# Communication with Feedback via Posterior Matching

Ofer Shayevitz
Tel Aviv University, Dept. of EE-Systems
Tel Aviv 69978, Israel
ofersha@eng.tau.ac.il

Meir Feder
Tel Aviv University, Dept. of EE-Systems
Tel Aviv 69978, Israel
meir@eng.tau.ac.il

***Abstract*— In this paper we describe a general algorithmic scheme for communication over any memoryless channel in the presence of noiseless feedback. The scheme is based on the idea of posterior matching, in which the information still missing at the receiver is extracted from the a-posteriori density function, and matched to any desirable input distribution. We analyze the error probability attained by this scheme for additive noise channels, and show that the well-known Schalkwijk-Kailath scheme for the AWGN channel with average power constraint and the Horstein scheme for the BSC, can be derived as special cases.**

## I. Introduction

Feedback cannot increase the capacity of memoryless channels, but it can significantly improve the error probability performance, and perhaps more importantly - it can drastically simplify the transmission schemes required to achieve capacity. Whereas complex coding techniques strive to approach capacity in the absence of feedback, the same goal can sometimes be obtained using noiseless feedback by simple deterministic schemes that work "*on the fly*".

Probably the first elegant feedback scheme ever to be presented is due to Horstein [1] for the Binary Symmetric Channel (BSC) with feedback. In that paper, a message point inside the unit interval was used to represent the data bits, and was conveyed to the receiver by always indicating whether it lies to the left or to the right of the receiver posterior's median, which is also known to the transmitter via feedback. That way, the transmitter always answers the most informative question that can be posed by the receiver based on the information the latter has. Remarkably, this simple technique is enough to attain the capacity of the BSC, and is easily adopted to any Discrete Memoryless Channel (DMC) with feedback.

A few years later, two landmark papers by Schalkwijk-Kailath [2] and Schalkwijk [3] presented an elegant capacity achieving feedback scheme for the Additive White Gaussian Noise (AWGN) channel with an average power constraint. The Schalkwijk-Kailath scheme has a "parameter estimation" spirit. At each time point, it finds the Minimum Mean Square Error (MMSE) estimate of the message point at the receiver, and transmits the MMSE error on the next channel use, amplified to match the permissible input power constraint. This 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.

Since the emergence of the Horstein and the Schalkwijk-Kailath schemes, it was evident that those were similar in some fundamental sense. Both schemes used the message point representation, and both always attempted to tell the receiver what it was still missing in order to "get it right". However, neither the precise correspondence between these schemes nor a generalization to other cases has ever been established. In this paper, we show that in fact there exists an underlying principal connecting these two methods. We present a general feedback transmission scheme for any memoryless channel and any required input distribution. Our scheme is simple and elegant, and manifests the idea of always transmitting what the receiver is missing. In the special cases of a BSC with uniform input distribution and an AWGN channel with a Gaussian input distribution, our scheme is reduced to those of Horstein and Schalkwijk-Kailath respectively.

The paper is organized as follows. In section II we present the new scheme. In section III we provide a preliminary analysis of the scheme and its error performance. In section IV we derive the Horstein and Schalkwijk schemes as a special case, and in section V we provide an illustrative example by applying our scheme to a simple setting. A discussion and some future research items are provided in section VI.

## II. The New Scheme

Consider a discrete-time memoryless channel with an instantaneous noiseless feedback and a transition probability law $W = W(y|x)$, and let $Q = Q(x)$ be any desirable input distribution[1,2] with finite expectancy and variance. Let $\theta_0 \in [0,1)$ be the *message point* whose binary expansion represents an infinite bitstream to be reliably conveyed to the receiver. The message point is selected according to a uniform distribution over the unit interval. Denote the transmitted signal at time $n$ by $x_n = g_n(\theta_0, y_1^{n-1})$ where $y_n$ is the corresponding channel output, and

$$g_n : [0,1) \times \mathbb{R}^{n-1} \mapsto \mathbb{R}$$

is a sequence of deterministic *transmission functions* known at both terminals. At each time point, the receiver calcu-

[1] For instance, $Q$ may be selected to be a capacity achieving distribution under some desirable input constraint.

[2] Our approach is valid for any channel/input distribution pair, but here we assume $Q, W$ are Probability Density Functions (PDFs).

lates the a-posteriori density function of the message point $f_n(\theta) = f_{\theta|y_1^n}(\theta\,|\,y_1^n)$, starting with $f_0(\theta)$ uniform over the unit interval. Thanks to the noiseless feedback, the transmitter can track $f_n(\theta)$ as well, and is assumed to do so. A proper selection of transmission functions will hopefully result in a fast concentration of the posterior $f_n(\theta)$ around $\theta_0$, rapidly reducing the uncertainty at the receiver.

For communications with a fixed rate $R$, the decoding rule we consider after $n$ channel uses is looking for an interval of size $2^{-nR}$ whose a-posteriori probability is maximal[3]. Alternatively, one can set the bit error probability at a threshold, and decode bits whenever their respective intervals accumulate enough probability, as in [1]. This variable rate approach possesses an error exponent inherently superior to that of the aforementioned fixed rate approach, however below we focus on the latter as it is easier to analyze.

What is the best selection of the functions $g_n$? We argue the following: Since $f_n(\theta)$ describes the receiver's knowledge (or lack of it) regarding $\theta_0$ at time $n$ given $y_1^n$, it is reasonable to *zoom-in* on $\theta_0$ by somehow "stretching" the posterior into the desired input distribution, and hence describe to the receiver in greater detail what it is still missing. Therefore, we suggest to use

$$g_{n+1}(\theta_0, y_1^n) = F_Q^{-1} \circ F_{\theta|y_1^n}\left(\theta_0|y_1^n\right) \tag{1}$$

where $F_{\theta|y_1^n}$ is the Cumulative Distribution Function (CDF) corresponding to the posterior $f_n(\theta)$, and $F_Q$ is the CDF of the desired input distribution $Q$. It is easy to see that $F_{\theta|y_1^n}(\theta_0|y_1^n)$, viewed as a random variable, is uniform over the unit interval given any value of $y_1^n$ and is therefore independent of $y_1^n$. Hence, $g_{n+1}$ is $Q$-distributed and independent of $y_1^n$ as well. Notice that the inputs are essentially produced in two steps. In the first step, the information regarding $\theta_0$ still missing at the receiver is "extracted", by deterministically generating a random variable independent of previous observations, that together with those observations uniquely determines $\theta_0$. In the second step, the distribution of that random variable is "matched" to the channel by transforming it into that of the desired input distribution $Q$.

This strategy admits a simpler recursive form. Define the *inverse channel* for $W$ with an input distribution $Q$ to be

$$V(x|y) = \frac{Q(x)W(y|x)}{\sum_x Q(x)W(y|x)}$$

and let $F_{X|Y}(x|y)$ be the CDF of $V(x|y)$ for a fixed value $y$, namely,

$$F_{X|Y}(x|y) = \int_{-\infty}^{x} V(\xi|y)d\xi$$

Define also

$$S(x,y) \stackrel{\triangle}{=} F_Q^{-1} \circ F_{X|Y}(x|y).$$

*Lemma 1:* The transmission functions (1) are also given by the recursive formula

$$\begin{aligned} g_1(\theta_0) &= F_Q^{-1}(\theta_0) \\ g_{n+1}(\theta_0, y_1^n) &= S\left(g_n(\theta_0, y_1^{n-1}), y_n\right) \end{aligned} \tag{2}$$

*Proof:* This Lemma can be proved directly by taking derivatives, as we shall verify in the sequel. Here we provide a more illuminating proof by induction, showing that (1) and (2) are the same as functions for any $n$. This is immediately true for $n = 1$. Assume now it is true for $n = k$. As shown above $g_k$ is independent of $y_1^{k-1}$. Since the channel is memoryless, $(g_k, y_k)$ are also independent of $y_1^{k-1}$. Therefore, $F_{X|Y}(g_k|y_k)$ is uniform given $y_1^k$, and applying $F_Q^{-1}$ transforms its distribution into $Q$. Thus, $g_{k+1} = S(g_k, y_k)$ is $Q$-distributed given $y_1^k$ and is obtained from $\theta_0$ by a composition of monotonic functions in $\theta_0$, which itself is monotonic in $\theta_0$. The same is true for $g_{k+1}$ obtained from (1). By the uniqueness of a monotonic transformation between distributions, $g_{k+1}$ generated by either (1) or (2) is identical for any $y_1^k$, and the proof is concluded. ■

The recursive form (2) provides a simple way for implementing our transmission scheme: The next channel input is given by a deterministic function of the previous input and previous output only, i.e., $x_{n+1} = S(x_n, y_n)$.

## III. Analysis

In this section we derive some basic properties of the suggested scheme. We shall focus henceforth on the case of an additive noise channel, but essentially the same results are expected to be valid in a general memoryless setting, under some regularity conditions. Both the noise and the input are assumed to have bounded first and second moments, and the input distribution is assumed to satisfy $Q(x) < Q_{max}$. We denote the noise sequence by $z_k$ and its PDF by $f_Z(\cdot)$. The mutual information of the channel $W$ with input distribution $Q$ is denoted by $I = I(Q, W)$. The dependence of $f_n(\theta)$ and $g_{n+1}(\theta)$ on $y_1^n$ is usually omitted for notational clarity.

*Lemma 2:* The posterior evaluated at the message point has the following asymptotic behavior

$$\lim_{n\to\infty} \frac{1}{n} \log f_n(\theta_0) = I(Q, W) \quad \text{with probability } 1 \tag{3}$$

*Proof:* Applying Bayes' law it is easily verified that the posterior satisfies the following recursion rule:

$$f_n(\theta) = \frac{f(y_n\,|\,\theta, y_1^{n-1})}{f(y_n\,|\,y_1^{n-1})}\, f_{n-1}(\theta)$$

Applying the recursion rule $n$ times and taking a logarithm, we get

$$\begin{aligned} \frac{1}{n}\log f_n(\theta) = &\frac{1}{n}\sum_{k=1}^{n} \log W(y_k\,|\,g_k(\theta, y_1^{k-1})) \\ &- \frac{1}{n}\sum_{k=1}^{n} \log f(y_k\,|\,y_1^{k-1}) \end{aligned}$$

[3] As in arithmetic coding, this interval may be positioned so that less than $nR$ bits are decoded. Similarly, the number of bits not decoded is expected to be small and independent of $n$ [4]. Alternatively, note that just a single extra bit is required to decode the rest, and it can be appended to the next block.

Evaluating the above at the message point, we use the fact that $g_k, y_k$ are independent of $y_1^{k-1}$ and the noise is additive, and apply the law of large numbers (LLN) to the i.i.d. sequences:

$$\begin{aligned}\frac{1}{n}\log f_n(\theta_0) =&\ \frac{1}{n}\sum_{k=1}^{n}\log f_Z(z_k) - \frac{1}{n}\sum_{k=1}^{n}\log f_Y(y_k)\\ \xrightarrow[n\to\infty]{}&\ -H(Z)+H(Y) = \ I(Q,W)\end{aligned}$$

with probability 1, as required. ■

*Lemma 3:* The derivative of the transmission function evaluated at the message point, has the following asymptotic behavior

$$\lim_{n\to\infty}\frac{1}{n}\log\left.\frac{\partial g_n(\theta,y_1^{n-1})}{\partial\theta}\right|_{\theta=\theta_0} \geq I(Q,W) \text{ with probability 1} \tag{4}$$

*Proof:* From (1) we easily find that

$$\int_0^{\theta} f_{n-1}(\theta')d\theta' = F_Q\left(g_n(\theta)\right)$$

which results in

$$\frac{\partial g_n(\theta)}{\partial\theta} = \frac{f_{n-1}(\theta)}{Q\left(g_n(\theta)\right)}$$

this can also be obtained from (2) by noticing that

$$S_1(x,y) \triangleq \frac{\partial S\left(x,y\right)}{\partial x} = \frac{f_{X|Y}(x|y)}{Q\left(S\left(x,y\right)\right)}$$

and then applying the chain rule for derivatives

$$\begin{aligned}\frac{\partial g_n(\theta,y_1^{n-1})}{\partial\theta} &= \frac{1}{Q(g_1(\theta))}\prod_{k=1}^{n-1}S_1\left(g_k(\theta),y_k\right) &(5)\\ &= \frac{1}{Q(g_n(\theta))}\prod_{k=1}^{n-1}\frac{f_{X|Y}(g_k(\theta)|y_k)}{Q(g_k(\theta))} = \frac{f_{n-1}(\theta)}{Q\left(g_n(\theta)\right)}\end{aligned}$$

verifying Lemma 1 again. We now immediately have that

$$\frac{1}{n}\log\frac{\partial g_n(\theta)}{\partial\theta} = \frac{1}{n}\log f_{n-1}(\theta) - \frac{1}{n}\log Q(g_n(\theta))$$

and using Lemma 2 together with the assumption $Q < Q_{max}$ we get the desired result. ■

The properties described above provide a good idea regarding the behavior of the posterior. Loosely speaking, the posterior has a peak of $2^{nI}$ at the message point, and since the derivative of $g_n(\theta)$ at that point is at least $2^{nI}$, the trajectory[4] of points that lie $2^{-n(I+\varepsilon)}$ close to $\theta_0$ is attracted to that of $\theta_0$, hence for such points we expect that $f_n(\theta)\approx 2^{nI}$. In contrast, the trajectory of points that lie $2^{-n(I-\varepsilon)}$ far from $\theta_0$ diverges from that of $\theta_0$, towards the boundaries of support$(Q)$. We therefore expect a probability mass approaching one to be concentrated in a $2^{-nR}$ vicinity of the message point for any $R < I$, which translates to reliable communications at any rate below the mutual information.

The following Lemma provides a useful expression for the error probability of our scheme, which is applied to the AWGN channel in the next section.

*Lemma 4:* For any rate $R$ our scheme attains an error probability upper bounded by

$$P_e \leq 1-\mathbb{E}\Big(F_Q(g_{n+1}(\theta_0+\Delta\theta))-F_Q(g_{n+1}(\theta_0-\Delta\theta))\Big) \tag{6}$$

where $\Delta\theta = 2^{-(nR+1)}$.

*Proof:* From (1) we easily find again that the posterior's integral is given by

$$\int_{\theta_1}^{\theta_2} f_n(\theta)d\theta = F_Q(g_{n+1}(\theta_2)) - F_Q(g_{n+1}(\theta_1)) \tag{7}$$

We therefore have the following expression for the error probability given $y_1^n$:

$$\begin{aligned}P_e(y_1^n) &= 1-\sup_{\theta_1}\int_{\theta_1}^{\theta_1+2^{-nR}} f_n(\theta)d\theta\\ &= 1-\sup_{\theta_1}\Big(F_Q(g_{n+1}(\theta_1+2\Delta\theta)) - F_Q(g_{n+1}(\theta_1))\Big)\\ &\leq 1-F_Q(g_{n+1}(\theta_0+\Delta\theta)) + F_Q(g_{n+1}(\theta_0-\Delta\theta))\end{aligned}$$

and the proof is concluded by taking the expectation on both sides to get the average error probability. ■

Lemma 4 demonstrates that the error probability is determined by two factors: The input CDF's tail behavior, and the sensitivity of the transmission functions to a $2^{-nR}$ perturbation in the assumed position of the message point, namely how fast is the resulting divergence of the trajectory towards the boundaries of support$(Q)$.

*Corollary 1:* Assume $\sup S_1(x,y) < \infty$, and so the divergence of the trajectory is exponential at best. If support$(Q) = \mathbb{R}$ and fixed-rate block decoding is used, then a necessary condition for a doubly-exponential error probability is for $Q$ to have an exponentially decaying tail.

## IV. Schalkwijk and Horstein Revisited

*Example 1 (The AWGN channel):* We now provide a sketch of the analysis for the AWGN setting, and show that our scheme in this particular case is essentially the same as the Schalkwijk-Kailath scheme [2][3]. Assume the noise is $\mathcal{N}(0,\sigma^2)$, the average power constraint is $P$, and denote $SNR = \frac{P}{\sigma^2}$. Set $Q\sim\mathcal{N}(0,P)$ (capacity achieving) and let $\phi_0 = F_Q^{-1}(\theta_0)$, which is the message point converted into a Gaussian distribution, and also the first channel input $g_1(\theta_0)$.

It is easily verified that (1) in this case is merely an affine transformation that transform the posterior into $\mathcal{N}(0,P)$, hence the transmission functions are given by

$$g_n(\theta_0,y_1^{n-1}) = (1+SNR)^{\frac{n}{2}}\left(\phi_0-\mathbb{E}(\phi_0\,|\,y_1^{k-1})\right)$$

Observe that in this case $g_n(\theta_0)$ is just the estimation error of an $MMSE$ estimator for $\phi_0$ (which represents $\theta_0$) given the observations, amplified to match the permissible input power. The recursive representation (2) in this case is simply

$$S(x,y) = \sqrt{1+SNR}\left(x-\frac{SNR}{1+SNR}\,y\right)$$

[4] The trajectory of a point $\theta$ is the sequence of values obtained by applying $g_k(\theta,y_1^{k-1})$ with increasing $k$. When calculating the a-posterior density, the receiver in fact tracks the trajectory of all possible message points.

which is exactly the transmission strategy of the Schalkwijk-Kailath scheme [3]. We now find an explicit expression upper bounding the error probability. Taking the derivative of the transmission function, we get

$$\frac{\partial g_n(\theta)}{\partial \theta} = \frac{(1+SNR)^{\frac{n}{2}}}{Q(F_Q^{-1}(\theta))} \geq \sqrt{2\pi P}(1+SNR)^{\frac{n}{2}}$$

and so

$$\begin{aligned} g_n(\theta_0 + 2^{-nR}) &\geq g_n(\theta_0) + \int_{\theta_0}^{\theta_0+2^{-nR}} \sqrt{2\pi P}(1+SNR)^{\frac{n}{2}} d\theta \\ &= g_n(\theta_0) + \sqrt{2\pi P}\cdot 2^{-n(C-R)} \end{aligned}$$

where $C = \frac{1}{2}\log(1+SNR)$ is the Gaussian channel capacity. Similarly,

$$g_n(\theta_0 - 2^{-nR}) \leq g_n(\theta_0) - \sqrt{2\pi P}\cdot 2^{-n(C-R)}$$

Applying Lemma 4, we bound each of the terms in (6) separately, using the fact that $g_n(\theta_0)$ is Gaussian. Denote by $a_n = \sqrt{2\pi P}\cdot 2^{n(C-R)}$, and we have

$$\begin{aligned} \mathbb{E}F_Q(g_n + a_n) &\geq \mathbb{P}(g_n > -\frac{a_n}{2})\mathbb{E}\left(F_Q(g_n+a_n)\big| g_n > -\frac{a_n}{2}\right) \\ &\geq F_Q^2(\frac{a_n}{2}) \end{aligned}$$

$$\begin{aligned} \mathbb{E}F_Q(g_n - a_n) &\leq \mathbb{E}\left(F_Q(g_n-a_n)\big| g_n < \frac{a_n}{2}\right) + \mathbb{P}(g_n > \frac{a_n}{2}) \\ &\leq 2(1 - F_Q(\frac{a_n}{2})) \end{aligned}$$

Putting the terms together we get asymptotically

$$\begin{aligned} P_e &\leq 1 - F_Q^2(\frac{a_n}{2}) + 2(1-F_Q(\frac{a_n}{2})) \approx 4(1-F_Q(\frac{a_n}{2})) \\ &= 4\left(1 - F_Q(\frac{1}{2}\sqrt{2\pi P}\cdot 2^{n(C-R)})\right) \approx 2\exp\left(-\frac{\pi}{4}2^{2n(C-R)}\right) \end{aligned}$$

where we have used the exponential approximation of the Gaussian CDF. We thus get the same double exponential decay as in the Schalkwijk-Kailath scheme

$$\lim_{n\to\infty} \frac{1}{n}\log\log\frac{1}{P_e} \geq 2(C-R) \tag{8}$$

via a slightly different analysis.

The difference between our general scheme and the ‘estimation error” approach of the Schalkwijk-Kailath scheme in a non-Gaussian setting should now be evident. For general additive noise the Schalkwijk-Kailath scheme transmits the linear MMSE estimation error given past observations, which is *uncorrelated* with those observations but *not independent* of them as in our scheme, except for the Gaussian case.

*Example 2 (The BSC channel):* We now consider the BSC setting with crossover probability $p$, and show that our scheme in this case is essentially the same as the Horstein scheme [1]. The discussion is easily adjusted to any DMC with feedback. According to our approach, the channel's input should be independent of previous outputs and distributed $\sim$ Ber$(\frac{1}{2})$ (capacity achieving). To that end, the function $g_n$ can be an indicator function of any subset with a-posteriori probability equal to $\frac{1}{2}$. One possibility is:

$$g_n(\theta_0, y_1^{n-1}) = \begin{cases} 0 & \theta_0 < median\{f_{n-1}(\theta)\} \\ 1 & o.w. \end{cases} \tag{9}$$

which is precisely the Horstein scheme. Applying (1) results in (9) as well, since $F_Q^{-1}$ corresponds to a selection of a median subset. Note that unlike the continuous alphabet case, there is an inherent loss of information in the “matching” step here, since the posterior is converted into a discrete distribution.

The posterior in this case is built by multiplying each side of the median by either $2p$ or $2(1-p)$ according to the received bit, and since the message point always lies on the correct side of the median, we get

$$f_n(\theta_0) = 2^n p^{n_1}(1-p)^{n-n_1}$$

where $n_1 \approx np$ is the number of crossovers that occurred during transmission. This immediately results in

$$\frac{1}{n}\log f_n(\theta_0) \quad \xrightarrow[n\to\infty]{} \quad 1 - h_b(p) = C \quad \text{with probability 1}$$

as expected. Notice that the posterior is quasi-constant over at most $n+1$ disjoint intervals, therefore the size of the interval containing the message point is no larger than $2^{-nC}$. These observations have been utilized before [5] for variable rate universal communications when the noise is an individual sequence. Due to the discrete nature of the setting, the error probability analysis differs from that described herein (for instance, Lemma 3 naturally does not apply) and is left out.

## V. Uniform Noise Example

Our suggested method generalizes previously proposed feedback schemes, and to demonstrate its application in cases not handled before, we provide a simple illustrative example of a uniform noise channel with a uniform input distribution. We shall see that the resulting transmission strategy in this case turns out to be a very intuitive one, which vividly demonstrates the *zoom-in* effect mentioned earlier.

*Example 3 (Uniform noise with uniform input distribution):* Consider a memoryless additive noise channel with $U(0,1)$ noise, and say we choose an input distribution which is also $U(0,1)$. What is our transmission strategy in this simple case? It is easy to verify that the inverse channel $V(x|y)$ is

$$V(x|y) \sim \begin{cases} U(0,y) & y \leq 1 \\ U(y-1,1) & y > 1 \end{cases}$$

Since the input distribution was set to be $U(0,1)$, the function $S(x,y)$ is merely the CDF of $V(x|y)$ and is given by

$$S(x,y) = \begin{cases} \Lambda(\frac{x}{y}) & y \leq 1 \\ \Lambda(\frac{x-y+1}{2-y}) & y > 1 \end{cases}$$

where $\Lambda(x) = \min(\max(x,0),1)$. This means that our transmission strategy in this case is very simple. We start by transmitting $g_1 = \theta_0$. Then, given $y_1$ we find the range of inputs that could have generated it, and apply to $g_1$ a transformation that linearly stretches this range to fill the entire unit interval, which provides us with $g_2$ to transmit. We now find the range of possible inputs given $y_2$, and apply the corresponding linear transformation to $g_2$, and so on. This is intuitively appealing since what we do in each iteration is just *zoom-in* on the remaining uncertainty region for $\theta_0$. Since the posterior is always uniform, this zooming-in is linear.

This transmission strategy results in a posterior which is uniform in an ever shrinking sequence of intervals. Consequently, in this case it is easier to look at a variable-rate decoding rule, by simply decoding the current interval $(a_n, b_n)$ within which the posterior is uniform. The size of that interval is

$$|b_n - a_n| = \prod_{k\in A} y_k \prod_{k\notin A} (2 - y_k)$$

where $A = \{k : y_k < 1\}$. This is a *zero-error* decoding rule which results in a variable rate that converges to

$$R = \frac{-\log|b_n - a_n|}{n} = \frac{1}{n}\sum_{k\in A}\log\frac{1}{y_k} + \frac{1}{n}\sum_{k\notin A}\log\frac{1}{2-y_k}$$

$$= \frac{1}{n}\sum_{k=1}^{n}\log\frac{f_{X|Y}(x_k|y_k)}{f_X(x_k)} \quad \xrightarrow[n\to\infty]{} \quad I \quad \text{with probability 1}$$

where $I$ is the corresponding mutual information. Note that in this example, every channel output actually produces bits in the amount corresponding to its individual mutual information.

## VI. Discussion

A sequential communications strategy for memoryless channels with feedback was described, providing in particular a unified view of the known Horstein and Schalkwijk-Kailath schemes. The core of the 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 a-posteriori cumulative distribution function at the message point, followed by a technical step of matching this quantity to the channel via an appropriate transformation. A preliminary analysis for additive noise channels was provided. The proposed scheme is expected to attain the capacity of general memoryless channels under suitable regularity conditions, an issue which is currently under investigation.

A known drawback of the Schalkwijk-Kailath scheme is that its peak power may become arbitrarily large. This problem was treated in [6] by ceasing transmission and declaring an error whenever the time averaged power exceeded some given threshold, at the cost of loosing the doubly-exponential error probability. However, our scheme allows for a much simpler solution, since the input distribution can be set (and optimized) to obey any required single letter peak constraint.

An interesting research direction could be the treatment of channels with memory within the same framework, possibly by modifying the channel matching step to depend on previous outputs. Another direction to be explored is the possible use of our method for universal communications with feedback. In a stochastic universal setting, the transmitter can estimate the channel with increasing accuracy, and match the transmission strategy accordingly. Although the receiver does not know the channel, it seems plausible that for a "not too rich" family of channels, the calculated posterior will have a significant peak only when "close enough" to the true channel, and will be flat otherwise. Furthermore, it should be examined whether the same method can be used in an individual noise setting as well, employing randomization techniques in the spirit of [5].

## VII. Acknowledgment

We gratefully acknowledge the useful comments made by an anonymous reviewer, which greatly improved the presentation.

## References

[1] M. Horstein, "Sequential transmission using noiseless feedback," *IEEE Trans. Info. Theory*, pp. 136–143, July 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 siganls," *IEEE Trans. Info. Theory*, vol. IT-12, pp. 183 – 189, 1966.

[4] O. Shayevitz, R. Zamir, and M. Feder, "Bounded expected delay in arithmetic coding," in *Proc. of the International Symposium on Information Theory*, 2006.

[5] O. Shayevitz and M. Feder, "Achieving the empirical capacity using feedback - part I: Memoryless additive models," *Submitted to the IEEE Trans. Info. Theory,* Available online at: http://www.eng.tau.ac.il/∼ofersha/empirical_capacity_part1.pdf.

[6] A.D. Wyner, "On the schalkwijk-kailath coding scheme with a peak energy constraint," *IEEE Trans. on Info. Theory*, vol. IT-14, no. 1, pp. 129–134, Jan. 1968.