Capacity-Achieving Schemes for Finite-State Channels

by

Jung Hyun Bae

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Electrical Engineering: Systems)
in The University of Michigan
2012

Doctoral Committee:
Associate Professor Achilleas Anastasopoulos, Chair
Professor George Michailidis
Professor Demosthenis Teneketzis
Associate Professor Sandeep P. Sadanandarao

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF ACRONYMS

ACOE	average cost optimality equation
BEC	binary erasure channel
CSI	channel state information
DMC	discrete memoryless channel
FSC	finite-state channel
FSMC	finite-state Markov channel
GE	Gilbert-Elliott
GEC	generalized erasure channel
ISI	inter-symbol interference
i.u.d.	independent and uniformly distributed
LDGM	low-density, generator matrix
LDPC	low-density, parity-check
MBIOS	memoryless, binary-input, output-symmetric
ML	maximum-likelihood
MQ	Markov quantizer
pmf	probability mass function
PMS	posterior matching scheme
SIR	symmetric information rate
SLLN	strong law of large numbers

CHAPTER 1

Introduction

Consider a typical point-to-point digital communication system as shown in Fig. 1.1. At the transmitter, information from the source is encoded for transmission over the channel. More specifically, the source data is first compressed by the source encoder to remove redundancy and then encoded by the channel encoder for error correction. Subsequently, the modulator generates appropriate waveforms that are transmitted through a noisy channel, and then decoded at the receiver. The main goal for the communication engineer is to design the encoder and decoder so that the system can transmit data reliably at the highest possible transmission rate.

To achieve this goal, channel coding, which is the focus of this thesis, strategically adds redundancy to the transmitted data. With this redundancy, the receiver can more accurately decode the data it receives. C. E. Shannon established in his seminal papers [1, 2] the maximum amount of information that can be transmitted over the channel called the capacity of the channel. Since then, coding theory has provided specific transmission schemes that approach the capacity for point-to-point links.

Historically, most of the aforementioned schemes are designed for the case of memoryless channels. For this class of channels, a multitude of remarkable results have been established relatively recently with the discovery of turbo codes in 1993 [3]

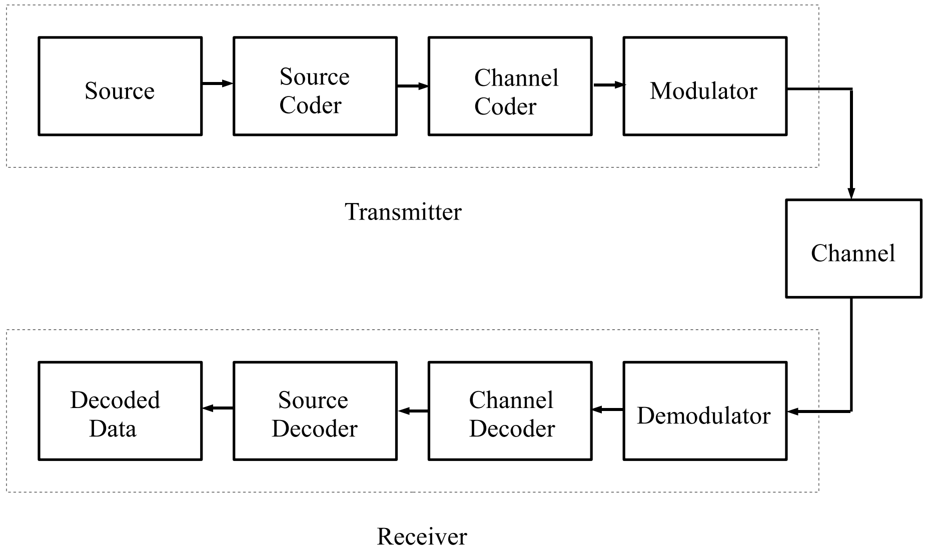

Figure 1.1: Digital communication system

and the subsequent rediscovery of low-density, parity-check (LDPC) codes [4–7]. For channels with memory, however, few results exist on capacity-achieving codes. *This is the first direction we follow in this thesis, i.e. finding capacity-achieving scheme for channels with memory.*

One of the greatest advantages that the aforementioned turbo codes have is low decoding complexity. While maximum-likelihood (ML) decoding (which is the optimal decoding algorithm) requires exponential complexity (with respect to the code length), iterative decoding with message passing, utilized with turbo codes, has much less decoding complexity. This fact has been shown to hold experimentally, although no definitive theoretical study of the overall complexity-performance tradeoff exists for iterative message passive decoding in noisy channels (other than the binary erasure channel (BEC)).

In our quest to find simple encoding and decoding strategies for channels with memory we turn our attention to channels with feedback. Communication in the presence of feedback has been a long studied problem which dates back to Shannon's early work [8]. Shannon showed that feedback cannot increase the capacity of memoryless channels, but can improve the error performance and/or simplify the transmission scheme. There have been several remarkable results on designing transmission schemes with feedback, but again, most of these results are for the case of memoryless channels. *The second direction in this thesis, therefore, is to design simple transmission schemes for channels with memory in the presence of feedback.*

1.1 Background

1.1.1 Capacity-achieving codes for channels without feedback

The capacity and the capacity-achieving input distribution of memoryless, binary-input, output-symmetric (MBIOS) channels, is well known [9, p. 94]. It is also well known that linear codes can achieve the capacity of MBIOS channels [10]. However, capacity achieving codes with low complexity were not known until the remarkable discovery of turbo codes in 1993 [3] and the subsequent rediscovery of LDPC codes [4–7] as mentioned above. Since then, codes defined on sparse graphs, collectively referred to as "turbo-like" or "LDPC-like" codes have attracted a large amount of attention in the quest to achieve channel capacity with small complexity.

For the special case of the BEC, it is known that capacity can be achieved with bounded decoding complexity per information bit. In particular, based on the density evolution (DE) technique, it was shown in [11–13] that LDPC-like codes achieve the BEC capacity with iterative decoding, while [14] showed that this can be done with bounded (and small) decoding complexity per information bit.

For general MBIOS channels, however, there is no conclusive answer regarding capacity achievability of turbo-like codes with iterative decoding. This is so due to the fact that DE is not amenable to theoretical analysis for MBIOS channels, since code performance can only be evaluated through an uncountably-infinite-dimensional non-linear recursive equation (there is a line of research studying iterative decoding performance [15–17] based on results from statistical physics). Thus, only numerical results exist that show how turbo-like codes can approach capacity using iterative decoding [18]. On the other hand, when ML decoding is assumed, the performance of turbo-like codes can be analyzed. For instance, it was shown that LDPC-like codes achieve the capacity of MBIOS channels with ML decoding in [19–22]. Moreover, it was shown in [23] that a family of turbo-like codes consisting of a serial concatenation of an outer LDPC code and an inner rate-one low-density, generator matrix (LDGM) code, can achieve the capacity of MBIOS channels with bounded graphical complexity (i.e., number of edges in the graph representing the code) per information bit when ML decoding is performed.

The aforementioned results were developed for memoryless channels. For channels with memory, few results exist on their capacity and the corresponding capacity-achieving input distribution, and even fewer on capacity-achieving codes. Regarding the former, the capacity of Gilbert-Elliott (GE) channels was studied in [24] and the capacity of more general finite-state Markov channel (FSMC)s was studied in [25, 26]. An efficient method for computing the information rate of a finite-state channel (FSC) whose input is a Markov process was proposed independently in [27–29] and extended in [30]. In [31] an optimization algorithm was proposed to numerically compute tight lower bounds on the capacity of FSCs, and the techniques were generalized in [32]. Upper bounds on the capacity of FSCs were developed in [33, 34]. The above results, demonstrated that a Markov input sequence provides larger information rate than the information rate achievable with independent and uniformly distributed (i.u.d.) input sequences, also called the symmetric information rate (SIR) of the channel. Furthermore, it was recently shown in [35] that a sequence of Markov sources asymptotically achieves the capacity of FSCs as the Markov order approaches infinity.

Regarding capacity-achieving codes for channels with memory, methods of constructing codes which induce Markov distribution on the transmitted sequence were proposed in [36–39], and their performance on partial response channels was evaluated using simulation. Pfister and Siegel [40] introduced a generalized erasure channel (GEC) as a simple model of a channel with memory and proved that LDPC-like codes achieve its SIR. Several authors have investigated the performance of LDPC-like codes with iterative decoding using numerical tools based on DE for some specific channels with memory. In particular, LDPC-like codes were analyzed using numerical tools based on DE for binary inter-symbol interference (ISI) channels in [41], for the GE channel in [42], and for FSMCs in [43]. Finally, the results derived for MBIOS channels in [20], were extended to establish an upper bound on the rate of LDPC codes for GE channels [44] and FSMCs [45].

1.1.2 Capacity-achieving schemes for channels with feedback

As mentioned above, communication in the presence of feedback has been a long studied problem which dates back to Shannon's early work [8]. Horstein [46] proposed a simple sequential transmission scheme which is capacity-achieving and provides larger error exponents than traditional fixed-length block-coding. Similarly, Schalkwijk and Kailath [47] showed that capacity and a double exponentially decreasing error probability can be achieved by a simple sequential transmission scheme for the additive white Gaussian noise channel (AWGNC) with average power constraint. Recently, Shayevitz and Feder [48–50] identified an underlying principle shared by the aforementioned Horstein and Schalkwijk-Kailath schemes and introduced a simple encoding scheme, namely the posterior matching scheme (PMS) for general memoryless channels. Furthermore, they showed that the PMS achieves the capacity of general discrete memoryless channel (DMC)s. Subsequently, Coleman [51] revisited the PMS and provided a proof of capacity achievability by reformulating the problem in a stochastic control framework.

The starting point of our investigation of simple transmission scheme for channels with memory is the derivation of a single-letter capacity expression for this channel. One of the first capacity results for FSCs was by Viswanathan [52], who found the capacity of a FSC with receiver channel state information (CSI) and delayed feedback where there is no ISI. Later, Chen and Berger [53] found the capacity of a FSC with ISI where current CSI is available at the transmitter and the receiver. Yang et. al. [34] used a stochastic control method to find the capacity of the ISI channel. Recently, Tatikonda and Mitter [54] provided a general stochastic control framework for evaluating the capacity of the FSC with feedback. In that paper, the capacity was characterized as the solution of a dynamic programming average cost optimality equation (ACOE). Como et. al. [55] used an approach similar to that in [54] to find the capacity of the FSC when current CSI is available at the transmitter and the receiver. An upper bound on the capacity of the FSC without ISI and CSI was found using dynamic programming by Huang et. al. [56].

1.2 Thesis contributions and outline

The general philosophy of this thesis and its main contribution on the conceptual level, is that a number of problems relating to information theory can be solved effectively if one takes the viewpoint of stochastic control. We identify and solve two such problems.

The first problem is that of evaluating the capacity of a FSC. We show that this off-line optimization problem can be formulated as a centralized sequential stochastic control problem and solved using Markov decision theory. This approach is general enough and we anticipate that it can be applied to other information theoretic problems involving the evaluation of the channel capacity (or the channel capacity region for multi-user scenaria).

The second problem involves the design of on-line transmission schemes for FSCs. When feedback is not present, a simple modification of the corresponding transmission schemes utilized for memoryless channels seems to be sufficient for achieving capacity. The case of channels with feedback is more interesting. In this case one may attempt to formulate the problem of finding the optimal (in the sense of achieving capacity) transmission scheme as a stochastic control problem involving the transmitter and the receiver. Such an approach may lead to a implicit characterization of the optimal transmission scheme (through an optimality equation). However we are following a different approach: we propose a transmission scheme (which generalizes previously proposed schemes for memoryless channels) and show that it achieves capacity. This is done by evaluating the asymptotic properties of this scheme through the study of an appropriate Markov chain. As in the case of the first problem mentioned above, the approach is quite general and it is anticipated that it can be applied to a larger class of problems.

In the following we detail the individual contributions on a per chapter basis.

In Chapter 2, capacity-achieving codes are constructed for FSCs (with or without ISI) with ML decoding when there is no feedback. The codes are derived from the corresponding capacity-achieving codes for MBIOS channels through simple modifications. In particular, it is shown that several LDPC-like coset codes which are capacity-achieving on MBIOS channels achieve the SIR of FSCs. Next, a family of *quantized coset codes* is constructed by using block-wise Markov quantization of LDPC-like codes. This technique generalizes the quantization technique presented in [57] for memoryless channels and results in a simple encoding and iterative decoding algorithm. The constructed quantized codes induce a k-th order Markov distribution on the channel input sequence and are shown to achieve the corresponding information rate. The basic analytical tool used to prove these results is an upper bound on the ensemble average of the corresponding modified LDPC-like ensembles. This bound is a non-trivial generalization of the "union plus Shulman and Feder" bound [19, 57, 58] which relates the average code error probability with the asymptotic growth rate of the average code weight enumerator, the random coding exponent for channels with memory and an appropriately defined Battacharrya-like parameter. The results of this chapter have been published in [59, 60].

In case when there is feedback, we generalize the PMS for FSCs where the channel state is affected both by nature and by the input sequence (thus introducing ISI), the CSI and output are available at the receiver, and the CSI and output are available at the transmitter with some finite delay through noiseless feedback. In Chapter 3, we derive a single-letter capacity expression for FSCs with ISI as the starting point of our investigation of simple transmission schemes. We point out that although the channel considered in this thesis is indeed a special case of the one considered in [54], our approach in deriving a single-letter capacity expression provides more intuition, and the resulting capacity expression is significantly simpler. We also consider a special no ISI case and obtain the result presented in [52] using considerably simpler approach. The results of this chapter have been published in [61]. Based on this simplified capacity, we propose generalized PMS for FSCs. In Chapter 4, we consider the case when there is no ISI. The PMS is a rather straightforward generalization of PMS for DMCs. We show that the proposed PMS achieves the capacity of FSCs with feedback and no ISI. In Chapter 5, we turn our attention to FSCs with feedback and ISI. Because of the ISI, the resulting PMS scheme becomes considerably different from its DMC counterpart, and construction of this scheme requires careful consideration of necessary conditions of capacity achievability. We also show that the proposed PMS achieves the capacity of FSCs with feedback and ISI. In proving capacity achievability in Chapters 4 and 5, we generalize the tools provided in [50]. Analysis consists of two parts. First, we prove zero-rate achievability, and based on this we show positive rate achievability up to capacity. Important steps in the first part are identifying appropriate Lyapunov functions and contraction mappings. The strong law of large numbers (SLLN) of Markov chains plays an important role in bridging the first and the second part. The results of Chapters 4 and 5 are prepared for publication in [62].

The rest of the thesis is organized as follows. In Chapter 2, quantized LDPC-like codes are constructed and it is shown that they are capacity-achieving for FSCs without feedback. A single-letter capacity expression for FSCs with feedback is derived in Chapter 3. The PMS for FSCs with feedback is defined, and its capacity achievability is proven in Chapter 4 for no ISI case and Chapter 5 for ISI case. The conclusions of this thesis including future research directions are presented in Chapter 6.

CHAPTER 2

Capacity achieving codes for finite-state channels without feedback

2.1 Channel model and preliminaries

Let $\{S_n\}_{n=1}^{\infty}$ with $s_n \in \mathcal{S} = \{1, 2, ..., K\}$ be a sequence representing the channel states at time n. It is assumed that the state space $\mathcal{S}$ corresponds to K different MBIOS channels. Let $\{X_n\}_{n=1}^{\infty}$ be the random process representing the channel input sequence, where $x_n \in \mathcal{X} = \{0, 1\}$. Let $\{Y_n\}_{n=1}^{\infty}$ be the random process representing the channel output sequence, where $y_n \in \mathcal{Y}$ and $\mathcal{Y}$ is assumed to be a discrete subset of the real line (a continuous subset can also be assumed throughout the chapter by changing summations over $\mathcal{Y}$ to integrations) symmetric around zero. Since for each state the channels are symmetric, it is true that $Q(y_n|x_n, s_n) = Q(-y_n|x_n \oplus 1, s_n)$, where $\oplus$ denotes modulo-2 addition. Let $P(\mathbf{x}^N, \mathbf{y}^N)$ be the joint pmf of $\mathbf{X}^N$ and $\mathbf{Y}^N$, where $\mathbf{x}^N$ denotes the length-N vector $(x_1, \ldots, x_N)$. Then,

$$P(\mathbf{x}^N, \mathbf{s}^N, \mathbf{y}^N) = Q(\mathbf{y}^N|\mathbf{x}^N, \mathbf{s}^N) P(\mathbf{s}^N|\mathbf{x}^N) P(\mathbf{x}^N) \tag{2.1a}$$

$$= P(\mathbf{x}^N) P(s_1) \prod_{n=1}^{N} Q(y_n|x_n, s_n) \prod_{n=1}^{N-1} P(s_{n+1}|s_n, x_n). \tag{2.1b}$$

Implicit in (2.1), is the fact that the considered channel states are affected by both nature and ISI. In this chapter, we also consider non-inverting channels, that is, channels with the property

$$\forall s \in \mathcal{S}, \qquad \sum_{y>0} Q(y|1,s) > \sum_{y<0} Q(y|1,s). \tag{2.2}$$

For any sequence of real-valued random variables $(Z_1, Z_2, Z_3, ...)$, define the *limit inferior in probability* $p - \liminf_{N\to\infty} Z_N$ as

$$p - \liminf_{N\to\infty} Z_N \triangleq \sup\{\alpha| \lim_{N\to\infty} Pr[Z_N < \alpha] = 0\}. \tag{2.3}$$

Then, the capacity of the aforementioned FSC when no channel state information is available at the transmitter and the receiver is defined as [63, sec. 3.2]

$$C \triangleq \sup_{\mathbf{X}} \underline{I}(\mathbf{X};\mathbf{Y}), \tag{2.4}$$

where

$$\underline{I}(\mathbf{X};\mathbf{Y}) = p - \liminf_{N\to\infty} \frac{1}{N} \log_2 \frac{Q(\mathbf{y}^N|\mathbf{x}^N)}{P(\mathbf{y}^N)}. \tag{2.5}$$

Since the focus of this thesis is not in finding this maximizing input distribution, in the following we will be interested in the maximum achievable rate for a certain sequence of input pmfs $P \triangleq \{P(\mathbf{x}^N)\}_N$, defined as

$$C_P \triangleq \underline{I}(\mathbf{X};\mathbf{Y}), \tag{2.6}$$

where $I(\mathbf{X};\mathbf{Y})$ is computed under $\{P(\mathbf{x}^N)\}_N$ using (2.1). The maximum achievable rate relative to independent and uniformly distributed (i.u.d.) inputs is also known as SIR (for binary input) and will be denoted as

$$SIR \triangleq C_{IUD}. \tag{2.7}$$

The achievability of C_P can be shown by the following coding theorem which is a modification of channel coding theorem using information spectrum methods [63, Th. 3.2.1] according to the idea in [58].

Proposition 2.1. *Consider an arbitrary discrete channel. Let $Q(\mathbf{y}^N|\mathbf{x}^N)$ be the conditional pmf for sequences of length $N \geq 1$ on this channel. Let $P(\mathbf{x}^N)$ be an arbitrary input pmf. Consider an ensemble of codes of size M, length N, and rate R, whose codewords $\mathbf{c}_m$, $1 \leq m \leq M$ satisfy the following properties*

$$Pr[\mathbf{c}_i = \mathbf{x}^N] = P(\mathbf{x}^N) \qquad \forall i \in \{1, \ldots, M\}, \tag{2.8}$$

$$Pr[\mathbf{c}_i = \mathbf{x}^N | \mathbf{c}_j = \mathbf{x}'^N] \leq \alpha P(\mathbf{x}^N)$$
$$\forall i, j \in \{1, \ldots, M\} \text{ with } i \neq j \text{ and } d(\mathbf{c}_i, \mathbf{c}_j) \in U^c, \tag{2.9}$$

where $U \subseteq \{1, 2, ..., N\}$. Note that $\alpha = 1$ for an ensemble of codes whose codewords are selected independently with pmf $P(\mathbf{x}^N)$. Suppose that an arbitrary message m, $1 \leq m \leq M$ enters the encoder and that ML decoding is employed. Then for any input pmf $P = P(\mathbf{x}^N)$ and for any $\epsilon > 0$, the average probability of decoding error over this ensemble of codes, is bounded as

$$\overline{P}_{e|m} \leq \overline{P}^U_{e|m} + Pr[\mathbf{X}^N \mathbf{Y}^N \notin T_N] + 2^{-N(C_P - R - \frac{\log_2 \alpha}{N} - \epsilon)}, \tag{2.10a}$$

where

$$\overline{P}^U_{e|m} = \sum_C Pr[C] P^U_{e|m}, \tag{2.10b}$$

$P^U_{e|m}$ *denote the probability that there exists some codeword* $\mathbf{c}_{m'}$ *such that* $Q(\mathbf{y}|\mathbf{c}_{m'}) \geq Q(\mathbf{y}|\mathbf{c}_m)$ *and* $d(\mathbf{c}_m, \mathbf{c}_{m'}) \in U$,

$$T_N = \left\{(\mathbf{x}^N, \mathbf{y}^N) \in \mathcal{X}^N \times \mathcal{Y}^N | \frac{1}{N} \log_2 \frac{Q(\mathbf{y}^N|\mathbf{x}^N)}{P(\mathbf{y}^N)} > C_P - \epsilon\right\}. \tag{2.10c}$$

Proof. See 2.4. □

Note that the presence of the factor α in (2.9) allows one to deal with structured ensembles for which strict independence of the codeword selection cannot be guaranteed, as long as one shows that the rate loss due to α as manifested in (2.10a) approaches zero asymptotically for long codes. In the following section we give specific expressions for α for several structured ensembles of interest.

We now summarize the bounding techniques developed in the literature for showing that certain turbo-like ensembles are capacity achieving for MBIOS channels. As it will become evident in the following sections, the bounds derived for the FSCs are similar in structure to their MBIOS counterparts. An upper-bound on the error probability of linear codes transmitted over MBIOS channels was derived in [19]. The bound involves two terms; the first term is based on the Shulman and Feder bound [58] for MBIOS channels and the second term is based on a union bound which involves the Battacharrya parameter for MBIOS channels. For a given ensemble $\mathcal{C}$ of codes with length N, we denote the average weight enumerator of the ensemble by by $\overline{A}_l$, $l = 0, 1, \ldots, N$. The error probability bound is presented in the following fact.

Fact 2.2. *Consider an ensemble* $\mathcal{C}$ *of linear codes with average weight enumerator* $\overline{A}_l$, *where each code is comprised of* M *codewords of length* N. *Let* $Q(y|x)$ *be an MBIOS channel, and let*

$$D_0 \triangleq \sum_y \sqrt{Q(y|0)Q(y|1)} \tag{2.11}$$

be the Battacharrya channel parameter. Denote the ensemble averaged ML decoding error probability by $\overline{P_e}$. Let $U \subseteq \{1, 2, ..., N\}$. Then

$$\overline{P}_e \leq \sum_{l \in U} \overline{A}_l D_0^l + 2^{-NE_r(R + \frac{\log_2 \alpha}{N})}, \tag{2.12}$$

where

$$\alpha \triangleq \max_{l \in U^c} \frac{\overline{A}_l}{M-1} \frac{2^N}{\binom{N}{l}}, \tag{2.13}$$

and where$E_r(\cdot)$ is the random coding error exponent

$$E_r(R) \triangleq \max_{P} \max_{0 \leq s \leq 1} \left\{-sR + E_0(s, P)\right\}, \tag{2.14a}$$

$$E_0(s, P) \triangleq -\log_2 \left\{ \sum_y \left\{ \sum_x P(x) Q(y|x)^{\frac{1}{1+s}} \right\}^{1+s} \right\}. \tag{2.14b}$$

As discussed in [19], the second part of (2.12) approaches $2^{-NE_r(R)}$ provided that the asymptotic growth rate of $\overline{A}_l$ approaches the asymptotic growth rate, $H(l/N) + R - 1$, of the random ensemble for $l \in U^c$. Therefore, in order to prove that a specific sequence of ensembles approaches capacity, we have to choose U such that the above statement is true and the first part of (2.12) approaches zero. Consider the following three ensembles with increasing degree of sophistication.

Ensemble A is the regular Gallager (N, d_v, d_c) LDPC ensemble [64] with column and row degrees d_v and d_c, respectively (we also assume an even d_c).

Ensemble B is the punctured LDPC ensemble introduced in [22] resulting from puncturing a sufficiently low-rate (N, d_v, d_c) Gallager ensemble to achieve the specified rate.

Ensemble C is the LDPC-GM ensemble introduced in [23] consisting of a a serial concatenation of an outer (N, d_v, d_c) Gallager LDPC code and an inner rate-one regular LDGM code with row and column degrees equal to d_c.

For all these three ensembles, we can find an appropriate set U and prove capacity achievability of MBIOS channels [22, 23] as can be seen in the following fact.

Fact 2.3. *For any $\epsilon > 0$, the sequence of the aforementioned ensembles A, B, and C, have vanishing average block error probability under ML decoding on MBIOS channels, and limiting (with respect to N) rate $(1-\epsilon)C$, where C is the channel capacity, when the following conditions are satisfied.*

Ensemble A: $d_c \geq d_c(\epsilon)$.

Ensemble B: $d_c > d_v \geq 5$, *d_c even, and $R_0 \leq R_0(\epsilon)$, where R_0 is the rate of the original code before puncturing.*

Ensemble C: $d_c \geq d_{c,min}$, *d_c even, and $d_v \geq 4$, where $d_{c,min}$ is independent of ϵ and bounded.*

Proof. The dependence of the upper bound in Fact 2.2 on the code ensemble is only through $\overline{A}_l$ and in particular only through the corresponding asymptotic growth rate. In [22, 23] upper bounds on these asymptotic growth rates have been developed for the ensembles A, B, and C. Combining Fact 2.2 with these upper bounds, and with an appropriate choice of U as in [19, 22, 23], proves the fact under the conditions mentioned above.

We present the choice of U, for ensemble A as an example. The choices for ensembles B and C are similar to that of the ensemble A. It was shown in [22] that $\overline{A}_l$ of the ensemble A satisfies

$$\begin{aligned} w_0(a) \leq (1-R)\log_2[1+(1-2a)^{d_c}] \\ + [H(a)-(1-R)], \qquad 0 \leq a \leq 1, \end{aligned} \tag{2.15}$$

where $w_0(a) \triangleq \lim_{N\to\infty} \frac{1}{N}\log_2 \overline{A}_{\lceil Na \rceil}$, and $R = 1 - d_v/d_c$. Furthermore, there exists a $\delta_0 \in (0, 1/2)$ such that

$$\sum_{l \in (0, N\delta_0)} \overline{A}_l = O(N^{-d_v+2}). \tag{2.16}$$

Finally, when d_c is even, $\overline{A}_l = \overline{A}_{N-l}$, for all $l \in \{0, 1, 2, ..., N\}$.

Let

$$U \triangleq \left\{ l : \frac{l}{N} \in (0, \delta_0) \cup (1 - \delta_0, 1] \right\}. \tag{2.17}$$

With this choice of U, the union bound term becomes

$$\sum_{l \in U} \overline{A}_l D_0^l = \sum_{l \in U \setminus \{N\}} \overline{A}_l D_0^l + \overline{A}_N D_0^N \tag{2.18a}$$

$$\leq \sum_{l \in U \setminus \{N\}} \overline{A}_l + 1 \times D_0^N = O(N^{-d_v+2}), \tag{2.18b}$$

and the rate loss in the error exponent is bounded above by $\log_2[1 + (1 - 2\delta_0)^{d_c}]$. As a result, for $d_v \geq 3$ and $\log_2[1 + (1 - 2\delta_0)^{d_c}] \leq \epsilon C$, vanishing error probability is ensured for all rates up to $(1 - \epsilon)C$. The last inequality is guaranteed by selecting a large enough $d_c \geq d_c(\epsilon) \stackrel{\text{def}}{=} \log_2(2^{\epsilon C} - 1)/\log_2(1 - 2\delta_0)$. □

2.2 SIR-achieving codes for FSCs

In order to analyze the performance of code ensembles on FSCs, we first derive an upper-bound on the error probability of linear coset codes transmitted over FSCs, which is similar to the bound in Fact 2.2. Since we are interested in turbo-like ensembles that achieve the SIR of FSCs, and since linear codes do not necessarily induce an i.u.d. input, we need to enlarge linear code ensembles in a way similar to [41]. This motivates the following definition.

Definition 2.4. Consider a code ensemble $\mathcal{C}$. The coset ensemble $\mathcal{C}'$ generated by $\mathcal{C}$ is defined as the ensemble generated from $\mathcal{C}$ by including, for each code $C \in \mathcal{C}$, codes of the form $C' = \{\mathbf{c} \oplus \mathbf{v} | \mathbf{c} \in C\}$ for all possible vectors $\mathbf{v}$. The measure on $\mathcal{C}'$ is the product of the uniform measure over all vectors $\mathbf{v}$ and the measure of the original ensemble $\mathcal{C}$.

In the following lemma, we establish an upper-bound for the pairwise error probability between two sequences. This will be needed for the union-bound term of the aforementioned bounding technique. We denote by $P_{e|\mathbf{x}^N,\mathbf{x}'^N}$ the pairwise error probability of decoding the sequence $\mathbf{x}'^N$ with ML decoding on a FSC conditioned on $\mathbf{x}^N$ being transmitted.

Lemma 2.5. *The pairwise error probability $P_{e|\mathbf{x}^N,\mathbf{x}'^N}$ is upper-bounded as*

$$P_{e|\mathbf{x}^N,\mathbf{x}'^N} \leq D^{d(\mathbf{x}^N,\mathbf{x}'^N)}, \tag{2.19}$$

where $d(\mathbf{x}^N, \mathbf{x}'^N)$ denotes the Hamming distance between $\mathbf{x}^N$ and $\mathbf{x}'^N$, and

$$D \triangleq \min_{s_0} \max_{s} \left[Q_s^+ \sqrt{\frac{Q_{s_0}^-}{Q_{s_0}^+}} + Q_s^- \sqrt{\frac{Q_{s_0}^+}{Q_{s_0}^-}} \right], \tag{2.20}$$

where

$$Q_s^+ = \frac{1}{2} Q(0|1,s) + \sum_{y>0} Q(y|1,s), \tag{2.21a}$$

$$Q_s^- = \frac{1}{2} Q(0|1,s) + \sum_{y<0} Q(y|1,s). \tag{2.21b}$$

Furthermore, $D < 1$.

Proof. See 2.5. □

In order to utilize the techniques in [58] and [19] for bounding the error probability, a special kind of linear code ensemble is considered. Let Π_N denote the set of all permutations of N numbers. We define a *permutation-invariant* code ensemble as follows.

Definition 2.6. Let $\mathcal{C}$ be an ensemble of length-N block codes. We say that $\mathcal{C}$ is a *permutation-invariant* ensemble if for all permutations $\pi \in \Pi_N$ and for all codes $C \in \mathcal{C}$, it is true that $\pi(C) \in \mathcal{C}$ and the codes $\pi(C)$ are selected with the same probability. Here $\pi(C)$ denotes the codebook constructed by permuting the order of the symbols in all the codewords of C according to π.

We are now ready to establish an upper bound on the average error probability of permutation-invariant coset ensembles on FSCs.

Proposition 2.7. *Consider a permutation-invariant ensemble $\mathcal{C}$ of binary linear codes with M codewords of length N, rate R, and average weight enumerator $\overline{A}_l$. Consider the coset ensemble $\mathcal{C}'$ generated by $\mathcal{C}$. Let $U \subseteq \{1, 2, ..., N\}$. Then, for any $\epsilon > 0$, the average (over $\mathcal{C}'$) error probability with ML decoding given the mth message is transmitted is upper-bounded as*

$$\overline{P}_{e|m} \leq \sum_{l \in U} \overline{A}_l D^l + Pr[\mathbf{X}^N \mathbf{Y}^N \notin T_N] + 2^{-N(C_{IUD} - R - \frac{\log_2 \alpha}{N} - \epsilon)}, \tag{2.22}$$

where D is as defined in Lemma 2.5 and T_N is as defined in Theorem 2.1,

$$\alpha \triangleq \max_{l \in U^c} \frac{\overline{A}_l}{M-1} \frac{2^N}{\binom{N}{l}}, \tag{2.23}$$

Proof. The proof of this theorem is omitted. This is a simplified version of the proof of Theorem 2.13 that follows. □

Since the aforementioned ensembles A, B and C are permutation-invariant, Theorem 2.7 can be applied for those ensembles to prove the SIR-achievability which is stated as the following corollary.

Corollary 2.8. *For any $\epsilon > 0$, there exists a sequence of the coset ensembles generated by the ensembles A, B, and C, have limiting (with respect to N) rate $(1-\epsilon)C_{IUD}$ and vanishing average block error probability under ML decoding on FSCs when the conditions stated in Fact 2.3 are satisfied.*

Proof. From the definition of T_N, we have $Pr[\mathbf{X}^N\mathbf{Y}^N \notin T_N] \to 0$ as $N \to \infty$. Since the bound in Fact 2.2 and the first and the third term of the bound in Theorem 2.7 have a similar form, the same technique used in Fact 2.3 can also be used to prove SIR-achievability for the sequence of ensembles A, B, and C satisfying Theorem 2.7. □

Some observations regarding the complexity of the aforementioned SIR-achieving ensembles are in order. Regarding ensemble A, the quantity $d_c(\epsilon)$ is diverging to ∞ as $\log(1/\epsilon)$ when ϵ approaches zero. Therefore, the density of the parity check matrix, and thus the number of edges in the graph per information bit approaches ∞, which implies an infinite complexity per information bit per iteration even when iterative decoding is applied. A similar conclusion can be reached for ensemble B, as detailed in [22]. The advantage of B over A is the universality of the former, i.e., a single mother code of low rate can be used to approach a wide range of channel SIRs for different FSCs. Finally, ensemble C achieves the SIR with bounded number of edges per information bit in the graph representing the code and thus it can be decoded with finite complexity per information bit per iteration when iterative decoding is performed. It should be noted however, that SIR-achievability is only guaranteed for ML decoding of these ensembles.

We also note that if we concentrate only on ensemble C, neither the symmetric channel assumption nor the non-inverting channel assumption is required for the proof of SIR achievability. This is so because these two assumptions only enter our arguments through Lemma 2.5. However, the first term of (2.22) can always be substituted by $\sum_{l \in U} \bar{A}_l$, which makes the overall upper bound independent of the Battacharrya parameter. Moreover, for ensemble C, for all $l \in U$ the asymptotic growth rate of the number of codewords with weight l is negative [23], and thus the above sum is converging to zero.

We finally point out that the three examples mentioned above are only examples of possible SIR-achieving ensembles. Other ensembles may also possess this property and may also possess additional properties that make them more desirable in practical applications.

2.3 Capacity-achieving codes on FSCs

In section 2.2, we proposed several LDPC-like coset code ensembles that achieve the SIR of FSCs. For general FSCs, however, the capacity could be greater than the SIR. Motivated by the need to achieve rates above the SIR, in this section we propose a simple quantization technique that induces a Markov distribution on the transmitted sequence and analyze its performance.

2.3.1 Construction of quantized coset code ensembles

Bennatan and Burshtein [57] presented a method of constructing codes for transmission over arbitrary memoryless channels by using a linear code followed by a simple memoryless quantization technique. Since memoryless quantization can only induce an i.i.d. (not necessarily uniform) on the input sequence, and since the capacity achieving input might not be i.i.d. for FSCs, we present a modified quantization technique that can induce a k-th order (stationary) Markov distribution on the input sequence. The block diagram of the proposed scheme is shown in Fig. 2.1. Other methods of constructing codes which induce Markov distribution on the transmitted

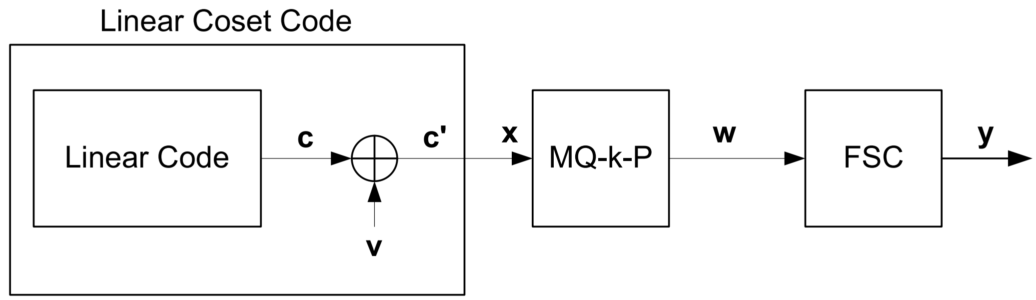

Figure 2.1: Capacity achieving transmission scheme with coset codes and Markov quantization.

sequence can be found in [37, 38].

Definition 2.9. Consider a sequence $\mathbf{x}^{NT}$ and some arbitrary function $f : \{0,1\}^T \times \{0,1\}^k \to \{0,1\}$. An order-$k$ Markov quantizer (denoted by MQ-k) is a mapping $\delta : \{0,1\}^{NT} \to \{0,1\}^N$ with $\delta(\mathbf{x}^{NT}) = \mathbf{w}^N$, with the following structure[1]

$$w_n = f(\mathbf{x}_{(n-1)T+1}^{nT}, \mathbf{w}_{n-k}^{n-1}), \qquad n = 1, 2, \ldots, N. \tag{2.24}$$

Consider now a pmf of a k-th order stationary Markov process $P(\mathbf{w}^N) = \prod_{n=1}^{N} P(w_n|\mathbf{w}_{n-k}^{n-1})$ for a binary sequence of length N. An order-k Markov quantizer with respect to P (denoted by MQ-k-P) is an MQ-k satisfying

$$\frac{|\{\mathbf{x}_{(n-1)T+1}^{nT} | f(\mathbf{x}_{(n-1)T+1}^{nT}, \mathbf{w}_{n-k}^{n-1}) = 0\}|}{2^T} = P\left(0 \mid \mathbf{w}_{n-k}^{n-1}\right),$$
$$\forall \mathbf{w}_{n-k}^{n-1} \text{ and } n = 1, 2, \ldots, N \tag{2.25}$$

More descriptively, an MQ$-k$ partitions a length-NT binary sequence into N blocks of length T each, and then quantizes each block into a bit using a mapping that depends also on the k previously produced bits, thus producing a length-N binary sequence. This is shown in Fig. 2.2. Furthermore, if the input sequence to

[1] For notational simplicity we do not specify precisely the quantizer for the first k symbols. These edge effects will be negligible for large N. In the following we use $\mathbf{x}_i^j$ to denote the subsequence $(x_i, x_{i+1} \ldots, x_j)$.

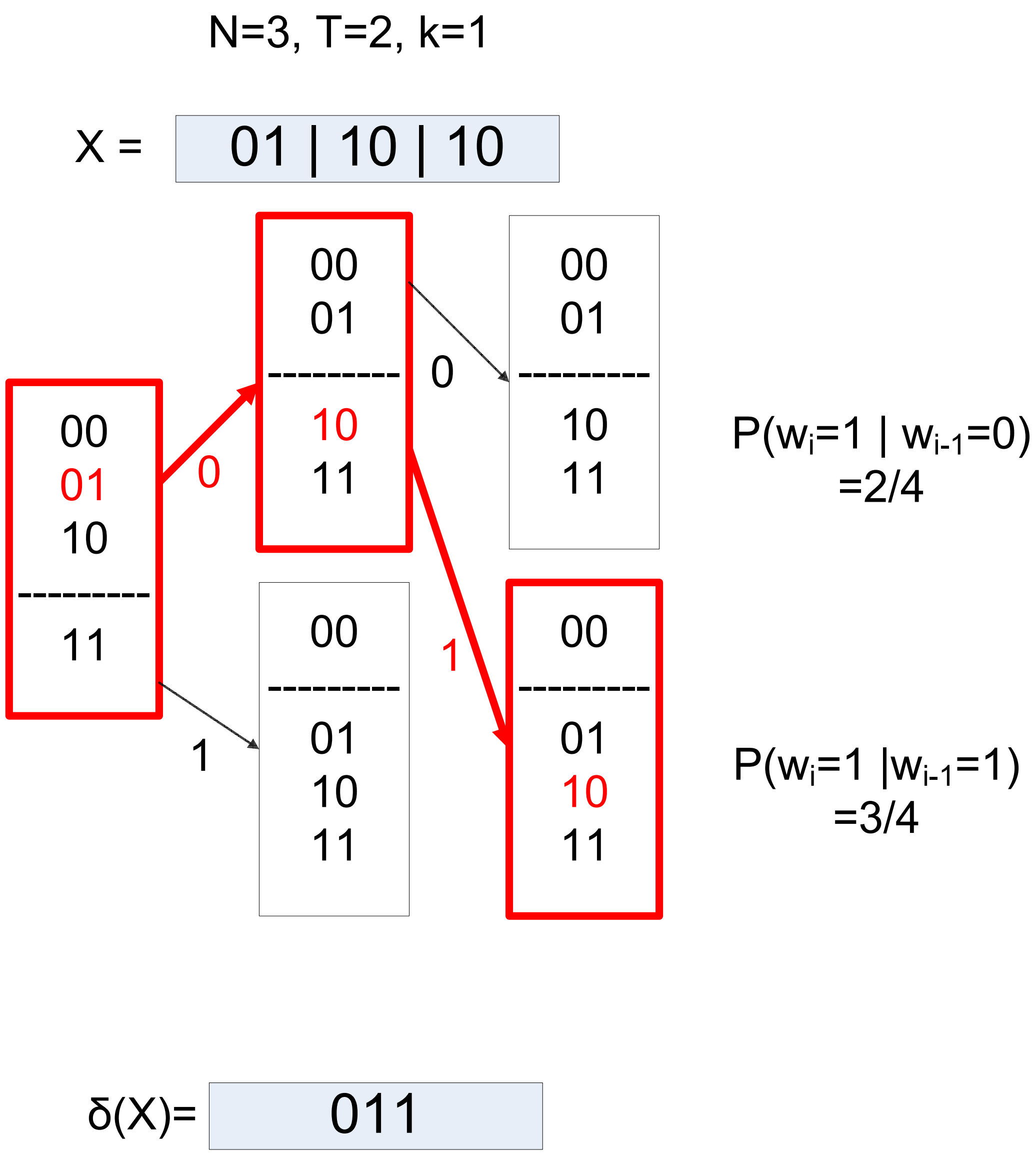

Figure 2.2: An example of Markov quantization ($k = 1$).

an MQ-k-P is i.u.d., then the quantizer performs quantization in a way that induces the pmf P on the transmitted sequence. Note that MQ-k-P does not exist for some pmfs, since there is a granularity of 2^{-T} in the quantization process.

In the following lemma we establish that if a codeword drawn from a coset ensemble is the input to an MQ-k-P, then the induced pmf on the transmitted sequence $\mathbf{w}$ is indeed P.

Lemma 2.10. *Consider a length-NT coset ensemble $\mathcal{C}'$. A code C' is picked from the ensemble and a codeword $\mathbf{c}' \in C'$ is picked uniformly from this code. Let P be a pmf of an arbitrary kth order Markov process for a binary sequence of length N for which an MQ-k-P exists. The codeword $\mathbf{c}'$ is quantized into $\mathbf{w}^N = \delta(\mathbf{c}')$ where $\delta(\cdot)$ is an MQ-k-P. Let $\hat{P}$ be the induced pmf on the transmitted sequence $\mathbf{w}^N$. Then, $P = \hat{P}$.*

Proof. Since $\mathcal{C}'$ is generated by adding a uniformly selected random vector to all codewords of all codes of $\mathcal{C}$ (the original ensemble from which $\mathcal{C}'$ is generated), we have

$$Pr[\mathbf{c}' = \mathbf{a}^{NT}] = 2^{-NT} \qquad \text{for all } \mathbf{a}^{NT} \tag{2.26a}$$

$$Pr[\mathbf{c}'^{iT}_{(i-1)T+1} = \mathbf{a}^T | \mathbf{w}^{i-1}_{i-j} = \mathbf{b}^T] = 2^{-T}$$
$$\text{for all } \mathbf{a}^T, \mathbf{b}^T, \text{ and } 1 \leq j \leq i \leq N. \tag{2.26b}$$

Therefore,

$$\hat{P}(\mathbf{w}^N) = \prod_{i=1}^{N} \hat{P}(w_i|\mathbf{w}^{i-1}) \tag{2.27a}$$

$$= \prod_{i=1}^{N} \sum_{\mathbf{a}^T} Pr[w_i, \mathbf{c}'^{iT}_{(i-1)T+1} = \mathbf{a}^T|\mathbf{w}^{i-1}] \tag{2.27b}$$

$$= \prod_{i=1}^{N} \sum_{\mathbf{a}^T} Pr[w_i|\mathbf{c}'^{iT}_{(i-1)T+1} = \mathbf{a}^T, \mathbf{w}^{i-1}] Pr[\mathbf{c}'^{iT}_{(i-1)T+1} = \mathbf{a}^T|\mathbf{w}^{i-1}] \tag{2.27c}$$

$$= \prod_{i=1}^{N} \sum_{\mathbf{a}^T} Pr[w_i|\mathbf{c}'^{iT}_{(i-1)T+1} = \mathbf{a}^T, \mathbf{w}^{i-1}] 2^{-T} \tag{2.27d}$$

$$= \prod_{i=1}^{N} \sum_{\mathbf{a}^T} Pr[w_i|\mathbf{c}'^{iT}_{(i-1)T+1} = \mathbf{a}^T, \mathbf{w}^{i-1}_{i-k}] 2^{-T} \tag{2.27e}$$

$$= \prod_{i=1}^{N} \sum_{\mathbf{a}^T} I\left\{ f(\mathbf{a}^T, \mathbf{w}^{i-1}_{i-k}) = w_i \right\} 2^{-T} \tag{2.27f}$$

$$= \prod_{i=1}^{N} P(w_i|\mathbf{w}^{i-1}_{i-k}) \tag{2.27g}$$

$$= P(\mathbf{w}^N), \tag{2.27h}$$

where (2.27d) is due to (2.26b), (2.27e) and (2.27f) is due to the definition of a quantizer (Definition 2.9), and (2.27g) is due to (2.25). Here $I\{\cdot\}$ denotes the indicator function of its argument. □

2.3.2 Analysis of Markov-quantized coset code ensembles

In this section, we derive an upper-bound on the error probability of quantized linear coset codes over FSCs. First, we establish an upper-bound on the pairwise error probability between two sequences which is similar to Lemma 2.5.

Definition 2.11. An MQ-k is called "robust" if the all-zeros block of length T, $\mathbf{0}^T$, and the all-ones block of length T, $\mathbf{1}^T$, are quantized to different values regardless of the quantizer memory, i.e.,

$$\forall \mathbf{w}^k \in \{0,1\}^k : \quad f(\mathbf{0}^T, \mathbf{w}^k) \neq f(\mathbf{1}^T, \mathbf{w}^k). \tag{2.28}$$

This property will be used below to establish a pairwise error probability upper bound between two codewords having Hamming distance NT before quantization.

Lemma 2.12. *Consider a code C. Let $\mathcal{C}'$ be an ensemble which consists of codes of the form $C' = \{\mathbf{c} \oplus \mathbf{v} | \mathbf{c} \in C\}$ for all uniformly selected vectors $\mathbf{v} \in \{0,1\}^{NT}$. A code $C' \in \mathcal{C}'$ is quantized using a "robust" MQ-k-P before transmission. Then, the ensemble-averaged pairwise ML decoding error probability $\overline{P}_{e|m,m'}$ of decoding message m' when m is transmitted on FSC when $d(\mathbf{c}_m, \mathbf{c}_{m'}) = NT$ is upper-bounded as*

$$\overline{P}_{e|m,m'} \leq D_1^N, \tag{2.29}$$

with

$$D_1 \triangleq \frac{2^T - 1 + D}{2^T}, \tag{2.30}$$

and D as defined in Lemma 2.5.

Furthermore, $D_1 < 1$.

Proof. See 2.6. □

It is noted that the above Lemma is more specialized than Lemma 2.5 in that it only establishes a bound for a pair of maximally separated codewords. As shown below, this is sufficient for our purpose of establishing capacity-achievability for the quantized ensembles A, B, C defined earlier. However, it might not be sufficient to prove capacity achievability for other quantized ensembles. As it turns out, this bound can be generalized to non-maximally separated codewords. However, due to space limitations we do not present this more general result.

We now state an error probability upper bound for Markov-quantized ensembles.

Proposition 2.13. *Consider a permutation-invariant ensemble $\mathcal{C}$ of binary linear codes with M codewords of length NT, rate R/T, and average weight enumerator $\overline{A}_l$. Let $\mathcal{C}'$ be the coset ensemble generated by $\mathcal{C}$. A code from $\mathcal{C}'$ is quantized using a "robust" MQ-k-P, and transmitted over an FSC. Let $U \subseteq \{1, 2, ..., NT\}$ and $NT \in U$. Then, for any $\epsilon > 0$, the average error probability with ML decoding given the $m$the message is transmitted, is upper-bounded as*

$$\overline{P}_{e|m} \leq \sum_{l \in U \backslash \{NT\}} \overline{A}_l + \overline{A}_{NT} D_1^N + Pr[\mathbf{X}^N \mathbf{Y}^N \notin T_N] + 2^{-N(C_P - R - \frac{\log_2 \alpha}{N} - \epsilon)}, \qquad (2.31)$$

where

$$\alpha = \max_{l \in U^c} \frac{\overline{A}_l}{M-1} \frac{2^{NT}}{\binom{NT}{l}} \qquad (2.32)$$

Proof. See 2.7. □

The following corollary proves that the coset ensembles generated by the ensembles A, B and C mentioned in Section 2.2 in conjunction with an MQ-k-P achieves C_P by using Theorem 2.13.

Corollary 2.14. *Consider a sequence of coset ensembles generated by the ensembles A, B, and C mentioned in Section 2.2 with length NT. For a given pmf P of a stationary ergodic Markov process, a "robust" MQ-k-P is used to quantize these ensembles before transmission. Then, for any $\epsilon > 0$, there exists a sequence of quantized coset ensembles generated by A, B, and C, have limiting (with respect to N) rate $(1-\epsilon)C_P$ and vanishing average block error probability under ML decoding on FSCs when conditions stated in Fact 2.3 are satisfied.*

Proof. The form of the bound in Theorem 2.13 is slightly different from that of the bound in Theorem 2.7. The union bound part with Battacharrya-like parameter in

Theorem 2.13 is established only for $l = NT$ instead of being established for the whole set U as in Theorem 2.7. However, this is not problematic, since $\sum_{l \in U \setminus \{NT\}} \overline{A}_l$ approaches 0 as $N \to \infty$ for all three ensembles. Hence we can proceed again as in the proof of Fact 2.3. $\square$

Since a sequence of stationary ergodic Markov sources asymptotically achieves the capacity of FSCs as the order k goes to infinity as in [35], a sequence of quantized coset code ensembles asymptotically achieves the capacity of FSCs for a large enough T. We note however, that [35] does not provide any useful bounds on how fast (with respect to k) a k-th order Markov process approaches the capacity of a FSC, and thus, we cannot make more accurate predictions about the requited order of the MQ-k-P.

As mentioned in the previous Section, if we concentrate on ensemble C, neither the symmetric assumption nor the non-inverting assumption about the channel is required to prove capacity achievability, since these two assumptions only enter our arguments through Lemma 2.12 which is not needed when proving capacity achievability for ensemble C. Thus, for a large enough k and T, the quantized coset ensemble generated by ensemble C can achieve the capacity of any binary-input FSC.

2.4 Proof of Theorem 2.1

For a specific code, let $P_{e|m}^{U^c}$ denote the probability that there exists some codeword $\mathbf{c}_{m'}$ such that $Q(\mathbf{y}|\mathbf{c}_{m'}) \geq Q(\mathbf{y}|\mathbf{c}_m)$ and $d(\mathbf{c}_m, \mathbf{c}_{m'}) \in U^c$. Let $\overline{P}_{e|m}^{U^c}$ denote the corresponding ensemble averaged probabilities. Then,

$$\overline{P}_{e|m} = \overline{P}_{e|m}^{U} + \overline{P}_{e|m}^{U^c}. \tag{2.33}$$

Consider now a new ensemble of codes generated by removing all codewords which satisfy $d(\mathbf{c}_m, \mathbf{c}_{m'}) \in U$. Let $\overline{P}'_{e|m}$ be the average error probability of the new ensemble when the mth codeword of the original ensemble is transmitted. Then, $\overline{P}^{U^c}_{e|m} \leq \overline{P}'_{e|m}$.

From now on, we will upper-bound $\overline{P}'_{e|m}$. Consider a decoder which declares that the ith message is transmitted if there exists a *unique* $\mathbf{c}_i$ satisfying

$$(\mathbf{c}_i, \mathbf{y}^N) \in T_N. \tag{2.34}$$

Otherwise, it declares an error. Then, the probability of error of the ML decoder is upper-bounded by the probability of error of this decoder. Let

$$E_i = \left\{(\mathbf{c}_i, \mathbf{y}^N) \in T_N\right\}. \tag{2.35}$$

Then,

$$P'_{e|m} \leq Pr\left[E_m^C \cup (\cup_{m' \neq m} E_{m'})\right] \tag{2.36a}$$

$$\leq Pr\left[E_m^C\right] + \sum_{m' \neq m : d(\mathbf{c}_m, \mathbf{c}_{m'}) \in U^c} Pr\left[E_{m'}\right]. \tag{2.36b}$$

where $\mathbf{y}^N$ is interpreted as an output corresponding to $\mathbf{c}_m$. We have

$$Pr[E_m^C] = \sum_{\mathbf{y}^N} Pr[(\mathbf{c}_m, \mathbf{y}^N) \notin T_N], \tag{2.37}$$

and for $m' \neq m$,

$$Pr[E_{m'}] = \sum_{\mathbf{y}^N} Pr[(\mathbf{c}_{m'}, \mathbf{y}^N) \in T_N]. \tag{2.38}$$

Note that

$$\overline{P}'_{e|m} \leq \sum_C Pr[C] Pr\left[E_m^C\right] + \sum_C Pr[C] \sum_{m' \neq m : d(\mathbf{c}_m, \mathbf{c}_{m'}) \in U^c} Pr\left[E_{m'}\right]. \tag{2.39}$$

We have

$$\sum_{C} Pr[C] Pr\left[E_m^C\right] = \sum_{\mathbf{y}^N}\sum_{\mathbf{x}^N} Pr[\mathbf{c}_m = \mathbf{x}^N] Pr[(\mathbf{x}^N, \mathbf{y}^N) \notin T_N] \tag{2.40a}$$

$$= \sum_{\mathbf{y}^N}\sum_{\mathbf{x}^N} P(\mathbf{x}^N) Pr[(\mathbf{x}^N, \mathbf{y}^N) \notin T_N] \tag{2.40b}$$

$$= Pr[\mathbf{X}^N \mathbf{Y}^N \notin T_N], \tag{2.40c}$$

and

$$\sum_{C} Pr[C] \sum_{m' \neq m : d(\mathbf{c}_m, \mathbf{c}_{m'}) \in U^c} Pr\left[E_{m'}\right] \tag{2.41a}$$

$$= \sum_{m' \neq m} \sum_{\mathbf{x}^N, \mathbf{x}'^N : d(\mathbf{x}^N, \mathbf{x}'^N) \in U^c} Pr[\mathbf{c}_m = \mathbf{x}^N] Pr[\mathbf{c}_{m'} = \mathbf{x}'^N | \mathbf{c}_m = \mathbf{x}^N] \tag{2.41b}$$

$$\times \sum_{\mathbf{y}^N} Pr[(\mathbf{x}'^N, \mathbf{y}^N) \in T_N)]$$

$$\leq \sum_{m' \neq m} \sum_{\mathbf{y}^N} \sum_{\mathbf{x}^N, \mathbf{x}'^N : d(\mathbf{x}^N, \mathbf{x}'^N) \in U^c} P(\mathbf{x}^N) \alpha P(\mathbf{x}'^N) Pr[(\mathbf{x}'^N, \mathbf{y}^N) \in T_N)] \tag{2.41c}$$

$$\leq \alpha \sum_{m' \neq m} \sum_{\mathbf{y}^N} \sum_{\mathbf{x}^N, \mathbf{x}'^N} P(\mathbf{x}^N) P(\mathbf{x}'^N) Pr[(\mathbf{x}'^N, \mathbf{y}^N) \in T_N)] \tag{2.41d}$$

$$\leq \alpha \sum_{m' \neq m} \sum_{\mathbf{x}^N, (\mathbf{x}'^N, \mathbf{y}^N) \in T_N} P(\mathbf{x}^N) P(\mathbf{x}'^N) Q(\mathbf{y}^N | \mathbf{x}^N) \tag{2.41e}$$

$$= \alpha \sum_{m' \neq m} \sum_{\mathbf{x}^N, (\mathbf{x}'^N, \mathbf{y}^N) \in T_N} P(\mathbf{x}^N, \mathbf{y}^N) P(\mathbf{x}'^N) \tag{2.41f}$$

$$= \alpha \sum_{m' \neq m} \sum_{(\mathbf{x}'^N, \mathbf{y}^N) \in T_N} P(\mathbf{x}'^N) P(\mathbf{y}^N), \tag{2.41g}$$

Note that we have, for $(\mathbf{x}'^N, \mathbf{y}^N) \in T_N$,

$$P(\mathbf{y}^N) \leq Q(\mathbf{y}^N | \mathbf{x}'^N) 2^{-N(C_P - \epsilon)}. \tag{2.42}$$

Hence,

$$
\begin{aligned}
&\sum_{C} Pr[C] \sum_{m' \neq m : d(\mathbf{c}_m, \mathbf{c}_{m'}) \in U^c} Pr\left[E_{m'}\right] \\
&\quad \leq \alpha \sum_{m' \neq m} \sum_{(\mathbf{x}'^N, \mathbf{y}^N) \in T_N} P(\mathbf{x}'^N) Q(\mathbf{y}^N | \mathbf{x}'^N) 2^{-N(C_P - \epsilon)} && (2.43a) \\
&\quad \leq \alpha \sum_{m' \neq m} 2^{-N(C_P - \epsilon)}. && (2.43b)
\end{aligned}
$$

Consequently,

$$
\begin{aligned}
\overline{P}'_{e|m} &\leq Pr[\mathbf{X}^N \mathbf{Y}^N \notin T_N] + (M-1)\alpha 2^{-N(C_P - \epsilon)} && (2.44a) \\
&\leq Pr[\mathbf{X}^N \mathbf{Y}^N \notin T_N] + 2^{-N(C_P - R - \frac{\log_2 \alpha}{N} - \epsilon)}. && (2.44b)
\end{aligned}
$$

Finally,

$$
\overline{P}_{e|m} \leq \overline{P}^U_{e|m} + Pr[\mathbf{X}^N \mathbf{Y}^N \notin T_N] + 2^{-N(C_P - R - \frac{\log_2 \alpha}{N} - \epsilon)}. \tag{2.45}
$$

2.5 Proof of Lemma 2.5

Consider a decoder with the following properties. First, the decoder assumes that the FSC stays at the state s_0 for the whole transmission of $\mathbf{x}$. Second, it quantizes every received value y to z in the following way

$$
z = \begin{cases} 1, & \text{when } y > 0 \\ -1, & \text{when } y < 0 \\ \pm 1 \text{ w.p. } 1/2, & \text{when } y = 0. \end{cases} \tag{2.46}
$$

Third, it decides that $\mathbf{x}^N$ is transmitted instead of $\mathbf{x}'^N$ if and only if $P(\mathbf{z}^N | \mathbf{x}^N, \mathbf{s}_0^N) \geq P(\mathbf{z}^N | \mathbf{x}'^N, \mathbf{s}_0^N)$, where $\mathbf{s}_0^N = (s_0, s_0, \ldots, s_0)$. Then, the pairwise error probability between $\mathbf{x}^N$ and $\mathbf{x}'^N$ with ML decoding is no greater than that with this decoder.

Therefore,

$$
\begin{aligned}
P_{e|\mathbf{x}^N,\mathbf{x}'^N} &\le \sum_{\mathbf{z}^N} P(\mathbf{z}^N|\mathbf{x}^N) I\left\{P(\mathbf{z}^N|\mathbf{x}^N,\mathbf{s}_0^N) < P(\mathbf{z}^N|\mathbf{x}'^N,\mathbf{s}_0^N)\right\} && \text{(2.47a)}\\
&\le \sum_{\mathbf{z}^N} P(\mathbf{z}^N|\mathbf{x}^N)\sqrt{\frac{P(\mathbf{z}^N|\mathbf{x}'^N,\mathbf{s}_0^N)}{P(\mathbf{z}^N|\mathbf{x}^N,\mathbf{s}_0^N)}} && \text{(2.47b)}\\
&= \sum_{\mathbf{z}^N}\sum_{\mathbf{s}^N} P(\mathbf{s}^N|\mathbf{x}^N)\prod_{i=1}^{N} P(z_i|x_i,s_i)\sqrt{\frac{P(z_i|x'_i,s_0)}{P(z_i|x_i,s_0)}} && \text{(2.47c)}\\
&= \sum_{\mathbf{s}^N} P(\mathbf{s}^N|\mathbf{x}^N)\prod_{i=1}^{N}\left\{\sum_{z} P(z|x_i,s_i)\sqrt{\frac{P(z|x'_i,s_0)}{P(z|x_i,s_0)}}\right\} && \text{(2.47d)}\\
&= \sum_{\mathbf{s}^N} P(\mathbf{s}^N|\mathbf{x}^N)\prod_{i:x_i\neq x'_i}\left\{\sum_{z} P(z|x_i,s_i)\sqrt{\frac{P(z|x'_i,s_0)}{P(z|x_i,s_0)}}\right\} && \text{(2.47e)}\\
&= \sum_{\mathbf{s}^N} P(\mathbf{s}^N|\mathbf{x}^N)\times\prod_{i:x_i=1}\left\{\sum_{z} P(z|1,s_i)\sqrt{\frac{P(z|0,s_0)}{P(z|1,s_0)}}\right\} && \text{(2.47f)}\\
&\qquad\times\prod_{i:x_i=0}\left\{\sum_{z} P(z|0,s_i)\sqrt{\frac{P(z|1,s_0)}{P(z|0,s_0)}}\right\}\\
&\overset{(a)}{=} \sum_{\mathbf{s}^N} P(\mathbf{s}^N|\mathbf{x}^N)\prod_{i:x_i\neq x'_i}\left\{\sum_{z} P(z|1,s_i)\sqrt{\frac{P(z|0,s_0)}{P(z|1,s_0)}}\right\} && \text{(2.47g)}\\
&\le \sum_{\mathbf{s}^N} P(\mathbf{s}^N|\mathbf{x}^N)\prod_{i:x_i\neq x'_i}\max_{s}\left\{\sum_{z} P(z|1,s)\sqrt{\frac{P(z|0,s_0)}{P(z|1,s_0)}}\right\} && \text{(2.47h)}\\
&= \left[\max_{s}\left\{\sum_{z} P(z|1,s)\sqrt{\frac{P(z|0,s_0)}{P(z|1,s_0)}}\right\}\right]^{d(\mathbf{x}^N,\mathbf{x}'^N)}, && \text{(2.47i)}
\end{aligned}
$$

where $I\{\cdot\}$ denotes the indicator function of its argument, and the equality in (a) is due to the fact that the channel at each state is symmetric. Since the above is true for any choice of the hypothesized state s_0, we have

$$\begin{aligned} & P_{e|\mathbf{x}^N,\mathbf{x}'^N} \\ & \leq \left[\min_{s_0}\max_{s}\left\{\sum_z P(z|1,s)\sqrt{\frac{P(z|0,s_0)}{P(z|1,s_0)}}\right\}\right]^{d(\mathbf{x}^N,\mathbf{x}'^N)} && \text{(2.48a)} \\ & = D^{d(\mathbf{x}^N,\mathbf{x}'^N)}. && \text{(2.48b)} \end{aligned}$$

In order to show that $D < 1$ we argue as follows. Since the channel at each state is non-inverting, we have

$$\sqrt{\frac{Q^+_{s_0}}{Q^-_{s_0}}} > 1 > \sqrt{\frac{Q^-_{s_0}}{Q^+_{s_0}}}. \tag{2.49}$$

Hence, for any s_0 we have

$$\begin{aligned} s^* & \triangleq \arg\max_s \left[Q^+_s\sqrt{\frac{Q^-_{s_0}}{Q^+_{s_0}}} + Q^-_s\sqrt{\frac{Q^+_{s_0}}{Q^-_{s_0}}}\right] && \text{(2.50a)} \\ & = \arg\max_s Q^-_s, && \text{(2.50b)} \end{aligned}$$

which means that the maximizer of (2.50) is independent of s_0. Thus choosing $s_0 = s^*$ we have

$$D \leq Q^+_{s^*}\sqrt{\frac{Q^-_{s^*}}{Q^+_{s^*}}} + Q^-_{s^*}\sqrt{\frac{Q^+_{s^*}}{Q^-_{s^*}}} = 2\sqrt{Q^+_{s^*}Q^-_{s^*}} < 1. \tag{2.51a}$$

2.6 Proof of Lemma 2.12

Consider a decoder with the following properties. First, the decoder assumes that the FSC stays at the state s_0 for the whole codeword transmission. Second, it quantizes every received value y to z according to (2.46). Third, it decides that message m is transmitted instead of m' if and only if $P(\mathbf{z}^N|\delta(\mathbf{c}'_m), \mathbf{s}_0^N) \geq P(\mathbf{z}^N|\delta(\mathbf{c}'_{m'}), \mathbf{s}_0^N)$, where $\mathbf{c}'_m,\ \mathbf{c}'_{m'} \in C'$, $\delta(\cdot)$ is the Markov quantizer mapping, and $\mathbf{s}_0^N = (s_0, s_0, \ldots, s_0)$. Then, the pairwise error probability between m and m' with ML decoding is no greater than that with this decoder. For a binary sequence $\mathbf{x}^{NT}$, let

$$\delta(\mathbf{x}^{NT}) \triangleq \left(\delta(\mathbf{x}^{NT})_1, \delta(\mathbf{x}^{NT})_2, ..., \delta(\mathbf{x}^{NT})_N\right). \tag{2.52}$$

Then,

$$\begin{aligned}
P_{e|m,m'} &\leq \sum_{\mathbf{z}^N} P(\mathbf{z}^N|\delta(\mathbf{c}'_m)) I\left\{P(\mathbf{z}^N|\delta(\mathbf{c}'_m), \mathbf{s}_0^N) < P(\mathbf{z}^N|\delta(\mathbf{c}'_{m'}), \mathbf{s}_0^N)\right\} && \text{(2.53a)}\\
&\leq \sum_{\mathbf{z}^N} P(\mathbf{z}^N|\delta(\mathbf{c}'_m)) \sqrt{\frac{P(\mathbf{z}^N|\delta(\mathbf{c}'_{m'}), \mathbf{s}_0^N)}{P(\mathbf{z}^N|\delta(\mathbf{c}'_m), \mathbf{s}_0^N)}} && \text{(2.53b)}\\
&= \sum_{\mathbf{z}^N} \sum_{\mathbf{s}^N} P(\mathbf{s}^N|\delta(\mathbf{c}'_m)) \prod_{i=1}^{N} \left\{P(z_i|\delta(\mathbf{c}'_m)_i, s_i) \sqrt{\frac{P(z_i|\delta(\mathbf{c}'_{m'})_i, s_0)}{P(z_i|\delta(\mathbf{c}'_m)_i, s_0)}}\right\} && \text{(2.53c)}\\
&\leq \prod_{i=1}^{N} \max_s \left\{\sum_z P(z|\delta(\mathbf{c}'_m)_i, s) \sqrt{\frac{P(z|\delta(\mathbf{c}'_{m'})_i, s_0)}{P(z|\delta(\mathbf{c}'_m)_i, s_0)}}\right\}. && \text{(2.53d)}
\end{aligned}$$

We can now average over the ensemble $\mathcal{C}'$, which is equivalent to averaging over all possible translation vectors, as follows

$$
\begin{aligned}
&\overline{P}_{e|m,m'} \\
&\le \mathbf{E}_{\mathbf{v}}\left[\prod_{i=1}^{N} \max_{s}\left\{\sum_{z} P(z|\delta(\mathbf{c}_m \oplus \mathbf{v})_i, s)\sqrt{\frac{P(z|\delta(\mathbf{c}_{m'} \oplus \mathbf{v})_i, s_0)}{P(z|\delta(\mathbf{c}_m \oplus \mathbf{v})_i, s_0)}}\right\}\right] && \text{(2.54a)} \\
&= \sum_{\mathbf{v}} \frac{1}{2^{NT}} && \text{(2.54b)} \\
&\quad \times \prod_{i=1}^{N} \max_{s}\left\{\sum_{z} P(z|\delta(\mathbf{c}_{m'} \oplus \mathbf{c}_m \oplus \mathbf{v})_i, s)\sqrt{\frac{P(z|\delta(\mathbf{v})_i, s_0)}{P(z|\delta(\mathbf{c}_{m'} \oplus \mathbf{c}_m \oplus \mathbf{v})_i, s_0)}}\right\} \\
&\le \sum_{\mathbf{v}} \frac{1}{2^{NT}} \prod_{i=1}^{N} \max_{s}\left\{\sum_{z} P(z|\delta(\mathbf{1}^{NT} \oplus \mathbf{v})_i, s)\sqrt{\frac{P(z|\delta(\mathbf{v})_i, s_0)}{P(z|\delta(\mathbf{1}^{NT} \oplus \mathbf{v})_i, s_0)}}\right\} && \text{(2.54c)} \\
&\le \sum_{\mathbf{v}} \frac{1}{2^{NT}} && \text{(2.54d)} \\
&\quad \times \prod_{i=1,\delta(\mathbf{v})_i \ne \delta(\mathbf{1}^{NT} \oplus \mathbf{v})_i}^{N} \max_{s}\left\{\sum_{z} P(z|\delta(\mathbf{1}^{NT} \oplus \mathbf{v})_i, s)\sqrt{\frac{P(z|\delta(\mathbf{v})_i, s_0)}{P(z|\delta(\mathbf{1}^{NT} \oplus \mathbf{v})_i, s_0)}}\right\} \\
&\le \sum_{\mathbf{v}} \frac{1}{2^{NT}} \prod_{i=1,\delta(\mathbf{v})_i \ne \delta(\mathbf{1}^{NT} \oplus \mathbf{v})_i}^{N} \max_{s}\left\{\sum_{z} P(z|0, s)\sqrt{\frac{P(z|1, s_0)}{P(z|0, s_0)}}\right\} && \text{(2.54e)}
\end{aligned}
$$

where (2.54b) is due to a change of variables in the summation over $\mathbf{v}$, (2.54c) is due to the fact that the two codewords in consideration are distance NT apart, (2.54e) is due to the state conditioned channel symmetry. Since the above is true for any choice of the hypothesized state s_0. we have

$$
\begin{aligned}
\overline{P}_{e|m,m'} &\le \sum_{\mathbf{v}} \frac{1}{2^{NT}} \prod_{i=1,\delta(\mathbf{v})_i \ne \delta(\mathbf{1}^{NT} \oplus \mathbf{v})_i}^{N} \min_{s_0} \max_{s}\left\{\sum_{z} P(z|0, s)\sqrt{\frac{P(z|1, s_0)}{P(z|0, s_0)}}\right\} && \text{(2.55a)} \\
&\le \sum_{\mathbf{v}} \frac{1}{2^{NT}} \prod_{i=1,\delta(\mathbf{v})_i \ne \delta(\mathbf{1}^{NT} \oplus \mathbf{v})_i}^{N} D && \text{(2.55b)}
\end{aligned}
$$

$$\leq \sum_{\mathbf{v}} \frac{1}{2^{NT}} \left[\prod_{\substack{i=1, \delta(\mathbf{v})_i \neq \delta(\mathbf{1}^{NT} \oplus \mathbf{v})_i \\ \mathbf{v}_{(i-1)T+1}^{iT} = \mathbf{0}^T}}^{N} D \right] \left[\prod_{\substack{i=1, \delta(\mathbf{v})_i \neq \delta(\mathbf{1}^{NT} \oplus \mathbf{v})_i, \\ \mathbf{v}_{(i-1)T+1}^{iT} \neq \mathbf{0}^T}}^{N} D \right]. \tag{2.55c}$$

Since $D < 1$ (due to Lemma 2.5), the last factor is upper bounded by one. Furthermore, since the quantizer is "robust", for every i for which $\mathbf{v}_{(i-1)T+1}^{iT} = \mathbf{0}^T$, it is also true that $\delta(\mathbf{1}^{NT} \oplus \mathbf{v})_i \neq \delta(\mathbf{v})_i$. Thus we can write

$$\overline{P}_{e|m,m'} \leq \sum_{\mathbf{v}} \frac{1}{2^{NT}} \prod_{i=1}^{N} D^{I\{\mathbf{v}_{(i-1)T+1}^{iT} = \mathbf{0}^T\}} \tag{2.56a}$$

$$= \left(\mathrm{E}\{ D^{I\{\mathbf{v}_{(i-1)T+1}^{iT} = \mathbf{0}^T\}} \} \right)^N \tag{2.56b}$$

$$= \left(\frac{2^T - 1}{2^T} + \frac{1}{2^T} D \right)^N \tag{2.56c}$$

$$= D_1^N, \tag{2.56d}$$

where $\mathrm{E}\{\cdot\}$ denotes expectation. Again, since $D < 1$, we deduce that $D_1 < 1$.

2.7 Proof of Theorem 2.13

We will use the bound in Theorem 2.1 which is stated as follows.

$$\overline{P}_{e|m} \leq \overline{P}_{e|m}^{U} + Pr[\mathbf{X}^N \mathbf{Y}^N \notin T_N] + 2^{-N(C_P - R - \frac{\log_2 \alpha}{N} - \epsilon)}, \tag{2.57}$$

First, we will bound $\overline{P}_{e|m}^{U}$ by using the union bound. Consider a code in $\mathcal{C}'$ which results from a code C in $\mathcal{C}$ having weight distribution A_l by adding a constant vector to all codewords. Let $P_{e|m,m'}$ be the pairwise error probability between messages m

and m'. Then,

$$P_{e|m}^{U} \leq \sum_{m' \neq m : d(\mathbf{c}'_m, \mathbf{c}'_{m'}) \in U} P_{e|m,m'} \tag{2.58a}$$

$$= \sum_{m' \neq m : d(\mathbf{c}'_m, \mathbf{c}'_{m'}) \in U \setminus \{NT\}} P_{e|m,m'} + \sum_{m' \neq m : d(\mathbf{c}'_m, \mathbf{c}'_{m'}) = NT} P_{e|m,m'} \tag{2.58b}$$

$$\leq \sum_{m' \neq m : d(\mathbf{c}'_m, \mathbf{c}'_{m'}) \in U \setminus \{NT\}} 1 + \sum_{m' \neq m : d(\mathbf{c}'_m, \mathbf{c}'_{m'}) = NT} P_{e|m,m'} \tag{2.58c}$$

$$= \sum_{l \in U \setminus \{NT\}} A_l + A_{NT} P_{e|m,m'}. \tag{2.58d}$$

Let $\tilde{P}_{e|m,C}^{U}$ be the average of $P_{e|m}^{U}$ over the coset ensemble generated by C. Then, from Lemma 2.12,

$$\tilde{P}_{e|m,C}^{U} \leq \sum_{l \in U \setminus \{NT\}} A_l + A_{NT} D_1^N. \tag{2.59}$$

Then, the average error probability (over the ensemble $\mathcal{C}'$) is

$$\overline{P}_{e|m}^{U} = \sum_{C \in \mathcal{C}} Pr(C) \tilde{P}_{e|m,C}^{U} \tag{2.60a}$$

$$\leq \sum_{l \in U \setminus \{NT\}} \overline{A}_l + \overline{A}_{NT} D_1^N. \tag{2.60b}$$

To apply Theorem 2.1, the code ensemble must satisfy (2.8) and (2.9). In [58], a code ensemble which satisfy (2.8) and (2.9) is generated from a certain linear code for MBIOS channel. In the following we point out how the derivation is different from the one in [58]. In [58, Lemma 1 and Th. 1], starting from an original code C, three ensembles are generated with increasing degree and randomness. The first ensemble $\mathcal{C}^1$ is generated by including $\sigma(C)$ in $\mathcal{C}^1$, for all $\sigma \in \mathcal{S}_M$, where $\sigma(C)$ denotes the code resulting by permuting the order of codewords of C according to σ. The second ensemble $\mathcal{C}^2$ is generated by including $\pi(C^1)$ in $\mathcal{C}^2$, for all $\pi \in \mathcal{S}_N$, and for

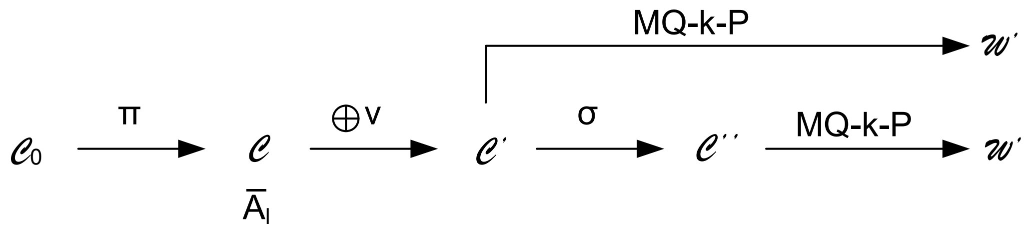

Figure 2.3: Relation between different code ensembles used in the proof of Theorem 2.13.

all $C^1 \in \mathcal{C}^1$, where $\pi(C^1)$ denotes the code resulting by permuting the order of the symbols of all codewords in C^1 according to π. The third ensemble $\mathcal{C}^3$ is generated by including codes of the form $\{\mathbf{c}^2 \oplus \mathbf{v} | \mathbf{c}^2 \in C^2\}$ in $\mathcal{C}^3$, for all $\mathbf{v} \in \{0,1\}^N$, and for all $C^2 \in \mathcal{C}^2$. Although in [58], $\mathcal{C}^3$ is generated starting from a specific code C, it is straightforward to see that the conclusions drawn in that paper still hold if $\mathcal{C}^3$ is generated starting from an ensemble of codes, with the only difference being that the average weight distribution over the original ensemble is used in place of the weight distribution of the original code. Furthermore, it is true that the order of the three operations which generate $\mathcal{C}^1$, $\mathcal{C}^2$, $\mathcal{C}^3$ is irrelevant. Therefore, if we apply σ to the coset permutation-invariant ensemble $\mathcal{C}'$ to generate a new ensemble $\mathcal{C}''$, then $\mathcal{C}''$ also satisfies

$$Pr[\mathbf{c}''_i = \mathbf{x}^{NT}] = 2^{-NT} \qquad \forall i \in \{1, \ldots, M\} \tag{2.61a}$$

$$Pr[\mathbf{c}''_i = \mathbf{x}^{NT} | \mathbf{c}''_j = \mathbf{x}'^{NT}] \leq \alpha Pr[\mathbf{c}''_i = \mathbf{x}^{NT}]$$
$$\forall i, j \in \{1, \ldots, M\} \text{ with } i \neq j \text{ and } d(\mathbf{c}''_i, \mathbf{c}''_j) \in U^c, \tag{2.61b}$$

with $\alpha = \max_{l \in U^c} \frac{\overline{A_l}}{M-1} \frac{2^{NT}}{\binom{NT}{l}}$ (recall that $\mathcal{C}'$ already includes all symbol-permuted and vector-translated codes. In other words, $\mathcal{C}''$ can be thought as generated from $\mathcal{C}_0$ by applying symbol permutation, vector translation, codeword permutation as depicted in Figure 2.3).

Now consider quantization of the ensemble $\mathcal{C}''$ that produces the Markov quantized ensemble $\mathcal{W}''$. We have,

$$Pr[\delta(\mathbf{c}''_i) = \mathbf{w}] = \sum_{\mathbf{x}:\delta(\mathbf{x})=\mathbf{w}} Pr[\mathbf{c}''_i = \mathbf{x}] \tag{2.62a}$$

$$\stackrel{(a)}{=} \sum_{\mathbf{x}:\delta(\mathbf{x})=\mathbf{w}} 2^{-NT} \tag{2.62b}$$

$$= P(\mathbf{w}), \tag{2.62c}$$

where the equality in (a) is due to (2.61a). Furthermore, for $\mathbf{c}''_i, \mathbf{c}''_j$ such that $d(\mathbf{c}''_i, \mathbf{c}''_j) \in U^c$,

$$Pr[\delta(\mathbf{c}''_i) = \mathbf{w} | \delta(\mathbf{c}''_j) = \mathbf{w}']$$

$$= \frac{Pr[\delta(\mathbf{c}''_i) = \mathbf{w}, \delta(\mathbf{c}''_j) = \mathbf{w}']}{Pr[\delta(\mathbf{c}''_j) = \mathbf{w}']} \tag{2.63a}$$

$$= \frac{\sum_{\mathbf{x}:\delta(\mathbf{x})=\mathbf{w}} \sum_{\mathbf{x}':\delta(\mathbf{x}')=\mathbf{w}'} Pr[\mathbf{c}''_i = \mathbf{x}, \mathbf{c}''_j = \mathbf{x}']}{\sum_{\mathbf{x}':\delta(\mathbf{x}')=\mathbf{w}'} Pr[\mathbf{c}''_j = \mathbf{x}']} \tag{2.63b}$$

$$\stackrel{(a)}{\leq} \frac{\sum_{\mathbf{x}:\delta(\mathbf{x})=\mathbf{w}} \sum_{\mathbf{x}':\delta(\mathbf{x}')=\mathbf{w}'} \alpha 2^{-NT} Pr[\mathbf{c}''_j = \mathbf{x}']}{\sum_{\mathbf{x}':\delta(\mathbf{x}')=\mathbf{w}'} Pr[\mathbf{c}''_j = \mathbf{x}']} \tag{2.63c}$$

$$= \sum_{\mathbf{x}:\delta(\mathbf{x})=\mathbf{w}} \alpha 2^{-NT} \tag{2.63d}$$

$$= \alpha P(\mathbf{w}), \tag{2.63e}$$

where the inequality in (a) is due to (2.61b). Note that the average error probability of the quantized ensemble $\mathcal{W}''$ of $\mathcal{C}''$ is the same as the average error probability of the quantized ensemble $\mathcal{W}'$ of the original coset permutation-invariant ensemble $\mathcal{C}'$ (see Fig. 2.3).

Even though (2.62) has a slightly different form than that of (2.9), we can still apply Theorem 2.1 for this quantized ensemble. Then, for any $\epsilon > 0$,

$$\overline{P}_{e|m} \leq \sum_{l \in U \setminus \{NT\}} \overline{A}_l + \overline{A}_{NT} D_1^N + Pr[\mathbf{X}^N \mathbf{Y}^N \notin T_N] + 2^{-N(C_P - R - \frac{\log_2 \alpha}{N} - \epsilon)}. \tag{2.64}$$

CHAPTER 3

A single-letter capacity expression for finite-state channels with feedback

3.1 Channel model and preliminaries

We consider channels with input $X_t \in \mathcal{X}$, output $Y_t \in \mathcal{Y}$ and state $S_t \in \mathcal{S}$ at time t. The corresponding input, output and state random processes are denoted by $(X_t)_{t=1}^{\infty}$, $(Y_t)_{t=1}^{\infty}$, $(S_t)_{t=1}^{\infty}$, respectively. Input, output and state alphabets are finite and of size $|\mathcal{X}| = K_x$, $|\mathcal{Y}| = K_y$, $|\mathcal{S}| = K_s$, respectively. At time t the receiver has access to the current channel output y_t and state s_t. The state s_t and output y_t are fed back to the transmitter with delay d. The state transition and the channel output stochastic kernel at time t are given as $Q(s_{t+1}|s_t, x_t)$ and $Q'(y_t|x_t, s_t)$, respectively. Note that channel state evolution is affected by both nature and ISI.

Definition 3.1. A sequence of joint measures $\{P(x^T, s^T, y^T)\}_{T=1}^{\infty}$ where v^T denotes the length-T vector $(v_1, \ldots, v_T)$ is *directed information stable* if

$$\lim_{T\to\infty} P\left(\left|\frac{\overrightarrow{i}(X^T; S^T, Y^T)}{I(X^T \to S^T, Y^T)} - 1\right| > \epsilon\right) = 0, \qquad \forall \epsilon > 0, \tag{3.1}$$

where $\overrightarrow{i}(X^T; S^T, Y^T) = \log \frac{P(x^T|s^T, y^T)}{\prod_{t=1}^{T} P(x^t|s^{t-1}, y^{t-1})}$ and $I(X^T \to S^T, Y^T) =$

$\sum_{t=1}^{T} I(X^t; S_t, Y_t | S^{t-1}, Y^{t-1})$.

Throughout the chapter we assume directed information stability.

In [54] the authors have developed a capacity expression for the general class of such channels with unit feedback delay (which includes the case in which only the channel output is fed back to the transmitter) in the form of

$$C = \sup_{\{\{P(x_t|x^{t-1}, s^{t-1}, y^{t-1})\}_{t=1}^{T}\}_{T=1}^{\infty}} \liminf_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} I(X^t; S_t, Y_t | S^{t-1}, Y^{t-1}). \tag{3.2}$$

This expression was further simplified in [54] to

$$\begin{aligned} C &= \sup_{\{\{P(x_t|\pi_t, \gamma_t, s_{t-1})\}_{t=1}^{T}\}_{T=1}^{\infty}} \liminf_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} I(X_t, \Pi_t; S_t, Y_t | S_{t-1}, \Gamma_t) && \text{(3.3a)} \\ &= \sup_{\{P(X|\Pi, \Gamma, S')\}_{\Pi, \Gamma, S'}} I(X, \Pi; S, Y | S', \Gamma) && \text{(3.3b)} \end{aligned}$$

where $\Pi_t \in \mathcal{P}(\mathcal{S})$ defined as $\Pi_t(s_t) \stackrel{\text{def}}{=} P(s_t | X^{t-1}, S^{t-1}, Y^{t-1})$, and $\Gamma_t \in \mathcal{P}(\mathcal{P}(\mathcal{S}))$ defined as $\Gamma_t(\pi_t) \stackrel{\text{def}}{=} P(\pi_t | S^{t-1}, Y^{t-1})$. In the above, $\mathcal{P}(\mathcal{S})$ is used to denote the set of probability measures on the set S.

Definition 3.2. Consider a Markov decision process specified by $(\mathcal{S}, \mathcal{A}, P, c)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, P is a stochastic kernel on $\mathcal{S}$ given $\mathcal{S} \times \mathcal{A}$, $c : \mathcal{S} \times \mathcal{A} \to \mathcal{R}$ is the instantaneous cost. An ACOE of this process is given by

$$\rho + h(i) = \min_{a \in \mathcal{A}} \left\{ c(i, a) + \sum_{j \in \mathcal{S}} P(j|i, a) h(j) \right\}, \tag{3.4}$$

where ρ is a scalar, $h : \mathcal{S} \to \mathcal{R}$. If a bounded solution (ρ, h) exists ρ is the minimum average cost.

It was further shown in [54] that the above capacity expression can in principle be evaluated as the solution of an appropriate ACOE [65, Th. 6.2., Th. 6.3.].

Clearly, the presence of the quantity Γ in (3.3a) renders this expression practically useless since a measure on the measures over a finite set needs to be considered. This complication is not surprising, since the above expressions were developed for general feedback patterns. Since in this work we are only interested in the special case where both state and output are fed back to the transmitter, it may be possible to get a significantly simpler capacity expression for such channels. We will pursue this direction in the following section, and we will show that a simpler capacity expression is possible even in the case of arbitrary finite feedback delay d.

3.2 A single-letter capacity expression for the FSC with arbitrary feedback delay

The expression given in (3.2) is valid when $d = 1$. In the following, we first derive a similar capacity expression for the FSC where the both channel output and state are fed back to the transmitter with arbitrary delay. This expression will serve as the starting point for the derivation of a simplified single-letter expression.

3.2.1 Capacity in the case of arbitrary feedback delay

Definition 3.3. A channel code-function is a sequence of T deterministic measurable maps $\{f_t\}_{t=1}^T$ such that $f_t : \mathcal{S}^{t-d} \times \mathcal{Y}^{t-d} \to \mathcal{X}$ which maps $(s^{t-d}, y^{t-d}) \to x_t$.

Suppose that code-functions $\{f_t\}_{t=1}^T$ are generated according to the measure

$P(f^T)$. Then, the induced joint measure on all relevant random variables is

$$
\begin{aligned}
&P(f^T, x^T, s^T, y^T) \\
&= \prod_{t=1}^{T} P(x_t|f_t, s^{t-d}, y^{t-d}) P(s_t, y_t|f^t, s^{t-1}, y^{t-1}) P(f_t|f^{t-1}, s^{t-1}, y^{t-1}) && \text{(3.5a)} \\
&= \prod_{t=1}^{T} \delta_{f_t(s^{t-d}, y^{t-d})}(x_t) Q'(y_t|f_t(s^{t-d}, y^{t-d}), s_t) \\
&\quad \times Q(s_t|s_{t-1}, f_{t-1}(s^{t-d-1}, y^{t-d-1})) P(f_t|f^{t-1}). && \text{(3.5b)}
\end{aligned}
$$

Lemma 3.4. *Given* $P(f^T)$ *and the induced joint measure* $P(f^T, x^T, s^T, y^T)$ *by* $P(f^T)$,

$$
I(F^T; S^T, Y^T) = I(X^T \to S^T, Y^T). \tag{3.6}
$$

Proof. For $x^t = f^t(s^{t-d}, y^{t-d})$

$$
\begin{aligned}
\frac{P(s^T, y^T|f^T)}{P(s^T, y^T)} &= \frac{\prod_{t=1}^{T} P(s_t, y_t|s^{t-1}, y^{t-1}, f^t)}{P(s^T, y^T)} && \text{(3.7a)} \\
&= \frac{\prod_{t=1}^{T} P(s_t, y_t|x^t, s^{t-1}, y^{t-1})}{P(s^T, y^T)} && \text{(3.7b)} \\
&= \frac{\prod_{t=1}^{T} P(s_t, y_t|x^t, s^{t-1}, y^{t-1})}{\prod_{t=1}^{T} P(s_t, y_t|s^{t-1}, y^{t-1})} && \text{(3.7c)} \\
&= \prod_{t=1}^{T} \frac{P(s_t, y_t|x^t, s^{t-1}, y^{t-1})}{P(s_t, y_t|s^{t-1}, y^{t-1})}. && \text{(3.7d)}
\end{aligned}
$$

After taking expectations on the logarithms of the above, we obtain the final result.

□

Following [54], we define the following quantities.

Definition 3.5. Let $graph(f_t) \overset{\text{def}}{=} \{(s^{t-d}, y^{t-d}, x_t) : f_t(s^{t-d}, y^{t-d}) = x_t\}$.

Definition 3.6. Let $\sigma(s^{t-d}, y^{t-d}, x_t) \overset{\text{def}}{=} \{f_t : (s^{t-d}, y^{t-d}, x_t) \in graph(f_t)\}$. We also denote $\sigma^t(s^{t-d}, y^{t-d}, x^t) \overset{\text{def}}{=} \{f^t : (s^{t'-d}, y^{t'-d}, x_{t'}) \in graph(f_{t'}),\ \text{for all } t' = 1, 2, \ldots, t\}$.

Lemma 3.7. *Given $P(f^T)$ and the induced joint measure $P(f^T, x^T, s^T, y^T)$, we have $P(x_t|x^{t-1}, s^{t-d}, y^{t-d}) = P_{f_t|f^{t-1}}(\sigma_t(s^{t-d}, y^{t-d}, x_t)|\sigma^{t-1}(s^{t-d-1}, y^{t-d-1}, x^{t-1}))$.*

Proof. We have

$$\begin{aligned} & P(x_t|x^{t-1}, s^{t-d}, y^{t-d}) \\ & \quad = P(x_t|\sigma^{t-1}(s^{t-d-1}, y^{t-d-1}, x^{t-1}), s^{t-d}, y^{t-d}) && \text{(3.8a)} \\ & \quad = P(\sigma_t(s^{t-d}, y^{t-d}, x_t)|\sigma^{t-1}(s^{t-d-1}, y^{t-d-1}, x^{t-1}), s^{t-d}, y^{t-d}) && \text{(3.8b)} \\ & \quad = P(\sigma_t(s^{t-d}, y^{t-d}, x_t)|\sigma^{t-1}(s^{t-d-1}, y^{t-d-1}, x^{t-1})). && \text{(3.8c)} \end{aligned}$$

□

Proposition 3.8. *The capacity of the FSC with arbitrary feedback delay is*

$$C_d = \sup_{\{\{P(x_t|x^{t-1}, s^{t-d}, y^{t-d})\}_{t=1}^T\}_{T=1}^\infty} \liminf_{T \to \infty} \frac{1}{T} \sum_{t=1}^T I(X^t; S_t, Y_t|S^{t-1}, Y^{t-1}). \tag{3.9}$$

Proof. What we want to show is

$$\begin{aligned} & \sup_{\{\{P(x_t|x^{t-1}, s^{t-d}, y^{t-d})\}_{t=1}^T\}_{T=1}^\infty} \liminf_{T \to \infty} \frac{1}{T} \sum_{t=1}^T I(X^t; S_t, Y_t|S^{t-1}, Y^{t-1}) \\ = & \sup_{\{\{P(f_t|f^{t-1})\}_{t=1}^T\}_{T=1}^\infty} \liminf_{T \to \infty} \frac{1}{T} I(F^T; S^T, Y^T). \end{aligned} \tag{3.10}$$

For a given $\{P(x_t|x^{t-1}, s^{t-d}, y^{t-d})\}_{t=1}^T$, we construct $P(f^T)$ as follows. For every t

$$P(f_t|f^{t-1}) = \prod_{(s^{t-d}, y^{t-d}, x_t) \in graph(f_t)} P(x_t|f^{t-1}(s^{t-d-1}, y^{t-d-1}), s^{t-d}, y^{t-d}). \tag{3.11}$$

Since we have Lemma 3.4, to show that (3.10) is true it only remains to show that the induced channel input distribution by $P(f^T)$ equals to $\{P(x_t|x^{t-1}, s^{t-d}, y^{t-d})\}_{t=1}^T$.

We have for every t

$$P_{f_t|f^{t-1}}(\sigma_t(s^{t-d}, y^{t-d}, x_t)|f^{t-1}) = P(x_t|f^{t-1}(s^{t-d-1}, y^{t-d-1}), s^{t-d}, y^{t-d}). \tag{3.12}$$

Consider now the induced channel input distribution $\{Q(x_t|x^{t-1}, s^{t-d}, y^{t-d})\}_{t=1}^T$ by $P(f^T)$. We have

$$\begin{aligned} Q(x_t|x^{t-1}, s^{t-d}, y^{t-d}) &= P(\sigma_t(s^{t-d}, y^{t-d}, x_t)|\sigma^{t-1}(s^{t-d-1}, y^{t-d-1}, x^{t-1})) && \text{(3.13a)} \\ &= P(x_t|x^{t-1}, s^{t-d}, y^{t-d}), && \text{(3.13b)} \end{aligned}$$

where the first equality follows from Lemma 3.7, and the second equality due to (3.12).

Hence (3.10) is true. Furthermore, by Lemmas 5.5, 5.6 and Theorems 5.2,5.3 in [54], the right-hand side of (3.10) is the capacity of this channel. □

We have found the capacity expression of the FSC with arbitrary feedback delay. In the next section, we will simplify this expression.

3.2.2 A simplified capacity expression

Consider the term $I(X^t; S_t, Y_t|S^{t-1}, Y^{t-1})$ with the channel input distribution $P(x_t|x^{t-1}, s^{t-d}, y^{t-d})$ in (3.9). The following theorem proves that the form of the optimal channel input distribution can be simplified.

Lemma 3.9. *For every T,*

$$\begin{aligned} &\sup_{\{P(x_t|x^{t-1}, s^{t-d}, y^{t-d})\}_{t=1}^T} \frac{1}{T} \sum_{t=1}^T I(X^t; S_t, Y_t|S^{t-1}, Y^{t-1}) \\ &= \sup_{\{P(x_t|x_{t-d}^{t-1}, s^{t-d}, y^{t-d})\}_{t=1}^T} \frac{1}{T} \sum_{t=1}^T I(X_{t-1}^t; S_t, Y_t|S^{t-1}, Y^{t-1}). \end{aligned} \tag{3.14}$$

Proof. See 3.4. □

The above lemma shows that in order to achieve capacity it is sufficient to restrict the channel input distributions to be of the form of $P(x_t|x_{t-d}^{t-1}, s^{t-d}, y^{t-d})$, i.e., the capacity expression becomes

$$C = \sup_{\{\{P(x_t|x_{t-d}^{t-1}, s^{t-d}, y^{t-d})\}_{t=1}^{T}\}_{T=1}^{\infty}} \liminf_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} I(X_{t-1}^{t}; S_t, Y_t | S^{t-1}, Y^{t-1}). \tag{3.15}$$

To further simplify the capacity expression, we will formulate a control problem which is equivalent to the problem of computing capacity.

Problem 3.10. Let $(X_{t-d}^{t-1}, S_{t-d}, Y_{t-d})$ be the system state at time t, and (S_{t-d}, Y_{t-d}) be the controller observation at time t. Let the control action at time t be $U_t : \mathcal{X}^d \to \mathcal{P}(\mathcal{X})$ defined as $U_t(x_t; x_{t-d}^{t-1}) \stackrel{\text{def}}{=} P(x_t|x_{t-d}^{t-1}, S^{t-d}, Y^{t-d})$. Further, define the instantaneous reward at time t to be $R_t = \log \frac{P(S_t, Y_t|S^{t-1}, Y^{t-1}, X_{t-1}^{t})}{P(S_t, Y_t|S^{t-1}, Y^{t-1})}$. The control problem is to determine the optimal policy $g = \{g_t\}_{t=1}^{\infty}$ (such that $u_t = g_t(s^{t-d}, y^{t-d})$) that maximizes the average expected reward $\liminf_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} E^g[R_t]$.

First we need to prove that the above control problem is equivalent to the problem of computing capacity as stated in (3.15). All that is required is to show that the sequence of measures $\{P(x_{t-d}^{t}, s^t, y^t)\}_{t=1}^{\infty}$ induced by the channel input distributions $\{P(x_t|x_{t-d}^{t-1}, s_{t-d}^{t}, y^{t-d})\}_{t=1}^{\infty}$ is equal to the sequence of measures $\{P^g(x_{t-d}^{t}, s^t, y^t)\}_{t=1}^{\infty}$ induced by the control policy g. This equivalence is established in the following lemma.

Lemma 3.11. *For every sequence of channel input distributions* $\{P(x_t|x_{t-d}^{t-1}, s^{t-d}, y^{t-d})\}_{t=1}^{\infty}$ *with resulting sequence of joint measures* $\{P(x_{t-d}^{t}, s^t, y^t)\}_{t=1}^{\infty}$ *there exists a policy* g *with resulting sequence of joint measures* $\{P^g(x_{t-d}^{t}, s^t, y^t)\}_{t=1}^{\infty}$ *such that for each* t*:* $P^g(x_{t-d}^{t}, s^t, y^t) = P(x_{t-d}^{t}, s^t, y^t)$.

Conversely, for every policy g with resulting sequence of joint measures $\{P^g(x_{t-d}^t, s^t, y^t)\}_{t=1}^{\infty}$ there exists a sequence of channel input distributions $\{P(x_t|x_{t-d}^{t-1}, s^{t-d}, y^{t-d})\}_{t=1}^{\infty}$ with resulting sequence of joint measures $\{P(x_{t-d}^t, s^t, y^t)\}_{t=1}^{\infty}$ such that for each t: $P(x_{t-d}^t, s^t, y^t) = P^g(x_{t-d}^t, s^t, y^t)$.

Proof. See 3.5. □

We are now ready to state and prove the main result of this chapter.

Proposition 3.12. *The capacity of the finite-state channel with output and state feedback with arbitrary delay is*

$$C_d = \sup_{\{P(X|X'^d, S_1', \Theta)\}_{X'^d, S_1', \Theta}} I(X, X_d'; S, Y | S'^d, Y'^{d-1}, \Theta) \tag{3.16}$$

where $\Theta \in \mathcal{P}(\mathcal{X}^d)$, and the mutual information is evaluated using the joint measure

$$\begin{aligned} &P(Y, S, X, X_d', S'^d, Y'^{d-1}, d\Theta) \\ &= \sum_{X'^{d-1}} Q'(Y|S, X) Q(S|S_d', X_d') \left(\prod_{i=2}^{d} Q'(Y_{i-1}'|S_i', X_i') Q(S_i'|S_{i-1}', X_{i-1}') \right) \\ &\quad \times P(X|X'^d, S_1', \Theta) \Theta(X'^d) P(S_1', d\Theta). \end{aligned} \tag{3.17}$$

The distribution $P(S, d\Theta)$ is the solution of the equation

$$\begin{aligned} &P(S, d\Theta') \\ &= \int_{S', \Theta} P(S', d\Theta) \sum_{Y, X, X'^d} \delta_{\omega(\Theta, P(X|X'^d, S', \Theta), Y, S, S')}(\Theta') \\ &\quad \times Q'(Y|X, S) Q(S|S', X_1') P(X|X'^d, S', \Theta) \Theta(X'^d), \end{aligned} \tag{3.18}$$

where the function $\omega(\cdot)$ is defined in the following proof.

Proof. See 3.6. □

To find the capacity using (3.16), we need to identify the stationary distribution of S and Θ for each choice of $P(X|X'^d, S'_1, \Theta)$ using (3.18), and evaluate the mutual information by using the joint measure specified in (3.17). Alternatively, we may use dynamic programming to find the capacity. In other words, the optimal value in (3.16) can be obtained by the solution of the following ACOE with some bounded function $\eta : \mathcal{S} \times \Theta \to \mathcal{R}$ [65, Th. 6.2., Th. 6.3.]

$$C_d + \eta(s, \theta) = \sup_u J(s, \theta, u), \tag{3.19}$$

where

$$\begin{aligned} & J(s, \theta, u) \\ & = \bar{r}(s, \theta, u) \\ & \quad + \left(\sum_{y', s'} \eta(s', \omega(\theta, u, y', s', s)) \sum_{x'} Q'(y'|s', x') \sum_{x^d} Q(s'|s, x_1) u(x'; x^d) \theta(x^d) \right), \end{aligned} \tag{3.20}$$

and $\bar{r}(s, \theta, u) = E[R_t | S_{t-d} = s, \Theta_{t-1} = \theta, U_t = u]$.

3.3 Special case: No ISI

So far we have considered the situation when there is ISI. Now let's consider the special case with no ISI. In this case, the state transition stochastic kernel at time t is given as $Q(s_{t+1}|s_t)$. Viswanathan found the capacity in this case in [52]. In the following we provide an alternative approach to showing this result using stochastic control.

Consider the term $I(X^t; S_t, Y_t|S^{t-1}, Y^{t-1})$ with the channel input distribution $P(x_t|x^{t-1}, s^{t-d}, y^{t-d})$ in (3.9). We can simplify the form of the optimal channel input distribution in similar ways to (3.14).

Lemma 3.13. *For every T,*

$$\begin{aligned} &\sup_{\{P(x_t|x^{t-1},s^{t-d},y^{t-d})\}_{t=1}^T} \frac{1}{T}\sum_{t=1}^{T} I(X^t; S_t, Y_t|S^{t-1}, Y^{t-1}) \\ &= \sup_{\{P(x_t|s^{t-d},y^{t-d})\}_{t=1}^T} \frac{1}{T}\sum_{t=1}^{T} I(X_t; Y_t|S^t, Y^{t-1}). \end{aligned} \tag{3.21}$$

Proof. See 3.7. □

The above lemma shows that in order to achieve capacity it is sufficient to restrict the channel input distributions to be the form of $P(x_t|s^{t-d}, y^{t-d})$, i.e., the capacity expression becomes

$$C = \sup_{\{\{P(x_t|s^{t-d},y^{t-d})\}_{t=1}^T\}_{T=1}^\infty} \liminf_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} I(X_t; Y_t|S^t, Y^{t-1}). \tag{3.22}$$

To further simplify the capacity expression, we will formulate a control problem which is equivalent to the problem of computing capacity.

Problem 3.14. Let (S_{t-d}, Y_{t-d}) be the system state at time t, and (S_{t-d}, Y_{t-d}) be the controller observation at time t. Let the control action at time t be $U_t \in \mathcal{P}(\mathcal{X})$ defined as $U_t(x_t) \stackrel{\text{def}}{=} P(x_t; S^{t-d}, Y^{t-d})$. Further, define the instantaneous reward at time t to be $R_t = \log \frac{P(Y_t|S^t,Y^{t-1},X_t)}{P(Y_t|S^t,Y^{t-1})}$. The control problem is to determine the optimal policy $g = \{g_t\}_{t=1}^\infty$ (such that $u_t = g_t(s^{t-d}, y^{t-d})$) that maximizes the average expected reward $\liminf_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} E^g[R_t]$.

In similar ways to Lemma 3.11, we can see that the above control problem is equivalent to the problem of computing capacity as stated in (3.22).

We are now ready to state and prove the main result which confirms the results in [52].

Proposition 3.15. *The capacity of the finite-state channel with output and state feedback with arbitrary delay is*

$$C_d = \sup_{\{P(X|S')\}_{S'}} I(X;Y|S,S') \tag{3.23}$$

where the mutual information is evaluated using the joint measure

$$\begin{aligned} &P(Y,S,S',X) \\ &\quad = Q'(Y|S,X) \sum_{S^{d-1}} Q(S|S_{d-1}) \prod_{t=2}^{d-1} Q(S_t|S_{t-1}) Q(S_1|S') P(X|S') P(S'). \end{aligned} \tag{3.24a}$$

The distribution $P(S)$ is the solution of the equation

$$P(S) = \sum_{S'} Q(S|S') P(S'). \tag{3.24b}$$

Proof. See 3.8. □

3.4 Proof of Lemma 3.9

First, note that we have for every t

$$\begin{aligned} &I(X^t;S_t,Y_t|S^{t-1},Y^{t-1}) \\ &\quad = I(X_{t-1}^t;S_t,Y_t|S^{t-1},Y^{t-1}) + I(X^{t-2};S_t,Y_t|S^{t-1},Y^{t-1},X_{t-1}^t) && (3.25a) \\ &\quad \overset{(a)}{=} I(X_{t-1}^t;S_t,Y_t|S^{t-1},Y^{t-1}), && (3.25b) \end{aligned}$$

where (a) is from the fact that S_t, Y_t is independent of X^{t-2} given S_{t-1}, X_{t-1}^t. Each of the terms $I(X_{t-1}^t; S_t, Y_t | S^{t-1}, Y^{t-1})$ in the summation is evaluated based on the joint distribution $P(x_{t-1}^t, s^t, y^t)$. We now proceed by induction to prove that the sequence of measures $\{P(x_{t-d}^t, s^t, y^t)\}_{t=1}^T$ induced by the sequence of channel input distributions $\{P(x_t|x^{t-1}, s^{t-d}, y^{t-d})\}_{t=1}^T$ equals to the sequence of measures $\{P_1(x_{t-d}^t, s^t, y^t)\}_{t=1}^T$ induced by an appropriately defined sequence of channel input distributions of the form $\{P_1(x_t|x_{t-d}^{t-1}, s^{t-d}, y^{t-d})\}_{t=1}^T$.

For $t = 1$ we set $P_1(x_1) = P(x)$ and have

$$P_1(x_1, s_1, y_1) = Q'(y_1|s_1, x_1)Q(s_1)P_1(x_1) \tag{3.26a}$$

$$= Q'(y_1|s_1, x_1)Q(s_1)P(x_1) = P(x_1, s_1, y_1). \tag{3.26b}$$

Now for $t+1$ we set $P_1(x_{t+1}|x_{t-d+1}^t, s^{t-d+1}, y^{t-d+1}) = P(x_{t+1}|x_{t-d+1}^t, s^{t-d+1}, y^{t-d+1})$ and have

$$\begin{aligned}
&P_1(x_{t-d+1}^{t+1}, s^{t+1}, y^{t+1}) \\
&= \left(\prod_{i=t-d+2}^{t+1} Q'(y_i|s_i, x_i)Q(s_i|s_{i-1}, x_{i-1})\right) P_1(x_{t+1}|x_{t-d+1}^t, s^{t-d+1}, y^{t-d+1}) \\
&\quad \times \sum_{x_{t-d}, s_{t-d+2}^t, y_{t-d+2}^t} P_1(x_{t-d}^t, s^t, y^t) && (3.27a) \\
&\overset{(a)}{=} \left(\prod_{i=t-d+2}^{t+1} Q'(y_i|s_i, x_i)Q(s_i|s_{i-1}, x_{i-1})\right) P(x_{t+1}|x_{t-d+1}^t, s^{t-d+1}, y^{t-d+1}) \\
&\quad \times \sum_{x_{t-d}, s_{t-d+2}^t, y_{t-d+2}^t} P(x_{t-d}^t, s^t, y^t) && (3.27b) \\
&= P(x_{t-d+1}^{t+1}, s^{t+1}, y^{t+1}), && (3.27c)
\end{aligned}$$

where (a) is due to the construction of $P_1(x_{t+1}|x_{t-d+1}^t, s^{t-d+1}, y^{t-d+1})$, and the induction hypothesis. The above equality implies that the equality in (3.14).

3.5 Proof of Lemma 3.11

We will use the notation $u_t = g_t(s^{t-d}, y^{t-d})$ and for convenience, we will write $u_t(x_t; x_{t-d}^{t-1}) = g_t[s^{t-d}, y^{t-d}](x_t; x_{t-d}^{t-1})$.

For the direct part, for each t we choose a policy g_t as

$$g_t[s^{t-d}, y^{t-d}](x_t; x_{t-d}^{t-1}) = P(x_t | x_{t-d}^{t-1}, s^{t-d}, y^{t-d}), \tag{3.28}$$

and proceed by induction.

For $t = 1$ we have

$$\begin{aligned} P^g(x_1, s_1, y_1) &= Q'(y_1|s_1, x_1)Q(s_1)g_1(x_1) &\text{(3.29a)}\\ &= Q'(y_1|s_1, x_1)Q(s_1)P(x_1) = P(x_1, s_1, y_1). &\text{(3.29b)}\end{aligned}$$

Now for $t + 1$ we have

$$\begin{aligned} &P^g(x_{t-d+1}^{t+1}, s^{t+1}, y^{t+1}) \\ &= \left(\prod_{i=t-d+2}^{t+1} Q'(y_i|s_i, x_i)Q(s_i|s_{i-1}, x_{i-1})\right) \\ &\quad \times g_{t+1}[s^{t-d+1}, y^{t-d+1}](x_{t+1}; x_{t-d+1}^{t}) \\ &\quad \times \sum_{x_{t-d}, s_{t-d+2}^{t}, y_{t-d+2}^{t}} P^g(x_{t-d}^{t}, s^t, y^t) &\text{(3.30a)}\\ &\overset{(a)}{=} \left(\prod_{i=t-d+2}^{t+1} Q'(y_i|s_i, x_i)Q(s_i|s_{i-1}, x_{i-1})\right) \\ &\quad \times P(x_{t+1}|x_{t-d+1}^{t}, s^{t-d+1}, y^{t-d+1}) \\ &\quad \times \sum_{x_{t-d}, s_{t-d+2}^{t}, y_{t-d+2}^{t}} P(x_{t-d}^{t}, s^t, y^t) &\text{(3.30b)}\\ &= P(x_{t-d+1}^{t+1}, s^{t+1}, y^{t+1}), &\text{(3.30c)}\end{aligned}$$

where (a) is due to the choice of the policy g_{t+1} and the induction hypothesis.

For the converse, for each t we choose a channel input distribution as

$$P(x_t|x_{t-d}^{t-1}, s^{t-d}, y^{t-d}) = P^g(x_t|x_{t-d}^{t-1}, s^{t-d}, y^{t-d}) \tag{3.31a}$$

$$= g_t[s^{t-d}, y^{t-d}](x_t; x_{t-d}^{t-1}). \tag{3.31b}$$

Then, for $t = 1$ we have

$$P(x_1, s_1, y_1) = Q'(y_1|s_1, x_1)Q(s_1)P(x_1) \tag{3.32a}$$

$$= Q'(y_1|s_1, x_1)Q(s_1)g_1(x_1) = P^g(x_1, s_1, y_1). \tag{3.32b}$$

Now for $t + 1$ we have

$$\begin{aligned}
&P(x_{t-d+1}^{t+1}, s^{t+1}, y^{t+1}) \\
&= \left(\prod_{i=t-d+2}^{t+1} Q'(y_i|s_i, x_i)Q(s_i|s_{i-1}, x_{i-1}) \right) \\
&\quad \times P(x_{t+1}|x_{t-d+1}^{t}, s^{t-d+1}, y^{t-d+1}) \\
&\quad \times \sum_{x_{t-d}, s_{t-d+2}^{t}, y_{t-d+2}^{t}} P(x_{t-d}^{t}, s^t, y^t) && (3.33a) \\
&\overset{(a)}{=} \left(\prod_{i=t-d+2}^{t+1} Q'(y_i|s_i, x_i)Q(s_i|s_{i-1}, x_{i-1}) \right) \\
&\quad \times g_{t+1}[s^{t-d+1}, y^{t-d+1}](x_{t+1}; x_{t-d+1}^{t}) \\
&\quad \times \sum_{x_{t-d}, s_{t-d+2}^{t}, y_{t-d+2}^{t}} P^g(x_{t-d}^{t}, s^t, y^t) && (3.33b) \\
&= P^g(x_{t-d+1}^{t+1}, s^{t+1}, y^{t+1}), && (3.33c)
\end{aligned}$$

where (a) is due to the construction of the channel input distributions and the induction hypothesis.

3.6 Proof of Proposition 3.12

Define the information state $\Theta_t \in \mathcal{P}(\mathcal{X}^d)$

with $\Theta_t(x_{t-d+1}^t) \overset{\text{def}}{=} P(x_{t-d+1}^t | S^{t-d+1}, Y^{t-d+1})$.

$$
\begin{aligned}
&\theta_t(x_{t-d+1}^t) \\
&\quad = P(x_{t-d+1}^t | S^{t-d+1}, Y^{t-d+1}) && \text{(3.34a)} \\
&\quad = \frac{P(x_{t-d+1}^t, s_{t-d+1}, y_{t-d+1} | s^{t-d}, y^{t-d})}{P(s_{t-d+1}, y_{t-d+1} | s^{t-d}, y^{t-d})} && \text{(3.34b)} \\
&\quad = \Bigg(\sum_{x_{t-d}} u_t(x_t; x_{t-d}^{t-1}) Q'(y_{t-d+1} | x_{t-d+1}, s_{t-d+1}) Q(s_{t-d+1} | s_{t-d}, x_{t-d}) \\
&\qquad\quad \times P(x_{t-d}^{t-1} | s^{t-d}, y^{t-d}) \Bigg) \Big/ \Bigg(P(s_{t-d+1}, y_{t-d+1} | s^{t-d}, y^{t-d}) \Bigg) && \text{(3.34c)} \\
&\quad = \Bigg(\sum_{x_{t-d}} u_t(x_t; x_{t-d}^{t-1}) Q'(y_{t-d+1} | x_{t-d+1}, s_{t-d+1}) Q(s_{t-d+1} | s_{t-d}, x_{t-d}) \\
&\qquad\quad \times \theta_{t-1}(x_{t-d}^{t-1}) \Bigg) \Big/ \Bigg(P(s_{t-d+1}, y_{t-d+1} | s^{t-d}, y^{t-d}) \Bigg), && \text{(3.34d)}
\end{aligned}
$$

which implies that θ_t can be recursively updated as

$$
\theta_t = \omega(\theta_{t-1}, u_t, y_{t-d+1}, s_{t-d+1}, s_{t-d}). \tag{3.35}
$$

We now show that $\{(S_{t-d}, \Theta_{t-1})\}_t$ is a controlled Markov chain with control U_t. Indeed,

$$
\begin{aligned}
&P(s_{t-d+1}, d\theta_t | s^{t-d}, \theta^{t-1}, u^t) \\
&\quad = \sum_{y_{t-d+1}, x_{t-d}^{t-d+1}} \delta_{\omega(\theta_{t-1}, u_t, y_{t-d+1}, s_{t-d+1}, s_{t-d})} Q'(y_{t-d+1} | x_{t-d+1}, s_{t-d+1}) \\
&\qquad\qquad \times Q(s_{t-d+1} | s_{t-d}, x_{t-d}) \theta_{t-1}(x_{t-d}^{t-d+1}) && \text{(3.36a)} \\
&\quad = P(s_{t-d+1}, d\theta_t | s_{t-d}, \theta_{t-1}, u_t). && \text{(3.36b)}
\end{aligned}
$$

Furthermore, the instantaneous reward r_t can be written as

$$
\begin{aligned}
&r_t \\
&= \log \frac{P(y_t, s_t|s^{t-1}, y^{t-1}, x_{t-1}^{t})}{P(y_t, s_t|s^{t-1}, y^{t-1})} && (3.37a) \\
&= \log \frac{Q'(y_t|x_t, s_t)Q(s_t|s_{t-1}, x_{t-1})P(y_{t-d+1}^{t-1}, s_{t-d+1}^{t-1}|s^{t-d}, y^{t-d})}{P(y_{t-d+1}^{t}, s_{t-d+1}^{t}|s^{t-d}, y^{t-d})} && (3.37b) \\
&= \log \frac{Q'(y_t|x_t, s_t)Q(s_t|s_{t-1}, x_{t-1})P(y_{t-d+1}^{t-1}, s_{t-d+1}^{t-1}|s^{t-d}, y^{t-d})}{\sum_{x_{t-d}^{t}} \Big(\prod_{i=t-d+1}^{t} Q'(y_i|x_i, s_i)Q(s_i|s_{i-1}, x_{i-1})\Big) u_t(x_t; x_{t-d}^{t-1})\theta_{t-1}(x_{t-d}^{t-1})} && (3.37c) \\
&= \log \Bigg\{ \Bigg(Q'(y_t|x_t, s_t)Q(s_t|s_{t-1}, x_{t-1}) \sum_{x_{t-d}^{t-1}} \Big(\prod_{i=t-d+1}^{t-1} Q'(y_i|x_i, s_i)Q(s_i|s_{i-1}, x_{i-1}) \Big) \\
&\qquad \times \theta_{t-1}(x_{t-d}^{t-1}) \Bigg) \Big/ \Bigg(\sum_{x_{t-d}^{t}} \Big(\prod_{i=t-d+1}^{t} Q'(y_i|x_i, s_i)Q(s_i|s_{i-1}, x_{i-1}) \Big) \\
&\qquad \times u_t(x_t; x_{t-d}^{t-1})\theta_{t-1}(x_{t-d}^{t-1}) \Bigg) \Bigg\}. && (3.37d)
\end{aligned}
$$

The expected reward at time t conditioned on the states and control actions up to time t is

$$
\begin{aligned}
&E\left[R_t|s^{t-d}, \theta^{t-1}, u^t\right] \\
&= E\Bigg[\log \Bigg\{ \Bigg(Q'(Y_t|X_t, S_t)Q(S_t|S_{t-1}, X_{t-1}) \\
&\qquad \times \sum_{x_{t-d}^{t-1}} \Big(\prod_{i=t-d+2}^{t-1} Q'(Y_i|x_i, S_i)Q(S_i|S_{i-1}, x_{i-1}) \Big) \\
&\qquad \times Q'(Y_{t-d+1}|x_{t-d+1}, S_{t-d+1})Q(S_{t-d+1}|s_{t-d}, x_{t-d})\theta_{t-1}(x_{t-d}^{t-1}) \Bigg) \\
&\quad \Big/ \Bigg(\sum_{x_{t-d}^{t}} \Big(\prod_{i=t-d+2}^{t} Q'(Y_i|x_i, S_i)Q(S_i|S_{i-1}, x_{i-1}) \Big) \\
&\qquad \times Q'(Y_{t-d+1}|x_{t-d+1}, S_{t-d+1}) \\
&\qquad \times Q(S_{t-d+1}|s_{t-d}, x_{t-d})u_t(x_t; x_{t-d}^{t-1})\theta_{t-1}(x_{t-d}^{t-1}) \Bigg) \Bigg\} \mid s^{t-d}, \theta^{t-1}, u^t \Bigg] && (3.38a)
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{y_{t-d+1}^t, x_{t-d}^t, s_{t-d+1}^t} \prod_{i=t-d+1}^{t} Q'(y_i|x_i, s_i) Q(s_i|s_{i-1}, x_{i-1}) u_t(x_t; x_{t-d}^{t-1}) \theta_{t-1}(x_{t-d}^{t-1}) \\
&\quad \times \log \Bigg\{ \bigg(Q'(y_t|x_t, s_t) Q(s_t|s_{t-1}, x_{t-1}) \sum_{x_{t-d}^{t-1}} \Big(\prod_{i=t-d+1}^{t-1} Q'(y_i|x_i, s_i) Q(s_i|s_{i-1}, x_{i-1}) \Big) \\
&\quad \times \theta_{t-1}(x_{t-d}^{t-1}) \bigg) \Big/ \bigg(\sum_{x_{t-d}^t} \Big(\prod_{i=t-d+1}^{t} Q'(y_i|x_i, s_i) Q(s_i|s_{i-1}, x_{i-1}) \Big) \qquad (3.38b) \\
&\quad \times u_t(x_t; x_{t-d}^{t-1}) \theta_{t-1}(x_{t-d}^{t-1}) \bigg) \Bigg\} \\
&= \overline{r}_t(s_{t-d}, \theta_{t-1}, u_t), \qquad (3.38c)
\end{aligned}
$$

which is only a function of the observed part of the state s_{t-d}, the belief on the unobserved part of the state θ_{t-1}, and the action u_t.

Note that the conditional expected reward at time t is does not depend on y_{t-d}. Furthermore, y_{t-d} does not affect the future evolution of the information state as seen in (3.34). Therefore, it can be shown that the optimal policy is a function of only s_{t-d}, θ_{t-1} (this can be shown for instance using the graphical modeling approach presented in [66]). Then, the optimal channel input distributions take the form $P(x_t|x_{t-d}^{t-1}, s_{t-d}, \theta_{t-1})$, and the capacity expression becomes

$$
C_d = \sup_{\{P(x_t|x_{t-d}^{t-1}, s_{t-d}, \theta_{t-1})\}_t} \liminf_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} I(X_{t-1}^t; S_t, Y_t | S_{t-d}^{t-1}, Y_{t-d+1}^{t-1}, \Theta_{t-1}). \qquad (3.39)
$$

Note that above described controlled Markov chain is time-homogenous, and hence the optimal channel input distribution is time-invariant, and consequently the capacity expression reduces to the one in (3.16).

3.7 Proof of Lemma 3.13

First, note that we have for every t

$$
\begin{aligned}
& I(X^t; S_t, Y_t | S^{t-1}, Y^{t-1}) \\
&= I(X_t; S_t, Y_t | S^{t-1}, Y^{t-1}) + I(X^{t-1}; S_t, Y_t | S^{t-1}, Y^{t-1}, X_t) && \text{(3.40a)} \\
&\overset{(a)}{=} I(X_t; S_t, Y_t | S^{t-1}, Y^{t-1}) && \text{(3.40b)} \\
&= I(X_t; Y_t | S^t, Y^{t-1}) + I(X_t; S_t | S^{t-1}, Y^{t-1}) && \text{(3.40c)} \\
&\overset{(b)}{=} I(X_t; Y_t | S^t, Y^{t-1}), && \text{(3.40d)}
\end{aligned}
$$

where (a) is from the fact that S_t, Y_t is independent of X^{t-1} given S_{t-1}, X_t, and (b) is from the fact that S_t is independent of X_t given S_{t-1}. Each of the terms $I(X_t; Y_t | S^t, Y^{t-1})$ in the summation is evaluated based on the joint distribution $P(x_t, s^t, y^t)$. We now proceed by induction to prove that the sequence of measures $\{P(x_t, s^t, y^t)\}_{t=1}^T$ induced by the sequence of channel input distributions $\{P(x_t | x^{t-1}, s^{t-d}, y^{t-d})\}_{t=1}^T$ equals to the sequence of measures $\{P_1(x^t, s^t, y^t)\}_{t=1}^T$ induced by an appropriately defined sequence of channel input distributions of the form $\{P_1(x_t | s^{t-d}, y^{t-d})\}_{t=1}^T$.

For $t = 1$ we set $P_1(x_1) = P(x)$ and have

$$
\begin{aligned}
P_1(x_1, s_1, y_1) &= Q'(y_1 | s_1, x_1) Q(s_1) P_1(x_1) && \text{(3.41a)} \\
&= Q'(y_1 | s_1, x_1) Q(s_1) P(x_1) = P(x_1, s_1, y_1). && \text{(3.41b)}
\end{aligned}
$$

Now for $t+1$ we set $P_1(x_{t+1}|s^{t-d+1}, y^{t-d+1}) = P(x_{t+1}|s^{t-d+1}, y^{t-d+1})$ and have

$$\begin{aligned} & P_1(x_{t+1}, s^{t+1}, y^{t+1}) \\ & \quad = Q'(y_{t+1}|x_{t+1}, s_{t+1})Q(s_{t+1}|s_t)P_1(x_{t+1}|s^{t-d+1}, y^{t-d+1}) \sum_{x_t} P_1(x_t, s^t, y^t) && \text{(3.42a)} \\ & \quad \overset{(a)}{=} Q'(y_{t+1}|x_{t+1}, s_{t+1})Q(s_{t+1}|s_t)P(x_{t+1}|s^{t-d+1}, y^{t-d+1}) \sum_{x_t} P(x_t, s^t, y^t) && \text{(3.42b)} \\ & \quad = P(x_{t+1}, s^{t+1}, y^{t+1}), && \text{(3.42c)} \end{aligned}$$

where (a) is due to the construction of $P_1(x_{t+1}|, s^t, y^t)$, and the induction hypothesis. The above equality implies that the equality in (3.21).

3.8 Proof of Proposition 3.15

First, note that $\{S_{t-d}\}_t$ is a controlled Markov chain with control U_t.

Furthermore, the instantaneous reward r_t can be written as

$$\begin{aligned} r_t &= \log \frac{P(y_t|s^t, y^{t-1}, x_t)}{P(y_t|s^t, y^{t-1})} && \text{(3.43a)} \\ &= \log \frac{Q'(y_t|x_t, s_t)}{\sum_{x_t} Q'(y_t|x_t, s_t)u_t(x_t)}. && \text{(3.43b)} \end{aligned}$$

The expected reward at time t conditioned on the states and control actions up to time t is

$$\begin{aligned}
&E\left[R_t|s^{t-d}u^t\right] \\
&= E\left[\log \frac{Q'(Y_t|X_t,S_t)}{\sum_{x_t} Q'(Y_t|x_t,S_t)u_t(x_t)} \mid s^{t-d},u^t\right] && \text{(3.44a)} \\
&= \sum_{y_t,x_t,s^t_{t-d+1}} Q'(y_t|x_t,s_t) \\
&\quad \times \prod_{t'=t-d+1}^{t} Q(s_{t'}|s_{t'-1})u_t(x_t) \log \frac{Q'(y_t|x_t,s_t)}{\sum_{x_t} Q'(y_t|x_t,s_t)u_t(x_t)}, && \text{(3.44b)} \\
&= \bar{r}_t(s_{t-d},u_t), && \text{(3.44c)}
\end{aligned}$$

which is only a function of s_{t-d} and the action u_t.

Note that the conditional expected reward at time t is does not depend on y_{t-1}. Furthermore, y_{t-1} does not affect the future evolution of $\{S_{t-d}\}_t$. Therefore, it can be shown that the optimal policy is the function of only s_{t-d} (this can be shown for instance using the graphical modeling approach presented in [66]). Then, the optimal channel input distributions take the form $P(x_t|s_{t-d})$, and the capacity expression becomes

$$C = \sup_{\{P(x_t|s_{t-d})\}_t} \liminf_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} I(X_t;Y_t|S^t_{t-1}). \tag{3.45}$$

Note that above described controlled Markov chain is time-homogenous, and hence the optimal channel input distribution is time-invariant, and consequently the capacity expression reduces to the one in (3.23).

CHAPTER 4

A capacity-achieving posterior matching scheme: No inter-symbol interference (ISI) case

In this chapter we describe a transmission scheme which achieves the capacity of the finite-state channel with unit feedback delay when there is no ISI. This case can be thought of as the special case of general FSC with ISI. In order to appreciate the difficulties involved in proving capacity achievability for the channels with memory we present in Appendix A the corresponding scheme for the discrete memoryless channel (DMC). In [49, 50] a proof of capacity achievability was established using Lyapunov function, contraction mapping and the strong law of large numbers (SLLN) of Markov chains. Our goal is to provide a extended version of this proof for the case of channels with memory. As will be seen in Chapter 5, overall situation changes significantly from the DMC case when there is ISI. Before going into that let's look at the no ISI case as a rather simple generalization of the DMC case.

4.1 A posterior matching scheme for channels with memory

In this section we describe a transmission scheme which achieves the capacity of the finite-state channel with unit feedback delay. The scheme which achieves the capacity of the finite-state channel with arbitrary feedback delay can be thought as a slight generalization of the scheme presented here. With unit feedback delay, the capacity expression in (3.23) becomes

$$C = \sup_{\{P(X|S')\}_{S'}} I(X;Y|S,S') \tag{4.1}$$

where the mutual information is evaluated using the joint measure

$$P(Y,S,X,S') = Q'(Y|S,X)Q(S|S')P(X|S')P(S'). \tag{4.2a}$$

The distribution $P(S)$ is the solution of the equation

$$P(S) = \sum_{S'} P(S')Q(S|S'). \tag{4.2b}$$

We assume that the capacity achieving distributions $\{\hat{P}(X|S')\}_{S'}$ have been found for all values of (S') and the corresponding steady-state distribution on X, $\hat{P}(X|S')$ and on Y_t conditioned on S_t, S_{t-1}, $\hat{P}(y|s,s') = \sum_x Q'(y|x,s)\hat{P}(x|s')$ have been evaluated. Define the random variable $F_t \in \mathcal{F}$ as $F_t(w) \stackrel{\text{def}}{=} F(w|Y^t,S^t)$, where $F(\cdot|Y^t,S^t)$ is the a-posteriori cdf of W conditioned on Y^t, S^t, and $\mathcal{F}$ is the set of all valid cdfs over $[0,1)$.

The channel input X_t is generated as

$$
\begin{aligned}
X_t &= F^{-1}_{\hat{P}(\cdot|S_{t-1})}(F_{t-1}(W)) && \text{(4.3a)} \\
&= x, \qquad \sum_{i=0}^{x-1} \hat{P}(i|S_{t-1}) < F_{t-1}(W) \leq \sum_{i=0}^{x} \hat{P}(i|S_{t-1}) \\
&\qquad x = 0, \ldots, K_X - 1 && \text{(4.3b)} \\
&\stackrel{\text{def}}{=} e(F_{t-1}(W), \hat{P}(\cdot|S_{t-1})). && \text{(4.3c)}
\end{aligned}
$$

where the inverse cdf $F^{-1}(y) \stackrel{\text{def}}{=} \inf\{x : F(x) \geq y\}$.

At the receiver, the message estimate is obtained as

$$\hat{W}_t = d(F_t, 2^{-Rt}/2), \tag{4.4}$$

where the message estimate function $d(F, \epsilon)$ is defined as

$$d(F, \epsilon) = \arg\max_{w} \{F(w + \epsilon) - F(w - \epsilon)\}. \tag{4.5}$$

Lemma 4.1. *For the PMS scheme we have*

$$
\begin{aligned}
P(X_t|Y^{t-1}, S^{t-1}) &= \hat{P}(X_t|S_{t-1}) && \text{(4.6a)} \\
P(Y_t|Y^{t-1}, S^{t}) &= \hat{P}(Y_t|S_t, S_{t-1}) && \text{(4.6b)} \\
F_t &= \phi(F_{t-1}, Y_t, S_t, S_{t-1}). && \text{(4.6c)}
\end{aligned}
$$

where $F_0 = Uniform[0, 1)$.

Proof.

$$
\begin{aligned}
&P(X_t|Y^{t-1}, S^{t-1}) \\
&= P(X_t|Y^{t-1}, S^{t-1}, F_{t-1}) && \text{(4.7a)} \\
&= \int_w P(X_t|Y^{t-1}, S^{t-1}, F_{t-1}, W = w)P(dw|Y^{t-1}, S^{t-1}, F_{t-1}) && \text{(4.7b)} \\
&= \int_w \delta_{e(F_{t-1}(w), S_{t-1})}(X_t) dF_{t-1}(w) && \text{(4.7c)} \\
&= \sum_x \delta_x(X_t)\hat{P}(x|S_{t-1}) && \text{(4.7d)} \\
&= \hat{P}(X_t|S_{t-1}), && \text{(4.7e)}
\end{aligned}
$$

and similarly

$$
\begin{aligned}
&P(Y_t|Y^{t-1}, S^t) \\
&\quad = \sum_x P(Y_t|Y^{t-1}, S^t, X_t = x)P(X_t = x|Y^{t-1}, S^{t-1}) && \text{(4.8a)} \\
&\quad = \sum_x Q'(Y_t|S_t, x)\hat{P}(x|S_{t-1}) && \text{(4.8b)} \\
&\quad = \hat{P}(Y_t|S_t, S_{t-1}). && \text{(4.8c)}
\end{aligned}
$$

Then, F_t is updated as

$$
F_t = \phi(F_{t-1}, Y_t, S_t, S_{t-1}). \tag{4.9}
$$

where ϕ is given implicitly through the corresponding pdf update

$$\begin{aligned}
&df_t(a) \\
&= \frac{P(y^t, s^t|a)da}{P(y^t, s^t)} && \text{(4.10a)} \\
&= \frac{Q'(y_t|s_t, e(f_{t-1}(a), s_{-1}))Q(s_t|s_{t-1})P(y^{t-1}, s^{t-1}|a)da}{P(y_t, s_t|y^{t-1}, s^{t-1})P(y^{t-1}, s^{t-1})} && \text{(4.10b)} \\
&= \frac{Q'(y_t|s_t, e(f_{t-1}(a), s_{t-1}))Q(s_t|s_{t-1})df_{t-1}(a)}{P(y_t, s_t|y^{t-1}, s^{t-1})} && \text{(4.10c)} \\
&\overset{(a)}{=} \frac{Q'(y_t|s_t, e(f_{t-1}(a), s_{t-1}))Q(s_t|s_{t-1})df_{t-1}(a)}{Q(s_t|s_{t-1})\hat{P}(y_t|s_t, s_{t-1})} && \text{(4.10d)} \\
&= \frac{Q'(y_t|s_t, e(f_{t-1}(a), s_{t-1}))df_{t-1}(a)}{\hat{P}(y_t|s_t, s_{t-1})}, && \text{(4.10e)}
\end{aligned}$$

where (a) is from (4.6b), and explicitly through

$$f_t(a) = \frac{\sum_{i=0}^{x-1} Q'(y_t|s_t, i)\hat{P}(i|s_{t-1}) + Q'(y_t|s_t, x)[f_{t-1}(a) - \sum_{i=0}^{x-1} \hat{P}(i|s_{t-1})]}{\hat{P}(y_t|s_t, s_{t-1})},$$
$$\sum_{i=0}^{x-1} \hat{P}(i|s_{t-1}) < f_{t-1}(a) \leq \sum_{i=0}^{x} \hat{P}(i|s_{t-1}), x = 0, \ldots, K_X - 1. \tag{4.11}$$

□

Observe that $f_t(a)$ is a function of f_{t-1} only through $f_{t-1}(a)$. In the following we will also use the notation $f_t(a) = \phi(f_{t-1}, y_t, s_t, s_{t-1})(w) = \phi(f_{t-1}(a), y_t, s_t, s_{t-1})$. Observe also from (4.3) that the transmitted symbol X_t is a function of W and F_{t-1} only through the quantity $F_{t-1}(W)$. This has important implications for the analysis of the PMS scheme.

Assuming that the channel and state transition probabilities $Q'(y|x, s)$ and $Q(s'|s)$ are non-zero for all x, y, s, s', the recursion (4.11) guarantees that for every realization of the random variables of interest, F_t will always have a pdf; in addition the pdf will be non-zero everywhere in $(0, 1]$.

4.2 Achievability Result

Let $\hat{W}_t$ be the message point estimate at the receiver at time t. Then, a transmission scheme achieves rate R if

$$\lim_{t \to \infty} P(|W - \hat{W}_t| > 2^{-tR}) = 0. \tag{4.12}$$

In particular, we say that a transmission schemes achieves zero rate if

$$\forall \epsilon > 0 \qquad \lim_{t \to \infty} P(|W - \hat{W}_t| > \epsilon) = 0. \tag{4.13}$$

4.2.1 Zero Rate Result

For a cdf $h : [0,1] \to [0,1]$ define a Lyapunov function V_λ as follows.

$$V_\lambda(h) = \int_0^1 \lambda(h(w))dw, \tag{4.14}$$

where $\lambda : [0,1] \to [0,1]$ is onto, strictly concave and symmetric about 0.5. This definition implies that $\lambda(x)$ is 0 at $x = 0,\ 1$ and 1 at $x = 1/2$. Furthermore, for a cdf $F \in \mathcal{F}$, $V_\lambda(F)$ is small if F resembles a step function (it is exactly 0 for a step function). A function $\xi : [0,1] \to [0,1]$ is called contraction if it is nonnegative, concave, and $\xi(x) < x$ for $x \in (0,1)$.

Definition 4.2. A channel is called fixed-point free if for any $f_t(w), s_t$

$$P\Big(\phi(f_t(w), Y_{t+1}, S_{t+1}, s_t) = f_t(w)\Big) < 1. \tag{4.15}$$

Lemma 4.3. *If the channel is fixed-point free, then for $\epsilon > 0$ and for all $f \in \mathcal{F}$,*

$$\lim_{t\to\infty} P(V_\lambda(F_t) > \epsilon | F_0 = f) = 0. \tag{4.16}$$

Proof. See 4.3. $\square$

The intuitive interpretation of the above lemma is that the probability of having an F_t that does not resemble a step function is zero at the limit of large t.

For any $t_2 > t_1 > 0$ we can write F_{t_2} as a function of F_{t_1-1} and the quantities $Y_{t_1}^{t_2}, S_{t_1-1}^{t_2}$ through a repeated application of the ϕ recursion, i.e.,
$F_{t_2} \stackrel{\text{def}}{=} \phi_{t_2-t_1}(F_{t_1-1}, Y_{t_1}^{t_2}, S_{t_1-1}^{t_2})$. Let F_{t_1,t_2}^f be the random variable defined as $F_{t_1,t_2}^f \stackrel{\text{def}}{=} \phi_{t_2-t_1}(f, Y_{t_1}^{t_2}, S_{t_1-1}^{t_2})$. Clearly $F_t = F_{1,t}^u$, where u denotes the uniform distribution over $(0,1)$. In addition, due to the recursion implied by the PMS, we will denote $F_{t_2}(a) \stackrel{\text{def}}{=} \phi_{t_2-t_1}(F_{t_1-1}, Y_{t_1}^{t_2}, S_{t_1-1}^{t_2})(a) \stackrel{\text{def}}{=} \phi_{t_2-t_1}(F_{t_1-1}(a), Y_{t_1}^{t_2}, S_{t_1-1}^{t_2})$ with some notational abuse. With the above notation, the function $\phi_{t_2-t_1}(\cdot, Y_{t_1}^{t^2}, S_{t_1-1}^{t_2})$ is monotonically increasing. We now prove the following lemma which is a stronger version of Lemma 4.3.

Lemma 4.4. *If the channel is fixed-point free, then for $\epsilon > 0$,*

$$\lim_{t\to\infty} P(\max_{1\le t'\le t} V_\lambda(F_{t'+1,(1+\alpha)t}^u) > \epsilon) = 0. \tag{4.17}$$

Proof. Let $V_{\lambda,t_1,t_2}^* = \sup_f V_\lambda(F_{t_1,t_2}^f)$. Note that V_{t_1,t_2}^* is a deterministic function of

$Y_{t_1}^{t_2}, S_{t_1-1}^{t_2}$, hence there exists a sequence of cdfs $\{f_{k,Y_{t_1}^{t_2},S_{t_1-1}^{t_2}}\}_{k=1}^{\infty}$ such that

$$V^*_{\lambda,t_1,t_2} = \lim_{k\to\infty} V_\lambda(F_{t_1,t_2}^{f_{k,Y_{t_1}^{t_2},S_{t_1-1}^{t_2}}}) \tag{4.18a}$$

$$= \lim_{k\to\infty} V_\lambda(F_{t_1+1,t_2}^{\phi(f_{k,Y_{t_1}^{t_2},S_{t_1-1}^{t_2}},Y_{t_1},S_{t_1},S_{t_1-1})}) \tag{4.18b}$$

$$\leq \sup_f V_\lambda(F_{t_1+1,t_2}^f) \tag{4.18c}$$

$$= V^*_{\lambda,t_1+1,t_2}. \tag{4.18d}$$

Note that there exists a sequence of cdfs $\{f_{k,Y_1^{\alpha t},S_0^{\alpha t}}\}_{k=1}^{\infty}$ such that $V^*_{\lambda,1,\alpha t} \geq V_\lambda(F_{1,\alpha t}^{f_{k,Y_1^{\alpha t},S_0^{\alpha t}}}) > V^*_{\lambda,1,\alpha t} - 1/k$. Therefore,

$$P(\max_{1\leq t'\leq t} V_\lambda(F^u_{t'+1,(1+\alpha)t} > \epsilon) \leq P(\max_{1\leq t'\leq t} V^*_{\lambda,t'+1,(1+\alpha)t} > \epsilon) \tag{4.19a}$$

$$= P(V^*_{\lambda,t+1,(1+\alpha)t} > \epsilon) \tag{4.19b}$$

$$= P(\sup_f V_\lambda(F^f_{t+1,(1+\alpha)t}) > \epsilon) \tag{4.19c}$$

$$\stackrel{(a)}{=} P(\sup_f V_\lambda(F^f_{1,\alpha t}) > \epsilon) \tag{4.19d}$$

$$= P(\lim_{k\to\infty} V_\lambda(F_{1,\alpha t}^{f_{k,Y_1^{\alpha t},S_0^{\alpha t}}}) > \epsilon) \tag{4.19e}$$

$$\stackrel{(b)}{\leq} P(V_\lambda(F_{1,\alpha t}^{f_{k',Y_1^{\alpha t},S_0^{\alpha t}}}) > \epsilon) \tag{4.19f}$$

$$\stackrel{(c)}{\to} 0, \tag{4.19g}$$

where (a) is due to the fact that $F^f_{t+1,(1+\alpha)t} = \phi_{\alpha t-1}(f, Y_{t+1}^{(1+\alpha)t}, S_t^{(1+\alpha)t})$, $F^f_{1,\alpha t} = \phi_{\alpha t-1}(f, Y_1^{\alpha t}, S_0^{\alpha t})$, and $(Y_{t+1}^{(1+\alpha)t}, S_t^{(1+\alpha)t})$, $(Y_1^{\alpha t}, S_0^{\alpha t})$ have the same statistics; (b) is true for $k' > 1/(V^*_{\lambda,1,\alpha t} - \epsilon)$; and (c) is due to the fact that Lemma 4.3 holds for any F_0. □

Observe that indeed this lemma is stronger than Lemma 4.3, since $P(V_\lambda(F_t) > \epsilon) = P(V_\lambda(F^u_{1,t}) > \epsilon) = P(V_\lambda(F^u_{1+t,t+t}) > \epsilon) \leq P(\max_{1\leq t'\leq t} V_\lambda(F^u_{t'+1,2t}) > \epsilon)$.

Proposition 4.5. *If the channel is fixed-point free, then for* $\epsilon, \delta > 0$

$$\lim_{t\to\infty} P\Big(\max_{1\le t'\le t} F^u_{t'+1,(1+\alpha)t}\Big(F_{t'}(W)-\delta\Big) > \epsilon\Big) = 0, \tag{4.20a}$$

$$\lim_{t\to\infty} P\Big(\max_{1\le t'\le t} F^u_{t'+1,(1+\alpha)t}\Big(F_{t'}(W)+\delta\Big) < 1-\epsilon\Big) = 0. \tag{4.20b}$$

Proof. See 4.4. □

The intuition behind the above proof is that for an error to occur, either the cdf F_t does not behave as a step function (first, third and fourth terms in (4.42d)) or the step does not occur at the transmitted message W (second term in (4.42d)).

4.2.2 Rate $R < C$ achievability

So far we have established zero-rate result. We now prove the rate R achievability.

Lemma 4.6. $(F_t(W), S_t^{t+1}, Y_{t+1})_t$ *is a Markov chain.*

Proof. Let $\Theta_t \overset{\text{def}}{=} F_t(W)$. We have

$$\begin{aligned}
&P(\theta_t, s_t^{t+1}, y_{t+1}|\theta^{t-1}, s'^t, y^t)\\
&\quad= Q'(y_{t+1}|e(\theta_t, s_t), s_{t+1})Q(s_{t+1}|s_t)P(\theta_t, s_t|\theta_{t-1}, s'^t, y^t) &(4.21a)\\
&\quad= Q'(y_{t+1}|e(\theta_t, s_t), s_{t+1})Q(s_{t+1}|s_t)\delta_{\phi(\theta_{t-1},y_t,s_t,s'_{t-1})}(\theta_t)\delta_{s'_t}(s_t) &(4.21b)\\
&\quad= P(\theta_t, s_t^{t+1}, y_{t+1}|\theta_{t-1}, s'^t_{t-1}, y_t). &(4.21c)
\end{aligned}$$

□

Lemma 4.7. $E\{\log \frac{dF_{t+1}(W)}{dF_t(W)}|\} = C$ *if the PMS is used as a transmission scheme.*

Proof. See 4.5. □

Definition 4.8. An invariant distribution P_Ψ of a Markov chain $\{\Psi_t\}_t$ is called ergodic if for every invariant set A either $P_\Psi(A) = 0$ or $P_\Psi(A) = 1$.

Lemma 4.9. *If a Markov chain $(F_t(W), S_t^{t+1}, Y_{t+1})_t$ has ergodic invariant distribution, then*

$$\lim_{t\to\infty} \frac{1}{t} \log dF_t(W) = C \ \ a.s. \tag{4.22}$$

Proof.

$$\frac{1}{t} \log dF_t(W) = \frac{1}{t} \sum_{s=1}^{t} \log \frac{dF_s(W)}{dF_{s-1}(W)}. \tag{4.23}$$

If $(F_t(W), S_t^{t+1}, Y_{t+1})_t$ is ergodic, then by the strong law of large numbers for Markov chains [67]

$$\lim_{t\to\infty} \frac{1}{t} \log dF_t(W) = E[\log \frac{dF_t(W)}{dF_{t-1}(W)}] \overset{(a)}{=} C \qquad \text{a.s.} \tag{4.24}$$

where (a) is from Lemma 4.7. □

Lemma 4.10. *If a Markov chain $(F_t(W), S_t^{t+1}, Y_{t+1})_t$ has ergodic invariant distribution, then for any $\delta > 0$ and rate $R < C - \delta$ there exists $\epsilon' > 0$ so that for all $\epsilon \leq \epsilon'$*

$$\lim_{t\to\infty} P\left(\bigcap_{s=0}^{t-1} \{F_s(W) - F_s(W - 2^{-tR}) < \epsilon\}\right) = 0 \tag{4.25a}$$

$$\lim_{t\to\infty} P\left(\bigcap_{s=0}^{t-1} \{F_s(W + 2^{-tR}) - F_s(W) < \epsilon\}\right) = 0. \tag{4.25b}$$

Proof. See 4.6. □

The above lemma guarantees that at some time before t there will be a jump of at least ϵ in the posterior message cdf in the interval of 2^{-tR} around W. Using this lemma we show the main result.

Proposition 4.11. *If a Markov chain $(F_t(W), S_t^{t+1}, Y_{t+1})_t$ has ergodic invariant distribution, then for $\delta, \alpha > 0$*

$$\lim_{t\to\infty} P(F_{(1+\alpha)t}(W - 2^{-tR}) > \delta) = 0, \tag{4.26a}$$

$$\lim_{t\to\infty} P(F_{(1+\alpha)t}(W + 2^{-tR}) < 1 - \delta) = 0. \tag{4.26b}$$

Proof. See 4.7. □

4.3 Proof of Lemma 4.3

$$E[F_t|s^{t-1}, y^{t-1}] = E[\phi(f_{t-1}, Y_t, S_t, s_{t-1}|s^{t-1}, y^{t-1})] \tag{4.27a}$$

$$= E[\phi(f_{t-1}, Y_t, S_t, s_{t-1})|s_{t-1}]. \tag{4.27b}$$

Then,

$$E[dF_t(w)|s^{t-1}, y^{t-1}]$$

$$= E[\phi(df_{t-1}(w), Y_t, S_t, s_{t-1})|s_{t-1}] \tag{4.28a}$$

$$= \sum_{y_t, s_t, x_t} Q'(y_t|s_t, x_t) Q(s_t|s_{t-1}) \hat{P}(x_t|s_{t-1}) \times \frac{Q'(y_t|s_t, e(f_{t-1}(w), s_{t-1})) df_{t-1}(w)}{\hat{P}(y_t|s_t, s_{t-1})} \tag{4.28b}$$

$$= df_{t-1}(w). \tag{4.28c}$$

Similarly, we get

$$E[F_t(w)|s^{t-1}, y^{t-1}] = E[\phi(f_{t-1}(w), Y_t, S_t, s_{t-1})|s_{t-1}] \tag{4.29a}$$

$$= f_{t-1}(w). \tag{4.29b}$$

We would like to find a contraction mapping ξ such that for every w and f_{t-1} we have $E[\lambda(\phi(f_{t-1}(w), Y_t, S_t, s_{t-1}|s_{t-1}))] \leq \xi(\lambda(f_{t-1}(w)))$. Let us assume for now that such a contraction mapping exists. We have

$$E[V_\lambda(\phi(f_{t-1}, Y_t, S_t, s_{t-1}))|s_{t-1}]$$

$$= E\Big[\int_0^1 \lambda(\phi(f_{t-1}(w), Y_t, S_t, s_{t-1}))|s_{t-1}\Big] \tag{4.30a}$$

$$< \int_0^1 \xi(\lambda(f_{t-1}(w)))dw \tag{4.30b}$$

$$\leq \xi(V_\lambda(f_{t-1})), \tag{4.30c}$$

where the first inequality is due to the assumption for the property of ξ and the second inequality is due to the concavity of ξ. Then

$$P(V_\lambda(F_t) > \epsilon) \leq \frac{E[V_\lambda(F_t)]}{\epsilon} \tag{4.31a}$$

$$= \frac{E[E[V_\lambda(F_t)|S^{t-1}, Y^{t-1}]]}{\epsilon} \tag{4.31b}$$

$$= \frac{E[E[V_\lambda(\phi(f_{t-1}, Y_t, S_t, S_{t-1}))|S_{t-1}]]}{\epsilon} \tag{4.31c}$$

$$\leq \frac{E[\xi(V_\lambda(F_{t-1}))]}{\epsilon} \tag{4.31d}$$

$$\leq \frac{\xi(E[V_\lambda(F_{t-1})])}{\epsilon} \cdots \leq \frac{\xi^t(E[V_\lambda(F_0)])}{\epsilon} \xrightarrow{(a)} 0, \tag{4.31e}$$

where the first inequality is the Markov inequality, the second inequality is due to (4.30), the third inequality is due to the concavity of ξ, the fourth inequality is due to repeated application of the above inequalities and the convergence to 0 is due to the property of the contraction [50, Lemma8]. Observe that the convergence is true for any initial distribution F_0, which implies that the convergence is uniform in the initial distribution.

It remains to find the contraction ξ with the property $E[\lambda(\phi(f_{t-1}(w), Y_t, S_t, s_{t-1}|s_{t-1}))] \leq \xi(\lambda(f_{t-1}(w)))$. To this end let $\lambda' : [0, 0.5] \to [0, 1]$ be a restriction of λ on $[0, 0.5]$. Then, λ' becomes one-to-one and onto hence it has inverse. Let $\tilde{\xi} : [0, 1] \to [0, 1]$ be $\tilde{\xi}(a) = \max_{s_{t-1}} E[\lambda(\phi(a, Y_t, S_t, s_{t-1}))|s_{t-1}]$. Consider now a following function.

$$\xi^*(a) = \max\left\{\tilde{\xi}(\lambda'^{-1}(a)), \tilde{\xi}(1 - \lambda'^{-1}(a))\right\}. \tag{4.32}$$

Clearly, $\xi^*(x) \geq 0$. We will now show that ξ^* satisfies the aforementioned property. Indeed, let $a \stackrel{\text{def}}{=} f_{t-1}(w)$. If $a \in [0, 1/2]$ then $\lambda'^{-1}(\lambda(a)) = a$ and the first term in the maximization on the r.h.s. of (4.32) equals $\max_{s_{t-1}} E[\lambda(\phi(a, Y_t, S_t, s_{t-1}|s_{t-1}))]$. If $a \in [1/2, 1]$ then $1 - \lambda'^{-1}(\lambda(a)) = a$ and the second term in the maximization of the r.h.s. of (4.32) equals $\max_{s_{t-1}} E[\lambda(\phi(a, Y_t, S_t, s_{t-1}|s_{t-1}))]$. Thus the property holds. We now need to show that $\xi^*(x) < x$ for all $x \in (0, 1)$. This is equivalent to showing that for every $x \in (0, 1)$,$E[\lambda(\phi(\lambda'^{-1}(x), Y_t, S_t, s_{t-1}|s_{t-1}))] < x$ and $E[\lambda(\phi(1-\lambda'^{-1}(x), Y_t, S_t, s_{t-1}|s_{t-1}))] < x$, which is equivalent to showing that for all $a \in (0, 1/2)$ we have $E[\lambda(\phi(a, Y_t, S_t, s_{t-1}|s_{t-1}))] < \lambda'(a)$ and $E[\lambda(\phi(1 - a, Y_t, S_t, s_{t-1}|s_{t-1}))] < \lambda'(a)$. This in turn is equivalent to showing that $E[\lambda(\phi(\lambda'^{-1}(x), Y_t, S_t, s_{t-1}|s_{t-1}))] < \lambda(a)$ for all $a \in (0, 1)$. To show this, since the channel is fixed-point free, $F_t(w)$ is not a.s. constant. Hence, using Jensen's inequality we get

$$
\begin{aligned}
&E[\lambda(\phi(f_{t-1}(w), Y_t, S_t, s_{t-1}))|s_{t-1}] \\
&\quad < \lambda(E[\phi(f_{t-1}(w), Y_t, S_t, s_{t-1})|s_{t-1}]) &\text{(4.33a)} \\
&\quad = \lambda(f_{t-1}(w)). &\text{(4.33b)}
\end{aligned}
$$

Hence,

$$\xi^*(a) < a. \tag{4.34}$$

Finally, we need to establish the concavity of ξ. Since this property does not hold for ξ^* we define ξ as the supremum of the convex hull of ξ^*. Let ξ be the upper convex envelope of ξ^*, i.e.,

$$
\begin{aligned}
\xi(a) &= \sup\{b : (a, b) \in L\}, &\text{(4.35a)} \\
L &= \mathrm{conv}\{(a, b) : a \in [0, 1], b \in (0, \xi^*(a))\}. &\text{(4.35b)}
\end{aligned}
$$

Then ξ is concave and from the definition of ξ

$$E[\lambda(\phi(f_{t-1}(w), Y_t, S_t, s_{t-1}))|s_{t-1}] \leq \xi(\lambda(f_{t-1}(w))). \tag{4.36}$$

For any $a \in (0, 1]$, there must exist some constant $\alpha \in [0, 1]$ such that $a = \alpha a_0 + (1 - \alpha)a_1$

$$\xi(a) \leq \alpha\xi^*(a_0) + (1 - \alpha)\xi^*(a_1) < \alpha a_0 + (1 - \alpha)a_1 = a, \tag{4.37}$$

where we used the definition of he upper convex envelope in the first inequality. Since

ξ is nonnegative, it is contraction.

4.4 Proof of Proposition 4.5

Using the symmetry of λ, we can write

$$\begin{aligned} & V_\lambda(F^u_{t'+1,(1+\alpha)t}) \\ & = \int_0^{W^*_{t',t}} \lambda(F^u_{t'+1,(1+\alpha)t}(w))dw + \int_{W^*_{t',t}}^1 \lambda(1 - F^u_{t'+1,(1+\alpha)t}(w))dw, \end{aligned} \tag{4.38}$$

$$\begin{aligned} & \max_{1\le t'\le t} V_\lambda(F^u_{t'+1,(1+\alpha)t}) \\ & = \max_{1\le t'\le t} \Big[\int_0^{W^*_{t',t}} \lambda(F^u_{t'+1,(1+\alpha)t}(w))dw + \int_{W^*_{t',t}}^1 \lambda(1 - F^u_{t'+1,(1+\alpha)t}(w))dw \Big], \end{aligned} \tag{4.39}$$

where $W^*_{t',t}$ is the unique solution of $F^u_{t'+1,(1+\alpha)t}(w) = 0.5$. Then, we have

$$\begin{aligned} & P(\max_{1\le t'\le t} F^u_{t'+1,(1+\alpha)t}(W^*_{t',t} - \delta) > \nu) \\ & \le P(\max_{1\le t'\le t} \lambda(F^u_{t'+1,(1+\alpha)t}(W^*_{t',t} - \delta)) > \nu) && \text{(4.40a)} \\ & \le P(\max_{1\le t'\le t} \int_{W^*_{t',t}-\delta}^{W^*_{t',t}} \lambda(F^u_{t'+1,(1+\alpha)t}(w))dw > \nu\delta) && \text{(4.40b)} \\ & \le P(\max_{1\le t'\le t} V_\lambda(F^u_{t'+1,(1+\alpha)t})) > \nu\delta). && \text{(4.40c)} \end{aligned}$$

Similarly,

$$\begin{aligned} & P(\max_{1\le t'\le t} \Big[1 - F^u_{t'+1,(1+\alpha)t}(W^*_{t',t} + \delta)\Big] < 1 - \nu) \\ & \le P(\max_{1\le t'\le t} V_\lambda(F^u_{t'+1,(1+\alpha)t})) > \nu\delta). \end{aligned} \tag{4.41}$$

For any $\eta \in (0, 0.5)$

$$P\bigg(\max_{1 \le t' \le t} \Big[\int_0^{F_{t'}(W)} F^u_{t'+1,(1+\alpha)t}(w)dw + \int_{F_{t'}(W)}^1 (1 - F^u_{t'+1,(1+\alpha)t}(w))dw \Big] > \nu \bigg) \tag{4.42a}$$

$$\le P\bigg(\max_{1 \le t' \le t} \Big[\int_0^{W^*_{t',t}} F^u_{t'+1,(1+\alpha)t}(w)dw + \int_{W^*_{t',t}}^1 (1 - F^u_{t'+1,(1+\alpha)t}(w))dw \Big] > \nu/2 \bigg) + P\bigg(\max_{1 \le t' \le t} |F_{t'}(W) - W^*_{t',t}| > \nu/2 \bigg) \tag{4.42b}$$

$$\le P\bigg(\max_{1 \le t' \le t} V_\lambda(F^u_{t'+1,(1+\alpha)t}) > \nu/2 \bigg) + P\bigg(\Big\{ F_{(1+\alpha)t}(W) < \max_{1 \le t' \le t} F^u_{t'+1,(1+\alpha)t}(W^*_{t',t} - \nu/2) \Big\} \cup \Big\{ F_{(1+\alpha)t}(W) > \min_{1 \le t' \le t} F^u_{t'+1,(1+\alpha)t}(W^*_{t',t} + \nu/2) \Big\} \bigg) \tag{4.42c}$$

$$\le P\bigg(\max_{1 \le t' \le t} V_\lambda(F^u_{t'+1,(1+\alpha)t}) > \nu/2 \bigg) + P\bigg(F_{(1+\alpha)t}(W) \notin (\eta, 1-\eta) \bigg) + P\bigg(\max_{1 \le t' \le t} F^u_{t'+1,(1+\alpha)t}(W^*_{t',t} - \nu/2) > \eta \bigg) + P\bigg(\max_{1 \le t' \le t} [1 - F^u_{t'+1,(1+\alpha)t}(W^*_{t',t} + \nu/2)] < 1 - \eta \bigg) \tag{4.42d}$$

$$\le P(\max_{1 \le t' \le t} V_\lambda(F^u_{t'+1,(1+\alpha)t})) > \nu/2) + 2\eta + 2P(\max_{1 \le t' \le t} V_\lambda(F^u_{t'+1,(1+\alpha)t})) > \nu\eta/2). \tag{4.42e}$$

Thus,

$$
\begin{aligned}
&P\bigg(\max_{1\le t'\le t} F^u_{t'+1,(1+\alpha)t}\Big(F_{t'}(W)-\delta\Big) > \epsilon\bigg) \\
&\le P\bigg(\max_{1\le t'\le t} \int_{F_{t'}(W)-\delta}^{F_{t'}(W)} F^u_{t'+1,(1+\alpha)t}(w)dw > \delta\epsilon\bigg) && \text{(4.43a)} \\
&\le P\bigg(\max_{1\le t'\le t} \int_{0}^{F_{t'}(W)} F^u_{t'+1,(1+\alpha)t}(w)dw > \delta\epsilon\bigg) && \text{(4.43b)} \\
&\le P\bigg(\max_{1\le t'\le t} \Big[\int_{0}^{F_{t'}(W)} F^u_{t'+1,(1+\alpha)t}(w)dw + \int_{F_{t'}(W)}^{1} (1-F^u_{t'+1,(1+\alpha)t}(w))dw\Big] \\
&\qquad > \delta\epsilon\bigg) && \text{(4.43c)} \\
&\le P(\max_{1\le t'\le t} V_\lambda(F^u_{t'+1,(1+\alpha)t})) > \delta\epsilon/2) \\
&\qquad + 2\eta + 2P(\max_{1\le t'\le t} V_\lambda(F^u_{t'+1,(1+\alpha)t})) > \delta\epsilon\eta/2). && \text{(4.43d)}
\end{aligned}
$$

From Lemma 4.3, we have for any $\nu > 0$

$$
\lim_{t\to\infty} P(\max_{1\le t'\le t} V_\lambda(F^u_{t'+1,(1+\alpha)t})) > \nu) = 0. \tag{4.44}
$$

Setting $\eta = \sqrt{\sup_{a\in[0,1]} \xi^t(a)}/(\delta\epsilon)$, together with Lemma 4.4 completes the proof of the first assertion of the proposition. The proof of the second assertion is similar.

4.5 Proof of Lemma 4.7

First we show that $I(W; Y_{t+1}, S_{t+1}|Y^t, S^t) = E\{\log \frac{dF_{t+1}(W)}{dF_t(W)}|\}$.

$$
\begin{aligned}
&I(W; Y_{t+1}, S_{t+1}|Y^t, S^t) \\
&\quad = H(Y_{t+1}, S_{t+1}|Y^t, S^t) - H(Y_{t+1}, S_{t+1}|W, Y^t, S^t) && \text{(4.45a)} \\
&\quad = E\left[\log \frac{1}{P(Y_{t+1}, S_{t+1}|Y^t, S^t)}\right] \\
&\qquad + E\bigg[\log \Big\{Q'(Y_{t+1}|S_{t+1}, e(F_t(W), S_t))Q(S_{t+1}|S_t)\Big\}\bigg] && \text{(4.45b)} \\
&\quad = E\bigg[\log \bigg\{\frac{\Big(Q'(Y_{t+1}|S_{t+1}, e(F_t(W), S_t))Q(S_{t+1}|S_t)\Big)}{\Big(P(Y_{t+1}, S_{t+1}|Y^t, S^t)\Big)}\bigg\}\bigg] && \text{(4.45c)} \\
&\quad = E\bigg[\log \bigg\{\frac{Q'(Y_{t+1}|S_{t+1}, e(F_t(W), S_t))}{\hat{P}(Y_{t+1}|S_{t+1}, S_t)}\bigg\}\bigg] && \text{(4.45d)} \\
&\quad = E\left[\log \frac{dF_{t+1}(W)}{dF_t(W)}\right], && \text{(4.45e)}
\end{aligned}
$$

where the last equality is due to (4.10c).

Now we show that $I(W; Y_t, S_t|Y^{t-1}, S^{t-1}) = C$. Note that for a given Y^{t-1}, S^{t-1}, $X_t = F^{-1}_{\hat{P}(\cdot|S_{t-1})}(F_{t-1}(W))$ is $\hat{P}(\cdot|S_{t-1})$ distributed and hence is independent of Y^{t-1}, S^{t-2}. Then,

$$
\begin{aligned}
&I(W; Y_t, S_t|Y^{t-1}, S^{t-1}) \\
&\quad = H(Y_t, S_t|Y^{t-1}, S^{t-1}) - H(Y_t, S_t|Y^{t-1}, S^{t-1}, W) && \text{(4.46a)} \\
&\quad \overset{(a)}{=} H(Y_t, S_t|Y^{t-1}, S^{t-1}) - H(Y_t, S_t|Y^{t-1}, S^{t-1}, W, X_t) && \text{(4.46b)} \\
&\quad \overset{(b)}{=} H(Y_t, S_t|S_{t-1}) - H(Y_t, S_t|S_{t-1}, X_t) && \text{(4.46c)} \\
&\quad = I(X_t; S_t, Y_t|S_{t-1}) && \text{(4.46d)} \\
&\quad = I(X_t; Y_t|S^t_{t-1}) && \text{(4.46e)} \\
&\quad = C, && \text{(4.46f)}
\end{aligned}
$$

where (a) is due to the fact that X_t is a function of Y^{t-1}, S^{t-1}, W; (b) is due to the channel characteristics and (4.6b); and the last equation is due to the fact that the channel input sequence for the PMS has distribution which is capacity-achieving.

4.6 Proof of Lemma 4.10

Let $\Theta_t \stackrel{\text{def}}{=} F_t(W)$ and $P_{Y_{t+1}|\Theta_t, S_t^{t+1}} = Q'_{Y_{t+1}|S_{t+1}, e(\Theta_t, S_t)}$. For any $\epsilon > 0$, define ${}^{-}P^{\epsilon}_{Y_{t+1}|\Theta_t, S_t^{t+1}}$ to be

$$ {}^{-}P^{\epsilon}_{Y_{t+1}|\Theta_t, S_t^{t+1}}(y, |\theta, s, s') = \inf_{\theta - \epsilon < \theta' < \theta} P_{Y_{t+1}|\Theta_t, S_t^{t+1}}(y|\theta', s, s'). \tag{4.47} $$

Similarly, define ${}^{+}P^{\epsilon}_{Y_{t+1}|\Theta_t, S_t^{t+1}}$ to be

$$ {}^{+}P^{\epsilon}_{Y_{t+1}|\Theta_t, S_t^{t+1}}(y|\theta, s, s') = \inf_{\theta < \theta' < \theta + \epsilon} P_{Y_{t+1}|\Theta_t, S_t^{t+1}}(y|\theta, s, s'). \tag{4.48} $$

From now on we prove the first assertion of lemma, the second assertion follows in a similar way. Define

$$ C_{\epsilon}^{-} = E\left[\log \left\{ \frac{{}^{-}P^{\epsilon}(Y_{t+1}|\Theta_t, S_t^{t+1})}{P(Y_{t+1}|S_t^{t+1})} \right\} \right] \tag{4.49a} $$

$$ = \int_{\theta_t} \sum_{y_{t+1}, s_t^{t+1}} P(y_{t+1}|\theta_t, s_t^{t+1}) P(\theta_t, s_t^{t+1}) \log \left\{ \frac{{}^{-}P^{\epsilon}(y_{t+1}|\theta_t, s_t^{t+1})}{P(y_{t+1}|s_t^{t+1})} \right\}. \tag{4.49b} $$

Note that

$$ C = E\left[\log \left\{ \frac{P(Y_{t+1}|\Theta_t, S_t^{t+1})}{P(Y_{t+1}|S_t^{t+1})} \right\} \right] \tag{4.50a} $$

$$ = \int_{\theta_t} \sum_{y_{t+1}, s_t^{t+1}} P(y_{t+1}|\theta_t, s_t^{t+1}) P(\theta_t, s_t^{t+1}) \log \left\{ \frac{P(y_{t+1}|\theta_t, s_t^{t+1})}{P(y_{t+1}|s_t^{t+1})} \right\}. \tag{4.50b} $$

Then we have,

$$C - C_\epsilon^- = D(P_{Y_{t+1}|\Theta_t, S_t^{t+1}} || {}^-P^\epsilon_{Y_{t+1}|\Theta_t, S_t^{t+1}} | P_{\Theta_t, S_t^{t+1}}) \geq 0. \tag{4.51}$$

Assume that

$$\inf_{\epsilon > 0} D(P_{Y_{t+1}|\Theta_t, S_t^{t+1}} || {}^-P^\epsilon_{Y_{t+1}|\Theta_t, S_t^{t+1}} | P_{\Theta_t, S_t^{t+1}}) < \infty. \tag{4.52}$$

Then, $-\infty < C_\epsilon^- \leq C$. Therefore, $P_{Y_{t+1}, S_{t+1}\Theta_t, S_t} \log \left\{ \frac{{}^-P^\epsilon_{Y_{t+1}|\Theta_t, S_t^{t+1}}}{P_{Y_{t+1}|S_t^{t+1}}} \right\}$ is finitely integrable and converges to $P_{Y_{t+1}, S_{t+1}, \Theta_t, S_t} \log \left\{ \frac{P_{Y_{t+1}|\Theta_t, S_t^{t+1}}}{P_{Y_{t+1}|S_t^{t+1}}} \right\}$ a.e. in a monotonically nondecreasing fashion as $\epsilon \to 0$. Hence, by the monotone convergence theorem,

$$\lim_{\epsilon \to 0} C_\epsilon = C, \tag{4.53}$$

and there exists ϵ'_δ such that for all $\epsilon \leq \epsilon'_\delta$

$$C_\epsilon^- > C - \delta/2. \tag{4.54}$$

Furthermore, we can apply the strong law of large numbers for Markov chains to get

$$\lim_{t \to \infty} \frac{1}{t} \sum_{s=0}^{t-1} \log \left\{ \frac{{}^-P^\epsilon(Y_{s+1}|\Theta_s, S_s^{s+1})}{P(Y_{s+1}|S_s^{s+1})} \right\} = C_\epsilon^- \qquad \text{a.s.} \tag{4.55}$$

Note that if $F_s(W) - F_s(W - 2^{-tR}) < \epsilon$ for $0 \leq s \leq t-1$

$$\inf_{W - 2^{-tR} < w < W} \frac{1}{t} \log dF_t(w) \geq \frac{1}{t} \sum_{s=0}^{t-1} \log \inf_{F_s(W - 2^{-tR}) < \theta_s < F_s(W)} \frac{P(Y_{s+1}|\theta_s, S_s^{s+1})}{P(Y_{s+1}|S_s^{s+1})} \tag{4.56a}$$

$$\geq \frac{1}{t} \sum_{s=0}^{t-1} \log \left\{ \frac{{}^-P^\epsilon(Y_{s+1}|\Theta_s, S_s^{s+1})}{P(Y_{s+1}|S_s^{s+1})} \right\}. \tag{4.56b}$$

Therefore,

$$
\begin{aligned}
&P\biggl(\bigcap_{s=0}^{t-1}\{F_s(W) - F_s(W-2^{-tR}) < \epsilon\}\biggr) \\
&\le P\biggl(\bigcap_{s=0}^{t-1}\{F_s(W) - F_s(W-2^{-tR}) < \epsilon\} \\
&\qquad \bigcap\{\frac{1}{t}\sum_{s=0}^{t-1}\log\left\{\frac{{}^{-}P^{\epsilon}(Y_{s+1}|\Theta_s, S_s^{s+1})}{P(Y_{s+1}|S_s^{s+1})}\right\} \ge C-\delta/2\}\biggr) \\
&\quad + P\biggl(\frac{1}{t}\sum_{s=0}^{t-1}\log\left\{\frac{{}^{-}P^{\epsilon}(Y_{s+1}|\Theta_s, S_s^{s+1})}{P(Y_{s+1}|S_s^{s+1})}\right\} < C-\delta/2\biggr) &&\text{(4.57a)}\\
&\overset{(a)}{\le} P\biggl(\inf_{W-2^{-tR}<w<W}\frac{1}{t}\log dF_t(w) \ge C-\delta/2\biggr) \\
&\quad + P\biggl(\frac{1}{t}\sum_{s=0}^{t-1}\log\left\{\frac{{}^{-}P^{\epsilon}(Y_{s+1}|\Theta_s, S_s^{s+1})}{P(Y_{s+1}|S_s^{s+1})}\right\} < C-\delta/2\biggr) &&\text{(4.57b)}\\
&\overset{(b)}{\le} P\biggl(\inf_{W-2^{-tR}<w<W}\frac{1}{t}\log dF_t(w) \ge C-\delta/2\biggr) + o(1) &&\text{(4.57c)}\\
&\le P\biggl(\int_{W-2^{-tR}}^{W} dF_t(w) \ge 2^{t(C-R-\delta/2)}\biggr) + o(1) &&\text{(4.57d)}\\
&\le P\biggl(\int_{W-2^{-tR}}^{W} dF_t(w) \ge 2^{t\delta/2}\biggr) + o(1) &&\text{(4.57e)}\\
&\overset{(c)}{\to} 0, &&\text{(4.57f)}
\end{aligned}
$$

where where (a) is due to (4.56), (b) is due to (4.55), and (c) is from the fact that $\int_{W-2^{-tR}}^{W} dF_t(w) \le 1$.

4.7 Proof of Proposition 4.11

For every $\epsilon > 0$ we have

$$\begin{aligned}
&P(F_{(1+\alpha)t}(W - 2^{-tR}) > \delta) \\
&\leq P\Bigg(\bigcap_{t'=1}^{t} \Big\{F_{t'}(W - 2^{-tR}) > F_{t'}(W) - \epsilon\Big\}\Bigg) \\
&\quad + P\Bigg(\Big\{F_{(1+\alpha)t}(W - 2^{-tR}) > \delta)\Big\} \\
&\qquad\qquad \cap \bigcup_{t'=1}^{t} \Big\{F_{t'}(W - 2^{-tR}) \leq F_{t'}(W) - \epsilon\Big\}\Bigg).
\end{aligned} \tag{4.58}$$

Regarding the first quantity we have

$$P\Bigg(\bigcap_{t'=1}^{t} \Big\{F_{t'}(W - 2^{-tR}) > F_{t'}(W) - \epsilon\Big\}\Bigg) \overset{(a)}{\to} 0, \tag{4.59}$$

where (a) is from Lemma 4.10.

Consider an event $\bigcup_{t'=1}^{t} \Big\{F_{t'}(W - 2^{-tR}) \leq F_{t'}(W) - \epsilon\Big\}$. This event is equivalent to $\bigcup_{t'=1}^{t} \{W - 2^{-tR} \leq W'\}$ where $W' = F_{t'}^{-1}(F_{t'}(W) - \epsilon)$. Then by the monotonicity of cdf it is equivalent to $\bigcup_{t'=1}^{t} \{F_t(W - 2^{-tR}) \leq F_t(W')\}$ which again is equivalent to $\Big\{F_t(W - 2^{-tR}) \leq \max_{1 \leq t' \leq t} F_t\Big(W'\Big)\Big)\Big\}$.

Therefore, regarding the second quantity we have

$$\begin{aligned}
&P\Bigg(\Big\{F_{(1+\alpha)t}(W - 2^{-tR}) > \delta)\Big\} \\
&\qquad \cap \bigcup_{t'=1}^{t} \Big\{F_{t'}(W - 2^{-tR}) \leq F_{t'}(W) - \epsilon\Big\}\Bigg) \\
&= P\Bigg(\Big\{F_{(1+\alpha)t}(W - 2^{-tR}) > \delta)\Big\} \\
&\qquad \cap \Big\{F_t(W - 2^{-tR}) \leq \max_{1 \leq t' \leq t} F_t\Big(F_{t'}^{-1}\Big(F_{t'}(W) - \epsilon\Big)\Big)\Big\}\Bigg)
\end{aligned} \tag{4.60a}$$

$$\overset{(a)}{\leq} P\Bigg(\Big\{F_{(1+\alpha)t}(W - 2^{-tR}) > \delta)\Big\}$$
$$\cap \Big\{F_{(1+\alpha)t}(W - 2^{-tR}) \leq \max_{1 \leq t' \leq t} F_{(1+\alpha)t}\Big(F_{t'}^{-1}\Big(F_{t'}(W) - \epsilon\Big)\Big)\Big\}\Bigg) \tag{4.60b}$$
$$= P\Bigg(\max_{1 \leq t' \leq t} F_{(1+\alpha)t}\Big(F_{t'}^{-1}\Big(F_{t'}(W) - \epsilon\Big)\Big) > \delta\Bigg) \tag{4.60c}$$
$$\overset{(b)}{=} P\Bigg(\max_{1 \leq t' \leq t} F^{u}_{t'+1,(1+\alpha)t}\Big(F_{t'}(W) - \epsilon\Big) > \delta\Bigg) \overset{(c)}{\to} 0. \tag{4.60d}$$

where (a) is from the monotonicity of cdf, (b) is due to the PMS update, and (c) is from Proposition 4.5. $P(F_{(1+\alpha)t}(W + 2^{-tR}) < 1 - \delta)$ can also be evaluated in a similar way. Note that the fixed-point free nature of the channel is required for zero-rate result, but it can be easily verified that ergodicity of invariant distribution of $(F_t(W), S_t^{t+1}, Y_{t+1})_t$ implies fixed-point free channel as in Lemma 14 of [50].

CHAPTER 5

A capacity-achieving posterior matching scheme: ISI case

In this chapter we describe a transmission scheme which achieves the capacity of the finite-state channel with unit feedback delay. The scheme which achieves the capacity of the finite-state channel with arbitrary feedback delay can be thought of as a slight generalization of the scheme presented here. With unit feedback delay, the capacity expression in (3.16) becomes

$$C = \sup_{\{P(X|X',S',\Theta)\}_{X',S',\Theta}} I(X, X'; S, Y|S', \Theta) \tag{5.1}$$

where $\Theta \in \mathcal{P}(\mathcal{X})$, and the mutual information is evaluated using the joint measure

$$\begin{aligned} &P(Y, S, X, X', S', d\Theta) \\ &= Q'(Y|S, X)Q(S|S', X')P(X|X', S', \Theta)\Theta(X')P(S', d\Theta). \end{aligned} \tag{5.2a}$$

The distribution $P(S, d\Theta)$ is the solution of the equation

$$\begin{aligned}
&P(S, d\Theta') \\
&= \int_{S',\Theta} P(S', d\Theta) \sum_{Y} \delta_{\omega(\Theta, P(X|X',S',\Theta),Y,S,S')}(\Theta') \\
&\quad \times \sum_{X} Q'(Y|X,S) \sum_{X'} Q(S|S',X') P(X|X',S',\Theta)\Theta(X'). \qquad (5.2b)
\end{aligned}$$

Let W be a random message point uniformly distributed over the unit interval. A transmission scheme is a sequence of functions $\{\tilde{e}_t : [0,1) \times \mathcal{Y}^{t-1} \times \mathcal{S}^{t-1} \to \mathcal{X}\}_{t=1}^{\infty}$ such that

$$X_t = \tilde{e}_t(W, Y^{t-1}, S^{t-1}). \qquad (5.3)$$

We now describe a simple sequential transmission scheme. We assume that the capacity achieving distributions $\{\hat{P}(X'|X,S,\Theta)\}_{X,S,\Theta}$ in (5.1) have been found for all values of (X, S, Θ) and the corresponding steady-state distribution on X, $\hat{P}(X|X,S,\Theta)$ and on Y_t, S_t conditioned on S_{t-1}, Θ_{t-1},
$\hat{P}(y,s|s',\theta) = \sum_{x,x'} Q'(y|x,s)Q(s|s',x')(\hat{P}(x|x',s',\theta)\theta(x')$ have been evaluated.

Define the random variable $F_t \in \mathcal{F}$ as $F_t(w) \stackrel{\text{def}}{=} F(w|Y^t, S^t)$, where $F(\cdot|Y^t, S^t)$ is the a-posteriori cdf of W conditioned on Y^t, S^t, and $\mathcal{F}$ is the set of all valid cdfs over $[0,1)$. Also define $\Theta_t \in \mathcal{P}(\mathcal{X})$ with $\Theta_t(x_t) \stackrel{\text{def}}{=} P(x_t|S^t, Y^t)$. At time $t = 1$ the channel input X_1 is generated as

$$X_1 = F^{-1}_{\hat{P}(\cdot|S_0)}(F_0(W)) \qquad (5.4a)$$

$$= e(F_0(W), S_0) \qquad (5.4b)$$

where the inverse cdf $F^{-1}(y) \stackrel{\text{def}}{=} \inf\{x : F(x) \geq y\}$, $F_0 = Uniform[0,1)$ and $S_0 \sim P(S_0)$.

Define $F'_{x,t}(w) = P(W \leq w|X_t = x, S^t, Y^t)$. Then $F_t(w) = \sum_{x\in\mathcal{X}} \Theta_t(x)F'_{x,t}(w)$.

Let $F'_t(w) = (F'_{0,t}(w), ..., F'_{K_X-1,t}(w))$. Note that $F'_1(w)$ can be found from $F_0(w)$ and S_0.

Assuming that the channel and state transition probabilities $Q'(y|x,s)$ and $Q(s'|s,x)$ are non-zero for all x, y, s, s', the recursion (5.40) guarantees that for every realization of the random variables of interest, F_t will always have a pdf; in addition the pdf will be non-zero everywhere in $(0, 1]$.

The channel input X_t for $t \geq 2$ is generated as

$$X_t = F^{-1}_{\hat{P}(\cdot|X_{t-1},S_{t-1},\Theta_{t-1})}(F'_{X_{t-1},t-1}(W)) \tag{5.5a}$$

$$= x, \quad \sum_{i=0}^{x-1} \hat{P}(i|X_{t-1}, S_{t-1}, \Theta_{t-1}) < F'_{X_{t-1},t-1}(W) \leq \sum_{i=0}^{x} \hat{P}(i|X_{t-1}, S_{t-1}, \Theta_{t-1})$$

$$x = 0, \ldots, K_X - 1 \tag{5.5b}$$

$$\stackrel{\text{def}}{=} e(F'_{X_{t-1},t-1}(W), \hat{P}(\cdot|X_{t-1}, S_{t-1}, \Theta_{t-1})), \tag{5.5c}$$

For $t \geq 2$, Θ_t are updated as

$$\Theta_t = \omega(\Theta_{t-1}, \hat{P}(\cdot|\cdot, S_{t-1}, \Theta_{t-1}), Y_t, S_t, S_{t-1}) \tag{5.6}$$

where ω is given in (3.34).

At the receiver, the message estimate is obtained as

$$\hat{W}_t = d(F_t, 2^{-Rt}/2), \tag{5.7}$$

where the message estimate function $d(F, \epsilon)$ is defined as

$$d(F, \epsilon) = \arg\max_w \{F(w + \epsilon) - F(w - \epsilon)\}. \tag{5.8}$$

Lemma 5.1. *Under the PMS the following is true for θ_{t-1} corresponding to s^{t-1}, y^{t-1}.*

$$P(X_t|X_{t-1}, Y^{t-1}, S^{t-1}) = \hat{P}(X_t|X_{t-1}, S_{t-1}, \Theta_{t-1}) \tag{5.9}$$

$$P(Y_t, S_t|Y^{t-1}, S^{t-1}) = \hat{P}(Y_t, S_t|S_{t-1}, \Theta_{t-1}) \tag{5.10}$$

$$F'_t = \phi(F'_{t-1}, Y_t, S_t, S_{t-1}, \Theta_{t-1}). \tag{5.11}$$

Proof. See 5.2. □

With slight abuse of notation we also say that F_t is updated as

$$F_t = \phi(F'_{t-1}, Y_t, S_t, S_{t-1}, \Theta_{t-1}) \tag{5.12}$$

because F_t is determined by F'_t and Θ_t.

Note that $f'_t(a)$ is a function of f'_{t-1} only through $f'_{t-1}(a)$. In the following we will also use the notation $f'_t(a) = \phi(f'_{t-1}, y_t, s_t, s_{t-1}, \theta_{t-1})(a) = \phi(f'_{t-1}(a), y_t, s_t, s_{t-1}, \theta_{t-1})$. Observe also from (5.5) that the transmitted symbol X_t is a function of W and F'_{t-1} only through the quantity $F'_{t-1}(W)$. This has important implications for the analysis of the PMS scheme.

5.1 Achievability Result

Let $\hat{W}_t$ be the message point estimate at the receiver at time t. Then, a transmission scheme achieves rate R if

$$\lim_{t\to\infty} P(|W - \hat{W}_t| > 2^{-tR}) = 0. \tag{5.13}$$

In particular, we say that a transmission schemes achieves zero rate if

$$\forall \epsilon > 0 \qquad \lim_{t\to\infty} P(|W - \hat{W}_t| > \epsilon) = 0. \tag{5.14}$$

5.1.1 Zero Rate Result

For a cdf $h : [0,1] \to [0,1]$ define a Lyapunov function V_λ as follows.

$$V_\lambda(h) = \int_0^1 \lambda(h(w))dw, \tag{5.15}$$

where $\lambda : [0,1] \to [0,1]$ is onto, strictly concave and symmetric about 0.5. This definition implies that $\lambda(x)$ is 0 at $x = 0,\ 1$ and 1 at $x = 1/2$. Furthermore, for a cdf $F \in \mathcal{F}$, $V_\lambda(F)$ is small if F resembles a step function (it is exactly 0 for a step function). A function $\xi : [0,1] \to [0,1]$ is called contraction if it is nonnegative, concave, and $\xi(x) < x$ for $x \in (0,1)$.

Definition 5.2. A channel is called fixed-point free if for any $f_t'(w), s_t, \theta_t$

$$P\Big(\phi(f_t'(w), Y_{t+1}, S_{t+1}, s_t, \theta_t) = f_t(w)\Big) < 1. \tag{5.16}$$

Lemma 5.3. *If the channel is fixed-point free, then for $\epsilon > 0$ and for all $f \in \mathcal{F}$,*

$$\lim_{t\to\infty} P(V_\lambda(F_t) > \epsilon | F_0 = f) = 0. \tag{5.17}$$

Proof. See 5.3. □

The intuitive interpretation of the above lemma is that the probability of having an F_t that does not resemble a step function is zero at the limit of large t.

Define the function $H_t : [0,1] \to \mathcal{X}$ as follows. Since there is no w and w' such that $F_1(w) = F_1(w')$ with $w \neq w'$ (that would imply that $dF_1(w)$ is zero in an interval), we can set $H_1(F_1(w)) = F^{-1}_{\hat{P}(x_1|S_0)}(F_0(w))$. For $t \geq 2$,

$$H_t(F_t(w)) = e(F'_{H_{t-1}(F_{t-1}(w)),t-1}(w), H_{t-1}(F_{t-1}(w)), S_{t-1}, \Theta_{t-1}), \tag{5.18a}$$

which implies that H_t can be recursively updated as

$$H_t = \phi'(H_{t-1}, S_{t-1}, F'_{t-1}, \Theta_{t-1}, S_t, Y_t). \tag{5.18b}$$

With slight abuse of notation, we say F_t is updated as

$$F_t = \phi(F_{t-1}, Y_t, S_t, S_{t-1}, \Theta_{t-1}, H_{t-1}) \tag{5.19}$$

because F'_{t-1} is determined by F_{t-1} and H_{t-1} as shown below.

$$dF_{x,t-1}(w) = P(dw|X_{t-1} = x, Y^{t-1}) \tag{5.20}$$

$$= P(X_{t-1} = x|W = w, Y^{t-1})P(dw|Y^{t-1}) \tag{5.21}$$

$$= \delta_{H_{t-1}(F_{t-1}(w))}(x)dF_{t-1}(w). \tag{5.22}$$

Furthermore we also say H_t is updated as

$$H_t = \phi'(H_{t-1}, S_{t-1}, F_{t-1}, \Theta_{t-1}) \tag{5.23}$$

from the same reason.

For any $t_2 > t_1 > 0$ we can write F_{t_2} as a function of F_{t_1-1} and the quantities $Y^{t_2}_{t_1}, S^{t_2}_{t_1-1}, \Theta^{t_2-1}_{t_1-1}, H^{t_2-1}_{t_1-1}$ through a repeated application of the ϕ recursion, i.e., $F_{t_2} \stackrel{\text{def}}{=} \phi_{t_2-t_1}(F_{t_1-1}, Y^{t_2}_{t_1}, S^{t_2}_{t_1-1}, \Theta^{t_2-1}_{t_1-1}, H^{t_2-1}_{t_1-1})$. Let $F^f_{t_1,t_2}$ be the random variable defined as $F^f_{t_1,t_2} \stackrel{\text{def}}{=} \phi_{t_2-t_1}(f, Y^{t_2}_{t_1}, S^{t_2}_{t_1-1}, \Theta^{t_2-1}_{t_1-1}, H^{t_2-1}_{t_1-1})$. Clearly $F_t = F^u_{1,t}$, where u denotes the uniform distribution over $(0,1)$. In addition, due to the recursion implied by the PMS, we will denote $F_{t_2}(a) \stackrel{\text{def}}{=} \phi_{t_2-t_1}(F_{t_1-1}, Y^{t_2}_{t_1}, S^{t_2}_{t_1-1}, \Theta^{t_2-1}_{t_1-1}, H^{t_2-1}_{t_1-1})(a) \stackrel{\text{def}}{=}$

$\phi_{t_2-t_1}(F_{t_1-1}(a), Y_{t_1}^{t_2}, S_{t_1-1}^{t_2}, \Theta_{t_1-1}^{t_2-1}, H_{t_1-1}^{t_2-1})$ with some notational abuse. With the above notation, the function $\phi_{t_2-t_1}(\cdot, Y_{t_1}^{t^2}, S_{t_1-1}^{t_2}, \Theta_{t_1-1}^{t_2-1}, H_{t_1-1}^{t_2-1})$ is monotonically increasing. We now prove the following lemma which is a stronger version of Lemma 5.3.

Lemma 5.4. *If the channel is fixed-point free, then for $\epsilon > 0$,*

$$\lim_{t\to\infty} P(\max_{1\le t'\le t} V_\lambda(F^u_{t'+1,(1+\alpha)t}) > \epsilon) = 0. \tag{5.24}$$

Proof. Let $V^*_{\lambda,t_1,t_2} = \sup_f V_\lambda(F^f_{t_1,t_2})$. Note that V^*_{λ,t_1,t_2} is a deterministic function of $Y_{t_1}^{t_2}, S_{t_1-1}^{t_2}, \Theta_{t_1-1}^{t_2-1}, H_{t_1-1}^{t_2-1}$, hence there exists a sequence of $\{f_{k,Y_{t_1}^{t_2},S_{t_1-1}^{t_2},\Theta_{t_1-1}^{t_2-1},H_{t_1-1}^{t_2-1}}\}_{k=1}^{\infty}$ such that

$$\begin{aligned}
V^*_{\lambda,t_1,t_2} &= \lim_{k\to\infty} V_\lambda(F_{t_1,t_2}^{f_{k,Y_{t_1}^{t_2},S_{t_1-1}^{t_2},\Theta_{t_1-1}^{t_2-1},H_{t_1-1}^{t_2-1}}}) && \text{(5.25a)}\\
&= \lim_{k\to\infty} V_\lambda\left(F_{t_1+1,t_2}^{\phi\left(f_{k,Y_{t_1}^{t_2},S_{t_1-1}^{t_2},\Theta_{t_1-1}^{t_2-1},H_{t_1-1}^{t_2-1}},Y_{t_1},S_{t_1-1}^{t_1},\Theta_{t_1-1},H_{t_1-1}\right)}\right) && \text{(5.25b)}\\
&\le \sup_f V_\lambda(F^f_{t_1+1,t_2}) && \text{(5.25c)}\\
&= V^*_{\lambda,t_1+1,t_2}. && \text{(5.25d)}
\end{aligned}$$

Note that there exists a sequence of cdfs $\{f_{k,Y_2^{\alpha t+1},S_1^{\alpha t+1}\Theta_1^{\alpha t},H_1^{\alpha t}}\}_{k=1}^{\infty}$ such that

$V^*_{\lambda,2,\alpha t+1} \geq V_\lambda(F_{2,\alpha t+1}^{f_{k,Y_2^{\alpha t+1},S_1^{\alpha t+1},\Theta_1^{\alpha t},H_1^{\alpha t}}}) > V^*_{\lambda,2,\alpha t+1} - 1/k$. Therefore,

$$
\begin{aligned}
P(\max_{1\leq t'\leq t} V_\lambda(F^u_{t'+1,(1+\alpha)t} > \epsilon) &\leq P(\max_{1\leq t'\leq t} V^*_{\lambda,t'+1,(1+\alpha)t} > \epsilon) && (5.26a)\\
&= P(V^*_{\lambda,t+1,(1+\alpha)t} > \epsilon) && (5.26b)\\
&= P(\sup_f V_\lambda(F^f_{t+1,(1+\alpha)t}) > \epsilon) && (5.26c)\\
&\overset{(a)}{=} P(\sup_f V_\lambda(F^f_{2,\alpha t+1}) > \epsilon) && (5.26d)\\
&= P(\lim_{k\to\infty} V_\lambda(F_{2,\alpha t+1}^{f_{k,Y_2^{\alpha t+1},S_1^{\alpha t+1}\Theta_1^{\alpha t},H_1^{\alpha t}}}) > \epsilon) && (5.26e)\\
&\overset{(b)}{\leq} P(V_\lambda(F_{2,\alpha t}^{f_{k',Y_2^{\alpha t+1},S_1^{\alpha t+1}\Theta_1^{\alpha t},H_1^{\alpha t}}}) > \epsilon) && (5.26f)\\
&\overset{(c)}{\to} 0, && (5.26g)
\end{aligned}
$$

where (a) is due to the fact that $F^f_{t+1,(1+\alpha)t} = \phi_{\alpha t-1}(f, Y_{t+1}^{(1+\alpha)t}, S_t^{(1+\alpha)t}, \Theta_t^{(1+\alpha)t-1}, H_t^{(1+\alpha)t-1})$, $F^f_{2,\alpha t+1} = \phi_{\alpha t-1}(f, Y_2^{\alpha t+1}, S_1^{\alpha t+1}, \Theta_1^{\alpha t}, H_1^{\alpha t})$, and $(Y_{t+1}^{(1+\alpha)t}, S_t^{(1+\alpha)t}, \Theta_t^{(1+\alpha)t-1}, H_t^{(1+\alpha)t-1})$, $(Y_2^{\alpha t+1}, S_1^{\alpha t+1}, \Theta_1^{\alpha t}, H_1^{\alpha t})$ have the same statistics; (b) is true for $k' > 1/(V^*_{\lambda,2,\alpha t+1} - \epsilon)$; and (c) is due to the fact that Lemma 5.3 holds for any F_0. □

Proposition 5.5. *If the channel is fixed-point free, then for $\epsilon, \delta > 0$*

$$\lim_{t\to\infty} P\left(\max_{1\leq t'\leq t} F^u_{t'+1,(1+\alpha)t}\Big(F_{t'}(W) - \delta\Big) > \epsilon\right) = 0, \tag{5.27}$$

$$\lim_{t\to\infty} P\left(\max_{1\leq t'\leq t} F^u_{t'+1,(1+\alpha)t}\Big(F_{t'}(W) + \delta\Big) < 1-\epsilon\right) = 0. \tag{5.28}$$

Proof. See 5.4. □

The intuition behind the above proof is that for an error to occur, either the cdf F_t does not behave as a step function (first, third and fourth terms in (5.57c)) or the step does not occur at the transmitted message W (second term in (5.57c)).

5.1.2 Rate $R < C$ achievability

Lemma 5.6. $E\{\log \frac{dF_{t+1}(W)}{dF_t(W)}|\} = C$.

Proof. See 5.5. $\square$

Lemma 5.7. $(F'_t(W), \Theta_t, X_t, S_t^{t+1}, Y_{t+1})_t$ *is a Markov chain.*

Proof.

$$\begin{aligned}
&P(df'_t(w), d\theta_t, x_t, s_t^{t+1}, y_{t+1} | f'^{t-1}(w), \theta^{t-1}, x^{t-1}, s'^t, y^t) \\
&\quad = Q'(y_{t+1}|e(f'_t(w), x_t, s_t, \theta_t), s_{t+1}) Q(s_{t+1}|s_t, x_t) \\
&\qquad\quad \times P(df'_t(w), d\theta_t, x_t, s_t | f'^{t-1}(w), \theta^{t-1}, x^{t-1}, s'^t, y^t) \qquad (5.29a)
\end{aligned}$$

$$\begin{aligned}
&= Q'(y_{t+1}|e(f'_t(w), x_t, s_t, \theta_t), s_{t+1}) Q(s_{t+1}|s_t, x_t) \\
&\qquad \times \delta_{\phi(f'_{t-1}(w), y_t, s_t, s'_{t-1}, \theta_{t-1})}(df'_t(w)) \delta_{\omega(\theta_{t-1}, \hat{P}(\cdot|\cdot, s'_{t-1}, \theta_{t-1}), y_t, s_t, s'_{t-1})}(d\theta_t) \\
&\qquad \times \delta_{e(f'_{x_{t-1}, t-1}(w), x_{t-1}, s'_{t-1}, \theta_{t-1})}(x_t) \delta_{s'_t}(s_t) \qquad (5.29b) \\
&= P(df'_t(w), d\theta_t, x_t, s_t^{t+1}, y_{t+1} | f'_{t-1}(w), \theta_{t-1}, x_{t-1}, s'^t_{t-1}, y_t). \qquad (5.29c)
\end{aligned}$$

$\square$

Lemma 5.8. *If a Markov chain* $(F'_t(W), \Theta_t, X_t, S_t^{t+1}, Y_{t+1})_t$ *has ergodic invariant distribution, then*

$$\lim_{t\to\infty} \frac{1}{t} \log dF_t(W) = C \ \ a.s. \qquad (5.30)$$

Proof.

$$\frac{1}{t} \log dF_t(W) = \frac{1}{t} \sum_{s=1}^{t} \log \frac{dF_s(W)}{dF_{s-1}(W)}. \qquad (5.31)$$

If $(F'_t(W), \Theta_t, X_t, S_t^{t+1}, Y_{t+1})_t$ has ergodic invariant distribution, then by the strong law of large numbers for Markov chains

$$\lim_{t\to\infty} \frac{1}{t} \log dF_t(W) = E[\log \frac{dF_t(W)}{dF_{t-1}(W)}] \overset{(a)}{=} C \qquad \text{a.s.} \tag{5.32}$$

where (a) is from Lemma 5.6. □

Lemma 5.9. *$(F'_t, \Theta_t, H_t, S_t^{t+1}, Y_{t+1}, W)_t$ is a Markov chain.*

Proof.

$$\begin{aligned}
&P(df'_t, d\theta_t, dh_t, s_t^{t+1}, y_{t+1}, dw | f'^{t-1}, \theta^{t-1}, h^{t-1}, s'^t, y^t, w') \\
&\quad = Q'(y_{t+1}|e(f'_{h_t(f_t(w')),t}(w'), h_t(f_t(w')), s_t, \theta_t), s_{t+1}) \\
&\qquad \times Q(s_{t+1}|s_t, h_t(f_t(w'))) \\
&\qquad \times P(df'_t, d\theta_t, dh_t, s_t, dw|f'^{t-1}, \theta^{t-1}, h^{t-1}, s'^t, y^t, w') && \text{(5.33a)} \\
&\quad = Q'(y_{t+1}|e(f'_{h_t(f_t(w')),t}(w'), h_t(f_t(w')), s_t, \theta_t), s_{t+1}) \\
&\qquad \times Q(s_{t+1}|s_t, h_t(f_t(w'))) \\
&\qquad \times \delta_{\phi(f'_{t-1}, y_t, s_t, s'_{t-1}, \theta_{t-1})}(df'_t)\delta_{\omega(\theta_{t-1}, \hat{P}(\cdot|\cdot, s'_{t-1}, \theta_{t-1}), y_t, s_t, s'_{t-1})}(d\theta_t) \\
&\qquad \times \delta_{\phi'(h_{t-1}, s'_{t-1}, f'_{t-1}, \theta_{t-1})}(dh_t)\delta_{s'_t}(s_t)\delta_{w'}(dw) && \text{(5.33b)} \\
&\quad = P(df'_t, d\theta_t, dh_t, s_t^{t+1}, y_{t+1}, dw|f'_{t-1}, \theta_{t-1}, h_{t-1}, s'^t_{t-1}, y_t, w'). && \text{(5.33c)}
\end{aligned}$$

□

Lemma 5.10. *If a Markov chain $(F'_t, \Theta_t, H_t, S_t^{t+1}, Y_{t+1}, W)_t$ has ergodic invariant distribution, then for any $\delta > 0$ and rate $R < C - \delta$ there exists $\epsilon' > 0$ so that for* *all* $\epsilon \leq \epsilon'$

$$\lim_{t\to\infty} P\left(\bigcap_{s=0}^{t-1}\{F_s(W) - F_s(W - 2^{-tR}) < \epsilon\}\right) = 0 \tag{5.34a}$$

$$\lim_{t\to\infty} P\left(\bigcap_{s=0}^{t-1}\{F_s(W + 2^{-tR}) - F_s(W) < \epsilon\}\right) = 0. \tag{5.34b}$$

Proof. See 5.6. □

The above lemma guarantees that at some time before t there will be a jump of at least ϵ in the posterior message cdf in the interval of 2^{-tR} around W. Note that assumption of ergodicity of invariant distribution of Markov chain $(F'_t, \Theta_t, H_t, S_t^{t+1}, Y_{t+1}, W)_t$ is stronger than the same assumption on $(F'_t(W), \Theta_t, X_t, S_t^{t+1}, Y_{t+1})_t$ or the corresponding assumption in no ISI case. Latter assumptions are related to recurrence of $F'_t(W)$ or $F_t(W)$ which implies fixed point free nature of the system which is violated only in several pathological cases. Using this lemma we show the main result.

Proposition 5.11. *If a Markov chain* $(F'_t, \Theta_t, H_t, S_t^{t+1}, Y_{t+1}, W)_t$ *has ergodic invariant distribution, then for* $\delta, \alpha > 0$

$$\lim_{t\to\infty} P(F_{(1+\alpha)t}(W - 2^{-tR}) > \delta) = 0, \tag{5.35a}$$

$$\lim_{t\to\infty} P(F_{(1+\alpha)t}(W + 2^{-tR}) < 1 - \delta) = 0. \tag{5.35b}$$

Proof. See 5.7. □

5.2 Proof of Lemma 5.1

$$
\begin{aligned}
&P(X_t|X_{t-1}, S^{t-1}, Y^{t-1}) \\
&\quad \overset{(a)}{=} \int_w \delta_{e(F'_{X_{t-1},t-1}(w), X_{t-1}, S_{t-1}, \Theta_{t-1})}(X_t) dF'_{X_{t-1},t-1}(w) && (5.36a) \\
&\quad \overset{(b)}{=} \int_{w=0}^{w_0^*} \delta_0(X_t) dF'_{X_{t-1},t-1}(w) + \int_{w=w_0^*}^{w_1^*} \delta_1(X_t) dF'_{X_{t-1},t-1}(w) \\
&\qquad + ... + \int_{w=w^*_{K_X-2}}^{1} \delta_{K_X-1}(X_t) dF'_{X_{t-1},t-1}(w) && (5.36b) \\
&\quad = \sum_{x=0}^{K_X-1} \delta_x(X_t) \hat{P}(x|X_{t-1}, S_{t-1}, \Theta_{t-1}) && (5.36c) \\
&\quad = \hat{P}(X_t|X_{t-1}, S_{t-1}, \Theta_{t-1}). && (5.36d)
\end{aligned}
$$

where w_x^* is the solution of $F'_{X_{t-1},t-1}(w) = \sum_{x'=0}^{x} \hat{P}(x'|X_{t-1}, S_{t-1}, \Theta_{t-1})$, (a) comes from the fact that F'_{t-1}, Θ_{t-1} are determined by S^{t-1}, Y^{t-1}, and (b) comes from (5.5), and similarly

$$
\begin{aligned}
&P(Y_t, S_t|Y^{t-1}, S^{t-1}) \\
&\quad = \sum_{x,x'} P(Y_t|Y^{t-1}, S^t, X_t = x) P(S_t|Y^{t-1}, S^{t-1}, X_t = x, X_{t-1} = x') \\
&\qquad \times (X_t = x|Y^{t-1}, S^{t-1}, X_{t-1} = x') P(X_{t-1} = x'|Y^{t-1}, S^{t-1}) && (5.37a) \\
&\quad = \sum_{x,x'} Q'(Y_t|S_t, x) Q(S_t|S_{t-1}, x') \hat{P}(x|x', S_{t-1}, \Theta_{t-1}) \Theta_{t-1}(x') && (5.37b) \\
&\quad = \hat{P}(Y_t, S_t|S_{t-1}, \Theta_{t-1}). && (5.37c)
\end{aligned}
$$

In order to define ϕ we need to look at how $dF'_{x_t,t}(w)$ is updated. Consider the following recursive expression for the realization $f'_{x_{t+1},t+1}$ of $F'_{x_{t+1},t+1}$

$$
\begin{aligned}
& df'_{x_t,t}(a) \\
&= \frac{P(a, y_t, s_t, x_t | y^{t-1}, s^{t-1})}{P(y_t, s_t, x_t | y^{t-1}, s^{t-1})} && \text{(5.38a)} \\
&= \Bigg(\sum_{x_{t-1}} Q'(y_t | s_t, x_t) Q(s_t | s_{t-1}, x_{t-1}) \\
&\quad \times \delta_{e(f'_{x_{t-1},t-1}(a), x_{t-1}, s_{t-1}, \theta_{t-1})}(x_t) \theta_{t-1}(x_{t-1}) df'_{x_{t-1},t-1}(a) \Bigg) \\
&\quad \Bigg/ \Bigg(P(y_t, s_t, x_t | y^{t-1}, s^{t-1}) \Bigg) && \text{(5.38b)} \\
&\overset{(a)}{=} \Bigg(\sum_{x_{t-1}} Q'(y_t | s_t, x_t) Q(s_t | s_{t-1}, x_{t-1}) \\
&\quad \times \delta_{e(f'_{x_{t-1},t-1}(a), x_{t-1}, s_{t-1}, \theta_{t-1})}(x_t) \theta_{t-1}(x_{t-1}) df'_{x_{t-1},t-1}(a) \Bigg) \\
&\quad \Bigg/ \Bigg(\theta_t(x_t) \hat{P}(y_t, s_t | s_{t-1}, \theta_{t-1}) \Bigg), && \text{(5.38c)}
\end{aligned}
$$

where (a) is from (5.10). This implies that $f'_{x_t,t}$ can be recursively updated as

$$
\begin{aligned}
& f'_{x_t,t}(a) \\
& = \int_{a'=0}^{a} df'_{x_t,t}(a') && (5.39a) \\
& = \Bigg(\sum_{x_{t-1}} Q'(y_t|s_t, x_t) Q(s_t|s_{t-1}, x_{t-1}) \theta_{t-1}(x_{t-1}) \\
& \qquad \times \int_{a'=0}^{a} \delta_{e(f'_{x_{t-1},t-1}(a'), x_{t-1}, s_{t-1}, \theta_{t-1})}(x_t) df'_{x_{t-1},t-1}(a') \Bigg) \\
& \quad \Bigg/ \Bigg(\theta_t(x_t) \hat{P}(y_t, s_t|s_{t-1}, \theta_{t-1}) \Bigg) && (5.39b) \\
& = \sum_{x_{t-1}} \frac{Q'(y_t|s_t, x_t) Q(s_t|s_{t-1}, x_{t-1}) \theta_{t-1}(x_{t-1})}{\theta_t(x_t) \hat{P}(y_t, s_t|s_{t-1}, \theta_{t-1})} \\
& \qquad \times \Bigg\{ 1\Bigg(f'_{x_{t-1},t-1}(a) \geq \sum_{x=0}^{x_t} \hat{P}(x_t|x_{t-1}, s_{t-1}, \theta_{t-1}) \Bigg) \Bigg(\hat{P}(x_t|x_{t-1}, s_{t-1}, \theta_{t-1}) \Bigg) \\
& \qquad\qquad + 1\Bigg(\sum_{x=0}^{x_t-1} \hat{P}(x_t|x_{t-1}, s_{t-1}, \theta_{t-1}) \leq f'_{x_{t-1},t-1}(a) \\
& \qquad\qquad\qquad \leq \sum_{x=0}^{x_t} \hat{P}(x_t|x_{t-1}, s_{t-1}, \theta_{t-1}) \Bigg) \\
& \qquad\qquad\quad \times \Bigg(f'_{x_{t-1},t-1}(a) - \sum_{x=0}^{x_t-1} \hat{P}(x_t|x_{t-1}, s_{t-1}, \theta_{t-1}) \Bigg) \Bigg\} && (5.39c) \\
& \overset{(a)}{=} \phi(f'_{t-1}(a), y_t, s_t, s_{t-1}, \theta_{t-1}), && (5.39d)
\end{aligned}
$$

where we used in (a) the fact that θ_t is determined by $y_t, s_t, s_{t-1}, \theta_{t-1}$. Then,

$$
f'_t = \phi(f'_{t-1}, y_t, s_t, s_{t-1}, \theta_{t-1}). \tag{5.40}
$$

5.3 Proof of Lemma 5.3

$$E[F_t|s^{t-1}, y^{t-1}] = E[\phi(f'_{t-1}, Y_t, S_t, s_{t-1}, \theta_{t-1})|s^{t-1}, y^{t-1}] \tag{5.41a}$$

$$= E[\phi(f'_{t-1}, Y_t, S_t, s_{t-1}, \theta_{t-1})|s_{t-1}, \theta_{t-1}]. \tag{5.41b}$$

Then,

$$E[dF_t(w)|s^{t-1}, y^{t-1}]$$

$$= \sum_{y_t}\sum_{s_t}\sum_{x_{t-1}^t} Q'(y_t|s_t, x_t)Q(s_t|s_{t-1}, x_{t-1})\hat{P}(x_t|x_{t-1}, s_{t-1}, \theta_{t-1})\theta_{t-1}(x_{t-1})$$

$$\times\Bigg(\sum_{\hat{x}_{t-1}} Q'(y_t|s_t, e(f_{\hat{x}_{t-1},t-1}(w), \hat{x}_{t-1}, s_{t-1}, \theta_{t-1}))$$

$$\times Q(s_t|s_{t-1}, \hat{x}_{t-1})\theta_{t-1}(\hat{x}_{t-1})df'_{\hat{x}_{t-1},t-1}(w)\Bigg)/\Bigg(\sum_{\tilde{x}_{t-1}^t} Q'(y_t|s_t, \tilde{x}_t)$$

$$\times Q(s_t|s_{t-1}, \tilde{x}_{t-1})\hat{P}(\tilde{x}_{t-1}|\tilde{x}_{t-1}, s_{t-1}, \theta_{t-1})\theta_{t-1}(\tilde{x}_{t-1})\Bigg) \tag{5.42a}$$

$$= \sum_{\hat{x}_{t-1}} \theta_{t-1}(\hat{x}_{t-1})df'_{\hat{x}_{t-1},t-1}(w) \tag{5.42b}$$

$$= df_{t-1}(w). \tag{5.42c}$$

Similarly, we get

$$E[F_t(w)|s^{t-1}, y^{t-1}]$$

$$= E[\phi(f'_{t-1}(w), Y_t, S_t, s_{t-1}, \theta_{t-1})|s_{t-1}, \theta_{t-1}] \tag{5.43a}$$

$$= f_{t-1}(w). \tag{5.43b}$$

We would like to find a contraction mapping ξ such that for every w and f_{t-1} we have $E[\lambda(\phi(f'_{t-1}(w), Y_t, S_t, s_{t-1}, \theta_{t-1}|s_{t-1}, \theta_{t-1}))] \leq \xi(\lambda(f_{t-1}(w)))$. Let us assume for now that such a contraction mapping exists. We have

$$\begin{aligned}
&E[V_\lambda(\phi(f'_{t-1}, Y_t, S_t, s_{t-1}, \theta_{t-1}))|s_{t-1}, \theta_{t-1}] \\
&= E\Big[\int_0^1 \lambda(\phi(f'_{t-1}(w), Y_t, S_t, s_{t-1}, \theta_{t-1}))|s_{t-1}, \theta_{t-1}\Big] && (5.44a) \\
&< \int_0^1 \xi(\lambda(f_{t-1}(w)))dw && (5.44b) \\
&\overset{(a)}{\leq} \xi(V_\lambda(f_{t-1})), && (5.44c)
\end{aligned}$$

where the first inequality is due to the assumption for the property of ξ and the second inequality is due to the concavity of ξ. Then

$$\begin{aligned}
P(V_\lambda(F_t) > \epsilon) &\leq \frac{E[V_\lambda(F_t)]}{\epsilon} && (5.45a) \\
&= \frac{E[E[V_\lambda(F_t)|S^{t-1}, Y^{t-1}]]}{\epsilon} && (5.45b) \\
&= \frac{E[E[V_\lambda(\phi(f'_{t-1}, Y_t, S_t, S_{t-1}, \Theta_{t-1}))|S_{t-1}, \Theta_{t-1}]]}{\epsilon} && (5.45c) \\
&\leq \frac{E[\xi(V_\lambda(F_{t-1}))]}{\epsilon} && (5.45d) \\
&\leq \frac{\xi(E[V_\lambda(F_{t-1})])}{\epsilon} \cdots \leq \frac{\xi^t(E[V_\lambda(F_0)])}{\epsilon} \overset{(a)}{\to} 0, && (5.45e)
\end{aligned}$$

where the first inequality is the Markov inequality, the second inequality is due to (5.44), the third inequality is due to the concavity of ξ, the fourth inequality is due to repeated application of the above inequalities and the convergence to 0 is due to the property of the contraction [50, Lemma8]. Observe that the convergence is true for any initial distribution F_0, which implies that the convergence is uniform in the initial distribution.

It remains to find the contraction ξ with the property $E[\lambda(\phi(f'_{t-1}(w), Y_t, S_t, s_{t-1}, \theta_{t-1}|s_{t-1}, \theta_{t-1}))] \leq \xi(\lambda(f_{t-1}(w)))$. To this end let λ' :

$[0, 0.5] \to [0, 1]$ be a restriction of λ on $[0, 0.5]$. Then, λ' becomes one-to-one and onto hence it has inverse. Let $\tilde{\xi} : [0, 1] \to [0, 1]$ be

$$\begin{aligned} &\tilde{\xi}(a) \\ &= \sup_{\substack{s_{t-1},\theta_{t-1},a_i,0\le i\le K_X-1 \\ :0\le a_i\le 1,\sum_i \theta_{t-1}(i)a_i=a}} E[\lambda(\phi(a_0, a_1, ..., a_{K_X-1}, Y_t, S_t, s_{t-1}, \theta_{t-1}))|s_{t-1}, \theta_{t-1}]. \end{aligned} \tag{5.46}$$

Consider the above supremum. Space of s_{t-1} is finite, and spaces of θ_{t-1} and $(a_0, ..., a_{K_X-1})$ are compact. Since ϕ is continuous with respect to θ_{t-1} and $(a_0, ..., a_{K_X-1})$ given that $\hat{P}_{X|X',S,\Theta}(\cdot|\cdot,\cdot,\theta)$ is continuous with respect to θ. Therefore the above supremum belongs to space of $s_{t-1}, \theta_{t-1}, (a_0, ..., a_{K_X-1})$, and hence $\tilde{\xi}(a) < a$ because of (5.48). Consider now a following function.

$$\xi^*(a) = \max\left\{\tilde{\xi}(\lambda'^{-1}(a)), \tilde{\xi}(1 - \lambda'^{-1}(a))\right\}. \tag{5.47}$$

Clearly, $\xi^*(x) \geq 0$. We will now show that ξ^* satisfies the aforementioned property. Indeed, let $a \overset{\text{def}}{=} f_{t-1}(w)$. If $a \in [0, 1/2]$ then $\lambda'^{-1}(\lambda(a)) = a$ and the first term in the maximization on the r.h.s. of (5.47) equals $\tilde{\xi}(a)$. If $a \in [1/2, 1]$ then $1-\lambda'^{-1}(\lambda(a)) = a$ and the second term in the maximization of the r.h.s. of (4.32) equals $\tilde{\xi}(a)$. Thus the property holds. We now need to show that $\xi^*(a) < a$ for all $a \in (0, 1)$. To show this, since the channel is fixed-point free, $F_t(w)$ is not a.s. constant. Hence, using Jensen's inequality we get

$$E[\lambda(F_t(w))|s^{t-1}, y^{t-1}]$$
$$< \lambda(E[F_t(w)|s^{t-1}, y^{t-1}]) \tag{5.48a}$$
$$= \lambda(f_{t-1}(w)). \tag{5.48b}$$

Therefore, $\tilde{\xi}(\lambda'^{-1}(a)) < a$ from (5.48) and the definition of $\tilde{\xi}$. We also have $\tilde{\xi}(1 -$

$\lambda'^{-1}(a)) < a$. Hence,

$$\xi^*(a) < a. \tag{5.49}$$

Finally, we need to establish the concavity of ξ. Since this property does not hold for ξ^* we define ξ as the supremum of the convex hull of ξ^*. Let ξ be the upper convex envelope of ξ^*, i.e.,

$$\xi(a) = \sup\{b : (a, b) \in L\}, \tag{5.50a}$$

$$L = \text{conv}\{(a, b) : a \in [0, 1], b \in (0, \xi^*(a))\}. \tag{5.50b}$$

Then ξ is concave and from the definition of ξ

$$E[\lambda(\phi(f'_{t-1}(w), Y_t, S_t, s_{t-1}, \theta_{t-1}))|s_{t-1}, \theta_{t-1}] \tag{5.51a}$$

$$\leq \xi(\lambda(f_{t-1}(w)))$$

$$< \lambda(f_{t-1}(w)). \tag{5.51b}$$

For any $a \in (0, 1]$, there must exist some constant $\alpha \in [0, 1]$ such that $a = \alpha a_0 + (1 - \alpha)a_1$

$$\xi(a) \leq \alpha\xi^*(a_0) + (1 - \alpha)\xi^*(a_1) < \alpha a_0 + (1 - \alpha)a_1 = a, \tag{5.52}$$

where we used the definition of he upper convex envelope in the first inequality. Since ξ is nonnegative, it is contraction.

5.4 Proof of Proposition 5.5

Using the symmetry of λ, we can write

$$
\begin{aligned}
&V_\lambda(F^u_{t'+1,(1+\alpha)t}) \\
&= \int_0^{W^*_{t',t}} \lambda(F^u_{t'+1,(1+\alpha)t}(w))dw + \int_{W^*_{t',t}}^1 \lambda(1 - F^u_{t'+1,(1+\alpha)t}(w))dw,
\end{aligned}
$$

$$
\begin{aligned}
&\max_{1\le t'\le t} V_\lambda(F^u_{t'+1,(1+\alpha)t}) && (5.53)\\
&= \max_{1\le t'\le t} \Big[\int_0^{W^*_{t',t}} \lambda(F^u_{t'+1,(1+\alpha)t}(w))dw + \int_{W^*_{t',t}}^1 \lambda(1 - F^u_{t'+1,(1+\alpha)t}(w))dw \Big], && (5.54)
\end{aligned}
$$

where $W^*_{t',t}$ is the unique solution of $F^u_{t'+1,(1+\alpha)t}(w) = 0.5$. Then, we have

$$
\begin{aligned}
&P(\max_{1\le t'\le t} F^u_{t'+1,(1+\alpha)t}(W^*_{t',t} - \delta) > \nu) \\
&\le P(\max_{1\le t'\le t} \lambda(F^u_{t'+1,(1+\alpha)t}(W^*_{t',t} - \delta)) > \nu) && (5.55a)\\
&\le P(\max_{1\le t'\le t} \int_{W^*_{t',t}-\delta}^{W^*_{t',t}} \lambda(F^u_{t'+1,(1+\alpha)t}(w))dw > \nu\delta) && (5.55b)\\
&\le P(\max_{1\le t'\le t} V_\lambda(F^u_{t'+1,(1+\alpha)t})) > \nu\delta). && (5.55c)
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
&P(\max_{1\le t'\le t} \Big[1 - F^u_{t'+1,(1+\alpha)t}(W^*_{t',t} + \delta)\Big] < 1 - \nu) \\
&\le P(\max_{1\le t'\le t} V_\lambda(F^u_{t'+1,(1+\alpha)t})) > \nu\delta). && (5.56)
\end{aligned}
$$

For any $\eta \in (0, 0.5)$

$$
\begin{aligned}
&P\bigg(\max_{1\le t'\le t}\Big[\int_0^{F_{t'}(W)} F^u_{t'+1,(1+\alpha)t}(w)dw + \int_{F_{t'}(W)}^1 (1 - F^u_{t'+1,(1+\alpha)t}(w))dw\Big] \\
&\qquad > \nu\bigg) \\
&\le P\bigg(\max_{1\le t'\le t}\Big[\int_0^{W^*_{t',t}} F^u_{t'+1,(1+\alpha)t}(w)dw + \int_{W^*_{t',t}}^1 (1 - F^u_{t'+1,(1+\alpha)t}(w))dw\Big] \\
&\qquad > \nu/2\bigg) + P\bigg(\max_{1\le t'\le t}|F_{t'}(W) - W^*_{t',t}| > \nu/2\bigg) && \text{(5.57a)} \\
&\le P\bigg(\max_{1\le t'\le t} V_\lambda(F^u_{t'+1,(1+\alpha)t}) > \nu/2\bigg) \\
&\quad + P\bigg(\Big\{F_{(1+\alpha)t}(W) < \max_{1\le t'\le t} F^u_{t'+1,(1+\alpha)t}(W^*_{t',t} - \nu/2)\Big\} \\
&\qquad\quad \cup \Big\{F_{(1+\alpha)t}(W) > \min_{1\le t'\le t} F^u_{t'+1,(1+\alpha)t}(W^*_{t',t} + \nu/2)\Big\}\bigg) && \text{(5.57b)} \\
&\le P\bigg(\max_{1\le t'\le t} V_\lambda(F^u_{t'+1,(1+\alpha)t}) > \nu/2\bigg) + P\bigg(F_{(1+\alpha)t}(W) \notin (\eta, 1-\eta)\bigg) \\
&\qquad + P\bigg(\max_{1\le t'\le t} F^u_{t'+1,(1+\alpha)t}(W^*_{t',t} - \nu/2) > \eta\bigg) \\
&\qquad + P\bigg(\max_{1\le t'\le t}[1 - F^u_{t'+1,(1+\alpha)t}(W^*_{t',t} + \nu/2)] < 1 - \eta\bigg) && \text{(5.57c)} \\
&\le P(\max_{1\le t'\le t} V_\lambda(F^u_{t'+1,(1+\alpha)t})) > \nu/2) \\
&\qquad + 2\eta + 2P(\max_{1\le t'\le t} V_\lambda(F^u_{t'+1,(1+\alpha)t})) > \nu\eta/2). && \text{(5.57d)}
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&P\Big(\max_{1\le t'\le t} F^u_{t'+1,(1+\alpha)t}\Big(F_{t'}(W)-\delta\Big) > \epsilon\Big) \\
&\le P\Big(\max_{1\le t'\le t} \int_{F_{t'}(W)-\delta}^{F_{t'}(W)} F^u_{t'+1,(1+\alpha)t}(w)dw > \delta\epsilon\Big) && \text{(5.58a)} \\
&\le P\Big(\max_{1\le t'\le t} \int_{0}^{F_{t'}(W)} F^u_{t'+1,(1+\alpha)t}(w)dw > \delta\epsilon\Big) && \text{(5.58b)} \\
&\le P\Big(\max_{1\le t'\le t} \Big[\int_{0}^{F_{t'}(W)} F^u_{t'+1,(1+\alpha)t}(w)dw + \int_{F_{t'}(W)}^{1} (1-F^u_{t'+1,(1+\alpha)t}(w))dw\Big] \\
&\qquad > \nu\Big) && \text{(5.58c)} \\
&\le P(\max_{1\le t'\le t} V_\lambda(F^u_{t'+1,(1+\alpha)t})) > \delta\epsilon/2) \\
&\qquad + 2\eta + 2P(\max_{1\le t'\le t} V_\lambda(F^u_{t'+1,(1+\alpha)t})) > \delta\epsilon\eta/2). && \text{(5.58d)}
\end{aligned}
$$

Setting $\eta = \sqrt{\sup_{a\in[0,1]} \xi^t(a)}/(\delta\epsilon)$, together with Lemma 5.4 completes the proof of the first assertion of the proposition. The proof of the second assertion is similar.

5.5 Proof of Lemma 5.6

First we show that $I(W; Y_{t+1}, S_{t+1}|Y^t, S^t) = E\{\log \frac{dF_{t+1}(W)}{dF_t(W)}|\}$.

$$
\begin{aligned}
&I(W; Y_{t+1}, S_{t+1}|Y^t, S^t) \\
&= H(Y_{t+1}, S_{t+1}|Y^t, S^t) - H(Y_{t+1}, S_{t+1}|W, Y^t, S^t) && \text{(5.59a)} \\
&= E\left[\log \frac{1}{P(Y_{t+1}, S_{t+1}|Y^t, S^t)}\right] \\
&\quad + E\Big[\log\Big\{Q'(Y_{t+1}|S_{t+1}, e(F'_{X_t,t}(W), X_t, S_t, \Theta_t))Q(S_{t+1}|S_t, X_t)\Big\}\Big] && \text{(5.59b)}
\end{aligned}
$$

$$= E\left[\log \frac{Q'(Y_{t+1}|S_{t+1}, e(F'_{X_t,t}(W), X_t, S_t, \Theta_t))Q(S_{t+1}|S_t, X_t)}{\hat{P}(Y_{t+1}, S_{t+1}|S_t, \Theta_t)}\right] \tag{5.59c}$$

$$= E\left[\log \frac{dF_{t+1}(W)}{dF_t(W)}\right]. \tag{5.59d}$$

Now we show that $I(W; Y_t, S_t|Y^{t-1}, S^{t-1}) = C$. Note that for a given Y^{t-1}, S^{t-1}, X_t is $\hat{P}(\cdot|X_{t-1}, S_{t-1}, \Theta_t)$ distributed and hence is independent of Y^{t-1}, S^{t-2}. Then,

$$I(W; Y_t, S_t|Y^{t-1}, S^{t-1})$$

$$= H(Y_t, S_t|Y^{t-1}, S^{t-1}) - H(Y_t, S_t|Y^{t-1}, S^{t-1}, W) \tag{5.60a}$$

$$\overset{(a)}{=} H(Y_t, S_t|Y^{t-1}, S^{t-1}, \theta_{t-1}) - H(Y_t, S_t|Y^{t-1}, S^{t-1}, W, X_{t-1}^t, \Theta_{t-1}) \tag{5.60b}$$

$$\overset{(b)}{=} H(Y_t, S_t|S_{t-1}, \Theta_{t-1}) - H(Y_t, S_t|S_{t-1}, X_{t-1}^t, \Theta_{t-1}) \tag{5.60c}$$

$$= I(X_{t-1}^t; S_t, Y_t|S_{t-1}, \Theta_{t-1}) \tag{5.60d}$$

$$= C, \tag{5.60e}$$

where (a) due to the fact that Θ_{t-1} is a function of Y^{t-1}, S^{t-1} and X_{t-1}^t is a function of Y^{t-1}, S^{t-1}, W; ((b) is due to the channel characteristics and (5.10); and the last equation is due to the fact that the channel input sequence for the PMS has distribution which is capacity-achieving.

5.6 Proof of Lemma 5.10

Consider

$P_{Y_{t+1},S_{t+1}|F'_t,H_t,S_t,\Theta_t,W} = Q'_{Y_{t+1}|S_{t+1},g(F'_{H_t(F_t(W)),t}(W),H_t(F_t(W)),S_t,\Theta_t)} Q_{S_{t+1}|S_t,H_t(F_t(W))}$. For any $\epsilon > 0$, define ${}^{-}P^{\epsilon}_{Y_{t+1},S_{t+1}|F'_t,H_t,S_t,\Theta_t,W}$ to be

$$\begin{aligned}
&{}^{-}P^{\epsilon}_{Y_{t+1},S_{t+1}|F'_t,H_t,S_t,\Theta_t,W}(y,s|f',h,s,\theta,w) \\
&\quad = \inf_{f^{-1}(f(w)-\epsilon)<w'<w} P_{Y_{t+1},S_{t+1}|F'_t,H_t,S_t,\Theta_t,W}(y,s|f',h,s,\theta,w'), \qquad (5.61)
\end{aligned}$$

where $f(w) = \sum_x \theta(x) f'_x(w)$. Similarly, define ${}^{+}P^{\epsilon}_{Y_{t+1},S_{t+1}|F'_t,H_t,S_t,\Theta_t,W}$ to be

$$\begin{aligned}
&{}^{-}P^{\epsilon}_{Y_{t+1},S_{t+1}|F'_t,H_t,S_t,\Theta_t,W}(y,s|f',h,s,\theta,w) \\
&\quad = \inf_{w<w'<f^{-1}(f(w)+\epsilon)} P_{Y_{t+1},S_{t+1}|F'_t,H_t,S_t,\Theta_t,W}(y,s|f',h,s,\theta,w'). \qquad (5.62)
\end{aligned}$$

From now on we prove the first assertion of lemma, the second assertion follows in a similar way. Define

$$\begin{aligned}
C^{-}_{\epsilon} &= E\left[\log\left\{\frac{{}^{-}P^{\epsilon}(Y_{t+1},S_{t+1}|F'_t,H_t,S_t,\Theta_t,W)}{P(Y_{t+1},S_{t+1}|S_t,\Theta_t)}\right\}\right] && (5.63a) \\
&= \int_{f'_t,h_t,\theta_t,w} \sum_{y_{t+1},s_t^{t+1}} P(y_{t+1},s_{t+1}|f'_t,h_t,s_t,\theta_t,w) P(f'_t,h_t,s_t,\theta_t,w) \\
&\qquad\qquad \times \log\left\{\frac{{}^{-}P^{\epsilon}(y_{t+1},s_{t+1}|f'_t,h_t,s_t,\theta_t,w)}{P(y_{t+1},s_{t+1}|s_t,\theta_t)}\right\}. && (5.63b)
\end{aligned}$$

Note that

$$\begin{aligned}
C &= E\left[\log\left\{\frac{P(Y_{t+1},S_{t+1}|F'_t,H_t,S_t,\Theta_t,W)}{P(Y_{t+1},S_{t+1}|S_t,\Theta_t)}\right\}\right] && (5.64a) \\
&= \int_{f'_t,h_t,\theta_t,w} \sum_{y_{t+1},s_t^{t+1}} P(y_{t+1},s_{t+1}|f'_t,h_t,s_t,\theta_t,w) P(f'_t,h_t,s_t,\theta_t,w) \\
&\qquad\qquad \times \log\left\{\frac{P(y_{t+1},s_{t+1}|f'_t,h_t,s_t,\theta_t,w)}{P(y_{t+1},s_{t+1}|s_t,\theta_t)}\right\}. && (5.64b)
\end{aligned}$$

Then we have,

$$D(P_{Y_{t+1},S_{t+1}|F'_t,H_t,S_t,\Theta_t,W}||^{-}P^{\epsilon}_{Y_{t+1},S_{t+1}|F'_t,H_t,S_t,\Theta_t,W}|P_{F'_t,H_t,S_t,\Theta_t,W})$$
$$= C - C^{-}_{\epsilon} \geq 0. \tag{5.65}$$

Assume that

$$\inf_{\epsilon>0} D(P_{Y_{t+1},S_{t+1}|F'_t,H_t,S_t,\Theta_t,W}||^{-}P^{\epsilon}_{Y_{t+1},S_{t+1}|F'_t,H_t,S_t,\Theta_t,W}|P_{F'_t,H_t,S_t,\Theta_t,W}) < \infty. \tag{5.66}$$

Then,$-\infty < C^{-}_{\epsilon} \leq C$. Therefore, $P_{Y_{t+1},S_{t+1},F'_t,H_t,S_t,\Theta_t,W} \log \left\{ \frac{^{-}P^{\epsilon}_{Y_{t+1},S_{t+1}|F'_t,H_t,S_t,\Theta_t,W}}{P_{Y_{t+1},S_{t+1}|S_t,\Theta_t}} \right\}$ is finitely integrable and converges to $P_{Y_{t+1},S_{t+1},F'_t,H_t,S_t,\Theta_t,W} \log \left\{ \frac{P_{Y_{t+1},S_{t+1}|F'_t,H_t,S_t,\Theta_t,W}}{P_{Y_{t+1},S_{t+1}|S_t,\Theta_t}} \right\}$ a.e. in a monotonically nondecreasing fashion as $\epsilon \to 0$. Hence, by the monotone convergence theorem,

$$\lim_{\epsilon \to 0} C_{\epsilon} = C, \tag{5.67}$$

and there exists ϵ'_{δ} such that for all $\epsilon \leq \epsilon'_{\delta}$

$$C^{-}_{\epsilon} > C - \delta/2. \tag{5.68}$$

Furthermore, we can apply the strong law of large numbers for Markov chains to get

$$\lim_{t\to\infty} \frac{1}{t} \sum_{s=0}^{t-1} \log \left\{ \frac{^{-}P^{\epsilon}(Y_{s+1}, S_{s+1}|F'_s, H_s, S_s, \Theta_s, W)}{P(Y_{s+1}, S_{s+1}|S_s, \Theta_s)} \right\} = C^{-}_{\epsilon} \qquad \text{a.s.} \tag{5.69}$$

Note that if $F_s(W) - F_s(W - 2^{-tR}) < \epsilon$ for $0 \le s \le t-1$

$$\begin{aligned}
&\inf_{W-2^{-tR}<w<W} \frac{1}{t} \log dF_t(w) \\
&\ge \frac{1}{t}\sum_{s=0}^{t-1} \log \inf_{W-2^{-tR}<w<W} \frac{P(Y_{s+1}, S_{s+1}|F'_s, H_s, S_s, \Theta_s, w)}{P(Y_{s+1}, S_{s+1}|S_s, \Theta_s)} && \text{(5.70a)} \\
&\ge \frac{1}{t}\sum_{s=0}^{t-1} \log \left\{ \frac{{}^-P^\epsilon(Y_{s+1}, S_{s+1}|F'_s, H_s, S_s, \Theta_s, W)}{P(Y_{s+1}, S_{s+1}|S_s, \Theta_s)} \right\}. && \text{(5.70b)}
\end{aligned}$$

Therefore,

$$\begin{aligned}
&P\Big(\bigcap_{s=0}^{t-1}\{F_s(W) - F_s(W - 2^{-tR}) < \epsilon\}\Big) \\
&\le P\Big(\bigcap_{s=0}^{t-1}\{F_s(W) - F_s(W - 2^{-tR}) < \epsilon\} \\
&\qquad \bigcap\{\frac{1}{t}\sum_{s=0}^{t-1} \log \left\{ \frac{{}^-P^\epsilon(Y_{s+1}, S_{s+1}|F'_s, H_s, S_s, \Theta_s, W)}{P(Y_{s+1}, S_{s+1}|S_s, \Theta_s)} \right\} \ge C - \delta/2\}\Big) \\
&\quad + P\Big(\frac{1}{t}\sum_{s=0}^{t-1} \log \left\{ \frac{{}^-P^\epsilon(Y_{s+1}, S_{s+1}|F'_s, H_s, S_s, \Theta_s, W)}{P(Y_{s+1}, S_{s+1}|S_s, \Theta_s)} \right\} < C - \delta/2\Big) && \text{(5.71a)} \\
&\overset{(a)}{\le} P\Big(\inf_{W-2^{-tR}<w<W} \frac{1}{t} \log dF_t(w) \ge C - \delta/2\Big) \\
&\quad + P\Big(\frac{1}{t}\sum_{s=0}^{t-1} \log \left\{ \frac{{}^-P^\epsilon(Y_{s+1}, S_{s+1}|F'_s, H_s, S_s, \Theta_s, W)}{P(Y_{s+1}, S_{s+1}|S_s, \Theta_s)} \right\} < C - \delta/2\Big) && \text{(5.71b)} \\
&\overset{(b)}{\le} P\Big(\inf_{W-2^{-tR}<w<W} \frac{1}{t} \log dF_t(w) \ge C - \delta/2\Big) + o(1) && \text{(5.71c)} \\
&\le P\Big(\int_{W-2^{-tR}}^{W} dF_t(w) \ge 2^{t(C-R-\delta/2)}\Big) + o(1) && \text{(5.71d)} \\
&\le P\Big(\int_{W-2^{-tR}}^{W} dF_t(w) \ge 2^{t\delta/2}\Big) + o(1) && \text{(5.71e)} \\
&\overset{(c)}{\to} 0, && \text{(5.71f)}
\end{aligned}$$

where (a) is due to (5.70), (b) is due to (5.69), and (c) is from the fact that $\int_{W-2^{-tR}}^{W} dF_t(w) \le 1$.

5.7 Proof of Proposition 5.11

For every $\epsilon > 0$ we have

$$
\begin{aligned}
&P(F_{(1+\alpha)t}(W - 2^{-tR}) > \delta) \\
&\le P\Bigg(\bigcap_{t'=1}^{t} \Big\{ F_{t'}(W - 2^{-tR}) > F_{t'}(W) - \epsilon \Big\} \Bigg) \\
&\qquad + P\Bigg(\Big\{ F_{(1+\alpha)t}(W - 2^{-tR}) > \delta) \Big\} \\
&\qquad\qquad \cap \bigcup_{t'=1}^{t} \Big\{ F_{t'}(W - 2^{-tR}) \le F_{t'}(W) - \epsilon \Big\} \Bigg).
\end{aligned} \tag{5.72}
$$

Regarding the first quantity we have

$$
P\Bigg(\bigcap_{t'=1}^{t} \Big\{ F_{t'}(W - 2^{-tR}) > F_{t'}(W) - \epsilon \Big\} \Bigg) \overset{(a)}{\to} 0, \tag{5.73}
$$

where (a) is from Lemma 5.10.

Consider an event $\bigcup_{t'=1}^{t} \Big\{ F_{t'}(W - 2^{-tR}) \le F_{t'}(W) - \epsilon \Big\}$. This event is equivalent to $\bigcup_{t'=1}^{t} \{W - 2^{-tR} \le W'\}$ where $W' = F_{t'}^{-1}(F_{t'}(W) - \epsilon)$. Then by the monotonicity of cdf it is equivalent to $\bigcup_{t'=1}^{t} \{F_t(W - 2^{-tR}) \le F_t(W')\}$ which again is equivalent to $\Big\{ F_t(W - 2^{-tR}) \le \max_{1 \le t' \le t} F_t\Big(W'\Big)\Big) \Big\}$.

Therefore, regarding the second quantity we have

$$
\begin{aligned}
&P\Bigg(\Big\{ F_{(1+\alpha)t}(W - 2^{-tR}) > \delta) \Big\} \\
&\qquad \cap \bigcup_{t'=1}^{t} \Big\{ F_{t'}(W - 2^{-tR}) \le F_{t'}(W) - \epsilon \Big\} \Bigg) \\
&= P\Bigg(\Big\{ F_{(1+\alpha)t}(W - 2^{-tR}) > \delta) \Big\} \\
&\qquad \cap \Big\{ F_t(W - 2^{-tR}) \le \max_{1 \le t' \le t} F_t\Big(F_{t'}^{-1}\Big(F_{t'}(W) - \epsilon \Big)\Big) \Big\} \Bigg)
\end{aligned} \tag{5.74a}
$$

$$
\begin{aligned}
&\overset{(a)}{\leq} P\Bigg(\Big\{F_{(1+\alpha)t}(W - 2^{-tR}) > \delta)\Big\} \\
&\qquad \cap \Big\{F_{(1+\alpha)t}(W - 2^{-tR}) \leq \max_{1\leq t'\leq t} F_{(1+\alpha)t}\Big(F_{t'}^{-1}\Big(F_{t'}(W) - \epsilon\Big)\Big)\Big\}\Bigg) \qquad (5.74\text{b}) \\
&= P\Bigg(\max_{1\leq t'\leq t} F_{(1+\alpha)t}\Big(F_{t'}^{-1}\Big(F_{t'}(W) - \epsilon\Big)\Big) > \delta\Bigg) \qquad (5.74\text{c}) \\
&\overset{(b)}{=} P\Bigg(\max_{1\leq t'\leq t} F^{u}_{t'+1,(1+\alpha)t}\Big(F_{t'}(W) - \epsilon\Big) > \delta\Bigg) \overset{(c)}{\to} 0. \qquad (5.74\text{d})
\end{aligned}
$$

where (a) is from the monotonicity of cdf, (b) is due to the PMS update, and (c) is from Proposition 5.5. $P(F_{(1+\alpha)t}(W + 2^{-tR}) < 1 - \delta)$ can also be evaluated in a similar way. Note that the fixed-point free nature of the channel is required for zero-rate result, but it can be easily verified that ergodicity of invariant distribution of $(F_t'(W), \Theta_t, X_t, S_t^{t+1}, Y_{t+1})_t$ implies fixed-point free channel as in Lemma 14 of [50], and ergodicity of invariant distribution of $(F_t'(W), \Theta_t, X_t, S_t^{t+1}, Y_{t+1})_t$ is implied by that of $(F_t', \Theta_t, H_t, S_t^{t+1}, Y_{t+1}, W)_t$.

CHAPTER 6

Conclusion

6.1 Summary of results and comments

In this chapter, a summary of the obtained results is provided, together with comments and future research directions.

6.1.1 Capacity achieving codes for finite-state channels without feedback

In Chapter 2, we presented SIR-achieving and capacity-achieving code ensembles for FSCs. We first established an upper bound on the average block error probability of coset code ensembles transmitted through FSCs. We used this bound to show that coset ensembles generated by regular LDPC, punctured LDPC, and LDPC-GM ensembles which achieve the capacity of MBIOS channels also achieve the SIR of FSCs. Next, we presented a method of quantization that enables the construction of code ensembles inducing a Markov distribution. We established an upper bound on the average block error probability of quantized coset code ensembles transmitted through FSCs. Using this bound, we showed that the sequences of quantized regular

LDPC, punctured LDPC, and LDPC-GM coset ensembles can achieve the capacity of FSCs as the order of the induced Markov distribution approaches infinity.

6.1.2 A single-letter capacity expression for finite-state channels with feedback

In Chapter 3, a single-letter expression for the capacity of the FSC with delayed feedback was derived. This study was motivated by the fact that the capacity expression for the general class of such channels given in (3.3a) and derived in [54] involves a measure on the space of channel states and a measure on the space of measures on channel states as well. Our derived expression given in (3.16) only involves a measure on the previously transmitted channel input. Although this is an infinite "state space" (even for finite state alphabets), this methodology results in a capacity expression that does not depend on an ever-expanding sequence of random variables X^t, Y^t, S^t.

The methodology was based on reformulating the capacity problem as a stochastic control problem. Although this methodology was based on the one proposed in [54], the key difference was to identify the information state and action directly instead of considering a stochastic control problem with a state dependent action space, as was done in [54]. We believe that our approach reflects the basic principles of Markov decision theory better and hence is more natural. This methodology is quite general and will likely be useful in finding single-letter expressions for the capacity of other channels, such as the multiple-access channel with feedback.

We finally applied this methodology to the evaluation of the capacity expression for the case with no ISI, and the resulting expression agrees with the previous result given in [52]. It turns out that the corresponding control problem is a simple Markov decision process problem, and our methodology is considerably simpler than that of [52].

6.1.3 A capacity-achieving posterior matching scheme

In Chapters 4 and 5, we proposed a feedback transmission scheme based on the simplified capacity expression found in Chapter 3. In both cases (i.e., with and without ISI), the proposed schemes can be thought of as generalization of the PMS proposed in [50]. We first considered the no ISI case in Chapter 4. As we saw earlier, the corresponding scheme is a rather straightforward generalization of its DMC counterpart. We followed the same methodology as in [50] to show capacity achievability: first establishing the zero-rate result and then the rate $R < C$ achievability. The zero-rate result is based on identifying an appropriate Lyapunov function V_λ whose convergence implies zero-rate achievability. The intuition can easily be seen from the definition of the Lyapunov function, i.e., the fact that a small $V_\lambda(f)$ means that f is almost a step function. As seen in this thesis, the same Lyapunov function can be used for channels with memory, and the main difficulty of proving the zero-rate result comes from identifying an appropriate contraction mapping to show convergence of the Lyapunov function.

In Chapter 5 we turn our attention to the case with ISI and observe that the differences from the previous cases are considerable. First, the PMS itself looks a lot different from its DMC (or no-ISI) counterpart. The reason for that is because the optimal transmitter must use information of the previous channel input when it generates the current channel input because information of the previous channel input is delivered to the receiver through the presence of ISI. The general structure of the proof of capacity achievability follows the structure of the no ISI case. Regarding the zero-rate result, the identification of an appropriate contraction mapping is the main difficulty. Regarding the rate $R < C$ result, however, a number of unique difficulties arose that were not present in the no-ISI case. The first is that unlike the DMC or no-ISI case the transmitted symbol X_t is not only a function of $F_t(W)$. This difficulty posed the dilemma of either studying a Markov process related to the entire cdf F_t, or construct (and study) a Markov chain involving an appropriate set of sufficient quantities. Another problem arose when considering the quantity $H_t(F_t(W))$ relating to the value of this function in the 2^{-tR} neighborhood of the message point W. We have not resolved this issue conclusively but we strongly believe this is the case.

6.2 Future research directions

First, consider the case when there is no feedback. Although we showed that our scheme achieves the capacity of FSC without feedback we have not studied the corresponding error exponents. Because error exponents give information on how fast the error probability converges to zero, this study is important in practical situations where we cannot assume infinitely long transmission time. Therefore, investigating error exponents for our proposed schemes would be a meaningful future direction. To follow this direction, one must come up with better upper bounds on error probability than the one we derived. The main objective and difficulty would be the derivation of a random coding bound in single letter form, and the corresponding development of a stronger SF bound.

In the case where there is feedback, we started our investigation by simplifying the capacity expression. As mentioned in the summary, the methodology used here is quite general, so it may be useful when trying to simplify the capacity expression for non point-to-point channels. One may have to combine information theoretical concepts and concepts from stochastic control in order to come up with a simplified expression, since both these areas have their own strengths and weaknesses.

Consider now the derivation of the capacity-achieving scheme for channels with memory and feedback, especially the analysis of generalized PMS. As discussed above, the current methodology might have weaknesses for more general channels with memory. One might follow a different approach to the analysis of the PMS: recently Coleman provided one example of such an alternative analysis in [51]. It would be interesting to see if Coleman's methodology resolves some of the unresolved issues at hand. Another direction of research would be investigation of error exponents of the PMS. This is partially done for the DMC case in [50], however the authors did not provide closed form expressions for the error exponents and did not show any kind of optimality of the exponents. One might consider a completely different PMS-like scheme that aims at optimizing the finite-length performance of the communication system as opposed to achieving capacity. Other future directions include applying the PMS-like schemes for non point-to-point channels. One such example can be found in [68] for broadcast channels.

APPENDICES

Appendix A

The posterior matching scheme (PMS) for the discrete memoryless channel (DMC)

Consider a DMC with input/output alphabets $\mathcal{X} = \{0, 1, \ldots, K_X - 1\}$, $\mathcal{Y} = \{0, 1, \ldots, K_Y - 1\}$, transition matrix $\{Q(y|x)\}_{x \in \mathcal{X}, y \in \mathcal{Y}}$ and unit delay noiseless feedback. Let W be a random message uniformly distributed over the unit interval. A transmission scheme over this channel is a sequence of functions $\{\tilde{e}_t : [0,1) \times \mathcal{Y}^{t-1} \to \mathcal{X}\}_{t=1}^{\infty}$ such that

$$X_t = \tilde{e}_t(W, Y^{t-1}). \tag{A.1}$$

Let $\hat{W}_t$ be the message point estimate at the receiver at time t. We now describe a simple sequential transmission scheme, namely the posterior matching scheme (PMS). We assume that the capacity achieving distribution $\hat{P}(X)$ has been found, and we denote the corresponding output distribution by $\hat{P}_Y(y) = \sum_x Q(y|x)\hat{P}(x)$. Define the random variable $F_t \in \mathcal{F}$ as $F_t(w) \stackrel{\text{def}}{=} F(w|Y^t)$, where $F(\cdot|Y^t)$ is the a-posteriori cdf of W conditioned on Y^t, and $\mathcal{F}$ is the set of all valid cdfs over $[0, 1)$.

Then, at time $t \geq 1$ the transmitter generates and transmits the channel input X_t as

$$
\begin{aligned}
X_t &= F_{\hat{P}}^{-1}(F_{t-1}(W)) && \text{(A.2a)}\\
&= x, \qquad \sum_{i=0}^{x-1} \hat{P}(i) < F_{t-1}(W) \leq \sum_{i=0}^{x} \hat{P}(i) \qquad x = 0, \ldots, K_X - 1 && \text{(A.2b)}\\
&\stackrel{\text{def}}{=} e(F_{t-1}(W)), && \text{(A.2c)}
\end{aligned}
$$

where the inverse cdf $F^{-1}(y) \stackrel{\text{def}}{=} \inf\{x : F(x) \geq y\}$.

Lemma A.1. *For the PMS scheme we have*

$$P(X_t|Y^{t-1}) = \hat{P}(X_t) \tag{A.3}$$

$$P(Y_t|Y^{t-1}) = \hat{P}_Y(Y_t) \tag{A.4}$$

$$F_t = \phi(F_{t-1}, Y_t), \quad t = 1, 2, \ldots, \tag{A.5}$$

where $F_0 = Uniform[0, 1)$.

Proof.

$$
\begin{aligned}
P(X_t|Y^{t-1}) &= P(X_t|Y^{t-1}, F_{t-1}) && \text{(A.6a)}\\
&= \int_w P(X_t|Y^{t-1}, F_{t-1}, W = w)P(dw|Y^{t-1}, F_{t-1}) && \text{(A.6b)}\\
&= \int_w \delta_{e(F_{t-1}(w))}(X_t) dF_{t-1}(w) && \text{(A.6c)}\\
&= \sum_x \delta_x(X_t)\hat{P}(x) && \text{(A.6d)}\\
&= \hat{P}(X_t), && \text{(A.6e)}
\end{aligned}
$$

and similarly

$$P(Y_t|Y^{t-1}) = \sum_x P(Y_t|Y^{t-1}, X_t = x)P(X_t = x|Y^{t-1}) \tag{A.7a}$$

$$= \sum_x Q(Y_t|x)\hat{P}(x) \tag{A.7b}$$

$$= \hat{P}_Y(Y_t). \tag{A.7c}$$

The transmitter subsequently updates the quantity F_{t-1} according

$$F_t = \phi(F_{t-1}, Y_t) \tag{A.8}$$

where ϕ is given implicitly through the corresponding pdf update

$$df_t(a) = \frac{P(y^t|a)da}{P(y^t)} \tag{A.9a}$$

$$= \frac{Q(y_t|e(f_{t-1}(a)))P(y^{t-1}|a)da}{P(y_t|y^{t-1})P(y^{t-1})} \tag{A.9b}$$

$$= \frac{Q(y_t|e(f_{t-1}(a)))df_{t-1}(a)}{P(y_t|y^{t-1})} \tag{A.9c}$$

$$= \frac{Q(y_t|e(f_{t-1}(a)))df_{t-1}(a)}{\int_{a'} Q(y_t|e(f_{t-1}(a')))df_{t-1}(a')} \tag{A.9d}$$

$$= \frac{Q(y_t|e(f_{t-1}(a)))df_{t-1}(a)}{\hat{P}_Y(y_t)}, \tag{A.9e}$$

and explicitly through

$$f_t(a) = \frac{\sum_{i=0}^{x-1} Q(y_t|i)\hat{P}(i) + Q(y_t|x)[f_{t-1}(a) - \sum_{i=0}^{x-1} \hat{P}(i)]}{\hat{P}_Y(y_t)}, \tag{A.10}$$

$$\sum_{i=0}^{x-1} \hat{P}(i) < f_{t-1}(a) \le \sum_{i=0}^{x} \hat{P}(i),$$

$$x = 0, \ldots, K_X - 1.$$

$\square$

Observe from the above equation that $f_t(a)$ is a function of f_{t-1} only through $f_{t-1}(a)$. In the following we will also use the notation $f_t(a) = \phi(f_{t-1}, y_t)(a) = \phi(f_{t-1}(a), y_t)$. Observe also from (A.2a) that the transmitted symbol X_t is a function of W and F_{t-1} only through the quantity $F_{t-1}(W)$. This has important implications for the analysis of the PMS.

Note that from Lemma A.1, Y_t's are i.i.d. random variables, each with distribution $\hat{P}_Y$.

Assuming that the channel transition probabilities $Q(y|x)$ are non-zero for all x, y, the recursion described by (A.5) and (A.9) guarantees that for every realization of the random variables of interest, F_t will always have a pdf; in addition the pdf will be piecewise constant and non-zero everywhere in $(0, 1]$.

At the receiver, the quantity F_t is updated according to (A.5), and the message estimate is obtained as

$$\hat{W}_t = d(F_t, 2^{-Rt}/2), \tag{A.11}$$

where the message estimate function $d(F, \delta)$ is defined as

$$d(F, \delta) = \arg\max_w \{F(w + \delta) - F(w - \delta)\}. \tag{A.12}$$

In the following figure, a directed acyclic graph is shown describing the generation of all random variables of interest.

A.0.1 Special Case: BSC

In the special case of BSC the capacity achieving input distribution is $\hat{P}(0) = \hat{P}(1) = 1/2$, and the encoding function becomes

$$X_t = F_{\hat{P}}^{-1}(F_{t-1}(W)) \tag{A.13a}$$

$$= \begin{cases} 0, & F_{t-1}(W) \leq 1/2 \\ 1, & \text{else} \end{cases} \tag{A.13b}$$

$$\stackrel{\text{def}}{=} e(F_{t-1}(W)). \tag{A.13c}$$

Also, the update function $\phi(\cdot,\cdot)$ simplifies to

$$F_t(a) = \begin{cases} 2Q(Y_t|0)F_{t-1}(a), & F_{t-1}(a) \leq 1/2 \\ Q(Y_t|0) + 2Q(Y_t|1)(F_{t-1}(a) - 1/2), & \text{else} \end{cases} \tag{A.14}$$

A.1 Capacity achievability of the PMS (according to [50])

This is a modified version of Shayevitz's original proof in [50]. The only difference from the original proof is notation. For simplicity we will consider DMC from now on. Shayevitz's original proof is for general memoryless channels.

A transmission scheme achieves rate R if

$$\lim_{t\to\infty} P(|W - \hat{W}_t| > 2^{-tR}) = 0. \tag{A.15}$$

In particular, we say that a transmission schemes achieves zero rate if

$$\forall \epsilon > 0 \qquad \lim_{t\to\infty} P(|W - \hat{W}_t| > \epsilon) = 0. \tag{A.16}$$

A.1.1 Zero rate result

Let $\phi_Y(\cdot) \stackrel{\text{def}}{=} \phi(\cdot, Y)$. Then $F_t = \phi_{Y_t}(\phi_{Y_{t-1}}(\cdots \phi_{Y_1}(F_0)))$. For a non-decreasing function $h : [0,1] \to [0,1]$ define a Lyapunov function V_λ as

$$V_\lambda(h) = \int_0^1 \lambda(h(w))dw, \tag{A.17}$$

where $\lambda : [0,1] \to [0,1]$ is onto, strictly concave and symmetric about 0.5. This definition implies that $\lambda(x)$ is 0 at $x = 0,\ 1$ and 1 at $x = 1/2$. Furthermore, for a cdf $F \in \mathcal{F}$, $V_\lambda(F)$ is small if F resembles a step function (it is exactly 0 for a step function). A function $\xi : [0,1] \to [0,1]$ is called contraction if it is nonnegative, concave, and $\xi(x) < x$ for $x \in (0,1)$.

Definition A.2. A channel is called fixed-point free if for any $f_t(w)$

$$P\Big(\phi_{Y_{t+1}}(f_t(w)) = f_t(w)\Big) < 1. \tag{A.18}$$

Lemma A.3. *If the channel is fixed-point free, then for $\epsilon > 0$, and for all $f \in \mathcal{F}$*

$$\lim_{t\to\infty} P(V_\lambda(F_t) > \epsilon | F_0 = f) = 0. \tag{A.19}$$

Proof. We would like to find a contraction mapping ξ such that for every w and f_{t-1} we have $E[\lambda(F_t(w))|f_{t-1}(w)] \leq \xi(\lambda(f_{t-1}(w)))$. Let us assume for now that such a contraction mapping exists. We have

$$\begin{aligned}
E[V_\lambda(F_t)|f_{t-1}] &= E[\int_0^1 \lambda(F_t(w))dw|f_{t-1}] && \text{(A.20a)}\\
&= \int_0^1 E[\lambda(F_t(w))|f_{t-1}(w)]dw && \text{(A.20b)}\\
&\leq \int_0^1 \xi(\lambda(f_{t-1}(w)))dw && \text{(A.20c)}\\
&\leq \xi(\int_0^1 \lambda(f_{t-1}(w))dw) && \text{(A.20d)}\\
&= \xi(V_\lambda(f_{t-1})), && \text{(A.20e)}
\end{aligned}$$

where the first inequality is due to the assumption for the property of ξ and the second inequality is due to the concavity of ξ. Then,

$$\begin{aligned}
P(V_\lambda(F_t) > \epsilon) &\leq \frac{E[V_\lambda(F_t)]}{\epsilon} && \text{(A.21a)}\\
&= \frac{E[E[V_\lambda(\phi_{Y_t}(F_{t-1}))|Y^{t-1}]]}{\epsilon} && \text{(A.21b)}\\
&= \frac{E_{F_{t-1}}[E_{Y_t}[V_\lambda(\phi_{Y_t}(F_{t-1}))|F_{t-1}]]}{\epsilon} && \text{(A.21c)}\\
&\leq \frac{E[\xi(V_\lambda(F_{t-1}))]}{\epsilon} && \text{(A.21d)}\\
&\leq \frac{\xi(E[V_\lambda(F_{t-1})])}{\epsilon} && \text{(A.21e)}\\
&\leq \cdots \leq \frac{\xi^t(E[V_\lambda(F_0)])}{\epsilon} \to 0, && \text{(A.21f)}
\end{aligned}$$

where the first inequality is the Markov inequality, the second inequality is due to (A.20), the third inequality is due to the concavity of ξ, the fourth inequality is due to repeated application of the above inequalities and the convergence to 0 is due to the property of the contraction. Observe that the convergence is true for any initial distribution F_0, which implies that the convergence is uniform in the initial distribution.

It remains to find the contraction ξ with the property $E[\lambda(\phi_{Y_t}(f_{t-1}(w)))] \leq \xi(\lambda(f_{t-1}(w)))$. To this end let $\lambda' : [0, 0.5] \to [0, 1]$ be a restriction of λ on $[0, 0.5]$. Then, λ' becomes one-to-one and onto hence it has inverse. We define ξ^* as

$$\xi^*(x) = \max\{E[\lambda(\phi_{Y_t}(\lambda'^{-1}(x)))], E[\lambda(\phi_{Y_t}(1 - \lambda'^{-1}(x)))]\}. \tag{A.22}$$

Clearly, $\xi^*(x) \geq 0$. We will now show that ξ^* satisfies the aforementioned property. Indeed, let $\theta \overset{\text{def}}{=} f_{t-1}(w)$. If $\theta \in [0, 1/2]$ then $\lambda'^{-1}(\lambda(\theta)) = \theta$ and the first term in the maximization on the r.h.s. of (A.22) equals $E[\lambda(\phi_{Y_t}(\theta))]$. If $\theta \in [1/2, 1]$ then $1 - \lambda'^{-1}(\lambda(\theta)) = \theta$ and the second term in the maximization of the r.h.s. of (A.22) equals $E[\lambda(\phi_{Y_t}(\theta))]$. Thus the property holds. We now need to show that $\xi^*(x) < x$ for all $x \in (0, 1)$. This is equivalent to showing that for every $x \in (0, 1)$, $E[\lambda(\phi_{Y_t}(\lambda'^{-1}(x)))] < x$ and $E[\lambda(\phi_{Y_t}(1-\lambda'^{-1}(x)))] < x$, which is equivalent to showing that for all $\theta \in (0, 1/2)$ we have $E[\lambda(\phi_{Y_t}(\theta))] < \lambda'(\theta)$ and $E[\lambda(\phi_{Y_t}(1 - \theta))] < \lambda'(\theta)$. This in turn is equivalent to showing that $E[\lambda(\phi_{Y_t}(\theta))] < \lambda(\theta)$ for all $\theta \in (0, 1)$. To show this, since the channel is fixed-point free, $F_t(w)$ is not a.s. constant. Hence, using Jensen's inequality we get

$$\begin{aligned} E[\lambda(\phi_{Y_t}(\theta))] &< \lambda(E[\phi_{Y_t}(\theta)]) && \text{(A.23a)} \\ &= \lambda(\sum_y \hat{P}_Y(y)\phi_y(\theta)) && \text{(A.23b)} \\ &= \lambda(\theta), && \text{(A.23c)} \end{aligned}$$

where the last equality is due to (A.10). Thus, ξ^* satisfies all requirements for a contraction mapping except concavity. To this end we define ξ to be the upper convex envelope of ξ^*, i.e.,

$$\xi(a) = \sup\{b : (a,b) \in L\}, \tag{A.24a}$$

$$L = \mathrm{conv}\{(a,b) : a \in [0,1], b \in (0, \xi^*(a))\}. \tag{A.24b}$$

Then, ξ is non-negative and concave. In addition, since $\xi^*(a) \le \xi(a)$, it is clear that $E[\lambda(\phi_{Y_t}(f_{t-1}(w)))] \le \xi(\lambda(f_{t-1}(w)))$. To show that $\xi(\theta) < \theta$, we observe that for any $\theta \in (0,1]$, there must exist some constant $\alpha \in [0,1]$ such that $\theta = \alpha\theta_0 + (1-\alpha)\theta_1$. Then

$$\xi(\theta) \le \alpha\xi^*(\theta_0) + (1-\alpha)\xi^*(\theta_1) < \alpha\theta_0 + (1-\alpha)\theta_1 = \theta, \tag{A.25}$$

where we used the definition of the upper convex envelope in the first inequality.

□

The intuitive interpretation of the above lemma is that the probability of having an F_t that does not resemble a step function is zero at the limit of large t.

We now state the zero rate result.

Proposition A.4. *If the channel is fixed-point free, then for $\epsilon, \delta > 0$*

$$\lim_{t\to\infty} P(F_t(W-\delta) > \epsilon) = 0, \qquad \lim_{t\to\infty} P(F_t(W+\delta) < 1-\epsilon) = 0. \tag{A.26}$$

Proof. Using the symmetry of λ, we can write

$$V_\lambda(F_t) = \int_0^{W_t^*} \lambda(F_t(w))dw + \int_{W_t^*}^1 \lambda(1 - F_t(w))dw, \tag{A.27}$$

where W_t^* is the unique solution of $F_t(w) = 0.5$. Then, for any $\nu > 0$ we have

$$\begin{aligned} P(F_t(W_t^* - \delta) > \nu) &\le P(\lambda(F_t(W_t^* - \delta)) > \nu) && \text{(A.28a)} \\ &\le P(\int_{W_t^*-\delta}^{W_t^*} \lambda(F_t(w))dw > \nu\delta) && \text{(A.28b)} \\ &\le P(V_\lambda(F_t) > \nu\delta), && \text{(A.28c)} \end{aligned}$$

and similarly,

$$P(F_t(W_t^* + \delta) < 1 - \nu) \le P(V_\lambda(F_t) > \nu\delta). \tag{A.29}$$

For any $\eta \in (0, 0.5)$ we have

$$\begin{aligned} &P\left(\int_0^W F_t(w)dw + \int_W^1 (1 - F_t(w))dw > \nu\right) \\ &\le P\left(\int_0^{W_t^*} F_t(w)dw + \int_{W_t^*}^1 (1 - F_t(w))dw > \nu/2\right) + P(|W - W_t^*| > \nu/2) && \text{(A.30a)} \\ &\le P(V_\lambda(F_t) > \nu/2) + P(F_t(W) \notin (\eta, 1 - \eta)) \\ &\qquad + P(F_t(W_t^* - \nu/2) > \eta) + P(F_t(W_t^* + \nu/2) < 1 - \eta) && \text{(A.30b)} \\ &\le P(V_\lambda(F_t) > \nu/2) + 2\eta + 2P(V_\lambda(F_t) > \nu\eta/2). && \text{(A.30c)} \end{aligned}$$

Thus

$$\begin{aligned} P(F_t(W - \delta) > \epsilon) &\le P\left(\int_{W-\delta}^W F_t(w)dw > \delta\epsilon\right) && \text{(A.31a)} \\ &\le P\left(\int_0^W F_t(w)dw > \delta\epsilon\right) && \text{(A.31b)} \\ &\le P\left(\int_0^W F_t(w)dw + \int_W^1 (1 - F_t(w))dw > \delta\epsilon\right) && \text{(A.31c)} \\ &\le P(V_\lambda(F_t) > \delta\epsilon/2) + 2\eta + 2P(V_\lambda(F_t) > \delta\epsilon\eta/2). && \text{(A.31d)} \end{aligned}$$

Setting $\eta = \sqrt{\sup_{a\in[0,1]} \xi^t(a)}/(\delta\epsilon)$, together with Lemma A.3 completes the proof of the first assertion of the lemma. The proof of the second assertion is similar. $\square$

The intuition behind the above proof is that for an error to occur, either the cdf F_t does not behave as a step function (first, third and fourth terms in (A.30b)) or the step does not occur at the transmitted message W (second term in (A.30b)).

Although the previous proposition is sufficient to show achievability of $R = 0$, in the following we provide a stronger version of it, that will be used in the next section for the proof of rate $R > 0$ achievability. For any $t_2 > t_1 > 0$ we can write F_{t_2} as a function of F_{t_1-1} and the observations $Y_{t_1}^{t_2}$ through a repeated application of the ϕ recursion, i.e., $F_{t_2} \stackrel{\text{def}}{=} \phi_{t_2-t_1}(F_{t_1-1}, Y_{t_1}^{t_2})$. Let F_{t_1,t_2}^f be the random variable defined as $F_{t_1,t_2}^f \stackrel{\text{def}}{=} \phi_{t_2-t_1}(f, Y_{t_1}^{t_2})$. Clearly $F_t = F_{1,t}^u$, where u denotes the uniform distribution over $(0,1)$. In addition, due to the recursion implied by the PMS, we will denote $F_{t_2}(a) \stackrel{\text{def}}{=} \phi_{t_2-t_1}(F_{t_1-1}, Y_{t_1}^{t_2})(a) \stackrel{\text{def}}{=} \phi_{t_2-t_1}(F_{t_1-1}(a), Y_{t_1}^{t_2})$ with some notational abuse. With the above notation, the function $\phi_{t_2-t_1}(\cdot, Y_{t_1}^{t^2})$ is monotonically increasing. We now prove the following lemma which is a stronger version of Lemma A.3.

Lemma A.5. *If the channel is fixed-point free, then for $\epsilon > 0$ and for any $\alpha > 0$,*

$$\lim_{t\to\infty} P(\max_{1\le t'\le t} V_\lambda(F_{t'+1,(1+\alpha)t}^u) > \epsilon) = 0. \tag{A.32}$$

Proof. Let $V_{\lambda,t_1,t_2}^* = \sup_f V_\lambda(F_{t_1,t_2}^f)$. Note that V_{t_1,t_2}^* is a deterministic function of

$Y_{t_1}^{t_2}$, hence there exists a sequence of cdfs $\{f_{k,Y_{t_1}^{t_2}}\}_{k=1}^{\infty}$ such that

$$V^*_{\lambda,t_1,t_2} = \lim_{k\to\infty} V_\lambda(F_{t_1,t_2}^{f_{k,Y_{t_1}^{t_2}}}) \tag{A.33a}$$

$$= \lim_{k\to\infty} V_\lambda(F_{t_1+1,t_2}^{\phi_{Y_{t_1}}(f_{k,Y_{t_1}^{t_2}})}) \tag{A.33b}$$

$$\leq \sup_f V_\lambda(F^f_{t_1+1,t_2}) \tag{A.33c}$$

$$= V^*_{\lambda,t_1+1,t_2}. \tag{A.33d}$$

Note that there exists a sequence of cdfs $\{f_{k,Y_1^{\alpha t}}\}_{k=1}^{\infty}$ such that $V^*_{\lambda,1,\alpha t} \geq V_\lambda(F_{1,\alpha t}^{f_{k,Y_1^{\alpha t}}}) > V^*_{\lambda,,\alpha t} - 1/k$. Therefore,

$$P(\max_{1\leq t'\leq t} V_\lambda(F^u_{t'+1,(1+\alpha)t} > \epsilon) \leq P(\max_{1\leq t'\leq t} V^*_{\lambda,t'+1,(1+\alpha)t} > \epsilon) \tag{A.34a}$$

$$= P(V^*_{\lambda,t+1,(1+\alpha)t} > \epsilon) \tag{A.34b}$$

$$= P(\sup_f V_\lambda(F^f_{t+1,(1+\alpha)t}) > \epsilon) \tag{A.34c}$$

$$\overset{(a)}{=} P(\sup_f V_\lambda(F^f_{1,\alpha t}) > \epsilon) \tag{A.34d}$$

$$= P(\lim_{k\to\infty} V_\lambda(F_{1,\alpha t}^{f_{k,Y_1^{\alpha t}}}) > \epsilon) \tag{A.34e}$$

$$\overset{(b)}{\leq} P(V_\lambda(F_{1,\alpha t}^{f_{k,Y_1^{\alpha t}}}) > \epsilon) \tag{A.34f}$$

$$\overset{(c)}{\to} 0, \tag{A.34g}$$

where (a) is due to the fact that $F^f_{t+1,(1+\alpha)t} = \phi_{\alpha t-1}(f, Y_{t+1}^{(1+\alpha)t})$, $F^f_{1,\alpha t} = \phi_{\alpha t-1}(f, Y_1^{\alpha t})$, and $Y_{t+1}^{(1+\alpha)t}$, $Y_1^{\alpha t}$ have the same statistics; (b) is true for $k' > 1/(V^*_{\lambda,1,\alpha t} - \epsilon)$; and (c) is due to the fact that Lemma A.3 holds for any F_0. □

Observe that indeed this lemma is stronger than Lemma A.3, since $P(V_\lambda(F_t) > \epsilon) = P(V_\lambda(F^u_{1,t}) > \epsilon) = P(V_\lambda(F^u_{1+t,t+t}) > \epsilon) \leq P(\max_{1\leq t'\leq t} V_\lambda(F^u_{t'+1,2t}) > \epsilon)$.

Proposition A.6. *If the channel is fixed-point free, then for* $\epsilon, \delta > 0$

$$\lim_{t\to\infty} P\Big(\max_{1\le t'\le t} F^u_{t'+1,(1+\alpha)t}\Big(F_{t'}(W) - \delta\Big) > \epsilon\Big) = 0, \tag{A.35}$$

$$\lim_{t\to\infty} P\Big(\max_{1\le t'\le t} F^u_{t'+1,(1+\alpha)t}\Big(F_{t'}(W) + \delta\Big) < 1 - \epsilon\Big) = 0. \tag{A.36}$$

Proof. Using the symmetry of λ, we can write

$$\begin{aligned} &\max_{1\le t'\le t} V_\lambda(F^u_{t'+1,(1+\alpha)t}) \\ &= \max_{1\le t'\le t} \Big[\int_0^{W^*_{t',t}} \lambda(F^u_{t'+1,(1+\alpha)t}(w))dw + \int_{W^*_{t',t}}^1 \lambda(1 - F^u_{t'+1,(1+\alpha)t}(w))dw\Big], \end{aligned} \tag{A.37}$$

where $W^*_{t',t}$ is the unique solution of $F^u_{t'+1,(1+\alpha)t}(w) = 0.5$. Then, we have

$$P(\max_{1\le t'\le t} F^u_{t'+1,(1+\alpha)t}(W^*_{t',t} - \delta) > \nu)$$

$$\le P(\max_{1\le t'\le t} \lambda(F^u_{t'+1,(1+\alpha)t}(W^*_{t',t} - \delta)) > \nu) \tag{A.38a}$$

$$\le P(\max_{1\le t'\le t} \int_{W^*_{t',t}-\delta}^{W^*_{t',t}} \lambda(F^u_{t'+1,(1+\alpha)t}(w))dw > \nu\delta) \tag{A.38b}$$

$$\le P(\max_{1\le t'\le t} V_\lambda(F^u_{t'+1,(1+\alpha)t}) > \nu\delta), \tag{A.38c}$$

and similarly

$$P(\max_{1\le t'\le t} \Big[1 - F^u_{t'+1,(1+\alpha)t}(W^*_{t',t} + \delta)\Big] < 1 - \nu)$$

$$\le P(\max_{1\le t'\le t} V_\lambda(F^u_{t'+1,(1+\alpha)t}) > \nu\delta). \tag{A.39}$$

For any $\eta \in (0, 0.5)$

$$
\begin{aligned}
& P\bigg(\max_{1 \le t' \le t} \Big[\int_0^{F_{t'}(W)} F^u_{t'+1,(1+\alpha)t}(w)dw + \int_{F_{t'}(W)}^1 (1 - F^u_{t'+1,(1+\alpha)t}(w))dw \Big] \\
& \qquad > \nu \bigg) \\
& \le P\bigg(\max_{1 \le t' \le t} \Big[\int_0^{W^*_{t',t}} F^u_{t'+1,(1+\alpha)t}(w)dw + \int_{W^*_{t',t}}^1 (1 - F^u_{t'+1,(1+\alpha)t}(w))dw \Big] \\
& \qquad > \nu/2 \bigg) \\
& \quad + P\bigg(\max_{1 \le t' \le t} |F_{t'}(W) - W^*_{t',t}| > \nu/2 \bigg) \qquad \text{(A.40a)}
\end{aligned}
$$

$$
\begin{aligned}
& \le P\bigg(\max_{1 \le t' \le t} V_\lambda(F^u_{t'+1,(1+\alpha)t}) > \nu/2 \bigg) \\
& \quad + P\bigg(\Big\{ F_{(1+\alpha)t}(W) < \max_{1 \le t' \le t} F^u_{t'+1,(1+\alpha)t}(W^*_{t',t} - \nu/2) \Big\} \\
& \qquad \cup \Big\{ F_{(1+\alpha)t}(W) > \min_{1 \le t' \le t} F^u_{t'+1,(1+\alpha)t}(W^*_{t',t} + \nu/2) \Big\} \bigg) \qquad \text{(A.40b)} \\
& \le P\bigg(\max_{1 \le t' \le t} V_\lambda(F^u_{t'+1,(1+\alpha)t}) > \nu/2 \bigg) + P\bigg(F_{(1+\alpha)t}(W) \notin (\eta, 1 - \eta) \bigg) \\
& \quad + P\bigg(\max_{1 \le t' \le t} F^u_{t'+1,(1+\alpha)t}(W^*_{t',t} - \nu/2) > \eta \bigg) \\
& \quad + P\bigg(\max_{1 \le t' \le t} [1 - F^u_{t'+1,(1+\alpha)t}(W^*_{t',t} + \nu/2)] < 1 - \eta \bigg) \qquad \text{(A.40c)} \\
& \le P(\max_{1 \le t' \le t} V_\lambda(F^u_{t'+1,(1+\alpha)t}) > \nu/2) \\
& \quad + 2\eta + 2P(\max_{1 \le t' \le t} V_\lambda(F^u_{t'+1,(1+\alpha)t}) > \nu\eta/2). \qquad \text{(A.40d)}
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&P\Big(\max_{1\le t'\le t} F^u_{t'+1,(1+\alpha)t}\Big(F_{t'}(W)-\delta\Big) > \epsilon\Big) \\
&\le P\Big(\max_{1\le t'\le t} \int_{F_{t'}(W)-\delta}^{F_{t'}(W)} F^u_{t'+1,(1+\alpha)t}(w)dw > \delta\epsilon\Big) && \text{(A.41a)} \\
&\le P\Big(\max_{1\le t'\le t} \int_{0}^{F_{t'}(W)} F^u_{t'+1,(1+\alpha)t}(w)dw > \delta\epsilon\Big) && \text{(A.41b)} \\
&\le P\Big(\max_{1\le t'\le t} \Big[\int_{0}^{F_{t'}(W)} F^u_{t'+1,(1+\alpha)t}(w)dw + \int_{F_{t'}(W)}^{1} (1-F^u_{t'+1,(1+\alpha)t}(w))dw\Big] \\
&\qquad > \delta\epsilon\Big) && \text{(A.41c)} \\
&\le P(\max_{1\le t'\le t} V_\lambda(F^u_{t'+1,(1+\alpha)t}) > \delta\epsilon/2) \\
&\quad + 2\eta + 2P(\max_{1\le t'\le t} V_\lambda(F^u_{t'+1,(1+\alpha)t}) > \delta\epsilon\eta/2). && \text{(A.41d)}
\end{aligned}
$$

Setting $\eta = \sqrt{\sup_{a\in[0,1]} \xi^t(a)}/(\delta\epsilon)$, together with Lemma A.5 completes the proof of the first assertion of the proposition. The proof of the second assertion is similar. □

A.1.2 Rate $R < C$ achievability

Lemma A.7. *$(F_t(W), Y_{t+1})_t$ is a Markov chain.*

Proof. Let $\Theta_t \stackrel{\text{def}}{=} F_t(W)$. We have

$$
\begin{aligned}
&P(\theta_t, y_{t+1}|\theta^{t-1}, y^t) \\
&\quad = Q(y_{t+1}|e(\theta_t))P(\theta_t|\theta_{t-1}, y^t) && \text{(A.42a)} \\
&\quad = Q(y_{t+1}|e(\theta_t))\delta_{\phi(\theta_{t-1},y_t)}(\theta_t) && \text{(A.42b)} \\
&\quad = P(\theta_t, y_{t+1}|\theta_{t-1}, y_t). && \text{(A.42c)}
\end{aligned}
$$

□

Lemma A.8. $E\{\log \frac{dF_{t+1}(W)}{dF_t(W)}|\} = C$ *if the PMS is used as the transmission scheme.*

Proof. First we show that $I(W;Y_{t+1}|Y^t) = E\left[\log \frac{dF_{t+1}(W)}{dF_t(W)}\right]$.

$$\begin{aligned}
&I(W;Y_{t+1}|Y^t) \\
&\quad = E\left[\log\left\{\frac{P(Y_{t+1}|Y^t,W}{P(Y_{t+1}|Y^t,)}\right\}\right] && \text{(A.43a)}\\
&\quad = E\left[\log\left\{\frac{Q(Y_{t+1}|e(F_t(W)))}{\hat{P}_Y(Y_{t+1})}\right\}\right] && \text{(A.43b)}\\
&\quad = E\left[\log\frac{dF_{t+1}(W)}{dF_t(W)}\right], && \text{(A.43c)}
\end{aligned}$$

where the last equality is due to (5.40).

Now we show that $I(W;Y_t|Y^{t-1}) = C$. Note that for a given Y^{t-1}, $X_t = F_{\hat{P}}^{-1}(F_{t-1}(W))$ is $\hat{P}$ distributed and hence is independent of Y^{t-1}. Then,

$$\begin{aligned}
&I(W;Y_t|Y^{t-1}) \\
&\quad = H(Y_t|Y^{t-1}) - H(Y_t|Y^{t-1},W) && \text{(A.44a)}\\
&\quad \overset{(a)}{=} H(Y_t|Y^{t-1}) - H(Y_t|Y^{t-1},W,X_t) && \text{(A.44b)}\\
&\quad \overset{(b)}{=} H(Y_t) - H(Y_t|X_t) && \text{(A.44c)}\\
&\quad = I(X_t;Y_t) && \text{(A.44d)}\\
&\quad = C, && \text{(A.44e)}
\end{aligned}$$

where (a) is due to the fact that X_t is a function of Y^{t-1}, W; (b) is due to the channel being DMC and (4.6b); and the last equation is due to the fact that the channel input sequence for the PMS has distribution which is capacity-achieving. □

Lemma A.9. *If the Markov chain $\{F_t(W), Y_{t+1}\}_1^\infty$ has ergodic invariant distribution, then*

$$\lim_{t\to\infty}\frac{1}{t}\log dF_t(W) = C \ \ a.s. \tag{A.45}$$

Proof.

$$\frac{1}{t}\log dF_t(W) = \frac{1}{t}\sum_{s=1}^{t}\log\frac{dF_s(W)}{dF_{s-1}(W)}. \tag{A.46}$$

If $(F_t(W), Y_{t+1})_t$ has ergodic invariant distribution, then by the strong law of large numbers for Markov chains

$$\lim_{t\to\infty}\frac{1}{t}\log dF_t(W) = E[\log\frac{dF_t(W)}{dF_{t-1}(W)}] \overset{(a)}{=} C \qquad \text{a.s.} \tag{A.47}$$

where (a) is from Lemma A.8. □

Lemma A.10. *If the Markov chain $\{F_t(W), Y_{t+1}\}_1^\infty$ has ergodic invariant distribution, then for any $\delta > 0$ and rate $R < C - \delta$ there exists $\epsilon' > 0$ so that for all $\epsilon \leq \epsilon'$*

$$\lim_{t\to\infty} P\left(\bigcap_{s=1}^{t}\{F_s(W) - F_s(W - 2^{-tR}) < \epsilon\}\right) = 0 \tag{A.48}$$

$$\lim_{t\to\infty} P\left(\bigcap_{s=1}^{t}\{F_s(W + 2^{-tR}) - F_s(W) < \epsilon\}\right) = 0. \tag{A.49}$$

Proof. Let $\Theta_t \overset{\text{def}}{=} F_t(W)$, and $P_{Y_{t+1}|\Theta_t} \overset{\text{def}}{=} Q'_{Y_{t+1}|e(\Theta_t)}$. For any $\epsilon > 0$, define ${}^-P^\epsilon_{Y_{t+1}|\Theta_t}$ to be

$${}^-P^\epsilon_{Y_{t+1}|\Theta}(y|\theta) \overset{\text{def}}{=} \inf_{\theta-\epsilon<\theta'<\theta} P_{Y_{t+1}|\Theta}(y|\theta'). \tag{A.50}$$

Similarly, define ${}^+P^\epsilon_{Y_{t+1}|\Theta_t}$ to be

$${}^+P^\epsilon_{Y_{t+1}|\Theta_t}(y|\theta) = \inf_{\theta<\theta'<\theta+\epsilon} P_{Y_{t+1}|\Theta_t}(y|\theta'). \tag{A.51}$$

From now on we prove the first assertion of the lemma; the second assertion follows in a similar way. Define

$$C_\epsilon^- = E\left[\log\left\{\frac{{}^-P^\epsilon(Y_{t+1}|\Theta_t)}{P(Y_{t+1})}\right\}\right] \tag{A.52a}$$

$$= \int_{\theta_t}\sum_{y_{t+1}} P(y_{t+1}|\theta_t)\log\left\{\frac{{}^-P^\epsilon(y_{t+1}|\theta_t)}{P(y_{t+1})}\right\}. \tag{A.52b}$$

Note that

$$C = E\left[\log\left\{\frac{P(Y_{t+1}|\Theta_t)}{P(Y_{t+1})}\right\}\right] \tag{A.53a}$$

$$= \int_{\theta_t}\sum_{y_{t+1}} P(y_{t+1}|\theta_t)\log\left\{\frac{P(y_{t+1}|\theta_t)}{P(y_{t+1})}\right\}. \tag{A.53b}$$

Then we have,

$$C - C_\epsilon^- = D(P_{Y_{t+1}|\Theta_t}||{}^-P^\epsilon_{Y_{t+1}|\Theta_t}|P_{\Theta_t}) \geq 0. \tag{A.54}$$

Assume that

$$\inf_{\epsilon>0} D(P_{Y_{t+1}|\Theta_t}||{}^-P^\epsilon_{Y_{t+1}|\Theta_t}|P_{\Theta_t}) < \infty. \tag{A.55}$$

Then,$-\infty < C_\epsilon^- \leq C$. Therefore, $P_{Y_{t+1},\Theta_t}\log\left\{\frac{{}^-P^\epsilon_{Y_{t+1}|\Theta_t}}{P_{Y_{t+1}}}\right\}$ is finitely integrable and converges to $P_{Y_{t+1},\Theta_t}\log\left\{\frac{P_{Y_{t+1}|\Theta_t}}{P_{Y_{t+1}}}\right\}$ a.e. in a monotonically nondecreasing fashion as $\epsilon \to 0$. Hence, by the monotone convergence theorem,

$$\lim_{\epsilon\to 0} C_\epsilon^- = C, \tag{A.56}$$

and for $\delta > 0$, there exists ϵ'_δ such that for all $\epsilon \leq \epsilon'_\delta$

$$C_\epsilon^- > C - \delta/2. \tag{A.57}$$

Furthermore, we can apply the strong law of large numbers for Markov chains to get

$$\lim_{t\to\infty} \frac{1}{t} \sum_{s=0}^{t-1} \log \left\{ \frac{{}^-P^\epsilon(Y_{s+1}|\Theta_s)}{P(Y_{s+1})} \right\} = C_\epsilon^- \qquad \text{a.s.} \tag{A.58}$$

Note that if $F_s(W) - F_s(W - 2^{-tR}) < \epsilon$ for $0 \le s \le t-1$

$$\begin{aligned} \inf_{W-2^{-tR}<w<W} \frac{1}{t} \log dF_t(w) &\ge \frac{1}{t} \sum_{s=0}^{t-1} \log \inf_{F_s(W-2^{-tR})<\theta_s<F_s(W)} \frac{P(Y_{s+1}|\theta_s)}{P(Y_{s+1})} & \text{(A.59a)} \\ &\ge \frac{1}{t} \sum_{s=0}^{t-1} \log \left\{ \frac{{}^-P^\epsilon(Y_{s+1}|\Theta_s)}{P(Y_{s+1})} \right\}. & \text{(A.59b)} \end{aligned}$$

Therefore,

$$\begin{aligned} & P\left(\bigcap_{s=0}^{t-1} \{F_s(W) - F_s(W - 2^{-tR}) < \epsilon\} \right) \\ & \le P\left(\bigcap_{s=0}^{t-1} \{F_s(W) - F_s(W - 2^{-tR}) < \epsilon\} \right. \\ & \qquad \left. \bigcap \{ \frac{1}{t} \sum_{s=0}^{t-1} \log \left\{ \frac{{}^-P^\epsilon(Y_{s+1}|\Theta_s)}{P(Y_{s+1})} \right\} \ge C - \delta/2 \} \right) \\ & \quad + P\left(\frac{1}{t} \sum_{s=0}^{t-1} \log \left\{ \frac{{}^-P^\epsilon(Y_{s+1}|\Theta_s)}{P(Y_{s+1})} \right\} < C - \delta/2 \right) \end{aligned} \tag{A.60a}$$

$$\overset{(a)}{\leq} P\Big(\inf_{W-2^{-tR}<w<W} \frac{1}{t} \log dF_t(w) \geq C - \delta/2 \Big)$$

$$+ P\Big(\frac{1}{t} \sum_{s=0}^{t-1} \log \Big\{ \frac{{}^{-}P^{\epsilon}(Y_{s+1}|\Theta_s)}{P(Y_{s+1})} \Big\} < C - \delta/2 \Big) \tag{A.60b}$$

$$\overset{(b)}{\leq} P\Big(\inf_{W-2^{-tR}<w<W} \frac{1}{t} \log dF_t(w) \geq C - \delta/2 \Big) + o(1) \tag{A.60c}$$

$$\leq P\Big(\int_{W-2^{-tR}}^{W} dF_t(w) \geq 2^{t(C-R-\delta/2)} \Big) + o(1) \tag{A.60d}$$

$$\leq P\Big(\int_{W-2^{-tR}}^{W} dF_t(w) \geq 2^{t\delta/2} \Big) + o(1) \tag{A.60e}$$

$$\overset{(c)}{\to} 0, \tag{A.60f}$$

where (a) is due to (A.59), (b) is due to (A.58), and (c) is from the fact that $\int_{W-2^{-tR}}^{W} dF_t(w) \leq 1$. □

The above lemma guarantees that at some time before t there will be a jump of at least ϵ in the posterior message cdf in the interval of 2^{-tR} around W. Using this lemma we show the main result.

Proposition A.11. *If a Markov chain $\{F_t(W), Y_{t+1}\}_1^\infty$ has ergodic invariant distribution, then for $\delta, \alpha > 0$*

$$\lim_{t\to\infty} P(F_{(1+\alpha)t}(W - 2^{-tR}) > \delta) = 0, \tag{A.61a}$$

$$\lim_{t\to\infty} P(F_{(1+\alpha)t}(W + 2^{-tR}) < 1 - \delta) = 0. \tag{A.61b}$$

Proof. For every $\epsilon > 0$ we have

$$\begin{aligned}
&P(F_{(1+\alpha)t}(W - 2^{-tR}) > \delta) \\
&\leq P\Bigg(\bigcap_{t'=1}^{t} \Big\{F_{t'}(W - 2^{-tR}) > F_{t'}(W) - \epsilon\Big\}\Bigg) \\
&\quad + P\Bigg(\Big\{F_{(1+\alpha)t}(W - 2^{-tR}) > \delta)\Big\} \\
&\qquad\qquad \cap \bigcup_{t'=1}^{t} \Big\{F_{t'}(W - 2^{-tR}) \leq F_{t'}(W) - \epsilon\Big\}\Bigg).
\end{aligned} \tag{A.62}$$

Regarding the first quantity we have

$$P\Bigg(\bigcap_{t'=1}^{t} \Big\{F_{t'}(W - 2^{-tR}) > F_{t'}(W) - \epsilon\Big\}\Bigg) \overset{(a)}{\to} 0, \tag{A.63}$$

where (a) is from Lemma A.10.

Consider an event $\bigcup_{t'=1}^{t} \Big\{F_{t'}(W - 2^{-tR}) \leq F_{t'}(W) - \epsilon\Big\}$. This event is equivalent to $\bigcup_{t'=1}^{t} \{W - 2^{-tR} \leq W'\}$ where $W' = F_{t'}^{-1}(F_{t'}(W) - \epsilon)$. Then by the monotonicity of cdf it is equivalent to $\bigcup_{t'=1}^{t} \{F_t(W - 2^{-tR}) \leq F_t(W')\}$ which again is equivalent to $\Big\{F_t(W - 2^{-tR}) \leq \max_{1 \leq t' \leq t} F_t\Big(W'\Big)\Big)\Big\}$.

Therefore, regarding the second quantity we have

$$
\begin{aligned}
&P\bigg(\Big\{F_{(1+\alpha)t}(W-2^{-tR})>\delta)\Big\} \\
&\qquad \cap \bigcup_{t'=1}^{t}\Big\{F_{t'}(W-2^{-tR}) \le F_{t'}(W)-\epsilon\Big\}\bigg) \\
=&P\bigg(\Big\{F_{(1+\alpha)t}(W-2^{-tR})>\delta)\Big\} \\
&\qquad \cap \Big\{F_t(W-2^{-tR}) \le \max_{1\le t'\le t} F_t\Big(F_{t'}^{-1}\Big(F_{t'}(W)-\epsilon\Big)\Big)\Big\}\bigg) && \text{(A.64a)} \\
\overset{(a)}{\le}& P\bigg(\Big\{F_{(1+\alpha)t}(W-2^{-tR})>\delta)\Big\} \\
&\qquad \cap \Big\{F_{(1+\alpha)t}(W-2^{-tR}) \le \max_{1\le t'\le t} F_{(1+\alpha)t}\Big(F_{t'}^{-1}\Big(F_{t'}(W)-\epsilon\Big)\Big)\Big\}\bigg) && \text{(A.64b)} \\
=& P\bigg(\max_{1\le t'\le t} F_{(1+\alpha)t}\Big(F_{t'}^{-1}\Big(F_{t'}(W)-\epsilon\Big)\Big)>\delta\bigg) && \text{(A.64c)} \\
\overset{(b)}{=}& P\bigg(\max_{1\le t'\le t} F^{u}_{t'+1,(1+\alpha)t}\Big(F_{t'}(W)-\epsilon\Big)>\delta\bigg) \overset{(c)}{\to} 0. && \text{(A.64d)}
\end{aligned}
$$

where (a) is from the monotonicity of cdf, (b) is due to the PMS update, and (c) is from Proposition 5.5. $P(F_{(1+\alpha)t}(W+2^{-tR})<1-\delta)$ can also be evaluated in a similar way. □

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] C. E. Shannon, "A mathematical theory of communication," *Bell System Tech. J.*, vol. 27, pp. 379–423, July 1948.

[2] ——, "A mathematical theory of communication," *Bell System Tech. J.*, vol. 27, pp. 623–656, Oct. 1948.

[3] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon limit error-correcting coding and decoding: turbo-codes," in *Proc. International Conf. Communications*, Geneva, Switzerland, May 1993, pp. 1064–1070.

[4] R. G. Gallager, *Low-Density Parity-Check Codes.* Cambridge, MA: MIT Press, 1963.

[5] D. J. C. MacKay and R. M. Neal, "Near Shannon limit performance of low density parity check codes," *IEE Electronics Letters*, vol. 32, no. 18, pp. 1645–1646, Aug. 1996.

[6] M. G. Luby, M. Mitzenmacher, M. A. Shokrollahi, D. A. Spielman, and V. Stemann, "Practical loss-resilient codes," El Paso, Texas, May 1997, pp. 150–159.

[7] T. J. Richardson and R. L. Urbanke, "The capacity of low-density parity-check codes under message-passing decoding," *IEEE Trans. Information Theory*, vol. 47, no. 2, pp. 599–618, Feb. 2001.

[8] C. E. Shannon, "The zero error capacity of a noisy channel," *IRE Trans. Information Theory*, vol. IT-2, no. 3, pp. 8–19, Sept. 1956.

[9] R. G. Gallager, *Information theory and Reliable Communications.* New York: John Wiley & Sons, 1968.

[10] P. Elias, "Coding for noisy channels," in *IRE Conv. Rec.*, 1955, pp. 37–46.

[11] M. A. Shokrollahi, "New sequences of linear time erasure codes approaching channel capacity," in *Proc. International Symposium on Information Theory and its Applications*, Honolulu, Hawaii, Nov. 1999, pp. 65–76.

[12] A. Shokrollahi, "Capacity-achieving sequences," in *Codes, Systems, and Graphical Models*, ser. IMA Volumes in Mathematics and its Applications, B. Marcus and J. Rosenthal, Eds. Springer-Verlag, 2000, vol. 123, pp. 153–166.

[13] P. Oswald and A. Shokrollahi, "Capacity-achieving sequences for the erasure channel," *IEEE Trans. Information Theory*, vol. 48, no. 12, pp. 3017–3028, Dec. 2002.

[14] H. D. Pfister, I. Sason, and R. Urbanke, "Capacity-achieving ensembles for the binary erasure channel with bounded complexity," *IEEE Trans. Information Theory*, vol. 51, no. 7, pp. 2352–2379, July 2005.

[15] A. Montanari, "Tight bounds for LDPC and LDGM codes under MAP decoding," *IEEE Trans. Information Theory*, vol. 51, no. 9, pp. 3221–3246, Sept. 2005.

[16] C. Measson, A. Montanari, and R. Urbanke, "The generalized area theorem and some of its consequences," *IEEE Trans. Information Theory*, (Submitted).

[17] N. Macris, "Sharp bounds on generalized EXIT functions," *IEEE Trans. Information Theory*, vol. 53, no. 7, pp. 2365–2375, July 2007.

[18] T. Richardson and R. Urbanke, *Modern Coding Theory.* Cambridge University Press, 2007.

[19] G. Miller and D. Burshtein, "Bounds on the maximum-likelihood decoding error probability of low-density parity-check codes," *IEEE Trans. Information Theory*, vol. 47, no. 7, pp. 2696–2710, Nov. 2001.

[20] I. Sason and R. Urbanke, "Parity-check density versus performance of binary linear block codes over memoryless symmetric channels," *IEEE Trans. Information Theory*, vol. 49, no. 7, pp. 1611–1635, July 2003.

[21] G. Wiechman and I. Sason, "Parity-check density versus performance of binary linear block codes over memoryless symmetric channels: New bounds and applications," *IEEE Trans. Information Theory*, vol. 53, no. 2, pp. 550–579, Feb. 2007.

[22] C.-H. Hsu and A. Anastasopoulos, "Capacity achieving LDPC codes through puncturing," *IEEE Trans. Information Theory*, Apr. 2008, (Accepted for publication).

[23] ——, "Capacity-achieving codes for noisy channels with bounded graphical complexity and maximum likelihood decoding," *IEEE Trans. Information Theory*, Feb. 2006, (Submitted).

[24] M. Mushkin and I. Bar-David, "Capacity and coding for the Gilbert-Elliott channels," *IEEE Trans. Information Theory*, vol. 35, no. 6, pp. 1277–1290, Nov. 1989.

[25] A. Goldsmith and P. Vaaiya, "Capacity, mutual information and coding for finfite-state Markov channels," *IEEE Trans. Information Theory*, vol. 42, pp. 868–886, May 1996.

[26] T. Holliday, A. Goldsmith, and P. Glynn, "Capacity of finite state channels based on Lyapunov exponents of random matrices," *IEEE Trans. Information Theory*, vol. 52, no. 8, pp. 3509–3532, Aug. 2006.

[27] D. M. Arnold and H.-A. Loeliger, "On the information rate of binary-input channels with memory," in *Proc. International Conf. Communications*, Helsinki, Finland, June 2001, pp. 2692–2695.

[28] V. Sharma and S. K. Singh, "Entropy and channel capacity in the regenerative setup with applications to Markov channels," in *Proc. International Symposium on Information Theory*, Washington DC, USA, June 2001, p. 283.

[29] H. D. Pfister, J. B. Soriaga, and P. H. Siegel, "On the achievable information rates of finite state ISI channels," in *Proc. Globecom Conf.*, San Antonio, USA, Nov. 2001, pp. 2992–2996.

[30] D. M. Arnold, H.-A. Loeliger, P. O. Vontobel, A. Kavcic, and W. Zeng, "Simulation-based computation of information rates for channels with memory," *IEEE Trans. Information Theory*, vol. 52, no. 8, pp. 3498–3508, Aug. 2006.

[31] A. Kavcic, "On the capacity of Markov sources over noisy channels," in *Proc. Globecom Conf.*, San Antonio, USA, Nov. 2001, pp. 2997–3001.

[32] P. O. Vontobel, A. Kavcic, D. M. Arnold, and H.-A. Loeliger, "A generalization of the Blahut-Arimoto algorithm to finite-state channels," *IEEE Trans. Information Theory*, vol. 54, no. 5, pp. 1887–1918, May 2008.

[33] P. O. Vontobel and D. M. Arnold, "An upper bound on the capacity of channels with memory and constraint input," in *Proc. Information Theory Workshop*, Cairns, Australia, Sept. 2001, pp. 147–149.

[34] S. Yang, A. Kavcic, and S. Tatikonda, "Feedback capacity of finite-state machine channels," *IEEE Trans. Information Theory*, vol. 51, no. 3, pp. 799–810, Mar. 2005.

[35] J. Chen and P. H. Siegel, "Markov processes asymptotically achieve the capacity of finite-state intersymbol interference channels," *IEEE Trans. Information Theory*, vol. 54, no. 3, pp. 1295–1303, Mar. 2008.

[36] X. Ma, N. Varnica, and A. Kavcic, "Matched information rate codes for binary ISI channels," in *Proc. International Symposium on Information Theory*, Lausanne, Switzerland, June 2002, p. 269.

[37] J. B. Soriaga and P. H. Siegel, “On near-capacity coding systems for partial-response channels,” in *Proc. International Symposium on Information Theory*, Chicago, USA, June 2004, p. 263.

[38] A. Kavcic, X. Ma, and N. Varnica, “Matched information rate codes for partial response channels,” *IEEE Trans. Information Theory*, vol. 51, no. 3, pp. 973–989, Mar. 2005.

[39] J. B. Soriaga, H. D. Pfister, and P. H. Siegel, “Determining and approaching achievable rates of binary intersymbol interference channels using multistage decoding,” *IEEE Trans. Information Theory*, vol. 53, no. 4, pp. 1416–1429, Apr. 2007.

[40] H. D. Pfister and P. H. Siegel, “Joint iterative decoding of ldpc codes for channels with memory and erasure noise,” *IEEE J. Select. Areas Commun.*, vol. 26, no. 2, pp. 320–327, Feb. 2008.

[41] A. Kavcic, X. Ma, and M. Mitzenmacher, “Binary intersymbol interference channels: Gallager codes, density evolution, and code performance bounds,” *IEEE Trans. Information Theory*, vol. 49, no. 7, pp. 1636–1652, July 2003.

[42] A. W. Eckford, F. R. Kschischang, and S. Pasupathy, “Analysis of low-density parity-check codes for the Gilbert-Elliott channel,” *IEEE Trans. Information Theory*, vol. 51, no. 11, pp. 3872–3889, Nov. 2005.

[43] ——, “On designing good LDPC codes for Markov channels,” *IEEE Trans. Information Theory*, vol. 53, no. 1, pp. 5–21, Jan. 2007.

[44] P. Grover and A. K. Chaturvedi, “Upper bounds on the rate of LDPC codes for Gilbert-Elliott channels,” in *Proc. Information Theory Workshop*, 2004, pp. 230–235.

[45] ——, “Upper bounds on the rate of LDPC codes for a class of finite-state Markov channels,” *IEEE Trans. Information Theory*, vol. 53, no. 2, pp. 794–804, Feb. 2007.

[46] M. Horstein, “Sequential transmission using noiseless feedback,” *IEEE Trans. Information Theory*, vol. IT-9, no. 4, pp. 136–143, July 1963.

[47] J. Schalkwijk and T. Kailath, “A coding scheme for additive noise channels with feedback–I: No bandwidth constraint,” vol. 12, no. 2, pp. 172–182, Apr 1966.

[48] O. Shayevitz and M. Feder, “Communication with feedback via posterior matching,” in *Proc. International Symposium on Information Theory*, Nice, France, June 2007.

[49] ——, “The posterior matching feedback scheme: Capacity achieving and error analysis,” in *Proc. International Symposium on Information Theory*, Toronto, Canada, July 2008.

[50] ——, “Optimal feedback communication via posterior matching,” *IEEE Trans. Information Theory*, vol. 57, no. 3, pp. 1186–1222, Mar. 2011.

[51] T. P. Coleman, “A stochastic control aproach to ’posterior matching’-style feedback communication schemes,” in *Proc. International Symposium on Information Theory*, Seoul, Korea, June 2009.

[52] H. Viswanathan, “Capacity of Markov channels with receriver CSI and delayed feedback,” *IEEE Trans. Information Theory*, vol. 45, no. 2, pp. 761–771, Mar. 1999.

[53] J. Chen and T. Berger, “The capacity of finite-state Markov channels with feedback,” *IEEE Trans. Information Theory*, vol. 51, no. 3, pp. 780–798, Mar. 2005.

[54] S. Tatikonda and S. Mitter, “The capacity of channels with feedback,” *IEEE Trans. Information Theory*, vol. 55, no. 1, pp. 323–349, Jan. 2009.

[55] G. Como, S. Yuksel, and S. Tatikonda, “The error exponent of variable-length codes over Markov channels with feedback,” *IEEE Trans. Information Theory*, vol. 55, no. 5, pp. 2139–2160, May 2009.

[56] X. Huang, A. Kavcic, X. Ma, and D. Mandic, “Upper bounds on the capacities of non-controllable finite-state channles using dynamic programming methods,” in *Proc. International Symposium on Information Theory*, Seoul, South Korea, June 2009.

[57] A. Bennatan and D. Burshtein, “On the application of LDPC codes to arbitrary discrete-memoryless channels,” *IEEE Trans. Information Theory*, vol. 50, no. 3, pp. 417–438, Mar. 2004.

[58] N. Shulman and M. Feder, “Random coding techniques for nonrandom codes,” *IEEE Trans. Information Theory*, vol. 45, no. 6, pp. 2101–2104, Sept. 1999.

[59] J. H. Bae and A. Anastasopoulos, “Capacity-achieving codes for channels with memory with maximum-likelihood decoding,” in *Proc. International Symposium on Information Theory*, Seoul, Korea, June 2009.

[60] ——, “Capacity-achieving codes for finite-state channels with maximum-likelihood decoding,” *IEEE J. Select. Areas Commun.*, vol. 27, no. 6, pp. 974–984, Aug. 2009.

[61] ——, “A posterior matching scheme for finite-state channels with feedback,” in *Proc. International Symposium on Information Theory*, Austin, TX, June 2010.

[62] ——, "Capacity-achieving schemes for finite-state channels with feedback," *IEEE Trans. Information Theory*, 2011, in preparation.

[63] T. S. Han, *Information-spectrum methods in information theory.* Springer-Verlag, 2003.

[64] R. G. Gallager, "Low density parity check codes," *IEEE Trans. Information Theory*, vol. 8, pp. 21–28, Jan. 1962.

[65] A. Arapostathis, V. Borkar, E. Fernandez-Gaucherand, M. Ghost, and S. Marcels, "Discrete-time controlled Markov processes with average cost criterion: A survey," *Siam Journal on Control and Optimization*, vol. 31, no. 2, pp. 282–344, Mar. 1993.

[66] A. Mahajan and S. Tatikonda, "A graphical modeling approach to simplifying sequential teams," in *Proc. International Symposium on Modeling and Optimization in Mobile , Ad Hoc, and Wireless Networks (WiOpt), Control over Communication Channels (ConCom) Workshop*, Seoul, South Korea, June 2009.

[67] O. Hernandez-Lerma and J. B. Lasserre, *Markov chains and invariant probabilities.* Birkhauser, 2003.

[68] S. K. Gorantla and T. P. Coleman, "A stochastic control approach to coding with feedback over degraded broadcast channels," in *Proc. IEEE Conf. on Decision and Control*, Sept. 2010.

ABSTRACT

Capacity-Achieving Schemes for Finite-State Channels

by

Jung Hyun Bae

Chair: Achilleas Anastasopoulos

The main goal for the communication engineer is to design the encoder and decoder so that the system can transmit data reliably at the highest possible transmission rate. To achieve this goal, channel coding, which is the focus of this thesis, strategically adds redundancy to the transmitted data, and coding theory has provided specific transmission schemes that approach the capacity for point-to-point links.

Historically, most of the aforementioned schemes are designed for the case of memoryless channels. For channels with memory, however, few results exist on capacity-achieving codes. This is the first direction explored in this thesis, i.e. finding capacity-achieving scheme for channels with memory. In particular, capacity-achieving codes are constructed for channels with memory when the receiver employs maximum likelihood decoding. The codes are derived from the corresponding capacity-achieving codes for memoryless channels by using block-wise Markov quantization. The constructed quantized codes induce Markov distribution on the channel

input sequence and are shown to achieve the corresponding information rate.

It has been well known that feedback can improve the error performance and/or simplify the transmission scheme, and may increase the capacity. There have been several remarkable results on designing transmission schemes with feedback, but again, most of these results are for the case of memoryless channels. The second direction in this thesis, therefore, is to design simple transmission schemes for channels with memory in the presence of feedback. As the starting point of the investigation, a single-letter capacity expression is derived for the channel in consideration. Based on this capacity expression and corresponding capacity-achieving distribution, a feedback transmission scheme which achieves the capacity of channels with memory and feedback is proposed. For the case of channels with no inter-symbol interference (ISI) it is shown that the proposed transmission scheme is a straightforward generalization of the one proposed for memoryless channels. For the case where ISI is present a substantially different scheme is proposed and shown to achieve the channel capacity.