The Posterior Matching Feedback Scheme: Capacity Achieving and Error Analysis

Ofer Shayevitz and Meir Feder
Department of EE-Systems, Tel Aviv University
Tel Aviv, Israel 69978
Email: {ofersha , meir}@eng.tau.ac.il

***Abstract*— Recently, we have introduced a sequential communication scheme for general memoryless channels with feedback based on the idea of *posterior matching*, providing a unified framework in which the known Horstein and Schalkwijk-Kailath schemes are special cases. In this paper, we show that the posterior matching scheme achieves the mutual information for a large family of channels and input distributions, and provide closed-form expressions for the attainable error probability over a range of rates. Moreover, we derive the achievable rates in a mismatched setting, where the scheme is designed according to the wrong channel model. In particular, our results hold for discrete memoryless channels, thereby confirming a longstanding conjecture that the Horstein scheme achieves capacity. The proof techniques employed utilize novel relations between information rates and convergence properties of iterated function systems.**

I. Introduction

Feedback is an invaluable resource for enhancing the performance of communication systems. Although falling short in increasing the capacity of memoryless channels, feedback can significantly reduce the complexity required to attain capacity, and boost the error performance simultaneously. Among the most notable feedback schemes possessing such merits are the Horstein scheme for the *Binary Symmetric Channel*[1] *(BSC)* [2], and the Schalkwijk-Kailath scheme for the *Additive White Gaussian Noise (AWGN)* channel [5]. Recently [7], we have identified an underlying principle shared by these specific schemes, and formalized it into a general transmission scheme for memoryless channels with feedback, dubbed *posterior matching*. This scheme is simple, sequential, and given explicitly by the channel law and the desired input distribution[2].

In this paper, we prove that the posterior matching scheme achieves rates up to the mutual information for a large family of channels and input distributions, specifically including any DMC and thereby positively settling the longstanding conjecture that the Horstein scheme achieves capacity. Furthermore, we analyze the error probability attained by the scheme, and provide two closed form expressions for a range of rates below the mutual information. We also find the achievable rates in a mismatched setting, where the scheme is designed according to the wrong channel model. Finally, several illustrative examples are given, including a new error exponent analysis for the Horstein scheme. Full proofs are omitted and will appear in [6], brief outlines are given where possible.

[1]The general Horstein scheme is applicable to any Discrete Memoryless Channel (DMC).

[2]The latter can be set e.g. as capacity achieving under some input constraint.

II. Notations and Preliminaries

Probability distributions (over $\mathbb{R}\,,\mathbb{R}^m$) may have both continuous and discrete parts, and are assumed to be equipped with a *Probability Density Function* (PDF), which is called *proper* when there is only a continuous part[3]. For a distribution $\mathcal{P}$ the PDF is denoted by $\mathcal{P}(\cdot)$, the *Cumulative Distribution Function* (CDF) by $F_{\mathcal{P}}(\cdot)$, and the inverse CDF is defined by $F_{\mathcal{P}}^{-1}(t)=\inf\{s:F_{\mathcal{P}}(s)>t\}$. The *support* of $\mathcal{P}$ is denoted by $\mathrm{supp}(\mathcal{P})$. Random variables (r.vs.) are in upper-case, their realizations in corresponding lower-case. The PDF/CDF of a r.v. X are denoted f_X and F_X respectively, and $X\sim\mathcal{P}$ means X is $\mathcal{P}$-distributed. For a measurable function $\eta:\mathbb{R}\mapsto\mathbb{R}$, we write $\mathbb{E}_{\mathcal{P}}\eta$ in short for $\mathbb{E}_{X\sim\mathcal{P}}\left(\eta(X)\right)$, i.e., the expectation of $\eta(X)$ where $X\sim\mathcal{P}$. We also write $x^n=(x_1,\ldots,x_n)$, $\mathrm{conv}(\cdot)$ for the *convex hull* operator, $|\Delta|$ for the length of an interval $\Delta\subseteq\mathbb{R}$, $\log$ for $\log_2$ and $\circ$ for function composition. The uniform distribution over $(0,1)$ is denoted by $\mathcal{U}$.

A. Information Theoretical Notions

We consider a *memoryless channel* defined by a conditional distribution $\mathcal{W}(y|x)$, and equipped with a noiseless instantaneous feedback. The input and output of the channel at time n are denoted by X_n,Y_n respectively, and satisfy $f_{Y_n|X^n,Y^{n-1}}(\cdot|x^n,y^{n-1})=\mathcal{W}(\cdot|\,x_n)$. The *input alphabet* of $\mathcal{W}$ is the set of all $x\in\mathbb{R}$ for which $\mathcal{W}(\cdot|x)$ is defined, and the *output alphabet* is $\cup_x\mathrm{supp}(\mathcal{W}(\cdot|x))$. $\mathcal{Q}(x)$ is said to be a (memoryless) *input distribution* for the channel $\mathcal{W}$ if the input alphabet of $\mathcal{W}$ contains $\mathrm{supp}(\mathcal{Q})$. In this case, $(\mathcal{Q},\mathcal{W})$ is said to be an *input/channel pair*. The corresponding *output distribution* for this pair is denoted throughout by $\mathcal{T}(y)$, and the distribution of the input given the output, namely the *inverse channel*, is denoted by $\mathcal{V}(x|y)$. We write $\mathcal{I}(\mathcal{Q},\mathcal{W})$ for the input-output *mutual information*, $\mathcal{H}(\mathcal{Q})$ for the *entropy* of $\mathcal{Q}$, and $\mathcal{D}(\mathcal{P}\|\mathcal{Q})$ for the *divergence* between $\mathcal{P}$ and $\mathcal{Q}$.

Let Θ be a random *message point* uniformly distributed over the unit interval, its binary expansion representing an infinite independent-identically-distributed (i.i.d) binary sequence to be reliably transmitted over the channel $\mathcal{W}$. A *transmission scheme* (with feedback) is a sequence of measurable *transmission functions* $\{g_n:(0,1)\times\mathbb{R}^{n-1}\mapsto\mathbb{R}\}_{n=1}^{\infty}$, so that the input to the channel $\mathcal{W}$ at time n is given by

$$X_n=g_n(\Theta,Y^{n-1}) \tag{1}$$

[3]In general, the PDF can be a mixture of an integrable function for the continuous part, and Dirac delta functions for the discrete part.

A *decoding rule* is a sequence of measurable maps $\{\Delta_n : \mathbb{R}^n \mapsto \mathfrak{J}\}_{n=1}^{\infty}$, mapping a channel output y^n to a *decoded interval* $\Delta_n(y^n)$, where $\mathfrak{J}$ is the set of all open intervals in $(0,1)$. The *error probability* and the *rate* associated with a transmission scheme and a decoding rule, are defined as

$$p_e(n) \stackrel{\triangle}{=} \mathbb{P}(\Theta \notin \Delta_n(Y^n)), \quad R_n \stackrel{\triangle}{=} -n^{-1}\log|\Delta_n(Y^n)|$$

Optimal decoding rules make use of the message point's posterior PDF $f_{\Theta|Y^n}(\theta \,|\, y^n)$. An *optimal fixed rate* R decoding rule decodes an interval of size 2^{-nR} with maximal a-posteriori probability[4], thereby minimizing $p_e(n)$ for a fixed $R_n = R$. An *optimal variable rate* decoding rule with a *target error probability* $p_e(n)$, decodes the minimal interval whose a-posteriori probability exceeds $1-p_e(n)$, thereby maximizing the instantaneous rate. In general, the rate is a random variable.

We say a transmission scheme together with a decoding rule *achieves* a rate R over $\mathcal{W}$ *within an input constraint* (η, u), if

$$\lim_{n\to\infty}\mathbb{P}(R_n < R) = 0\,, \qquad \lim_{n\to\infty} p_e(n) = 0$$
$$\lim_{n\to\infty} n^{-1}\sum_{k=1}^{n}\eta(X_k) \le u \quad \text{a.s. (element-wise)}$$

where $\eta : \mathbb{R} \mapsto \mathbb{R}^m$ is a measurable function and $u \in \mathbb{R}^m$. R is achieved *pointwise* if this holds for *any* $\Theta = \theta \in (0,1)$.

B. Markov Chains

A Markov chain $\{\Pi_n\}_{n=1}^{\infty}$ over a state space $\mathfrak{F} \subseteq \mathbb{R}^m$ evolves according to a conditional distribution $\mathcal{P}(\pi_{n+1} \,|\, \pi_n)$. The chain is called *Positive Harris Recurrent (PHR)* if it has a unique invariant distribution $\mathcal{M}$ over $\mathfrak{F}$, and every measurable set $A \subseteq \mathfrak{F}$ with $\mathcal{M}(A) > 0$ is reached from any starting point infinitely often. For a variety of PHR conditions, see e.g. [1]. Let $\mathcal{P}_s$ be the distribution induced by $\mathcal{P}$ over $\mathfrak{F}^{\infty}$ with $\Pi_1 = s$.

Lemma 1 (SLLN for Markov chains, from [1]): If $\{\Pi_n\}$ has a unique invariant distribution $\mathcal{M}$, then for any measurable function $\eta : \mathfrak{F} \mapsto \mathbb{R}$ with $\mathbb{E}_{\mathcal{M}}|\eta| < \infty$ and for $\mathcal{M}$-a.a. $s \in \mathfrak{F}$

$$\lim_{n\to\infty} n^{-1}\sum_{k=1}^{n}\eta(\Pi_k) = \mathbb{E}_{\mathcal{M}}\eta \qquad \mathcal{P}_s\text{ - a.s.}$$

If the chain is PHR, then the above holds for any $s \in \mathfrak{F}$.

C. Iterated Function System (IFS)

Let $\{Y_n\}_{n=1}^{\infty}$ be an i.i.d. sequence with some distribution $Y_n \sim \mathcal{P}$. Let $\mathfrak{F}$ be a measurable space, $\omega : \mathbb{R}\times\mathfrak{F} \mapsto \mathfrak{F}$ a measurable function, and write $\omega_y(\cdot) \stackrel{\triangle}{=} \omega(y,\cdot)$ for $y \in \mathbb{R}$. An *IFS* $\{S_n(s)\}_{n=1}^{\infty}$ is a stochastic process over $\mathfrak{F}$, defined by[5]

$$S_1 = s \in \mathfrak{F}\,, \quad S_{n+1}(s) = \omega_{Y_n} \circ \omega_{Y_{n-1}} \circ \cdots \circ \omega_{Y_1}(s) \qquad (2)$$

A *Reversed IFS (RIFS)* $\{\widetilde{S}_n(s)\}_{n=1}^{\infty}$ is the stochastic process obtained by a reversed order composition:[6]

$$\widetilde{S}_1 = s \in \mathfrak{F}\,, \quad \widetilde{S}_{n+1}(s) = \omega_{Y_1} \circ \omega_{Y_2} \circ \cdots \circ \omega_{Y_n}(s) \qquad (3)$$

The (R)IFS is *generated* by $\omega_y(\cdot)$ and *controlled* by $\{Y_n\}_{n=1}^{\infty}$.

[4]This decoded interval may be positioned so that less than nR bits are decoded, e.g. containing $1/2$. Note that just a single extra bit is required to decode the rest, and it can be appended to the next transmission.

[5]We call the process itself an IFS. In the literature sometimes ω_y is the IFS and the process is defined separately. Note the IFS is a Markov chain over $\mathfrak{F}$.

[6]An RIFS is not a Markov chain. It is a standard tool in IFS analysis (cf. [8]), but in our case it turns out to have an independent significance.

We now state some useful convergence Lemmas. A function $\xi : [0,1] \mapsto [0,1]$ is called a (generally nonlinear) *contraction* if it is nonnegative, $\cap$-convex, monotonically increasing, and $\xi(x) < x$ for any $x \in (0,1]$. The simplest example is a linear contraction $\xi(x) = rx$ for $r \in (0,1)$. Note that the n-fold iteration $\xi^{(n)}(\cdot) \to 0$ pointwise. A function $\psi : \mathfrak{F} \mapsto [0,1]$ is called a *length function* if it is continuous and surjective.

Lemma 2: Consider the IFS defined in (2). Suppose there exist a length function $\psi(\cdot)$ and a contraction $\xi(\cdot)$ so that

$$\mathbb{E}_{Y\sim\mathcal{P}}\big[\psi(\omega_Y(s))\big] \le \xi(\psi(s))\,, \quad \forall s \in \mathfrak{F}$$

Then $\psi(S_n(s)) \to 0$ in probability, for any $s \in \mathfrak{F}$.

In the sequel, we consider an IFS over the space $\mathfrak{F}_c$ of CDFs over $(0,1)$, i.e., the space[7] of all monotone non-decreasing functions $h : (0,1) \mapsto (0,1)$ so that $\text{conv}(\text{range}(h)) = (0,1)$. We define the following family of length functions on $\mathfrak{F}_c$:

$$\psi_\lambda(h) \stackrel{\triangle}{=} \int_0^1 \lambda(h(x))dx\,, \quad h \in \mathfrak{F}_c\,, \quad \lambda : [0,1] \mapsto [0,1]$$

where $\lambda(\cdot)$ is surjective, $\cap$-convex and symmetric about $\frac{1}{2}$.

For the remainder of the section assume $\mathfrak{F} \subseteq \mathbb{R}$. Define

$$D_{s,t}(h) \stackrel{\triangle}{=} \frac{|h(s)-h(t)|}{|s-t|}\,, \quad D_s(h) \stackrel{\triangle}{=} \limsup_{t\to s} D_{s,t}(h)$$

for any $h : \mathfrak{F} \mapsto \mathfrak{F}$ and $s,t \in \mathfrak{F}$. $D_{s,t}(\cdot)$ and $D_s(\cdot)$ are called *global* and *local Lipschitz operators* respectively.

Lemma 3: Consider the RIFS (3) with $\mathfrak{F} = (a,b)$. Assume $r \stackrel{\triangle}{=} \sup_{s\neq t\in\mathfrak{F}}\mathbb{E}_{Y\sim\mathcal{P}}\left[D_{s,t}(\omega_Y)\right]^q < 1$ for $q > 0$. Then $\forall\varepsilon > 0$

$$\mathbb{P}(|\widetilde{S}_n(s) - \widetilde{S}_n(t)| > \varepsilon) \le \varepsilon^{-q}|s-t|^q r^n \qquad \forall s,t \in \mathfrak{F}$$

Lemma 4 (From [8]): Consider the RIFS defined in (3) with $\mathfrak{F} = (0,1)$, and let $\rho : (0,1) \mapsto [1,\infty)$ be a continuous function. Define $J(s;t) \stackrel{\triangle}{=} \sup\{\rho(\text{conv}\{s,t\})\}$ and $K_s \stackrel{\triangle}{=} \mathbb{E}_{Y\sim\mathcal{P}}[\,J(s;\omega_Y(s))]$. If

$$r \stackrel{\triangle}{=} \sup_{s\in(0,1)}\mathbb{E}_{Y\sim\mathcal{P}}\left[\frac{\rho(\omega_Y(s))}{\rho(s)}\,D_s(\omega_Y)\right] < 1$$

then defining $\Psi(s,t,r) \stackrel{\triangle}{=} \frac{K_s+K_t}{1-r} + 2J(s;t)$, we have $\forall\varepsilon > 0$

$$\mathbb{P}\left(\left|\widetilde{S}_n(s) - \widetilde{S}_n(t)\right| > \varepsilon\right) \le \varepsilon^{-1}\Psi(s,t,r)\,r^n \quad \forall s,t \in (0,1)$$

III. The Posterior Matching Scheme

In a recent paper [7], we have introduced the following *posterior matching* transmission scheme:

$$X_{n+1} = F_Q^{-1} \circ F_{\Theta|Y^n}(\Theta|Y^n) \qquad (4)$$

This scheme corresponds to the transmission functions $g_{n+1}(\theta, y^n) = F_Q^{-1} \circ F_{\Theta|Y^n}(\theta|y^n)$. The inputs are produced in two steps: First, the information missing at the receiver is "extracted" from the a-posteriori distribution, by generating a r.v. independent of past observations, that together with them uniquely determines Θ. Then, the distribution of that r.v. is "matched" to the channel by transforming it into the desired input distribution $\mathcal{Q}$. This results in $X_n \sim \mathcal{Q}$ independent of an i.i.d output sequence Y^{n-1}. It turns out that the posterior matching scheme admits a simple recursive form.

[7]The space $\mathfrak{F}_c$ is equipped with the Borel σ-algebra associated with the topology of pointwise convergence.

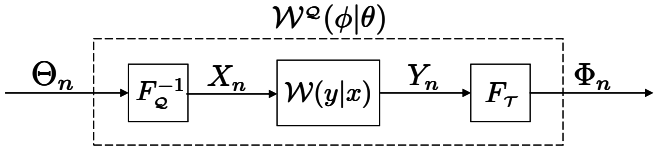

Fig. 1. The normalized channel $\mathcal{W}^{\mathcal{Q}}$

Lemma 5 (from [7]): If $\mathcal{Q}(\cdot)$ is a proper PDF, then the posterior matching scheme (4) is given by

$$X_1 = F_{\mathcal{Q}}^{-1}(\Theta)\,, \quad X_{n+1} = F_{\mathcal{Q}}^{-1} \circ F_{\mathcal{V}}(X_n|Y_n) \tag{5}$$

where $\mathcal{V}$ is the inverse channel corresponding to $(\mathcal{Q},\mathcal{W})$.

Obviously, this recursive representation is invalid in many interesting cases, including DMCs in particular. In order to treat discrete, continuous and mixed alphabet channels within a common framework, we define for any input/channel pair $(\mathcal{Q},\mathcal{W})$ a corresponding *normalized channel* $\mathcal{W}^{\mathcal{Q}}(\phi|\theta)$, by viewing the matching operator $F_{\mathcal{Q}}^{-1}(\cdot)$ as part of the channel, and connecting the output CDF operator $F_{\mathcal{T}}(\cdot)$ to the channel's output[8], as depicted in Figure 1. The normalized channel maps the unit interval to itself, and moreover maps an input $\Theta\sim\mathcal{U}$ to an output $\Phi\sim\mathcal{U}$, namely preserves the uniform distribution over $(0,1)$. It is easy to see that the posterior matching scheme applied to the pair $(\mathcal{U},\mathcal{W}^{\mathcal{Q}})$, is equivalent to the posterior matching scheme for $(\mathcal{Q},\mathcal{W})$, and that the mutual information is conserved, i.e., $\mathcal{I}(\mathcal{U},\mathcal{W}^{\mathcal{Q}}) = \mathcal{I}(\mathcal{Q},\mathcal{W})$. However, the normalized channel allows a unified recursive representation, via the inverse normalized channel $\mathcal{V}^{\mathcal{Q}}$ corresponding to $(\mathcal{U},\mathcal{W}^{\mathcal{Q}})$.

Corollary 1: The posterior matching scheme over the normalized channel is given by

$$\Theta_1 = \Theta\,, \quad \Theta_{n+1} = F_{\mathcal{V}^{\mathcal{Q}}}(\Theta_n|\Phi_n) \tag{6}$$

Corollary 2: The sequence $\{(X_n,Y_n)\}_{n=1}^{\infty}$ is a Markov chain with a state space $\mathrm{supp}(\mathcal{Q}\mathcal{W}) \subseteq \mathbb{R}^2$, and by construction, an invariant distribution $\mathcal{Q}(x)\mathcal{W}(y|x)$. This chain *emulates* the "correct" input marginal $X_n\sim\mathcal{Q}$ and an i.i.d. output sequence $Y_n\sim\mathcal{T}$. Similarly, $\{(\Theta_n,\Phi_n)\}_{n=1}^{\infty}$ is a Markov chain with a state space $(0,1)^2$, and an invariant distribution $\mathcal{W}^{\mathcal{Q}}(\phi|\theta)\mathcal{U}(\theta) = \mathcal{W}^{\mathcal{Q}}(\phi|\theta)$. The chain emulates a uniform input marginal $\Theta_n\sim\mathcal{U}$ and an i.i.d output sequence $\Phi_n\sim\mathcal{U}$. These observations play a central role in the sequel.

IV. Main Results

A. Achievability of the Mutual Information

We now prove that the posterior matching scheme achieves the mutual information for a large family of channels and input distributions. Let Ω be the set of all input/channel pairs $(\mathcal{Q},\mathcal{W})$ with the following properties:

- **(Regularity)** $\mathcal{I}(\mathcal{Q},\mathcal{W}) < \infty$, the normalized channel $\mathcal{W}^{\mathcal{Q}}(\cdot|\theta)$ is a proper PDF for any $\theta\in(0,1)$.
- **(Invariance)** The invariant distribution $\mathcal{W}^{\mathcal{Q}}$ for the Markov chain $\{(\Theta_n,\Phi_n)\}_{n=1}^{\infty}$ defined by (6), is unique.
- **(Contraction)**[9] There is a contraction ξ and a length function ψ_λ over $\mathfrak{F}_c$ such that for any $h\in\mathfrak{F}_c$

$$\mathbb{E}_{\,\Phi\sim\mathcal{U}}\Big(\psi_\lambda\big[F_{\mathcal{V}^{\mathcal{Q}}}(\cdot\,|\Phi)\circ h\big]\Big) \;\le\; \xi\big(\psi_\lambda(h)\,\big) \tag{7}$$

[8]In fact, whenever $F_{\mathcal{T}}(\cdot)$ has a jump discontinuity, the output is randomly selected uniformly over the jump span.

[9]Note that (7) holds if $\xi(\cdot)$ is the identity function ("almost" a contraction).

Let Ω_C be the set of all input/channel pairs $(\mathcal{Q},\mathcal{W})$ satisfying the regularity condition for Ω, for which $\mathcal{Q}(x)\mathcal{W}(y|x)$ is a bounded proper PDF, continuous over an open convex support. Let Ω_{DM} be the set of all input/channel pairs $(\mathcal{Q},\mathcal{W})$ where $\mathcal{W}$ is a DMC (with finite input/output alphabets taken over $\mathbb{R}$, without loss of generality). The following Theorem states that the posterior matching scheme achieves the mutual information for all pairs in Ω, including Ω_C,Ω_{DM} in particular.

Theorem 1 (Achievability): Let $(\mathcal{Q},\mathcal{W})\in\Omega$ (resp. Ω_C). The corresponding posterior matching scheme with a fixed/variable rate optimal decoding rule, achieves (resp. pointwise achieves) any rate $R<\mathcal{I}(\mathcal{Q},\mathcal{W})$ over the channel $\mathcal{W}$ within an input constraint $(\eta,\mathbb{E}_{\mathcal{Q}}\eta)$, provided that $\mathbb{E}_{\mathcal{Q}}|\eta|<\infty$. Furthermore, $\Omega_C\cup\Omega_{DM}\subset\Omega$.

Proof outline. Let $F_n(\cdot)\stackrel{\triangle}{=}F_{\Theta|\Phi^n}(\cdot|\Phi^n)$ be the posterior CDF at time n, so $\Theta_{n+1}=F_n(\Theta)$. We first show that zero rate is achievable, i.e., $|F_n(\Theta+\varepsilon)-F_n(\Theta-\varepsilon)|\to 1$ in probability for any $\varepsilon>0$. To this end we note that $F_n(\cdot)$ is an IFS over the CDF space $\mathfrak{F}_c$, generated by $F_{\mathcal{V}^{\mathcal{Q}}}(\cdot|\phi)$ and controlled by $\{\Phi_n\}_{n=1}^{\infty}$. Using the contraction property (7) and Lemma 2, $\psi_\lambda(F_n(\cdot))\to 0$ in probability which implies that $F_n(\cdot)$ tends to a unit step function about Θ, verifying $R=0$ is achievable.

For $R>0$, let $f_n(\cdot)\stackrel{\triangle}{=}f_{\Theta|\Phi^n}(\cdot|\Phi^n)$ be the posterior PDF at time n, and expand it at the message point using corollary 2 as $\log f_n(\Theta)=\sum_k\log\mathcal{W}^{\mathcal{Q}}(\Phi_k|\Theta_k)$. Using the invariance property for Ω and Lemma 1, we get $n^{-1}\log f_n(\Theta)\to\mathbb{E}_{\mathcal{W}^{\mathcal{Q}}}\log\mathcal{W}^{\mathcal{Q}}(\Phi|\Theta)=\mathcal{I}(\mathcal{U},\mathcal{W}^{\mathcal{Q}})=\mathcal{I}(\mathcal{Q},\mathcal{W})$ a.s., or roughly $f_n(\Theta)\approx 2^{n\mathcal{I}(\mathcal{Q},\mathcal{W})}$. Suppose that $|F_k(\Theta\pm 2^{-nR})-F_k|<\varepsilon$ for all $k\le n$. It can be shown that this implies $f_n(\Theta\pm 2^{-nR})\approx 2^{n(\mathcal{I}(\mathcal{Q},\mathcal{W})-\delta)}$ and $\varepsilon\to 0\Rightarrow\delta\to 0$. For $R<\mathcal{I}(\mathcal{Q},\mathcal{W})-\delta$ this means that $\int f_n(\vartheta)d\vartheta\to\infty$, in contradiction. Thus with high probability $|F_{k_0}(\Theta\pm 2^{-nR})-F_{k_0}|>\varepsilon$ for some $k_0\le n$. We can now consider a transmission starting at k_0 with a message point Θ_{k_0}, and utilizing the zero rate result we conclude the proof, where Lemma 1 verifies the input constraint $(\eta,\mathbb{E}_{\mathcal{Q}}\eta)$ is satisfied. For pairs in Ω_C we can prove the chain is PHR, replacing achievability with pointwise achievability. ■

B. Error Probability Analysis

In this subsection, we provide two sufficient conditions on the target error probability, allowing to achieve a given rate using the corresponding optimal variable rate decoding rule. The approach is different and the results are applicable only to rates below some thresholds $R^*,R^\dagger\le\mathcal{I}(\mathcal{Q},\mathcal{W})$. In some cases an equality holds, but unfortunately we do not know whether this is true in general or not.

The basic idea is the following. Say the receiver has an estimate $\widehat{\theta}_n$ for Θ_n at time n. Then $(\widehat{\theta}_n,\Phi^{n-1})$ corresponds to a unique estimate $\widehat{\theta}_0$ of the message point which is recovered by *reversing* the transmission scheme, i.e., running a RIFS over $(0,1)$ generated by $\omega_\phi(\cdot)\stackrel{\triangle}{=}F_{\mathcal{V}^{\mathcal{Q}}}^{-1}(\cdot|\phi)$ and controlled by $\{\Phi_n\}_{n=1}^{\infty}$. In practice, the receiver decodes an interval and therefore to attain a specific target error probability $p_e(n)$, one tentatively decodes a subinterval of $(0,1)$ in which Θ_n lies with probability $1-p_e(n)$, which since $\Theta_n\sim\mathcal{U}$, is any interval of length $1-p_e(n)$. This interval is then "rolled back" to recover the decoded interval w.r.t. the message point $\Theta=\Theta_1$. The

convergence rate of the RIFS will determine the information rate R. A similar idea works for the original channel, with a RIFS over $\mathrm{supp}(\mathcal{Q})$ generated by $\omega_y(\cdot) \stackrel{\triangle}{=} F_{\mathcal{V}}^{-1}(\cdot|y) \circ F_{\mathcal{Q}}$ and controlled by $\{Y_n\}_{n=1}^{\infty}$.

We now make this notion precise. Suppose $\mathcal{Q}$ has a support over an open (possibly infinite) interval. The *tail function* $T_{\mathcal{Q}} : \mathbb{R}^+ \mapsto [0,1]$ of $\mathcal{Q}$ is defined by

$$T_{\mathcal{Q}}(\ell) \stackrel{\triangle}{=} \inf\left\{p \,:\, p = F_{\mathcal{Q}}(x_1) - F_{\mathcal{Q}}(x_0)\,,\, x_1 - x_0 = \ell\right\}$$

Let $\rho : \mathrm{supp}(\mathcal{Q}) :\mapsto (a,b)$ be differentiable and bijective (a, b may be infinite), and define $\rho[\mathcal{Q}]$ to be the distribution of a r.v. obtained by applying $\rho(\cdot)$ to a $\mathcal{Q}$-distributed r.v.. Such $\rho(\cdot)$ is called *appropriate for* $\mathcal{Q}$ if $|\mathcal{H}(\rho[\mathcal{Q}])| < \infty$. The *family* of appropriate functions for $\mathcal{Q}$ is denoted $\mathfrak{A}(\mathcal{Q})$. Recall also the Lipschitz operators $D_{s,t}, D_s$ defined in subsection II-C.

Theorem 2 (Error Probability I): Let $(\mathcal{Q},\mathcal{W})$ be an input/channel pair with a bounded joint PDF, continuous over an open convex support, and let $\omega_y(\cdot) \stackrel{\triangle}{=} F_{\mathcal{V}}^{-1}(\cdot|y) \circ F_{\mathcal{Q}}$. Define

$$R^* \stackrel{\triangle}{=} \sup_{\rho\in\mathfrak{A}(\mathcal{Q})} \inf_{\substack{s,t\in\mathrm{range}(\rho)\\ s\neq t}} \Big(-\mathbb{E}_{Y\sim\mathcal{T}} \log D_{s,t}\big[\rho\circ\omega_Y\circ\rho^{-1}\big]\Big)$$

and let $R^*(\rho)$ be the infimum above for any specific $\rho \in \mathfrak{A}$. If $R^* > 0$, then the posterior matching scheme with an optimal variable rate decoding rule achieves any rate $R < R^*$, by setting the target error probability to

$$p_e(n) = T_{\rho[\mathcal{Q}]}\left(2^{\,n(R^*(\rho)-R-\varepsilon_n)}\right) \tag{8}$$

for any $\rho \in \mathfrak{A}(\mathcal{Q})$ satisfying $R^*(\rho) > R$, $\varepsilon_n \to 0\,, n\varepsilon_n \to \infty$. Specifically, if ρ has a bounded range then any rate $R < R^*(\rho)$ can be achieved with zero error probability.

Proof: Apply Lemma 3 to the (differentiable) RIFS $\omega_y(\cdot)$ and take the limit $q \to 0$. ∎

Theorem 3 (Error Probability II): Let $(\mathcal{Q},\mathcal{W})$ be an input/channel pair, and let $\omega_\phi(\cdot) \stackrel{\triangle}{=} F_{\mathcal{V}^{\mathcal{Q}}}^{-1}(\cdot|\phi)$. Define

$$R^\dagger \stackrel{\triangle}{=} \sup_\rho R^\dagger(\rho) = \sup_\rho \left(-\log \sup_{s\in(0,1)} \mathbb{E}_{\Phi\sim\mathcal{U}} \frac{\rho(\omega_\Phi(s))}{\rho(s)} D_s\big(\omega_\Phi\big)\right)$$

with the supremum above taken over all continuous functions $\rho : (0,1) \mapsto [1,\infty)$, and $R^\dagger(\rho)$ implicitly defined. 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

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

for any $\alpha \in (0,1)$ and any ρ such that $R^*(\rho) > R$, where Ψ is defined in Lemma 4.

Proof: Apply Lemma 4 to the RIFS $\omega_\phi(\cdot)$. ∎

C. Channel Model Mismatch

Consider a model mismatch case, where the scheme is designed according to the wrong channel model. To that end, for any pair $(\mathcal{Q},\mathcal{W}) \in \Omega_C$, define a *mismatch set* $\Omega_C^{\mathrm{mis}}(\mathcal{Q},\mathcal{W})$ consisting of all input/channel pairs $(\mathcal{Q}^*,\mathcal{W}^*)$, with a corresponding output distribution $\mathcal{T}^*$, satisfying:

- **(Regularity I)** The joint PDF $\mathcal{Q}^*\mathcal{W}^*$ is bounded and continuous over an open convex $\mathrm{supp}(\mathcal{Q}^*\mathcal{W}^*) \subseteq \mathrm{supp}(\mathcal{Q}\mathcal{W})$.
- **(Regularity II)** $\mathcal{I}(\mathcal{Q}^*,\mathcal{W}^*) < \infty$, $\mathcal{D}(\mathcal{W}^*\|\mathcal{W}\,|\,\mathcal{Q}^*) < \infty$.
- **(Contraction)** There is a contraction ξ and a length function ψ_λ over $\mathfrak{F}_c$ such that for every $h \in \mathfrak{F}_c$

$$\mathbb{E}_{Y\sim\mathcal{T}^*}\Big(\psi_\lambda\big[F_{\mathcal{V}}(\cdot\,|Y)\circ F_{\mathcal{Q}}^{-1}\circ h\big]\Big) \;\le\; \xi\big(\psi_\lambda(h)\,\big)$$

- **(Invariance)** Let $X^* \sim \mathcal{Q}^*$ be the input to the channel $\mathcal{W}^*$ and Y^* the corresponding output. Then,

$$F_{\mathcal{Q}}^{-1}(F_{\mathcal{V}}(X^*|Y^*)) \sim \mathcal{Q}^* \tag{9}$$

- **(Expansion)** For any $x \in \mathrm{supp}(\mathcal{Q}^*)$ and $Y \sim \mathcal{W}^*(\cdot|x)$, the r.v. $F_{\mathcal{Q}}^{-1}(F_{\mathcal{V}}(x|Y))$ has a nonzero PDF within some open neighborhood of x.

Remark 1: $(\mathcal{Q},\mathcal{W}) \in \Omega_C \;\Rightarrow\; (\mathcal{Q},\mathcal{W}) \in \Omega_C^{\mathrm{mis}}(\mathcal{Q},\mathcal{W})$.

Theorem 4 (Mismatch Achievability): Let $(\mathcal{Q},\mathcal{W}) \in \Omega_C$, and suppose the corresponding posterior matching scheme (5) is used over a different channel $\mathcal{W}^*$ (unknown on both terminals). If there exists an input distribution $\mathcal{Q}^*$ such that $(\mathcal{Q}^*,\mathcal{W}^*) \in \Omega_C^{\mathrm{mis}}(\mathcal{Q},\mathcal{W})$, then the mismatched scheme with a fixed/variable rate decoding rule matched to $(\mathcal{Q},\mathcal{W})$, pointwise achieves any rate

$$R \;<\; \mathcal{I}(\mathcal{Q}^*,\mathcal{W}^*) - \Big(\mathcal{D}(\mathcal{W}^*\|\mathcal{W}\,|\,\mathcal{Q}^*) - \mathcal{D}(\mathcal{T}^*\|\mathcal{T})\Big) \tag{10}$$

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

V. Examples

*Example 1 (**AWGN**):* Consider an AWGN channel with noise variance N, set $\mathcal{Q} \sim \mathcal{N}(0,P)$ (capacity achieving for input power constraint P), and let $\mathrm{SNR} \stackrel{\triangle}{=} \frac{\mathrm{P}}{\mathrm{N}}$. As already noted [7], the posterior matching scheme (5) reduces in this case to the Schalkwijk-Kailath scheme $x_{n+1} = \sqrt{1+\mathrm{SNR}}\left(x_n - \frac{\mathrm{SNR}}{1+\mathrm{SNR}}\,y_n\right)$, and since this input/channel pair $\in \Omega_C$, Theorem 1 reconfirms the well known fact that this scheme pointwise achieves any rate below $C = \mathcal{I}(\mathcal{Q},\mathcal{W}) = \frac{1}{2}\log(1+\mathrm{SNR})$. Let us derive the error probability using Theorem 2. Inverting the recursive formula above we get $\omega_y(s) = \frac{s}{\sqrt{1+\mathrm{SNR}}} + \frac{\mathrm{SNR}}{1+\mathrm{SNR}}\;y$. Setting $\rho(a) = a$ and using the linearity of ω_y we easily get:

$$R^*(\rho) = \inf_{s\neq t\in\mathbb{R}} \left(-\mathbb{E}_{Y\sim\mathcal{T}} \log D_{s,t}\left[\omega_Y\right]\right) = \frac{1}{2}\log(1+\mathrm{SNR})$$

so in this case $R^* = C$. Plugging the Gaussian tail function $T_{\mathcal{Q}}(\ell) \approx \exp\left(-\ell^2/(8P)\right)$ into (8), we find that a rate $R < C$ is achieved for a target error probability decay given by $p_e(n) \;\approx\; \exp\left(-2^{\,2n(C-R-\varepsilon_n)}\right)$, recovering the well known double-exponential behavior. Note that unlike the traditional Schalkwijk-Kailath scheme where the block length is fixed, our version is sequential and can attain this error performance *on the fly*. A reminiscent sequential Schalkwijk-Kailath scheme with a double-exponential *anytime (delay universal) reliability*, is mentioned in [4].

*Example 2 (**Uniform Input/Noise**):* Let $\mathcal{W}$ be an additive noise channel with noise $\sim \mathcal{U}$, and set $\mathcal{Q} \sim \mathcal{U}$ as well, so $(\mathcal{Q},\mathcal{W}) \in \Omega_C$. In [7], we derived the corresponding posterior matching scheme which is given by $x_{n+1} = \frac{x_n}{y_n}$ for $y_n \in (0,1]$, and $x_{n+1} = \frac{x_n - y_n + 1}{2 - y_n}$ for $y_n \in (1,2)$. By Theorem 1 this simple scheme achieves the mutual information. Using

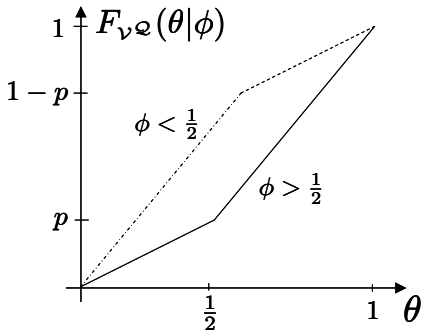

Fig. 2. Recursive transmission functions for the normalized BSC

Theorem 2 with $\rho(a) = a$ for $a \in (0,1)$ and practicing some algebra yields $R^* = \mathcal{I}(\mathcal{Q},\mathcal{W})$, reverifying that the mutual information is achievable. Since $\mathrm{range}(\rho) = (0,1)$ is bounded zero error decoding is possible, as already observed in [7].

*Example 3 (**Exponential Input/Noise**):* Consider an additive noise channel $\mathcal{W}$ with $\mathrm{Exp}(1)$ noise, and set $\mathcal{Q} \sim \mathrm{Exp}(1)$ as well, so $(\mathcal{Q},\mathcal{W}) \in \Omega_C$. The mutual information in this case is $\mathcal{I}(\mathcal{Q},\mathcal{W}) \approx 0.8327$, and by Theorem 1 it is achievable using the posterior matching scheme, which in this case is simply given by $x_{n+1} = \ln\left(\frac{y_n}{y_n - x_n}\right)$. As for the error probability, this time taking $\rho_1(a) = a$ results in $R^*(\rho_1) = -\mathbb{E}\log Y \approx -0.61 < 0$, and we must look for a different function. Setting $\rho_2(a) = \frac{1}{\sqrt{a}}$ results in $R^*(\rho_2) = \frac{1}{2}\,\mathbb{E}\log Y \approx 0.305$. The tail function is bounded by $T_{\rho_2[\mathcal{Q}]}(\ell) \le 1 - e^{-\frac{1}{\ell^2}}$, and any rate $R < R^*(\rho_2)$ is achieved by our scheme with a variable decoding rule and a target error probability set to $p_e(n) = 1 - \exp\left(-2^{-2n(R^*(\rho_2)-R-\varepsilon_n)}\right)$ which provides an error exponent

$$\lim_{n\to\infty} n^{-1}\log\left[1/p_e(n)\right] \;=\; 2(R^*(\rho_2) - R) \approx 0.61 - 2R$$

Although the mutual information is achievable, our error analysis (limited by the selection ρ_2) is valid only up to the rate $R^*(\rho_2) \approx 0.305 < \mathcal{I}(\mathcal{Q},\mathcal{W})$.

*Example 4 (**BSC**):* As already demonstrated in [7], the non-recursive posterior matching scheme (4) for the BSC with $\mathcal{Q} \sim \mathrm{Ber}(\frac{1}{2})$, is precisely the Horstein scheme. Therefore, since $BSC \subseteq \Omega_{DM}$, we conclude that the Horstein scheme is capacity achieving as a corollary of Theorem 1; this widely believed fact lacked a rigorous proof thus far. Note that a similar claim for a block-coding version of the Horstein scheme seems to appear in [3], which presents a lossless joint source-channel coding scheme with feedback, combining a block-Horstein scheme with arithmetic coding.

The original scheme cannot be stated recursively, since the input is quantized, but it nonetheless admits a recursive form (6) relative to the corresponding normalized channel, as depicted in Figure 2 (p is the crossover probability). The corresponding RIFS with $\omega_\phi(\cdot) = F^{-1}_{\mathcal{V}\mathcal{Q}}(\cdot|\phi)$ is therefore supported on two Lipschitz functions, depending on whether $\phi \lessgtr \frac{1}{2}$ (corresponds to $y = 0,1$ in the discrete setting) with $D_s\left(\omega_\phi\right) = (2p)^{-1}$ or $(2(1-p))^{-1}$. The smoothness condition of Theorem 2 is not satisfied, and we must use Theorem 3 for the error probability. For $\beta > 1$, set let $\rho_\beta(\cdot)$ be symmetric about $\frac{1}{2}$, with $\rho_\beta(a) = a^{-\beta}$ for $a \in (0,p\,]$, and $\rho_\beta(a) = p^{-\beta}$ for $a \in (p,\frac{1}{2}]$. Simple manipulations yield

$$R^\dagger(\rho_\beta) = 1 - \log\left[(2p)^{\beta-1} + (2(1-p))^{\beta-1}\right]$$

Unfortunately, $\sup_\beta R^\dagger(\rho_\beta) < 1 - h_b(p)$ for any $p \in (0,1)$. Nevertheless, the permissible error probability can be computed via Theorem 3 and some algebra:

$$\Psi\left((1-\alpha)p_e(n), 1-\alpha p_e(n), 2^{-R^\dagger(\rho_\beta)}\right) = \;\Psi(p,\alpha,\beta)\cdot p_e^{-\beta}$$
$$\Rightarrow \quad \lim_{n\to\infty} -n^{-1}\log p_e(n) \ge \beta^{-1}\left(R^\dagger(\rho_\beta) - R\right)$$

valid for $R < R^\dagger(\rho_\beta)$. Although limited, this is the first rigorous error exponent characterization for the Horstein scheme.

*Example 5 (**Robustness of Schalkwijk-Kailath**):* We now utilize Theorem 4 to demonstrate how the Schalkwijk-Kailath scheme is *robust* to changes in the noise statistics. This property was already mentioned in [9], but only for a change in the variance of the Gaussian noise (SNR mismatch).

Suppose that the Schalkwijk-Kailath scheme designed for an AWGN channel with noise $Z \sim \mathcal{N}(0,N)$ and input $X \sim \mathcal{N}(0,P)$, is used over a *generally non-Gaussian* additive noise channel with noise Z^* having zero mean and a variance N^*. Let $X^* \sim \mathcal{Q}^*$ where $(\mathcal{Q}^*,\mathcal{W}^*) \in \Omega_C^{mis}(\mathcal{Q},\mathcal{W})$ (assuming it exists), and let $Y = X + Z$ and $Y^* = X^* + Z^*$. Looking at variances in the invariance property (9) we have

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

which immediately results in $\mathrm{SNR}^* \stackrel{\triangle}{=} \frac{\mathrm{P}^*}{\mathrm{N}^*} = \mathrm{SNR}$, so the SNR is conserved despite the mismatch. Now applying Theorem 4 and some standard manipulations, we find that the mismatched scheme attains any rate R satisfying

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

Therefore, the mismatched scheme achieves any rate below the Gaussian capacity it was designed for, despite the different noise statistics. The input power is automatically scaled to maintain the same SNR for which the scheme was designed.

References

[1] O. Hernández-Lerma and J.B. Lasserre. *Markov Chains and Invariant Probabilities*. Birkhäuser Verlag, 2003.

[2] M. Horstein. Sequential transmission using noiseless feedback. *IEEE Trans. Info. Theory*, pages 136–143, July 1963.

[3] R. Manakkal and B. Rimoldi. A source-channel coding scheme for discrete memoryless channels with feedback. In *Proc. of ISIT*, 2005.

[4] 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 Info. Theory*, 52(8):3369–3395, Aug. 2006.

[5] J. P. M. Schalkwijk. A coding scheme for additive noise channels with feedback part II: Band-limited signals. *IEEE Trans. Info. Theory*, IT-12:183 – 189, 1966.

[6] O. Shayevitz and M. Feder. *IEEE Trans. Info. Theory*. to be submitted.

[7] O. Shayevitz and M. Feder. Communication with feedback via posterior matching. In *Proc. of ISIT*, 2007.

[8] D. Steinsaltz. Locally contractive iterated function systems. *Ann. of Prob.*, 27(4):1952–1979, Oct 1999.

[9] T. Weissman. Robustness and sensitivity of the Schalkwijk-Kailath scheme. In *The Kailath Colloquium*, 2006.