Posterior Matching Variants and Fixed-Point Elimination

Ofer Shayevitz
University of California, San Diego
La Jolla, CA 92093, USA
Email: ofersha@ucsd.edu

***Abstract*— The posterior matching scheme is known to achieve capacity for a large class of memoryless channels with noiseless feedback. In this contribution, it is shown that whenever the posterior matching kernel admits a fixed point, the corresponding scheme is not ergodic and cannot achieve any positive rate. The source of the problem is traced back to the input ordering implicit in the definition of the scheme. It is then shown how for any discrete memoryless channel, a simple (and easily computable) input permutation eliminates the fixed point phenomena and allows a corresponding variant of the scheme to achieve capacity. This notion is then systematically extended to the case of continuous alphabet memoryless channels.**

I. Introduction

The idea of iterative refinement for communication over memoryless channels with feedback goes back to the classical schemes of Horstein for the Binary Symmetric Channel (BSC) [1], and of Schalkwijk and Kailath for the Additive White Gaussian Noise channel [2][3]. This idea has been recently formalized under the so-called *posterior matching scheme*, a simple and sequential transmission scheme defined for any memoryless channel and any desirable input distribution, achieving the corresponding mutual information under general conditions [4][5][6]. The posterior matching scheme was also shown to admit a natural stochastic control interpretation [7].

In this paper, we point out that whenever the posterior matching recursive kernel admits a fixed-point, then the corresponding scheme is not ergodic (in a sense defined below), and cannot achieve any positive rate. We then show that the fixed-point phenomena is in many cases just an artifact of the implicit input ordering induced by the selection of the c.d.f. in the posterior matching rule, in conjunction with a matching insensitivity of the channel to that ordering. For Discrete Memoryless Channels (DMC), we demonstrate how any finite number of fixed points can be eliminated, by applying a suitable (and easily computable) input permutation, and considering the posterior matching scheme corresponding to the input-permuted channel. Hence, for any DMC and any input distribution, we find a suitable variant of the posterior matching scheme that achieves the corresponding mutual information, within the same input constraints. We then generalize this notion to arbitrary memoryless channels by defining a parameterized family of scheme variants that facilitate the elimination of fixed-points, and explicitly derive the corresponding recursive kernels. This expands the set of channels and input distributions for which the posterior matching scheme, or a suitable variant, is optimal.

The paper is organized as follows. In Section II necessary notations and definitions are provided. The posterior matching scheme and some previous results are described in Section III. In Section IV fixed-point free kernels are defined and shown to be a necessary condition for achievability. In Section V a systematic way to circumvent the fixed-point is developed. Two examples are discussed in Section VI.

II. Preliminaries

A. Notations

Random variables (r.v's) are denoted by upper-case letters, their realizations by corresponding lower-case letters. A (real) r.v. X is associated with a probability distribution $P_X(\cdot)$ over $\mathbb{R}$, and we write $X \sim P_X$. The *cumulative distribution function* (c.d.f.) of X is given by $F_X(x) = P_X\big((-\infty, x\,]\big)$, and the inverse c.d.f. is defined by $F_X^{-1}(t) \triangleq \inf\{x : F_X(x) > t\}$. The *support* of X is the intersection of all closed sets A for which $P_X(\mathbb{R} \setminus A) = 0$, and is denoted $\mathrm{supp}(X)$. We assume that all r.v. are either continuous, discrete, or a mixture of the two. We write $\mathbb{E}(\cdot)$ for expectation and $\mathbb{P}(\cdot)$ for the probability of an event within the parentheses. The uniform probability distribution over $(0,1)$ is denoted throughout by $\mathcal{U}$. A 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 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 $P_X \prec_d P_Y$. We use $\circ$ for function composition, and the indicator function over a set A is denoted by $\mathbb{1}_A(\cdot)$.

B. Information Theoretic Notions

The mutual information between two r.v's X and Y is denoted $I(X;Y)$. A *memoryless channel* is defined via (and 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 $\mathcal{Y} \triangleq \bigcup_{x\in\mathcal{X}} \mathrm{supp}(Y|X = x) \subseteq \mathbb{R}$. A sequence of real r.v. pairs $\{(X_n, Y_n)\}_{n=1}^{\infty}$ is said to be an *input/output sequence* for the memoryless channel $P_{Y|X}$ if for all $n \in \mathbb{N}$,

$$P_{Y_n|X^nY^{n-1}}(\cdot|x^n, y^{n-1}) = P_{Y|X}(\cdot|x_n) \tag{1}$$

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 $\mathcal{Y} = \{0, 1, \ldots, |\mathcal{Y}| - 1\}$.

Let $\Theta_0 \sim \mathcal{U}$ be a random *message point*, its binary expansion representing an infinite i.i.d $\sim \mathrm{Bern}(\frac{1}{2})$ sequence to be reliably conveyed by a transmitter to a receiver over the channel $P_{Y|X}$. A *transmission scheme* is a sequence of *transmission functions* $g_n : (0,1) \times \mathbb{R}^{n-1} \mapsto \mathbb{R}$, so that the input to the channel generated by the transmitter is given by

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

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.

A *decoding rule* is a sequence of mappings $\{\Delta_n : \mathbb{R}^n \mapsto \mathcal{E}\}_{n=1}^{\infty}$, 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

$$p_e(n) \triangleq \mathbb{P}(\Theta_0 \notin \Delta_n(Y^n))$$

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

$$R_n \triangleq -\frac{1}{n} \log |\Delta_n(Y^n)|$$

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

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

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

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

where $\eta : \mathbb{R} \mapsto \mathbb{R}^m$ is a function and $u \in \mathbb{R}^m$. Accordingly, a rate R is called *achievable* over a channel $P_{Y|X}$ within an input constraint (η, u) if there exist a transmission scheme and a decoding rule achieving it. These nonstandard definitions for channel coding with feedback were adopted for a cleaner analysis, and it can be shown that achievability as defined above implies achievability in the standard sense of [8], see [6].

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

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

where ties are broken arbitrarily. This decoding rule minimizes the error probability $p_e(n)$ for a fixed $R_n = R$. This rule uses the posterior distribution of the message point $P_{\Theta_0|Y^n}(\cdot|y^n)$, which can be calculated online at both terminals.

III. Posterior Matching – Background

In this section, we provide a brief account of the posterior matching scheme and corresponding optimality claims. For any given input/channel pair $(P_X, P_{Y|X})$, the posterior matching scheme generates the next input as follows [6]:

$$X_{n+1} = F_X^{-1} \circ F_{\Theta_0|Y^n}(\Theta_0|Y^n) \tag{4}$$

The high level idea here is that evaluating the posterior c.d.f. at the message point, generates a $\mathcal{U}$-distributed r.v. that is independent of Y^n, and together with Y^n uniquely determines Θ_0. Intuitively, we think of that r.v. as representing the information still missing at the receiver in order to reconstruct the message point. The inverse c.d.f. F_X^{-1} is then applied to this r.v., shaping it to match the desired input distribution P_X.

In order to treat discrete, continuous and mixed alphabet channels within a common framework, we define for any input/channel pair $(P_X, P_{Y|X})$ a corresponding *normalized channel* $P_{\Phi|\Theta}$ with $(0,1)$ as a common input/output alphabet, and uniform input/output distributions $\Theta \sim \mathcal{U}, \Phi \sim \mathcal{U}$. 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.[1] This construction is simply formalized by

$$P_{Y|\Theta}(\cdot|\theta) = P_{Y|X}(\cdot|F_X^{-1}(\theta)) \quad \Phi = F_Y(Y) - P_Y(\{Y\}) \cdot \Lambda$$

where $\Lambda \sim \mathcal{U}$ is statistically independent of Θ. Note in particular that $I(\Theta\,;\Phi) = I(X;Y)$.

The normalized channel allows a unified recursive representation for the posterior matching scheme via the inverse normalized channel $P_{\Theta|\Phi}$ corresponding to $(P_\Theta, P_{\Phi|\Theta})$.

Theorem 1 (From [6]): The posterior matching scheme for the normalized channel is given by the recursive relation:

$$\Theta_1 = \Theta_0\,, \quad \Theta_{n+1} = F_{\Theta|\Phi}(\Theta_n|\Phi_n) \tag{5}$$

The sequence of input/output pairs $\{(\Theta_n, \Phi_n)\}_{n=1}^{\infty}$ constitutes a Markov chain over $(0,1)^2$ with an invariant distribution $P_{\Theta\Phi}$. Furthermore, (5) is equivalent to the posterior matching scheme (4) 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 $\{(X_n, Y_n)\}_{n=1}^{\infty}$.

We will refer to $F_{\Theta|\Phi}$ as the (normalized) *posterior matching kernel*. It can be shown [6] that the posterior c.d.f. evolves as an iterated function system generated by this kernel and controlled by the output sequence Φ^n, i.e.,

$$F_{\Theta_0|\Phi^{n+1}}(\cdot|\phi^{n+1}) = F_{\Theta|\Phi}(\cdot|\phi_{n+1}) \circ F_{\Theta_0|\Phi^n}(\cdot|\phi^n) \tag{6}$$

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

(A1) $(P_X, P_{Y|X})$ is regular (see [6]).

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

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

Let Ω be the set of all input/channel pairs satisfying properties (A1) and (A2). We have the following optimality result.

Theorem 2 ([From [6]):] Let $(P_X, P_{Y|X}) \in \Omega$. The corresponding posterior matching scheme with the optimal fixed rate decoding rule, achieves any rate $R < I(X;Y)$ over the channel $P_{Y|X}$ within an input constraint $(\eta, \mathbb{E}\eta(X))$, provided that $\mathbb{E}|\eta(X)| < \infty$.

IV. Fixed-Point Kernels

In this section we establish a necessary achievability condition for the posterior matching scheme, and discuss some important implications. We say that the (normalized) posterior matching kernel $P_{\Theta|\Phi}$ has a *fixed point* at θ_f if

$$\mathbb{P}\left(F_{\Theta|\Phi}(\theta_f|\Phi) = \theta_f\right) = 1 \tag{7}$$

If no fixed-point exists, we say the kernel is *fixed-point free*. Stated differently, a fixed-point at θ_f means the normalized channel output Φ is independent of $\mathbb{1}_{(0,\theta_f]}(\Theta)$, or loosely speaking, provides no information as to whether $\Theta \lessgtr \theta_f$.

Lemma 1: If the posterior matching kernel has a fixed point, then property (A2) does not hold, and the corresponding scheme cannot achieve any positive rate.

Proof: The posterior c.d.f. is obtained by an iterated composition with the kernel, as given in (6). Thus, the fixed point at θ_f induces a fixed point for the posterior c.d.f at θ_f as well, since

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

This immediately implies that no positive rate can be achieved, since the posterior probability of the interval $(0, \theta_f)$ remains a.s. fixed at θ_f. 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 $0 < P_{\Theta\Phi}\left((0, \theta_f) \times (0,1)\right) = \theta_f < 1$. ∎

Loosely speaking, the problem stems from the fact that the posterior matching scheme tries to "encode information" in an input feature to which the channel output is insensitive, namely in $\Theta \lessgtr \theta_f$. This point is perhaps best demonstrated by (the somewhat synthetic) Example 2 in Section VI.

Suppose our kernel has L fixed points $0 < \theta_f^1 < \theta_f^2 < \cdots < \theta_f^L < 1$, and so the unit interval can be partitioned into a disjoint union of $L+1$ contiguous *invariant intervals* $\{J_k = [\theta_f^k, \theta_f^{k+1})\}_{k=0}^{L}$ (where we define $\theta_f^0 = 0$, $\theta_f^{L+1} = 1$), i.e., such that

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

Define the r.v.

$$B \triangleq \sum_{k=0}^{L} k \mathbb{1}_{J_k}(\Theta) \tag{8}$$

namely, B is equal to k if and only if $\Theta \in J_k$.

Lemma 2: $I(\Theta;\Phi|B) = I(\Theta;\Phi)$.

Proof: B is a function of Θ, hence it suffices to show that B and Φ are statistically independent. To that end, for $\mathcal{U}$-a.a. $\phi \in (0,1)$ and any $k \in \{0, \ldots, L+1\}$ we have

$$\begin{aligned} P_{B|\Phi}(k|\phi) &= P_{\Theta|\Phi}(J_k|\phi) = F_{\Theta|\Phi}(\theta_f^{k+1}|\phi) - F_{\Theta|\Phi}(\theta_f^k|\phi) \\ &= \theta_f^{k+1} - \theta_f^k = P_\Theta(J_k) = P_B(k) \end{aligned}$$

where we have used the fixed-point assumption in the third equality. ∎

Now, one brute-force way to try and handle the fixed-point problem is to decode a disjoint union of $L+1$ exponentially small intervals (one per J_k) in which the message point lies with high probability, and then resolve the remaining ambiguity using some simple non-feedback zero-rate code (e.g., repetition). This seems reasonable, yet there are two caveats. First, the maximal rate supported by J_k is $I(\Theta;\Phi|B=k)$, and according to Lemma 2 this can be generally smaller than $I(\Theta;\Phi)$, 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 B).

A better idea is to map the message point into the invariant interval that can support the maximal rate, rather than to use the entire $(0,1)$. Using Lemma 2 again, we have $\max_k I(\Theta;\Phi|B=k) \geq I(\Theta;\Phi)$ hence now there is no loss in rate. Note that using only the maximal rate interval is precisely equivalent to a posterior matching scheme with a different input distribution, induced by the restriction of the kernel to that interval. Since each invariant interval corresponds only to a subset of channel inputs, this in particular means that when a fixed-point exists one can always achieve the same mutual information (at least) using a properly selected subset of the inputs. For DMC this observation is closely related to the fact that whenever $|\mathcal{X}| > |\mathcal{Y}|$, using $|\mathcal{Y}|$ inputs is always sufficient to achieve the unconstrained capacity [9][10], see a brief discussion in Example 1.

We conclude that in the absence of input constraints, any finite number of fixed-points can be eliminated by changing the input distribution (in a way that is easy to compute) without any penalty in rate. However, changing the input distribution will generally breach any imposed input constraints, and so we must seek a different type of solution.

V. Equivalent Channels and μ-Variants

In this section we provide a systematic solution to the fixed-point kernel problem. We start by discussing the DMC case, and then extend the ideas developed to general alphabet channels.

We first note that for a DMC $P_{Y|X}$ with a discrete input distribution P_X, the normalized posterior matching kernel is supported over a finite number of $|\mathcal{Y}|$ functions which are all quasi-linear over a fixed partition of the unit interval corresponding to the input distribution. Precisely, fixing $\phi \in (0,1)$ we have that for any $x \in \mathcal{X}$, the normalized posterior matching kernel evaluated at $\theta = F_X(x)$ is given by

$$F_{\Theta|\Phi}(F_X(x)|\phi) = F_{X|Y}(x|F_Y^{-1}(\phi)) \tag{9}$$

and by a linear interpolation in between these points.

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 for any $i \in \mathcal{X}, j \in \mathcal{Y}$

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

In particular, equivalent pairs have the same mutual information, i.e., $I(X;Y) = I(X^*;Y^*)$.

Let Ω_{DM}^+ be the family of all input/DMC pairs $(P_X, P_{Y|X})$ with nonzero transition probabilities, for which $I(X;Y) > 0$. Let Ω_{DM}^{f} be the family of all pairs in Ω_{DM}^+ that admit a fixed-point free kernel. Recall that by Lemma 1 we already know that $\Omega_{DM}^+ \not\subset \Omega$.

Lemma 3: $\Omega_{DM}^{\mathrm{f}} \subset \Omega$, and for any pair $(P_X, P_{Y|X}) \in \Omega_{DM}^+$ there exists an equivalent pair $(P_{X^*}, P_{Y^*|X^*}) \in \Omega_{DM}^{\mathrm{f}}$.

To prove the Lemma, the following result is found useful.

Lemma 4: 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. ∎

Proof of Lemma 3: For the first statement, see [6]. Consider the case where $(P_X, P_{Y|X}) \in \Omega_{DM}^+$ and let us show there exists an equivalent input/channel pair in Ω_{DM}^{f}. Since $I(X;Y) > 0$ there must exist some $y_0, y_1 \in \mathcal{Y}$ with $P_Y(y_1), P_Y(y_2) > 0$ so that $P_{X|Y}(\cdot|y_0) \neq P_{X|Y}(\cdot|y_1)$. By Lemma 4 there exists a permutation operator σ such that $\sigma(P_{X|Y}(\cdot|y_1)) \prec_d \sigma(P_{X|Y}(\cdot|y_0))$. Thus, combining that with (9) and the quasi-linearity of DMC kernels, it is easily observed that applying σ to the input results in an equivalent input/channel pair for which the corresponding posterior matching kernel has two curves (corresponding to y_1, y_2) where one dominates the other, hence non-intersecting. Therefore, the kernel is fixed-point free and the equivalent input/channel pair is a member of Ω_{DM}^{f} as required. ∎

In the DMC case, the fixed points phenomena is therefore seen to be artifact of the specific input ordering induced by the selection of the c.d.f. in the posterior matching rule. By properly choosing an input permutation inducing a different ordering, fixed points can be eliminated and the posterior matching scheme for the equivalent channel achieves the mutual information under any input constraints encapsulated in P_X. Note that this scheme is immediately translated into an equivalent optimal scheme for the original channel.

Let us now extend the notion of equivalence between input/channel pairs (i.e., the notion input/output permutation) 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. Moreover, for any u.p.f. μ define the *μ-variant* of the posterior matching scheme to be

$$X_{n+1} = F_X^{-1} \circ \mu \circ F_{\Theta_0|Y^n}(\Theta_0|Y^n) \tag{10}$$

where the baseline scheme (4) is recovered by setting μ to be the identity function. It is also easily observed that the μ-variant scheme for an input/channel pair is equivalent to the baseline scheme for a μ-related channel.

The following is an extension of Theorem 2 to the μ-variant case, the proof follows easily along the same lines.

Theorem 3: For any input/channel pair $(P_X, P_{Y|X})$ and any u.p.f. μ, the corresponding μ-variant posterior matching scheme (10) 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 $(P_X, P_{Y|X}) \in \mu(\Omega)$, the scheme achieves any rate $R < I(X;Y)$ within an input constraint $(\eta, \mathbb{E}\eta(X))$, provided that $\mathbb{E}|\eta(X)| < \infty$.

Theorem 3 expands the set of input/channel pairs for which some variant of the posterior matching scheme achieves the mutual information. Note that for a DMC, an input permutation is just a u.p.f. μ permuting intervals that correspond to the discrete inputs, and following our discussion above there always exists such a μ so that the μ-variant kernel is fixed-point free.

VI. Examples

Example 1: For binary input DMC ($|\mathcal{X}| = 2$), the posterior matching kernel is always fixed-point free since any two nonidentical binary distributions can be ordered by dominance. The simplest case of a fixed-point kernel for a DMC is therefore a channel with $|\mathcal{X}| = 3, |\mathcal{Y}| = 2$. A sketch of a typical posterior matching kernel admitting fixed point for the 3×2 channel appears in Figure 1. At a first glance, it might seem that perhaps the occurrence of such a fixed point is pathological, namely that the parameters of the channel and the input distribution must be carefully tuned to that end. However, this is not a rare event at all, in the following sense: Let P_{XY} be picked randomly and uniformly over the simplex of 3×2 probability matrices, and let $P_{\Theta|\Phi}$ be the corresponding posterior matching kernel. Then

$$\mathbb{P}\left(F_{\Theta|\Phi} \text{ has a fixed-point}\right) = \frac{1}{3}$$

To prove this, define (as in Lemmas 3 and 4)

$$\delta_k \triangleq P_{X|Y}(k|1) - P_{X|Y}(k|0) \tag{11}$$

and note that $\sum \delta_k = 0$. Now suppose for a moment that

$$|\delta_0| \leq |\delta_2| \leq |\delta_1| \tag{12}$$

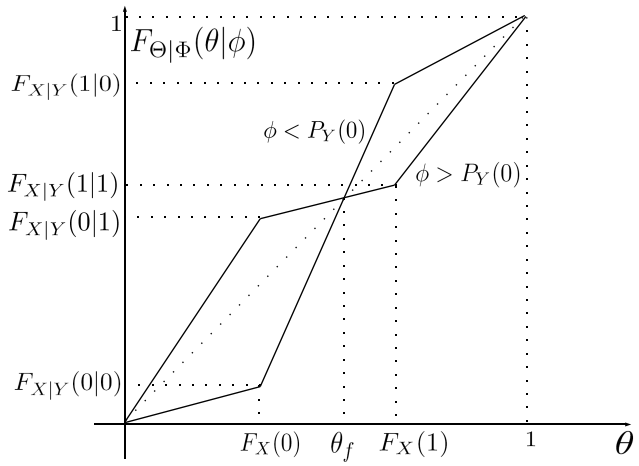

Fig. 1. A fixed-point kernel for a 3×2 channel

which means that δ_0, δ_2 have the same sign, and $\delta_1 = -(\delta_0 + \delta_2)$. Using (9) we see that the two curves $F_{\Theta|\Phi}(\theta|\phi)$ corresponding to the two discrete output values (or equivalently, to $\phi < P_Y(0)$ and $\phi > P_Y(0)$) must intersect at some point $\theta \in (F_X(0), F_X(1))$, since one is below the other for $\theta = F_X(0)$ and then the vice versa for $\theta = F_X(1)$, and both are continuous (in fact linear) in this region. The fact that this intersection is also a fixed point stems from the fact that there are only two curves, and that for any $\theta \in (0,1)$

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

Now, if the δ_k do not satisfy relation (12), there must exist a (unique) permutation (relabeling) of the inputs for which (12) does hold. Moreover, the mirror case ($\delta_0 \leftrightarrow \delta_2$) also yields a fixed point. The two permutations which order the δ_k's as in Lemma 3 cannot of course result in a fixed point, and it is easily verified that so cannot the remaining two. Hence, precisely two out of six possible input permutations yield a fixed point. From the symmetry of drawing P_{XY} (or more precisely, by exchangeability) there is no preference of one permutation over the other, and the result follows.

Note that the above discussion together with Lemma 2 immediately imply (the well known fact) that for any 3×2 channel, the unconstrained capacity can always be achieved using only two inputs. Furthermore, if capacity can be achieved by an input distribution using all three inputs, then it can also be achieved by removing either one of the two inputs with the lowest $|\delta_k|$.

The same type of argument can be applied to a $N \times 2$ channel, where any permutation for which the partial sums sequence of δ_k changes sign will result in a fixed point. Accordingly, if P_{XY} is drawn uniformly as above, the probability of a fixed point approaches 1 as N grows. The analysis becomes much more tricky for $|\mathcal{Y}| > 2$, as one requires all the $\mathcal{Y}$ quasi-linear curves to intersect at the same point, which now cannot be guaranteed simply by requiring the partial sums sequences of δ_k's corresponding to all pairs of outputs to change sign at the same time.

Example 2: 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 1 the scheme cannot attain any positive rate. Intuitively, this stems from the fact that information is being coded in the sign of the input, 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

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

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.

REFERENCES

[1] M. Horstein, "Sequential transmission using noiseless feedback," *IEEE Trans. Info. Theory*, vol. IT-9, pp. 136–143, Jul 1963.

[2] J. P. M. Schalkwijk and T. Kailath, "A coding scheme for additive noise channels with feedback part I: No bandwidth constraint," *IEEE Trans. Info. Theory*, vol. IT-12, pp. 172 – 182, 1966.

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

[4] O. Shayevitz and M. Feder, "Communication with feedback via posterior matching," in *Proc. of the International Symposium on Information Theory*, 2007.

[5] O. Shayevitz and M. Feder, "The posterior matching feedback scheme: Capacity achieving and error analysis," in *Proc. of the International Symposium on Information Theory*, 2008.

[6] O. Shayevitz and M. Feder, "Optimal feedback communication via posterior matching," Submitted to IEEE Trans. Info. Theory, available online at arXiv:0909.4828 [cs.IT].

[7] T. P. Coleman, "A stochastic control viewpoint on 'posterior matching-style' communication schemes," in *Proc. of the International Symposium on Information Theory*, June 2009.

[8] T.M. Cover and J.A Thomas, *Elements of Information Theory*, John Wiley & Sons, Inc., 1991.

[9] C. E. Shannon, "Some geometrical results in channel capacity," *Nachrichtentech. Z.*, vol. 10, 1957.

[10] R. Gallager, *Information Theory and Reliable Communication*, John Wiley & Sons, Inc., 1968.