Optimal Feedback Communication
Via Posterior Matching

Ofer Shayevitz, *Member, IEEE*, and Meir Feder, *Fellow, IEEE*

Abstract—In this paper, we introduce a fundamental principle for optimal communication over general memoryless channels in the presence of noiseless feedback, termed *posterior matching*. Using this principle, we devise a (simple, sequential) generic feedback transmission scheme suitable for a large class of memoryless channels and input distributions, achieving any rate below the corresponding mutual information. This provides a unified framework for optimal feedback communication in which the Horstein scheme (BSC) and the Schalkwijk-Kailath scheme (AWGN channel) are special cases. Thus, as a corollary, we prove that the Horstein scheme indeed attains the BSC capacity, settling a longstanding conjecture. We further provide closed form expressions for the error probability of the scheme over a range of rates, and derive the achievable rates in a mismatch setting where the scheme is designed according to the wrong channel model. Several illustrative examples of the posterior matching scheme for specific channels are given, and the corresponding error probability expressions are evaluated. The proof techniques employed utilize novel relations between information rates and contraction properties of iterated function systems.

Index Terms—Channel capacity, feedback communication, Horstein scheme, iterated function system, Schalkwijk-Kailath scheme.

I. INTRODUCTION

FEEDBACK cannot increase the capacity of memoryless channels [1], [2], but can significantly improve error probability performance, and perhaps more importantly—can drastically simplify capacity achieving transmission schemes. Whereas complex coding techniques strive to approach capacity in the absence of feedback, that same goal can sometimes be attained using noiseless feedback via simple deterministic schemes that work "*on the fly*". Probably the first elegant feedback scheme in that spirit is due to Horstein [3] for the Binary Symmetric Channel (BSC). In that work, information is represented by a uniformly distributed *message point* over the unit interval, its binary expansion representing an infinite random binary sequence. The message point is then conveyed to the receiver in an increasing resolution by always indicating whether it lies to the left or to the right of its posterior distribution's median, which is also available to the transmitter via feedback. Loosely speaking, using this strategy the transmitter always answers the most informative binary question that can be posed by the receiver based on the information the latter has. Bits from the binary representation of the message point are decoded by the receiver whenever their respective intervals accumulate a sufficient posterior probability mass. The Horstein scheme was conjectured to achieve the capacity of the BSC, but this claim was verified only for a discrete set of crossover probability values for which the medians exhibit regular behavior [4], [5], and otherwise not rigorously established hitherto[1].

A few years later, two landmark papers by Schalkwijk-Kailath [7] and Schalkwijk [8] presented an elegant capacity achieving feedback scheme for the additive white Gaussian noise (AWGN) channel with an average power constraint. The Schalkwijk-Kailath scheme is "parameter estimation" in spirit, and its simplest realization is described as follows: Fixing a rate R and a block length n, the unit interval is partitioned into 2^{nR} equal length subintervals, and a (deterministic) message point is selected as one of the subintervals' midpoints. The transmitter first sends the message point itself, which is corrupted by the additive Gaussian noise in the channel and so received with some bias. The goal of the transmitter is now to refine the receiver's knowledge of that bias, thereby zooming-in on the message point. This is achieved by computing the minimum mean square error (MMSE) estimate of the bias given the output sequence observed thus far, and sending the error term amplified to match the permissible input power constraint, on each channel use. At the end of transmission the receiver uses a nearest neighbor decoding rule to recover the message point. This linear scheme is strikingly simple and yet achieves capacity; in fact at any rate below capacity it has an error probability decaying double-exponentially with the block length, as opposed to the single exponential attained by non-feedback schemes. A clean analysis of the Schalkwijk-Kailath scheme can be found in [9] and a discussion of a sequential delay-universal variant is given in [10].

Since the emergence of the Horstein and the Schalkwijk-Kailath schemes, it was evident that these are similar in some fundamental sense. Both schemes use the message point representation, and both attempt to "steer" the receiver in the right direction by transmitting what is still missing in order to "get it right". However, neither the precise correspondence nor a generalization to other cases has ever been established. In this paper, we show that in fact there exists an underlying principle, which we term *posterior matching*, that connects these two schemes. Applying this principle, we present a simple

---

[1]The rate and error exponent analysis in the original papers [3], [6], while intuitively appealing, are widely considered to be nonrigorous.

recursive feedback transmission scheme that can be tailored to any memoryless channel and any desired input distribution (e.g., capacity achieving under some input constraints), and is optimal in the sense of achieving the corresponding mutual information, under general conditions. Loosely speaking, the new scheme operates as follows: At each time instance, the transmitter computes the posterior distribution of the message point given the receiver's observations. According to the posterior, it "shapes" the message point into a random variable that is independent of the receiver's observations and has the desired input distribution, and transmits it over the channel. Intuitively, this random variable captures the information still missing at the receiver, described in a way that best matches the channel input. In the special cases of a BSC with uniform input distribution and an AWGN channel with a Gaussian input distribution, the posterior matching scheme is reduced to those of Horstein and Schalkwijk-Kailath respectively, thereby also proving the Horstein conjecture as a corollary.

The paper is organized as follows. In Section II, notations and necessary mathematical background are provided. In Section III, the posterior matching principle is introduced and the corresponding transmission scheme is derived. Technical regularity conditions for channels and input distributions are discussed in Section IV. The main result of this paper, the achievability of the mutual information via posterior matching, is presented in Section V. Error probability analysis is addressed in Section VI, where closed-form expressions are provided for a range of rates (sometimes strictly) below the mutual information. Some extensions including variants of the baseline scheme, and the penalty in rate incurred by a channel model mismatch, are addressed in Section VII. A discussion and some future research items appear in Section VIII. Several illustrative examples are discussed and revisited throughout the paper, clarifying the ideas developed.

II. Preliminaries

In this section, we provide some necessary mathematical background. Notations and definitions are given in Section II-A. Information theoretic notions pertaining to the setting of communication with feedback are described in Section II-B. An introduction to the main mathematical tools used in the paper, continuous state-space Markov chains and iterated function systems, is given in Sections II-C and II-D.

A. Notations and Definitions

Random variables (r.v.'s) are denoted by upper-case letters, their realizations by corresponding lower-case letters. A real-valued r.v. X is associated with a probability distribution $P_X(\cdot)$ defined on the usual Borel σ-algebra over $\mathbb{R}$, and we write $X \sim P_X$. The *cumulative distribution function* (c.d.f.) of X is given by $F_X(x) = P_X((-\infty, x])$, and the inverse c.d.f. is defined to be $F_X^{-1}(t) \stackrel{\text{def}}{=} \inf\{x : F_X(x) > t\}$. Unless otherwise stated, we assume that any real-valued r.v. X is either continuous, discrete, or a mixture of the two[2]. Accordingly, X ad-

[2]This restricts F_X to be the sum of an absolutely continuous function (continuous part) and a jump function (discrete part). This is to say we avoid the case of a *singular part*, where P_X assigns positive probability to some uncountable set of zero Lebesgue measure.

mits a (wide sense) *probability density function* (p.d.f.) $f_X(x)$, which can be written as a mixture of a Lebesgue integrable function (continuous part) and Dirac delta functions (discrete part). If there is only a continuous part then X and its distribution/c.d.f./p.d.f. are called *proper*. The *support* of X is the intersection of all closed sets A for which $P_X(\mathbb{R} \setminus A) = 0$, and is denoted $\operatorname{supp}(X)$.[3] For brevity, we write $P_X(x)$ for $P_X(\{x\})$, and $x \in \operatorname{supp}(X)$ is called a *mass point* if $P_X(x) > 0$. The discrete part of the support is the set of all mass points, and the continuous part the complement set. The *interior* of the support is denoted by $\underline{\operatorname{supp}}(X)$ for short. A vector of real-valued r.v.'s $X^n = (X_1, X_2, \ldots, X_n)$ is similarly associated with $P_{X^n}, F_{X^n}, f_{X^n}$ and with $\operatorname{supp}(X^n)$, where the p.d.f. is now called proper if all the scalar conditional distributions are a.s. (almost surely) proper. We write $\mathbb{E}(\cdot)$ for expectation and $\mathbb{P}(\cdot)$ for the probability of a measurable event within the parentheses. The uniform probability distribution over $(0, 1)$ is denoted throughout by $\mathcal{U}$. A measurable bijective function $\mu : (0, 1) \mapsto (0, 1)$ is called a *uniformity preserving function* (u.p.f.) if $\Theta \sim \mathcal{U}$ implies that $\mu(\Theta) \sim \mathcal{U}$.

A scalar distribution P_X is said to be (strictly) *dominated* by another distribution P_Y if $F_X(x) < F_Y(x)$ whenever $F_Y(x) \in (0, 1)$, and the relation is denoted by $P_X \prec_d P_Y$. A distribution P_X is called *absolutely continuous* w.r.t. another distribution P_Y, if $P_Y(A) = 0$ implies $P_X(A) = 0$ for every $A \in \mathfrak{B}$, where $\mathfrak{B}$ is the corresponding σ-algebra. This relation is denoted $P_X \ll P_Y$. If both distributions are absolutely continuous w.r.t. each other, then they are said to be *equivalent*. The *total variation distance* between P_X and P_Y is defined as

$$d_{TV}(P_X, P_Y) = \sup_{A \in \mathfrak{B}} |P_X(A) - P_Y(A)|.$$

A statement is said to be satisfied for P_X-a.a. *(almost all)* x, if the set of x's for which it is satisfied has probability one under P_X.

In what follows, we use $\operatorname{conv}(\cdot)$ for the *convex hull* operator, $|\Delta|$ for the length of an interval $\Delta \subseteq \mathbb{R}$, $\log$ for $\log_2$, $\operatorname{range}(f)$ for the range of a function f, and $\circ$ for function composition. The indicator function over a set A is denoted by $\mathbb{1}_A(\cdot)$. A set $A \subseteq \mathbb{R}^m$ is said to be *convex in the direction* $u \in \mathbb{R}^m$, if the intersection of A with any line parallel to u is a connected set (possibly empty). Note that A is convex if and only if it is convex in any direction.

The following simple lemma states that (up to discreteness issues) any real-valued r.v. can be shaped into a uniform r.v. or vice versa, by applying the corresponding c.d.f or its inverse, respectively. This fact is found very useful in the sequel.

Lemma 1: Let $X \sim P_X, \Theta \sim \mathcal{U}$ be statistically independent. Then
  (i) $F_X^{-1}(\Theta) \sim P_X$.
  (ii) $F_X(X) - \Theta \cdot P_X(X) \sim \mathcal{U}$. Specifically, if X is proper then $F_X(X) \sim \mathcal{U}$.
  Proof: See Appendix A. ∎

[3]This coincides with the usual definitions of support for continuous and discrete r.v.'s.

A proper real-valued r.v. X is said to have *a regular tail* if there exists some $\gamma \in (0, \frac{1}{2}]$ and positive constants $c_0, c_1, \alpha_0, \alpha_1$, such that

$$c_0 f_X^{\alpha_0}(x) \leq \min(F_X(x), 1 - F_X(x)) \leq c_1 f_X^{\alpha_1}(x)$$

for any $x \in \mathrm{supp}(X)$ satisfying $\min(F_X(x), 1 - F_X(x)) \leq \gamma$.

Lemma 2: Let X be proper with $\mathrm{supp}(X) = \mathbb{R}$ and a bounded unimodal p.d.f. f_X. Each of the following conditions implies that X has a regular tail:

(i) $f_X(x) = O(|x|^{-a})$ and $f_X(x) = \Omega(|x|^{-b})$ as $|x| \to \infty$, for some $b \geq a > 1$.

(ii) $f_X(x) = O(e^{-b|x|^a})$ and $f_X(x) = \Omega(e^{-b|x|^a})$ as $|x| \to \infty$, for some $a \geq 1, b > 0$.

Proof: See Appendix C.   ∎

Example 1: If X is either Gaussian, Laplace, or Cauchy distributed then X has a regular tail.

B. Information Theoretic Notions

The relative entropy between two distributions P_X and P_Y is denoted by $D(P_X \| P_Y)$. The mutual information between two r.v.'s X and Y is denoted $I(X; Y)$, and the differential entropy of a continuous r.v. X is denoted $h(X)$. A *memoryless channel* is defined via (and usually identified with) a conditional probability distribution $P_{Y|X}$ on $\mathbb{R}$. The *input alphabet* $\mathcal{X}$ of the channel is the set of all $x \in \mathbb{R}$ for which the distribution $P_{Y|X}(\cdot|x)$ is defined, the output alphabet of the channel is the set $\mathcal{Y} \stackrel{\text{def}}{=} \bigcup_{x \in \mathcal{X}} \mathrm{supp}(Y|X = x) \subseteq \mathbb{R}$. A sequence of real-valued r.v. pairs $\{(X_n, Y_n)\}_{n=1}^{\infty}$ taking values in $\mathcal{X} \times \mathcal{Y}$ is said to be an *input/output sequence* for the memoryless channel $P_{Y|X}$ if

$$P_{Y_n|X^n Y^{n-1}}(\cdot|x^n, y^{n-1}) = P_{Y|X}(\cdot|x_n), \quad n \in \mathbb{N}. \quad (1)$$

A probability distribution P_X is said to be a (memoryless) *input distribution* for the channel $P_{Y|X}$ if $\mathrm{supp}(X) \subseteq \mathcal{X}$. The pair $(P_X, P_{Y|X})$ induces an *output distribution* P_Y over the output alphabet, a joint input/output distribution P_{XY}, and an *inverse channel* $P_{X|Y}$. Such a pair $(P_X, P_{Y|X})$ is called an *input/channel pair* if $I(X; Y) < \infty$.

A channel for which both the input and output alphabets $\mathcal{X}, \mathcal{Y}$ are finite sets is called a *discrete memoryless channel (DMC)*. Note that the numerical values of the inputs/outputs are practically irrelevant for a DMC, and hence in this case one can assume without loss of generality that $\mathcal{X} = \{0, 1, \ldots, |\mathcal{X}| - 1\}$ and $\mathcal{Y} = \{0, 1, \ldots, |\mathcal{Y}| - 1\}$. Moreover, two input/DMC pairs $(P_X, P_{Y|X})$ and $(P_{X^*}, P_{Y^*|X^*})$ are said to be *equivalent* if one can be obtained from the other by input and output permutations, i.e., there exist permutations $\sigma_1 : \mathcal{X} \mapsto \mathcal{X}$ and $\sigma_2 : \mathcal{Y} \mapsto \mathcal{Y}$ such that

$$P_X(i) = P_{X^*}(\sigma_1(i)), \quad P_{Y|X}(j|i) = P_{Y^*|X^*}(\sigma_2(j)|\sigma_1(i))$$

for all $i \in \mathcal{X}, j \in \mathcal{Y}$. In particular, equivalent pairs have the same mutual information.

Let Θ_0 be a random *message point* uniformly distributed over the unit interval, with its binary expansion representing an infinite independent-identically-distributed (i.i.d.) Bernoulli($\frac{1}{2}$)

sequence to be reliably conveyed by a transmitter to a receiver over the channel $P_{Y|X}$. A *transmission scheme* is a sequence of a-priori agreed upon measurable *transmission functions* $g_n : (0, 1) \times \mathcal{Y}^{n-1} \mapsto \mathcal{X}$, so that the input to the channel generated by the transmitter is given by

$$X_n = g_n(\Theta_0, Y^{n-1}), \quad n \in \mathbb{N}.$$

A transmission scheme induces a distribution $P_{X_n|X^{n-1}Y^{n-1}}$ which together with (1) uniquely defines the joint distribution of the input/output sequence. In the special case where g_n does not depend on y^{n-1}, the transmission scheme is said to work *without feedback* and is otherwise said to work *with feedback*.

A *decoding rule* is a sequence of measurable mappings $\{\Delta_n : \mathcal{Y}^n \mapsto \mathcal{E}\}_{n=1}^{\infty}$, where $\mathcal{E}$ is the set of all open intervals in $(0, 1)$. We refer to $\Delta_n(y^n)$ as the *decoded interval*. The *error probability* at time n associated with a transmission scheme and a decoding rule, is defined as

$$p_e(n) \stackrel{\text{def}}{=} \mathbb{P}(\Theta_0 \notin \Delta_n(Y^n))$$

and the corresponding *rate* at time n is defined to be

$$R_n \stackrel{\text{def}}{=} -\frac{1}{n} \log |\Delta_n(Y^n)|.$$

We say that a transmission scheme together with a decoding rule *achieve* a rate R over a channel $P_{Y|X}$ if

$$\lim_{n \to \infty} \mathbb{P}(R_n < R) = 0, \quad \lim_{n \to \infty} p_e(n) = 0. \quad (2)$$

The rate is achieved *within an input constraint* (η, u), if in addition

$$\lim_{n \to \infty} n^{-1} \sum_{k=1}^{n} \eta(X_k) \leq u \quad \text{a.s.(element-wise)} \quad (3)$$

where $\eta : \mathcal{X} \mapsto \mathbb{R}^m$ is a measurable function and $u \in \mathbb{R}^m$. A scheme and a decoding rule are also said to *pointwise achieve* a rate R if for all $\theta_0 \in (0, 1)$

$$\lim_{n \to \infty} \mathbb{P}(R_n < R | \Theta_0 = \theta_0) = 0$$
$$\lim_{n \to \infty} \mathbb{P}(\Theta_0 \notin \Delta_n(Y^n) | \Theta_0 = \theta_0) = 0$$

and to do the above within an input constraint (η, u) if (3) is also satisfied (given $\Theta_0 = \theta_0$). Clearly, pointwise achievability implies achievability but not vice versa. Accordingly, a rate R is called (pointwise) achievable over a channel $P_{Y|X}$ within an input constraint (η, u) if there exist a transmission scheme and a decoding rule (pointwise) achieving it. The *capacity* (with feedback) $C(P_{Y|X}, \eta, u)$ of the channel under the input constraint is the supremum of all the corresponding achievable rates[4]. It is well known that the capacity is given by [11]

$$C(P_{Y|X}, \eta, u) = \sup_{\substack{P_X : \mathbb{E}\eta(X) \leq u \\ \mathrm{supp}(X) \subseteq \mathcal{X}}} I(X; Y). \quad (4)$$

---

[4]A pointwise capacity can be defined as well, and may be smaller than (4) depending on the channel. However, we do not pursue this direction.

Furthermore, the capacity without feedback (i.e., considering only schemes that work without feedback) is given by the above as well. The *unconstrained capacity* (i.e., when no input constraint is imposed) is denoted $C(P_{Y|X})$ for short.

An *optimal fixed rate* decoding rule with rate R is one that decodes an interval of length 2^{-nR} whose a-posteriori probability is maximal, i.e.,

$$\Delta_n(y^n) = \underset{\{J \in \mathcal{E}: |J| = 2^{-nR}\}}{\operatorname{argmax}} P_{\Theta_0|Y^n}(J|y^n)$$

where ties are broken arbitrarily. This decoding rule minimizes the error probability $p_e(n)$ for a fixed $R_n = R$. An *optimal variable rate* decoding rule with a *target error probability* $p_e(n) = \delta_n$ is one that decodes a minimal-length interval whose accumulated a-posteriori probability exceeds $1 - \delta_n$, i.e.,

$$\Delta_n(y^n) = \underset{\{J \in \mathcal{E}: P_{\Theta_0|Y^n}(J|y^n) \geq 1 - \delta_n\}}{\operatorname{argmin}} |J|$$

where ties are broken arbitrarily, thereby maximizing the instantaneous rate for a given error probability. Both decoding rules make use of the posterior distribution of the message point $P_{\Theta_0|Y^n}(\cdot|y^n)$, which can be calculated online at both terminals.

It should be noted that the main reason we adopt the above nonstandard definitions for channel coding with feedback, is that they result in a much cleaner analysis. It may not be immediately clear how this corresponds to the standard coding framework [12], and in particular, how achievability as defined above translates into the actual reliable decoding of messages at a desired rate. The following Lemma justifies this alternative formalization.

Lemma 3: Achievability as defined in (2) and (3) above, implies achievability in the standard framework.

Proof: See Appendix A. Loosely speaking, a rate R is achievable in our framework if the posterior distribution $P_{\Theta_0|Y^n}$ concentrates in an interval of size $\approx 2^{-nR}$ around Θ_0, as n grows large. This intuitively suggests that nR bits from the message point representation could be reliably decoded, or, more accurately, that the unit interval can be partitioned into $\approx 2^{nR}$ intervals such that the one containing Θ_0 can be identified with high probability. ∎

C. Markov Chains

A *Markov chain* $\{\Psi_n\}_{n=1}^{\infty}$ over a measurable *state space* $\mathfrak{F}$, is a stochastic process defined via an initial distribution P_{Ψ_1} on $\mathfrak{F}$, and a *stochastic kernel* (conditional probability distribution) $\mathcal{P}$, such that

$$P_{\Psi_n|\Psi^{n-1}}(\cdot|\psi^{n-1}) = P_{\Psi_n|\Psi_{n-1}}(\cdot|\psi_{n-1}) \overset{\text{def}}{=} \mathcal{P}(\cdot|\psi_{n-1}).$$

We say $s \in \mathfrak{F}$ is the *initial point* of the chain if $P_{\Psi_1}(s) = 1$, and denote the probability distribution induced over the chain for an initial point s by $\mathcal{P}_s$. The Markov chain generated by sampling the original chain in steps of m is called the *m-skeleton*, and its kernel is denoted by $\mathcal{P}^m$. The chain is said to be P_Ψ-*irreducible* for a distribution P_Ψ over $\mathfrak{F}$, if any set $A \in \mathfrak{B}$ with $P_\Psi(A) > 0$ is reached in a finite number of steps with a positive

probability for any initial point, where $\mathfrak{B}$ is the corresponding σ-algebra over $\mathfrak{F}$. P_Ψ is said to be *maximal* for the chain if any other irreducibility distribution is absolutely continuous w.r.t. P_Ψ. A maximal P_Ψ-irreducible chain is said to be *recurrent* if for any initial point, the expected number of visits to any set $A \in \mathfrak{B}$ with $P_\Psi(A) > 0$, is infinite. The chain is said to be *Harris recurrent*, if any such set is visited infinitely often for any initial point. Thus, Harris recurrence implies recurrence but not vice versa. A set $A \in \mathfrak{B}$ is called *invariant* if $\mathcal{P}(A|s) = 1$ for any $s \in A$. An *invariant distribution* P_Ψ is one for which $P_{\Psi_{n-1}} = P_\Psi$ implies $P_{\Psi_n} = P_\Psi$. Such an invariant distribution is called *ergodic* if for every invariant set A either $P_\Psi(A) = 0$ or $P_\Psi(A) = 1$. A chain which has (at least one) invariant distribution is called *positive*. For short, we use the acronym *p.h.r.* to indicate positive Harris recurrence. A chain is said to have a *d-cycle* if its state space can be partitioned into d disjoint sets amongst which the chain moves cyclicly a.s. The largest d-cycle possible is called a *period*, and a chain is called *aperiodic* if its period equals one.

The following results are taken from [13] and [14]. We will assume here that $\mathfrak{F}$ is an open/closed subset of $\mathbb{R}^m$ associated with the usual Borel σ-algebra $\mathfrak{B}$, although the claims hold under more general conditions.

Lemma 4: An irreducible chain that has an invariant distribution is (positive) recurrent, and the invariant distribution is unique (and hence ergodic).

Lemma 5 (p.h.r. Conditions): Consider a chain with a kernel $\mathcal{P}$. Each of the following conditions implies p.h.r.:
  (i) The chain has a unique invariant distribution P_Ψ, and $\mathcal{P}(\cdot|s) \ll P_\Psi$ for any $s \in \mathfrak{F}$.
  (ii) Some m-skeleton $\mathcal{P}^m$ is p.h.r.

Lemma 6 (p.h.r. Convergence): Consider an aperiodic p.h.r. chain with a kernel $\mathcal{P}$ and an invariant distribution P_Ψ. Then for any $s \in \mathfrak{F}$

$$\lim_{n \to \infty} d_{TV}\left(\mathcal{P}^n(\cdot|s), P_\Psi\right) = 0.$$

Lemma 7 (Strong Law of Large Numbers (SLLN)): If P_Ψ is an ergodic invariant distribution for the Markov chain $\{\Psi_n\}_{n=1}^{\infty}$ with kernel $\mathcal{P}$, then for any measurable function $\eta : \mathfrak{F} \mapsto \mathbb{R}$ satisfying $\mathbb{E}|\eta(\Psi)| < \infty$ and P_Ψ-a.a. initial point $s \in \mathfrak{F}$

$$\lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} \eta(\Psi_k) = \mathbb{E}\eta(\Psi) \quad \mathcal{P}_s\text{-a.s.}$$

Furthermore, if the chain is p.h.r. then the above holds for any $s \in \mathfrak{F}$.

D. Iterated Function Systems

Let $\mathfrak{F}$ be a measurable space, $\omega : \mathbb{R} \times \mathfrak{F} \mapsto \mathfrak{F}$ a measurable function[5], and write $\omega_y(\cdot) \overset{\text{def}}{=} \omega(y, \cdot)$ for any $y \in \mathbb{R}$. Let $\{Y_n\}_{n=1}^{\infty}$ be an i.i.d. sequence of real-valued r.v.'s. An *Iterated*

---

[5] $\mathbb{R}$ is equipped with the usual Borel σ-algebra, and $\mathbb{R} \times \mathfrak{F}$ is equipped with the corresponding product σ-algebra.

Function system (IFS) $\{S_n(s)\}_{n=1}^{\infty}$ is a stochastic process over $\mathfrak{F}$, defined by[6]

$$S_1 = s \in \mathfrak{F}, \quad S_{n+1}(s) = \omega_{Y_n} \circ \omega_{Y_{n-1}} \circ \cdots \circ \omega_{Y_1}(s). \quad (5)$$

A *Reversed IFS (RIFS)* $\{\tilde{S}_n(s)\}_{n=1}^{\infty}$ is a stochastic process over $\mathfrak{F}$, obtained by a reversed order composition

$$\tilde{S}_1 = s \in \mathfrak{F}, \quad \tilde{S}_{n+1}(s) = \omega_{Y_1} \circ \omega_{Y_2} \circ \cdots \circ \omega_{Y_n}(s). \quad (6)$$

We say that the (R)IFS is *generated* by the (R)IFS *kernel* $\omega_y(\cdot)$, *controlled* by the sequence $\{Y_n\}_{n=1}^{\infty}$, and s is its initial point. Note that an IFS is a Markov chain over the state space $\mathfrak{F}$, and in fact a large class of Markov chains can be represented by a suitable IFS [15]. In contrast, a RIFS is not a Markov chain but it is however useful in the analysis of the corresponding IFS,[7] see, e.g., [16]–[18]. However, in what follows, the RIFS will turn out to have an independent significance.

A function $\xi : [0, 1] \mapsto [0, 1]$ is called a (generally nonlinear) *contraction* if it is nonnegative, $\cap$-convex, and $\xi(x) < x$ for any $x \in (0, 1]$.

Lemma 8: For any contraction $\xi(\cdot)$

$$r(n) \overset{\text{def}}{=} \sup_{x \in [0,1]} \xi^{(n)}(x), \quad \lim_{n \to \infty} r(n) = 0$$

where $\xi^{(n)}$ is the n-fold iteration of ξ. The sequence $r(n)$ is called the *decay profile* of ξ.

   Proof: See Appendix A.   ∎

Example 2: The function $\xi(x) = rx$ is a (linear) contraction for $0 < r < 1$, with an exponential decay profile $r(n) = r^n$.

Example 3: The function $\xi(x) = x - \alpha x^{\beta}$ is a contraction for $\alpha < \frac{1}{\beta}$ and $\beta > 1$, with a polynomial decay profile $r(n) = O(n^{\frac{1}{1-\beta}})$.

In what follows, a measurable and surjective function $\psi : \mathfrak{F} \mapsto [0, 1]$ is called a *length function*. We now state some useful convergence Lemmas for (R)IFS.

Lemma 9: Consider the IFS defined in (5), and suppose there exist a length function $\psi(\cdot)$ and a contraction $\xi(\cdot)$ with a decay profile $r(n)$, such that

$$\mathbb{E}[\psi(\omega_{Y_1}(s))] \leq \xi(\psi(s)), \quad \forall s \in \mathfrak{F} \quad (7)$$

Then for any $s \in \mathfrak{F}$ and any $\varepsilon > 0$

$$\mathbb{P}(\psi(S_n(s)) > \varepsilon) \leq \varepsilon^{-1} r(n)$$

   Proof: See Appendix A.   ∎

In the sequel, we consider an IFS over the space $\mathfrak{F}_c$ of all c.d.f. functions over the open unit interval[8], i.e., all monotone nondecreasing functions $h : (0, 1) \mapsto (0, 1)$ for which

$\text{conv}(\text{range}(h)) = (0, 1)$. Furthermore, we define the following family of length functions over $\mathfrak{F}_c$:

$$\psi_{\lambda}(h) \overset{\text{def}}{=} \int_0^1 \lambda(h(x))dx, \quad h \in \mathfrak{F}_c \quad (8)$$

where $\lambda : [0, 1] \mapsto [0, 1]$ is surjective, $\cap$-convex and symmetric about $\frac{1}{2}$.

   For any $h : \mathbb{R} \mapsto \mathbb{R}$ and $s, t \in \mathbb{R}$, define

$$D_{s,t}(h) \overset{\text{def}}{=} \frac{|h(s) - h(t)|}{|s - t|}, \quad D_s(h) \overset{\text{def}}{=} \limsup_{t \to s} D_{s,t}(h) \quad (9)$$

$D_{s,t}(\cdot)$ and $D_s(\cdot)$ are called *global* and *local Lipschitz operators*, respectively.

Lemma 10: Consider the RIFS in (6) over some interval $\mathfrak{F} \subseteq \mathbb{R}$, and suppose the following condition holds for some $q > 0$:

$$r \overset{\text{def}}{=} \sup_{s \neq t \in \mathfrak{F}} \mathbb{E}[D_{s,t}(\omega_{Y_1})]^q < 1. \quad (10)$$

Then for any $\varepsilon > 0$

$$\mathbb{P}(|\tilde{S}_n(s) - \tilde{S}_n(t)| > \varepsilon) \leq \varepsilon^{-q}|s - t|^q r^n \quad s, t \in \mathfrak{F}$$

   Proof: See Appendix A.   ∎

Lemma 11 (From [17]): Consider the RIFS in (6) over the interval $\mathfrak{F} = (0, 1)$. Let $\rho : (0, 1) \mapsto [1, \infty)$ be a continuous function, and define

$$J(s; t) \overset{\text{def}}{=} \sup\{\rho(\text{conv}\{s, t\})\},$$
$$K_s \overset{\text{def}}{=} \mathbb{E}[J(s; \omega_{Y_1}(s))]$$
$$\Psi(s, t, r) \overset{\text{def}}{=} \frac{K_s + K_t}{1 - r} + 2J(s; t).$$

If

$$r \overset{\text{def}}{=} \sup_{s \in \mathfrak{F}} \mathbb{E}\left[\frac{\rho(\omega_{Y_1}(s))}{\rho(s)} D_s(\omega_{Y_1})\right] < 1$$

then for any $s, t \in (0, 1)$ and any $\varepsilon > 0$

$$\mathbb{P}\left(\left|\tilde{S}_n(s) - \tilde{S}_n(t)\right| > \varepsilon\right) \leq \varepsilon^{-1}\Psi(s, t, r) \cdot r^n.$$

III. POSTERIOR MATCHING

In this section, we introduce the idea of posterior matching and develop the corresponding framework. In Section III-A, a new fundamental principle for optimal communication with feedback is presented. This principle is applied in Section III-B, to devise a general transmission scheme suitable for any given input/channel pair $(P_X, P_{Y|X})$,[9]. This scheme will later be shown (in Section V) to achieve any rate below the corresponding mutual information $I(X; Y)$, under general conditions. A recursive representation of the scheme in a continuous alphabet setting is developed, where the recursion rule is given as a simple function of the input/channel pair $(P_X, P_{Y|X})$. A

---

[6]We call the process itself an IFS. In the literature sometimes ω_y is the IFS and the process is defined separately

[7]The idea is that it is relatively simple to prove (under suitable contraction conditions) that the RIFS converges to a unique random fixed point a.s., and since the IFS and the RIFS have the same marginal distribution, the distribution of that fixed point must be the unique stationary distribution of the IFS.

[8]$\mathfrak{F}_c$ is associated with the topology of pointwise convergence, and the corresponding Borel σ-algebra.

[9]For instance, P_X may be selected to be capacity achieving for $P_{Y|X}$, possibly under some desirable input constraints.

common framework for discrete, continuous and mixed alphabets is introduced in Section III-C, and a corresponding unified recursive representation is provided. Several illustrative examples are discussed throughout the section, where in each the corresponding scheme is explicitly derived. In the special cases of the AWGN channel with a Gaussian input, and the BSC with a uniform input, it is demonstrated how the scheme reduces to the Schalkwijk-Kailath and Horstein schemes, respectively.

A. The Basic Principle

Suppose the receiver has observed the output sequence Y^n, induced by a message point Θ_0 and an arbitrary transmission scheme used so far. The receiver has possibly gained some information regarding the value of Θ_0 via Y^n, but what is the information it is still missing? We argue that a natural candidate is any r.v. U with the following properties:

  (I) U *is statistically independent of Y^n.*
  (II) *The message point Θ_0 can be a.s. uniquely recovered from (U, Y^n).*

Intuitively, the first requirement guarantees that U represents "new information" not yet observed by the receiver, while the second requirement makes sure this information is "relevant" in terms of describing the message point. Following this line of thought, we suggest a simple principle for generating the next channel input:

The transmission function g_{n+1} should be selected so that X_{n+1} is P_X-distributed, and is a fixed function[10] of some r.v. U satisfying properties (I) and (II).

That way, the transmitter attempts to convey the missing information to the receiver, while at the same time satisfying the input constraints encapsulated in P_X[11]. We call this the *posterior matching principle* for reasons that will become clear immediately. Note that any transmission scheme adhering to the posterior matching principle, satisfies

$$
\begin{aligned}
I(\Theta_0; Y_{n+1}|Y^n) &= I(\Theta_0, Y^n; Y_{n+1}) - I(Y_{n+1}; Y^n) \\
&= I(X_{n+1}; Y_{n+1}) - I(Y_{n+1}; Y^n) \\
&= I(X; Y).
\end{aligned} \tag{11}
$$

The second equality follows from the memorylessness of the channel and the fact that X_{n+1} is a function of (Θ_0, Y^n). The last equality holds since $X_{n+1} \sim P_X$, and since Y_{n+1} is independent of Y^n, where the latter is implied by property (I) together with the memorylessness of the channel. Loosely speaking, a transmission scheme satisfying the posterior matching principle therefore conveys, on each channel use, "new information" pertaining to the message point that is equal to the associated one-shot mutual information. This is intuitively appealing, and gives some idea as to why such a scheme may be good. However, this property does not prove nor directly implies anything regarding achievability. It merely indicates that we have done "information lossless" processing when converting the one-shot channel into an n-shot channel,

an obvious necessary condition. In fact, note we did not use property (II), which turns out to be important[12].

The rest of this paper is dedicated to the translation of the posterior matching principle into a viable transmission scheme, and to its analysis. As we shall see shortly, there are infinitely many transmission functions that satisfy the posterior matching principle. There is, however, one baseline scheme which is simple to express and analyze.

B. The Posterior Matching Scheme

Theorem 1 (Posterior Matching Scheme): The following transmission scheme satisfies the posterior matching principle for any n:

$$
g_{n+1}(\theta, y^n) = F_X^{-1} \circ F_{\Theta_0|Y^n}(\theta|y^n). \tag{12}
$$

Based on the above transmission functions, the input to the channel is a sequence of r.v.'s given by

$$
X_{n+1} = F_X^{-1} \circ F_{\Theta_0|Y^n}(\Theta_0|Y^n). \tag{13}
$$

Proof: Assume $P_{\Theta_0|Y^n}(\cdot|y^n)$ is proper for any $y^n \in \mathcal{Y}^n$. Then Lemma 1 claim (ii) implies that $F_{\Theta_0|Y^n}(\Theta_0|y^n) \sim \mathcal{U}$, and since this holds for all y^n then $F_{\Theta_0|Y^n}(\Theta_0|Y^n) \sim \mathcal{U}$ and is statistically independent of Y^n. It is easy to see that for any y^n, the mapping $F_{\Theta_0|Y^n}(\cdot|y^n)$ is injective when its domain is restricted to $\mathrm{supp}(P_{\Theta_0|Y^n}(\cdot|y^n))$, thus Θ_0 can be a.s. uniquely recovered from $(F_{\Theta_0|Y^n}(\Theta_0|Y^n), Y^n)$. Hence, we conclude that $F_{\Theta_0|Y^n}(\Theta_0|y^n)$ satisfies properties (I) and (II) required by the posterior matching principle. By Lemma 1 claim (i), applying the inverse c.d.f. F_X^{-1} merely shapes the uniform distribution into the distribution P_X. Therefore, X_{n+1} is P_X-distributed and since it is also a deterministic function of $F_{\Theta_0|Y^n}(\Theta_0|Y^n)$, the posterior matching principle is satisfied. See Appendix A to eliminate the properness assumption. ∎

Following the above, it is now easy to derive a plethora of schemes satisfying the posterior matching principle.

Corollary 1: Let $\{\mu_n\}_{n=1}^\infty$ be a sequences of u.p.f's, and let $\{\varsigma_n : (0,1) \mapsto (0,1)\}_{n=1}^\infty$ be a sequence of measurable bijective functions. The transmission scheme given by

$$
g_{n+1}(\theta, y^n) = F_X^{-1} \circ \mu_n \circ P_{\Theta_0|Y^n}\left(\varsigma_n^{-1}\left((0,\varsigma_n(\theta)]\right)|y^n\right)
$$

satisfies the posterior matching principle for any n. In particular, a scheme obtained by fixing $\mu_n = \mu$ and ς_n to be the identity function[13] for all n, is called a μ-*variant*. The transmission scheme corresponding to a μ-variant is thus given by

$$
g_{n+1}(\theta, y^n) = F_X^{-1} \circ \mu \circ F_{\Theta_0|Y^n}(\theta|y^n). \tag{14}
$$

Finally, the baseline scheme (12) is recovered by setting μ to be the identity function.

We note that the different schemes described above have a similar flavor. Loosely speaking, the message point is described

---

[10]By *fixed* we mean that the function cannot depend on the outputs y^n, so that X_{n+1} is still independent of Y^n.

[11]The extra degree of freedom in the form of a deterministic function is in fact significant only when P_X has a discrete part, in which case a quantization of U may void property (II).

[12]One can easily come up with useless schemes for which only property (I) holds. A simple example is uncoded transmission: Transmit the binary representation of Θ_0 bit by bit over a BSC, independent of the feedback.

[13]In fact, letting ς_n be any sequence of monotonically increasing functions results in the same scheme. This fact is used in the error probability analysis on Section VI, to obtain tighter bounds.

each time at a resolution determined by the current uncertainty at the receiver, by somehow stretching and redistributing the posterior probability mass so that it matches the desired input distribution (we will later see that the "stretching rate" corresponds to the mutual information). This interpretation explains the posterior matching moniker. From this point forward we mostly limit our discussion to the baseline scheme described by (12) or (13), which is henceforth called the *posterior matching scheme*. The μ-*variants* (14) of the scheme will be discussed in more detail on Section VII-A.

As it turns out, the posterior matching scheme may sometimes admit a simple recursive form.

Theorem 2 (Recursive Representation I): If P_{XY} is proper, then the posterior matching scheme (12) is also given by

$$g_1(\theta) = F_X^{-1}(\theta),$$
$$g_{n+1}(\theta|y^n) = \left(F_X^{-1} \circ F_{X|Y}(\cdot \,|\, y_n)\right) \circ g_n(\theta|y^{n-1}).$$
(15)

Moreover, the corresponding sequence of input/output pairs $\{(X_n, Y_n)\}_{n=1}^{\infty}$ constitute a Markov chain over a state space $\underline{\mathrm{supp}}(X, Y) \subseteq \mathbb{R}^2$, with an invariant distribution P_{XY}, and satisfy the recursion rule

$$X_1 = F_X^{-1}(\Theta_0), \quad X_{n+1} = F_X^{-1} \circ F_{X|Y}(X_n|Y_n). \quad (16)$$

Proof: The initialization $g_1(\theta) = F_X^{-1}(\theta)$ results immediately from (12), recalling that Θ_0 is uniform over the unit interval. To prove the recursion relation, we notice that since P_{XY} is proper then the transmission functions $g_n(\theta, y^{n-1})$ are continuous when restricted to the support of the posterior, and strictly increasing in θ for any fixed y^{n-1}. Therefore, we have the following set of equalities:

$$F_{\Theta_0|Y^n}(\theta|y^n) = \mathbb{P}(\Theta_0 \le \theta|Y^n = y^n)$$
$$\stackrel{(a)}{=} \mathbb{P}(g_n(\Theta_0, y^{n-1}) \le g_n(\theta, y^{n-1})|Y^n = y^n)$$
$$= \mathbb{P}(X_n \le g_n(\theta, y^{n-1})|Y^n = y^n)$$
$$\stackrel{(b)}{=} \mathbb{P}(X_n \le g_n(\theta, y^{n-1})|Y_n = y_n)$$
$$= F_{X|Y}(g_n(\theta, y^{n-1})|y_n) \quad (17)$$

where in (a) we used the continuity and monotonicity of the transmission functions, and in (b) we used the facts that the channel is memoryless and that by construction X_n is statistically independent of Y^{n-1}, which also imply that Y^n is an i.i.d. sequence. The recursive rule (15) now results immediately by combining (12) and (17).

Now, using (13) we obtain

$$X_{n+1} = F_X^{-1} \circ F_{\Theta_0|Y^n}(\Theta_0|Y^n)$$
$$= F_X^{-1} \circ F_{X|Y}(g_n(\Theta_0, Y^{n-1})|Y_n)$$
$$= F_X^{-1} \circ F_{X|Y}(X_n|Y_n)$$

yielding relation (16). Since Y_n is generated from X_n via a memoryless channel, the Markovity of $\{(X_n, Y_n)\}_{n=1}^{\infty}$ is established. The distribution P_{XY} is invariant since by construction $(X_n, Y_n) \sim P_{XY}$ implies $X_{n+1} \sim P_X$, and then Y_{n+1} is generated via the memoryless channel $P_{Y|X}$. Taking the state space

to be $\underline{\mathrm{supp}}(X, Y)$ is artificial here since $P_{XY}(\underline{\mathrm{supp}}(X, Y) \setminus \underline{\mathrm{supp}}(X, Y)) = 0$, and is done for reasons of mathematical convenience to avoid having trivial invariant distributions (this is not true when P_{XY} is not proper). Note that the chain *emulates* the "correct" input marginal and the "correct" joint (i.i.d.) output distribution; this interpretation is further discussed in Section VIII. ∎

In the sequel, we refer to the function $F_X^{-1} \circ F_{X|Y}$ appearing in the recursive representation as the *posterior matching kernel*. Let us now turn to consider several examples, which are frequently revisited throughout the paper.

Example 4 (AWGN Channel): Let $P_{Y|X}$ be an AWGN channel with noise variance N, and let us set a Gaussian input distribution $X \sim \mathcal{N}(0, \mathrm{P})$, which is capacity achieving for an input power constraint P. We now derive the posterior matching scheme in this case, and show it reduces to the Schalkwijk-Kailath scheme. Let $\mathrm{SNR} \stackrel{\text{def}}{=} \frac{\mathrm{P}}{\mathrm{N}}$. Standard manipulations yield the following posterior distribution

$$X|Y = y \sim \mathcal{N}\left(\frac{\mathrm{SNR}}{1 + \mathrm{SNR}} \cdot y, \frac{1}{1 + \mathrm{SNR}} \cdot P\right). \quad (18)$$

The joint p.d.f. f_{XY} is Gaussian and hence proper, so the recursive representation of Theorem 2 is valid. By definition, the corresponding posterior matching kernel satisfies

$$F_X^{-1} \circ F_{X|Y}(x \,|\, y) = \{z : F_X(z) = F_{X|Y}(x \,|\, y)\}. \quad (19)$$

However, from Gaussianity and (18) we know that

$$F_{X|Y}(x \,|\, y) = F_X\left(\sqrt{1 + \mathrm{SNR}}\left(x - \frac{\mathrm{SNR}}{1 + \mathrm{SNR}} \cdot y\right)\right). \quad (20)$$

Combining (19) and (20), the posterior matching kernel for the AWGN channel setting is given by

$$F_X^{-1} \circ F_{X|Y}(x \,|\, y) = \sqrt{1 + \mathrm{SNR}}\left(x - \frac{\mathrm{SNR}}{1 + \mathrm{SNR}} \cdot y\right) \quad (21)$$

and hence the posterior matching scheme is given by

$$X_1 = F_X^{-1}(\Theta_0),$$
$$X_{n+1} = \sqrt{1 + \mathrm{SNR}}\left(X_n - \frac{\mathrm{SNR}}{1 + \mathrm{SNR}} Y_n\right). \quad (22)$$

From the above, we see that at time $n + 1$, the transmitter sends the error term pertaining to the MMSE estimate of X_n from Y_n, scaled to match the permissible input power P. In fact, it can be verified either directly or using the equivalence stated in Theorem 2 that X_{n+1} is the scaled MMSE term of X_n given the entire output sequence Y^n. Therefore, the posterior matching scheme in this case is an infinite-horizon, variable-rate variant of the Schalkwijk-Kailath scheme. This variant is in fact even somewhat simpler than the original scheme [8], since the initial matching step of the random message point makes transmission start at a steady-state. The fundamental difference between the posterior matching principle and the Schalkwijk-Kailath "parameter estimation" approach in a non-Gaussian setting, is now

evident. According to Schalkwijk-Kailath one should transmit a scaled linear MMSE term given past observations, which is *uncorrelated* with these observations but *not independent* of them as dictated by the posterior matching principle; the two notions thus coincide only in the AWGN case. In fact, it can be shown that following the Schalkwijk-Kailath approach when the additive noise is not Gaussian results in achieving only the corresponding "Gaussian equivalent" capacity, see Example 12.

Example 5 (BSC): Let $P_{Y|X}$ be a BSC with crossover probability p, and set a capacity achieving input distribution $X \sim$ Bernoulli$(\frac{1}{2})$, i.e., $f_X(x) = \frac{1}{2}(\delta(x) + \delta(x-1))$. We now derive the posterior matching scheme for this setting, and show it reduces to the Horstein scheme [3]. The conditions of Theorem 2 are not satisfied since the input distribution is discrete, and we therefore use the original nonrecursive representation (12) for now. It is easy to see that the matching step F_X^{-1} acts as a quantizer above/below $\frac{1}{2}$, and so we get

$$X_{n+1} = F_X^{-1} \circ F_{\Theta_0|Y^n}(\Theta_0|Y^n)$$
$$= \begin{cases} 0 & \Theta_0 < \text{median}\{f_{\Theta_0|Y^n}(\theta|Y^n)\} \\ 1 & o.w. \end{cases}$$

which is precisely the Horstein scheme. The posterior matching principle is evident in this case, since slicing the posterior distribution at its median results in an input $X_{n+1} \sim$ Bernoulli$(\frac{1}{2})$ given any possible output $Y^n = y^n$, and is hence independent of Y^n and Bernoulli$(\frac{1}{2})$-distributed. We return to the BSC example later in this section, after we develop the necessary tools to provide an alternative (and more useful) recursive representation for the Horstein scheme.

Example 6 (Uniform Input/Noise): Let $P_{Y|X}$ be an additive noise channel with noise uniformly distributed over the unit interval, and set the input $X \sim \mathcal{U}$, i.e., uniform over the unit interval as well. Let us derive the posterior matching scheme in this case. It is easy to verify that the inverse channel's p.d.f. is given by

$$f_{X|Y}(x|y) = \begin{cases} y^{-1}\mathbb{1}_{(0,y)}(x) & y \in (0,1] \\ (2-y)^{-1}\mathbb{1}_{(y-1,1)}(x) & y \in (1,2) \end{cases}.$$

Since the conditions of Theorem 2 are satisfied, we can use the recursive representation. We note that since the input distribution is $\mathcal{U}$, the matching step is trivial and the posterior matching kernel is given by

$$F_X^{-1} \circ F_{X|Y}(x|y)$$
$$= F_{X|Y}(x|y)$$
$$= \begin{cases} \frac{x}{y} \cdot \mathbb{1}_{(0,y)}(x) + \mathbb{1}_{[y,\infty)}(x) & y \in (0,1] \\ \frac{x-y+1}{2-y} \cdot \mathbb{1}_{(y-1,1)}(x) + \mathbb{1}_{[1,\infty)}(x) & y \in (1,2) \end{cases} \quad (23)$$

and therefore the posterior matching scheme is given by

$$X_1 = \Theta_0,$$
$$X_{n+1} = \frac{X_n}{Y_n} \cdot \mathbb{1}_{(0,1]}(Y_n) + \frac{X_n - Y_n + 1}{2 - Y_n} \cdot \mathbb{1}_{(1,2)}(Y_n). \quad (24)$$

The above has in fact a very simple interpretation. The desired input distribution is uniform, so we start by transmitting the message point $X_1 = \Theta_0$. Then, given Y_1, we determine the range of inputs that could have generated this output value, and

find an affine transformation that stretches this range to fill the entire unit interval. Applying this transformation to X_1 generates X_2. We now determine the range of possible inputs given Y_2, and apply the corresponding affine transformation to X_2, and so on. This is intuitively appealing since what we do in each iteration is just *zoom-in* on the remaining uncertainty region for Θ_0. Since the posterior distribution is always uniform, this zooming-in is linear.

The posterior distribution induced by this transmission strategy is uniform in an ever shrinking sequence of intervals. Therefore, a zero-error variable-rate decoding rule would be to simply decode at time n the (random) posterior support interval J_n. The size of that interval is

$$|J_n| = \prod_{k \in J} Y_k \prod_{k \notin J} (2 - Y_k)$$

where $J = \{k : 1 \leq k \leq n, Y_k < 1\}$. Denoting the channel noise sequence by $Z_n \sim P_Z$, the corresponding rate is

$$R_n = -\frac{1}{n} \log |J_n| = \frac{1}{n} \sum_{k \in J} \log \frac{1}{Y_k} + \frac{1}{n} \sum_{k \notin J} \log \frac{1}{2 - Y_k}$$
$$= \frac{1}{n} \sum_{k=1}^n \log \frac{f_{X|Y}(X_k|Y_k)}{f_X(X_k)} = \frac{1}{n} \sum_{k=1}^n \log \frac{f_Z(Z_k)}{f_Y(Y_k)}$$
$$\xrightarrow[n \to \infty]{} \mathbb{E} \log f_Z(Z) - \mathbb{E} \log f_Y(Y) \quad \text{a.s.}$$
$$= I(X;Y) = \frac{1}{2} \log e$$

where we have used the SLLN for the i.i.d. sequences Z^n, Y^n. Therefore, in this simple case we were able to directly show that the posterior matching scheme, in conjunction with a simple variable rate decoding rule, achieves the mutual information with zero-error probability. In the sequel, the achievability of the mutual information and the tradeoff between rate, error probability and transmission period obtained by the posterior matching scheme are derived for a general setting. We then revisit this example and provide the same results as above from this more general viewpoint.

Example 7 (Exponential Input/Noise): Consider an additive noise channel $P_{Y|X}$ with $\sim$ Exponential(1) noise, and set the input $X \sim$ Exponential(1) as well. This selection is not claimed to be capacity achieving under any reasonable input constraints, yet it is instructive to study due to the simplicity of the resulting scheme. We will return to the exponential noise channel in Example 10 after developing the necessary tools, and analyze it using the capacity achieving distribution under an input mean constraint.

It is easy to verify that for the above simple selection, the input given the output is uniformly distributed, i.e., the inverse channel p.d.f./c.d.f. are given by

$$f_{X|Y}(x|y) = \frac{1}{y} \cdot \mathbb{1}_{(0,y)}(x)$$
$$F_{X|Y}(x|y) = \frac{x}{y} \cdot \mathbb{1}_{(0,y)}(x) + \mathbb{1}_{[y,\infty)}(x).$$

The input's inverse c.d.f. is given by

$$F_X^{-1}(s) = \ln\left(\frac{1}{1-s}\right).$$
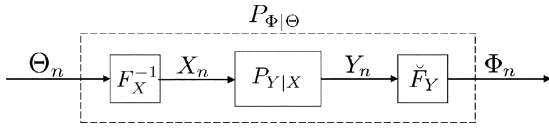
Fig. 1.   Normalized channel $P_{\Phi \mid \Theta}$.

Therefore, the posterior matching kernel is given by

$$F_X^{-1} \circ F_{X \mid Y}(x \mid y) = \ln \left(\frac{y}{y - x} \right) \qquad (25)$$

and the posterior matching scheme in this case is simply given by

$$X_1 = \ln \left(\frac{1}{1 - \Theta_0} \right), \quad X_{n+1} = \ln \left(\frac{Y_n}{Y_n - X_n} \right). \qquad (26)$$

C. The Normalized Channel

The recursive representation provided in Theorem 2 is inapplicable in many interesting cases, including DMCs in particular. In order to treat discrete, continuous, and mixed alphabet inputs/channels within a common framework, we define for any input/channel pair $(P_X, P_{Y \mid X})$ a corresponding *normalized channel* $P_{\Phi \mid \Theta}$ with $(0, 1)$ as a common input/output alphabet, and a uniform input distribution $\Theta \sim \mathcal{U}$. The normalized channel is obtained by viewing the matching operator $F_X^{-1}(\cdot)$ as part of the original channel, and applying the output c.d.f. operator $F_Y(\cdot)$ to the channel's output, with the technical exception that whenever $F_Y(\cdot)$ has a jump discontinuity the output is randomly selected uniformly over the jump span.[14] This is depicted in Fig. 1, where $\breve{F}_Y(\cdot)$ stands for the aforementioned possibly random mapping. This construction is most simply formalized by

$$P_{Y \mid \Theta}(\cdot \mid \theta) = P_{Y \mid X} \left(\cdot \mid F_X^{-1}(\theta) \right), \quad \Phi = F_Y(Y) - P_Y(Y) \cdot \Lambda \qquad (27)$$

where $\Theta \sim \mathcal{U}$, and $\Lambda \sim \mathcal{U}$ is statistically independent of (Θ, Y).

Lemma 12 (Normalized Channel Properties): Let $(P_\Theta, P_{\Phi \mid \Theta})$ be the normalized input/channel pair corresponding to the pair $(P_X, P_{Y \mid X})$. The following properties are satisfied:

  (i) $\Phi \sim \mathcal{U}$, i.e., $P_{\Phi \mid \Theta}$ preserves the uniform distribution over the unit interval.
  (ii) The mutual information is preserved, i.e.,

$$I(\Theta; \Phi) = I(X; Y)$$

  (iii) The joint distribution $P_{\Theta \Phi}$ is proper.
  (iv) The normalized kernel $F_{\Theta \mid \Phi}(\theta \mid \phi)$ is continuous in θ for P_Φ-a.a. $\phi \in (0, 1)$.
  Proof:
  (i) By Lemma 1 claim (i) we have $F_X^{-1}(\Theta) \sim P_X$, and so $Y \sim P_Y$ in (27). The result now follows from Lemma 1 claim (ii).

  (ii) An easy exercise using the relations in Fig. 1, and noting that X, Y are always uniquely recoverable from Θ, Φ, respectively.
  (iii) See Appendix A.
  (iv) Follows easily from (iii).   ∎

The posterior matching scheme over the normalized channel with a uniform input, is given by

$$\bar{g}_{n+1}(\theta, \phi^n) = F_{\Theta_0 \mid \Phi^n} (\theta \mid \phi^n).$$

The properties of the normalized channel allows for a unified recursive representation of the above scheme via the inverse normalized channel $P_{\Theta \mid \Phi}$ corresponding to $(P_\Theta, P_{\Phi \mid \Theta}) = (\mathcal{U}, P_{\Phi \mid \Theta})$, i.e., in terms of the *normalized posterior matching kernel* $F_{\Theta \mid \Phi}$.

Theorem 3 (Recursive Representation II): The posterior matching scheme for the normalized channel is given by the recursive relation:

$$\bar{g}_1(\theta) = \theta, \quad \bar{g}_{n+1}(\theta \mid \phi^n) = F_{\Theta \mid \Phi}(\cdot \mid \phi_n) \circ \bar{g}_n(\theta \mid \phi^{n-1}). \qquad (28)$$

The corresponding sequence of input/output pairs $\{(\Theta_n, \Phi_n)\}_{n=1}^\infty$ constitutes a Markov chain over a state space $\underline{\mathrm{supp}}(\theta, \Phi) \subseteq (0, 1)^2$, with an invariant distribution $P_{\Theta \Phi}$, and satisfy the recursion rule

$$\Theta_1 = \Theta_0, \quad \Theta_{n+1} = F_{\Theta \mid \Phi}(\Theta_n \mid \Phi_n). \qquad (29)$$

Furthermore, (29) is equivalent to the posterior matching scheme (13) in the sense that the distribution of the sequence $\{F_X^{-1}(\Theta_n), F_Y^{-1}(\Phi_n)\}_{n=1}^\infty$ coincides with the distribution of the sequence $\{(X_n, Y_n)\}_{n=1}^\infty$.

Proof: By Lemma 12, the joint distribution $P_{\Theta \Phi}$ is proper, hence Theorem 2 is applicable and the recursive representations and Markovity follow immediately. Once again, taking the state space to be $\underline{\mathrm{supp}}(\Theta, \Phi)$ and not $\mathrm{supp}(\Theta, \Phi)$ is artificial and is done for reasons of mathematical convenience, to avoid having the trivial invariant distributions $P_0 \times P_{\Phi \mid \Theta}(\cdot \mid 0)$ and $P_1 \times P_{\Phi \mid \Theta}(\cdot \mid 1)$, where $P_0(0) = 1, P_1(1) = 1$. The distribution $P_{\Theta \Phi}$ is invariant by construction, and the equivalence to the original scheme is by definition.   ∎

In the sequel, an initial point for the aforementioned Markov chain will be given by a fixed value $\theta_0 \in (0, 1)$ of the message point only[15]. Notice also that the Theorem above reveals an interesting fact: Whenever F_X^{-1} is not injective, the sequence of input/output pairs pertaining to the original posterior matching scheme (13) is a *hidden Markov process*. In particular, this is true for the BSC and the Horstein scheme.

Example 5 (BSC, Continued): The normalized channel's p.d.f. corresponding to a BSC with crossover probability p and a Bernoulli($\frac{1}{2}$) input distribution is given by $f_{\Phi \mid \Theta}(\phi \mid \theta) = 2(1 - p)$ when θ, ϕ are either both smaller than $\frac{1}{2}$, or both larger than $\frac{1}{2}$, and $f_{\Phi \mid \Theta}(\phi \mid \theta) = 2p$ otherwise. Following

---

[14]The output mapping is of a lesser importance, and is introduced mainly to provide a common framework.

[15]This is an abuse of notations, since an initial point is properly given by a pair (θ_1, ϕ_1). However, it can be justified since $\Theta_1 = \Theta_0$ and Φ_1 is generated via a memoryless channel. Hence, any statement that holds for a.a/all initial points (θ_1, ϕ_1) also holds in particular for a.a./all θ_0.
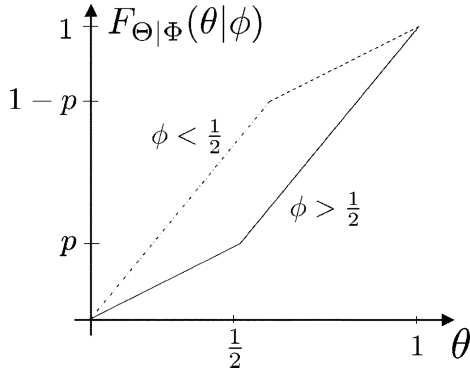
Fig. 2. BSC normalized posterior matching kernel.

Theorem 3 and simple manipulations, the corresponding normalized posterior matching kernel is given by

$$
F_{\Theta \mid \Phi}(\theta \mid \phi)
$$
$$
= \begin{cases}
2(1-p)\theta & \theta \in (0, \frac{1}{2}), \phi \in (0, \frac{1}{2}) \\
2p\theta + (1-2p) & \theta \in [\frac{1}{2}, 1), \phi \in (0, \frac{1}{2}) \\
2p\theta & \theta \in (0, \frac{1}{2}), \phi \in [\frac{1}{2}, 1) \\
2(1-p)\theta - (1-2p) & \theta \in [\frac{1}{2}, 1), \phi \in [\frac{1}{2}, 1)
\end{cases} \quad (30)
$$

and for a fixed ϕ is supported on two functions of θ, depending on whether $\phi \lessgtr \frac{1}{2}$ which corresponds to $y = 0, 1$ in the original discrete setting, see Fig. 2. Therefore, the posterior matching scheme (which is equivalent to the Horstein scheme in this case) is given by the following recursive representation:

$$
\Theta_1 = \Theta_0
$$
$$
\Theta_{n+1}
$$
$$
= \begin{cases}
2(1-p)\Theta_n & \Theta_n \in (0, \frac{1}{2}), \Phi_n \in (0, \frac{1}{2}) \\
2p\Theta_n + (1-2p) & \Theta_n \in [\frac{1}{2}, 1), \Phi_n \in (0, \frac{1}{2}) \\
2p\Theta_n & \Theta_n \in (0, \frac{1}{2}), \Phi_n \in [\frac{1}{2}, 1) \\
2(1-p)\Theta_n - (1-2p) & \Theta_n \in [\frac{1}{2}, 1), \Phi_n \in [\frac{1}{2}, 1)
\end{cases}.
$$

. The hidden Markov process describing the original Horstein scheme is recovered from the above by setting

$$
X_k = F_X^{-1}(\Theta_k) = \mathbb{1}_{[\frac{1}{2}, 1)}(\Theta_k)
$$
$$
Y_k = F_Y^{-1}(\Phi_k) = \mathbb{1}_{[\frac{1}{2}, 1)}(\Phi_k).
$$

Example 8 (The Binary Erasure Channel (BEC)): The binary erasure channel is defined over the input alphabet $\mathcal{X} = \{0, 1\}$ and the output alphabet $\mathcal{Y} = \{0, 1, 2\}$. Given any input, the output is equal to that input with probability p, and equal to 2 with probability $1 - p$. Using the capacity achieving distribution $P_X = \text{Bernoulli}(\frac{1}{2})$, it is easy to see from the nonrecursive representation (12) that the posterior matching scheme in this case is exactly the simple repetition rule—transmit the first bit of Θ_0 until it is correctly received, then continue to the next bit and so on. This scheme clearly achieves the capacity $1 - p$. The recursive representation w.r.t. the normalized channel is very simple and intuitive here as well. The normalized posterior matching kernel is supported on three functions—the identity function corresponding to the erasure output 2, and the functions $2\theta, 2\theta - 1$ that correspond to the outputs $0, 1$, respectively.

Example 9 (General DMC): The case where $P_{Y \mid X}$ is a DMC and P_X is a corresponding discrete input distribution is a simple extension of the BSC/BEC settings. The normalized posterior matching kernel is supported over a finite number of $|\mathcal{Y}|$ continuous functions, which are all quasi-affine relative to a fixed partition of the unit interval into subintervals corresponding to the input distribution. Precisely, for any $x \in \mathcal{X}$ the normalized posterior matching kernel evaluated at $\theta = F_X(x)$ is given by

$$
F_{\Theta \mid \Phi}(F_X(x) \mid \phi) = F_{X \mid Y}\left(x \mid F_Y^{-1}(\phi)\right) \quad (31)
$$

and by a linear interpolation in between these points. Hence, the corresponding kernel slopes are given by $\frac{P_{X \mid Y}(x \mid y)}{P_X(x)}$.

Example 10 (Exponential Noise, Input Mean Constraint): Consider an additive noise channel $P_{Y \mid X}$ with $\sim \text{Exponential}(b)$ noise, but now instead of arbitrarily assuming an exponential input distribution as in Example 7, let us impose an input mean constraint (x, a), i.e.,

$$
\lim_{n \to \infty} n^{-1} \sum_{k=1}^{n} X_k \leq a \quad \text{a.s.}
$$

The capacity achieving distribution under this input constraint was determined in [19] to be a *mixture* of a deterministic distribution and an exponential distribution, with the following generalized p.d.f.:

$$
f_X(x) = \frac{b}{a+b}\,\delta(x) + \frac{a}{(a+b)^2} \exp\left(-\frac{x}{a+b}\right).
$$

Under this input distribution the output is $Y \sim \text{Exponential}(a + b)$, and the capacity can be expressed in closed form

$$
C = I(X; Y) = \log\left(1 + \frac{a}{b}\right)
$$

in a remarkable resemblance to the AWGN channel with an input power constraint. Interestingly, in this case the posterior matching scheme can also be written in closed form, and as stated later, also achieves the channel capacity under the input mean constraint.

To derive the scheme, we must resort to the normalized representation since the input distribution is not proper. The input's inverse c.d.f. and the output's c.d.f. are given by

$$
F_X^{-1}(\theta) = (a+b) \ln\left(\frac{a}{(a+b)(1-\theta)}\right) \cdot \mathbb{1}_{[\frac{b}{a+b}, 1)}(\theta)
$$
$$
F_Y(y) = 1 - \exp\left(-\frac{y}{a+b}\right).
$$

Using the normalized representation and practicing some algebra, we find that the normalized posterior matching kernel is given by

$$
F_{\Theta \mid \Phi}(\theta \mid \phi) = (1-\phi)^{\frac{a}{b}} \left(\frac{a+b}{b} \cdot \theta \cdot \mathbb{1}_{(0, \frac{b}{a+b})}(\theta) \right.
$$
$$
\left. + \left(\frac{a}{a+b} \cdot \frac{1}{1-\theta}\right)^{\frac{a}{b}} \mathbb{1}_{[\frac{b}{a+b}, 1 - \frac{a(1-\phi)}{a+b})}(\theta) \right)
$$
$$
+ \mathbb{1}_{(1 - \frac{a(1-\phi)}{a+b}, \infty)}(\theta). \quad (32)
$$

Thus, the posterior matching scheme in this case is given by

$$\Theta_1 = \Theta_0,$$
$$\Theta_{n+1} = \begin{cases} \frac{a+b}{b} \cdot \Theta_n \cdot (1 - \Phi_n)^{\frac{a}{b}} & \Theta_n \leq \frac{b}{a+b} \\ \left(\frac{a}{a+b} \cdot \frac{1 - \Phi_n}{1 - \Theta_n} \right)^{\frac{a}{b}} & \Theta_n > \frac{b}{a+b} \end{cases} \quad (33)$$

where the original channel's input/output pairs are given by

$$X_n = (a+b) \ln \left(\frac{a}{(a+b)(1 - \Theta_n)} \right) \mathbb{1}_{[\frac{b}{a+b}, 1)}(\Theta_n)$$
$$Y_n = (a+b) \ln \frac{1}{1 - \Phi_n}$$

and constitute a hidden Markov process. Note that since we have $\Theta_n \in (0, 1 - \frac{a(1 - \Phi_n)}{a+b})$ a.s., then $\Theta_{n+1} \in (0, 1)$ a.s. and we need not worry about the rest of the thresholds appearing in (32).

IV. REGULARITY CONDITIONS FOR INPUT/CHANNEL PAIRS

In Section V, we prove the optimality of the posterior matching scheme. However, to that end we first need to introduce several regularity conditions, and define some well behaved families of input/channel pairs.

For any fixed $\phi \in (0, 1)$, define θ_ϕ^- and θ_ϕ^+ to be the unique solutions of

$$F_{\Theta \mid \Phi}(\theta_\phi^- | \phi) = \frac{1}{2} F_{\Theta \mid \Phi}(\theta | \phi)$$
$$1 - F_{\Theta \mid \Phi}(\theta_\phi^+ | \phi) = \frac{1}{2} \left(1 - F_{\Theta \mid \Phi}(\theta | \phi) \right)$$

respectively. For any $\varepsilon > 0$, define the *left-ε-measure* $^-P_{\Phi \mid \Theta}^\varepsilon(\cdot | \theta)$ of $P_{\Phi \mid \Theta}(\cdot | \theta)$ to have a density $^-f_{\Phi \mid \Theta}^\varepsilon$ given by

$$^-f_{\Phi \mid \Theta}^\varepsilon(\phi | \theta) \stackrel{\text{def}}{=} \inf_{\xi \in J_\varepsilon^-(\phi, \theta)} f_{\Phi \mid \Theta}(\phi | \xi)$$

where the interval $J_\varepsilon^-(\phi, \theta)$ is defined to be

$$J_\varepsilon^-(\phi, \theta) \stackrel{\text{def}}{=} (\max(\theta_\phi^-, \theta - \varepsilon), \theta). \quad (34)$$

Note that the left-ε-measure is not a probability distribution since in general $^-P_{\Phi \mid \Theta}^\varepsilon((0, 1) | \theta) < 1$. Similarly, define *right-ε-measure* $^+P_{\Phi \mid \Theta}^\varepsilon(\cdot | \theta)$ of $P_{\Phi \mid \Theta}(\cdot | \theta)$ to have a density $^+f_{\Phi \mid \Theta}^\varepsilon$ given by

$$^+f_{\Phi \mid \Theta}^\varepsilon(\phi | \theta) \stackrel{\text{def}}{=} \inf_{\xi \in J_\varepsilon^+(\phi, \theta)} f_{\Phi \mid \Theta}(\phi | \xi)$$

where the interval $J_\varepsilon^+(\phi, \theta)$ is defined to be

$$J_\varepsilon^+(\phi, \theta) \stackrel{\text{def}}{=} (\theta, \min(\theta + \varepsilon, \theta_\phi^+)). $$

Note that $\lim_{\varepsilon \to 0} {}^-f_{\Phi \mid \Theta}^\varepsilon = \lim_{\varepsilon \to 0} {}^+f_{\Phi \mid \Theta}^\varepsilon = f_{\Phi \mid \Theta}$ a.e. over $(0, 1)^2$. Following these definitions, an input/channel pair $(P_X, P_{Y \mid X})$ is said to be *regular*, if the corresponding normalized channel satisfies

$$\inf_{\varepsilon > 0} \left[D(P_{\Phi \mid \Theta} \| ^-P_{\Phi \mid \Theta}^\varepsilon | P_\Theta) + D(P_{\Phi \mid \Theta} \| ^+P_{\Phi \mid \Theta}^\varepsilon | P_\Theta) \right]$$
$$< \infty.$$

Loosely speaking, the regularity property guarantees that the *sensitivity* of the channel law $P_{Y \mid X}$ to input perturbations is not too high, or is at least attenuated by a proper selection of the input distribution P_X. Regularity is satisfied in many interesting cases, as demonstrated in the following Lemma.

Lemma 13: Each of the following conditions implies that the input/channel pair $(P_X, P_{Y \mid X})$ is regular:

(i) $h(\Theta, \Phi)$ is finite, $\underline{\text{supp}}(\Theta, \Phi)$ is convex in the θ-direction, and $f_{\Theta\Phi}$ is bounded away from zero over $\underline{\text{supp}}(\Theta, \Phi)$.

(ii) P_{XY} is proper, $\underline{\text{supp}}(X, Y)$ is convex in the x-direction, f_X is bounded, and $f_{X \mid Y}$ has a uniformly bounded max-to-min ratio, i.e.,

$$\sup_{y \in \underline{\text{supp}}(Y)} \left(\frac{\sup_{x \in \underline{\text{supp}}(X \mid Y = y)} f_{X \mid Y}(x | y)}{\inf_{x \in \underline{\text{supp}}(X \mid Y = y)} f_{X \mid Y}(x | y)} \right) < \infty.$$

(iii) P_{XY} is proper, and $f_{X \mid Y}(x | y)$ is unimodal with a regular tail and a bounded variance, uniformly over $y \in \underline{\text{supp}}(Y)$.

(iv) $P_{Y \mid X}$ is a DMC with nonzero transition probabilities.
  Proof: See Appendix C.   ∎

For an input/channnel pair $(P_X, P_{Y \mid X})$, define the following set of properties:

(A1) $(P_X, P_{Y \mid X})$ *is regular.*

(A2) *The invariant distribution* $P_{\Theta\Phi}$ *for the Markov chain* $\{(\Theta_n, \Phi_n)\}_{n=1}^\infty$, *is ergodic.*

(A3) $F_{\Theta \mid \Phi}$ *is fixed-point free, i.e., for any* $\theta \in (0, 1)$

$$\mathbb{P}(F_{\Theta \mid \Phi}(\theta | \Phi) = \theta) < 1. \quad (35)$$

(A4) P_X *achieves the unconstrained capacity over* $P_{Y \mid X}$, *i.e.,* $I(X; Y) = C(P_{Y \mid X})$.[16]

The following is easily observed.

Lemma 14: (A2) $\Rightarrow$ (A3).
  Proof: See proof of Lemma 21.   ∎

Let Ω_A be the family of all input/channel pairs satisfying properties (A1) and (A2). Let Ω_B be the family of all input/channel pairs satisfying properties (A1), (A3) and (A4). In the sequel, we show that for members in $\Omega_A \cup \Omega_B$ the corresponding posterior matching scheme achieves the mutual information. However, while Lemma 13 provides means to verify the regularity Property (A1), and Properties (A3) and (A4) are easy to check, the ergodicity property (A2) may be difficult to verify in general. Therefore, we introduce the following more tractable property:

(A5) f_{XY} *is bounded and continuous over* $\underline{\text{supp}}(X, Y)$, *where the latter is connected and convex in the y-direction.*

We now show that (A3) and (A5) together imply a stronger version of (A2). In fact, to that end a weaker version of (A3) is sufficient, which we state (for convenience) in terms of the nonnormalized kernel:

(A3*) *For any* $x \in \underline{\text{supp}}(X)$ *there exists* $y \in \underline{\text{supp}}(Y)$, *such that* $F_X^{-1} \circ F_{X \mid Y}(x | y) \neq x$.

---

[16]Since an input/channel pair has finite mutual information, (A4) implies that $C(P_{Y \mid X}) < \infty$. The unconstrained capacity is finite for discrete input and/or output channels, but can be finite under other input alphabet constraints (e.g., an amplitude constraint).

Lemma 15: $(A3^*) \land (A5) \Rightarrow \{(\Theta_n, \Phi_n)\}_{n=1}^{\infty}$ is p.h.r. and aperiodic $\Rightarrow (A2)$.

Proof: For the first implication, see Appendix B. The second implication is immediate since p.h.r. implies in particular a unique invariant distribution, which is hence ergodic. ∎

Following that, let us define Ω_C to be the family of all input/channel pairs $(P_X, P_{Y|X})$ satisfying properties $(A1), (A3^*)$ and $(A5)$.

Corollary 2: $\Omega_C \subset \Omega_A$.

Turning to the discrete case, let $(P_X, P_{Y|X})$ be an input/DMC pair. Without loss of generality, we will assume throughout that $\min_{x \in \mathcal{X}} P_X(x) > 0$, as otherwise the unused input can be removed. Define the following set of properties:

(B1) $\min_{x \in \mathcal{X}, y \in \mathcal{Y}} P_{Y|X}(y|x) > 0$

(B2) At least one of the following holds:
 (i) $\exists y \in \mathcal{Y}, P_Y(y) > 0,$ s.t. $P_X \prec_d P_{X|Y}(\cdot|y)$ or $P_{X|Y}(\cdot|y) \prec_d P_X.$
 (ii) $\exists y_0, y_1 \in \mathcal{Y}, P_Y(y_0) > 0, P_Y(y_1) > 0,$ s.t. $P_{X|Y}(\cdot|y_0) \prec_d P_{X|Y}(\cdot|y_1).$

(B3) $\forall x \in \mathcal{X}, \exists y_0, y_1 \in \mathcal{Y}$ s.t. $0 > \frac{\beta_0}{\beta_1} \notin \mathbb{Q}$, where[17]

$$\beta_i \overset{\text{def}}{=} \log\left(\frac{P_{X|Y}(x|y_i)}{P_X(x)}\right), \quad i \in \{0, 1\}.$$

Lemma 16: Let $(P_X, P_{Y|X})$ be an input/DMC pair. Then:
 (i) $(B1) \Rightarrow (A1)$.
 (ii) $(B2) \Rightarrow (A3)$.
 (iii) $(B1) \land (B3) \land (A3) \Rightarrow (A2)$.
 (iv) $|\mathcal{X}| = 2 \Rightarrow (B2)$.
 (v) $(B1) \land I(X;Y) > 0 \Rightarrow$ there exists an equivalent pair $(P_{X^*}, P_{Y^*|X^*})$ satisfying $(B1) \land (B2)$.
 (vi) $\forall \varepsilon > 0, \exists P'_X$ s.t. $d_{TV}(P_X, P'_X) < \varepsilon$, and $(P'_X, P_{Y|X})$ is an input/DMC pair satisfying $(B3)$.

Proof: Claim (i) follows immediately from condition (iv) of Lemma 13. Claim (iv) holds since any two nonidentical binary distributions can be ordered by dominance. For the remaining claims, see Appendix A. ∎

Remark 1: The equivalent pair in Lemma 16, claim (v), is obtained via an input permutation only, which is given explicitly in the proof and can be simply computed.

V. ACHIEVING THE MUTUAL INFORMATION

Our main theorem is presented in Section V-A, establishing the achievability of the mutual information via posterior matching for a large family of input/channel pairs. The examples of Section III are then revisited, and the applicability of the theorem is verified in each. Section V-B is dedicated to the proof of the Theorem.

A. Main Result

Theorem 4 (Achievability): Consider an input/channel pair $(P_X, P_{Y|X}) \in \Omega_A \cup \Omega_B$ (resp. Ω_C). The corresponding posterior matching scheme with a fixed/variable rate optimal decoding rule, achieves (resp. pointwise achieves) any rate $R < I(X;Y)$ over the channel $P_{Y|X}$. Furthermore, if

---

[17] $\mathbb{Q}$ is the set of rational numbers. Note that there always exists a pair for which $\frac{\beta_0}{\beta_1} < 0$, but the quotient is not necessarily irrational.

$(P_X, P_{Y|X}) \in \Omega_A$ (resp. Ω_C), then R is achieved (resp. pointwise achieved) within an input constraint $(\eta, \mathbb{E}\eta(X))$, for any measurable $\eta : \mathcal{X} \mapsto \mathbb{R}$ satisfying $\mathbb{E}|\eta(X)| < \infty$.

Example 4 (AWGN, Continued): P_{XY} is proper (jointly Gaussian), and the inverse channel's p.d.f. $f_{X|Y}(x|y)$ is Gaussian with a variance independent of y, hence by Lemma 2 condition (ii) has a regular tail uniformly in y. Therefore, by condition (iii) of Lemma 13, the Gaussian input/AWGN channel pair $(P_X, P_{Y|X})$ is regular and Property $(A1)$ is satisfied. It is easy to see that the linear posterior matching kernel (21) is fixed-point free, and so Property $(A3^*)$ is satisfied as well. Finally, f_{XY} is continuous and bounded over a $\mathbb{R}^2$ support, so Property $(A5)$ is also satisfied. Therefore $(P_X, P_{Y|X}) \in \Omega_C$, and Theorem 4 verifies the well known fact that the Schalkwijk-Kailath scheme (pointwise) achieves any rate below the capacity $I(X;Y) = \frac{1}{2}\log(1 + \text{SNR})$.

Example 5 (BSC, Continued): The pair of a Bernoulli($\frac{1}{2}$) input P_X and a BSC $P_{Y|X}$ with any nontrivial crossover probability $p \neq 0, 1$, satisfies properties $(A4)$ and $(B1)$. Properties $(A1)$ and $(A3)$ follow from claims (i), (ii) and (iv) of Lemma 16. Hence $(P_X, P_{Y|X}) \in \Omega_B$ and Theorem 4 implies that the posterior matching scheme, which coincides in this case with the Horstein scheme, indeed achieves the capacity $I(X;Y) = 1 - h_b(p)$. *This settles in the affirmative a longstanding conjecture.*

Remark 2: In the BSC Example above, it also holds [via Lemma 16, claim (iii)] that $(P_X, P_{Y|X}) \in \Omega_A$ for a.a. crossover probabilities p, except perhaps for the countable set $S = \{p : \frac{1+\log p}{1+\log(1-p)} \in \mathbb{Q}\}$ where property $(B3)$ is not satisfied. In these cases the ergodicity property $(A2)$ is not guaranteed, though this may be an artifact of the proof (see Remark 7). Therefore, although capacity is achieved for any p (via Ω_B), Theorem 4 guarantees the empirical distribution of the input sequence X^n to approach P_X only for $p \notin S$. However, since P_X is the *unique* capacity achieving distribution, this sample-path property of the input sequence holds for $p \in S$ nonetheless (see Remark 6).

Remark 3: Interestingly, for $p \in S$ the Horstein medians exhibit "regular behavior", meaning that any median point can always be returned to in a fixed number of steps. In fact, for the subset of S where $\frac{1+\log p}{1+\log(1-p)} = -k$ for some positive integer $k \geq 2$, the Horstein scheme can be interpreted as a simple finite-state constrained encoder that precludes subsequences of more than k consecutive 0's or 1's, together with an insertion mechanism repeating any erroneously received bit $k+1$ times. This fact was identified and utilized in [4] to prove achievability in this special case.

Example 6 (Uniform Input/Noise, Continued): P_{XY} is proper with a bounded p.d.f. over the convex support $\underline{\text{supp}}(X, Y) = (0, 1) \times (0, 2)$, the marginal p.d.f. f_X is bounded, and the inverse channel's p.d.f. is uniform hence has a bounded max-to-min ratio. Therefore, condition (ii) of Lemma 13 holds, and properties $(A1)$ and $(A5)$ are satisfied. It is readily verified that the kernel (23) is fixed-point free, and so property $(A3^*)$ is satisfied as well. Therefore $(P_X, P_{Y|X}) \in \Omega_C$, and Theorem 4 reverifies that the simple posterior matching scheme (24) pointwise achieves the mutual information $I(X;Y) = \frac{1}{2}\log e$, as previously established by direct calculation. In fact, we have already

seen that (variable-rate) zero-error decoding is possible in this case, and in the next section we arrive at the same conclusion from a different angle.

Example 7 (Exponential Input/Noise, Continued): P_{XY} is proper with a bounded p.d.f. over the convex support $\underline{\text{supp}}(X, Y) = \mathbb{R}^+ \times \mathbb{R}^+$, the marginal p.d.f. f_X is bounded, and the inverse channel's p.d.f. is uniform hence has a bounded max-to-min ratio. Therefore, condition (ii) of Lemma 13 holds, and properties (A1) and (A5) are satisfied. It is readily verified that the kernel (25) is fixed-point free, and so property (A3*) is satisfied as well. Therefore $(P_X, P_{Y|X}) \in \Omega_C$, and so by Theorem 4 the posterior matching scheme (26) pointwise achieves the mutual information, which is this case is $I(X; Y) \approx 0.8327$.

Example 9 (General DMC, Continued): It has already been demonstrated that the posterior matching scheme achieves the capacity of the BSC. We now show that the same holds true for a general DMC, up to some minor resolvable technicalities. Let $P_{Y|X}$ be a DMC with nonzero transition probabilities, and set P_X to be capacity achieving (unconstrained). Hence, properties (B1) and (A4) are satisfied, and by Lemma 16, claim (i), property (A1) holds as well. The corresponding posterior matching scheme in this case is equivalent to a generalized Horstein scheme, which was conjectured to achieve the unconstrained capacity when there are no fixed points, namely when property (A3) is satisfied [6, Section 4.6]. Since in this case $(P_X, P_{Y|X}) \in \Omega_B$, Theorem 4 verifies that this conjecture indeed holds. Moreover, the restriction of not having fixed points is in fact superfluous, since by Lemma 16, claim (v), there always exists an equivalent input/DMC pair (obtained simply by an input permutation) for which the posterior matching scheme is capacity achieving. This scheme can be easily translated into an equivalent optimal scheme for the original channel $P_{Y|X}$, which is in fact one of the many μ-variants satisfying the posterior matching principle mentioned in Corollary 1, where the u.p.f. μ plays the role of the input permutation. This observation is further discussed and generalized in Section VII-A.

More generally, let P_X be any input distribution for $P_{Y|X}$, e.g., capacity achieving under some input constraints. If the associated kernel is fixed-point free ((A3) holds) and (B3) is satisfied, then by Lemma 16, claim (iii), we have that (A2) holds as well. This implies $(P_X, P_{Y|X}) \in \Omega_A$, and hence by Theorem 4 the associated posterior matching scheme achieves rates up to the corresponding mutual information $I(X; Y)$, within any input constraints encapsulated in P_X. Again, the fixed-point requirement is superfluous, and achievability within the same input constraints can be guaranteed via a posterior matching scheme for an equivalent channel (or the corresponding μ-variant), for which the kernel is fixed-point free.

It is worth noting that requiring property (B3) to hold is practically nonrestrictive. For any fixed alphabet sizes $|\mathcal{X}|, |\mathcal{Y}|$, there is only a countable number of input/channel pairs that fail to satisfy this property. Moreover, even if $(P_X, P_{Y|X})$ does not satisfy (B3), then by Lemma 16, claim (iii), we can find an input distribution P'_X arbitrarily close (in total variation) to P_X, such that (B3) does hold for $(P'_X, P_{Y|X})$. Hence, the posterior matching scheme (or a suitable variant, if there are fixed points) for $(P'_X, P_{Y|X})$ achieves rates arbitrarily close to

$I(X; Y)$ while maintaining any input constraint encapsulated in P_X arbitrarily well.

Remark 4: For input/DMC pairs such that $(P_X, P_{Y|X}) \in \Omega_B$ but where (B3) does not hold, ergodicity is not guaranteed (see also Remark 7). Therefore, although the (unconstrained) capacity is achieved, the empirical distribution of the input sequence X^n will not necessarily approach P_X, unless P_X is the *unique* capacity achieving distribution for $P_{Y|X}$ (see Remark 6).

Remark 5: The nonzero DMC transition probabilities restriction (B1) is mainly intended to guarantee that the regularity property (A1) is satisfied (although this property holds under somewhat more general conditions, e.g., for the BEC.). However, regularity can be defined in a less restricting fashion so that this restriction could be removed. Roughly speaking, this can be done by redefining the left-ε-measure and right-ε-measure of Section IV so that the neighborhoods over which the infimum is taken shrink near some finite collection of points in (0, 1), and not only near the endpoints, thereby allowing "holes" in the conditional densities. For simplicity of exposition, this extension was left out.

Example 10 (Exponential Noise With an Input Mean Constraint, Continued): This example is not immediately covered by the Lemmas developed. However, studying the input/channel pair $(P_{Y|\Theta}, P_\Theta)$ (namely, the normalized pair but without the artificial output transformation), we see that $P_{\Theta Y}$ satisfies property (A5), and the corresponding posterior matching kernel (which is easily derived from (32)) is fixed-point free, hence property (A3*) is also satisfied. Proving that this is a regular pair is straightforward but requires some work. Loosely speaking, it stems from the fact that $f_{Y|\Theta}(y|\theta)$ is monotonically decreasing in y for any fixed θ, and has a one-sided regular tail. Therefore, the posterior matching scheme (33) pointwise achieves any rate below the mean-constrained capacity $I(X; Y) = \log\left(1 + \frac{a}{b}\right)$.

B. Proof of Theorem 4

Let us start by providing a rough outline of the proof. First, we show that zero rate is achievable, i.e., any fixed interval around the message point accumulates a posterior probability mass that tends to one. This is done by noting that the time evolution of the posterior c.d.f. $F_{\Theta_0|\Phi^n}$ can be represented by an IFS over the space $\mathfrak{F}_c$, generated by the inverse channel's c.d.f. via function composition, and controlled by the channel outputs. Showing that the inverse channel's c.d.f. is contractive on the average (Lemma 17), we conclude that the posterior c.d.f. tends to a unit step function about Θ_0 (Lemma 18) which verifies zero-rate achievability. For positive rates, we use the SLLN for Markov chains to show that the posterior p.d.f. at the message point is $\approx 2^{nI(X;Y)}$ (Lemma 19). Loosely speaking, a point that cannot be distinguished from Θ_0 must induce, from the receiver's perspective, about the same input sequence as does the true message point. Since the normalized inputs are just the posterior c.d.f. sequence evaluated at the message point, this means that such points will also have about the same c.d.f. sequence as Θ_0 does, hence also will have a posterior p.d.f. $\approx 2^{nI(X;Y)}$. But that is only possible within an interval no larger

than $\approx 2^{-nI(X;Y)}$ around Θ_0, since the posterior p.d.f. integrates to unity. Thus, points that cannot be distinguished from Θ_0 must be $2^{-nI(X;Y)}$ close to it. This is more of a converse, but essentially the same ideas can be applied (Lemma 20) to show that for any $R < I(X;Y)$, a 2^{-nR} neighborhood of the message point accumulates (with high probability) a posterior probability mass exceeding some fixed $\varepsilon > 0$ at some point during the first n channel uses. This essentially reduces the problem to the zero-rate setting, which was already solved.

We begin by establishing the required technical Lemmas.

Lemma 17: Let $(P_X, P_{Y|X})$ satisfy property (A3). Then there exist a contraction $\xi(\cdot)$ and a length function $\psi_\lambda(\cdot)$ as in (8) over $\mathfrak{F}_c$, such that for any $h \in \mathfrak{F}_c$

$$\mathbb{E}\left(\psi_\lambda\left[F_{\Theta|\Phi}(\cdot\,|\,\Phi)\circ h\right]\right) \leq \xi(\psi_\lambda(h)). \tag{36}$$

Proof: See Appendix A. ∎

Define the stochastic process $\{\bar{G}_n\}_{n=1}^\infty$,

$$\bar{G}_n(\cdot) \overset{\text{def}}{=} \bar{g}_n(\cdot, \Phi^{n-1}).$$

Since $\bar{g}_n(\theta, \phi^{n-1}) = F_{\Theta_0|\Phi^{n-1}}(\theta\,|\,\phi^{n-1})$, $\bar{G}_n$ is the posterior c.d.f. of the message point after observing the i.i.d. output sequence Φ^{n-1}, and is a r.v. taking values in the c.d.f. space $\mathfrak{F}_c$. Moreover, by (28), we have that

$$\bar{G}_{n+1} = F_{\Theta|\Phi}(\cdot\,|\,\Phi_n)\circ \bar{G}_n \tag{37}$$

and therefore $\{\bar{G}_n\}_{n=1}^\infty$ is an IFS over $\mathfrak{F}_c$, generated by the normalized posterior matching kernel $F_{\Theta|\Phi}(\cdot\,|\,\phi)$ (via function composition) and controlled by the outputs $\{\Phi_n\}_{n=1}^\infty$. Since the message point is uniform, the IFS initializes at $\bar{G}_1(\theta) = \theta\mathbb{1}_{(0,1)}(\theta) + \mathbb{1}_{[1,\infty)}(\theta)$ (the uniform c.d.f.). Recall that the normalized kernel is continuous in θ for P_Φ-a.a. ϕ (Lemma 12, claim (iv)), hence $\bar{G}_n$ is a.s. continuous.

We find it convenient to define the δ-*positive trajectory* $\{^+\Theta_k^\delta\}_{k=1}^\infty$ and δ-*negative trajectory* $\{^-\Theta_k^\delta\}_{k=1}^\infty$, as follows:

$$^+\Theta_k^\delta \overset{\text{def}}{=} \bar{G}_k(\Theta_0 + \Delta_\delta^+), \quad \Delta_\delta^+ = \min\left(\delta, \frac{1-\Theta_0}{2}\right)$$

$$^-\Theta_k^\delta \overset{\text{def}}{=} \bar{G}_k(\Theta_0 - \Delta_\delta^-), \quad \Delta_\delta^- = \min\left(\delta, \frac{\Theta_0}{2}\right). \tag{38}$$

These trajectories are essentially the posterior c.d.f. evaluated after k steps at a δ perturbation from Θ_0 (up to edge issues), or alternatively the induced normalized input sequence for such a perturbation from the point of view of the receiver. The true normalized input sequence, which corresponds to the c.d.f. evaluated at the message point itself, is $\Theta_k = \bar{G}_k(\Theta_0)$.

The next Lemma shows that for a zero rate, the trajectories diverge towards the boundaries of the unit interval with probability approaching one, hence our scheme has a vanishing error probability in this special case.

Lemma 18: Let $(P_X, P_{Y|X})$ satisfy property (A3). Then for any $\varepsilon > 0, \delta > 0$

$$\mathbb{P}\left(^-\Theta_n^\delta > \varepsilon\right) = O(\sqrt[8]{r(n)})$$

$$\mathbb{P}\left(^+\Theta_n^\delta < 1 - \varepsilon\right) = O(\sqrt[8]{r(n)})$$

where $r(n)$ is the decay profile of the contraction $\xi(\cdot)$ from Lemma 17.

Proof: Let ψ_λ and ξ be the length function and contraction from Lemma 17 corresponding to the pair $(P_X, P_{Y|X})$, and let

$r(n)$ be the decay profile of ξ. By the contraction property (36) and Lemma 9, we immediately have that for any $\nu > 0$

$$\mathbb{P}\left(\psi_\lambda(\bar{G}_n) > \nu\right) \leq \nu^{-1}r(n). \tag{39}$$

Define the (random) *median point* of $\bar{G}_n$

$$\Theta_n^* \overset{\text{def}}{=} \inf\left\{\theta \in (0,1) : \bar{G}_n(\theta) \geq \frac{1}{2}\right\}.$$

Since $\bar{G}_n$ is a.s. continuous, $\bar{G}_n(\Theta_n^*) = \frac{1}{2}$ is a.s. satisfied. Using the symmetry of the function $\lambda(\cdot)$, we can write

$$\psi_\lambda(\bar{G}_n) = \int_0^{\Theta_n^*} \lambda(\bar{G}_n(\theta))\,d\theta + \int_{\Theta_n^*}^1 \lambda(1 - \bar{G}_n(\theta))\,d\theta \quad \text{a.s.} \tag{40}$$

and then:

$$\mathbb{P}\left(\bar{G}_n(\Theta_n^* - \delta) > \nu\right) \overset{(a)}{\leq} \mathbb{P}\left(\lambda\left(\bar{G}_n(\Theta_n^* - \delta)\right) > \nu\right)$$

$$\overset{(b)}{\leq} \mathbb{P}\left(\int_{\Theta_n^* - \delta}^{\Theta_n^*} \lambda\left(\bar{G}_n(\theta)\right)\,d\theta > \nu\delta\right)$$

$$\overset{(c)}{\leq} \mathbb{P}(\psi_\lambda(\bar{G}_n) > \nu\delta)$$

where (a) holds since $\lambda(\theta) > \theta$ for any $\theta \in (0, \frac{1}{2})$, in (b) we use the monotonicity of $\bar{G}_n$, and (c) follows from (40). Using (39) this leads to

$$\mathbb{P}\left(\bar{G}_n(\Theta_n^* - \delta) > \nu\right) \leq \frac{1}{\nu\delta}r(n) \tag{41}$$

and similarly

$$\mathbb{P}\left(\bar{G}_n(\Theta_n^* + \delta) < 1 - \nu\right) \leq \frac{1}{\nu\delta}r(n). \tag{42}$$

Now set any $\eta \in (0, \frac{1}{2})$, and write

$$\mathbb{P}\left(\int_0^{\Theta_0} \bar{G}_n(\theta)d\theta + \int_{\Theta_0}^1 \left(1 - \bar{G}_n(\theta)\right)d\theta > \nu\right)$$

$$\overset{(a)}{\leq} \mathbb{P}\left(\left\{\int_0^{\Theta_n^*} \bar{G}_n(\theta)d\theta + \int_{\Theta_n^*}^1 \left(1 - \bar{G}_n(\theta)\right)d\theta > \frac{\nu}{2}\right\}\right.$$

$$\left.\cup\left\{|\Theta_n^* - \Theta_0| > \frac{\nu}{2}\right\}\right)$$

$$\overset{(b)}{\leq} \mathbb{P}\left(\psi_\lambda(\bar{G}_n) > \frac{\nu}{2}\right) + \mathbb{P}\left(|\Theta_n^* - \Theta_0| > \frac{\nu}{2}\right)$$

$$\overset{(c)}{\leq} \frac{2}{\nu}r(n) + \mathbb{P}\left(|\Theta_n^* - \Theta_0| > \frac{\nu}{2}\right)$$

$$\overset{(d)}{=} \frac{2}{\nu}r(n) + \mathbb{P}\left(\left\{\bar{G}_n(\Theta_0) > \bar{G}_n\left(\Theta_n^* + \frac{\nu}{2}\right)\right\}\right.$$

$$\left.\cup\left\{\bar{G}_n(\Theta_0) < \bar{G}_n\left(\Theta_n^* - \frac{\nu}{2}\right)\right\}\right)$$

$$\overset{(e)}{\leq} \frac{2}{\nu}r(n) + \mathbb{P}\left(\bar{G}_n(\Theta_0) \notin (\eta, 1-\eta)\right)$$

$$+ \mathbb{P}\left(\bar{G}_n\left(\Theta_n^* - \frac{\nu}{2}\right) > \eta\right)$$

$$+ \mathbb{P}\left(\bar{G}_n\left(\Theta_n^* + \frac{\nu}{2}\right) < 1 - \eta\right)$$

$$\overset{(f)}{\leq} \frac{2}{\nu}r(n) + 2\eta + \frac{4}{\nu\eta}r(n) \tag{43}$$

where in (a) we use the fact that integrals differ only over the interval between Θ_0 and Θ_0^* and the integrands are bounded by unity, in (b) we use the union bound, and then (39) by noting that applying $\lambda(\cdot)$ can only increase the integrands, in (c) we use (39), and (d) holds by the continuity and monotonicity of $\bar{G}_n$. These properties are applied again together with the union bound in (e), and the inequality holds for any $\eta \in (0, \frac{1}{2})$. Finally, in (f) we use (41)–(42) the fact that $\bar{G}_n(\Theta_0) = \Theta_n$ is uniformly distributed over the unit interval. Choosing $\eta = \sqrt{r(n)}$, we get

$$\mathbb{P}\left(\int_0^{\Theta_0} \bar{G}_n(\theta)\, d\theta + \int_{\Theta_0}^1 (1 - \bar{G}_n(\theta))\, d\theta > \nu\right)$$
$$\leq c\nu^{-1}\sqrt{r(n)}$$

for some $c > 0$. The same bound clearly holds separately for each of the two integrals above. Define the set

$$\Pi_\nu \overset{\text{def}}{=} \left\{ \theta \in (0, 1) : \mathbb{P}\left(\int_0^{\Theta_0} \bar{G}_n(\theta)d\theta > \nu \mid \Theta_0 = \theta\right) \right.$$
$$\left. \geq c\nu^{-1}\sqrt[4]{r(n)}\right\}.$$

Then

$$\mathbb{P}\left(\int_0^{\Theta_0} \bar{G}_n(\theta)d\theta > \nu\right) = \mathbb{E}\mathbb{P}\left(\int_0^{\Theta_0} \bar{G}_n(\theta)d\theta > \nu \mid \Theta_0\right)$$
$$\geq \mathbb{P}(\Theta_0 \in \Pi_\nu) \cdot c\nu^{-1}\sqrt[4]{r(n)}$$

and we get $\mathbb{P}(\Theta_0 \in \Pi_\nu) \leq \sqrt[4]{r(n)}$. Let us now set $\nu_n = \sqrt[8]{r(n)}$, and suppose n is large enough so that $\nu_n < \frac{\varepsilon\delta}{2}$. Recalling the definition of the negative trajectory $^-\Theta_n^\delta$, we have

$$\mathbb{P}\left(^-\Theta_n^\delta > \varepsilon\right) = \mathbb{E}\mathbb{P}\left(^-\Theta_n^\delta > \varepsilon \mid \Theta_0\right)$$
$$\leq \int_0^1 \mathbb{P}\left(\int_0^{\Theta_0} \bar{G}_n(\theta)d\theta > \frac{\varepsilon}{2}\cdot\min\{\delta, \theta\} \,\middle|\, \Theta_0 = \theta\right) d\theta$$
$$\leq \int_0^1 \mathbb{P}\left(\int_0^{\Theta_0} \bar{G}_n(\theta)d\theta > \min\left\{\nu_n, \frac{\varepsilon\theta}{2}\right\} \,\middle|\, \Theta_0 = \theta\right) d\theta$$
$$\leq \mathbb{P}(\Theta_0 \in \Pi_{\nu_n}) + \int_0^{2\nu_n\varepsilon^{-1}} d\theta + \int_{2\nu_n\varepsilon^{-1}}^1 c\nu_n^{-1}\sqrt[4]{r(n)}d\theta$$
$$\leq \sqrt[4]{r(n)} + 2\nu_n\varepsilon^{-1} + c\nu_n^{-1}\sqrt[4]{r(n)} = O(\sqrt[8]{r(n)}).$$

The result for $\mathbb{P}(^+\Theta_n^\delta < 1 - \varepsilon)$ is proved via the exact same arguments. ∎

Lemma 19: Let $(P_X, P_{Y\mid X})$ satisfy property (A2). Then the posterior p.d.f. evaluated at the message point satisfies

$$\lim_{n\to\infty} \frac{1}{n}\log f_{\Theta_0\mid\Phi^n}(\Theta_0\mid\Phi^n) = I(X;Y) \quad \text{a.s.} \quad (44)$$

Proof: Since the p.d.f.'s involved are all proper, we can use Bayes law to obtain the following recursion rule:

$$f_{\Theta_0\mid\Phi^n}(\theta\mid\phi^n)$$
$$= \frac{f_{\Phi_n\mid\Theta_0,\Phi^{n-1}}(\phi_n\mid\theta, \phi^{n-1})}{f_{\Phi_n\mid\Phi^{n-1}}(\phi_n\mid\phi^{n-1})} f_{\Theta_0\mid\Phi^{n-1}}(\theta\mid\phi^{n-1})$$
$$= f_{\Phi\mid\Theta}(\phi_n\mid\bar{g}_n(\theta, \phi^{n-1})) \cdot f_{\Theta_0\mid\Phi^{n-1}}(\theta\mid\phi^{n-1}) \quad (45)$$

where in the second equality we have used the memoryless channel property and the fact that the output sequence Φ^∞ is an i.i.d. sequence with marginal $\mathcal{U}$. Applying the recursion rule n times, taking a logarithm and evaluating at the message point, we obtain

$$\frac{1}{n}\log f_{\Theta_0\mid\Phi^n}(\Theta_0\mid\Phi^n)$$
$$= \frac{1}{n}\sum_{k=1}^n \log f_{\Phi\mid\Theta}\left(\Phi_k\mid g_k(\Theta_0, \Phi^{k-1})\right)$$
$$= \frac{1}{n}\sum_{k=1}^n \log f_{\Phi\mid\Theta}(\Phi_k\mid\Theta_k).$$

Now by property (A2) the invariant distribution $P_{\Theta\Phi}$ is ergodic, and so we can use the SLLN for Markov chains (Lemma 7) which asserts in this case that for P_Θ-a.a. $\theta_0 \in (0, 1)$

$$\lim_{n\to\infty} \frac{1}{n}\log f_{\Theta_0\mid\Phi^n}(\theta_0\mid\Phi^n)$$
$$= \mathbb{E}\left(\log \frac{f_{\Phi\mid\Theta}(\Phi\mid\Theta)}{f_\Phi(\Phi)}\right) = I(\Theta;\Phi)$$
$$= I(X;Y) \quad \mathcal{P}_{\theta_0}\text{-a.s.}$$

Since $\Theta_0 \sim P_\Theta$, (44) is established. ∎

For short, let us now define the (n, R)-*positive trajectory* $^+\Theta_k^{n,R}$ and the (n, R)-*negative trajectory* $^-\Theta_k^{n,R}$ as the corresponding trajectories in (38) with $\delta = 2^{-nR}$. Accordingly, we also write $\Delta_{n,R}^+, \Delta_{n,R}^-$ in lieu of $\Delta_\delta^+, \Delta_\delta^-$, respectively. The following Lemma uses the SLLN to demonstrate how, for rates lower than the mutual information, these two trajectories eventually move away from some small and essentially fixed neighborhood of the input, with probability approaching one. This is achieved by essentially proving a more subtle version of Lemma 19, showing that it roughly holds at the vicinity of the message point.

Lemma 20: Let $(P_X, P_{Y\mid X})$ satisfy properties (A1) and (A2). Then for any rate $R < I(X;Y)$ there exists $\varepsilon > 0$ small enough such that

$$\lim_{n\to\infty} \mathbb{P}\left(\bigcap_{k=1}^n \left\{\Theta_k - ^-\Theta_k^{n,R} < \min\left(\varepsilon, \frac{\Theta_k}{2}\right)\right\}\right) = 0$$
$$\lim_{n\to\infty} \mathbb{P}\left(\bigcap_{k=1}^n \left\{^+\Theta_k^{n,R} - \Theta_k < \min\left(\varepsilon, \frac{1-\Theta_k}{2}\right)\right\}\right) = 0.$$
$$(46)$$

Proof: We prove the first assertion of (46), the second assertion follows through essentially the same way. Let $\delta > 0$ be such that $R < I(X;Y) - \delta$. Let $^-P_{\Phi\mid\Theta}^\varepsilon(\cdot\mid\theta)$ be the left-ε-measure corresponding to $P_{\Phi\mid\Theta}(\cdot\mid\theta)$, as defined in Section IV. Define

$$I_\varepsilon^- \overset{\text{def}}{=} \mathbb{E}\log {}^-f_{\Phi\mid\Theta}^\varepsilon(\Phi\mid\Theta)$$
$$= \iint_{\text{supp}(\Theta,\Phi)} f_{\Phi\mid\Theta}(\phi\mid\theta)\log {}^-f_{\Phi\mid\Theta}^\varepsilon(\phi\mid\theta)\, d\theta d\phi. \quad (47)$$

We have that

$$0 \le I(X;Y) - I_\varepsilon^- = I(\Theta;\Phi) - I_\varepsilon^-$$
$$= D(P_{\Phi\,|\,\Theta}\|^- P_{\Phi\,|\,\Theta}^\varepsilon\,|\,P_\Theta)$$

and since by property (A1) the input/channel is regular then $\inf_{\varepsilon>0} D(P_{\Phi\,|\,\Theta}\|^- P_{\Phi\,|\,\Theta}^\varepsilon\,|\,P_\Theta) < \infty$, hence for any ε small enough

$$-\infty < I_\varepsilon^- \le I(X;Y).$$

We have therefore established that the function $f_{\Phi\,|\,\Theta} \log{}^- f_{\Phi\,|\,\Theta}^\varepsilon$ is finitely integrable for any $\varepsilon > 0$ small enough, and converges to $f_{\Phi\,|\,\Theta} \log f_{\Phi\,|\,\Theta}$ a.e in a monotonically nondecreasing fashion, as $\varepsilon \to 0$. Applying Levi's monotone convergence Theorem [20], we can exchange the order of the limit and the integration to obtain

$$\lim_{\varepsilon\to 0} I_\varepsilon^- = I(X;Y)$$

Let us set ε hereinafter so that

$$I_\varepsilon^- - \varepsilon > I(X;Y) - \frac{\delta}{2}.$$

Since I_ε^- is finite we can once again apply (using property (A2)) the SLLN for Markov chains (Lemma 7) to obtain

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \log{}^- f_{\Phi\,|\,\Theta}^\varepsilon(\Phi_k|\Theta_k) = I_\varepsilon^- \quad \text{a.s.} \qquad (48)$$

The above intuitive relation roughly means that if the receiver, when considering the likelihood of a wrong message point, obtains an induced input sequence which is always ε-close to the true input sequence, then the posterior p.d.f. at this wrong message point will be close to that of the true message point given in Lemma 19.

Define the following two sequences of events:

$$E_{n,\varepsilon} \overset{\text{def}}{=} \bigcap_{k=1}^{n} \left\{ {}^- \Theta_k^{n,R} \in J_\varepsilon^-(\Theta_k, \Phi_k) \right\}$$
$$\tilde{E}_{n,\varepsilon} \overset{\text{def}}{=} \bigcap_{k=1}^{n} \left\{ \Theta_k - {}^- \Theta_k^{n,R} < \min\left(\varepsilon, \frac{\Theta_k}{2}\right) \right\} \qquad (49)$$

where the neighborhood J_ε^- is defined in (34). Let us now show that $\lim_{n\to\infty} \mathbb{P}(E_{n,\varepsilon}) = 0$. This fact will then be shown to imply $\lim_{n\to\infty} \mathbb{P}(\tilde{E}_{n,\varepsilon}) = 0$, which is precisely the first assertion in (46). Define the following sequence of events:

$$T_{n,\varepsilon} \overset{\text{def}}{=} \{\Theta_0 > 2^{-nR}\}$$
$$\cap \left\{ \frac{1}{n} \sum_{k=1}^{n} \log{}^- f_{\Phi\,|\,\Theta}^\varepsilon(\Phi_k|\Theta_k) \ge I(X;Y) - \frac{\delta}{2} \right\}.$$

Using (48) and the fact that the message point is uniform over the unit interval, it is immediately clear that $\mathbb{P}(T_{n,\varepsilon}) \to 1$. For

short, define the random interval $J_{n,R} \overset{\text{def}}{=} (\Theta_0 - \Delta_{n,R}^-, \Theta_0)$, and consider the following chain of inequalities:

$$0 \ge \log \mathbb{E}\left(\Theta_n - {}^- \Theta_n^{n,R}\right)$$
$$= \log \mathbb{E}\left(\bar{G}_n(\Theta_0) - \bar{G}_n(\Theta_0 - \Delta_{n,R}^-)\right)$$
$$\ge \log \mathbb{E}\left(\Delta_{n,R}^- \cdot \inf_{\theta\in J_{n,R}} \left(f_{\Theta_0\,|\,\Phi^{n-1}}\left(\theta|\Phi^{n-1}\right)\right)\right)$$
$$\ge \log \mathbb{P}(E_{n,\varepsilon} \cap T_{n,\varepsilon})$$
$$\quad + \log \mathbb{E}\left(\Delta_{n,R}^-\right.$$
$$\quad \left. \cdot \inf_{\theta\in J_{n,R}} f_{\Theta_0\,|\,\Phi^{n-1}}\left(\theta|\Phi^{n-1}\right) | E_{n,\varepsilon} \cap T_{n,\varepsilon}\right)$$
$$\overset{(a)}{\ge} \log \mathbb{P}(E_{n,\varepsilon} \cap T_{n,\varepsilon}) + \mathbb{E}(\log \Delta_{n,R}^- \,|\, E_{n,\varepsilon} \cap T_{n,\varepsilon})$$
$$\quad + \mathbb{E}\left(\log \inf_{\theta\in J_{n,R}} \prod_{k=1}^{n} f_{\Phi\,|\,\Theta}(\Phi_k|\bar{G}_k(\theta)) \,|\, E_{n,\varepsilon} \cap T_{n,\varepsilon}\right)$$
$$\overset{(b)}{\ge} \log \mathbb{P}(E_{n,\varepsilon} \cap T_{n,\varepsilon}) - nR - 1$$
$$\quad + \mathbb{E}\left(\sum_{k=1}^{n} \log \inf_{\xi\in(-\Theta_k^{n,R},\Theta_k)} f_{\Phi\,|\,\Theta}(\Phi_k|\xi) \,|\, E_{n,\varepsilon} \cap T_{n,\varepsilon}\right)$$
$$\overset{(c)}{\ge} \log \mathbb{P}(E_{n,\varepsilon} \cap T_{n,\varepsilon}) - nR - 1$$
$$\quad + \mathbb{E}\left(\sum_{k=1}^{n} \log{}^- f_{\Phi\,|\,\Theta}^\varepsilon(\Phi_k|\Theta_k) \,|\, E_{n,\varepsilon} \cap T_{n,\varepsilon}\right)$$
$$\overset{(d)}{\ge} \log \mathbb{P}(E_{n,\varepsilon} \cap T_{n,\varepsilon}) - nR - 1 + n\left(I(X;Y) - \frac{\delta}{2}\right).$$

In (a), we use Jensen's inequality and the expansion of the posterior p.d.f. given in (45), in (b) we use the definition and monotonicity of $\bar{G}_k$, (c) holds due to $E_{n,\varepsilon}$, and (d) due to $T_{n,\varepsilon}$. Therefore

$$\mathbb{P}(E_{n,\varepsilon} \cap T_{n,\varepsilon}) \le 2^{-n\left(I(X;Y)-\frac{\delta}{2}-R\right)-1}$$
$$\le 2^{-n\frac{\delta}{2}-1} \underset{n\to\infty}{\longrightarrow} 0 \qquad (50)$$

where the last inequality holds since $R < I(X;Y) - \delta$. Now, since $\mathbb{P}(T_{n,\varepsilon}) \to 1$, then for any $\eta > 0$ we have $\mathbb{P}(T_{n,\varepsilon}) > 1-\eta$ for n large enough. Using that and (50), we bound $\mathbb{P}(E_{n,\varepsilon})$ simply as follows:

$$\mathbb{P}(E_{n,\varepsilon}) \le \mathbb{P}(E_{n,\varepsilon} \cap T_{n,\varepsilon}) + (1 - \mathbb{P}(T_{n,\varepsilon}))$$
$$\le 2^{-n\frac{\delta}{2}-1} + \eta \le 2\eta \qquad (51)$$

where the last two inequalities are true for n large enough. Since (51) holds for any $\eta > 0$, we conclude that $\mathbb{P}(E_{n,\varepsilon}) \to 0$, as desired.

To finalize the proof, note that $\tilde{E}_{n,\varepsilon}$ implies that for any $1 \le k \le n-1$

$$\Theta_k - {}^- \Theta_k^{n,R} < \varepsilon, \quad {}^- \Theta_k^{n,R} > \frac{\Theta_k}{2}$$

and the rightmost inequality implies

$$F_{\Theta\,|\,\Phi}\left({}^- \Theta_{k-1}^{n,R}|\Phi_{k-1}\right) > \frac{1}{2} F_{\Theta\,|\,\Phi}\left(\Theta_{k-1}|\Phi_{k-1}\right).$$

The above constraints imply that for any $1 \leq k \leq n-2$

$$^-\Theta_k^{n,R} \in J_\varepsilon^-(\Phi_k, \Theta_k)$$

establishing the implication $\tilde{E}_{n,\varepsilon} \Rightarrow E_{n-1,\varepsilon}$. Consequently, $\mathbb{P}(\tilde{E}_{n,\varepsilon}) \leq \mathbb{P}(E_{n-1,\varepsilon})$, thus $\lim_{n\to\infty} \mathbb{P}(\tilde{E}_{n,\varepsilon}) = 0$. ∎

We are now finally in a position to prove Theorem 4 for the family Ω_A. Loosely speaking, we build on the simple fact that since the chain is stationary by construction, one can imagine transmission to have started at any time m with a message point Θ_m replacing Θ_0. With some abuse of notations, we define

$$^-\Theta_k^\varepsilon(\Theta_m) \stackrel{\text{def}}{=} F_{\Theta\,|\,\Phi}(\cdot\,|\,\Phi_{m+k-1}) \circ \cdots \circ F_{\Theta\,|\,\Phi}(\cdot\,|\,\Phi_{m+1})$$
$$\circ F_{\Theta\,|\,\Phi}\left(\Theta_m - \min\left(\varepsilon, \frac{\Theta_m}{2}\right)\,|\,\Phi_m\right).$$

Namely, the ε-negative trajectory when starting at time m from Θ_m. Note that in particular, we have $^-\Theta_k^\varepsilon(\Theta_0) = ^-\Theta_k^\varepsilon$. Since the chain is stationary, the distribution of $^-\Theta_k^\varepsilon(\Theta_m)$ is independent of m. The corresponding positive trajectory $^+\Theta_k^\varepsilon(\Theta_m)$ can be defined in the same manner.

Now recall the event $\tilde{E}_{n,\varepsilon}$ defined in (49), which by Lemma 20 satisfies $\mathbb{P}(\tilde{E}_{n,\varepsilon}) \to 0$ for any $\varepsilon > 0$ small enough. Note that the complementary event $\tilde{E}_{n,\varepsilon}^c$ implies that at some time $m \leq n$, the (n,R)-negative trajectory $^-\Theta_m^{n,R}$ is below the ε-neighborhood of Θ_m, namely $^-\Theta_m^{n,R} \leq \Theta_m - \min(\varepsilon, \frac{\Theta_m}{2})$ for some m. Using the monotonicity of the transmission functions, this in turn implies that the (n,R)-negative trajectory at time n lies below the corresponding ε-negative trajectory starting from Θ_m, namely $^-\Theta_n^{n,R} \leq ^-\Theta_{n-m}^\varepsilon(\Theta_m)$. Thus, we conclude that $\mathbb{P}(^-\Theta_n^{n,R} > \max(^-\Theta_{n-m}^\varepsilon(\Theta_m))) \to 0$ for any fixed $\varepsilon > 0$ small enough, where the maximum is taken over $1 \leq m \leq n$.

Fixing any $\alpha > 0$, we now show that $^-\Theta_{(1+\alpha)n}^{n,R} \to 0$ in probability:

$$\mathbb{P}\left(^-\Theta_{(1+\alpha)n}^{n,R} > \delta\right)$$
$$\leq \mathbb{P}\left(^-\Theta_n^{n,R} > \max_{1\leq m\leq n} {}^-\Theta_{n-m}^\varepsilon(\Theta_m)\right)$$
$$+ \mathbb{P}\left(\left\{^-\Theta_{(1+\alpha)n}^{n,R} > \delta\right\}\right.$$
$$\left.\cap\left\{^-\Theta_n^{n,R} \leq \max_{1\leq m\leq n} {}^-\Theta_{n-m}^\varepsilon(\Theta_m)\right\}\right)$$
$$\leq o(1) + \mathbb{P}\left(\max_{1\leq m\leq n} {}^-\Theta_{(1+\alpha)n-m}^\varepsilon(\Theta_m) > \delta\right)$$
$$\stackrel{(a)}{\leq} o(1) + \sum_{m=1}^n \mathbb{P}\left(^-\Theta_{(1+\alpha)n-m}^\varepsilon(\Theta_m) > \delta\right)$$
$$\stackrel{(b)}{=} o(1) + \sum_{m=1}^n \mathbb{P}\left(^-\Theta_{(1+\alpha)n-m}^\varepsilon(\Theta_0) > \delta\right)$$
$$\stackrel{(c)}{=} o(1) + \delta^{-1} O(n\sqrt[8]{r(\alpha n)}). \tag{52}$$

In (a) we used the union bound, in (b) the fact that the chain is stationary, and Lemma 18 was invoked in (c), where we recall that $(A2) \Rightarrow (A3)$ by Lemma 14. Therefore, if $\sqrt[8]{r(n)} = o(n^{-1})$ then $^-\Theta_{(1+\alpha)n}^{n,R} \to 0$ in probability is established. However,

we note that this mild constraint[18] is in fact superfluous. This stems from the fact that the union bound in (a) is very loose since the trajectories are all controlled by the same output sequence, and from the uniformity in the initial point in Lemma 9. In Appendix A, Lemma 25, we show that in fact

$$\mathbb{P}\left(\max_{1\leq m\leq n} {}^-\Theta_{(1+\alpha)n-m}^\varepsilon(\Theta_m) > \delta\right) = O(\sqrt[8]{r(\alpha n)}).$$

Following the derivations in (52), this in turn implies that $^-\Theta_{(1+\alpha)n}^{n,R} \to 0$ in probability without the additional constraint on the decay profile.

The same derivation applies to the positive trajectory, resulting in $^+\Theta_{(1+\alpha)n}^{n,R} \to 1$ in probability. Therefore, for any $\delta > 0$

$$\mathbb{P}\left(^+\Theta_{(1+\alpha)n}^{n,R} - {}^-\Theta_{(1+\alpha)n}^{n,R} < 1 - 2\delta\right)$$
$$\leq \mathbb{P}\left(^-\Theta_{(1+\alpha)n}^{n,R} > \delta\right) + \mathbb{P}\left(^+\Theta_{(1+\alpha)n}^{n,R} < 1 - \delta\right) = o(1)$$

and so the posterior probability mass within a 2^{-nR} symmetric neighborhood of Θ_0 (up to edge issues) after $(1+\alpha)n$ iterations, approaches one in probability as $n \to \infty$. We can therefore find a sequence $\delta_n \to 0$ such that the probability this mass exceeds $1 - \delta_n$ tends to zero. Using the optimal variable rate decoding rule and setting the target error probability to $p_e(n) = \delta_n$, we immediately have that $\mathbb{P}(R_n < (1+\alpha)^{-1}R) \to 0$. This holds for any $R < I(X;Y)$, and since $\alpha > 0$ can be arbitrarily small, any rate below the mutual information is achievable.

To prove achievability using the optimal fixed-rate decoding rule, note that any variable-rate rule achieving some rate $R > 0$ induces a fixed-rate rule achieving an arbitrarily close rate $R-\varepsilon$, by extending the variable-sized decoded interval into a larger one of a fixed size $2^{-n(R-\varepsilon)}$ whenever the former is smaller, and declaring an error otherwise. Therefore, any rate $R < I(X;Y)$ is achievable using the optimal fixed rate decoding rule. The fact that the input constraint is satisfied follows immediately from the SLLN since the marginal invariant distribution for the input is P_X. This concludes the achievability proof for the family Ω_A.

Extending the proof to the family Ω_B requires reproving Lemma 19 and a variation of Lemma 20, where the ergodicity property $(A2)$ is replaced with the maximality property $(A4)$. This is done via the ergodic decomposition [14] for the associated stationary Markov chain. The proof appears in Appendix A, Lemma 26. Achievability for the family Ω_C has already been established, since $\Omega_C \subset \Omega_A$. The stronger pointwise achievability statement for Ω_C is obtained via p.h.r. properties of the associated Markov chain, by essentially showing that Lemmas 18, 19 and 20 hold given any fixed message point. The proof appears in Appendix B, Lemma 27.

Remark 6: For $(P_X, P_{Y\,|\,X}) \in \Omega_B \backslash \Omega_A$, although the unconstrained capacity $C(P_{Y\,|\,X})$ is achieved, there is no guarantee on the sample path behavior of the input, which may generally differ from the expected behavior dictated by P_X, and depend on the ergodic component the chain lies in. However, if P_X is

---

[18]An exponentially decaying $r(n)$ can in fact be guaranteed by requiring the normalized posterior matching kernel to be fixed-point free in a somewhat stronger sense than that implied by property $(A2)$, which also holds in particular in all the examples considered in this paper.

the *unique* input distribution[19] such that $I(X;Y) = C(P_{Y|X})$, then the sample path behavior will nevertheless follow P_X independent of the ergodic component. This is made precise in Appendix A, Lemma 26.

VI. ERROR PROBABILITY ANALYSIS

In this section, we provide two sufficient conditions on the target error probability facilitating the achievability of a given rate using the corresponding optimal variable rate decoding rule. The approach here is substantially different from that of the previous subsection, and the derivations are much simpler. However, the obtained result is applicable only to rates below some thresholds $R^*, R^\dagger$. Unfortunately, it is currently unknown under what conditions do these thresholds equal the mutual information, rendering the previous section indispensable.

Loosely speaking, the basic idea is the following. After having observed Φ^n, say the receiver has some estimate $\hat{\theta}_{n+1}$ for the next input Θ_{n+1}. Then $(\hat{\theta}_{n+1}, \Phi^n)$ correspond to a unique estimate $\hat{\theta}_0$ of the message point which is recovered by *reversing* the transmission scheme, i.e., running a RIFS over (0,1) generated by the kernel $\omega_\phi(\cdot) \overset{\text{def}}{=} F_{\Theta|\Phi}^{-1}(\cdot|\phi)$ (the functional inverse of the normalized posterior matching kernel), controlled by the output sequence Φ^n, and initialized at $\hat{\theta}_{n+1}$. In practice however, the receiver decodes an interval and therefore to attain a specific target error probability $p_e(n)$, one can tentatively decode a subinterval of (0, 1) in which Θ_{n+1} lies with probability $1 - p_e(n)$, which since $\Theta_{n+1} \sim \mathcal{U}$, is any interval of length $1 - p_e(n)$. The endpoints of this interval are then "rolled back" via the RIFS to recover the decoded interval w.r.t. the message point Θ_0. The target error probability decay which facilitates the achievability of a given rate is determined by the convergence rate of the RIFS, which also corresponds to the maximal information rate supported by this analysis.

This general principle relating rate and error probability to the convergence properties of the corresponding RIFS, facilitates the use of any RIFS contraction condition for convergence. The only limitation stems from the fact that $\omega_\phi(\cdot)$ generating the RIFS is an inverse c.d.f. over the unit interval and hence never globally contractive, so only contraction on the average conditions can be used. The Theorems appearing below make use of the principle above in conjunction with the contraction Lemmas mentioned in Section II-D, to obtain two different expressions tying error probabilities, rate and transmission period. The discussion above is made precise in the course of the proofs.

Denote the family of all continuous functions $\rho : (0,1) \mapsto [1,\infty)$ by $\mathcal{C}$.

Theorem 5: Let $(P_X, P_{Y|X})$ be an input/channel pair, $(P_\Theta, P_{\Phi|\Theta})$ the corresponding normalized pair, and let $\omega_\phi(\cdot) \overset{\text{def}}{=} F_{\Theta|\Phi}^{-1}(\cdot|\phi)$. For any $\rho \in \mathcal{C}$, define

$$R^\dagger(\rho) \overset{\text{def}}{=} -\log \sup_{s \in (0,1)} \mathbb{E}\left\{ \frac{\rho(\omega_\Phi(s))}{\rho(s)} D_s(\omega_\Phi) \right\}$$

---

[19]Uniqueness of the capacity achieving distribution for $P_{Y|X}$ does not generally imply the same for the corresponding normalized channel $P_{\Phi|\Theta}$. For example, the normalized channel for a $\mathrm{BSC}/\mathrm{Bernoulli}(\frac{1}{2})$ pair, has an uncountably infinite number of capacity achieving distributions.

where $D_s(\cdot)$ is defined in (9), and let

$$R^\dagger \overset{\text{def}}{=} \sup_{\rho \in \mathcal{C}} R^\dagger(\rho).$$

If $R^\dagger > 0$, then the posterior matching scheme with an optimal variable rate decoding rule achieves any rate $R < R^\dagger$, by setting the target error probability to satisfy $p_e(n) \to 0$ under the constraint

$$\Psi\left((1-\alpha)p_e(n), 1 - \alpha p_e(n), 2^{-R^\dagger(\rho)}\right) = o\left(2^{n(R^\dagger(\rho)-R)}\right). \tag{53}$$

for some $\alpha \in (0,1)$, and some $\rho \in \mathcal{C}$ such that $R^\dagger(\rho) > R$, where Ψ is defined in Lemma 11.

Proof: Let $\tilde{S}_n(s)$ be the RIFS generated by $\omega_\phi(\cdot) \overset{\text{def}}{=} F_{\Theta|\Phi}^{-1}(\cdot|\phi)$ and the control sequence $\{\Phi_k\}_{k=1}^\infty$, initialized at $s \in (0,1)$. Select a fixed interval $J_1 = (s,t) \subseteq (0,1)$ as the decoded interval w.r.t. Θ_{n+1}. Since $\Theta_{n+1} \sim \mathcal{U}$, we have that

$$\mathbb{P}(\Theta_{n+1} \in J_1) = |J_1|.$$

Define the corresponding interval at the origin to be

$$J_n \overset{\text{def}}{=} (\tilde{S}_n(s), \tilde{S}_n(t))$$

and set it to be the decoded interval, i.e., $\Delta_n(\Phi^n) = J_n$. Note that the endpoints of J_n are r.v.'s. Since $F_{\Theta|\Phi}^{-1}(\cdot|\phi)$ is invertible for any ϕ, the interval J_n corresponds to Θ_1, namely

$$\begin{aligned} \mathbb{P}(\Theta_1 \in J_n) &= \mathbb{E}\mathbb{P}(\Theta_1 \in J_n \mid \Phi^n) \\ &= \mathbb{E}\mathbb{P}(\Theta_{n+1} \in J_1 \mid \Phi^n) \\ &= \mathbb{P}(\Theta_{n+1} \in J_1) = |J_1| \end{aligned}$$

and then in particular (recall that $\Theta_0 = \Theta_1$)

$$p_e(n) = \mathbb{P}(\Theta_0 \notin \Delta(\Phi^n)) = 1 - |J_1|.$$

For a variable rate decoding rule, the target error probability is set in advance. Therefore, given $p_e(n)$ the length of the interval J_1 is constrained to be $|J_1| = 1 - p_e(n)$, and so without loss of generality we can parameterize the endpoints of J_1 by

$$(s,t) = ((1-\alpha)p_e(n), 1 - \alpha p_e(n))$$

for some $\alpha \in (0,1)$.

Now let $\rho \in \mathcal{C}$, and define

$$r(\rho) \overset{\text{def}}{=} \sup_{s \in (0,1)} \mathbb{E}\left\{ \frac{\rho(\omega_\Phi(s))}{\rho(s)} D_s(\omega_\Phi) \right\}.$$

Note that the expectation above is taken w.r.t. $\Phi \sim \mathcal{U}$. Using Lemma 11, if $r(\rho) < 1$ then

$$\begin{aligned} \mathbb{P}(|J_1| > \varepsilon) &= \mathbb{P}\left(\left|\tilde{S}_n(s) - \tilde{S}_n(t)\right| > \varepsilon\right) \\ &\leq \varepsilon^{-1}\Psi(s,t,r(\rho)) \cdot r^n(\rho). \end{aligned}$$

To find the probability that the decoded interval is larger than 2^{-nR}, we substitute $\varepsilon = 2^{-nR}$ and obtain

$$\begin{aligned} \mathbb{P}(R_n < R) &= \mathbb{P}\left(|J_1| > 2^{-nR}\right) \\ &\leq 2^{nR} \cdot \Psi((1-\alpha)p_e(n), 1 - \alpha p_e(n), r(\rho)) \\ &\quad \cdot 2^{n \log r(\rho)}. \end{aligned}$$

Following the above and defining $R^\dagger(\rho) = -\log r(\rho)$, a sufficient condition for $\mathbb{P}(R_n < R) \to 0$ for $R < R^\dagger(\rho)$ is given by (53). The proof is concluded by taking the supremum over $\rho \in \mathcal{C}$, and noting that if no ρ results in a contraction, then $R^\dagger \le 0$.   ∎

Theorem 5 is very general in the sense of not imposing any constrains on the input/channel pair. It is however rather difficult to identify a *weight function* ρ that will result in $R^\dagger(\rho) > 0$. Our next error probability result is less general (e.g., does not apply to discrete alphabets), yet is much easier to work with. Although it also involves an optimization step over a set of functions, it is usually easier to find a function which results in a positive rate as the examples that follow demonstrate.

The basic idea is similar only now we essentially work with the original chain and so the RIFS evolves over $\underline{\mathrm{supp}}(X)$, generated by the kernel $\omega_y(\,\cdot\,) \stackrel{\text{def}}{=} F_{X|Y}^{-1}(\cdot\,|\,y) \circ F_X$ (the functional inverse of the posterior matching kernel), and controlled by the i.i.d. output sequence $\{Y_k\}_{k=1}^\infty$. To state the result, we need some definitions first. Let P_X be some input distribution and let $\rho : \underline{\mathrm{supp}}(X) \mapsto (\mathrm{a}, \mathrm{b})$ be differentiable and monotonically increasing (a, b may be infinite). The family of all such functions ρ for which $f_{\rho(X)}$ is bounded is denoted by $\mathcal{F}(X)$. Furthermore, for a proper r.v. X with a support over a (possibly infinite) interval, we define the *tail function* $\mathcal{T}_X : \mathbb{R}^+ \mapsto [0, 1]$ to be

$$\mathcal{T}_X(\ell) \stackrel{\text{def}}{=} 1 - \sup\{P_X((x, x+\ell)) : x \in \mathbb{R}\}.$$

Namely, $\mathcal{T}_X(\ell)$ is the minimal probability that can be assigned by P_X outside an interval of length ℓ.

Theorem 6: Let $(P_X, P_{Y|X})$ be an input/channel pair with f_{XY} continuous over $\underline{\mathrm{supp}}(X, Y)$, and let $\omega_y(\,\cdot\,) \stackrel{\text{def}}{=} F_{X|Y}^{-1}(\cdot\,|\,y) \circ F_X$. For any $\rho \in \mathcal{F}(X)$, define $R^*(\rho)$ as shown at the bottom of the page, and let

$$R^* \stackrel{\text{def}}{=} \sup_{\rho \in \mathcal{F}(X)} R^*(\rho).$$

The following statements hold:
(i) The posterior matching scheme with an optimal variable rate decoding rule achieves any rate $R < R^*$, by setting the target error probability to satisfy $p_e(n) \to 0$ under the constraint

$$p_e(n) = \mathcal{T}_{\rho(X)}\left(o\left(2^{n(R^*(\rho)-R)}\right)\right) \tag{54}$$

for some $\rho \in \mathcal{F}(X)$ satisfying $R^*(\rho) > R$.

(ii) If $|\mathrm{range}(\rho)| < \infty$ then any rate $R < R^*(\rho)$ can be achieved with zero error probability.[20]
(iii) If it is possible to write $\rho \circ \omega_y \circ \rho^{-1}(s) = u(s)v(y) + q(y)$, then

$$R^*(\rho) = -\mathbb{E}\log|v(Y)| - \log \sup_{s \in \mathrm{range}(\rho)} |u'(s)|$$

whenever the right-hand-side exists.

Proof: We first prove the three statements in the special case where P_X has a support over a (possibly infinite) interval, and considering only the identity function $\rho_1 : \underline{\mathrm{supp}}(X) \mapsto \underline{\mathrm{supp}}(X)$ over the support, i.e., discussing the achievability of $R^*(\rho_1)$ exclusively. We therefore implicitly assume here that $\rho_1 \in \mathcal{F}(X)$. Let $\tilde{S}_n(s)$ be the RIFS generated by $\omega_y(\,\cdot\,) \stackrel{\text{def}}{=} F_{X|Y}^{-1}(\cdot\,|\,y) \circ F_X$ and the control sequence $\{Y_k\}_{k=1}^\infty$, which evolves over the space $\underline{\mathrm{supp}}(X)$. Select a fixed interval $J_1 = (s, t) \subseteq \underline{\mathrm{supp}}(X)$ as the decoded interval w.r.t. X_{n+1}. Since $X_{n+1} \sim P_X$, we have that

$$\mathbb{P}(X_{n+1} \in J_1) = P_X(J_1)$$

Define the corresponding interval at the origin to be

$$J_n \stackrel{\text{def}}{=} (\tilde{S}_n(s), \tilde{S}_n(t))$$

and following the same lines as in the proof of the preceding Theorem, J_n is set to be the decoded interval w.r.t. $X_1 = F_X(\Theta_0)$, and so the decoded interval for Θ_0 is set to be $\Delta_n(Y^n) = F_X(J_n)$. Thus

$$\begin{aligned}p_e(n) &= \mathbb{P}(\Theta_0 \notin F_X(J_n)) \\ &= \mathbb{P}(X_1 \notin J_n) = 1 - P_X(J_1).\end{aligned}$$

.

For any $q > 0$ define

$$r_q \stackrel{\text{def}}{=} \sup_{s \ne t \in \underline{\mathrm{supp}}(X)} \mathbb{E}\left[D_{s,t}(\omega_Y)\right]^q.$$

Using Jensen's inequality, we have that for any $0 < q \le p$

$$\begin{aligned}r_q &= \sup_{s \ne t} \mathbb{E}\left[D_{s,t}(\omega_Y)\right]^q = \sup_{s \ne t} \mathbb{E}\left[D_{s,t}(\omega_Y)\right]^{p\frac{q}{p}} \\ &\le \sup_{s \ne t} \left(\mathbb{E}\left[D_{s,t}(\omega_Y)\right]^p\right)^{\frac{q}{p}} \\ &= \left(\sup_{s \ne t} \mathbb{E}\left[D_{s,t}(\omega_Y)\right]^p\right)^{\frac{q}{p}} = (r_p)^{\frac{q}{p}}. \tag{55}\end{aligned}$$

[20]This is not a standard zero-error achievability claim, since the rate is generally random. If a fixed rate must be guaranteed, then the error probability will be equal to the probability of "outage", i.e., the probability that the variable decoding rate falls below the rate threshold.

$$R^*(\rho) \stackrel{\text{def}}{=} \lim_{q \to 0^+} \inf_{\substack{s,t \in \mathrm{range}(\rho) \\ s \ne t}} \left(-q^{-1}\log \mathbb{E}\left[D_{s,t}(\rho \circ \omega_Y \circ \rho^{-1})\right]^q\right)$$

Now suppose there exists some $q^* > 0$ so that $r_{q^*} < 1$. Using (55), we conclude that $r_q < 1$ for any $0 < q \leq q^*$, and using Lemma 10 we have that for any $0 < q \leq q^*$ and any $\varepsilon > 0$

$$\mathbb{P}(|\tilde{S}_n(s) - \tilde{S}_n(t)| > \varepsilon) \leq \varepsilon^{-q}|s-t|^q r_q^n$$

and thus

$$\begin{aligned}
\mathbb{P}(R_n < R) &= \mathbb{P}\left(P_X((\tilde{S}_n(s), \tilde{S}_n(t))) > 2^{-nR}\right) \\
&\leq \mathbb{P}(M \cdot |J_n| > 2^{-nR}) \leq M^{-q} 2^{nRq} |J_1|^q r_q^n
\end{aligned}$$

where $M \overset{\text{def}}{=} \sup f_X(x)$. A sufficient condition for $\mathbb{P}(R_n < R) \to 0$ is given by

$$|J_1| = o\left(2^{n(q^{-1}\log r_q^{-1} - R)}\right).$$

Since the above depends only on the length of J_1, we can optimize over its position to obtain $p_e(n) = 1 - P_X(J_1) = \mathcal{T}_X(|J_1|)$, or arbitrarily close to that. Therefore, any rate $R < q^{-1}\log r_q^{-1}$ is achievable by setting $p_e(n) \to 0$ under the constraint

$$p_e(n) = \mathcal{T}_X(|J_1|) = \mathcal{T}_X\left(o\left(2^{n(q^{-1}\log r_q^{-1} - R)}\right)\right).$$

We would now like to maximize the term $q^{-1}\log r_q^{-1}$ over the choice of $0 < q \leq q^*$. Using (55) we obtain

$$q^{-1}\log (r_q)^{-1} \geq q^{-1}\log (r_p)^{-\frac{q}{p}} \geq p^{-1}\log (r_p)^{-1}$$

and so $q^{-1}\log r_q^{-1}$ is nonincreasing with q, thus

$$\sup_{0 < q \leq q^*} q^{-1}\log r_q^{-1} = \lim_{q \to 0^+} q^{-1}\log r_q^{-1} = R^*(\rho_1)$$

where ρ_1 is the identity function over $\underline{\text{supp}}(X)$. From the discussion above it is easily verified that $R^*(\rho_1) > 0$ iff $r_{q^*} < 1$ for some $q^* > 0$. Moreover, if $|\underline{\text{supp}}(X)| = M_0 < \infty$ then $\mathcal{T}_X(\ell) = 0$ for any $\ell > M_0$, therefore in this case $p_e(n) = 0$ for any n large enough. Note that since ρ_1 is defined only over $\underline{\text{supp}}(X)$, we have that $|\text{range}(\rho_1)| = |\underline{\text{supp}}(X)| = M_0$. Thus, statements (i) and (ii) are established for an input distribution with support over an interval, and the specific choice of the identity function $\rho = \rho_1$.

As for statement (iii), note first that since f_{XY} is continuous then $\omega_y(s)$ is jointly differentiable in y, s. Suppose that $\omega_y(s) = u(s)v(y) + q(y)$, and so u, v, q are all differentiable. In this separable case, we have

$$\begin{aligned}
r_q &= \sup_{s \neq t} \mathbb{E}\left(|v(Y)| \cdot \frac{|u(t) - u(s)|}{|t-s|}\right)^q \\
&= \mathbb{E}|v(Y)|^q \cdot \sup_{s \neq t}\left|\frac{u(t) - u(s)}{t-s}\right|^q \\
&= \mathbb{E}|v(Y)|^q \cdot \sup_s |u'(s)|^q
\end{aligned}$$

and

$$\begin{aligned}
q^{-1}\log r_q^{-1} &= -q^{-1}\log \mathbb{E}|v(Y)|^q - \sup_s \log |u'(s)| \\
&\leq -\mathbb{E}\log |v(Y)| - \sup_s \log |u'(s)|
\end{aligned}$$

where we have used Jensen's inequality in the last transition. We now show that the limit of the left-hand-side above as $q \to 0^+$ in fact attains the right-hand-side bound (assuming it exists), which is similar to the derivation of the Shannon entropy as a limit of Rényi entropies. Since $\mathbb{E}\log |v(Y)|$ is assumed to exist, then we have $\log \mathbb{E}|v(Y)|^q \to 0$ as $q \to 0^+$, and so to take the limit we need to use L'Hospital's rule. To that end, for any $0 < q \leq q^*$

$$\begin{aligned}
\frac{d}{dq}\mathbb{E}|v(Y)|^q &= \frac{d}{dq}\int f_Y(y)|v(Y)|^q \, dy \\
&= \int \frac{\partial}{\partial q} f_Y(y)|v(Y)|^q \, dy \\
&= \log e \cdot \int f_Y(y)|v(Y)|^q \log |v(Y)| \, dy \\
&= \log e \cdot \mathbb{E}\left(|v(Y)|^q \log |v(Y)|\right)
\end{aligned}$$

and thus

$$\begin{aligned}
&R^*(\rho_1) \\
&= \lim_{q \to 0^+}\left(-q^{-1}\log \mathbb{E}|v(Y)|^q - \sup_s \log |u'(s)|\right) \\
&= \lim_{q \to 0^+}\left(-\frac{d}{dq}\log \mathbb{E}|v(Y)|^q\right) - \sup_s \log |u'(s)| \\
&= \lim_{q \to 0^+}\left(-\frac{\mathbb{E}(|v(Y)|^q \log |v(Y)|)}{\mathbb{E}|v(Y)|^q}\right) - \sup_s \log |u'(s)| \\
&= -\mathbb{E}\log |v(Y)| - \sup_s \log |u'(s)|.
\end{aligned}$$

Which established statement (iii) in the special case under discussion. The derivations above all hold under the assumption that the right-hand-side above exists.

Treating the general case is now a simple extension. Consider a general input distribution P_X (with a p.d.f. continuous over its support), and a differentiable and monotonically increasing function $\rho : \underline{\text{supp}}(X) \mapsto (a, b)$. Let us define a ρ-normalized channel $P_{Y^\rho \mid X^\rho}$ by connecting the operator $\rho^{-1}(\cdot)$ to the channel's input. Let us Consider the posterior matching scheme for the ρ-normalized input/channel pair $(P_{\rho(X)}, P_{Y^\rho \mid X^\rho})$. Using the monotonicity of ρ, the corresponding input and inverse channel c.d.f.'s are given by

$$F_{\rho(X)} = F_X \circ \rho^{-1}, \quad F_{X^\rho \mid Y^\rho}(\cdot \mid y) = F_{X \mid Y}(\cdot \mid y) \circ \rho^{-1}$$

The posterior matching kernel is therefore given by

$$F_{\rho(X)}^{-1} \circ F_{X^\rho \mid Y^\rho}(\cdot \mid y) = \rho \circ \left(F_X^{-1} \circ F_{X \mid Y}(\cdot \mid y)\right) \circ \rho^{-1}$$

and the corresponding RIFS kernel is the functional inverse of the above, i.e.,

$$\begin{aligned}
&\left(F_{\rho(X)}^{-1} \circ F_{X^\rho \mid Y^\rho}(\cdot \mid y)\right)^{-1} \\
&= \rho \circ \left(F_{X \mid Y}^{-1}(\cdot \mid y) \circ F_X\right) \circ \rho^{-1} \\
&= \rho \circ \omega_y \circ \rho^{-1}.
\end{aligned}$$

Now, using the monotonicity of ρ it is readily verified that the input/channel pairs $(P_X, P_{Y \mid X})$ and $(P_{\rho(X)}, P_{Y^\rho \mid X^\rho})$

correspond to the same normalized channel. Hence, the corresponding posterior matching schemes are equivalent, in the sense that $\{(X_k, Y_k)\}_{k=1}^{\infty}$ and $\{(\rho^{-1}(X_k^{\rho}), Y_k^{\rho})\}_{k=1}^{\infty}$ have the same joint distribution. Therefore, the preceding analysis holds for the input/channel pair $(P_{\rho(X)}, P_{Y^{\rho}|X^{\rho}})$, and the result follows immediately. ∎

Loosely speaking, the optimization step in both Theorems has a similar task—changing the scale by which distances are measured so that the RIFS kernel appears contractive. In Theorem 5, the weight functions multiply the local slope of the RIFS. In Theorem 6 the approach is in a sense complementing, since the functions are applied to the RIFS kernel itself, thereby shaping the slopes directly. These functions will therefore be referred to as *shaping functions*.

Example 4 (AWGN, Continued): Returning to the AWGN channel setting with a Gaussian input, we can now determine the tradeoff between rate, error probability and transmission period obtained by the Schalkwijk-Kailath scheme. Inverting the kernel (21) we obtain the RIFS kernel

$$\omega_y(s) = F_{X|Y}^{-1}(F_X(s)|y) = \frac{s}{\sqrt{1+\text{SNR}}} + \frac{\text{SNR}}{1+\text{SNR}}y$$

Setting the identity shaping function $\rho_1(s) = s$, the condition of Theorem 6 statement (iii) holds and so

$$R^*(\rho_1) = -\log\sup_{s \in \mathbb{R}}\left|\frac{d}{ds}\left(\frac{s}{\sqrt{1+\text{SNR}}}\right)\right|$$
$$= \frac{1}{2}\log(1+\text{SNR})$$
$$= C$$

so in this case $R^* = C$, and statement (i) reconfirms that the Schalkwijk-Kailath scheme achieves capacity. Using standard bounds for the Gaussian distribution, the Gaussian tail function (for the input distribution) satisfies

$$\mathcal{T}_X(\ell) = O\left(e^{-\frac{\ell^2}{8P}}\right).$$

Plugging the above into (54), we find that a rate $R < C$ is achievable by setting the target error probability to

$$-\log p_e(n) = -\log \mathcal{T}_X\left(o\left(2^{n(R^*-R)}\right)\right) = o\left(2^{2n(C-R)}\right)$$

recovering the well known double-exponential behavior. Note that since the interval contraction factor in this case is independent of the output sequence, the variable-rate decoding rule is in fact fixed-rate, hence the same double-exponential performance is obtained using a fixed-rate decoding rule.

We mention here the well known fact that for the AWGN channel, the error probability can be made to decay as a higher order exponential in the block length, via adaptations of the Schalkwijk-Kailath scheme [21], [22]. These adaptations exploit the discreteness of the message set especially at the last stages of transmission, and are not directly applicable within our framework, since we define error probability in terms of intervals and not discrete messages. They can only be applied to the equivalent standard scheme obtained via Lemma 3.

Example 5 (BSC, Continued): The conditions of Theorem 6 are not satisfied in the BSC setting, and we resort to Theorem 5. Inverting the posterior matching kernel (30) pertaining to the corresponding normalized channel, we obtain the RIFS kernel

$$\omega_{\phi}(s) = F_{\Theta|\Phi}^{-1}(s|\phi)$$
$$= \begin{cases} \frac{s}{2(1-p)} & s \in (0, 1-p), \phi \in \left(0, \frac{1}{2}\right) \\ \frac{s-(1-2p)}{2p} & s \in [1-p, 1), \phi \in \left(0, \frac{1}{2}\right) \\ \frac{s}{2p} & s \in (0, p), \phi \in \left[\frac{1}{2}, 1\right) \\ \frac{s+(1-2p)}{2(1-p)} & s \in [p, 1), \phi \in \left[\frac{1}{2}, 1\right) \end{cases}$$

and
$$D_s(\omega_{\phi}) = \begin{cases} \frac{1}{2(1-p)} & s \in (0, 1-p), \phi \in \left(0, \frac{1}{2}\right) \\ \frac{1}{2p} & s \in [1-p, 1), \phi \in \left(0, \frac{1}{2}\right) \\ \frac{1}{2p} & s \in (0, p), \phi \in \left[\frac{1}{2}, 1\right) \\ \frac{1}{2(1-p)} & s \in [p, 1), \phi \in \left[\frac{1}{2}, 1\right) \end{cases}.$$

Using a constant weight function (i.e., no weights) does not work in this case, since the average of slopes for (say) $s \in (0, p)$, is

$$\mathbb{E}D_s(\omega_{\Phi}) = \frac{1}{2}\left(\frac{1}{2p} + \frac{1}{2(1-p)}\right) \geq 1.$$

In fact, any bounded weight function will result in the same problem for $s > 0$ small enough, which suggests that the weight function should diverge to infinity as $s \to 0$. Setting $\rho(s) = s^{-\beta}$ for $\beta > 1$ is a good choice for $s \in (0, p)$ since in that case

$$\mathbb{E}\left(\frac{\rho(\omega_{\Phi}(s))}{\rho(s)} \cdot D_s(\omega_{\Phi})\right) = \frac{1}{2}\left((2p)^{\beta-1} + (2(1-p))^{\beta-1}\right)$$

which can be made smaller than unity by properly choosing β. Setting ρ symmetric around $\frac{1}{2}$ duplicates the above to $s \in (1-p, 1)$. However, this choice (and some variants) do not seem to work in the range $s \in (p, \frac{1}{2})$, for which $\beta \leq 1$ is required. Finding a weight function ρ for which $R^{\dagger}(\rho) > 0$ (if exists at all) seems to be a difficult task, which we were unable to accomplish thus far.

Example 6 (Uniform Input/Noise, Continued): We have already seen that achieving the mutual information with zero error decoding is possible in the uniform noise/input setting. Let us now derive this fact via Theorem 6. The output p.d.f. is given by

$$f_Y(y) = y\mathbb{1}_{(0,1]}(y) + (2-y)\mathbb{1}_{(1,2)}(y).$$

The RIFS kernel is obtained by inverting the posterior matching kernel (23), which yields

$$\omega_y(s) = F_{X|Y}^{-1}(F_X(s)|y)$$
$$= s\left(y\mathbb{1}_{(0,1]}(y) + (2-y)\mathbb{1}_{(1,2)}(y)\right)$$
$$\quad + (y-1)\mathbb{1}_{(1,2)}(y)$$
$$= sf_Y(y) + (y-1)\mathbb{1}_{(1,2)}(y).$$

Using the identity shaping function ρ_1 again (but now restricted to $\text{supp}(X) = (0, 1)$), the condition of statement (iii) holds and therefore

$$R^*(\rho_1) = -\mathbb{E}\log f_Y(Y) - \sup_{s \in (0,1)}\log 1$$
$$= h(Y) = I(X; Y)$$

and we have $R^* = R^*(\rho_1) = I(X;Y)$, thereby verifying once again that the mutual information is achievable. Since $\mathrm{range}(\rho_1) = (0,1)$ is bounded, statement (ii) reconfirms that variable-rate zero error decoding is possible.

Example 7 (Exponential Input/Noise, Continued): Let us return to the additive noise channel with an exponentially distributed noise and input. We have already seen that the posterior matching scheme (26) achieves the mutual information, which in this case is $I(X;Y) \approx 0.8327$. The p.d.f. of the corresponding output is

$$f_Y(y) = ye^{-y}\mathbf{1}_{(0,\infty)}(y)$$

It is easily verified that $F_{X|Y}^{-1}(s\,|\,y) = sy$, and so the RIFS kernel is given by

$$\omega_y(s) = F_{X|Y}^{-1}(F_X(s)\,|\,y) = y(1 - e^{-s})$$

Now, using Theorem 6 with the identity shaping function ρ_1 restricted to $\underline{\mathrm{supp}}(X) = (0,\infty)$, the condition of statement (iii) holds and therefore

$$
\begin{aligned}
R^*(\rho_1) &= -\mathbb{E}\log Y - \log \sup_{s\in(0,\infty)} \left| \frac{d}{ds}\left(1 - e^{-s}\right) \right| \\
&= -\mathbb{E}\log Y \\
&\approx -0.61 < 0.
\end{aligned}
$$

Thus, the identity function is not a good choice in this case, and we must look for a different shaping function. Let us set $\rho_2(s) = s^{-\frac{1}{2}}$, which results in a p.d.f. and c.d.f.

$$
\begin{aligned}
f_{\rho_2(X)}(s) &= \frac{1}{2s^3}\exp\left(-s^{-2}\right) \\
F_{\rho_2(X)}(s) &= \exp\left(-s^{-2}\right)
\end{aligned}
$$

and

$$\rho_2 \circ \omega_y \circ \rho_2^{-1}(s) = \left[y\left(1 - \exp\left(-s^{-2}\right)\right)\right]^{-\frac{1}{2}}.$$

Since $f_{\rho_2(X)}$ is bounded and the above again satisfies the condition of statement (iii), we obtain

$$
\begin{aligned}
R^*(\rho_2) &= \frac{1}{2}\mathbb{E}\log Y \\
&\quad - \log \sup_{s\in(0,\infty)} \left| \frac{d}{ds}\left[\left(1 - \exp\left(-s^{-2}\right)\right)\right]^{-\frac{1}{2}} \right| \\
&= \frac{1}{2}\mathbb{E}\log Y + \inf_{s\in[0,\infty)}\left(\frac{\log e}{s^2} + 3\log s \right. \\
&\qquad \left. + \frac{3}{2}\log\left(1 - \exp(-s^{-2})\right)\right) \\
&= \frac{1}{2}\mathbb{E}\log Y \approx 0.305
\end{aligned}
$$

where the infimum above is attained as $s \to \infty$. The tail function of $P_{\rho_2(X)}$ is bounded by

$$\mathcal{T}_{\rho_2(X)}(\ell) \leq 1 - \exp\left(-\ell^{-2}\right) \leq \ell^{-2}.$$

Thus, any rate $R < R^*(\rho_2) \approx 0.305$ is achieved by the posterior matching scheme (26) using a variable decoding rule if the target error probability is set to

$$p_e(n) = \frac{1}{o\left(2^{2n(R^*(\rho_2)-R)}\right)}$$

and so the following error exponent is achievable:

$$\lim_{n\to\infty} \frac{1}{n}\log\frac{1}{p_e(n)} = 2(R^*(\rho_2) - R) \approx 0.61 - 2R.$$

Although we know from Theorem 4 that any rate up to the mutual information is achieved in this case, $\rho_2(\cdot)$ is the best shaping function we have found, and so our error analysis is valid only up to the rate $R^*(\rho_2) \approx 0.305 < I(X;Y)$.

VII. Extensions

A. The μ-Variants of the Posterior Matching Scheme

In this subsection, we return to discuss the μ-variants (14) of the baseline posterior matching scheme addressed thus far. To understand why these variants are of interest, let us first establish the necessity of a fixed-point free kernel (thereby also proving Lemma 14).

Lemma 21: If (A3) does not hold, then (A2) does not hold either and the corresponding scheme cannot achieve any positive rate.

Proof: By the assumption in the Lemma, there must exists some fixed-point $\theta_f \in (0,1)$ such that

$$\mathbb{P}\left(F_{\Theta|\Phi}(\theta_f|\Phi) = \theta_f\right) = 1.$$

The posterior c.d.f. $F_{\Theta_0|\Phi^n}(\theta|\phi^n)$ is obtained by an iterated composition of the kernel $P_{\Theta|\Phi}(\theta|\phi)$ controlled by the i.i.d. output sequence Φ^n. Thus, the fixed point at θ_f induces a fixed point for the posterior c.d.f at θ_f as well, since

$$
\begin{aligned}
\mathbb{P}\left(F_{\Theta_0|\Phi^n}(\theta_f|\Phi^n) = \theta_f\right) &\geq \prod_{k=1}^{n}\mathbb{P}\left(F_{\Theta_0|\Phi}(\theta_f|\Phi_k) = \theta_f\right) \\
&= 1.
\end{aligned}
$$

This immediately implies that no positive rate can be achieved, since the posterior probability of the interval $(0,\theta_f)$ remains fixed at θ_f. Stated differently, this simply means that the output sequence provides no information regarding whether $\Theta_0 < \theta_f$ or not. For practically the same reason, the invariant distribution $P_{\Theta\Phi}$ for the Markov chain $\{(\Theta_n, \Phi_n)\}_{n=1}^{\infty}$ is not ergodic, since the set $(0,\theta_f) \times (0,1)$ is invariant yet $0 < P_{\Theta\Phi}((0,\theta_f) \times (0,1)) = \theta_f < 1$. ∎

Suppose our kernel has L fixed points, and so following the above the unit interval can be partitioned into a total of $L+1$ corresponding *invariant intervals*. One exogenous way to try and handle the fixed-point problem is to decode a disjoint union of $L+1$ exponentially small intervals (one per invariant interval) in which the message point lies with high probability, and then resolve the remaining ambiguity using some simple nonfeedback zero-rate code. This seems reasonable, yet there are two caveats.

First, the maximal achievable rate in an invariant interval may generally be smaller than the mutual information, incurring a penalty in rate. Second, the invariant distribution $P_{\Theta\Phi}$ is not ergodic, and it is likely that any encapsulated input constraints will not be satisfied (i.e., not pathwise but only in expectation over invariant intervals). A better idea is to map our message into the invariant interval with the maximal achievable rate, which is always at least as high as the mutual information. This corresponds to a posterior matching scheme with a different input distribution (using only some of the inputs), and resolves the rate problem, but not the input constraint problem. We must therefore look for a different type of solution.

Fortunately, it turns out that the fixed points phenomena is in many cases just an artifact of the specific ordering imposed on the inputs, induced by the selection of the posterior c.d.f. in the posterior matching rule. In many cases, imposing a different ordering can eliminate this artifact altogether. We have already encountered that in the DMC setting (Example 9 in Section V, using Lemma 16), where in the case a fixed point exists, a simple input permutation was shown to be sufficient in order for the posterior matching scheme (matched to the equivalent input/channel pair) to achieve capacity. This permutation can be interpreted as inducing a different order over the inputs, and the scheme for the equivalent pair can be interpreted as a specific μ-variant of the original scheme.

These observations provide motivation to extend the notion of equivalence between input/channel pairs from the discrete case to the general case. Two input/channel pairs $(P_X, P_{Y|X})$ and $(P_{X^*}, P_{Y^*|X^*})$ are said to be *equivalent* if there exist u.p.f's μ, σ such that the corresponding normalized channels satisfy

$$P_{\Phi^*|\Theta^*}(\cdot|\theta) = P_{\Phi|\Theta}(\sigma(\cdot)|\mu(\theta))$$

for any $\theta \in (0,1)$. This practically means that the asterisked normalized channel is obtained by applying μ and σ^{-1} to the input and output of the asterisk-free normalized channel, respectively, and in this case we also say that the pair $(P_X, P_{Y|X})$ is μ-*related* to the pair $(P_{X^*}, P_{Y^*|X^*})$. Again, equivalent input/channel pairs have the same mutual information. Following this, for every u.p.f. μ and every set of input/channel pairs Γ, we define $\mu(\Gamma)$ to be the set of all input/channel pairs to which some pair in Γ is μ-related. The following result follows through immediately from the developments in Sections III and V, and the discussion above.

Theorem 7: For any input/channel pair $(P_X, P_{Y|X})$ and any u.p.f. μ, the corresponding μ-variant posterior matching scheme (14) has the following properties:
  (i) It admits a recursive representation w.r.t. the normalized channel, with a kernel $\mu \circ F_{\mu^{-1}(\Theta)|\Phi}(\cdot|\phi) \circ \mu^{-1}$, i.e.,

$$\Theta_1 = \mu(\Theta_0), \quad \Theta_{n+1} = \mu \circ F_{\mu^{-1}(\Theta)|\Phi}(\cdot|\Phi_n) \circ \mu^{-1}(\Theta_n)$$

  (ii) If $(P_X, P_{Y|X}) \in \mu(\Omega_A \cup \Omega_B)$ (resp. $\mu(\Omega_C)$), the scheme achieves (resp. pointwise achieves) any rate $R < I(X;Y)$ over the channel $P_{Y|X}$. Furthermore, if $(P_X, P_{Y|X}) \in \mu(\Omega_A \cup \Omega_C)$ then this is achieved within an input constraint $(\eta, \mathbb{E}\eta(X))$, for any measurable $\eta : \mathcal{X} \mapsto \mathbb{R}$ satisfying $\mathbb{E}|\eta(X)| < \infty$.

Theorem 7 expands the set of input/channel pairs for which some variant of the posterior matching scheme achieves the mutual information, by allowing different orderings of the inputs to eliminate the fixed point phenomena. For the DMC case, we have already seen that considering μ-variants is sometimes crucial for achieving capacity. Next we describe perhaps a more lucid (although very synthetic) example, making the same point for continuous alphabets.

Example 11: Let the memoryless channel $P_{Y|X}$ be defined by the following input to output relation:

$$Y = X^2 + Z$$

where the noise Z is statistically independent of the input X. Suppose that some input constraints are imposed so that the capacity is finite, and also such that the capacity achieving distribution does not have a mass point at zero. Now assume that an input zero mean constraint is additionally imposed. It is easy to see that the capacity achieving distribution P_X is now symmetric around zero, i.e., $P_X((-\infty, 0)) = P_X((0, \infty)) = \frac{1}{2}$. It is immediately clear that the output of the channel provides no information regarding the sign of the input, hence the corresponding posterior matching kernel $F_X^{-1} \circ F_{X|Y}(\cdot|y)$ has a fixed point at the origin, and equivalently, the normalized kernel $F_{\Theta|\Phi}(\cdot|\phi)$ has a fixed point at $\theta = \frac{1}{2}$. Thus, by Lemma 21 the scheme cannot attain any positive rate. Intuitively, this stems from the fact that information has been coded in the sign of the input, or the most-significant-bit of the message point, which cannot be recovered. To circumvent this problem we can change the ordering of the input, which is effectively achieved by using one of the μ-variants of the posterior matching scheme. For example, set

$$\mu(\theta) = \begin{cases} \theta + \frac{1}{3} & \theta \in \left(0, \frac{1}{3}\right] \\ \theta - \frac{1}{3} & \theta \in \left(\frac{1}{3}, \frac{2}{3}\right] \\ \theta & \theta \in \left(\frac{2}{3}, 1\right) \end{cases}$$

and use the corresponding μ-variant scheme. This maintains the same input distribution while breaking the symmetry around $\frac{1}{2}$, and eliminating the fixed point phenomena. This μ-variant scheme can therefore achieve the mutual information, assuming all the other conditions are satisfied.

B. Channel Model Mismatch

In this subsection, we discuss the model mismatch case, where the scheme is designed according to the wrong channel model. We assume that the transmitter and receiver are both unaware of the situation, or at least do not take advantage of it. To that end, for any pair $(P_X, P_{Y|X}) \in \Omega_C$, we define a *mismatch set* $\Omega_C^{\mathrm{mis}}(P_X, P_{Y|X})$ consisting of all input/channel pairs $(P_{X^*}, P_{Y^*|X^*})$, with a corresponding normalized channel $P_{\Phi^*|\Theta^*}$, that admit the following properties:
(C1) $(P_{X^*}, P_{Y^*|X^*})$ *satisfies* (A5), *and*

$$\inf_{\varepsilon > 0} \Big[D(P_{\Phi^*|\Theta^*} \| ^- P_{\Phi|\Theta}^\varepsilon \mid P_\Theta)$$

$$+ D(P_{\Phi^*|\Theta^*} \| ^+ P_{\Phi|\Theta}^\varepsilon \mid P_\Theta) \Big] < \infty.$$

(C2) $D(P_{Y^*|X^*} \| P_{Y|X} \mid P_{X^*}) < \infty, D(P_{Y^*} \| P_Y) < \infty.$
(C3) $F_X^{-1}(F_{X|Y}(X^*|Y^*)) \sim P_{X^*}.$

(C4) *Let* $\{Y_n^*\}_{n=1}^{\infty}$ *be the channel output sequence when the posterior matching scheme for* $(P_X, P_{Y\,|\,X})$ *is used over* $P_{Y^*\,|\,X^*}$ *and initialized with* $X_1 \sim P_{X^*}$. *There is a contraction* ξ *and a length function* ψ_λ *over* $\mathfrak{F}_c$, *such that for every* $h \in \mathfrak{F}_c$ *and* $n \in \mathbb{N}$,

$$\sup_{y^{n-1}} \mathbb{E}\left(\psi_\lambda\left[F_{X\,|\,Y}(\cdot\,|\,Y_n^*) \circ F_X^{-1} \circ h\right] \,|\, Y^{*\,n-1} = y^{n-1}\right)$$
$$\leq \xi(\psi_\lambda(h)).$$

(C5) *Let* $Z = F_X^{-1}(F_{X\,|\,Y}(X^*|Y^*))$. *For any* $x^* \in \underline{\mathrm{supp}}(\mathrm{X}^*)$ *the set* $\underline{\mathrm{supp}}(\mathrm{Z}|\mathrm{X}^* = \mathrm{x}^*)$ *contains some open neighborhood of* x^*.

The properties above are not too difficult to verify, with the notable exception of the contraction condition (*C4*) which is not "single letter". This stems from the fact that the output distribution under mismatch is generally not i.i.d. Clearly, for any $(P_X, P_{Y\,|\,X}) \in \Omega_C$, we have $(P_X, P_{Y\,|\,X}) \in \Omega_C^{\mathrm{mis}}(P_X, P_{Y\,|\,X})$ in particular. Moreover, if the posterior matching kernels for the pairs $(P_X, P_{Y\,|\,X})$ and $(P_{X^*}, P_{Y^*\,|\,X^*})$ happen to coincide, then we trivially have $(P_{X^*}, P_{Y^*\,|\,X^*}) \in \Omega_C^{\mathrm{mis}}(P_X, P_{Y\,|\,X})$ and any rate $R < I(X^*; Y^*) = I(X; Y)$ is pointwise achievable, hence there is no rate loss due to mismatch (although satisfaction of input constraints may be affected, see below). Note that the initialization step (i.e., transforming the message point into the first channel input) is in general different even when the kernels coincide. Nevertheless, identical kernels imply a common input support and so using a different initialization amounts to a one-to-one transformation of the message point, which poses no problem due to pointwise achievability.

The channel model mismatch may incur a rate loss in general, as quantified in the following Theorem.

Theorem 8 (Mismatch Achievability): Let $(P_X, P_{Y\,|\,X}) \in \Omega_C$, and suppose the corresponding posterior matching scheme (16) is used over a channel $P_{Y^*\,|\,X^*}$ (unknown on both terminals). If there exists an input distribution P_{X^*} such that $(P_{X^*}, P_{Y^*\,|\,X^*}) \in \Omega_C^{\mathrm{mis}}(P_X, P_{Y\,|\,X})$, then P_{X^*} is unique and the mismatched scheme with a fixed/variable rate optimal decoding rule matched to $(P_X, P_{Y\,|\,X})$, pointwise achieves any rate

$$R < I(X^*; Y^*)$$
$$- \left(D(P_{Y^*\,|\,X^*}\|P_{Y\,|\,X}\,|\,P_{X^*}) - D(P_{Y^*}\|P_Y)\right) \quad (56)$$

within an input constraint $(\eta, \mathbb{E}\eta(X^*))$, provided that $\mathbb{E}|\eta(X^*)| < \infty$.

Proof: See Appendix B. ∎

The difference between relative entropies in (56) constitutes the penalty in rate due to the mismatch, relative to what could have been achieved for the induced input distribution P_{X^*}. Note that this term is always nonnegative[21], and vanishes when there is no mismatch.

For the next example, we need the following Lemma. The proof (by direct calculation) is left out.

---

[21]The non-negativity of the penalty term is not immediately clear from the expressions. However it holds since (by the converse to the channel coding Theorem) no rate above $I(X^*; Y^*)$ is achievable with an input marginal P_{X^*}, which is the marginal induced by our scheme (see the proof of the theorem).

Lemma 22: Let U, V be a pair of continuous, zero mean, finite variance r.v.'s, and suppose V is Gaussian. Then

$$D(P_U\|P_V) = h(V) - h(U) + \frac{\log e}{2}\left(\frac{\mathbb{E}\,U^2}{\mathbb{E}\,V^2} - 1\right).$$

Example 12 (Robustness of the Schalkwijk-Kailath Scheme): Suppose that the Schalkwijk-Kailath scheme (22) designed for an AWGN channel $P_{Y\,|\,X}$ with noise $Z \sim \mathcal{N}(0, \mathrm{N})$ and input $X \sim \mathcal{N}(0, \mathrm{P})$, is used over an AWGN channel with noise variance N^*. Since the scheme depends on the channel and input only through the $\mathrm{SNR} = \frac{\mathrm{P}}{\mathrm{N}}$, then the scheme's kernel coincides with the Schalkwijk-Kailath kernel for an input $X^* \sim \mathcal{N}(0, \frac{\mathrm{N}^*}{\mathrm{N}}\mathrm{P})$ over the mismatch channel. Therefore, following the remark preceding Theorem 8, there is no rate loss, and the input power is automatically scaled to maintain the same SNR for which the scheme was designed. This robustness of the Schalkwijk-Kailath scheme to changes in the Gaussian noise (SNR mismatch) was already mentioned [23].

However, Theorem 8 can be used to demonstrate how the Schalkwijk-Kailath scheme is robust to more general perturbations in the noise statistics. Suppose the scheme is used over a *generally non-Gaussian* additive noise channel $P_{Y^*\,|\,X^*}$ with noise Z^* having zero mean and a variance N^*. Suppose there exists an input distribution P_{X^*} such that $(P_{X^*}, P_{Y^*\,|\,X^*}) \in \Omega_C^{\mathrm{mis}}(P_X, P_{Y\,|\,X})$. We have $Y = X + Z$ and $Y^* = X^* + Z^*$ for the original channel and the mismatch channel respectively. Plugging (22) into the invariance property (*C3*) and looking at the variance, we have

$$\mathrm{P}^* = \mathbb{E}\left(\frac{X^*}{\sqrt{1+\mathrm{SNR}}} + \frac{\mathrm{SNR}\cdot Z^*}{\sqrt{1+\mathrm{SNR}}}\right)^2$$
$$= \frac{\mathrm{P}^* + \mathrm{SNR}^2 \cdot \mathrm{N}^*}{1+\mathrm{SNR}}$$

which immediately results in $\mathrm{SNR}^* \stackrel{\mathrm{def}}{=} \frac{\mathrm{P}^*}{\mathrm{N}^*} = \mathrm{SNR}$, so the SNR is conserved despite the mismatch. Now applying Theorem 8 and some simple manipulations, we find that the mismatched scheme pointwise achieves any rate R satisfying

$$R < h(Y^*) - h(Z^*) - (D(P_{Z^*}\|P_Z) - D(P_{Y^*}\|P_Y))$$
$$= h(Y^*) - h(Z^*) - \Big(h(Z) - h(Z^*) - h(Y) + h(Y^*)$$
$$+ \frac{\log e}{2}\left(\frac{\mathbb{E}(Z^*)^2}{\mathbb{E}Z^2} - \frac{\mathbb{E}(Y^*)^2}{\mathbb{E}Y^2}\right)\Big)$$
$$= h(Y) - h(Z) + \frac{\log e}{2}\left(\frac{\mathrm{P}^* + \mathrm{N}^*}{P+N} - \frac{\mathrm{N}^*}{N}\right)$$
$$= I(X;Y) + \frac{\log e}{2}\cdot\frac{\mathrm{N}^*}{N}\left(\frac{1+\mathrm{SNR}^*}{1+\mathrm{SNR}} - 1\right)$$
$$= I(X;Y) = \frac{1}{2}\log(1+\mathrm{SNR})$$

where we have used Lemma 22 in the first equality. Therefore, the mismatched scheme can attain any rate below the Gaussian capacity it was designed for, despite the fact that the noise is not Gaussian, and the input power is automatically scaled to maintain the same SNR for which the scheme was designed. Invoking [24], we can now claim that the Schalkwijk-Kailath scheme is

universal for communication over a memoryless additive noise channel (within the mismatch set) with a given variance and an input power constraint, in the sense of loosing at most half a bit in rate w.r.t. the channel capacity.

VIII. DISCUSSION

An explicit feedback transmission scheme tailored to any memoryless channel and any input distribution was developed, based on a novel principle of posterior matching. In particular, this scheme was shown to provide a unified view of the well known Horstein and Schalkwijk-Kailath schemes. The core of the transmission strategy lies in the constantly refined representation of the message point's position *relative* to the uncertainty at the receiver. This is accomplished by evaluating the receiver's posterior c.d.f. at the message point, followed by a technical step of matching this quantity to the channel via an appropriate transformation. A recursive representation of the scheme renders it very simple to implement, as the next channel input is a fixed function of the previous input/output pair only. This function is explicitly given in terms of the channel and the selected input distribution. The posterior matching scheme was shown to achieve the mutual information for pairs of channels and input distributions under very general conditions. This was obtained by proving a concentration result of the posterior p.d.f. around the message point, in conjunction with a contraction result for the posterior c.d.f. over a suitable function space. In particular, achievability was established for discrete memoryless channels, thereby also proving that the Horstein scheme is capacity achieving.

The error probability performance of the scheme was analyzed, by casting the variable-rate decoding process as the evolution of a reversed iterated function system (RIFS), and interpreting the associated contraction factors as information rates. This approach yielded two closed form expressions for the exponential decay of the target error probability which facilitates the achievability of a given rate, then used to provide explicit results in several examples. However, the presented error analysis is preliminary and should be further pursued. First, the obtained expressions require searching for good weight or shaping functions, which in many cases may be a difficult task. In the same vein, it is yet unclear under what conditions the error analysis becomes valid for rates up to the mutual information. Finally, the basic technique is quite general and allows for other RIFS contraction lemmas to be plugged in, possibly to yield improved error expressions.

We have seen that a fixed-point free kernel is a necessary condition for achieving any positive rate. We have also demonstrated how fixed points can sometimes be eliminated by considering an equivalent channel, or a corresponding μ-variant scheme. But can this binary observation be refined? From the error probability analysis of Section VI, it roughly seems that the "closer" the kernel is to having a fixed point, the worst the error performance should be. It would be interesting to quantify this observation, and to characterize the best μ-variant scheme for a given input/channel pair, in terms of minimizing the error probability.

We have derived the rate penalty incurred in a channel model mismatch setting, where a posterior matching scheme devised according to one channel model (and input distribution) is used over a different channel. However, the presence of feedback allows for an adaptive transmission scheme to be used in order to possibly reduce or even eliminate this penalty. When the channel is known to belong to some parametric family, there exist universal feedback transmission schemes that can achieve the capacity of the realized channel if the family is not too rich [25], and sometimes even attain the optimal error exponent [26]. However, these results involve random coding arguments, and so the associated schemes are neither explicit nor simple. It would therefore be interesting to examine whether an *adaptive* posterior matching scheme, in which the transmitter modifies its strategy online based on channel estimation, can be proven universal for some families of memoryless channels. It seems plausible that if the family is not too rich (e.g., in the sense of [27]) then the posterior will have a significant peak only when "close enough" to the true channel, and be flat otherwise. Another related avenue of future research is the universal communication problem in an individual/adversarial setting with feedback. For discrete alphabets, it was already demonstrated that the *empirical capacity* relative to a modulo-additive memoryless model can be achieved using a randomized sequential transmission strategy that builds on the Horstein scheme [28]. It remains to be explored whether this result can be extended to general alphabets by building on the posterior matching scheme, where the empirical capacity is defined relative to some parametric family of channels.

An extension of the suggested scheme to channels with memory is certainly called for. However, the posterior matching principle needs to be modified to take the channel's memory into account, since it is clear that a transmission independent of previous observations is not always the best option in this case. In hindsight, this part of the principle could have been phrased differently: *The transmission functions should be selected so that the input sequence has the correct marginal distribution, and the output sequence has the correct joint distribution*. In the memoryless case, this is just to say that $X_n \sim P_X$, and Y^n is i.i.d. with the marginal P_Y induced by $(P_X, P_{Y|X})$, which coincides with the original principle. However, when the channel has memory the revised principle seems to lead to the correct generalization. For instance, consider a setting where the channel is Markovian of some order d, and the "designed" input distribution is selected to be Markovian of order d as well[22]. According to the revised principle, the input to the channel should be generated in such a way that any d consecutive input/output pairs have the correct (designed) distribution[23], and the joint output distribution is the one induced by the designed input distribution and the channel, so the receiver cannot "tell the difference". To emulate such a behavior, a $d + 1$ order (or higher) kernel is required, since any lower order will result in some deterministic dependence between any d consecutive pairs. This also implies that a $d + 1$ dimensional message point is generally required in order to provide the necessary degrees of freedom in terms of randomness. It can

---

[22]By that, we mean that $Y_n - X_{n-d}^n Y_{n-d}^{n-1} - X^{n-d-1} Y^{n-d-1}$ and $X_n - X_{n-d}^{n-1} Y_{n-d}^{n-1} - X^{n-d-1} Y^{n-d-1}$ are Markov chains.

[23]We interpret "marginal" here as pertaining to the degrees of freedom suggested by the designed input distribution.

be verified that whenever such a procedure is feasible, then under some mild regularity conditions the posterior p.d.f. at the message point is $\approx 2^{I(X^n \to Y^n)}$, where $I(X^n \to Y^n)$ is the *directed information* pertaining to the designed input distribution and the channel [31]. This is encouraging, since for channels with feedback the directed information usually plays the same role as mutual information does for channels without feedback [31]–[35]. Note also that the randomness degrees of freedom argument for a multidimensional message point, provides a complementary viewpoint on the more analytic argument as to why the additional dimensions are required in order to attain the capacity of an auto-regressive Gaussian channel via a generalized Schalkwijk-Kailath scheme [36]. It is expected that a scheme satisfying the revised principle and its analysis should follow through via a similar approach to that appearing in this paper.

APPENDIX A
MAIN PROOFS

Proof of Lemma 1: For the first claim, let us find the c.d.f. of $F_X^{-1}(\Theta)$:

$$\mathbb{P}(F_X^{-1}(\Theta) \leq x) = \mathbb{P}(\inf\{z : F_X(z) > \Theta\} \leq x)$$
$$\overset{(a)}{=} \mathbb{P}(F_X(x) \geq \Theta) = F_X(x)$$

where (a) holds since a c.d.f. is nondecreasing and continuous from the right, and so the result follows. For the second claim, define $\Phi = F_X(X) - \Theta \cdot P_X(X)$ and let $\phi \in (0,1)$ be such that there exists $x_0 \in \text{supp}(X)$ for which $F_X(x_0) = \phi$. Then

$$F_\Phi(\phi) \geq \mathbb{P}(F_X(X) \leq F_X(x_0)) = \mathbb{P}(X \leq x_0)$$
$$= F_X(x_0) = \phi$$

and on the other hand

$$F_\Phi(\phi) \leq \mathbb{P}(F_X(X) - P_X(X) \leq F_X(x_0)) = \mathbb{P}(X \leq x_0)$$
$$= F_X(x_0) = \phi$$

hence $F_\Phi(\phi) = \phi$. If such an x_0 does not exists then there must exist a jump point x_1 such that

$$F_X(x_1) - P_X(x_1) \leq \phi < F_X(x_1) \overset{\text{def}}{=} \phi_1$$

and so

$$F_\Phi(\phi) = F_\Phi(\phi_1) - P_\Phi((\phi, \phi_1])$$
$$= \phi_1 - \mathbb{P}(X = x_1, \Theta \cdot P_X(x_1) \leq \phi_1 - \phi)$$
$$= \phi_1 - P_X(x_1) \cdot \frac{\phi_1 - \phi}{P_X(x_1)} = \phi.$$

For a proper X, there are no mass points hence the simpler result follows immediately. ∎

Proof of Lemma 3: Assume we are given a transmission scheme g_n and a decoding rule Δ_n which are known to achieve a rate R_0. For simplicity, we assume the decoding rule is fixed rate, (i.e., $|\Delta(y^n)| = 2^{-nR_0}$ for all y^n), since any variable rate rule can be easily mapped into a fixed rate rule that achieves

the same rate. It is easy to see that in order to prove the above translates into achievability of some rate $R < R_0$ in the standard framework, it is enough to show we can find a sequence $\Gamma_n = \{\theta_{i,n} \in (0,1)\}_{i=1}^{\lfloor 2^{nR} \rfloor}$ of message point sets, such that $\theta_{i+1,n} - \theta_{i,n} \geq 2^{-nR_0}$ for any $1 \leq i < \lfloor 2^{nR} \rfloor$, and such that we have uniform achievability over Γ_n, i.e.,

$$\lim_{n \to \infty} \max_{\theta \in \Gamma_n} \mathbb{P}(\theta \notin \Delta_n(Y^n) | \Theta_0 = \theta) = 0$$

We now show how Γ_n can be constructed for any $R < R_0$. Let $p_e(n)$ be the (average) error probability associated with our scheme and the fixed rate R_0 decoding rule. Define

$$A_n = \{\theta \in (0,1) : \mathbb{P}(\Theta_0 \notin \Delta(Y^n) | \Theta_0 = \theta) > \sqrt{p_e(n)}\}$$

and write

$$p_e(n) = \int \mathbb{P}(\Theta_0 \notin \Delta(Y^n) | \Theta_0 = \theta) d\theta$$
$$> \sqrt{p_e(n)} \int \mathbb{1}_{A_n}(\theta) d\theta$$

and so we have that $\int \mathbb{1}_{A_n}(\theta) d\theta < \sqrt{p_e(n)}$. It is now easy to see that if we want to select Γ_n such that $\Gamma_n \cap A_n = \emptyset$, and also $\theta_{i+1,n} - \theta_{i,n} \geq 2^{-nR_0}$, then a sufficient condition is that $\frac{1}{|\Gamma_n|}(1 - \sqrt{p_e(n)} - \tau_n) \geq 2^{-nR_0}$ for some positive $\tau_n \to 0$. This condition can be written as

$$\frac{1}{n} \log |\Gamma_n| \leq R_0 + \frac{1}{n} \log(1 - \sqrt{p_e(n)} - \tau_n) = R_0 + o(1).$$

At the same time, we also have by definition

$$\lim_{n \to \infty} \max_{\theta \in \Gamma_n} \mathbb{P}(\theta \notin \Delta(Y^n) | \Theta_0 = \theta) \leq \lim_{n \to \infty} \sqrt{p_e(n)} = 0$$

and the proof is concluded. ∎

Proof of Lemma 8: Since ξ is $\cap$-convex over $[0,1]$, it has a unique maximal value attained at some (not necessarily unique) point x^*. Moreover, convexity implies ξ is continuous over $(0,1)$, and since it is nonnegative and upper bounded by $\xi(x) < x$, it is also continuous at $x = 0$ and $\xi(0) = 0$. Now, define the sequence $s_n = \xi^{(n)}(x^*)$. Since $\xi(x) < x$ the sequence s_n is monotonically decreasing, and since ξ is nonnegative it is also bounded from below. Therefore, s_n converges to a limit $s_\infty \in [0,1)$, and we can write

$$\lim_{n \to \infty} s_n = s_\infty, \quad \lim_{n \to \infty} \xi(s_n) = \lim_{n \to \infty} s_{n+1} = s_\infty$$

Since ξ is continuous over $[0,1)$ the above implies that $\xi(s_\infty) = s_\infty$, i.e., s_∞ is a fixed point of ξ. Thus, we either have $\xi \equiv 0$ in which case $s_\infty = 0$, or $\xi \not\equiv 0$ in which case the only fixed point for ξ is zero and so again $s_\infty = 0$. We now note that $\xi(x) \leq \xi(x^*) \leq x^*$ for any $x \in [0,1]$, and also that ξ is nondecreasing over $[0, x^*]$ and hence so is $\xi^{(n)}$. We therefore have

$$\lim_{n \to \infty} r(n) = \lim_{n \to \infty} \sup_{x \in [0,1]} \xi^{(n)}(x) \leq \lim_{n \to \infty} \xi^{(n-1)}(x^*)$$
$$= \lim_{n \to \infty} s_n = 0.$$

∎

Proof of Lemma 9: For any $\varepsilon > 0$

$$
\begin{aligned}
\mathbb{P}\left(\psi(S_n(s)) > \varepsilon\right) &\overset{(a)}{\leq} \varepsilon^{-1}\mathbb{E}[\psi(S_n(s))] \\
&= \varepsilon^{-1}\mathbb{E}\left(\mathbb{E}[\psi(S_n(s))\,|\,Y^{n-1}]\right) \\
&= \varepsilon^{-1}\mathbb{E}\left(\mathbb{E}[\psi(\omega_{Y_n}\circ S_{n-1}(s))\,|\,Y^{n-1}]\right) \\
&\overset{(b)}{\leq} \varepsilon^{-1}\mathbb{E}\,\xi\left(\psi(S_{n-1}(s))\right) \\
&\overset{(c)}{\leq} \varepsilon^{-1}\xi\left(\mathbb{E}\psi(S_{n-1}(s))\right) \leq \cdots \\
&\overset{(d)}{\leq} \varepsilon^{-1}\xi^{(n)}\left(\psi(s)\right) \\
&\overset{(e)}{\leq} \varepsilon^{-1}r(n).
\end{aligned}
$$

Markov's inequality was used in (a), the contraction relation (7) in (b) and Jensen's inequality in (c). Inequality (d) is a recursive application of the preceding transitions, and the definition of the decay profile was used in (e). ∎

Proof of Lemma 10: For any $\varepsilon > 0$

$$
\begin{aligned}
\mathbb{P}(|\tilde{S}_n(s) &- \tilde{S}_n(t)| > \varepsilon) \\
&= \mathbb{P}(|\tilde{S}_n(s) - \tilde{S}_n(t)|^q > \varepsilon^q) \\
&\overset{(a)}{\leq} \varepsilon^{-q}\mathbb{E}|\tilde{S}_n(s) - \tilde{S}_n(t)|^q \\
&= \varepsilon^{-q}\mathbb{E}(\mathbb{E}(|\tilde{S}_n(s) - \tilde{S}_n(t)|^q\,|\,Y_2^n)) \\
&= \varepsilon^{-q}\mathbb{E}(\mathbb{E}(|\omega_{Y_1}\circ\cdots\circ\omega_{Y_n}(s) - \omega_{Y_1}\circ\cdots\circ\omega_{Y_n}(t)|^q|Y_2^n)) \\
&\overset{(b)}{\leq} \varepsilon^{-q}r\cdot\mathbb{E}(|\omega_{Y_2}\circ\cdots\circ\omega_{Y_n}(s) - \omega_{Y_2}\circ\cdots\circ\omega_{Y_n}(t)|^q) \\
&\leq \cdots \overset{(c)}{\leq} \varepsilon^{-q}r^n|s-t|^q.
\end{aligned}
$$

Where in (a) we use Markov's inequality, in (b) we use the contraction (10), and (c) is a recursive application of the preceding transitions. ∎

Proof of Theorem 1 [completed]: We prove by induction that for any $n \in \mathbb{N}$, $P_{\Theta_0|Y^n}(\cdot\,|\,y^n)$ is proper for P_{Y^n}-a.a. $y^n \in \mathcal{Y}^n$, and the rest of the proof remains the same. First, this property is satisfied for $n = 0$ since P_{Θ_0} is proper. Now assume the property holds for any $1 \leq n \leq k-1$. By our previous derivations, this implies that $X_n \sim P_X$ for any $1 \leq n \leq k$, and thus by the definition of an input/channel pair we have in particular $I(X_n;Y_n) = I(X;Y) < \infty$ for any such n. Now suppose the property does not hold for $n = k$. This implies there exists a measurable set $A \subseteq \mathcal{Y}^k$ with $P_{Y^k}(A) > 0$ so that $P_{\Theta_0|Y^k}(\cdot\,|\,y^k) \not\ll P_{\Theta_0}$ for any $y^k \in A$. Therefore, it must be that $I(\Theta_0;Y^k) = \infty$. However standard manipulations using the fact that the channel is memoryless result in $I(\Theta_0;Y^k) \leq \sum_{n=1}^{k} I(X_n;Y_n) < \infty$, in contradiction. ∎

Proof of Lemma 12, Claim (iii): Since $\Theta \sim \mathcal{U}$, it is enough to show that $P_{\Phi|\Theta}(\cdot\,|\,\theta)$ is proper for $\mathcal{U}$-a.a. $\theta \in (0,1)$. Define the discrete part of the output support to be $\mathcal{Y}_D = \{y \in \text{supp}(Y) : P_Y(y) > 0\}$, which is a countable set. Define also the set $\mathfrak{Y}_D \overset{\text{def}}{=} \text{supp}(\Phi\,|\,Y \in \mathcal{Y}_D)$ which is a countable union of disjoint intervals inside the unit interval, corresponding to the "jump spans" introduced by F_Y over $\mathcal{Y}_D$. Furthermore, for any $x \in \text{supp}(X)$ define $\mathcal{Y}_{D,x}$ to be the set of mass points for $P_{Y|X}(\cdot\,|\,x)$. Since $I(X;Y) < \infty$, then it

must be that $P_{Y|X}(\cdot\,|\,x) \ll P_Y$ for P_X-a.a. $x \in \text{supp}(X)$. Therefore, there exists a set $A \subseteq \text{supp}(X)$ of full measure $P_X(A) = 1$, so that $\mathcal{Y}_{D,x} \subseteq \mathcal{Y}_D$ for any $x \in A$. Therefore, for any $x \in A$, $P_{Y|X}(\cdot\,|\,x)$ restricted to $\text{supp}(Y)\setminus\mathcal{Y}_D$ has a proper p.d.f., which implies that $P_{\Phi|X}(\cdot\,|\,x)$ restricted to $(0,1)\setminus\mathfrak{Y}_D$ has a proper p.d.f. as well, since Φ is obtained from Y by applying a continuous and bounded function. $P_{\Phi|X}(\cdot\,|\,x)$ restricted to any one of the countable number of intervals composing $\mathfrak{Y}_D$ is uniform, hence admits a proper p.d.f. as well. We therefore conclude that $P_{\Phi|X}(\cdot\,|\,x)$ is proper for any $x \in A$. To conclude, define the set $B = \{\theta \in (0,1) : F_X^{-1}(\theta) \in A\}$, which by Lemma 1 is of full measure $\mathcal{U}(B) = 1$, and from the discussion above $P_{\Phi|\Theta}(\cdot\,|\,\theta)$ is proper for any $\theta \in B$. ∎

Proof of Lemma 16, Claim (ii): Suppose there exists some $y_0 \in \mathcal{Y}$ so that $P_Y(y_0) > 0$ and $P_X \prec_d P_{X|Y}(\cdot\,|\,y_0)$. Define the set $A_0 = \{\phi \in (0,1) : F_Y^{-1}(\phi) = y_0\}$. For any $x \in \mathcal{X}$ and $\phi \in A_0$, the normalized posterior matching kernel evaluated at $\theta = F_X(x)$ satisfies

$$
F_{\Theta|\Phi}(F_X(x)|\phi) = F_{X|Y}(x|y_0) \geq F_X(x)
$$

where the last inequality is due to the dominance assumption above, and is *strict* for $x \in \{0,\ldots,|\mathcal{X}|-2\}$. Moreover, the normalized posterior matching kernel evaluated in between this finite set of points is simply a linear interpolation. Thus, for any $\theta \in (0,1)$ and any $\phi \in A_0$, we have $F_{\Theta|\Phi}(\theta|\phi) > \theta$, and so

$$
\mathbb{P}(F_{\Theta|\Phi}(\theta|\Phi) = \theta) \leq 1 - P_\Phi(A_0) = 1 - P_Y(y_0) < 1
$$

which implies the fixed-point free property (A3). The case where $P_{X|Y}(\cdot\,|\,y_0) \prec_d P_X$ follows by symmetry. The case where $P_{X|Y}(\cdot\,|\,y_0) \prec_d P_X(\cdot\,|\,y_1)$ is trivial. ∎

Proof of Lemma 16, Claim (iii): We find it simpler here to consider the normalized input Θ but the original output Y, namely to prove an equivalent claim stating that the invariant distribution $P_{\Theta Y}$ for the Markov chain (Θ_n, Y_n), is ergodic. To that end, we show that if $S \subseteq (0,1) \times \mathcal{Y}$ is an invariant set, then $P_{\Theta Y}(S) \in \{0,1\}$. Let us write S as a disjoint union:

$$
S = \bigcup_{y\in\mathcal{Y}} A_y \times \{y\}, \quad A_y \subseteq (0,1)
$$

The posterior matching kernel deterministically maps a pair (θ, y) to the input $\hat{\theta} = F_{\Theta|Y}(\theta|y)$, and then the corresponding output is determined via $P_{Y|\Theta}(\cdot\,|\,\hat{\theta})$. Since by (B1) all transition probabilities are nonzero, then each possible output in $\mathcal{Y}$ is seen with a nonzero probability given any input. Thus, denoting the stochastic kernel of the Markov chain by $\mathcal{P}$, we have that $\mathcal{P}(\cdot\,|(\theta,y))$ has support on the discrete set $\{F_{\Theta|Y}(\theta|y)\}\times\mathcal{Y}$ for any $(\theta, y) \in S$. Since S is an invariant set, this implies that

$$
S' \overset{\text{def}}{=} \bigcup_{y\in\mathcal{Y}} F_{\Theta|Y}(A_y|y) \times \mathcal{Y} \subseteq S
$$

where by $F_{\Theta|Y}(A_y|y)$ we mean the image set of A_y under $F_{\Theta|Y}(\cdot\,|\,y)$. This in turn implies that

$$
\bigcup_{y\in\mathcal{Y}} F_{\Theta|Y}(A_y|y) \subseteq \bigcap_{y\in\mathcal{Y}} A_y \overset{\text{def}}{=} A. \tag{57}
$$

Now, defining

$$\bar{S} \stackrel{\text{def}}{=} A \times \mathcal{Y}$$

we have that $S' \subseteq \bar{S} \subseteq S$, and hence $\bar{S}$ is also an invariant set. Going through the same derivations as for S, the invariance of $\bar{S}$ implies that

$$\bigcup_{y \in \mathcal{Y}} F_{\Theta|Y}(A|y) \subseteq A \qquad (58)$$

and hence

$$\mathcal{U}(A) \geq \max_{y \in \mathcal{Y}} \mathcal{U}(F_{\Theta|Y}(A|y)) \geq \sum_{y \in \mathcal{Y}} \mathcal{U}(F_{\Theta|Y}(A|y)) P_Y(y)$$
$$= \sum_{y \in \mathcal{Y}} P_{\Theta|Y}(A|y) P_Y(y) = P_\Theta(A) = \mathcal{U}(A).$$

To avoid contradiction, it must be that $\mathcal{U}(F_{\Theta|Y}(A|y)) = \mathcal{U}(A)$ for all $y \in \mathcal{Y}$, and together with (58) it immediately follows that for all $y \in \mathcal{Y}$

$$F_{\Theta|Y}(A|y) = A \setminus N_y, \quad \mathcal{U}(N_y) = 0. \qquad (59)$$

Namely, for any output value, the set A remains the same after applying the posterior matching kernel, up to a $\mathcal{U}$-null set.

Let us now prove the implication

$$\mathcal{U}(A) \in \{0,1\} \Rightarrow P_{\Theta Y}(S) \in \{0,1\}. \qquad (60)$$

To that end, we show that $0 < P_{\Theta Y}(S) < 1$ implies $0 < \mathcal{U}(A) < 1$. The upper bound follows from $\mathcal{U}(A) = P_\Theta(A) = P_{\Theta Y}(\bar{S}) \leq P_{\Theta Y}(S) < 1$. For the lower bound, we note that $P_{\Theta Y}(S) > 0$ implies there exists at least one $y_0 \in \mathcal{Y}$ such that $\mathcal{U}(A_{y_0}) > 0$. Recall that for a DMC, the normalized posterior matching kernel for any fixed output is a quasi-affine function with slopes given by $\frac{P_{X|Y}(x|y)}{P_X(x)} = \frac{P_{Y|X}(y|x)}{P_Y(y)}$. Since by $(B1)$ all the transition probabilities are nonzero, these slopes are all positive, and denote their minimal value by $\alpha > 0$. Therefore, it must be that $\mathcal{U}(F_{\Theta|Y}(A_{y_0}|y_0)) > \alpha \mathcal{U}(A_{y_0}) > 0$, which by (57) implies $\mathcal{U}(A) > 0$.

After having established (60), we proceed to show that $\mathcal{U}(A) \in \{0,1\}$ which will verify Property $(A2)$. It is easily observed that if A is an interval, (59) holds if and only if the endpoints of the interval are both either fixed points of the kernel or endpoints of $(0, 1)$. For A a finite disjoint union of intervals, (59) holds if and only if all nonshared endpoints are both either fixed points of the kernel or endpoints of $(0, 1)$. Hence for such A, since we assumed the kernel does not have any fixed points, (59) holds if and only if $\mathcal{U}(A) \in \{0,1\}$.

Let us now extend this argument to any $A \in \mathfrak{B}$. Under $(B3)$, there exist two output symbols $y_0, y_1 \in \mathcal{Y}$ such that

$$0 > \frac{\beta_0}{\beta_1} \notin \mathbb{Q}$$

where

$$\beta_i = \log \left(\frac{P_{X|Y}(0|y_i)}{P_X(0)} \right), \quad i \in \{0,1\}.$$

Define the set

$$B \stackrel{\text{def}}{=} \left\{ b \in (0,1) : \exists n_0, n_1 \in \mathbb{N}, b = 2^{n_0 \beta_0 + n_1 \beta_1} \right\}.$$

Lemma 23: B is dense in (0,1):
   Proof: Without loss of generality, we assume $\beta_0 < 0 < \beta_1$. We prove equivalently that the set $\log B$ is dense in $(-\infty, 0)$. Let $b \in (-\infty, 0)$. Define

$$b_n \stackrel{\text{def}}{=} n\beta_0 + \left\lfloor \frac{b - n\beta_0}{\beta_1} \right\rfloor \beta_1, \quad n \in \mathbb{N}.$$

It is easy to see that $\{b_n\}_{n=n'}^\infty \subset \log B$, if n' is taken to be large enough. Let $\{x\} \stackrel{\text{def}}{=} x - \lfloor x \rfloor$ be the fractional part of x. Write:

$$r_n \stackrel{\text{def}}{=} \frac{b - b_n}{\beta_1} = \left\{ \frac{b}{\beta_1} + n \left(-\frac{\beta_0}{\beta_1} \right) \right\}.$$

Since $\frac{\beta_1}{\beta_0} \notin \mathbb{Q}$, r_n can be though of as an irrational rotation on the unit circle, hence is dense in (0,1) [37]. In particular, this implies that r_n has a subsequence $r_{k_n} \to 0$, hence $b_{k_n} \to b$. ∎

For $\theta \in (0,1)$, let $A(\theta) \stackrel{\text{def}}{=} A \cap (0,\theta)$. For brevity, let $p \stackrel{\text{def}}{=} P_X(0)$. Define A_{n_0,n_1} be the set obtained starting from $A(p)$ and applying $F_{\Theta|Y}(\cdot|y_0)$ n_0 times, and then applying $F_{\Theta|Y}(\cdot|y_1)$ n_1 times. $F_{\Theta|Y}(\cdot|y_i)$ is linear over $(0,p)$ with a slope 2^{β_i}, hence assuming that $2^{n_0 \beta_0 + n_1 \beta_1} \leq 1$, we have

$$\mathcal{U}(A_{n_0,n_1}) = 2^{n_0 \beta_0 + n_1 \beta_1} \cdot \mathcal{U}(A(p)). \qquad (61)$$

On the other hand, (59) together with the aforementioned linearity imply that A_{n_0,n_1} and $A(p \cdot 2^{n_0 \beta_0 + n_1 \beta_1})$ are equal up to a $\mathcal{U}$-null set. Combining this with (61) and Lemma 23, we find that for any $\theta \in (0,p)$

$$\mathcal{U}(A(\theta)) = \theta p^{-1} \mathcal{U}(A(p)).$$

We note that $\mathcal{U}(A(\theta))$ is the indefinite Lebesgue integral of $\mathbb{1}_{A(p)}(\theta)$. Invoking the Lebesgue differentiation Theorem [20], the derivative $\frac{d\mathcal{U}(A(\theta))}{d\theta} = p^{-1} \mathcal{U}(A(p))$ must be equal to $\mathbb{1}_{A(p)}(\theta)$ for a.a. $\theta \in (0,p)$, which implies $p^{-1} \mathcal{U}(A(p)) \in \{0,1\}$. Hence, $A(p)$ is either of full measure or a null set.

Let us now show that this implies the same for $A = A(1)$. Define the function

$$\bar{F}(\theta) \stackrel{\text{def}}{=} \max_{y \in \mathcal{Y}} F_{\Theta|Y}(\theta)$$

Let us establish some properties of $\bar{F}$.
   a) *$\bar{F}$ is Lipschitz, monotonically increasing, and maps (0,1) onto (0,1):* Trivial.
   b) *$\bar{F}^{(n)}(\theta) \to 1$ monotonically as $n \to \infty$ for any $\theta \in (0,1)$:* Observe that

$$\mathbb{E}\left(F_{\Theta|Y}(\theta|Y) \right) = \mathbb{E}\mathbb{P}(\Theta \leq \theta|Y) = \mathbb{P}(\Theta \leq \theta) = \theta.$$

Hence $\bar{F}(\theta) \geq \theta$ with equality if and only if θ is a fixed point, which contradicts property $(A3)$. Thus, it must hold that $\bar{F}(\theta) > \theta$ for any $\theta \in (0,1)$, hence $\bar{F}^{(n)}(\theta)$ is increasing with n. $\bar{F} \leq 1$ and therefore a limit exists and

is at most 1. $\bar{F}$ is continuous, hence the limit cannot be smaller than 1 as this will violate $\bar{F}(\theta) > \theta$.

c) $\bar{F}(A) = A$ *up to a* $\mathcal{U}$-*null set:* it is easily observed that

$$\bigcap_{y \in \mathcal{Y}} F_{\Theta|Y}(A|y) \subseteq \bar{F}(A) \subseteq \bigcup_{y \in \mathcal{Y}} F_{\Theta|Y}(A|y).$$

The property follows by applying (59).

Combining (a) and (c) it follows that for any $n \geq 1$, $\bar{F}^{(n)}(A(p)) = A(\bar{F}^{(n)}(p))$ up to a $\mathcal{U}$-null set. Furthermore, since $A(p)$ is either of full measure or null, then property (a) implies the same for $\bar{F}^{(n)}(A(p))$, and so either $\mathcal{U}(A(\bar{F}^{(n)}(p))) = 0$ for all n, or $\mathcal{U}(A(\bar{F}^{(n)}(p))) = \bar{F}^{(n)}(p)$. Using (b), we get:

$$\mathcal{U}(A) = \mathcal{U}\left(\bigcup_{n=1}^{\infty} A(\bar{F}^{(n)}(p)) \right) \in \left\{ 0, \lim_{n \to \infty} \bar{F}^{(n)}(p) \right\}$$
$$= \{0, 1\}.$$

Hence (*A*2) holds.

Remark 7: The proof only requires an irrational ratio to be found for $x = 0$ (or similarly, for $x = |\mathcal{X}| - 1$), hence a weaker version of property (*B*3) suffices. It is unclear if even this weaker property is required for ergodicity to hold. The proof fails whenever the leftmost interval $(0, P_X(0))$ cannot be densely covered by a repeated application the posterior matching kernel (starting from the right endpoint), without ever leaving the interval. This argument leans only on the linearity of the kernel within that interval, and does not use the entire nonlinear structure of the kernel. It therefore seems plausible that condition (*B*3) could be further weakened, or perhaps even completely removed. ∎

Proof of Lemma 16, Claim (v): (*B*1) trivially holds for any equivalent input/channel pair. Let us show there exists one satisfying (*B*2). To that end, the following Lemma is found useful.

Lemma 24: Let p^n, q^n be two distinct probability vectors. Then there exists a permutation operator $\sigma : \mathbb{R}^n \mapsto \mathbb{R}^n$ such that $\sigma(q^n) \prec_d \sigma(p^n)$.

Proof: Let δ^n be the element-wise difference of p^n and q^n, i.e., $\delta_k = p_k - q_k$. Define σ to be a permutation operator such that $\sigma(\delta^n)$ is in descending order. Then since $p^n \neq q^n$ and $\sum_{i=1}^n \delta_i = 0$ we have that any partial sum of $\sigma(\delta^n)$ is positive, i.e., $\sum_{i=1}^k \{\sigma(\delta^n)\}_i > 0$ for any $k < n$, which implies the result. ∎

Now, since $I(X; Y) > 0$ there must exist some $y_0 \in \mathcal{Y}$ so that $P_{X|Y}(\cdot|y_0) \neq P_X$. Viewing distributions as probability vectors, then by Lemma 24 above there exists a permutation operator σ such that $\sigma(P_X) \prec_d \sigma(P_{X|Y}(\cdot|y_0))$. Thus, applying σ to the input results in an equivalent input/channel pair for which (*B*2) holds. ∎

Proof of Lemma 16, Claim (vi): Let $(\mathsf{P}(\mathcal{X}), d_{TV})$ be the space of probability distributions over the alphabet $\mathcal{X}$, equipped with the total variation metric. For a fixed channel $P_{Y|X}$, the set S of input distributions not satisfying property (*B*3) is clearly of countable cardinality. Since any open ball around any point in $\mathsf{P}(\mathcal{X})$ is of uncountable cardinality, then $\mathsf{P}(\mathcal{X}) \backslash S$ must be dense in $(\mathsf{P}(\mathcal{X}), d_{TV})$, and the claim follows. ∎

Proof of Lemma 17: Let $\lambda : [0, 1] \mapsto [0, 1]$ be any surjective, strictly $\cap$-convex function symmetric about $\frac{1}{2}$. This implies in particular that $\lambda(\cdot)$ is continuous, its restriction to $[0, \frac{1}{2}]$ is injective, and $\lambda(0) = \lambda(1) = 0, \lambda(\frac{1}{2}) = 1$. Let $\lambda^{-1} : [0, 1] \mapsto [0, \frac{1}{2}]$ be the inverse of λ restricted to the $[0, \frac{1}{2}]$ branch. Let ψ_λ be the corresponding length function over $\mathfrak{F}_c$, as defined in (8). Define the function $\xi^* : [0, 1] \mapsto [0, 1]$ as follows:

$$\xi^*(\theta) \overset{\text{def}}{=} \max \left\{ \mathbb{E}\lambda \left(F_{\Theta|\Phi}(\lambda^{-1}(\theta)|\Phi) \right), \right.$$
$$\left. \mathbb{E}\lambda \left(F_{\Theta|\Phi}(1 - \lambda^{-1}(\theta)|\Phi) \right) \right\}.$$

We now establish the following two properties:

a) $\xi^*(\cdot)$ *is continuous over* $[0, 1]$: Fix any $\theta' \in [0, 1]$, and let $\{\theta_n\}_{n=1}^{\infty}$ be a sequence in $[0, 1]$ such that $\theta_n \to \theta'$. Define $q(\theta, \phi) \overset{\text{def}}{=} \lambda(F_{\Theta|\Phi}(\lambda^{-1}(\theta)|\phi))$, and $q_n(\phi) \overset{\text{def}}{=} q(\theta_n, \phi)$. By Lemma 12 claim (iv), $F_{\Theta|\Phi}(\theta|\phi)$ is continuous in θ for P_Φ-a.a. $\phi \in (0, 1)$. Since $\lambda(\cdot), \lambda^{-1}(\cdot)$ are continuous, we have that $q(\theta, \phi)$ is continuous in θ for P_Φ-a.a. $\phi \in (0, 1)$, and therefore $q_n(\phi) \to q(\theta', \phi)$ for P_Φ-a.a. $\phi \in (0, 1)$. Furthermore, $|q_n(\phi)| \leq 1$. Thus, invoking the bounded convergence Theorem [20], we get

$$\lim_{n \to \infty} \mathbb{E}(q_n(\Phi)) = \mathbb{E}(q(\theta', \Phi))$$

Reiterating for $q(\theta, \phi) \overset{\text{def}}{=} \lambda(F_{\Theta|\Phi}(1 - \lambda^{-1}(\theta)|\phi))$, we conclude that $\xi^*(\theta_n) \to \xi^*(\theta')$.

b) $0 \leq \xi^*(\theta) < \theta$ *for* $\theta \in (0, 1]$: The lower bound is trivial. For the upper bound, we note again that

$$\mathbb{E}\left(F_{\Theta|\Phi}(\theta|\Phi) \right) = \mathbb{E}\mathbb{P}(\Theta \leq \theta|\Phi) = \mathbb{P}(\Theta \leq \theta) = \theta,$$

and since by the fixed-point free property (*A*3) we also have $\mathbb{P}(F_{\Theta|\Phi}(\theta|\Phi) = \theta) < 1$ for any $\theta \in (0, 1)$, then $F_{\Theta|\Phi}(\theta|\Phi)$ is not a.s. constant. Combining that with the fact that $\lambda(\cdot)$ is *strictly* $\cap$-convex, a strict Jensen's inequality holds

$$\mathbb{E}\lambda(F_{\Theta|\Phi}(\lambda^{-1}(\theta)|\Phi)) < \lambda \left(\mathbb{E} \left(F_{\Theta|\Phi}(\lambda^{-1}(\theta)|\Phi) \right) \right)$$
$$= \lambda(\lambda^{-1}(\theta)) = \theta$$

Similarly, using the symmetry of $\lambda(\cdot)$,

$$\mathbb{E}\lambda(F_{\Theta|\Phi}(1 - \lambda^{-1}(\theta)|\Phi)) < \lambda(1 - \lambda^{-1}(\theta))$$
$$= \lambda(\lambda^{-1}(\theta)) = \theta$$

Now, define $\xi(\cdot)$ to be the upper convex envelope of $\xi^*(\cdot)$. Let us show that $\xi(\cdot)$ is a contraction. $\xi(\cdot)$ is trivially $\cap$-convex and nonnegative, hence it remains to prove that $\xi(\theta) < \theta$ for $\theta \in (0, 1]$. Define the function

$$\delta(\theta) \overset{\text{def}}{=} \inf_{\phi \in [\theta, 1]} (\phi - \xi^*(\phi))$$

Property (b) implies that $\delta(0) = 0$. Combining properties (a) and (b), we observe that $\phi - \xi^*(\phi)$ is continuous and positive over $[\theta, 1]$ for any fixed $\theta \in (0, 1)$, hence attains a positive infimum over that interval. We conclude that $\delta(\theta)$ is continuous

and monotonically nondecreasing over $[0,1]$, and positive over $(0,1]$. Fixing any $\theta' \in (0,1]$, we use the definition of the upper convex hull and the above properties of $\delta(\cdot)$ to write

$$
\begin{aligned}
\xi(\theta') &= \sup\{\alpha\xi^*(\theta_0) + (1-\alpha)\xi^*(\theta_1)\} \\
&\leq \sup\{\alpha(\theta_0 - \delta(\theta_0)) + (1-\alpha)(\theta_1 - \delta(\theta_1))\} \\
&\leq \theta' - \inf\{\alpha\delta(\theta_0) + (1-\alpha)\delta(\theta')\} \quad (62)
\end{aligned}
$$

where the supremums and the infimum are taken over all $\{\theta_0, \theta_1, \alpha\}$ such that $0 \leq \theta_0 \leq \theta' \leq \theta_1 \leq 1$, and such that θ' is the convex combination $\theta' = \alpha\theta_0 + (1-\alpha)\theta_1$. Thus, since $\delta(\theta') > 0$, a necessary condition for $\xi(\theta') \geq \theta'$ is for the infimum in (62) to be attained as $\alpha \to 1$ and $\delta(\theta_0) \to 0$. By continuity and positivity, the latter implies $\theta_0 \to 0$. However, the convex combination for θ' can be maintained as $\alpha \to 1$ and $\theta_0 \to 0$ if and only if $\theta' = 0$, in contradiction. Hence $\xi(\theta') < \theta'$.

To conclude the proof, we demonstrate that $\xi(\cdot)$ and ψ_λ satisfy (36):

$$
\begin{aligned}
\mathbb{E}(\psi_\lambda[F_{\Theta|\Phi}(\cdot|\Phi) \circ h]) &= \int_0^1 \mathbb{E}\lambda(F_{\Theta|\Phi}(h(\theta)|\Phi))d\theta \\
&\overset{(a)}{\leq} \int_0^1 \xi^*(\lambda(h(\theta)))d\theta \\
&\overset{(b)}{\leq} \int_0^1 \xi(\lambda(h(\theta)))d\theta \\
&\overset{(c)}{\leq} \xi\left(\int_0^1 \lambda(h(\theta))d\theta\right) \\
&= \xi(\psi_\lambda(h))
\end{aligned}
$$

where (a) holds by the definition of ξ^* and the symmetry of $\lambda(\cdot)$, (b) holds since $\xi \geq \xi^*$, and (c) holds by Jensen's inequality.   ∎

Lemma 25: Let $(P_X, P_{Y|X})$ satisfy property (A3). Then for any $\alpha > 0$, $\varepsilon > 0$ and $\delta > 0$,

$$
\mathbb{P}\left(\max_{1\leq m\leq n} {}^-\Theta^\varepsilon_{(1+\alpha)n-m}(\Theta_m) > \delta\right) = O(\sqrt[8]{r(\alpha n)})
$$

where $r(n)$ is the decay profile of the contraction $\xi(\cdot)$ from Lemma 17.

Proof: For any $g \in \mathfrak{F}_c$ and any $m, n \in \mathbb{N}$ where $m \leq n$, define

$$
\bar{G}_{m,m}(\cdot) \overset{\text{def}}{=} g(\cdot), \quad \bar{G}^g_{m,n}(\cdot) \overset{\text{def}}{=} F_{\Theta|\Phi}(\cdot|\Phi_n) \circ \bar{G}_{m,n-1}(\cdot).
$$

Then for any fixed m and g, $\{G^g_{m,n}\}_{n=m}^\infty$ is an IFS over $\mathfrak{F}_c$. Let $g_u(\theta) = \theta$ be the uniform c.d.f., and define the following r.v.'s:

$$
L_{m,n} \overset{\text{def}}{=} \psi_\lambda\left(\bar{G}^{g_u}_{m,n}\right) \quad L^*_{m,n} \overset{\text{def}}{=} \sup_{g\in\mathfrak{F}_c} \psi_\lambda\left(\bar{G}^g_{m,n}\right)
$$

where ψ_λ is the associated length function from Lemma 17. Clearly, $L_{m,n} \leq L^*_{m,n}$. Furthermore, $L^*_{m,n} \leq L^*_{m+1,n}$ for any $m \leq n - 1$. To see that, we note that $L^*_{m,n}$ is a deterministic function of $\Phi^n_m \overset{\text{def}}{=} (\Phi_m, \ldots, \Phi_n)$, hence there exists a sequence

of functions $\{g_k(\theta; \phi^n_m)\}_{k=1}^\infty$ such that $g_k(\cdot; \phi^n_m) \in \mathfrak{F}_c$ for any sequence $\phi^n_m \in (0,1)^{n-m+1}$, and

$$
\begin{aligned}
L^*_{m,n} &= \lim_{k\to\infty} \psi_\lambda\left(\bar{G}^{g_k(\cdot;\Phi^n_m)}_{m,n}\right) \\
&= \lim_{k\to\infty} \psi_\lambda\left(\bar{G}^{F_{\Theta|\Phi}(\cdot|\Phi_m)\circ g_k(\cdot;\Phi^n_m)}_{m+1,n}\right) \\
&\leq \sup_{g\in\mathfrak{F}_c} \psi_\lambda\left(\bar{G}^g_{m+1,n}\right) = L^*_{m+1,n}
\end{aligned}
$$

Therefore, for any $\nu > 0$ we have

$$
\begin{aligned}
\mathbb{P}\left(\max_{1\leq m\leq n} L_{m,(1+\alpha)n} > \nu\right) &\leq \mathbb{P}\left(\max_{1\leq m\leq n} L^*_{m,(1+\alpha)n} > \nu\right) \\
&= \mathbb{P}\left(L^*_{n,(1+\alpha)n} > \nu\right) \\
&\leq \nu^{-1}r(\alpha n)
\end{aligned}
$$

where we have used Lemmas 9 and 17 for the last inequality, noting that the former lemma holds for any IFS initialization. The proof now follows that of Lemma 18, with the proper minor modifications.   ∎

Lemma 26: Let $(P_X, P_{Y|X})$ satisfy (A1) and (A4). Then (44) holds, and for any rate $R < I(X;Y)$

$$
\lim_{\varepsilon\to 0} \limsup_{n\to\infty} \mathbb{P}\left(\bigcap_{k=1}^n \left\{\Theta_k - {}^-\Theta^{n,R}_k < \min\left(\varepsilon, \frac{\Theta_k}{2}\right)\right\}\right) = 0
$$

$$
\lim_{\varepsilon\to 0} \limsup_{n\to\infty} \mathbb{P}\left(\bigcap_{k=1}^n \left\{{}^+\Theta^{n,R}_k - \Theta_k < \min\left(\varepsilon, \frac{1-\Theta_k}{2}\right)\right\}\right) = 0. \quad (63)
$$

Furthermore, if P_X is also the unique input distribution for $P_{Y|X}$ such that $I(X;Y) = C(P_{Y|X})$, then

$$
\lim_{n\to\infty} n^{-1}\sum_{k=1}^n \eta(X_k) = \mathbb{E}(\eta(X)) \quad \text{a.s.} \quad (64)
$$

for any measurable $\eta : \mathcal{X} \mapsto \mathbb{R}$ satisfying $\mathbb{E}(|\eta(X)|) < \infty$.

Proof: Without the ergodicity property (A2), we cannot directly use the SLLN which was a key tool in deriving (44) and (48). Instead, we use the *ergodic decomposition* for Markov chains[24] [14, Section 5.3] to write the invariant distribution $P_{\Theta\Phi}$ as a mixture of ergodic distributions. We then apply the SLLN to each ergodic component, and use the maximality property (A4) to control the behavior of the chain within each component. For clarity of exposition, we avoid some of the more subtle measure theoretic details for which the reader is referred to [14].

Let $\mathcal{P}$ denote the Markov stochastic kernel associated with the posterior matching scheme. The ergodic decomposition implies that there exists a r.v. Γ taking values in $(0, 1)$, such that $\Gamma = \chi(\Theta)$ for some measurable function $\chi : (0,1) \mapsto (0,1)$, and $P_{\Theta\Phi|\Gamma}(\cdot|\gamma)$ is ergodic for $\mathcal{P}$, for P_Γ-a.a. γ. Let us first show that Φ and Γ are statistically independent. For any $S \in \mathfrak{B}$, it is clear that $P_{\Theta|\Gamma}(\cdot|S)$ is an invariant distribution for $\mathcal{P}$, being a mixture of ergodic distributions. Hence the set $\chi^{-1}(S)$ must be invariant for the posterior matching kernel, i.e.,

$$
F_{\Theta|\Phi}(\chi^{-1}(S)|\phi) = \chi^{-1}(S) \quad (65)
$$

[24] The chain has at least one invariant distribution, and evolves over a locally compact state space $(0,1)^2$, hence admits an ergodic decomposition.

up to a P_Θ-null set, for P_Φ-a.a. ϕ. Define $Z \overset{\text{def}}{=} F_{\Theta|\Phi}(\Theta|\Phi)$. For any $S, T \in \mathfrak{B}$:

$$
\begin{aligned}
P_{\Gamma\Phi}(S,T) = P_{\Theta\Phi}(\chi^{-1}(S),T) &\overset{(a)}{=} P_{Z\Phi}(\chi^{-1}(S),T) \\
&\overset{(b)}{=} P_Z(\chi^{-1}(S)) \cdot P_\Phi(T) \overset{(c)}{=} P_\Theta(\chi^{-1}(S)) \cdot P_\Phi(T) \\
&= P_\Gamma(S) P_\Phi(T)
\end{aligned}
$$

where (a) follows from (65) and the fact that (Z, Φ) a.s. determines Θ, (b) holds since Z is independent of Φ, and (c) holds since $Z \sim P_\Theta$ (i.e., uniform).

We can now apply the SLLN (Lemma 7) to each ergodic component $\chi^{-1}(\gamma)$. For P_Γ-a.a. γ and $P_{\Theta|\Gamma}(\cdot|\gamma)$-a.a. message points $\theta_0 \in \chi^{-1}(\gamma)$

$$
\begin{aligned}
&\lim_{n\to\infty} \frac{1}{n} \log f_{\Theta_0|\Phi^n}(\theta_0|\Phi^n) \\
&= \mathbb{E}\left(\log \frac{f_{\Phi|\Theta}(\Phi|\Theta)}{f_\Phi(\Phi)} \,\middle|\, \Gamma = \gamma \right) \quad \text{a.s.} \\
&= \mathbb{E}\left(\log \frac{f_{\Phi|\Theta\Gamma}(\Phi|\Theta,\gamma)}{f_{\Phi|\Gamma}(\Phi|\gamma)} \,\middle|\, \Gamma = \gamma \right) \\
&= I(\Theta; \Phi|\Gamma = \gamma).
\end{aligned} \tag{66}
$$

Now, for any γ

$$
I(\Theta; \Phi|\Gamma = \gamma) \le C(P_{\Phi|\Theta}) = C(P_{Y|X})
$$

where the inequality holds by the definition of the unconstrained capacity and since $\Gamma - \Theta - \Phi$ is a Markov chain, and the equality holds since the normalized channel preserves the mutual information (Lemma 12). Furthermore, using the independence of Φ and Γ and the Markov relation above again, together with property (A4), leads to

$$
I(\Theta; \Phi|\Gamma) = I(\Theta; \Phi) = C(P_{Y|X})
$$

Combining the above we conclude[25] that for P_Γ-a.a. γ

$$
I(\Theta; \Phi|\Gamma = \gamma) = C(P_{Y|X}) \tag{67}
$$

Substituting the above into (66) yields

$$
\lim_{n\to\infty} \frac{1}{n} \log f_{\Theta_0|\Phi^n}(\theta_0|\Phi^n) = C(P_{Y|X}) \quad \text{a.s.}
$$

for P_Θ-a.a. θ_0. This in turn implies (44).

Establishing (63) follows the same line of argument, proving a weaker version of (48). By the ergodic decomposition, for P_Γ-a.a. γ and $P_{\Theta|\Gamma}(\cdot|\gamma)$-a.a. message points $\theta_0 \in \chi^{-1}(\gamma)$

$$
\begin{aligned}
&\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^n \log {}^- f^\varepsilon_{\Phi|\Theta}(\Phi_k|\Theta_k) \\
&= \mathbb{E}\left(\log {}^- f^\varepsilon_{\Phi|\Theta}(\Phi|\Theta) \,\middle|\, \Gamma = \gamma \right) \quad \text{a.s.} \\
&\overset{\text{def}}{=} \mathcal{L}_\varepsilon(\gamma).
\end{aligned} \tag{68}
$$

The function $\mathcal{L}_\varepsilon(\gamma)$ satisfies

$$
\mathbb{E}\mathcal{L}_\varepsilon(\Gamma) = \mathbb{E}\left(\log {}^- f^\varepsilon_{\Phi|\Theta}(\Phi|\Theta) \right) = I^-_\varepsilon
$$

[25]Note that (67) does not hold in general if property (A4) is not satisfied, as there may be variations in the limiting values between ergodic components.

and since ${}^- f^\varepsilon_{\Phi|\Theta} \le f_{\Phi|\Theta}$, then

$$
\mathcal{L}_\varepsilon(\gamma) \le C(P_{Y|X})
$$

for P_Γ-a.a. γ. Now since $I(X;Y) = C(P_{Y|X})$ under property (A4), then

$$
\lim_{\varepsilon\to 0} I^-_\varepsilon = C(P_{Y|X}). \tag{69}
$$

It is therefore clear that for small ε values $\mathcal{L}_\varepsilon(\gamma)$ must be close to I^-_ε for a set of high P_Γ probability. Precisely

$$
A_{\varepsilon,\nu} \overset{\text{def}}{=} \{\gamma \in \text{supp}(\Gamma) :
$$
$$
\mathcal{L}_\varepsilon(\gamma) > I^-_\varepsilon - \nu^{-1}(C(P_{Y|X}) - I^-_\varepsilon)\}.
$$

Then

$$
\begin{aligned}
I^-_\varepsilon = \mathbb{E}\mathcal{L}_\varepsilon(\Gamma) \le\ & P_\Gamma(A_{\varepsilon,\nu}) C(P_{Y|X}) \\
&+ (1 - P_\Gamma(A_{\varepsilon,\nu}))(I^-_\varepsilon - \nu^{-1}(C(P_{Y|X}) - I^-_\varepsilon)).
\end{aligned}
$$

Rearranging, we get

$$
P_\Gamma(A_{\varepsilon,\nu}) \ge \frac{1}{1+\nu}. \tag{70}
$$

Combining (68), (69) and (70), we conclude that for any $\nu > 0$ and any $\varepsilon > 0$ small enough

$$
\mathbb{P}\left(\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^n \log {}^- f^\varepsilon_{\Phi|\Theta}(\Phi_k|\Theta_k) > I^-_\varepsilon - \frac{\delta(\varepsilon)}{\nu} \right)
$$
$$
\ge \frac{1}{1+\nu} \tag{71}
$$

for P_Θ-a.a. message points θ_0, where $\delta(\varepsilon) \to 0$ as $\varepsilon \to 0$. The remainder of the proof follows that of Lemma 19, with some minor adaptations.

Finally, suppose P_X is the unique capacity achieving input distribution for $P_{Y|X}$. For P_Γ-a.a. γ

$$
I(X;Y|\Gamma = \gamma) = I(\Theta; \Phi|\Gamma = \gamma) = C(P_{Y|X}). \tag{72}
$$

Thus, since $\Gamma - X - Y$ is a Markov chain and from the uniqueness of P_X as capacity achieving, it must be that $P_{X|\Gamma}(\cdot|\gamma) = P_X(\cdot)$ for P_Γ-a.a. γ. Applying the SLLN to each ergodic component, we find that for P_Γ-a.a. γ and $P_{\Theta|\Gamma}(\cdot|\gamma)$-a.a. message points $\theta_0 \in \chi^{-1}(\gamma)$

$$
\begin{aligned}
\lim_{n\to\infty} n^{-1} \sum_{k=1}^n \eta(X_k) &= \lim_{n\to\infty} n^{-1} \sum_{k=1}^n \eta\left(F_X^{-1}(\Theta_k) \right) \\
&= \mathbb{E}(\eta\left(F_X^{-1}(\Theta) \right)|\Gamma = \gamma) \quad \text{a.s.} \\
&= \mathbb{E}(\eta(X)|\Gamma = \gamma) = \mathbb{E}\eta(X)
\end{aligned}
$$

establishing (64).

Remark 8: It is instructive to point out that the proof of the Lemma holds also when property (A3) is not satisfied, namely when the posterior matching kernel has fixed points. In that case, each ergodic component must lie strictly inside an invariant interval (i.e., an interval between adjacent fixed points), which results in a decoding ambiguity as the receiver cannot distinguish between the ergodic components. As discussed in Section VII-A, this exact phenomena prevents any positive rate

from being achieved, and generally requires using a posterior matching variant. The fact that capacity is nonetheless achieved under $(A4)$ in the absence of fixed-points even when the chain is not ergodic, suggests that in this case almost any ergodic component, in addition to being capacity achieving in the sense of (67), is also dense in $(0, 1)$. The intuitive interpretation is that in that case any interval intersects with almost all of the ergodic components, hence the receiver, interested in decoding intervals, is "indifferent" to the specific component the chain lies in. ∎

APPENDIX B
POINTWISE ACHIEVABILITY PROOFS

Proof of Lemma 15: Property $(A5)$ implies in particular that $F_X(x), F_Y(y)$ are continuous and bijective over $\underline{\mathrm{supp}}(X), \underline{\mathrm{supp}}(Y)$ respectively, and that $F_{XY}(x|y)$ is jointly continuous in x, y over $\underline{\mathrm{supp}}(X, Y)$. The normalized posterior matching kernel is therefore given by

$$F_{\Theta|\Phi}(\theta|\phi) = F_{X|Y}\left(F_X^{-1}(\theta)\,\middle|\,F_Y^{-1}(\phi)\right)$$

and is jointly continuous in θ, ϕ over $\underline{\mathrm{supp}}(\Theta, \Phi)$ (note that for the family Ω_A this does not hold in general). Thus, property $(A3^*)$ implies (by continuity) that for any $\theta \in (0, 1)$ there exists some $\phi \in (0, 1)$ so that $F_{\Theta|\Phi}(\theta|\phi) \neq \theta$.[26]

We first show that the chain is $P_{\Theta\Phi}$-irreducible. Let $\mathfrak{B}$ here denote the usual Borel σ-algebra corresponding to the open unit interval. Since Θ_n is a deterministic function of $(\Theta_{n-1}, \Phi_{n-1})$, and since Φ_n is generated from Θ_n via a memoryless channel, it follows (by arguments similar to those given in the proof of Lemma 16) that to establish irreducibility it suffices to consider only the Θ_n component of the chain, and (since $P_{\Theta\Phi}$ has a proper p.d.f.) to show that any set $\Delta \in \mathfrak{B}$ with $\mathcal{U}(\Delta) > 0$ is reached in a finite time with a positive probability starting from any fixed message point $\Theta_0 = \theta_0 \in (0, 1)$.

Define the set mapping $\pi : \mathfrak{B} \mapsto \mathfrak{B}$

$$\pi(A) \stackrel{\text{def}}{=} \{\xi \in (0, 1) : \xi = F_{\Theta|\Phi}(\theta|\phi), \theta \in A,$$
$$\phi \in \underline{\mathrm{supp}}(\Phi \,|\, \Theta = \theta))\}$$

namely, the set of all points that are "reachable" from the set A in a single iteration. If A is an interval (or a single point), then $\pi(A)$ is also an interval, since it is a continuous image of the set $A' = \underline{\mathrm{supp}}(\Theta, \Phi) \cap \{A \times (0, 1)\}$, which by property $(A5)$ is a connected set[27]. For any $\theta_0 \in (0, 1)$ it holds that

[26] Note that continuity also implies there is an interval for which this holds, and since $\Phi \sim \mathcal{U}$, the stronger property $(A3)$ holds.

[27] This is proved as follows: since F_X, F_Y are continuous, the set $\underline{\mathrm{supp}}(\Theta, \Phi)$ inherits the properties of $\underline{\mathrm{supp}}(X, Y)$, namely it is connected (and open, hence path-connected) and convex in the ϕ-direction. Therefore, any two points in $a, b \in A'$ can be connected by a path in $\underline{\mathrm{supp}}(\Theta, \Phi)$. If this path does not lie entirely in A', then consider a new path that starts from a in a straight line connecting to the last point in the original path which has the same θ coordinate as a, then merges with the original path until reaching the first point with the same θ coordinate as b, and continuing in a straight line to b. Since $\underline{\mathrm{supp}}(\Theta, \Phi)$ is convex in the ϕ-direction this new path is completely within A'.

$\mathbb{E}(F_{\Theta|\Phi}(\theta_0|\Phi)) = \theta_0$, and together with property $(A3^*)$ it must also be that

$$\inf_{\phi \in \underline{\mathrm{supp}}(\Phi \,|\, \Theta = \theta_0)} F_{\Theta|\Phi}(\theta_0|\phi) < \theta_0$$
$$< \sup_{\phi \in \underline{\mathrm{supp}}(\Phi \,|\, \Theta = \theta_0)} F_{\Theta|\Phi}(\theta_0|\phi). \tag{73}$$

Thus, θ_0 is an interior point of the interval $\pi(\{\theta_0\})$. The arguments above regarding π can be applied to all points within the set $\pi(\{\theta_0\})$, and then recursively to obtain

$$\theta_0 \in \pi(\{\theta_0\}) \subseteq \pi^{(2)}(\{\theta_0\}) \subseteq \cdots \subseteq \pi^{(n)}(\{\theta_0\}) \subseteq \cdots \tag{74}$$

where $\pi^{(n)}$ is the n-fold iteration of π. Therefore, $\{\pi^{(n)}(\{\theta_0\})\}_{n=1}^\infty$ is a sequence of *expanding intervals* containing θ_0 as an interior point. Note also that $\pi^{(n)}(\{\theta_0\}) = \underline{\mathrm{supp}}(\Theta_n \,|\, \Theta_0 = \theta_0)$. Consider the set

$$A_{\theta_0} = \bigcup_{n=0}^\infty \pi^{(n)}(\{\theta_0\}).$$

Let us show that $A_{\theta_0} = (0, 1)$. First, it is easy to see that A_{θ_0} is an open interval, since it is a union of nested intervals, and if it had contained one of its endpoints then that endpoint would have been contained in $\pi^{(n)}(\{\theta_0\})$ for some n, which by the expansion property above is an interior point of $\pi^{(n+1)}(\{\theta_0\}) \subseteq A_{\theta_0}$, in contradiction. Now, suppose that $A_{\theta_0} = (\theta_1, \theta_2)$ for $\theta_1 > 0$. Using (73) and the continuity of $F_{\Theta|\Phi}(\theta|\phi)$ once again, we have

$$\lim_{\theta \to \theta_1^+} \inf_{\phi \in \underline{\mathrm{supp}}(\Phi \,|\, \Theta = \theta)} F_{\Theta|\Phi}(\theta|\phi)$$
$$= \inf_{\phi \in \underline{\mathrm{supp}}(\Phi \,|\, \Theta = \theta_1)} F_{\Theta|\Phi}(\theta_1|\phi) < \theta_1$$

in contradiction. The same argument applies for θ_2, establishing $A_{\theta_0} = (0, 1)$. As a result, for any set $\Delta \in \mathfrak{B}$ with $\mathcal{U}(\Delta) > 0$ we have that $\mathcal{U}(\Delta \cap \pi^{(n)}(\{\theta_0\})) \to \mathcal{U}(\Delta)$ as $n \to \infty$. Therefore, there exists a finite n for which $\mathcal{U}(\Delta \cap \pi^{(n)}(\{\theta_0\})) > 0$, and since $\mathcal{U} \ll P_{\Theta_n|\Theta_0}$ when restricted to $\pi^{(n)}(\{\theta_0\})$, it must be that $P_{\Theta_n|\Theta_0}(\Delta|\theta_0) > 0$. Thus, the normalized chain is $P_{\Theta\Phi}$-irreducible. It was already verified that $P_{\Theta\Phi}$ is an invariant distribution, hence by Lemma 4 the chain is also recurrent, $P_{\Theta\Phi}$ is unique and ergodic, and so property $(A2)$ holds.

Let $\mathcal{P}$ denote the stochastic kernel of our Markov chain. To establish p.h.r., we would like to use condition (i) of Lemma 5. However, Θ_{n+1} is a deterministic function of (Θ_n, Φ_n), and thus $\mathcal{P}(\cdot|(\theta, \phi)) \not\ll P_{\Theta\Phi}$ (as the former is supported on a $P_{\Theta\Phi}$-null set). Nevertheless, it is easy to see that due to the expansion property, the 2-skeleton of the chain (which is also recurrent with the same invariant distribution) admits a proper p.d.f. over a subset of $\underline{\mathrm{supp}}(\Theta, \Phi)$ and therefore $\mathcal{P}^2(\cdot|(\theta, \phi)) \ll P_{\Theta\Phi}$ for any $(\theta, \phi) \in \underline{\mathrm{supp}}(\Theta, \Phi)$. Thus, by condition (i) of Lemma 5 the 2-skeleton is p.h.r., which in turn implies the chain itself is p.h.r. via condition (ii) of Lemma 5.

To establish aperiodicity, we use the expansion property (74) once again. Suppose the chain has period $d > 1$ and let $\{D_i\}_{i=0}^{d-1}$ be the corresponding partition of the state space $\underline{\mathrm{supp}}(\Theta, \Phi)$. From our previous discussion we already know that for any

(θ_0, ϕ_0), the set $\underline{\mathrm{supp}}(\Theta_n | \Theta_0 = \theta_0, \Phi_0 = \phi_0)$ is an interval that expands into $(0, 1)$ as $n \to \infty$. Since we have the Markov relation $\Phi_n - \Theta_n - \Theta^{n-1}\Phi^{n-1}$, the set $\underline{\mathrm{supp}}(\Theta_n, \Phi_n | \Theta_0 = \theta_0, \Phi_0 = \phi_0)$ expands into $\underline{\mathrm{supp}}(\Theta, \Phi)$ in the sense that it contains any open subset of $\underline{\mathrm{supp}}(\Theta, \Phi)$ for any n large enough. Therefore, by definition of periodicity for any $n \in \mathbb{N}$ and $i \in \{0, \ldots, d-1\}$ we have $\mathbb{P}((\Theta_{nd+i}, \Phi_{nd+i}) \in D_i | (\Theta_0, \Phi_0) \in D_0) = 1$, and since $P_{\Theta\Phi} \ll \mathcal{U} \times \mathcal{U}$, then it must be that $(\mathcal{U} \times \mathcal{U})(\underline{\mathrm{supp}}(\Theta, \Phi) \setminus D_i) = 0$ for any $i \in \{0, 1, \ldots, d-1\}$. However, this cannot be satisfied by $d > 1$ disjoint sets. ∎

Lemma 27: Suppose $(P_X, P_{Y|X}) \in \Omega_C$. Then Lemmas 19 and 20 hold for any fixed message point $\Theta_0 = \theta_0 \in (0, 1)$. Furthermore, for any $\varepsilon > 0, \delta > 0$ and $\theta_0 \in (0, 1)$:

$$\lim_{n \to \infty} \mathbb{P}(^-\Theta_n^\delta > \varepsilon | \Theta_0 = \theta_0)$$
$$= \lim_{n \to \infty} \mathbb{P}(^+\Theta_n^\delta < 1 - \varepsilon | \Theta_0 = \theta_0)$$
$$= 0.$$

Proof: The proofs of Lemmas 19 and 20 remain virtually the same, only now using the SLLN for p.h.r. chains (Lemma 7) to obtain convergence for any fixed message point.

Since by Lemma 15 the normalized chain is p.h.r. and aperiodic, Lemma 6 guarantees that the marginal distribution converges to the invariant distribution $P_{\Theta\Phi}$ in total variation, for any initial condition and hence any fixed message point. Loosely speaking, we prove the result by reducing the fixed message point setting for large enough n, to the already analyzed case of a uniform message point in Lemma 18.

First, let $\{\tilde{\Phi}_n\}_{n=1}^\infty$ be a sequence of r.v.'s such that $P_{\tilde{\Phi}_n}$ tends to $\mathcal{U}$ in total variation. Then the result of Lemma 17 can be rewritten as

$$\lim_{n \to \infty} \mathbb{E}(\psi_\lambda [F_{\Theta|\Phi}(\cdot | \tilde{\Phi}_n) \circ h]) \le \xi(\psi_\lambda(h)) \qquad (75)$$

which holds since the expectation is taken over a bounded function.

Now, consider the k-fold chain $\{\Theta_n^{n+k-1}, \Phi_n^{n+k-1}\}_{n=1}^\infty$ for some fixed k. It is immediately seen that this chain is also p.h.r., and its invariant distribution is $P_{\Theta\Phi}^k$, the k-fold cartesian product of $P_{\Theta\Phi}$. Thus, by Lemma 6 the k-fold chain approaches this invariant distribution in total variation for any initial condition. In particular, this implies that

$$\lim_{n \to \infty} d_{TV}\left(P_{\Phi_n^{n+k-1} | \Theta_0}(\cdot | \theta_0), \mathcal{U}^k\right) = 0$$

where $\mathcal{U}^k$ is the k-fold cartesian product of $\mathcal{U}$. Namely, the distribution of k consecutive outputs tends to i.i.d. uniform in total variation. Using (75) and a trivial modification of Lemma 9 for an asymptotically i.i.d. control sequence, we have that for any fixed k

$$\lim_{n \to \infty} \mathbb{P}\left(\psi_\lambda(\bar{G}_n(\theta)) > \nu | \Theta_0 = \theta_0\right) \le \frac{1}{\nu} r(k) \qquad (76)$$

where $r(\cdot)$ is the decay profile of ξ. Let n_k be the smallest integer such that for any $n \ge n_k$

$$\mathbb{P}(\psi_\lambda(\bar{G}_n(\theta)) > \nu | \Theta_0 = \theta_0) \le \frac{1}{\nu}\sqrt{r(k)}$$

holds, which must exist by (76). Then

$$\lim_{k \to \infty} \mathbb{P}\left(\psi_\lambda\left(\bar{G}_{n_k}(\theta)\right) > \nu | \Theta_0 = \theta_0\right) \le \frac{1}{\nu} \lim_{k \to \infty} \sqrt{r(k)} = 0.$$

Now, the proof of the Lemma follows through by working with (k, n_k) in lieu of n, and in (43) using the fact that the distribution of $\bar{G}_n(\theta_0)$ tends to $\mathcal{U}$ in total variation. ∎

Proof of Theorem 8: Let us first make the distinction between the Markov chain generated by the posterior matching scheme for $(P_X, P_{Y|X})$ when operating over the channel $P_{Y|X}$, according to whose law the transmitter and receiver encode/decode, and the chain generated by the same scheme when operating over the channel $P_{Y^*|X^*}$, which describes what actually takes place during transmission. We refer to the former as the *primary chain* denoting its input/output sequence as usual by (X_n, Y_n), and to the latter as the *mismatch chain*, denoting its input/outptut sequence by (X_n^*, Y_n^*). The same monikers and notations are used for the normalized counterparts.

Property $(C5)$ guarantees that the expansion property holds for the mismatch chain, and since by Property $(C3)$ $P_{X^*Y^*}$ is an invariant distribution, a similar derivation as in Lemma 15 implies that the mismatch chain is p.h.r., which in particular also guarantees the uniqueness of $P_{X^*Y^*}$. We would now like to obtain an analogue of Lemma 19. Let us expand posterior p.d.f. w.r.t. the primary chain, using the fact that it induces an i.i.d. output distribution is (this does not necessarily hold for the mismatch chain) and the channel is memoryless

$$f_{\Theta_0|Y^n}(\theta|y^n) = \frac{f_{Y_n|\Theta_0, Y^{n-1}}(y_n|\theta, y^{n-1})}{f_{Y_n|Y^{n-1}}(y_n|y^{n-1})} f_{\Theta_0|Y^{n-1}}(\theta|y^{n-1})$$
$$= \frac{f_{Y|X}(y_n|g_n(\theta, y^{n-1}))}{f_Y(y_n)} f_{\Theta_0|Y^{n-1}}(\theta|y^{n-1}).$$

Applying the recursion rule n times, taking a logarithm and evaluating the above at the message point, we obtain

$$\frac{1}{n}\log f_{\Theta_0|Y^n}(\theta|y^n) = \frac{1}{n}\sum_{k=1}^n \log \frac{f_{Y|X}(y_k|g_k(\theta, y^{k-1}))}{f_Y(y_k)}$$

Now we can evaluate this posterior of the primary chain using the inputs/outputs of the mismatch chain, and apply the p.h.r. SLLN (Lemma 19) for the mismatch chain using its invariant distribution $P_{\Theta^*\Phi^*}$

$$\lim_{n \to \infty} \frac{1}{n}\log f_{\Theta_0|Y^n}(\Theta_0|Y^{*n})$$
$$= \lim_{n \to \infty} \frac{1}{n}\sum_{k=1}^n \log \frac{f_{Y|X}(Y_k^*|g_k(\Theta_0, Y^{*k-1}))}{f_Y(Y_k^*)}$$
$$\stackrel{(a)}{=} \lim_{n \to \infty} \frac{1}{n}\sum_{k=1}^n \log \frac{f_{Y|X}(Y_k^*|X_k^*)}{f_Y(Y_k^*)}$$
$$= \mathbb{E}\left(\log \frac{f_{Y|X}(Y^*|X^*)}{f_Y(Y^*)}\right) \qquad \mathcal{P}_{\theta_0}^* - \text{a.s.}$$
$$= \mathbb{E}\left(\log \frac{f_{Y|X}(Y^*|X^*)}{f_{Y^*|X^*}(Y^*|X^*)} + \log \frac{f_{Y^*}(Y^*)}{f_Y(Y^*)}\right.$$
$$\left. + \log \frac{f_{Y^*|X^*}(Y^*|X^*)}{f_{Y^*}(Y^*)}\right)$$
$$\stackrel{(b)}{=} I(X^*; Y^*) - \left(D(P_{Y^*|X^*}\|P_{Y|X}|P_{X^*}) - D(P_{Y^*}\|P_Y)\right)$$
$$\stackrel{\text{def}}{=} R^{\text{mis}}(X, Y; X^*, Y^*)$$

where in (a) we used the definition of the channel input, and in (b) the finiteness of the terms is guaranteed by Property (C2) and the fact that $(P_{X^*}, P_{Y^*|X^*})$ is an input/channel pair. The same analysis using normalized chains results in

$$\lim_{n\to\infty} \frac{1}{n} \log f_{\Theta_0|\Phi^n}(\Theta_0|\Phi^{*n}) = \mathbb{E}\log f_{\Phi|\Theta}(\Phi^*|\Theta^*)$$
$$= R^{\mathrm{mis}} \quad P_{\theta_0}-\text{a.s.}$$

where the last equality is due to the invertibility of the chain normalization, which is guaranteed by property (A5). Now we can define the analogue of I_ε^- in (47) as follows:

$$R_\varepsilon^{\mathrm{mis}} \stackrel{\text{def}}{=} \mathbb{E}\log{}^- f_{\Phi|\Theta}^\varepsilon(\Phi^*|\Theta^*)$$

Therefore

$$0 \leq R^{\mathrm{mis}} - R_\varepsilon^{\mathrm{mis}}$$
$$= D\left(P_{\Phi^*|\Theta^*} \left\|{}^- P_{\Phi|\Theta}^\varepsilon \,|P_\Theta\right.\right) - D(P_{\Phi^*|\Theta^*}\|P_{\Phi|\Theta}|P_\Theta)$$

The second term on the right-hand-side above is finite due to Property (C2), and by Property (C1) we have that $\inf_{\varepsilon>0} D(P_{\Phi^*|\Theta^*}\|{}^- P_{\Phi|\Theta}^\varepsilon |P_\Theta) < \infty$. Thus, for any $\varepsilon > 0$ small enough

$$-\infty < R_\varepsilon^{\mathrm{mis}} \leq R^{\mathrm{mis}}.$$

We can now continue as in the proof of Lemma (20), to show that (46) holds in this case for any rate $R < R^{\mathrm{mis}}$.

The contraction Property (C4) implies the equivalent of Lemma 18 for the mismatch chain, since although the output sequence Y_n^* is not necessarily i.i.d. even when we start in the invariant distribution, we have a contraction uniformly given any conditioning. Tied together with the above and repeating the last steps of Theorem 4, the achievability of (56) is established. ∎

APPENDIX C
MISCELLANEOUS PROOFS

Proof of Lemma 2: For simplicity we assume that f_X is symmetric around its maximum, the general unimodal case follow through essentially the same way. Since the property of having a regular tail is shift invariant, we can further assume without loss of generality that f_X attains its maximum at (and is symmetric around) $x = 0$.

(i) By the assumption, there exist $m_0, m_1 > 0$, $b \geq a > 1$ and $x_0 > 1$ so that for any $|x| > x_0$

$$m_0|x|^{-b} \leq f_X(x) \leq m_1|x|^{-a}$$

Thus, for any $x > x_0$

$$1 - F_X(x) \leq m_1 \int_x^\infty y^{-a} dy = \frac{m_1}{a-1} x^{1-a}$$
$$\leq \frac{m_0^{\frac{1-a}{b}} m_1}{1-a} f_X^{\frac{a-1}{b}}(x)$$

and similarly

$$1 - F_X(x) \geq \frac{m_1^{\frac{1-b}{a}} m_0}{1-b} f_X^{\frac{b-1}{a}}(x).$$

Identical derivations hold for $F_X(x)$ and $x < -x_0$, and thus setting $\gamma = 1 - F_X(x_0)$ the tail regularity is established.

(ii) By the assumption, there exist $0 < m_0 < m_1$, $a \geq 1$, $b > 0$ and $x_0 > 1$ so that for any $|x| > x_0$

$$m_0 e^{-b|x|^a} \leq f_X(x) \leq m_1 e^{-b|x|^a}.$$

Thus, for any $x > x_0$

$$1 - F_X(x) \leq m_1 \int_x^\infty e^{-by^a} dy$$
$$\leq m_1 \int_x^\infty \left(\frac{y}{x}\right)^{a-1} e^{-by^a} dy$$
$$\stackrel{(z=y^a)}{\leq} m_1 \int_{x^a}^\infty \frac{1}{ax^{a-1}} e^{-bz} dz$$
$$= \frac{m_1}{abx^{a-1}} e^{-bx^a}$$
$$\leq \frac{m_1}{m_0 ab} f_X(x)$$

and on the other hand

$$(1 - F_X(x))(ab + (a-1)x^{-a})$$
$$\geq m_0 \int_x^\infty (ab + (a-1)x^{-a}) e^{-by^a} dy$$
$$\geq m_0 \int_x^\infty (ab + (a-1)y^{-a}) e^{-by^a} dy$$
$$\stackrel{(a)}{=} -m_0 \frac{e^{-by^a}}{y^{a-1}} \bigg|_x^\infty$$
$$= m_0 \frac{e^{-bx^a}}{x^{a-1}}$$

where (a) is easily verified by differentiation. Thus, for any $x > x_0$

$$1 - F_X(x) \geq \frac{m_0 x}{abx^a + a - 1} e^{-bx^a} \geq m_0 m_1^{-\frac{\beta}{b}} f_X^{\frac{\beta}{b}}(x)$$

where the last inequality holds for $x > x_0$ with suitable selection of $\beta > b$. Identical derivations hold for $F_X(x)$ and $x < -x_0$, and thus setting $\gamma = 1 - F_X(x_0)$ the tail regularity is established. ∎

Proof of Lemma 13:

(i) Let $0 < M \stackrel{\text{def}}{=} \inf_{\underline{\mathrm{supp}}(\Theta,\Phi)} f_{\Phi|\Theta}(\phi|\theta)$. Since $\underline{\mathrm{supp}}(\Theta, \Phi)$ is convex in the θ-direction and $f_{\Phi|\Theta}(\phi|\theta) = f_{\Theta|\Phi}(\theta|\phi)$, we have that ${}^- f_{\Phi|\Theta}^\varepsilon(\phi|\theta) \geq M$ over $\underline{\mathrm{supp}}(\Theta, \Phi)$. Therefore:

$$0 \leq D(P_{\Phi|\Theta}\|{}^- P_{\Phi|\Theta}^\varepsilon|P_\Theta)$$
$$= \iint_{\underline{\mathrm{supp}}(\Theta,\Phi)} f_{\Phi|\Theta}(\phi|\theta) \log \frac{f_{\Phi|\Theta}(\phi|\theta)}{{}^- f_{\Phi|\Theta}^\varepsilon(\phi|\theta)} d\theta d\phi$$
$$\leq \iint_{\underline{\mathrm{supp}}(\Theta,\Phi)} f_{\Phi|\Theta}(\phi|\theta) \log \frac{f_{\Phi|\Theta}(\phi|\theta)}{M} d\theta d\phi$$
$$= -(h(\Phi|\Theta) + \log M) < \infty$$

where in the last inequality we used the finiteness of the joint entropy and $\Theta \sim \mathcal{U}$. The same holds for $D(P_{\Phi|\Theta}\|{}^+ P_{\Phi|\Theta}^\varepsilon|P_\Theta)$, concluding the proof.

(ii) Since both F_X, F_Y are now bijective, we have that

$$F_{\Phi|\Theta}(\phi|\theta) = F_{Y|X}\left(F_Y^{-1}(\phi) \,|F_X^{-1}(\theta)\right)$$

and thus

$$f_{\Phi|\Theta}(\phi|\theta) = \frac{\partial}{\partial \phi}\left(F_{Y|X}\left(F_Y^{-1}(\phi)\,\big|\,F_X^{-1}(\theta)\right)\right)$$

$$= \frac{f_{Y|X}\left(F_Y^{-1}(\phi)\,\big|\,F_X^{-1}(\theta)\right)}{f_Y\left(F_Y^{-1}(\phi)\right)}$$

$$= \frac{f_{X|Y}\left(F_X^{-1}(\theta)\,\big|\,F_Y^{-1}(\phi)\right)}{f_X(F_X^{-1}(\theta))}$$

We can therefore write

$$^-f_{\Phi|\Theta}^\varepsilon(\phi|\theta) = \inf_{\xi \in J_\varepsilon^-(\phi,\theta)} \frac{f_{X|Y}\left(F_X^{-1}(\xi)\,\big|\,F_Y^{-1}(\phi)\right)}{f_X\left(F_X^{-1}(\xi)\right)}$$

$$\geq m^{-1}\cdot \inf_{\xi \in J_\varepsilon^-(\phi,\theta)} f_{X|Y}\left(F_X^{-1}(\xi)\,\big|\,F_Y^{-1}(\phi)\right)$$

where $m \overset{\text{def}}{=} \sup f_X(x) < \infty$. Denote the max-to-min ratio bound by

$$M = \sup_{y \in \underline{\text{supp}}(Y)} \left(\frac{\sup_{x \in \underline{\text{supp}}(X|Y=y)} f_{X|Y}(x|Y=y)}{\inf_{x \in \underline{\text{supp}}(X|y)} f_{X|Y}(x|y)}\right).$$

The relative entropy $D(f_{\Phi|\Theta}||^-f_{\Phi|\Theta}^\varepsilon)$ is now upper bounded as follows:

$$D\left(P_{\Phi|\Theta}\,\Big\|\,^-P_{\Phi|\Theta}^\varepsilon\,\Big|\,P_\Theta\right)$$

$$\leq \iint_{\underline{\text{supp}}(\Theta,\Phi)} d\theta d\phi \frac{f_{X|Y}\left(F_X^{-1}(\theta)\,\big|\,F_Y^{-1}(\phi)\right)}{f_X\left(F_X^{-1}(\theta)\right)}$$

$$\cdot \log \frac{m\cdot f_{X|Y}\left(F_X^{-1}(\theta)\,\big|\,F_Y^{-1}(\phi)\right)}{f_X\left(F_X^{-1}(\theta)\right)\cdot \inf_{\xi \in J_\varepsilon^-(\phi,\theta)} f_{X|Y}\left(F_X^{-1}(\xi)\right)\,\big|\,F_Y^{-1}(\phi)\right)}$$

$$= \log(m) + \iint_{\underline{\text{supp}}(X,Y)} dx dy\, f_{X|Y}(x|y) f_Y(y)$$

$$\cdot \log\left(\frac{1}{f_X(x)}\cdot \frac{f_{X|Y}(x|y)}{\inf_{z \in \widehat{J}_\varepsilon^-(y,x)} f_{X|Y}(z|y)}\right)$$

$$\leq \log(m) + h(X) + \log M < \infty \qquad (77)$$

where a straightforward change of variables was performed, and $\widehat{J}_\varepsilon^-(y,x)$ is the counterpart of $J_\varepsilon^-(\phi,\theta)$. In the last inequality we used the fact that $\underline{\text{supp}}(X,Y)$ is convex in the y-direction, which implies that $\widehat{J}_\varepsilon^-(y,x) \subseteq \underline{\text{supp}}(X|Y=y)$. Furthermore, $h(X)$ is finite since f_X is proper and bounded.

(iii) We prove the claim under the lenient assumption that $f_{X|Y}(x|y)$ is also symmetric for any fixed y. The argument for the general claim is a similar yet more tedious version of this proof. We need the following Lemma:

Lemma 28: Suppose X is proper with a symmetric unimodal p.d.f., a finite variance σ^2, and a regular tail with parameters γ, c_i, α_i. Define

$$f_X^*(x) \overset{\text{def}}{=} \inf_{z \in (e(x),x)} f_X(x), \quad e(x) \overset{\text{def}}{=} F_X^{-1}\left(\frac{1}{2}F_X(x)\right)$$

and let

$$\gamma^* \overset{\text{def}}{=} \min\left(\gamma, \frac{1}{3}\right), \quad M \overset{\text{def}}{=} \sup f_X(x),$$

$$M_1 \overset{\text{def}}{=} \frac{\gamma^*}{4}\left(\sqrt{\frac{2}{\gamma^*}}\sigma - \frac{1-\gamma^*}{2M}\right).$$

Then

$$D(f_X||f_X^*) \leq \alpha_1^{-1}\log\frac{2c_1}{c_0}$$

$$+ (1 + \alpha_1 - \alpha_0)\log M + \log M_1$$

Proof: Without loss of generality we can assume that $\gamma < \frac{1}{3}$, since a larger value implies a regular tail for any smaller value. Define $x_2 < x_1 < x_0 < 0$ to be

$$x_0 = F_X^{-1}(\gamma), \quad x_1 = F_X^{-1}\left(\frac{\gamma}{2}\right), \quad x_2 = F_X^{-1}\left(\frac{\gamma}{4}\right).$$

It is easy to see that $e(x_0) = x_1$ and $e(x_1) = x_2$. Defining $M = \sup f_X(x)$ we can lower bound $|x_1|$ using symmetry:

$$2|x_1|M \geq 1 - \gamma \quad \Rightarrow \quad |x_1| \geq \frac{1-\gamma}{2M}.$$

Using Chebyshev's inequality and symmetry, we can upper bound $|x_2|$ by

$$2\int_{|x_2|}^\infty f_X(x)dx = \frac{\gamma}{2} \leq \frac{\sigma^2}{x_2^2} \quad \Rightarrow \quad |x_2| \leq \sqrt{\frac{2}{\gamma}}\sigma$$

Combining the above and using the monotonicity of f for $x < 0$, we have

$$f_X(x_1)\cdot(|x_2| - |x_1|) \geq \frac{\gamma}{4}$$

which yields a lower bound for $f_X(x_1)$

$$f_X(x_1) \geq \frac{\gamma}{4(|x_2|-|x_1|)} \geq \frac{\gamma}{4}\left(\sqrt{\frac{2}{\gamma}}\sigma - \frac{1-\gamma}{2M}\right) = M_1$$

and since f_X is symmetric and unimodal and by the assumption $\gamma < \frac{1}{3}$, it is readily verified that

$$\begin{array}{ll} f_X^*(x) = f_X(e(x)) & x \in (-\infty, x_0) \\ f_X^*(x) \geq f_X(x_1) \geq M_1 & x \in (x_0, |x_0|) \\ f_X^*(x) = f_X(x) & x \in (|x_0|, \infty). \end{array} \qquad (78)$$

Now, recall that f has a regular tail, which is this symmetric case means that (recall that $x_0 < 0$)

$$c_0 f^{\alpha_0}(x) \leq F(x) \leq c_1 f^{\alpha_1}(x) \quad |x| > |x_0|$$

Let us upper bound the relative entropy between f_X, f_X^* using the above, together with (78), as

$$
\begin{aligned}
D(f_X \| f_X^*) \\
= \int_{-\infty}^{x_0} f_X(x) \log \frac{f_X(x)}{f_X^*(x)} dx \\
+ \int_{x_0}^{|x_0|} f_X(x) \log \frac{f_X(x)}{f_X^*(x)} dx \\
+ \int_{|x_0|}^{\infty} f_X(x) \log \frac{f_X(x)}{f_X^*(x)} dx \\
\leq \int_{-\infty}^{x_0} f_X(x) \log \frac{f_X(x)}{f_X(e(x))} dx \\
+ \int_{x_0}^{|x_0|} f_X(x) \log \frac{M}{M_1} dx \\
+ \int_{|x_0|}^{\infty} f_X(x) \log 1 \, dx \\
\leq \alpha_1^{-1} \int_{-\infty}^{x_0} f_X(x) \log \left(\frac{f_X^{\alpha_1}(x)}{F_X(x)} \frac{2 F_X(e(x))}{f_X^{\alpha_1}(e(x))} \right) dx \\
+ \log M - \log M_1 \\
\leq \alpha_1^{-1} \int_{-\infty}^{x_0} f_X(x) \log \left(\frac{f_X^{\alpha_1}(x)}{c_0 f_X^{\alpha_0}(x)} \frac{2 c_1 f_X^{\alpha_1}(e(x))}{f_X^{\alpha_1}(e(x))} \right) dx \\
+ \log M - \log M_1 \\
\leq \alpha_1^{-1} \log \frac{2 c_1}{c_0} + (1 + \alpha_1 - \alpha_0) \log M - \log M_1 \\
< \infty.
\end{aligned}
$$

∎

Returning to the pursued claim, let γ, c_i, α_i be the common tail parameters of $f_{X|Y}(\cdot|y)$, let $M = \sup f_{X|Y}(x|y)$ and let σ^2 be an upper bound on the variance of $f_{X|Y}(\cdot|y)$ for all y. It follows from definition that for any y

$$
\inf_{z \in \widehat{J_\varepsilon^-}(y,x)} f_{X|Y}(z|y) \geq f_{X|Y}^*(x|y)
$$

where $f_{X|Y}^*$ is defined as in Lemma 28. We now follow the derivations of the previous claim (ii) up to (77), and use the above inequality and Lemma 28 to obtain:

$$
\begin{aligned}
D \left(P_{\Phi|\Theta} \left\| ^- P_{\Phi|\Theta}^\varepsilon \right| P_\Theta \right) \\
\leq \log(m) + h(X) \\
+ \int_{\underline{\mathrm{supp}}(Y)} f_Y(y) dy \int_{\underline{\mathrm{supp}}(X|Y=y)} f_{X|Y}(x|y) \\
\times \log \frac{f_{X|Y}(x|y)}{f_{X|Y}^*(x|y)} dx \\
= \log(m) + h(X) \\
+ \int_{\underline{\mathrm{supp}}(Y)} f_Y(y) D(f_{X|Y}(\cdot|y) \| f_{X|Y}^*(\cdot|y)) dy \\
= \log(m) + h(X) + \alpha_1^{-1} \log \frac{2 c_1}{c_0} \\
+ (1 + \alpha_1 - \alpha_0) \log M - \log M_1 < \infty.
\end{aligned}
$$

The same proof holds for $D(P_{\Phi|\Theta} \|^+ P_{\Phi|\Theta}^\varepsilon | P_\Theta)$.

(iv) A direct consequence of (i).

∎

REFERENCES

[1] C. E. Shannon, "The zero-error capacity of a noisy channel," *IRE Trans. Inf. Theory*, vol. IT-2, pp. 8–19, Sep. 1956.

[2] T. T. Kadota, M. Zakai, and J. Ziv, "Capacity of a continuous memoryless channel with feedback," *IEEE Trans. Inf. Theory*, vol. IT-17, pp. 372–378, Jul. 1971.

[3] M. Horstein, "Sequential transmission using noiseless feedback," *IEEE Trans. Inf. Theory*, vol. IT-9, pp. 136–143, Jul. 1963.

[4] J. P. M. Schalkwijk, "A class of simple and optimal strategies for block coding on the binary symmetric channel with noiseless feedback," *IEEE Trans. Inf. Theory*, vol. IT-17, no. 3, pp. 283–287, May 1971.

[5] J. P. M. Schalkwijk and K. A. Post, "On the error probability for a class of binary recursive feedback strategies," *IEEE Trans. Inf. Theory*, vol. IT-19, pp. 498–511, Jul. 1973.

[6] M. Horstein, Sequential Transmission of Digital Information With Feedback Research Laboratory of Electronics, Cambridge, MA, Tech. Rep. 375, 1960, MIT.

[7] J. P. M. Schalkwijk and T. Kailath, "A coding scheme for additive noise channels with feedback part I: No bandwidth constraint," *IEEE Trans. Inf. Theory*, vol. IT-12, pp. 172–182, Apr. 1966.

[8] J. P. M. Schalkwijk, "A coding scheme for additive noise channels with feedback part II: Band-limited signals," *IEEE Trans. Inf. Theory*, vol. IT-12, pp. 183–189, Apr. 1966.

[9] Y. H. Kim, A. Lapidoth, and T. Weissman, "On reliability of Gaussian channels with noisy feedback," in *Proc. 44th Allerton Conf. Communication, Control, and Computing*, Sep. 2006.

[10] A. Sahai and S. Mitter, "The necessity and sufficiency of anytime capacity for stabilization of a linear system over a noisy communication link—Part I: Scalar systems," *IEEE Trans. on Inf. Theory*, vol. 52, no. 8, pp. 3369–3395, Aug. 2006.

[11] R. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.

[12] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[13] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*. Berlin/New York: Springer–Verlag, 1993.

[14] O. Hernández-Lerma and J. B. Lasserre, *Markov Chains and Invariant Probabilities*. Berlin: Birkhäuser Verlag, 2003.

[15] Y. Kifer, *Ergodic Theory of Random Transformations*. Boston, MA: Birkhäuser, 1986.

[16] P. Diaconis and D. Freedman, "Iterated random functions," *SIAM Rev.*, vol. 41, no. 1, pp. 45–76, 1999.

[17] D. Steinsaltz, "Locally contractive iterated function systems," *Ann. of Prob.*, vol. 27, no. 4, pp. 1952–1979, Oct. 1999.

[18] Ö. Stenflo, "Ergodic theorems for markov chains represented by iterated function systems," *Bull. Polish Acad. Sci. Math.*, vol. 49, no. 1, pp. 27–43, 2001.

[19] S. Verdú, "The exponential distribution in information theory," *Prob. Inform. Transmiss.*, vol. 32, no. 1, pp. 86–95, Jan.–Mar. 1996.

[20] A. N. Kolmogorov and S. V. Fomin, *Introductory Real Anal.*. New York: Dover Publications, 1975.

[21] A. Kramer, "Improving communication reliability by use of an intermittent feedback channel," *IEEE Trans. Inf. Theory*, vol. IT-15, no. 1, pp. 52–60, Jan. 1969.

[22] R. G. Gallager and B. Nakiboglu, "Variations on a theme by Schalkwijk and Kailath," *IEEE Trans. Inf. Theory*, vol. 56, no. 1, pp. 6–17, Jan. 2010.

[23] T. Weissman, "Robustness and sensitivity of the Schalkwijk-Kailath scheme," *The Kailath Colloq.*, 2006.

[24] R. Zamir and U. Erez, "A Gaussian input is not too bad," *IEEE Trans. Inf. Theory*, vol. 50, no. 6, pp. 1362–1367, Jun. 2004.

[25] J. M. Ooi and G. W. Wornell, "Fast iterative coding techniques for feedback channels," *IEEE Trans. Inf. Theory*, vol. IT-44, pp. 2960–2976, Nov. 1998.

[26] A. Tchamkerten and I. E. Telatar, "Variable length coding over unknown channels," *IEEE Trans. Inf. Theory*, vol. 52, no. 5, pp. 2126–2145, May 2006.

[27] M. Feder and A. Lapidoth, "Universal decoding for channels with memory," *IEEE Trans. Inf. Theory*, vol. 44, no. 5, pp. 1726–1745, Sep. 1998.

[28] O. Shayevitz and M. Feder, "Achieving the empirical capacity using feedback: Memoryless additive models," *IEEE Trans. Inf. Theory*, vol. 55, no. 3, pp. 1269–1295, Mar. 2009.

[29] O. Shayevitz and M. Feder, "Communication with feedback via posterior matching," in *Proc. Int. Symp. Information Theory*, Jun. 2007.

[30] O. Shayevitz and M. Feder, "The posterior matching feedback scheme: Capacity achieving and error analysis," in *Proc. Int. Symp. Information Theory*, Jul. 2008.

[31] J. Massey, "Causality, feedback and directed information," in *Proc. Int. Symp. Information Theory and Its Applications*, 1990.

[32] S. C. Tatikonda, "Control Under Communication Constraints," doctoral dissertation, , Sep. 2000, MIT.

[33] Y. H. Kim, "A coding theorem for a class of stationary channels with feedback," *IEEE Trans. Inf. Theory*, vol. 54, no. 4, pp. 1488–1499, Apr. 2008.

[34] S. Tatikonda and S. Mitter, "The capacity of channels with feedback," *IEEE Trans. Inf. Theory*, vol. 55, no. 1, pp. 323–349, Jan. 2009.

[35] H. H. Permuter, T. Weissman, and A. J. Goldsmith, "Finite state channels with time-invariant deterministic feedback," *IEEE Trans. Inf. Theory*, vol. 55, no. 2, pp. 644–662, Feb. 2009.

[36] Y.-.H. Kim, "Feedback capacity of stationary Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 56, no. 1, pp. 57–85, Jan. 2010.

[37] P. Walters, *An Introduction to Ergodic Theory*.   New York: Springer-Verlag, 1982.

Ofer Shayevitz received the B.Sc. degree (*summa cum laude*) from the Technion Institute of Technology, Haifa, Israel, in 1997 and the M.Sc. and Ph.D. degrees from the Tel-Aviv University, Tel Aviv, Israel, in 2004 and 2009, respectively, all in electrical engineering,

He is currently a Postdoctoral Fellow in the Information Theory and Applications (ITA) Center, University of California at San Diego. Prior to his graduate studies, he served as an Engineer and Team Leader in the Israeli Defense Forces (1997–2003), and as an Algorithms Engineer with CellGuide, a high-tech company developing low-power GPS navigation solutions (2003–2004).

Dr. Shayevitz is the recipient of the ITA Postdoctoral Fellowship (2009-2011), the Adams Fellowship (2006-2008) awarded by the Israel Academy of Sciences and Humanities, the Advanced Communication Center (ACC) Feder Family award for an outstanding Ph.D. thesis (2009), and the Weinstein Prize (2006-2009) for research and publications in signal processing.

Meir Feder received the B.Sc. and M.Sc. degrees from Tel-Aviv University, Tel Aviv, Israel, in 1980 and 1984, respectively, and the D.Sc. degree from the Massachusetts Institute of Technology (MIT), Cambridge, MA, and the Woods Hole Oceanographic Institution, Woods Hole, MA, in 1987, all in electrical engineering.

After being a Research Associate and Lecturer in MIT, he joined the Department of Electrical Engineering—Systems, Tel-Aviv University, where he is now a Professor. He had visiting appointments at the Woods Hole Oceanographic Institution, Scripps Institute, Bell Laboratories, and was a Visiting Professor at MIT during 1995/1996. He is also extensively involved in the high-tech industry and cofounded several companies, including Peach Networks, a developer of a unique server-based interactive TV solution which was acquired on March 2000 by Microsoft, and Amimon, a leading provider of ASICs for wireless high-definition A/V connectivity at the home.

Dr. Feder is a co-recipient of the 1993 IEEE Information Theory Best Paper Award. He also received the 1978 "creative thinking" award of the Israeli Defense Forces, the 1994 Tel-Aviv University prize for Excellent Young Scientists, the 1995 Research Prize of the Israeli Electronic Industry, and the research prize in applied electronics of the Ex-Serviceman Association, London, awarded by Ben-Gurion University. Between June 1993–June 1996, he served as an Associate Editor for Source Coding of the IEEE TRANSACTIONS ON INFORMATION THEORY.