International Zurich Seminar on Communications (IZS), February 26 – 28, 2014

An Alternative Coding Theorem for Posterior Matching via Extrinsic Jensen–Shannon Divergence

Tara Javidi
Electrical and Computer Engineering
University of California San Diego
Email: tjavidi@ucsd.edu

Michèle Wigger
Communications and Electronics
Telecom ParisTech
Email: michele.wigger@telecom-paristech.fr

Mohammad Naghshvar
Qualcomm Technologies, Inc.
Corporate R&D
Email: mnaghshvar@qti.qualcomm.com

***Abstract*—This paper considers the problem of coding over a discrete memoryless channel (DMC) with noiseless feedback. The paper provides a stochastic control view of a variable-length version of the posterior matching scheme which is analyzed via a recently proposed symmetrized divergence, termed Extrinsic Jensen–Shannon (EJS) divergence. In particular, under the variable-length posterior matching scheme, the EJS divergence can be lower bounded by the Shannon capacity of the DMC, which can be used for a relatively simple proof that the variable-length posterior matching scheme achieves capacity.**

I. Introduction

In [1], [2], see also [3], a sequential, one-phase scheme for transmission over a BSC with noiseless feedback was proposed. This scheme is briefly explained next. Each message is represented as a subinterval of size $\frac{1}{M}$ of the unit interval. After each transmission and given the channel output, the posterior probability of all subintervals are updated. In the next time slot, the transmitter sends 0 if the true message's corresponding subinterval is below the current median, or 1 if it is above. If the current median lies within the true message's subinterval, then the transmitter sends 0 with probability equal to the fraction of the interval above the median and 1 otherwise. As the rounds of transmission proceed, the posterior probability of the true message's subinterval most likely grows larger than $\frac{1}{2}$, which pushes the median within the message's subinterval and thus leads to a randomized encoding. Although this simple one-phase scheme was believed to achieve the capacity of a BSC, a rigorous proof remained illusive prior to the work by Shayevitz and Feder [3]. They generalized the described scheme to arbitrary DMCs (satisfying some mild conditions) and proved that their general scheme, named *posterior matching scheme*, achieves capacity [3]. Recently, Li and El Gamal proposed a related scheme [4] with a greatly improved error-exponent, i.e. with exponentially smaller probability of error than the posterior matching scheme.

In [5], we introduced the *Extrinsic Jensen–Shannon (EJS) divergence* as a tool to analyze error exponents and achievable rates for variable-length schemes. In this paper we show that this tool allows for a relatively simple proof that a variable-length version of the posterior matching scheme achieves the capacity of DMCs.

We finish this section with some notation.

Notation: Let $[x]^+ = \max\{x, 0\}$. The i^{th} element of vector $\boldsymbol{v}$ is denoted by v_i. The notations A^t and a^t stand for the tuples $[A_0, \ldots, A_t]$ and $[a_0, \ldots, a_t]$, respectively, for positive integer t. For any set $\mathcal{S}$, $|\mathcal{S}|$ denotes the cardinality of $\mathcal{S}$ and $\mathcal{S}^t$ its t-fold Cartesian product. All logarithms are in base 2. The entropy function on a vector $\boldsymbol{\rho} = [\rho_1, \rho_2, \ldots, \rho_M] \in [0,1]^M$ is defined as $H(\boldsymbol{\rho}) := \sum_{i=1}^M \rho_i \log \frac{1}{\rho_i}$, with the convention that $0 \log \frac{1}{0} = 0$. We denote the conditional probability $P(Y|X = x)$ by P_x.

II. Preliminaries: Symmetric Divergences

We first recall that the *Kullback–Leibler (KL) divergence* between two probability distributions P_Y and P'_Y over a finite set $\mathcal{Y}$ is defined as $D(P_Y \| P'_Y) := \sum_{y \in \mathcal{Y}} P_Y(y) \log \frac{P_Y(y)}{P'_Y(y)}$ with the convention $0 \log \frac{0}{a} = 0$ and $b \log \frac{b}{0} = \infty$ for $a, b \in [0,1]$ with $b \neq 0$. The KL divergence is *not* symmetric, i.e., in general $D(P_Y \| P'_Y) \neq D(P'_Y \| P_Y)$.

The *J divergence* [6] and *L divergence* [7] symmetrize the KL divergence:

$$J(P_1, P_2) := D(P_1 \| P_2) + D(P_2 \| P_1), \tag{1}$$

$$L(P_1, P_2) := D\Big(P_1 \| \frac{1}{2}P_1 + \frac{1}{2}P_2\Big) + D\Big(P_2 \| \frac{1}{2}P_1 + \frac{1}{2}P_2\Big). \tag{2}$$

The *Jensen–Shannon (JS) divergence* [7], [8] is defined similarly to the L divergence but for general $M \geq 2$ probability distributions. Given M probability distributions $P_1, P_2 \ldots, P_M$ over a set $\mathcal{Y}$ and a vector of a priori weights $\boldsymbol{\rho} = [\rho_1, \rho_2, \ldots, \rho_M]$, where $\boldsymbol{\rho} \in [0,1]^M$ and $\sum_{i=1}^M \rho_i = 1$, the JS divergence is defined as [7], [8]:

$$\begin{aligned} JS(\boldsymbol{\rho}; P_1, \ldots, P_M) &:= \sum_{i=1}^M \rho_i D\Big(P_i \| \sum_{j=1}^M \rho_j P_j\Big) \\ &= I(\theta; Y) \end{aligned} \tag{3}$$

where θ is a random variable that takes values in $\{1, 2, \ldots, M\}$ and has probability mass function $\boldsymbol{\rho}$ and $Y \sim P_\theta$ (which implies that $\Pr(Y = y) = \sum_{i=1}^M \rho_i P_i(y)$).

Similarly, one can consider the *Extrinsic Jensen–Shannon (EJS) divergence* [5] which extends the J divergence to general $M \geq 2$ probability distributions. For distributions $P_1, \ldots, P_M$ and an M-dimensional weight vector $\boldsymbol{\rho}$,

$$EJS(\boldsymbol{\rho}; P_1, \ldots, P_M) := \sum_{i=1}^M \rho_i D\Big(P_i \| \sum_{j \neq i} \frac{\rho_j}{1 - \rho_i} P_j\Big), \tag{4a}$$

when $\rho_i < 1$ for all $i \in \{1, \ldots, M\}$, and

$$EJS(\boldsymbol{\rho}; P_1, \ldots, P_M) := \max_{j \neq i} D(P_i \| P_j) \tag{4b}$$

when $\rho_i = 1$ for some $i \in \{1, \ldots, M\}$.

III. Coding over DMC with Noiseless Feedback

A. The Problem Setup

Consider the problem of variable-length coding over a discrete memoryless channel (DMC) with noiseless feedback as depicted in Fig. 1. The DMC is described by finite input

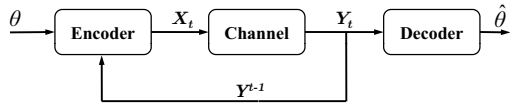

Fig. 1. A noisy memoryless channel with a noiseless causal feedback link.

and output sets $\mathcal{X}$ and $\mathcal{Y}$, and a collection of conditional probabilities $P(Y|X)$. To simplify notation, and without loss of generality, we assume that

$$\mathcal{X} = \{0, 1, \ldots, |\mathcal{X}| - 1\} \tag{5}$$

and

$$\mathcal{Y} = \{0, 1, \ldots, |\mathcal{Y}| - 1\}. \tag{6}$$

Let τ denote the total transmission time (or equivalently the total length of the code). In this paper, our focus is on variable-length coding, i.e., the case where τ is a random stopping time decided at the receiver as a function of the observed channel outputs. (A specific stopping rule is described later in this section.) Thanks to the noiseless feedback, the transmitter is also informed of the channel outputs and the stopping time.

The transmitter wishes to communicate a message θ to the receiver, where the message is uniformly distributed over a message set

$$\Omega := \{1, 2, \ldots, M\}. \tag{7}$$

To this end, it produces channel inputs X_t for $t = 0, 1, \ldots, \tau - 1$, which it can compute as a function of the message θ and (thanks to the feedback) also of the past channel outputs $Y^{t-1} := [Y_0, Y_1, \ldots, Y_{t-1}]$:

$$X_t = e_t(\theta, Y^{t-1}), \quad t = 0, 1, \ldots, \tau - 1, \tag{8}$$

for some encoding function $e_t \colon \Omega \times \mathcal{Y}^t \to \mathcal{X}$.

To describe the encoding process, we shall also use the functions $\{\gamma_{y^{t-1}}\}$ for $y^{t-1} \in \mathcal{Y}^t$ and $t \in \{0, 1, \ldots, \tau - 1\}$ where

$$\gamma_{y^{t-1}} \colon \Omega \to \mathcal{X} \tag{9a}$$

$$i \mapsto e_t(i, y^{t-1}). \tag{9b}$$

Where it is clear from the context and to simplify notation, we omit the subscript y^{t-1} and simply write γ.

We will particularly be interested in *randomized* encoding rules. In this case the encoding is described by the *random* encoding functions $\{\Gamma_{y^{t-1}}\}$ whose realizations $\gamma_{y^{t-1}}$ are of the form in (9). Again, for notational convenience we omit the subscript y^{t-1} where it is clear from the context.

After observing the τ channel outputs $Y_0, Y_1, \ldots, Y_{\tau-1}$, the receiver performs optimum maximum-likelihood decoding and produces as its guess the message with the highest posterior:

$$\hat{\theta} = \arg\max_{i \in \Omega} \rho_i(\tau), \tag{10}$$

where for each positive t and each $i \in \Omega$:

$$\rho_i(t) := \Pr(\theta = i | Y^{t-1}). \tag{11}$$

The probability of error is

$$\mathrm{Pe} := \Pr(\hat{\theta} \neq \theta). \tag{12}$$

For a fixed DMC and for a given $\epsilon > 0$, the goal is to find an encoding rule as in (8) and a stopping rule τ such that combined with the decoding rule in (10) the probability of error satisfies $\mathrm{Pe} \leq \epsilon$ and the expected number of channel uses $\mathbb{E}[\tau]$ is minimized.

Throughout the paper we assume that the receiver applies the following possibly suboptimal stopping rule

$$\tau := \min\{t : \max_{i \in \Omega} \rho_i(t) \geq 1 - \epsilon\}, \tag{13}$$

where $\epsilon > 0$ is the desired probability of error.

The main interest in this paper is in achieving the *capacity* of DMCs with a variable-length scheme. The capacity is defined as follows. If for any small numbers $\delta > 0$, $0 \leq \epsilon < 1$ and all sufficiently large positive integers ℓ an encoding scheme γ (or Γ) can transmit one out of M_ℓ equiprobable messages so that with the ML decoder in (10) and an appropriate stopping rule τ,

$$\mathrm{Pe} \leq \epsilon \tag{14a}$$

$$M_\ell \geq 2^{\ell(R-\delta)} \tag{14b}$$

$$\mathbb{E}[\tau] \leq \ell, \tag{14c}$$

for some positive real number R, then we say that the scheme achieves rate R. The capacity is the supremum over all achievable rates and is given by

$$C := \max_{P_X} I(X; Y), \tag{15}$$

as in the case of fixed-length coding.

B. Stochastic Control View

The problem of coding with noiseless feedback is a decentralized team problem with two agents (the encoder and the decoder) and non-classical information structure [9]. Appealing to [10], the problem can be interpreted as a special case of active hypothesis testing in which a (fictitious) Bayesian decision-maker is responsible to enhance his information about the correct message in a speedy manner by sequentially sampling from conditionally independent observations at the output of the channel (given the input). Here the (fictitious) decision maker has access to the channel output symbols causally (common observations) and is responsible to control the conditional distribution of the observations given the true message (private observation) by selecting encoding functions for the encoder which map the message θ to the input symbols

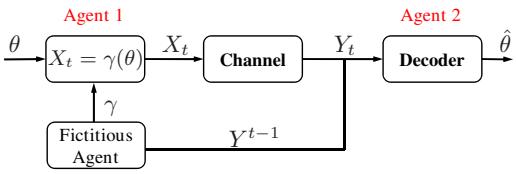

Fig. 2. Two-agent problem with common and private observations from the point of view of the fictitious agent.

of the channel. In other words, as also observed in [11], the problem can be viewed as a (centralized) partially observable Markov decision problem (POMDP) with (static) state space Ω and the observation space $\mathcal{Y}$. Let $\mathcal{E} := \{\gamma(\cdot) : \Omega \to \mathcal{X}\}$ be the set of all mappings from Ω to $\mathcal{X}$. The action space (for the fictitious agent) becomes $\mathcal{E} \cup \{T\}$ where T denotes the termination of the transmission phase, hence the realization of the stopping time τ.

Casting the problem as a POMDP allows for the structural characterization of the information state, also known as sufficient statistics: The decision maker's posteriors about the messages collectively,

$$\boldsymbol{\rho}(t) := [\rho_1(t), \rho_2(t), \ldots, \rho_M(t)], \tag{16}$$

form a sufficient statistics for our (fictitious) Bayesian decision maker. Furthermore, this (fictitious) decision maker's posterior at any time t coincides with the receiver's posterior and, thanks to the perfect feedback, is available to the transmitter. In other words, the selection of encoding and decoding rules as a function of this posterior incurs no loss of optimality [12].

We also note that the dynamics of the information state, i.e. the posterior, follows Bayes' rule. More specifically, given an encoding function γ at time t and an information state $\boldsymbol{\rho}$, the conditional distribution of the next channel output Y_t, given the past observation Y^{t-1}, is

$$P_{\boldsymbol{\rho}}(y) = \sum_{i=1}^{M} \rho_i P(Y = y | X = \gamma(i)).$$

Similarly, given also the output symbol $Y_t = y$, according to Bayes' rule, the posterior at time $t+1$ is:

$$\boldsymbol{\rho}(t+1) = \left[\frac{\rho_1 P_{\gamma(1)}(y)}{P_{\boldsymbol{\rho}}(y)}, \ldots, \frac{\rho_M P_{\gamma(M)}(y)}{P_{\boldsymbol{\rho}}(y)} \right].$$

This stochastic control view of the problem, suggests an achievability analysis which generalizes the approach of [11] beyond mutual information and is based on symmetric divergence associate with the belief state $\boldsymbol{\rho}$ and $\{P_x\}_{x \in \mathcal{X}}$. In the sections that follow, we utilize this approach with respect to the EJS divergence induced by the posterior matching. This allows us to provide a concise achievability analysis for variable-length posterior matching.

IV. Main Result

Consider the variable-length version of the posterior matching encoding in [3]:

At each time $t = 0, 1, \ldots, \tau - 1$, if $\theta = i$ and given the posterior vector $\boldsymbol{\rho}(t)$, the input $X(t)$ takes value in the set

$$\mathcal{X}_i(t) := \Big\{ x \in \mathcal{X} \colon \sum_{i'=1}^{i-1} \rho_{i'}(t) < \sum_{x' \le x} \pi^\star_{x'} \text{ and } \sum_{x' < x} \pi^\star_{x'} \le \sum_{i'=1}^{i} \rho_{i'}(t) \Big\};$$

where each value $x \in \mathcal{X}_i(t)$ is taken with probability

$$\Pr\big(X(t) = x | \theta = i, Y^{t-1} = y^{t-1}\big) = \frac{\min\Big\{ \sum_{i'=1}^{i} \rho_{i'}(t), \sum_{x' \le x} \pi^\star_{x'} \Big\} - \max\Big\{ \sum_{i'=1}^{i-1} \rho_{i'}(t), \sum_{x' < x} \pi^\star_{x'} \Big\}}{\rho_i(t)}.$$

We show that the described posterior matching encoding Γ^{PM} combined with the ML decoding in (10) and stopping rule (13) achieves capacity for all DMCs satisfying a mild condition. Let C_1 be the KL divergence between the two most distinguishable inputs of the DMC:

$$C_1 := \max_{x, x' \in \mathcal{X}} D(P(Y|X = x) \| P(Y|X = x')). \tag{18}$$

Theorem 1. *The described posterior matching encoding* Γ^{PM} *combined with the optimal ML decoder* (10) *and stopping rule* (13) *achieve the capacity of any DMC where* C *and* C_1 *are positive and finite.*[1]

Proof: For a fixed encoding rule γ and given a sequence of channel outputs y^{t-1} with corresponding posteriors $\boldsymbol{\rho}(t)$, we define $EJS(\boldsymbol{\rho}(t), \gamma)$ to be the EJS divergence between the conditional output distributions $P_{\gamma(1)}, \ldots, P_{\gamma(M)}$ with weight vector $\boldsymbol{\rho}(t)$:

$$EJS(\boldsymbol{\rho}(t), \gamma) := EJS\big(\boldsymbol{\rho}(t); P_{\gamma(1)}, \ldots, P_{\gamma(M)}\big). \tag{19}$$

For a randomized encoding function Γ, we use

$$EJS(\boldsymbol{\rho}(t), \Gamma) := \sum_{\gamma \in \mathcal{E}} \Pr(\Gamma = \gamma | Y^{t-1} = y^{t-1}) EJS(\boldsymbol{\rho}(t), \gamma)$$

where recall that $\mathcal{E}$ denotes the set of all possible encoding functions. Let $\tilde{\rho} := 1 - (1 + \max\{\log M, \log \frac{1}{\epsilon}\})^{-1}$.

Our proof is based on the following theorem from [5]:

Theorem 2 (Corollary 2 in [5])**.** *Consider a DMC with* $C > 0$ *and* $C_1 < \infty$ *and a variable-length encoding* Γ *combined with the ML decoding in* (10) *and the stopping rule* (13)*. If for any time* $t < \tau$ *and for any posterior vector* $\boldsymbol{\rho}(t)$*,*

$$EJS(\boldsymbol{\rho}(t), \Gamma) \ge C, \tag{20a}$$

then the scheme achieves the capacity C *of the channel. Furthermore, if also,*

$$EJS(\boldsymbol{\rho}(t), \Gamma) \ge \tilde{\rho} C_1 \qquad \text{if } \max_{i \in \Omega} \rho_i(t) \ge \tilde{\rho}, \tag{20b}$$

then it also achieves the channel's optimal reliability function.

[1] Notice that $C \le C_1$ and $C_1 < \infty$ if, and only if, $P(Y = y|X = x)$ is positive for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.

Theorem 1 can thus be shown by proving that Condition (20a) is satisfied for the posterior matching encoding $\Gamma = \Gamma^{\mathrm{PM}}$:

$$EJS(\boldsymbol{\rho}(t), \Gamma^{\mathrm{PM}}) \geq C. \tag{21}$$

Fix a time instant t and $Y^{t-1} = y^{t-1}$. For ease of notation, in the following we drop the time index t for $\rho_i(t)$ and simply write ρ_i. Let

$$\lambda_\gamma := \Pr(\Gamma^{\mathrm{PM}} = \gamma | Y^{t-1} = y^{t-1}), \qquad \gamma \in \mathcal{E}.$$

Define for each $i \in \Omega$ and $x \in \mathcal{X}$:

$$\Lambda_{i,x} := \sum_{\gamma\colon \gamma(i)=x} \lambda_\gamma = \Pr(X = x | \theta = i, Y^{t-1} = y^{t-1}) \tag{22}$$

and

$$\hat{\rho}_{i,x} := \rho_i \Lambda_{i,x} = \Pr(X = x, \theta = i | Y^{t-1} = y^{t-1}). \tag{23}$$

For a fixed posterior distribution, the various messages are mapped into inputs of the channel independently of each other and hence, for $x, x' \in \mathcal{X}$ and $i, j \in \Omega$ where $i \neq j$

$$\sum_{\gamma\colon \substack{\gamma(i)=x \\ \gamma(j)=x'}} \lambda_\gamma = \Lambda_{i,x}\Lambda_{j,x'}. \tag{24}$$

Let $\pi_0^\star, \ldots, \pi_{|\mathcal{X}|-1}^\star$ denote the capacity-achieving input distribution, i.e., the maximizer of (15). Rearranging terms and using Jensen's inequality and the convexity of the KL divergence, we have

$$\begin{aligned}
&EJS(\boldsymbol{\rho}(t), \Gamma^{\mathrm{PM}}) \\
&= \sum_{\gamma\in\mathcal{E}} \lambda_\gamma \sum_{i=1}^{M} \rho_i D\Big(P_{\gamma(i)} \Big\| \sum_{j\neq i} \frac{\rho_j}{1-\rho_i} P_{\gamma(j)}\Big) \\
&= \sum_{i=1}^{M} \rho_i \sum_{x\in\mathcal{X}} \sum_{\gamma\colon \gamma(i)=x} \lambda_\gamma D\Big(P_x \Big\| \sum_{j\neq i} \frac{\rho_j}{1-\rho_i} P_{\gamma(j)}\Big) \\
&\geq \sum_{i=1}^{M} \sum_{x\in\mathcal{X}} \rho_i \Lambda_{i,x} D\Big(P_x \Big\| \sum_{j\neq i} \frac{\rho_j}{1-\rho_i} \sum_{\gamma\colon \gamma(i)=x} \frac{\lambda_\gamma}{\Lambda_{i,x}} P_{\gamma(j)}\Big) \\
&= \sum_{i=1}^{M} \sum_{x\in\mathcal{X}} \hat{\rho}_{i,x} D\Big(P_x \Big\| \sum_{j\neq i} \frac{\rho_j}{1-\rho_i} \sum_{x'\in\mathcal{X}} \sum_{\gamma\colon \substack{\gamma(i)=x \\ \gamma(j)=x'}} \frac{\lambda_\gamma}{\Lambda_{i,x}} P_{x'}\Big) \\
&\overset{(a)}{=} \sum_{i=1}^{M} \sum_{x\in\mathcal{X}} \hat{\rho}_{i,x} D\Big(P_x \Big\| \frac{\sum_{j\neq i}\sum_{x'\in\mathcal{X}} \rho_j \Lambda_{j,x'} P_{x'}}{1-\rho_i}\Big) \\
&= \sum_{i=1}^{M} \sum_{x\in\mathcal{X}} \hat{\rho}_{i,x} D\Big(P_x \Big\| \frac{\sum_{x'\in\mathcal{X}} (\pi_{x'}^\star P_{x'} - \hat{\rho}_{i,x'} P_{x'})}{1-\rho_i}\Big), \\
&= \sum_{i=1}^{M} \sum_{x\in\mathcal{X}} \hat{\rho}_{i,x} D\Big(P_x \Big\| \frac{\sum_{x'\in\mathcal{X}} (\pi_{x'}^\star P_{x'} - \hat{\rho}_{i,x'} P_{x'})}{1-\rho_i}\Big) \\
&\quad + \sum_{i=1}^{M} \sum_{x\in\mathcal{X}} \hat{\rho}_{i,x} \frac{\rho_i}{1-\rho_i} D\Big(P_x \Big\| \frac{\sum_{x'} \hat{\rho}_{i,x'} P_{x'}}{\rho_i}\Big) \\
&\quad - \sum_{i=1}^{M} \sum_{x\in\mathcal{X}} \hat{\rho}_{i,x} \frac{\rho_i}{1-\rho_i} D\Big(P_x \Big\| \frac{\sum_{x'} \hat{\rho}_{i,x'} P_{x'}}{\rho_i}\Big)
\end{aligned}$$

$$\begin{aligned}
&\geq \sum_{i=1}^{M} \sum_{x\in\mathcal{X}} \frac{\hat{\rho}_{i,x}}{1-\rho_i} D\Big(P_x \Big\| \sum_{x'\in\mathcal{X}} \pi_{x'}^\star P_{x'}\Big) \\
&\quad - \sum_{i=1}^{M} \frac{\rho_i^2}{1-\rho_i} \sum_{x\in\mathcal{X}} \Lambda_{i,x} D\Big(P_x \Big\| \sum_{x'\in\mathcal{X}} \Lambda_{i,x'} P_{x'}\Big) \\
&\overset{(b)}{\geq} \sum_{i=1}^{M} \sum_{x\in\mathcal{X}} \frac{\hat{\rho}_{i,x}}{1-\rho_i} C - \sum_{i=1}^{M} \frac{\rho_i^2}{1-\rho_i} C \\
&= \sum_{i=1}^{M} \frac{\rho_i}{1-\rho_i} C - \sum_{i=1}^{M} \frac{\rho_i^2}{1-\rho_i} C \\
&= C
\end{aligned} \tag{25}$$

where (a) follows from (24); and where (b) follows from [13, Theorem 4.5.1] because $\hat{\rho}_{i,x} > 0$ only if $\pi_x^\star > 0$ and from the fact that $\sum_{x\in\mathcal{X}} \Lambda_{i,x} D\Big(P_x \Big\| \sum_{x'\in\mathcal{X}} \Lambda_{i,x'} P_{x'}\Big) = I(X;Y) \leq C$ when X denotes an input with probability mass function $\{\Lambda_{i,x}\}_{x\in\mathcal{X}}$ and Y the output produced by the channel. ■

Future Work

In future work, using large-deviation analysis, we plan to extend our EJS-divergence based proof technique to the original fixed-length posterior matching scheme.

References

[1] M. Horstein, "Sequential transmission using noiseless feedback," *IEEE Transactions on Information Theory*, vol. 9, no. 3, pp. 136–143, July 1963.

[2] M. V. Burnashev and K. S. Zigangirov, "An interval estimation problem for controlled observations," *Problemy Peredachi Informatsii*, vol. 10, no. 3, pp. 51–61, 1974.

[3] O. Shayevitz and M. Feder, "Optimal feedback communication via posterior matching," *IEEE Transactions on Information Theory*, vol. 57, no. 3, pp. 1186–1222, March 2011.

[4] C. T. Li and A. E. Gamal, "An efficient feedback coding scheme with low error probability for discrete memoryless channels," Nov. 2013, available on http://arxiv.org/pdf/1311.0100v2.

[5] M. Naghshvar, T. Javidi, and M. Wigger, "Extrinsic Jensen–Shannon divergence: Applications to variable-length coding," available on arXiv: 1307.0067.

[6] H. Jeffreys, "An invariant form for the prior probability in estimation problems," *Proc. Roy. Soc. London. Ser. A.*, vol. 186, pp. 453–461, 1946.

[7] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 145 –151, January 1991.

[8] J. Burbea and C. R. Rao, "On the convexity of some divergence measures based on entropy functions," *IEEE Trans. Inform. Theory*, vol. 28, no. 3, pp. 489–495, 1982.

[9] H. S. Witsenhausen, "A counterexample in stochastic optimum control," *SIAM Journal on Control*, vol. 6, no. 1, pp. 131–147, 1968.

[10] A. Mahajan, A. Nayyar, and D. Teneketzis, "Identifying tractable decentralized problems on the basis of information structures," in *Proceedings of the 46th Allerton conference on communication, control, and computing*, 2008, pp. 1440–1449.

[11] T. P. Coleman, "A stochastic control viewpoint on 'posterior matching'-style feedback communication schemes," in *IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2009, pp. 1520–1524.

[12] D. Bertsekas and J. Tsitsiklis, *Neuro-Dynamic Programming*. Athena Scientific, Belmont, Massachusetts, 1996.

[13] R. G. Gallager, *Information theory and reliable communication*. John Wiley & Sons, Inc., New York, 1968.