergence decision rule is as good as any other sequence of decision rules.

To formalize this claim, let us denote the normalized reward of the cheating player under the sequence of decision rules $(g_n)_{n=1}^{\infty}$ as $R((g_n)_{n=1}^{\infty})$. Let us also define the sequence of empirical divergence metrics as, $g_n^{ed}(X[1:n], \tilde{Y}[1:n]) = D(P_{X,\tilde{Y}}; V_{X,\tilde{Y}^{(r)}})$ where $P_{X[1:n],\tilde{Y}[1:n]}$ is the type of the realization $(X[1:n], \tilde{Y}[1:n])$.

The following proposition guarantees the asymptotic optimality of the empirical divergence decision rule.

Proposition 6.9. *For any sequence of metrics* $(g_n)_{i=1}^{\infty}$,

$$R((g_n^{ed})_{n=1}^{\infty}) \leq R((g_n)_{n=1}^{\infty}). \tag{6.53}$$

Proof. From proposition 6.8, one way to prove the statement is to show that,

$$\mathrm{Cl}(G^{-1}_{(g_n^{ed})_{n=1}^{\infty}}(K)) \subseteq \mathrm{ClInCl}(G^{-1}_{(g_n)_{n=1}^{\infty}}(K)). \tag{6.54}$$

The $\mathrm{ClInCl}(A)$ is the closure of the interior of closure of set A. If the inclusion holds, then the upper bound part of proposition 6.8 applied on $R((g_n^{ed})_{n=1}^{\infty})$ is smaller than the lower bound part of proposition 6.8 applied on $R((g_n)_{n=1}^{\infty})$. This is the approach that we will take.

First, observe that for every n and P_X we have,

$$\begin{aligned}
\{P'_{X,\tilde{Y}} : P'_X = P_X, g_n^{ed}(P'_{X,\tilde{Y}}) \leq & \sup_{\substack{P^*_{X,\tilde{Y}}: \\ D(P^*_{X,\tilde{Y}}; V_{X,\tilde{Y}^{(r)}}) < K}} g_n^{ed}(P^*_{X,\tilde{Y}})\} \\
& \overset{(*)}{=} \{P'_{X,\tilde{Y}} : P'_X = P_X, D(P'_{X,\tilde{Y}}; V_{X,\tilde{Y}^{(r)}}) < K\} \\
& \subseteq \{P'_{X,\tilde{Y}} : P'_X = P_X, g_n(P'_{X,\tilde{Y}}) \\
& \leq \sup_{\substack{P^*_{X,\tilde{Y}}: \\ D(P^*_{X,\tilde{Y}}; V_{X,\tilde{Y}^{(r)}}) < K}} g_n(P^*_{X,\tilde{Y}})\}.
\end{aligned} \tag{6.55}$$

Which imply that for every P_X,

$$G^{-1}_{(g_n^{ed})_{n=1}^{\infty}}(P_X, K) \subseteq G^{-1}_{(g_n)_{n=1}^{\infty}}(P_X, K) \tag{6.56}$$

and taking the appropriate lim gives us,

$$G^{-1}_{(g_n^{ed})_{n=1}^{\infty}}(K) = \lim_{\epsilon\to 0} \bigcup_{\substack{P_X: \\ |V_X - P_X| < \epsilon}} G^{-1}_{(g_n^{ed})_{n=1}^{\infty}}(P_X, K) \subseteq \lim_{\epsilon\to 0} \bigcup_{\substack{P_X: \\ |V_X - P_X| < \epsilon}} G^{-1}_{(g_n)_{n=1}^{\infty}}(P_X, K) = G^{-1}_{(g_n)_{n=1}^{\infty}}(K) \tag{6.57}$$

Furthermore, taking the ClIntCl in both sides gives us,

$$\mathrm{ClIntCl}(G^{-1}_{(g_n^{ed})_{n=1}^{\infty}}(K)) \subseteq \mathrm{ClIntCl}(G^{-1}_{(g_n)_{n=1}^{\infty}}(K)). \tag{6.58}$$

So we only need to show that,

$$\mathrm{ClIntCl}(G^{-1}_{(g_n^{ed})_{n=1}^{\infty}}(K)) = \mathrm{Cl}(G^{-1}_{(g_n^{ed})_{n=1}^{\infty}}(K)) \tag{6.59}$$

which is equivalent to showing that $\mathrm{Cl}(G^{-1}_{(g_n^{ed})_{n=1}^{\infty}}(K))$ does not contain any isolated points. Let

$$\mathscr{A} = \{P'_{X,\tilde{Y}} : D(P'_{X,\tilde{Y}}; V_{X,\tilde{Y}^{(r)}}) < K\}.$$

So we have $G^{-1}_{(g_n^{ed})_{n=1}^{\infty}}(K) = \mathscr{P}_{\infty} \bigcap \mathscr{A}$. As $\mathscr{P}_{\infty}$ is dense, then $\mathrm{Cl}(G^{-1}_{(g_n^{ed})_{n=1}^{\infty}}(K)) = \mathrm{Cl}(\mathscr{A}) = \{P'_{X,\tilde{Y}} : D(P'_{X,\tilde{Y}}; V_{X,\tilde{Y}^{(r)}}) \leq K\}$ which is a closed convex set with no isolated points.

□

Corollary 6.2. *We have*

$$R((g_n^{ed})_{n=1}^{\infty}) = \max_{\substack{V_{Y^{(c)}|Z^{(c)}}: \\ D(V_{X,\tilde{Y}^{(c)}}; V_{X,\tilde{Y}^{(r)}}) \leq K}} \mathbb{E}[R(X, Y^{(c)})]. \tag{6.60}$$

Proof. Due to the continuity of divergence function and $D(\cdot; V_{X,\tilde{Y}^{(r)}})$ is a convex function, the closure of $G^{-1}_{(g_n^{ed})_{n=1}^{\infty}}(K)$ does not have isolated points. Hence the lower bound and the upper bound in Proposition 6.8 coincide. Finally, one obtains the RHS by substituting the empirical divergence metric to the definition of $G^{-1}_{(g_n^{ed})_{n=1}^{\infty}}(K)$. □

Proposition 6.9 not only gives us the minimizing metric. It also shows that this minimizing metric is independent of the policy of the cheating player. For the purpose of summarizing the results, we state the following theorem.

Theorem 6.1. *For the* product*-regime, we have,*

$$\limsup_{n\to\infty} \frac{1}{n} \max_{f^{(c)}} \min_{\hat{H}\in\hat{\mathscr{H}}} \mathbb{E}[\mathbb{1}\{\hat{H} \neq H\} R(X[1:n], Y_H[1:n])] = \max_{\substack{V_{Y^{(c)}|Z^{(c)}}: \\ D(V_{X,\tilde{Y}^{(c)}}; V_{X,\tilde{Y}^{(r)}}) \leq K}} \mathbb{E}[R(X, Y^{(c)})]. \tag{6.61}$$

This result does not necessarily imply that the empirical divergence metric induces the monitor decision strategy at the Nash equilibrium for every finite value of n. But one can deduce that the difference between the normalized reward of the Nash equilibrium strategy and the empirical divergence strategy is sub-linear in n.

6.4 A Coin Guessing Game

As an illustration, let us consider a specific instance of the problem. In this instance, $X[1:n]$ is a sequence of i.i.d. binary Bernoulli(1/2) random variables. These are observed by the

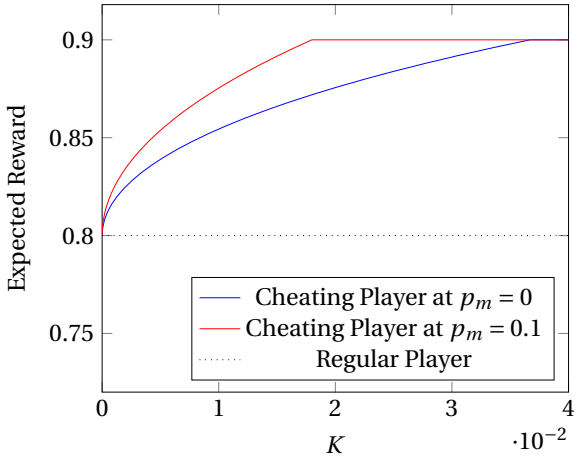


Figure 6.2: Trade-off between the normalized reward and the growth rate K of the number of players. The plots are for $p_c = 0.1$ and $p_r = 0.2$.

cheating player and the regular players through different memoryless channels. The regular player channel $V_{Z^{(r)}|X}$ is $BSC(p_r)$. The cheating player has a stochastically upgraded channel $BSC(p_c)$ with $p_c < p_r \leq 1/2$. The monitoring channel is $BSC(p_m)$, $p_m \leq 1/2$. The goal of the players is to guess the values of $X[i]$'s, so we have $R(X[1:n], Y[1:n]) = \sum_{i=1}^{n} \mathbb{1}\{X[i] = Y[i]\}$.

The optimization problem in Theorem 6.1 is a convex optimization problem, as the objective function is a linear function of $P_{Y^{(c)}|Z^{(c)}}$. The constraint is also convex as $V_{\tilde{Y}^{(c)}|X}$ is linear w.r.t. to $P_{Y^{(c)}|Z^{(c)}}$, while $D(\cdot; V_{\tilde{Y}^{(r)}|X}|V_X)$ is also convex.

For the regular player, the optimal strategy is to assign $Y^{(r)} = Z^{(r)}$. This strategy leads to the expected reward of $\mathbb{E}[R(X, Y^{(r)})] = 1 - p_r$. We can see that this problem is symmetric with respect to X, so the optimal $P_{Y^{(c)}|Z^{(c)}}$ is also BSC with certain flip probability. For this problem, we can express the optimal reward of cheating player in Theorem 6.1 as

$$\max_{\substack{p\in[0,1]:\\ d_2\left(p_c * p * p_m; p_r * p_m\right) \leq K}} 1 - (p * p_c) \tag{6.62}$$

where $p * p' := p(1-p') + p'(1-p)$ and $d_2(\cdot;\cdot)$ is the binary KL divergence.

A numerical example of the trade-off for a specific value of p_c and p_r is given in Figure 6.2. We can observe several properties,

- Even if $p_m = 0$ (i.e., the monitor has perfect knowledge of all players' actions), the cheating player can still improve its normalized reward by exploiting its information advantage if the number of regular players is large enough.
- There is a cut-off K^*, after which the cheating player does not get any further advantage from larger K. This K^* corresponds to the conditional KL divergence between the optimal distribution of cheating player and the regular player.

7 Conclusion and Future Works

In the first part of this thesis, we delved into the problem of information velocity. We showed that the strong data processing inequality enforced a recurrence relation on the F-information between the relays' observations and the original message. To build this recurrence relation, we also introduced the notion of tilted F-divergences. We proved a converse result by modifying standard converse arguments in information theory to bound the error probability of any relaying scheme strictly away from 0 if the F-information vanishes. We showed that if the asymptotic ratio between the length of an arbitrary relaying scheme and the number of its channel uses is above the channel contraction coefficient, then the F-information between the final relay's observations and the original message vanishes. Hence, it establishes the channel contraction coefficient as an upper bound to the information velocity of arbitrary relaying scheme.

We also introduced two achievability schemes which achieve positive information velocity. These schemes capture two different intuitions on how to prove a positive information velocity on the entire chain based on proving single-hop results.

The first idea is based on the notion of eventual reliability, namely the property of a channel coding scheme which ensures that each channel use improves the reliability of all past transmissions. This notion has some connections with the idea of anytime capacity, where we can bound the error probability of a communication scheme based on the delay that we are willing to accept. Using tree codes on BEC channels, we achieved eventual reliability while controlling the average delay needed to transmit each bit through a single hop. We then proved that the total transmission time, which can be decomposed into the transmission time of the first bit and the time interval between the reception of the first bit and the last bit, is dominated by the former term. Hence, we managed to bound the number of channel uses by the relaying scheme based on the concentration of the sum of realized delays at each hop.

The second idea is what we call a "control" perspective of information velocity. Intuitively, the idea is for each relay to maintain an estimate of the message, and the role of the communication scheme at each hop is simply to reduce the difference between the transmitter's

estimate of the message and the receiver's estimate of the message. The aim is to ensure that we can control the difference between the estimate of the final relay and the actual message by controlling the difference at each hops. In our case, we exploited this idea for AWGN channels with feedback using Schalkwijk-Kailath(SK) scheme, where the goal of the SK scheme is simply to reduce the L_2 difference between the transmitter's and the receiver's estimates of the message at each hop.

However, there are several directions that we did not manage to pursue, especially in the area around information velocity.

1. In chapter 3, we repeatedly drew the analogue between the bound that we derived and the transport equations; as we cannot help but notice that the equation that arises in the upper bound closely resembles the discretized version of transport equations under the finite-element method. An interesting line of investigation would be to see whether this connection can inspire a method to upper bound information velocity in other network topologies (think of a regular grid or Rényi graph for example). We expect that we can use the results of previous research on bounding the end-to-end channel contraction coefficient in a network. Several of the open questions here would be on: how to properly define the notion of information velocity in a network of relays, how to define the "direction" of data processing inequality in this network, and what is the correct asymptotic regime such that an approximation by some differential-equation-like structure is feasible.

2. In chapter 4, we leaned heavily on the fact that the relays can measure their local progress in BEC, hence it can simulate a queueing model on its tape abstraction. It would be interesting to study what can be done if the relays do not have this information, especially for channels with undetected errors such as BSC or AWGN channels. One idea that comes to mind is for the relay to use some kind of reliability criterion to measure its decoding progress, e.g., it decides that a particular branch in its tree code is the correct branch if the sum of its posterior probability and its closest siblings' posterior probability is above a certain threshold.

3. In chapter 5, the role of the communication scheme is to reduce the local difference of the transmitter's and the receiver's message estimates at each hop. An interesting line of research would be to study whether the posterior matching approach (Shayevitz and Feder, 2007) allows us to develop similar communication scheme for more general channels. This might not be a far-fetched idea, as previously there has been an analysis of posterior matching scheme in terms of the reduction in KL divergence between the transmitter's and receiver's estimates of the message (Coleman, 2009). The main challenge here is on how to adapt the posterior matching scheme to the relaying problem. In posterior matching, it is assumed that the transmitter has a complete information about the message, whereas in relaying problems, the transmitter only has "soft" information about the message. Hence it is not clear what is the correct divergence metric that

we should be tracking, and how to incorporate this "soft" information in a posterior matching scheme.

4. Also in chapter 5, we remarked that the scheme that we developed is most likely to be optimal if we assume that the relaying scheme is communicating at channel capacity. This raises a question on the nature of the trade-off between channel capacity and information velocity, e.g., does it ever makes sense to communicate at a rate less than channel capacity if it allows for larger information velocity? On the converse side, we expect that the answer to this question is related to the notion of the channel's input-dependent mutual contraction coefficient. The achievability side is more complicated as there is currently a big gap between the achievability results and the converse results for information velocity.

In the second part of the thesis, we explored an adversarial version of hypothesis testing problem where an adversarial data-generating process is hidden among regular data-generating processes. This adversarial data-generating process has to navigate the trade-off between deviating from the regular data-generating processes to exploit its private information and trying to be indistinguishable from the regular data-generating process from the perspective of the tester. We showed that classical information-theoretic arguments, such as arguments based on type method, is powerful enough to allow us to derive a non-trivial bound of what can be achieved by the adversarial data-generating process in this problem. Surprisingly, we can even recover the optimal testing metric, which is the empirical KL divergence.

Currently, our results only hold for the regime where the deviation of the adversarial data-generating process is "constant"(in the sense that that the total variation distance between the adversarial data-generating process and regular data-generating processes converges to a constant). It will be interesting to study other asymptotic regimes of the problem where the deviation in terms of total variation distance vanishes as the number of observations increases, which might pave the way for the case where the number of regular data-generating processes scales sub-exponentially.

Bibliography

Aggarwal, A., Corwin, I., and Ghosal, P. (2023). The ASEP speed process. *Advances in Mathematics*, 422:109004.

Ahlswede, R. and Gacs, P. (1976). Spreading of sets in product spaces and hypercontraction of the markov operator. *The Annals of Probability*, 4(6):925 – 939.

Ali, S. M. and Silvey, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society. Series B (Methodological)*, 28(1):131–142.

Arikan, E. (2009). Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels. *IEEE Transactions on Information Theory*, 55(7):3051–3073.

Barni, M. and Tondi, B. (2013). The source identification game: An information-theoretic perspective. *IEEE Transactions on Information Forensics and Security*, 8(3):450–463.

Berlekamp, E., McEliece, R., and van Tilborg, H. (1978). On the inherent intractability of certain coding problems (corresp.). *IEEE Transactions on Information Theory*, 24(3):384–386.

Blahut, R. (1974). Hypothesis testing and information theory. *IEEE Transactions on Information Theory*, 20(4):405–417.

Bose, R. and Ray-Chaudhuri, D. (1960). On a class of error correcting binary group codes. *Information and Control*, 3(1):68–79.

Coleman, T. P. (2009). A stochastic control viewpoint on 'posterior matching'-style feedback communication schemes. In *2009 IEEE International Symposium on Information Theory*, pages 1520–1524.

Csiszár, I. (1967). Information-type measures of difference of probability distributions and indirect observation. *Studia Scientiarum Mathematicarum Hungarica*, 2:229–318.

Csiszár, I. (1974). Information measures: A critical survey. In *Trans. 7th Prague Conf. Inform. Theory*, Prague.

Csiszár, I. and Körner, J. (2011). *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press, 2 edition.

Dembo, A. and Zeitouni, O. (1998). *Large Deviations Techniques and Applications*. Springer.

Domanovitz, E., Khina, A., Philosof, T., and Kochman, Y. (2023). Information velocity of cascaded Gaussian channels with feedback. preprint:https://arxiv.org/abs/2311.14223.

Domanovitz, E., Philosof, T., and Khina, A. (2022). The information velocity of packet-erasure links. *IEEE INFOCOM 2022 - IEEE Conference on Computer Communications*, pages 190–199.

Dritsoula, L., Loiseau, P., and Musacchio, J. (2017). A game-theoretic analysis of adversarial classification. *IEEE Transactions on Information Forensics and Security*, 12(12):3094–3109.

Fano, R. (1961). *Transmission of Information: A Statistical Theory of Communications*. The MIT Press, Cambridge, MA.

Feder, M. and Merhav, N. (2002). Universal composite hypothesis testing: a competitive minimax approach. *IEEE Transactions on Information Theory*, 48(6):1504–1517.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*, volume 1. Wiley.

Flajolet, P. and Sedgewick, R. (2009). *Analytic Combinatorics*. Cambridge University Press, USA, 1 edition.

Gallager, R. (1962). Low-density parity-check codes. *IRE Transactions on Information Theory*, 8(1):21–28.

Golay, M. J. E. (1949). Notes on digital coding. *Proceedings of the IRE*, 37(657):147–160.

Guo, N. and Kostina, V. (2023). Reliability function for streaming over a DMC with feedback. *IEEE Transactions on Information Theory*, 69(4):2165–2192.

Hamming, R. W. (1950). Error detecting and error correcting codes. *The Bell System Technical Journal*, 29(2):147–160.

Harremoës, P. and Vajda, I. (2010). Joint range of f-divergences. In *2010 IEEE International Symposium on Information Theory*, pages 1345–1349.

Hoeffding, W. (1965). Asymptotically optimal tests for multinomial distributions. *The Annals of Mathematical Statistics*, 36(2):369 – 401.

Horstein, M. (1963). Sequential transmission using noiseless feedback. *IEEE Transactions on Information Theory*, 9(3):136–143.

Huleihel, W., Polyanskiy, Y., and Shayevitz, O. (2019). Relaying one bit across a tandem of binary-symmetric channels. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 2928–2932.

Jensen, J. L. W. V. (1906). Sur les fonctions convexes et les inégualités entre les valeurs moyennes.

Jimenez Felstrom, A. and Zigangirov, K. (1999). Time-varying periodic convolutional codes with low-density parity-check matrix. *IEEE Transactions on Information Theory*, 45(6):2181–2191.

Jin, Y. and Lai, L. (2021). On the adversarial robustness of hypothesis testing. *IEEE Transactions on Signal Processing*, 69:515–530.

Karlin, S. and Rubin, H. (1956). The theory of decision procedures for distributions with monotone likelihood ratio. *The Annals of Mathematical Statistics*, 27(2):272 – 299.

Kaul, S., Yates, R., and Gruteser, M. (2012). Real-time status: How often should one update? In *2012 Proceedings IEEE INFOCOM*, pages 2731–2735.

Kayaalp, M., İnan, Y., Koivunen, V., Telatar, E., and Sayed, A. H. (2023). On the fusion strategies for federated decision making. In *2023 IEEE Statistical Signal Processing Workshop (SSP)*, pages 270–274.

Khina, A., Halbawi, W., and Hassibi, B. (2016). (Almost) practical tree codes. In *2016 IEEE International Symposium on Information Theory (ISIT)*, page 2404–2408. IEEE Press.

Kullback, S. (1959). *Information Theory and Statistics*. Wiley, New York.

Lalitha, A., Javidi, T., and Sarwate, A. D. (2018). Social learning and distributed hypothesis testing. *IEEE Transactions on Information Theory*, 64(9):6161–6179.

Ling, Y. H. and Scarlett, J. (2022). Simple coding techniques for many-hop relaying. *IEEE Trans. Inf. Theory*, 68(11):7043–7053.

Makur, A. and Zheng, L. (2015). Bounds between contraction coefficients. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1422–1429.

Massey, J. (1994). Guessing and entropy. In *Proceedings of 1994 IEEE International Symposium on Information Theory*, page 204.

Massey, J. L., Costello, D. J., and Justesen, J. (1973). Polynomial weights and code constructions. *IEEE Transactions on Information Theory*, 19(1):101–110.

Newey, W. K. and West, K. D. (1987). Hypothesis testing with efficient method of moments estimation. *International Economic Review*, 28(3):777–787.

Neyman, J., Pearson, E. S., and Pearson, K. (1933). Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337.

Polyanskiy, Y., Poor, H. V., and Verdu, S. (2010). Channel coding rate in the finite blocklength regime. *IEEE Transactions on Information Theory*, 56(5):2307–2359.

Polyanskiy, Y., Poor, H. V., and Verdu, S. (2011). Feedback in the non-asymptotic regime. *IEEE Transactions on Information Theory*, 57(8):4903–4925.

Polyanskiy, Y. and Wu, Y. (2015). Strong data-processing of mutual information: beyond Ahlswede and Gács. In *Proc. Information Theory and Applications Workshop*.

Polyanskiy, Y. and Wu, Y. (2016). Dissipation of information in channels with input constraints. *IEEE Transactions on Information Theory*, 62(1):35–55.

Rajagopalan, S. and Schulman, L. (1994). A coding theorem for distributed computation. In *Proceedings of the Twenty-Sixth Annual ACM Symposium on Theory of Computing*, STOC '94, page 790–799, New York, NY, USA. Association for Computing Machinery.

Reed, I. (1954). A class of multiple-error-correcting codes and the decoding scheme. *Transactions of the IRE Professional Group on Information Theory*, 4(4):38–49.

Reed, I. S. and Solomon, G. (1960). Polynomial codes over certain finite fields. *Journal of the Society for Industrial and Applied Mathematics*, 8(2):300–304.

Richardson, T. and Urbanke, R. (2008). *Modern Coding Theory*. Cambridge University Press.

Sahai, A. and Mitter, S. (2006). The necessity and sufficiency of anytime capacity for stabilization of a linear system over a noisy communication link—part i: Scalar systems. *IEEE Transactions on Information Theory*, 52(8):3369–3395.

Sanov, I. N. (1958). On the probability of large deviations of random variables (translated ver.). In *Institute of Statistics mimeo series 192*. North Carolina State University. Dept. of Statistics.

Sason, I. and Verdú, S. (2016). f -divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006.

Schalkwijk, J. and Kailath, T. (1966). A coding scheme for additive noise channels with feedback–i: No bandwidth constraint. *IEEE Transactions on Information Theory*, 12(2):172–182.

Schulman, L. J. (1993). Deterministic coding for interactive communication. In *Proceedings of the Twenty-Fifth Annual ACM Symposium on Theory of Computing*, STOC '93, page 747–756, New York, NY, USA. Association for Computing Machinery.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.

Shannon, C. E. (1949). Communication theory of secrecy systems. *Bell System Technical Journal*, 28(4):656–715.

Shannon, C. E. (1961). Two way communication channels. *Claude E. Shannon: Collected Papers*, pages 351–384.

Shayevitz, O. and Feder, M. (2007). Communication with feedback via posterior matching. In *2007 IEEE International Symposium on Information Theory*, pages 391–395.

Tondi, B., Barni, M., and Merhav, N. (2015). Detection games with a fully active attacker. In *2015 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6.

von Neumann, J., Morgenstern, O., and Rubinstein, A. (1944). *Theory of Games and Economic Behavior (60th Anniversary Commemorative Edition)*. Princeton University Press.

Wozencraft, J. R. (1957). Sequential decoding for reliable communication. Technical Report 325, Massachusetts Institute of Technology (Boston).

Wu, Y. and Zou, W. (1995). Orthogonal frequency division multiplexing: a multi-carrier modulation scheme. *IEEE Transactions on Consumer Electronics*, 41(3):392–399.

Yasodharan, S. and Loiseau, P. (2019). Nonzero-sum adversarial hypothesis testing games. In *Advances in Neural Information Processing Systems*, volume 32.

Reka Inovan

+41-767605985 | reka.inovan@gmail.com | Avenue du 24-Janvier, 1020 Renens VD, Switzerland

Education

École Polytechnique Fédérale de Lausanne — Lausanne, Switzerland
PhD in Computer and Communication Sciences — *Sep. 2018 – Present*

École Polytechnique Fédérale de Lausanne — Lausanne, Switzerland
MSc in Communication Systems — *Feb. 2015 – Sep. 2017*

Universitas Gadjah Mada — Yogyakarta, Indonesia
BSc in Electrical Engineering — *Aug. 2009 – May 2014*

Experiences

Academic Assistant — Nov. 2017 – Jun. 2018
Universitas Gadjah Mada — *Yogyakarta, Indonesia*

Research Assistant — Nov. 2017 – Jan. 2018
Universitas Telkom — *Bandung, Indonesia*

R&D Intern — Aug 2016 – Aug 2017
Sony Europe Technology Center — *Stuttgart, Germany*

Awards

- MSc. Fellowship from Indonesia Endownment Fund for Education
- Fellowship Ph.D. candidate at EPFL IC doctoral school

Teaching Assistant Experiences

- **EPFL** : Foundation of Data Science (2021), Information Theory and Coding (2019, 2020), Principles of Digital Communications (2019, 2021, 2022)
- **UGM** : Linear Algebra (2018), Optimization (2018)

Publications and Preprints

- Y. İnan, R. Inovan, and E. Telatar, "Optimal policies for age and distortion in a discrete-time model," 2022. [Full version]. Preprint: https://arxiv.org/abs/2210.12086.
- R. Inovan, and E. Telatar, "Safety in numbers: asymptotic analysis of a monitoring problem," in Proc. IEEE Information Theory Workshop (ITW), 2022
- Y. İnan, R. Inovan, and E. Telatar, "Optimal policies for age and distortion in a discrete-time model," in Proc. IEEE Information Theory Workshop (ITW), 2021