[20] G. L. Besenerais, J.-F. Bercher, and G. Demoment, "A new look at entropy for solving linear inverse problems," *IEEE Trans. Inf. Theory*, vol. 45, no. 5, pp. 1565–1577, Jul. 1999.
[21] I. Csiszár, F. Gamboa, and E. Gassiat, "MEM pixel correlated solution for generalized moment and interpolation problems," *IEEE Trans. Inf. Theory*, vol. 45, no. 7, pp. 2253–2270, Nov. 1999.
[22] S.-P. Han and O. L. Mangasarian, "Exact penalty functions in nonlinear programming," *Math. Prog.*, vol. 17, pp. 251–269, 1979.
[23] P. E. Gill, W. Murray, and M. H. Wright, *Practical Optimization*. San Diego, CA: Academic, 1981.
[24] O. L. Mangasarian, *Nonlinear Programming*. ser. Classics in Applied Mathematics, G. Golub, Ed. Philadelphia, PA: SIAM, 1994, vol. 10. Originally published: New York, McGraw-Hill, 1969.
[25] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York: Cambridge Univ. Press, 2004.
[26] T. Pietrzykowski, "An exact potential method for constrained maxima," *SIAM J. Numer. Anal.*, vol. 6, pp. 262–304, 1969.
[27] D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*. New York: Academic, 1982.
[28] R. Fletcher, *Practical Methods of Optimization*, 2nd ed. New York: Wiley, 1987.
[29] C. S. Peirce, *The Philosophy of Peirce: Selected Writings*. London, U.K.: Jarrolds, 1956.
[30] W. Kneale, *Probability and Induction*. Oxford, U.K.: Clarendon, 1949.
[31] M. Gupta, "An Information Theory Approach to Supervised Learning," Ph.D. dissertation, Stanford Univ., Stanford, CA, 2003.
[32] M. R. Gupta, R. M. Gray, and R. A. Olshen, "Nonparametric supervised learning with linear interpolation and maximum entropy," *IEEE Trans. Pattern Anal. Machine Intell.*. Available [Online] at ee.washington.edu/research/guptalab/publications.html, to be published.
[33] M. R. Gupta and R. M. Gray, "Reducing bias in supervised learning," in *Proc. IEEE Workshop on Statistical Signal Processing*, St. Louis, MO, Sep. 2003, pp. 482–485.
[34] D. O'Brien, M. Gupta, and R. M. Gray, "Analysis and classification of internal pipeline images," in *Proc. Int. Conf. Image Proc.*, Barcelona, Spain, Sep. 2003.
[35] M. R. Gupta, "Inverting color transforms," in *Proc. SPIE Conf. Computational Imaging*, vol. 5299, San Jose, CA, Jan. 2004, pp. 83–93.
[36] M. R. Gupta, S. Upton, and J. Bowen, "Simulating the effect of illumination using color transformations," in *Proc. SPIE Conf. Computational Imaging*, vol. 5674, San Jose, CA, Jan. 2005, pp. 248–258.
[37] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. New York: Springer-Verlag, 2001.
[38] S. Kulkarni, G. Lugosi, and S. S. Venkatesh, "Learning pattern classification—A survey," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2178–2206, Oct. 1998.
[39] E. Fix and J. L. Hodges, "Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties," uSAF School of Aviation Medicine, TX, Tech. Rep. 4, 1951.
[40] L. Devroye, L. Gyorfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. New York: Springer-Verlag, 1996.
[41] C. Stone, "Consistent nonparametric regression," *Ann. Statist.*, vol. 5, no. 4, pp. 595–645, 1977.
[42] N. N. Čencov, *Statistical Decision Rules and Optimal Inference*. Providence, RI: Amer. Math. Soc., 1982. Translated from Russian by the Israel Program for Scientific Translations. Translation edited by L. J. Leifman.

Error Exponents for Finite-Hypothesis Channel Identification

Patrick Mitran, *Student Member, IEEE,* and Aleksandar Kavčić, *Member, IEEE*

***Abstract*—We consider the problem of designing optimal probing signals for finite-hypothesis testing. Equivalently, we cast the problem as the design of optimal channel input sequences for identifying a discrete channel under observation from a finite set of known channels. The optimality criterion that we employ is the exponent of the Bayesian probability of error. In our study, we consider a feedforward scenario where there is no feedback from the channel output to the signal selector at the channel input and a feedback scenario where the past channel outputs are revealed to the signal selector.**

In the feedforward scenario, only the type of the input sequence matters and our main result is an expression for the error exponent in terms of the limiting distribution of the input sequence. In the feedback case, we show that when discriminating between two channels, the optimal scheme in the first scenario is simultaneously the optimal time-invariant Markov feedback policy of *any* order.

***Index Terms*—Bayesian hypothesis testing, classification, detection theory, Chernoff's theorem, error exponent, feedback, method of types, sequential detection, signal selection, waveform selection.**

I. Introduction

In traditional hypothesis testing, we are given a set of hypotheses $\mathcal{H}$. For each hypothesis $h \in \mathcal{H}$, we know the probability law for an observable variable Y, i.e., we know $P_Y^h[y] \triangleq P_{Y \mid H}[y \mid h]$. We make n observations $y_1^n = [y_1, y_2, \ldots, y_n]$ and based on these observations, we need to infer the hypothesis $h \in \mathcal{H}$. This is a well-known problem with well-known solutions in the context of Bayesian and Neyman–Pearson decision making [14]. Furthermore, the type-II error exponent (for Neyman–Pearson) or average error exponent (for Bayesian) detection is well known and may be derived by the method of types [5], [4] or large deviation theory [8].

Let us now suppose that the observable variable Y is obtained as the response to an input variable X, which we control. In particular, each hypothesis h, drawn from a finite set of hypotheses $\mathcal{H}$, may be viewed as a memoryless channel $P_{Y \mid X}^h[y \mid x] \triangleq P_{Y \mid X,H}[y \mid x, h]$. We refer to this type of problem as a *finite-hypothesis channel identification* or *channel detection* problem. The objective is to choose a set of input signals $x_1^n = [x_1, x_2, \ldots, x_n]$ according to some policy. We will consider two broad classes of policies.

Open-Loop Policies: We transmit these n signals x_1^n and only after all signals are transmitted, we observe the n responses $y_1^n = [y_1, \ldots, y_n]$. We make a decision on $h \in \mathcal{H}$ after we observe all outputs y_1^n.

Feedback Policies: This case may be described as follows. At time $t = 1$, an input x_1 is chosen according to some policy and sent over the channel. Based on the observation of the response output y_1, a new input x_2 is chosen. The signal x_2 is transmitted and a response y_2 is observed. Based on knowledge of x_1, x_2, y_1, and y_2, an input x_3 is

Manuscript received June 4, 2005; revised September 13, 2005. The materiral in this correspondence was presented in part at the IEEE International Symposium on Information Theory, Chicago, IL, June/July 2004, and the International Symposium on Information Theory and Its Applications, Parma, Italy, October 2004.

The authors are with the Division of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138 USA (e-mail: mitran@deas.harvard.edu; kavcic@hrl.harvard.edu).

Communicated by X. Wang, Associate Editor for Detection and Estimation.
Digital Object Identifier 10.1109/TIT.2005.860468

generated and so on. Thus, after each observed output y_{t-1}, there is a chance to refine the transmitted input x_t.

The objective of this correspondence is to determine policies for generating input sequences x_1^n that optimize some measure of hypothesis detection error for the cases outlined above. The measure of error that we use is the exponent of the average probability of error. In this light, our results are related to the error exponent for Bayesian hypothesis testing (Chernoff's theorem [4]).

We contrast this with the related, but different, problem of traditional channel identification. In traditional channel identification, one is typically given a class of channels, *parameterized* by some continuous vector (e.g., an unknown impulse response) (see [1], [11], [13]). There, the objective is to *estimate* these parameters for a particular channel under observation.

We also contrast the second class of policies (feedback policies) with sequential detection [3]. In a traditional sequential detection framework, one seeks to minimize the probing time for a given probability of error by designing appropriate stopping rules. In this work, the probing time is fixed and we seek to minimize the probability of error by appropriate design of probing signals. This may be more appropriate in scenarios where the phenomenon we are probing is only available for a limited time beyond our control, e.g., a hostile target. Furthermore, traditional sequential detection does not produce a signal selection strategy. Nevertheless, Chernoff has suggested a signal selection strategy in the context of sequential detection [3], [12]. In the context of discriminating between two hypotheses h_1 and h_2, this strategy may be described as follows. If at time t hypothesis h_1 is more likely given the past observations, choose the input x_t that maximizes the Kullback–Leibler divergence $D(P^{h_1}_{Y\mid X}[Y \mid x_t] \,\|\, P^{h_2}_{Y\mid X}[Y \mid x_t])$. A similar procedure is employed for h_2. However, since this algorithm was not explicitly designed to minimize the Bayesian error probability for a fixed probing time, it is unlikely to be optimal in the Bayesian setting. In particular, for a fixed x_t, the Kullback–Leibler divergence $D(P^{h_1}_{Y\mid X}[Y \mid x_t] \,\|\, P^{h_2}_{Y\mid X}[Y \mid x_t])$ is the optimal error exponent for *incorrectly* detecting h_1 given h_2 is true in a setting where a small but nonzero probability of error is tolerated for incorrectly detecting h_2 given h_1 is true.

To illustrate this point, we shall see by way of an example that in the context of Bayesian discrimination between two hypotheses, Chernoff's approach may be outperformed by a signal selection strategy that effectively employs no feedback at all. This strategy is correctly predicted by our results.

Furthermore, while it may be argued that a Bayesian setting puts undue importance on the prior probabilities of the hypotheses, this is in fact not the case with our formulation. While we assume the existence of prior probabilities in our derivations, the exponents that we derive (and the policies that achieve them) are in fact *independent* of the actual prior probabilities (provided none is zero).

This correspondence is structured as follows. In Section II, we introduce the notation, review the method of types, and define the error exponents for our two problems. As we employ the method of types, our results in the main body are directly applicable to discrete inputs and outputs only. In Section III, we prove one of our main results, a theorem on the error exponent for the open-loop case. We also provide an efficient numerical algorithm for the computation of the exponent and the asymptotic sequences that achieve these exponents. In Section IV, we investigate the feedback error exponent and derive the exact performance for a broad class of time-invariant Markov policies (this expression also applies to a class of Markov hypotheses). We show that this expression is an upper bound for *any* time-invariant Markov policy. We then show that in the case of two hypotheses, this expression is itself upper-bounded by the open-loop exponent in Section III. For the case of distinguishing between two hypotheses, this bound is achievable. In Section V, we compare a simple two-hypothesis scenario employing our strategy to that suggested by Chernoff; simulation results clearly show that our scheme attains a better error exponent (without employing any feedback). Section VI provides some closing remarks. Finally, the Appendix provides supporting lemmas of analytic nature relevant to the derivation of the main results.

II. Preliminaries

We consider the inputs x_n and outputs y_n to belong to finite discrete sets $\mathcal{X} = \{x^{(1)}, x^{(2)}, \ldots, x^{(M)}\}$ and $\mathcal{Y} = \{y^{(1)}, y^{(2)}, \ldots, y^{(N)}\}$, respectively. We further assume that each channel is memoryless and time invariant, i.e., $P^h_{Y\mid X}[y_1^n \mid x_1^n] = \prod_{j=1}^n P^h_{Y\mid X}[y_j \mid x_j]$. Also, we assume that $P^h_{Y\mid X}[y \mid x] > 0$ for all $h \in \mathcal{H}, y \in \mathcal{Y}, x \in \mathcal{X}$, which for future reference we abbreviate as $P^h_{Y\mid X} > 0$.

A. The Method of Types

Since sequence types will be of central importance in this correspondence, we now develop our notation for them. We will be interested in two different kinds of types: memoryless and Markov types.

We loosely follow the notation of [4]. Let $\mathcal{E}_X$ be the space of all distributions on X (with topology induced by the usual topology on $\mathbb{R}^n$). We denote by $Q_{x_1^n}$ the memoryless type (empirical distribution) of x_1^n, i.e.,

$$Q_{x_1^n}[a] \triangleq \frac{1}{n}|\{i : x_i = a\}|. \tag{1}$$

Also, $\mathcal{Q}_X^n$ will denote the set of all types $Q_{x_1^n}$. For $Q_X \in \mathcal{Q}_X^n$, let $T(Q_X) = \{x_1^n \in \mathcal{X}^n : Q_{x_1^n} = Q_X\}$ denote the set of all sequences whose type is Q_X. We note that a memoryless type $Q_X \in \mathcal{Q}_X$ is also a distribution on X, i.e., $Q_X \in \mathcal{E}_X$.

The entropy $H(Q_X)$ of a distribution Q_X and the Kullback–Leibler divergence $D(Q_X \,\|\, P_X)$ between two distributions Q_X and P_X are defined as

$$H(Q_X) \triangleq -\sum_{x\in\mathcal{X}} Q_X[x] \log Q_X[x] \tag{2}$$

$$D(Q_X \,\|\, P_X) \triangleq \sum_{x\in\mathcal{X}} Q_X[x] \log \frac{Q_X[x]}{P_X[x]}. \tag{3}$$

The following theorem enumerates some of the key properties of memoryless types.

Theorem 1: Let $Q_X \in \mathcal{Q}_X^n$ be the type of the length-n sequence x_1^n. Let $P_X[x_1^n]$ be a memoryless and time-invariant distribution on x_1^n, i.e., $P_X[x_1^n] = \prod_{i=1}^n P_X[x_i]$. Then

1) $|\mathcal{Q}_X^n| \le (n+1)^{|\mathcal{X}|}$;
2) $P_X[x_1^n] = 2^{-n[D(Q_X \,\|\, P_X)+H(Q_X)]}$;
3) $\frac{1}{|\mathcal{Q}_X^n|} 2^{nH(Q_X)} \le |T(Q_X)| \le 2^{nH(Q_X)}$. □

Proof: The proof is given in [6]. ■

Similarly, we denote by Q_{XY} and $\mathcal{E}_{XY}$ the joint type of two sequences (x_1^n, y_1^n) and the space of joint distributions on (X, Y), respectively. Likewise, we define $\mathcal{E}_{Y\mid X}$ to be the space of conditional distributions on Y given X and $Q_{Y\mid X}$ to be the conditional distribution on Y given X obtained from Q_{XY}. We shall use notation like

$$T(Q_{Y\mid X})(x_1^n) = \{y_1^n : (x_1^n, y_1^n) \in T(Q_{XY})\}$$

where the parent Q_{XY} of $Q_{Y\mid X}$ will always be clear from the context.

The Kullback–Leibler divergence between a distribution Q_{XY} and a conditional distribution $P_{Y\mid X}$ is defined as

$$D(Q_{XY}\,\|\,P_{Y\mid X}) \triangleq \sum_{x\in\mathcal{X},y\in\mathcal{Y}} Q_{XY}[x,y]\log\frac{Q_{XY}[x,y]}{P_{Y\mid X}[y\mid x]\sum_{y'\in\mathcal{Y}}Q_{XY}[x,y']}. \quad (4)$$

If Q_{XY} has a known factorization $Q_{XY}=Q_{Y\mid X}\times Q_X$, this will frequently be written as $D(Q_{Y\mid X}\,\|\,P_{Y\mid X}\mid Q_X)\triangleq D(Q_{XY}\,\|\,P_{Y\mid X})$.

The following theorem enumerates two key properties of joint types.

Theorem 2: Let $Q_{XY}\in\mathcal{Q}^n_{XY}$ be the joint type of the length-n sequences x_1^n and y_1^n. Let $P_{Y\mid X}[y_1^n\mid x_1^n]$ be a memoryless and time-invariant distribution, i.e.,

$$P_{Y\mid X}[y_1^n\mid x_1^n]=\prod_{i=1}^n P_{Y\mid X}[y_i\mid x_i].$$

Then

1) $P_{Y\mid X}[y_1^n\mid x_1^n]=2^{-n[D(Q_{XY}\,\|\,P_{Y\mid X})+H(Q_{Y\mid X})]}$;
2) $\frac{1}{|\mathcal{Q}^n_X\,\|\,\mathcal{Q}^n_Y|}2^{nH(Q_{Y\mid X})}\le|T(Q_{Y\mid X})|\le 2^{nH(Q_{Y\mid X})}$;

where for notational convenience, we write $H(Q_{Y\mid X})$ in place of $H(Q_{XY})-H(Q_X)$. □

Proof: The proof is given in [6]. ■

In Section IV, we will also employ the notion of circular Markov types [7]. In particular, if we denote by $U_{X_0,\ldots,X_k}=U_{X_0^k}$, the kth-order circular Markov type[1] of a sequence x_1^n, then $U_{X_0^k}$ is a probability mass function (PMF) defined by the relative frequencies

$$U_{X_0^k}[a_0,\ldots,a_k]=\frac{1}{n}\left|\left\{i: x_i^{i+k}=a_0^k, 1\le i\le n\right\}\right| \quad (5)$$

with the cyclic convention that $x_{i+n}=x_i$.

We denote by $\mathcal{U}^n_{X_0^k}$ the set of all kth-order circular Markov types for X sequences of length n. Likewise, $T(U_{X_0^k})$ denotes the set of all length-n sequences X whose Markov type is $U_{X_0^k}$. Finally, we denote by $\mathcal{F}_{X_0^k}$ the space of all distributions $F_{X_0^k}$ on the tuple $(X_0,\ldots,X_k)$ whose marginalization satisfies $F_{X_0^{k-1}}[x_0^{k-1}]=F_{X_1^k}[x_0^{k-1}]$ with topology induced by the usual topology on $\mathbb{R}^n$. We immediately note that $U_{X_0^k}\in\mathcal{F}_{X_0^k}$.

For any $F_{X_0^k}$ in the interior of $(\mathcal{F}_{X_0^k})$ ($\mathcal{F}_{X_0^k}$ viewed as a subset of $\mathbb{R}^n$), we have that $F_{X_0^k}>0$. Hence, the corresponding $F_{X_k\mid X_0^{k-1}}>0$ and $F_{X_k\mid X_0^{k-1}}$ represents an irreducible (hence ergodic, since the state space is finite [2], [9]) kth-order Markov chain. Conversely, for any ergodic Markov chain $F_{X_k\mid X_0^{k-1}}$ there is a unique invariant distribution $F_{X_0^{k-1}}$. Hence, each interior point of $\mathcal{F}_{X_0^k}$ is in one-to-one correspondence with an ergodic Markov chain $F_{X_k\mid X_0^{k-1}}$. Furthermore, for any ergodic Markov chain $F_{X_k\mid X_0^{k-1}}$, we may associate a unique $F_{X_0^k}\in\mathcal{F}_{X_0^k}$ (which need not be in the interior).

If $\Gamma_{X_k\mid X_0^{k-1}}[x_k\mid x_0^{k-1}]=0$ for some x_0^k implies that $U_{X_0^k}[x_0^k]=0$, we shall denote this by the short-hand notation $U_{X_0^k}\ll\Gamma_{X_k\mid X_0^{k-1}}$.

Bounds on the number of Markov types have been proven for first-order Markov types [7]. As stated in Csiszár [5], these readily generalize to any order. We now state the relevant bounds for arbitrary orders, the proof of which is a simple extension of the first order results in [7].

[1]Sometimes Markov types are referred to as higher order types. A $k+1$ higher order type is a kth-order Markov type. We prefer the notation "kth-order Markov type" as it is the type of interest when the sequence is generated by a kth-order Markov chain.

Theorem 3: (Davisson, Longo, and Sgarro): Let x_1^n be a sequence with kth-order Markov type $U_{X_0^k}\in\mathcal{U}^n_{X_0^k}$ and $P[x_1^n]$ a probability mass function on $\mathcal{X}^n$ defined by

$$P[x_1^n]=\mu_{X_1^k}\left[x_1^k\right]\prod_{m=k+1}^n\Gamma_{X_k\mid X_0^{k-1}}\left[x_m\;\middle|\;x_{m-k}^{m-1}\right] \quad (6)$$

with $\mu_{X_1^k}>0$ and $\Gamma_{X_k\mid X_0^{k-1}}>0$. Then, for some $\alpha>0$ and $\beta>0$ (which depend on $\mu_{X_1^k}$ and $\Gamma_{X_k\mid X_0^{k-1}}$, respectively), the following hold:

1) $|\mathcal{U}^n_{X_0^k}|\le(n+1)^{|\mathcal{X}|^{k+1}}$;
2) $$n^{-|\mathcal{X}|^k}(n+1)^{-|\mathcal{X}|^{k+1}}2^{nH(X_k\mid X_0^{k-1})}\le|T(U_{X_0^k})|\le|\mathcal{X}|^k2^{nH(X_k\mid X_0^{k-1})};$$
3) $$\alpha 2^{-n[D(U_{X_0^k}\|\Gamma_{X_k\mid X_0^{k-1}})+H(X_k\mid X_0^{k-1})]}\le P[x_1^n]\le\beta^{-k}2^{-n[D(U_{X_0^k}\|\Gamma_{X_k\mid X_0^{k-1}})+H(X_k\mid X_0^{k-1})]}$$

where for notational convenience, we write $H(X_k\mid X_0^{k-1})$ in place of $H(U_{X_0^k})-H(U_{X_0^{k-1}})$ and $U_{X_0^{k-1}}$ is a marginalization of $U_{X_0^k}$. Furthermore, if $\Gamma_{X_k\mid X_0^{k-1}}[x_k\mid x_0^{k-1}]=0$ for some x_0^k, then the lower bound in 3) still holds. Finally, if $\Gamma_{X_k\mid X_0^{k-1}}$ is irreducible and $U_{X_0^k}\ll\Gamma_{X_k\mid X_0^{k-1}}$, then the upper bound in 3) still holds. □

Unfortunately, Theorem 3 does not provide an upper bound on the probability of a sequence x_1^n when $\Gamma_{X_k\mid X_0^{k-1}}$ is irreducible and $U_{X_0^k}$, the type of x_1^n, does not satisfy $U_{X_0^k}\ll\Gamma_{X_k\mid X_0^{k-1}}$. The difficulty here is that the sequence x_1^n (or its cyclic extension) contains transitions which are forbidden by $\Gamma_{X_k\mid X_0^{k-1}}$. There are two scenarios in which these may occur. First, a forbidden transition may occur in the sequence x_1^n and is not due to the cyclic extension by which $U_{X_0^k}$ is derived. In this case, the upper bound in part 3) of Theorem 3 is still valid since $P[x_1^n]=0$. In the case that the only forbidden transitions in $U_{X_0^k}$ are due to the cyclic extension, clearly we have $P[x_1^n]>0$, yet the right-hand side in part 3) of Theorem 3 is 0.

Despite this, it is noted in [7] that for such a type $U_{X_0^k}$ we may find another type $U'_{X_0^k}$ with no forbidden transitions and for which $U'_{X_0^k}$ is sufficiently "close" to $U_{X_0^k}$ that the probability of a sequence of type $U'_{X_0^k}$ is "close" to the probability of x_1^n. In particular, for any irreducible Markov chain with $K=|\mathcal{X}|^k$ states, there is a sequence of at most K allowable transitions between any two states. Hence, by replacing the (at most) last K transitions of x_1^n appropriately and possibly shortening the sequence by up to K terms, we may eliminate the forbidden transitions entirely. Furthermore, since we only replaced (and removed) up to a fixed number K of terms in the sequence, we can upper-bound the ratio between the probability of the sequence x_1^n to that of a sequence of type $U'_{X_0^k}$. We may thus complement the upper bound in Theorem 3 with the following.

Theorem 4: Let x_1^n be a sequence of type $U_{X_0^k}$ and $P[x_1^n]$ a PMF on $\mathcal{X}^n$ defined by

$$P[x_1^n]=\mu_{X_1^k}\left[x_1^k\right]\prod_{m=k+1}^n\Gamma_{X_k\mid X_0^{k-1}}\left[x_m\;\middle|\;x_{m-k}^{m-1}\right] \quad (7)$$

with $\mu_{X_1^k}>0$ and $\Gamma_{X_k\mid X_0^{k-1}}$ irreducible. Let $K=|\mathcal{X}|^k$. Then there exists a type

$$U'_{X_0^k}\in\mathcal{U}^n_{X_0^k}\bigcup\cdots\bigcup\mathcal{U}^{n-K}_{X_0^k}$$

which depends only on $U_{X_0^k}$ and the forbidden transitions of $\Gamma_{X_k \mid X_0^{k-1}}$ such that

1) $|T(U_{X_0^k})| \leq 2^{n\rho_n} 2^{nH(X_k \mid X_0^{k-1})|_{U'}}$;
2) $P[x_1^n] \leq 2^{n\sigma_n} 2^{-n[D(U'_{X_0^k} \| \Gamma_{X_k \mid X_0^{k-1}}) + H(X_k \mid X_0^{k-1})|_{U'}]}$;

with $\rho_n \to 0$ and $\sigma_n \to 0$ as $n \to \infty$ and these do not depend on $U_{X_0^k}$. □

Proof:

1) Follows from the fact that we have only changed (at most) K terms in the sequence x_1^n. Hence, by the triangle inequality

$$\|U_{X_0^k} - U'_{X_0^k}\| \leq 2(K+k)/n + K/n$$

and the result follows by the uniform continuity of $H(X_k \mid X_0^{k-1})$ on $\mathcal{F}_{X_0^k}$.

2) If the sequence x_1^n has a forbidden transition that is not due to the cyclic extension of x_1^n, then the statement is clearly true for any $U'_{X_0^k}$ since $P[x_1^n] = 0$. If $U_{X_0^k} \ll \Gamma_{X_k \mid X_0^{k-1}}$ then we may take $U'_{X_0^k} = U_{X_0^k}$ and by Theorem 3, the bound holds.

If the only forbidden transitions are due to the cyclic extension of x_1^n, then by replacing up to the last K transitions of the cyclic sequence x_1^n, to form a new cyclic sequence $\hat{x}_1^{\hat{n}}$ with no forbidden transitions, we have $P[x_1^n] \leq \alpha^{-1}\beta^{-(K+k)} P[\hat{x}_1^{\hat{n}}]$ (α is the smallest term in $\mu_{X_1^k}$ and β is the smallest nonzero transition probability of $\Gamma_{X_k \mid X_0^{k-1}}$). The type $U'_{X_0^k}$ of the sequence $\hat{x}_1^{\hat{n}}$ thus derived depends only on the sequence x_1^n and the forbidden transitions of $\Gamma_{X_k \mid X_0^{k-1}}$, not just its type $U_{X_0^k}$. However, due to the choice of α and β, the bound derived for a particular x_1^n holds for all $x_1^n \in T(U_{X_0^k})$. Finally, since $U'_{X_0^k} \ll \Gamma_{X_k \mid X_0^{k-1}}$, the fact that $\hat{n}$ may be up to K terms smaller than n can be universally absorbed by the constants σ_n since the exponent in the upperbound of 2) is continuous over such $U'_{X_0^k}$ and the set of such $U'_{X_0^k}$ is compact. ■

Finally, we have the following result, which is a direct extension of Natarajan's result for first-order Markov types [10].

Lemma 1 (Natarajan): Given any $F_{X_0^k} \in \mathcal{F}_{X_0^k}$, there exists a sequence of kth-order Markov types $U^n_{X_0^k} \in \mathcal{U}^n_{X_0^k}$ such that $U^n_{X_0^k} \to F_{X_0^k}$ as $n \to \infty$. □

Proof: It suffices to prove the result for $F_{X_0^k} \in \mathbf{i}(\mathcal{F}_{X_0^k})$, the interior of $\mathcal{F}_{X_0^k}$. Since the latter is associated with an ergodic Markov chain $F_{X_k \mid X_0^{k-1}}$, by the law of large numbers, a sequence of types $U^n_{X_0^k} \in \mathcal{U}^n_{X_0^k}$ convergent to $F_{X_0^k}$ must exist. ■

B. Error Exponents and Definitions

If a sequence x_1^n is known to generate y_1^n, then the maximum *a posteriori* (MAP) probability detector (which maximizes the probability of correct detection) chooses the hypothesis

$$\hat{h}(x_1^n, y_1^n) \triangleq \arg\max_h P[h \mid x_1^n, y_1^n].$$

This may be evaluated using Bayes' rule. The probability of correct detection is then $P[H = \hat{h}(x_1^n, y_1^n) \mid x_1^n, y_1^n]$.

If the input sequence x_1^n is not randomized (i.e., it is fixed in advance), then, it is clear that the average probability of correct detection for that sequence using Bayes' rule is

$$P_c^n(x_1^n) \triangleq E_{Y_1^n} P\left[H = \hat{h}(x_1^n, Y_1^n) \,\middle|\, x_1^n, Y_1^n\right].$$

Furthermore, by the memoryless channel assumption, $P_c^n(x_1^n)$ depends only on the *type* of the sequence x_1^n. If we randomize the sequence X_1^n, then we average the performance over several types. Clearly, by the memoryless and time-invariant assumptions, it is best to choose the input sequence x_1^n from the best sequence type.

We are now in a position to define our error exponents. We have the following definitions.

Definition 1: Let $Q_n \in \mathcal{Q}_X^n$, $n = 1, \ldots, \infty$, be a sequence of types for X with $Q_n \to Q \in \mathcal{E}_X$ as $n \to \infty$.

1) The *type error exponent* of Q is defined to be

$$E_t(Q) = \lim_{n\to\infty} -\frac{1}{n} \log(1 - P_c^n(x_1^n))|_{x_1^n \in T(Q_n)}.$$

2) The *optimal type error exponent* is

$$E_t = \max_{Q \in \mathcal{E}_X} E_t(Q). \qquad \square$$

We note that the type error exponent is defined for any distribution $Q \in \mathcal{E}_X$. This should not be interpreted as implying that the input x_1^n is chosen randomly according to Q. As the definition says, we choose sequences x_1^n such that as $n \to \infty$, the type of x_1^n converges to Q. Thus, for all but a special set of $Q \in \mathcal{E}_X$, this sequence must be generated by a time-varying (and open loop) signal selection rule.

In the feedback case, each successive input x_t is allowed to depend on all previous inputs x_1^{t-1} and outputs y_1^{t-1} up to that time. In general, such a policy is given by the time-varying conditional probabilities $\{Q_t[X_t \mid X_1^{t-1}, Y_1^{t-1}]\}$. Employing such a policy, if a decision is made at time n, the probability of correct detection is then

$$P_{c,f}^n \triangleq E_{X_1^n, Y_1^n} P\left[H = \hat{h}(X_1^n, Y_1^n) \,\middle|\, X_1^n, Y_1^n\right] \tag{8}$$

where each successive x_t is chosen according to $Q_t[X_t \mid X_1^{t-1}, Y_1^{t-1}]$.

Definition 2: The *feedback error exponent* of a policy $\{Q_t\}$ is defined to be

$$E_f(\{Q_t\}) = \lim_{n\to\infty} -\frac{1}{n} \log\left(1 - P_{c,f}^n\right) \tag{9}$$

should the limit exist. The policies for which the limit exists will be called *error exponent stable*. □

We note that it is rather easy to construct a policy which is not error exponent stable by cycling between "good" and "bad" policies.

Finally, inSections III and IV, we will find it convenient to generalize the usual max and min operators.

Definition 3 (minN, maxN): Consider an M-tuple of real numbers $(a_1, \ldots, a_M)$ which may be (not necessarily uniquely) ordered as $a_{i_1} \leq a_{i_2} \leq \cdots \leq a_{i_M}$. Let N be an integer between 1 and M. Then, we define

$$\min\mathrm{N}\{a_1, \ldots, a_M\} \triangleq a_{i_N} \tag{10}$$

$$\max\mathrm{N}\{a_1, \ldots, a_M\} \triangleq a_{i_{M-N+1}}. \tag{11}$$

□

We note that if the numbers $a_1, \ldots, a_M$ are all distinct, then the $\min\mathrm{N}$ and $\max\mathrm{N}$ operators may be interpreted as evaluating the Nth smallest and Nth largest number, respectively.

III. The Type Error Exponent

In this section, we evaluate the type error exponent and we also provide a numerical method to determine these exponents (and hence the optimal exponent).

A. The Type Error Exponent $E_t(Q)$

Theorem 5: For any $Q \in \mathcal{E}_X$, the type error exponent is given by

$$E_t(Q) = \min_{\substack{h,h' \in \mathcal{H} \\ h \neq h'}} \min_{Q_{Y\mid X} \in \mathcal{E}_{Y\mid X}} \max\Big\{D\left(Q_{Y\mid X} \,\middle\|\, P^h_{Y\mid X} \,\middle|\, Q\right), D\left(Q_{Y\mid X} \,\middle\|\, P^{h'}_{Y\mid X} \,\middle|\, Q\right)\Big\}$$

where $P^h_{Y\mid X}[y\mid x]$ is the channel under hypothesis $h\in\mathcal{H}$. Furthermore, the minimizing $Q_{Y\mid X}$ results in equality of the two terms inside the max. □

Proof: The proof is divided into two parts. In the first part, we show that an alternate expression for $E_t(Q)$ is given by

$$E_t(Q)=\min_{Q_{Y\mid X}\in\mathcal{E}_{Y\mid X}}\min\nolimits_{h\in\mathcal{H}}2\left\{D\left(Q_{Y\mid X}\,\middle\|\,P^h_{Y\mid X}\,\middle|\,Q\right)\right\}$$

where the $\min 2$ operator was defined in Section II-B. In the second part, we extend this expression to the desired result. We also argue that the minimizing $Q_{Y\mid X}$ results in equality of the two terms inside the max.

Part 1: We first expand the expression for P_c^n using Bayes' rule. We assume that for all $n, x_1^n\in T(Q_n)$

$$P_c^n\triangleq E_{Y_1^n}\left[\max_{h\in\mathcal{H}}P\left[H=h\mid x_1^n,Y_1^n\right]\right] \tag{12}$$

$$=\sum_{y_1^n}\max_h\left\{P^h_{Y\mid X}\left[y_1^n\mid x_1^n\right]P[H=h]\right\}. \tag{13}$$

Hence, using Part 1 of Theorem 2 and the fact that

$$1=\sum_{h\in\mathcal{H}}\sum_{y_1^n}P^h_{Y\mid X}\left[y_1^n\mid x_1^n\right]P[H=h] \tag{14}$$

$$=\sum_{N=1}^{|\mathcal{H}|}\sum_{y_1^n}\operatorname*{maxN}_h\left\{P^h_{Y\mid X}\left[y_1^n\mid x_1^n\right]P[H=h]\right\} \tag{15}$$

we obtain

$$1-P_c^n=\sum_{N=2}^{|\mathcal{H}|}\sum_{y_1^n}\operatorname*{maxN}_h\left\{P^h_{Y\mid X}\left[y_1^n\mid x_1^n\right]P[H=h]\right\} \tag{16}$$

$$=\sum_{N=2}^{|\mathcal{H}|}\left[\sum_{\substack{Q_{XY}\in\mathcal{Q}^n_{XY}\\ :\,Q_X=Q_n}}|T(Q_{Y\mid X})|\right.$$

$$\left.\times\ \operatorname*{maxN}_h\left\{P[H=h]2^{-n\left[D(Q_{XY}\,\|\,P^h_{Y\mid X})+H(Q_{Y\mid X})\right]}\right\}\right]. \tag{17}$$

We now proceed to lowerbound the exponent. By applying part 2 of Theorem 2 we get

$$1-P_c^n\le\sum_{\substack{Q_{XY}\in\mathcal{Q}^n_{XY}\\ :\,Q_X=Q_n}}|\mathcal{H}||T(Q_{Y\mid X})|$$

$$\times\max_h 2\left\{2^{-n\left[D(Q_{XY}\,\|\,P^h_{Y\mid X})+H(Q_{Y\mid X})\right]}\right\} \tag{18}$$

$$\le\sum_{\substack{Q_{XY}\in\mathcal{Q}^n_{XY}\\ :\,Q_X=Q_n}}|\mathcal{H}|\max_h 2\left\{2^{-nD(Q_{XY}\,\|\,P^h_{Y\mid X})}\right\} \tag{19}$$

$$\le|\mathcal{H}|\,|\mathcal{Q}^n_{XY}|\max_{\substack{Q_{XY}\in\mathcal{E}_{XY}\\ :\,Q_X=Q_n}}\max_h 2\left\{2^{-nD(Q_{XY}\,\|\,P^h_{Y\mid X})}\right\}. \tag{20}$$

Hence,

$$-\frac{1}{n}\log(1-P_c^n)\ge-\frac{1}{n}\log(|\mathcal{H}|\,|\mathcal{Q}^n_{XY}|)$$

$$+\min_{\substack{Q_{XY}\in\mathcal{E}_{XY}\\ :\,Q_X=Q_n}}\min\nolimits_h 2\left\{D\left(Q_{XY}\,\middle\|\,P^h_{Y\mid X}\right)\right\}. \tag{21}$$

By Part 1 of Theorem 1, $|\mathcal{H}||\mathcal{Q}^n_{XY}|$ is polynomial in n and the first term in (21) vanishes as $n\to\infty$. Hence, from (21)

$$E_t(Q)\ge\lim_{n\to\infty}\min_{Q_{Y\mid X}\in\mathcal{E}_{Y\mid X}}\min\nolimits_h 2\left\{D\left(Q_{Y\mid X}\,\middle\|\,P^h_{Y\mid X}\,\middle|\,Q_n\right)\right\}$$

$$=\min_{Q_{Y\mid X}\in\mathcal{E}_{Y\mid X}}\min\nolimits_h 2\left\{D\left(Q_{Y\mid X}\,\middle\|\,P^h_{Y\mid X}\,\middle|\,Q\right)\right\} \tag{22}$$

where we have used Lemma 2 (see Appendix) with $\mathcal{V}=\mathcal{E}_{Y\mid X}, \mathcal{U}=\mathcal{E}_X, u_n=Q_n, u=Q, v=Q_{Y\mid X}$, and

$$f(u,v)=\max 2_h\left\{D(Q_{Y\mid X}\,\|\,P^h_{Y\mid X}\,|\,Q)\right\}$$

to obtain (22).

We now upper-bound the exponent. Starting from (17), we observe that for any sequence of joint types $Q^n_{XY}\in\mathcal{Q}^n_{XY}$ whose marginalization satisfies $Q^n_X=Q_n$, and using Part 2 of Theorem 2

$$1-P_c^n$$

$$\ge\left|T\left(Q^n_{Y\mid X}\right)\right|\max_h 2\left\{P[H=h]2^{-n\left[D(Q^n_{XY}\,\|\,P^h_{Y\mid X})+H(Q^n_{Y\mid X})\right]}\right\}$$

$$\ge|\mathcal{Q}^n_X|^{-1}\,|\mathcal{Q}^n_Y|^{-1}\max_h 2\left\{P[H=h]2^{-n\,D(Q^n_{XY}\,\|\,P^h_{Y\mid X})}\right\}$$

Hence,

$$-\frac{1}{n}\log(1-P_c^n)\le\frac{1}{n}\log|\mathcal{Q}^n_X|\,|\mathcal{Q}^n_Y|$$

$$+\min_h 2\left\{-\frac{\log P[H=h]}{n}+D\left(Q^n_{XY}\,\middle\|\,P^h_{Y\mid X}\right)\right\}. \tag{23}$$

Furthermore, by Lemma 3 (see the Appendix), for any $Q_{Y\mid X}\in\mathcal{E}_{Y\mid X}$ and a sequence of types $Q_n\in\mathcal{Q}^n_X$ such that $Q_n\to Q\in\mathcal{E}_X$, we have a sequence of joint types Q^n_{XY} in $\mathcal{Q}^n_{XY}$ such that $Q^n_{XY}\to Q_{Y\mid X}Q_X$ and the X marginal of Q^n_{XY} is Q_n. Evaluating (23) for such a sequence, the two terms with $\frac{1}{n}$ vanish and since $D(Q^n_{XY}\,\|\,P^h_{Y\mid X})$ is continuous in Q^n_{XY} (since $P^h_{Y\mid X}[y\mid x]>0$), we may bring the limit $n\to\infty$ into the argument of the divergence, from which we conclude that for all $Q_{Y\mid X}$

$$E_t(Q)\le\min_h 2\left\{D\left(Q_{Y\mid X}\,\middle\|\,P^h_{Y\mid X}\,\middle|\,Q\right)\right\}.$$

Part 2: We observe the following chain of equalities:

$$E_t(Q)=\min_{Q_{Y\mid X}\in\mathcal{E}_{Y\mid X}}\min\nolimits_{h\in\mathcal{H}}2\left\{D\left(Q_{Y\mid X}\,\middle\|\,P^h_{Y\mid X}\,\middle|\,Q\right)\right\} \tag{24}$$

$$=\min_{Q_{Y\mid X}\in\mathcal{E}_{Y\mid X}}\min_{\substack{h,h'\in\mathcal{H}\\ h\neq h'}}\max\left\{D\left(Q_{Y\mid X}\,\middle\|\,P^h_{Y\mid X}\,\middle|\,Q\right),\right.$$

$$\left.D\left(Q_{Y\mid X}\,\middle\|\,P^{h'}_{Y\mid X}\,\middle|\,Q\right)\right\} \tag{25}$$

$$=\min_{\substack{h,h'\in\mathcal{H}\\ h\neq h'}}\min_{Q_{Y\mid X}\in\mathcal{E}_{Y\mid X}}\max\left\{D\left(Q_{Y\mid X}\,\middle\|\,P^h_{Y\mid X}\,\middle|\,Q\right),\right.$$

$$\left.D\left(Q_{Y\mid X}\,\middle\|\,P^{h'}_{Y\mid X}\,\middle|\,Q\right)\right\}. \tag{26}$$

Finally, by Lemma 4 (see the Appendix), we have that the minimizing $Q_{Y\mid X}$ results in equality of the two terms inside the max. ■

Corollary 1: In the case of a binary hypotheses scenario, the optimal error exponent $E_t=\max_{Q_X\in\mathcal{E}_X}E_t(Q_X)$ is achieved by a vertex of the simplex $\mathcal{E}_X$ (i.e., the optimal input sequence is constant). □

Proof: In the case of detection with only two hypotheses with $\mathcal{H}=\{a,b\}$

$$E_t(Q_X)=\min_{Q_{Y\mid X}\in\mathcal{E}_{Y\mid X}}\max\left\{D\left(Q_{Y\mid X}\,\middle\|\,P^a_{Y\mid X}\,\middle|\,Q_X\right),\right.$$

$$\left.D\left(Q_{Y\mid X}\,\middle\|\,P^b_{Y\mid X}\,\middle|\,Q_X\right)\right\}. \tag{27}$$

Let V_i be the vertices of the simplex $\mathcal{E}_X$. Then, based on (27) and the fact that the minimizing $Q_{Y\,|\,X}$ results in equality of the two terms inside the max, we have for each $x^{(i)}$ a minimizing $Q^*_{Y\,|\,x^{(i)}}$.

For any convex combination $Q_X = \sum_i \lambda_i V_i$, we also have the following chain of inequalities,

$$\sum_i \lambda_i E_t(V_i) = D(Q^*_{Y\,|\,X} \,\|\, P^a_{Y\,|\,X} \,|\, Q_X) = D(Q^*_{Y\,|\,X} \,\|\, P^b_{Y\,|\,X} \,|\, Q_X) \geq E_t(Q_X)$$

where the last inequality is because the minimizing $Q_{Y\,|\,X}$ in $E_t(Q_X)$ also results in equality of the two terms inside the max of (27). Hence, by the relation $\sum_i \lambda_i E_t(V_i) \geq E_t(Q_X)$, it follows that the optimal exponent is achieved by a vertex. ∎

B. Evaluating $E_t(Q)$

In this subsection, we propose a method to efficiently evaluate $E_t(Q)$ numerically. The method is based on Lagrange multipliers, a mathematical tool which is well known to yield the Bayesian error exponent in traditional hypothesis testing [4]. The maximizing Q may then be found using standard numerical techniques. Given two distinct hypotheses $h_1, h_2 \in \mathcal{H}$, it will suffice to be able to efficiently evaluate D_{h_1,h_2}, defined as follows:

$$D_{h_1,h_2}(Q_X) \triangleq \min_{Q_{Y\,|\,X} \in \mathcal{E}_{Y\,|\,X}} \max\left\{ D\left(Q_{Y\,|\,X} \,\middle\|\, P^{h_1}_{Y\,|\,X} \,\middle|\, Q_X\right), D\left(Q_{Y\,|\,X} \,\middle\|\, P^{h_2}_{Y\,|\,X} \,\middle|\, Q_X\right)\right\}. \quad (28)$$

Theorem 6: If $D(P^{h_2}_{Y\,|\,X} \,\|\, P^{h_1}_{Y\,|\,X} \,|\, Q_X) = 0$, then $Q_{Y\,|\,X} = P^{h_2}_{Y\,|\,X}$ is a global minimizer of (28). Otherwise, the global minimizing $Q_{Y\,|\,X}$ in (28) has the form

$$Q^{(\lambda)}_{Y\,|\,X}[y\,|\,x] = \frac{P^{h_1}_{Y\,|\,X}[y\,|\,x]^\lambda P^{h_2}_{Y\,|\,X}[y\,|\,x]^{1-\lambda}}{\sum_{y'} P^{h_1}_{Y\,|\,X}[y'\,|\,x]^\lambda P^{h_2}_{Y\,|\,X}[y'\,|\,x]^{1-\lambda}} \quad (29)$$

for some $\lambda = \lambda^*$. Furthermore, this λ^* is the unique zero of

$$M(\lambda) = D(Q^{(\lambda)}_{Y\,|\,X} \,\|\, P^{h_1}_{Y\,|\,X} \,|\, Q_X) - D(Q^{(\lambda)}_{Y\,|\,X} \,\|\, P^{h_2}_{Y\,|\,X} \,|\, Q_X)$$

in the interval $[0.0, 1.0]$. Hence, λ^* may be computed efficiently using the bisection method. □

Proof: By Theorem 5, we know that the minimizing $Q_{Y\,|\,X}$ results in equality between

$$D(Q_{Y\,|\,X} \,\|\, P^{h_1}_{Y\,|\,X} \,|\, Q_X) \quad \text{and} \quad D(Q_{Y\,|\,X} \,\|\, P^{h_2}_{Y\,|\,X} \,|\, Q_X).$$

This suggests the use of the Lagrangian (where λ and γ_x are the multipliers)

$$J(Q_{Y\,|\,X}) = D\left(Q_{Y\,|\,X} \,\middle\|\, P^{h_2}_{Y\,|\,X} \,\middle|\, Q_X\right) + \lambda\left[D\left(Q_{Y\,|\,X} \,\middle\|\, P^{h_2}_{Y\,|\,X} \,\middle|\, Q_X\right) - D\left(Q_{Y\,|\,X} \,\middle\|\, P^{h_1}_{Y\,|\,X} \,\middle|\, Q_X\right)\right] + \sum_x \gamma_x \sum_y (Q_{Y\,|\,X}(y\,|\,x) - 1). \quad (30)$$

It is now straightforward to verify that the minimizing $Q_{Y\,|\,X}$ has the form specified in (29). Furthermore, the difference $M(\lambda)$ may be expanded as

$$M(\lambda) = \sum_x Q_X[x] \times \sum_y \frac{P^{h_1}_{Y\,|\,X}[y\,|\,x]^\lambda P^{h_2}_{Y\,|\,X}[y\,|\,x]^{1-\lambda}}{\sum_{y'} P^{h_1}_{Y\,|\,X}[y'\,|\,x]^\lambda P^{h_2}_{Y\,|\,X}[y'\,|\,x]^{1-\lambda}} \log \frac{P^{h_2}_{Y\,|\,X}[y\,|\,x]}{P^{h_1}_{Y\,|\,X}[y\,|\,x]}. \quad (31)$$

Then $M(0) = D(P^{h_2}_{Y\,|\,X} \,\|\, P^{h_1}_{Y\,|\,X} \,|\, Q_X) \geq 0$. If equality holds, then it is easy to see that $Q_{Y\,|\,X} = P^{h_2}_{Y\,|\,X}$ is a global minimizer of the right-hand side of (28).

Now, assume that $M(0) > 0$. Then

$$M(1) = -D(P^{h_1}_{Y\,|\,X} \,\|\, P^{h_2}_{Y\,|\,X} \,|\, Q_X) < 0$$

and by continuity, there is a λ^* in $(0.0, 1.0)$ such that $M(\lambda^*) = 0$. Furthermore, with the substitutions

$$a(y\,|\,x) \triangleq \frac{P^{h_1}_{Y\,|\,X}[y\,|\,x]}{P^{h_2}_{Y\,|\,X}[y\,|\,x]} \quad (32)$$

$$c(y\,|\,x) \triangleq P^{h_2}_{Y\,|\,X}[y\,|\,x] \quad (33)$$

we may rewrite the difference $M(\lambda)$ as

$$M(\lambda) = \sum_x Q_X[x] \sum_y \frac{-c(y\,|\,x)a(y\,|\,x)^\lambda}{\sum_{y'} c(y'\,|\,x)a(y'\,|\,x)^\lambda} \log a(y\,|\,x). \quad (34)$$

The derivative of the inner summation with respect to λ is then

$$\frac{-\left[\sum_y c(y\,|\,x)a(y\,|\,x)^\lambda (\log a(y\,|\,x))^2\right]\left[\sum_{y'} c(y'\,|\,x)a(y'\,|\,x)^\lambda\right]}{\left(\sum_{y'} c(y'\,|\,x)a(y'\,|\,x)^\lambda\right)^2} + \frac{\left[\sum_y c(y\,|\,x)a(y\,|\,x)^\lambda \log a(y\,|\,x)\right]^2}{\left(\sum_{y'} c(y'\,|\,x)a(y'\,|\,x)^\lambda\right)^2}. \quad (35)$$

We note that the denominator is always strictly greater than 0, hence, the sign of the overall expression equals to the sign of the numerator

$$-\left[\sum_y c(y\,|\,x)a(y\,|\,x)^\lambda (\log a(y\,|\,x))^2\right]\left[\sum_y c(y\,|\,x)a(y\,|\,x)^\lambda\right] + \left[\sum_y c(y\,|\,x)a(y\,|\,x)^\lambda \log a(y\,|\,x)\right]^2. \quad (36)$$

However, by the Cauchy–Schwartz inequality, the latter is always ≤ 0. Equality holds, provided there is a constant $k(x)$ (dependent on x) such that

$$c(y\,|\,x)a(y\,|\,x)^\lambda = k(x)c(y\,|\,x)a(y\,|\,x)^\lambda (\log a(y\,|\,x))^2 \quad (37)$$

for all y. However, if equality (37) were true for some λ and all x such that $Q_X[x] > 0$, then it would be true for all λ and such x. It follows that $M(\lambda)$ would be constant in λ, contradicting $M(0) > 0 > M(1)$. Hence, $dM(\lambda)/d\lambda < 0$ and λ^* is unique and the global minimizing $Q_{Y\,|\,X} = Q^{(\lambda^*)}_{Y\,|\,X}$. ∎

IV. Bayesian Hypothesis Testing With Markov Feedback

In this section, we derive the exact exponent for a particular class of feedback policies. In particular, for any integer $k \geq 0$, we shall consider Markov policies, i.e., policies $\{Q_t\}$ for which

$$Q_{X_t\,|\,X_1^{t-1},Y_1^{t-1}}\left[x_t \,\middle|\, x_1^{t-1}, y_1^{t-1}\right] > 0 \quad (38)$$

when $t \leq k$ and

$$Q_{X_t\,|\,X_1^{t-1},Y_1^{t-1}}\left[x_t \,\middle|\, x_1^{t-1}, y_1^{t-1}\right] = Q_{X_k\,|\,X_0^{k-1},Y_0^{k-1}}\left[x_t \,\middle|\, x_{t-k}^{t-1}, y_{t-k}^{t-1}\right] \quad (39)$$

otherwise. We note that this includes deterministic policies of the form $x_t = \xi(x_{t-k}^{t-1}, y_{t-k}^{t-1})$ as a special subset. Hence, by studying the class of k-memory Markov policies, we are studying all deterministic

input policies which employ an arbitrarily large (but finite) amount of memory.

For notational purposes, we *abbreviate* the feedback exponent $E_f(Q_{X_k \mid X_0^{k-1}, Y_0^{k-1}})$ of such policy by $E_f(Q)$. For the moment, we further restrict ourselves to policies $Q_{X_t \mid X_{t-k}^{t-1}, Y_{t-k}^{t-1}}$ for which

$$W^h_{(X,Y)_k \mid (X,Y)_0^{k-1}} \triangleq P^h_{Y_k \mid X_k} Q_{X_k \mid X_0^{k-1}, Y_0^{k-1}}$$

is irreducible (policies for which this is not the case are discussed in the second remark following Theorem 8). We note that since for all h we have that $P^h_{Y_k \mid X_k} > 0$ it follows that if $W^h_{(X,Y)_k \mid (X,Y)_0^{k-1}}$ is irreducible for a particular h, then it is for every $h \in \mathcal{H}$. Furthermore, all W^h have the same state transitions; they differ only in the probabilities of these transitions. Because of this, we label such a policy $Q_{X_k \mid X_0^{k-1}, Y_0^{k-1}}$ as irreducible even though strictly speaking it is the $W^h_{(X,Y)_k \mid (X,Y)_0^{k-1}}$ that are irreducible.

Although each successive input is based on the previous k inputs and outputs only, the MAP decision is based on the entire realizations of x_1^n and y_1^n. We now state and prove the following theorem.

Theorem 7: For a given k-memory time-invariant irreducible policy $Q_{X_k \mid X_0^{k-1}, Y_0^{k-1}}$, and channels $P^h_{Y_k \mid X_k} > 0$, define

$$E_f^*(Q) \triangleq \min_{Q_{(X,Y)_0^k} \in \mathcal{F}_{(X,Y)_0^k}} \min_h 2 \left\{ D\left(Q_{(X,Y)_0^k} \,\middle\|\, W^h_{(X,Y)_k \mid (X,Y)_0^{k-1}} \right) \right\}. \tag{40}$$

Then, the error exponent $E_f(Q) = E_f^*(Q)$. Furthermore, the minimizing $Q_{(X,Y)_0^k}$ results in equality between

$$\min_h \{ D(Q_{(X,Y)_0^k} \,\|\, W^h_{(X,Y)_k \mid (X,Y)_0^{k-1}}) \}$$

and

$$\min_h 2 \{ D(Q_{(X,Y)_0^k} \,\|\, W^h_{(X,Y)_k \mid (X,Y)_0^{k-1}}) \} \qquad \square$$

Proof: We note that by Lemma 5 (see the Appendix), the minimizing $Q_{(X,Y)_0^k}$ must result in equality of

$$\min_h \{ D(Q_{(X,Y)_0^k} \,\|\, W^h_{(X,Y)_k \mid (X,Y)_0^{k-1}}) \}$$

and

$$\min 2_h \{ D(Q_{(X,Y)_0^k} \,\|\, W^h_{(X,Y)_k \mid (X,Y)_0^{k-1}}) \}$$

since otherwise one could find a $Q_{(X,Y)_0^k}$ that did better (in the minimization of (40)).

If we let $Z_i = (X_i, Y_i)$, then Z_i is a Markov chain of order k.

Case 1: We will first consider the case $Q_{X_k \mid X_0^{k-1}, Y_0^{k-1}}$ is irreducible and show that $E_f(Q) \geq E_f^*(Q)$. Expanding the expression for the probability of correct detection yields

$$P_c^n = \sum_{(x_1^n, y_1^n)} \sum_h P^h [x_1^n, y_1^n] P[h] \max_{h'} \left\{ \frac{P^{h'} [x_1^n, y_1^n] P[h']}{\sum_{h''} P^{h''} [x_1^n, y_1^n] P[h'']} \right\} \tag{41}$$

$$= \sum_{z_1^n \in \mathcal{Z}^n} \max_h \left\{ W^h [z_1^n] P[h] \right\} \tag{42}$$

where the summation in (41) is over all sequences $z_1^n = (x_1^n, y_1^n)$ with strictly positive probability under W^h. In (42), we may sum over all sequences z_1^n since those with probability 0 contribute nothing. Employing Theorem 4, the probability of error may be upper-bounded as

$$1 - P_c^n \leq |\mathcal{H}| \sum_{z_1^n} \max 2 \left\{ W^h [z_1^n] P[h] \right\} \tag{43}$$

$$\leq |\mathcal{H}| 2^{n\sigma_n} \sum_{U_{Z_0^k} \in \mathcal{U}^n_{Z_0^k}} \left| T\left(U_{Z_0^k} \right) \right| \times \max_h 2 \left\{ 2^{-n [D(U'_{Z_0^k} \,\|\, W^h_{Z_k \mid Z_0^{k-1}}) + H(Z_k \mid Z_0^{k-1})_{\mid U'}]} \right\} \tag{44}$$

where we have employed Part 2 of Theorem 4 and $U'_{Z_0^k}$ is an implicit function of $U_{Z_0^k}$ as discussed in Theorem 4. Continuing and using Part 1 of Theorem 4

$$1 - P_c^n \leq |\mathcal{H}| 2^{n(\sigma_n + \rho_n)} \sum_{U_{Z_0^k} \in \mathcal{U}^n_{Z_0^k}} \times \max_h 2 \left\{ 2^{-n D(U'_{Z_0^k} \,\|\, W^h_{Z_k \mid Z_0^{k-1}})} \right\} \tag{45}$$

$$\leq |\mathcal{H}| 2^{n(\sigma_n + \rho_n)} |\mathcal{U}^n_{Z_0^k}| \times \max_{U_{Z_0^k} \in \mathcal{F}_{Z_0^k}} \max_h 2 \left\{ 2^{-n D(U_{Z_0^k} \,\|\, W^h_{Z_k \mid Z_0^{k-1}})} \right\} \tag{46}$$

where the last inequality follows for the fact that $U'_{Z_0^k} \in \mathcal{F}_{Z_0^k}$. Furthermore, we are justified in writting max as opposed to sup since on the subset A of $\mathcal{F}_{Z_0^k}$ defined by $U_{Z_0^k} \ll W^h_{Z_k \mid Z_0^{k-1}}$, the function $D(U_{Z_0^k} \,\|\, W^h_{Z_k \mid Z_0^{k-1}})$ is continuous and the subset A is compact while on $\mathcal{F}_{Z_0^k} \backslash A$, we have $D(U_{Z_0^k} \,\|\, W^h_{Z_k \mid Z_0^{k-1}}) = \infty$. Hence, since $\sigma_n \to 0$ and $\rho_n \to 0$

$$E_f(Q) \geq \min_{U_{Z_0^k} \in \mathcal{F}_{Z_0^k}} \min_h 2 \left\{ D\left(U_{Z_0^k} \,\middle\|\, W^h_{Z_k \mid Z_0^{k-1}} \right) \right\}. \tag{47}$$

Finally, from (47), this shows that (40) is a lower bound to $E_f(Q)$.

Case 2: We consider $Q_{X_k \mid X_0^{k-1}, Y_0^{k-1}} > 0$ and show that $E_f(Q) \leq E_f^*(Q)$. From (42), we may lower-bound the probability of error as

$$P_e^n \triangleq 1 - P_c^n \geq \sum_{z_1^n} \max_h 2 \left\{ W^h [z_1^n] P[h] \right\} \tag{48}$$

$$\geq \alpha \sum_{U_{Z_0^k} \in \mathcal{U}^n_{Z_0^k}} \left| T\left(U_{Z_0^k} \right) \right| \times \max 2 \left\{ 2^{-n [D(U_{Z_0^k} \,\|\, W^h_{Z_k \mid Z_0^{k-1}}) + H(Z_k \mid Z_0^{k-1})]} \right\} \tag{49}$$

$$\geq \alpha n^{-|\mathcal{Z}|^k} (n+1)^{-|\mathcal{Z}|^{k+1}} \times \max_{U_{Z_0^k} \in \mathcal{U}^n_{Z_0^k}} \max 2 \left\{ 2^{-n D(U_{Z_0^k} \,\|\, W^h_{Z_k \mid Z_0^{k-1}})} \right\} \tag{50}$$

where (49) follows from Part 3 of Theorem 3 and (50) follows from Part 2 of Theorem 3. One then obtains

$$E_f(Q) \leq \lim_{n \to \infty} \min_{U_{Z_0^k} \in \mathcal{U}^n_{Z_0^k}} \min_h 2 \left\{ D\left(U_{Z_0^k} \,\middle\|\, W^h_{Z_k \mid Z_0^{k-1}} \right) \right\} \tag{51}$$

$$= \min_{U_{Z_0^k} \in \mathcal{F}_{Z_0^k}} \min_h 2 \left\{ D\left(U_{Z_0^k} \,\middle\|\, W^h_{Z_k \mid Z_0^{k-1}} \right) \right\} \tag{52}$$

where the last equality follows from Lemma 1 and the continuity of $D(U_{Z_0^k} \| W^h_{Z_k \mid Z_0^{k-1}})$ in $U_{Z_0^k}$. Again, (52) shows that (40) is an upper bound to $E_f(Q)$.

Comment: The proofs in case 1 and case 2 combine to show that $E_f(Q) = E_f^*(Q)$ if $Q_{X_k \mid X_0^{k-1}, Y_0^{k-1}} > 0$. Now we look at the case $Q_{X_k \mid X_0^{k-1}, Y_0^{k-1}} \not> 0$.

Case 3: In this case, $W^h_{Z_k \mid Z_0^{k-1}}[z_k \mid z_0^{k-1}] = 0$ for some z_0^k, but $W^h_{Z_k \mid Z_0^{k-1}}$ is assumed irreducible (and hence ergodic [2], [9]). Then the expression in (51) is still valid but insufficient to derive (52). We note that since we are performing MAP detection, $P_e^n = \max\{P_e^n, \ldots, P_e^{n+K}\}$, $K = |\mathcal{Z}|^k$. It then follows that (51) is still valid if we minimize over $U_{Z_0^k} \in \mathcal{U}^n_{Z_0^k} \cup \cdots \cup \mathcal{U}^{n+K}_{Z_0^k}$ instead. We will show that evaluation yields (52). This is not entirely trivial as $D(U_{Z_0^k} \| W^h_{Z_k \mid Z_0^{k-1}})$ is not continuous on $\mathcal{F}_{Z_0^k}$. In particular, since all $W^h_{Z_k \mid Z_0^{k-1}}[z_k \mid z_0^{k-1}]$ have the same allowable state transitions (i.e., transitions with strictly positive probability), evaluation of (52) for any $U_{Z_0^k}$ associated with a $U_{Z_k \mid Z_0^{k-1}}$ with exactly the same allowable state transitions as $W^h_{Z_k \mid Z_0^{k-1}}$ will result in a finite bound on $E_f(Q)$. Consider $U'_{Z_0^k}$ to be any such distribution.

Now, let $U^*_{Z_0^k}$ be a minimizer in (52). $U^*_{Z_0^k}$ is not necessarily associated with an irreducible $U^*_{Z_k \mid Z_0^{k-1}}$. However, by convexity of divergence, for $0 \le \lambda \le 1$, the distribution

$$U^\lambda_{Z_0^k} = (1-\lambda)U^*_{Z_0^k} + \lambda U'_{Z_0^k}$$

results in a continuous (in λ) valuation of the expression $D(U^\lambda_{Z_0^k} \| W^h_{Z_k \mid Z_0^{k-1}})$.

Now, by employing methods similar to Lemma 1 and appropriately changing (and removing) up to the last K terms of any generated sequence with forbidden transitions in its cyclic extension, for each $\lambda > 0$, we can find a sequence of circular types

$$U^n_{Z_0^k} \in \mathcal{U}^n_{Z_0^k} \cup \cdots \cup \mathcal{U}^{n+K}_{Z_0^k}$$

which converges to $U^\lambda_{Z_0^k}$ and for which (for all n greater than some N) both $U^n_{Z_0^k}$ and $U^\lambda_{Z_0^k}$ share exactly the same allowable transitions. Hence, for each $\lambda > 0$

$$E_f(Q) \le \min_h 2\{D(U^\lambda_{Z_0^k} \| W^h_{Z_k \mid Z_0^{k-1}}\}$$

and (52) follows by continuity in λ. ∎

Remark: If the channels were Markov of order at most k, Theorem 7 would still apply as $Z_i = (X_i, Y_i)$ is then still a Markov chain of order k.

Remark: Since every finite state time-invariant Markov chain can be decomposed into ergodic classes, by suitably redefining our notion of state, we may evaluate the error exponent associated with each class. It is then clear that an optimal policy Q can always be reduced to one class (the best class if there is more than one class) and all other states transit to the class in at most a finite number of transitions, all of which are deterministic. In this latter case, direct evaluation of $E_f^*(Q)$ still provides the exponent of the class and hence the policy. If a policy Q^* has several classes, then direct evaluation of $E_f^*(Q^*)$ provides a lower bound to the exponent of any of its classes. This fact will be employed to derive the structure on an optimal Markov policy in Theorem 8.

Proposition 1: For any $Q \in \mathcal{E}_X$ we have that

$$\min_{Q_{Y \mid X} \in \mathcal{E}_{Y \mid X}} \min_h 2\left\{D\left(Q_{Y \mid X} Q \,\middle\|\, P^h_{Y \mid X}\right)\right\} \ge \min_{Q_{XY} \in \mathcal{E}_{XY}} \min_h 2\left\{D\left(Q_{XY} \,\middle\|\, P^h_{Y \mid X} Q\right)\right\}$$

Proof: Obvious. ∎

Theorem 8: In the case of discriminating between two memoryless hypotheses, the optimal k-memory Markov policy has its error exponent upper-bounded by the optimal open-loop exponent of Corollary 1. Furthermore, the bound is tight and the optimal Markov policy of *any* order consists of repeating the same input, regardless of past outputs. □

Proof: First, we remark that

$$D\left(Q_{Z_0^k} \,\middle\|\, W^h_{Z_k \mid Z_0^{k-1}}\right) = \sum_{z_0^{k-1}} Q_{Z_0^{k-1}}\left[z_0^{k-1}\right] \times D\left(Q_{Z_k \mid Z_0^{k-1}}\left[Z_k \mid z_0^{k-1}\right] \,\middle\|\, W^h_{Z_k \mid Z_0^{k-1}}\left[Z_k \,\middle|\, z_0^{k-1}\right]\right) \tag{53}$$

$$\le \max_{z_0^{k-1}} D\left(Q_{Z_k \mid Z_0^{k-1}}\left[Z_k \,\middle|\, z_0^{k-1}\right] \,\middle\|\, W^h_{Z_k \mid Z_0^{k-1}}\left[Z_k \,\middle|\, z_0^{k-1}\right]\right). \tag{54}$$

For notational convenience, we shall denote by

$$\max_{z_0^{k-1}} D(Q_{Z_k \mid Z_0^{k-1}} \| W^h_{Z_k \mid Z_0^{k-1}})$$

the right-hand side of (54). From the remark following Theorem 7, it suffices to consider policies with a single ergodic class. Starting with the result of Theorem 7 and assuming $\mathcal{H} = \{h_1, h_2\}$, we can bound the exponent $E_f(Q)$ of such a Markov policy $Q_{X_k \mid X_0^{k-1}, Y_0^{k-1}}$ by

$$E_f(Q) \le \min_{Q_{Z_0^k} \in \mathcal{F}^n_{Z_0^k}} \max_{h \in \{h_1, h_2\}} \left\{D\left(Q_{Z_0^k} \,\middle\|\, W^h_{Z_k \mid Z_0^{k-1}}\right)\right\} \tag{55}$$

$$\le \min_{Q_{Z_0^k} \in \mathcal{F}^n_{Z_0^k}} \max_{h \in \{h_1, h_2\}} \left\{\max_{z_0^{k-1}} D\left(Q_{Z_k \mid Z_0^{k-1}} \,\middle\|\, W^h_{Z_k \mid Z_0^{k-1}}\right)\right\} \tag{56}$$

$$= \min_{Q_{Z_k \mid Z_0^{k-1}} \in \mathcal{E}_{Z_k \mid Z_0^{k-1}}} \max_{h \in \{h_1, h_2\}} \times \left\{\max_{z_0^{k-1}} D\left(Q_{Z_k \mid Z_0^{k-1}} \,\middle\|\, W^h_{Z_k \mid Z_0^{k-1}}\right)\right\} \tag{57}$$

$$= \min_{Q_{Z_k \mid Z_0^{k-1}} \in \mathcal{E}_{Z_k \mid Z_0^{k-1}}} \max_{z_0^{k-1}} \max \times \left\{D\left(Q_{Z_k \mid Z_0^{k-1}} \,\middle\|\, W^{h_1}_{Z_k \mid Z_0^{k-1}}\right), D\left(Q_{Z_k \mid Z_0^{k-1}} \,\middle\|\, W^{h_2}_{Z_k \mid Z_0^{k-1}}\right)\right\}. \tag{58}$$

We denote by $\hat{z}_0^{k-1}$ a maximizing z_0^{k-1} in (58)

$$Q_Z = Q_{Z_k \mid Z_0^{k-1}}(Z_k \mid \hat{z}_0^{k-1})$$
$$W^h_Z = W^h_{Z_k \mid Z_0^{k-1}}(Z_k \mid \hat{z}_0^{k-1})$$

and define $Q'_X \in \mathcal{E}_X$ by the relation $W^h_Z = P^h_{Y \mid X} Q'_X$. Then, we can continue the chain of inequalities as

$$E_f(Q) \le \min_{Q_Z \in \mathcal{E}_Z} \max\left\{D\left(Q_Z \,\middle\|\, W^{h_1}_Z\right), D\left(Q_Z \,\middle\|\, W^{h_2}_Z\right)\right\} \tag{59}$$

$$= \min_{Q_Z \in \mathcal{E}_Z} \max_{h \in \{h_1, h_2\}} \left\{D\left(Q_Z \,\middle\|\, W^h_Z\right)\right\} \tag{60}$$

$$= \min_{Q_{XY} \in \mathcal{E}_{XY}} \max_{h \in \{h_1, h_2\}} \left\{D\left(Q_{XY} \,\middle\|\, P^h_{Y \mid X} Q'_X\right)\right\} \tag{61}$$

$$\le \min_{Q_{Y \mid X} \in \mathcal{E}_{Y \mid X}} \max_{h \in \{h_1, h_2\}} \left\{D\left(Q_{Y \mid X} Q'_X \,\middle\|\, P^h_{Y \mid X} Q'_X\right)\right\} \tag{62}$$

$$= E_t(Q'_X), \tag{63}$$

where (62) follows from (61) by Proposition 1. Furthermore, in the case of two hypotheses, the optimal input distribution Q'_X which maximizes

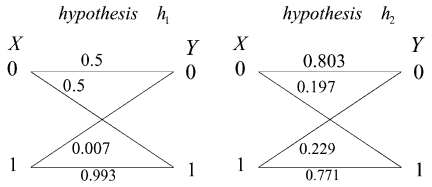

Fig. 1. The binary channels simulated in Fig. 2.

TABLE I
THE KULLBACK–LEIBLER DIVERGENCE [BASE 10] UNDER DIFFERENT INPUTS

	$x=0$	$x=1$	Best input x
$D(P_{Y\|X}^{h_1}(Y\|x)\|\|P_{Y\|X}^{h_2}(Y\|x))$	0.09938	0.09852	0
$D(P_{Y\|X}^{h_2}(Y\|x)\|\|P_{Y\|X}^{h_1}(Y\|x))$	0.08553	0.26215	1

$E_t(Q'_X)$ is a vertex of the simplex $\mathcal{E}_X$ and hence, a time-invariant finite memory policy exists which achieves this exponent (all inequalities hold with equality). ■

We note that the sequence of inequalities leading to (61) shows that when discriminating between two memoryless hypotheses, Markov policies of *any* order (deterministic or otherwise) perform no better than memoryless policies.

V. AN EXAMPLE

In this section, we numerically study a binary channel hypothesis scenario. In particular, consider the pair of channels illustrated in Fig. 1 with binary input and output alphabets.

Chernoff's signal selection strategy [3] tells us that if, based on all observations at time t, we have that the *a posteriori* probability of hypothesis h_1 is greater than that of h_2, we should select the next signal x_{t+1} that maximizes $D(P_{Y\mid X}^{h_1}[Y\mid x_{t+1}]\,\|\,P_{Y\mid X}^{h_2}[Y\mid x_{t+1}])$. Table I lists these values in base 10 and we see that in such a case, the best input is $x_{t+1}=0$. Likewise, if the *a posteriori* probability of h_2 is greater than that of h_1, one should select the next signal x_{t+1} that maximizes $D(P_{Y\mid X}^{h_2}[Y\mid x_{t+1}]\,\|\,P_{Y\mid X}^{h_1}[Y\mid x_{t+1}])$. This yields $x_{t+1}=1$. Thus, in this case, Chernoff's strategy is likely to yield alternating input signals.

We compare this signal selection strategy with the following rule: for all time t, select the signal x that maximizes

$$\min_{Q_Y\in\mathcal{E}_Y}\max\left\{D\left(Q_Y\;\middle\|\;P_{Y\mid X}^{h_1}[Y\mid x]\right),D\left(Q_Y\;\middle\|\;P_{Y\mid X}^{h_2}[Y\mid x]\right)\right\}. \tag{64}$$

In the example of Fig. 1, this rule says that a constant input $x=1$ should be chosen for all time. We have shown that this strategy is optimal in the following two ways.

1) From Theorem 5 and Corollary 1, it is the optimal (time-varying) open loop policy.
2) From Theorem 8, it is the optimal (time-invariant) Markov feedback policy for any order of Markov feedback.

Fig. 2 shows the average probability of decision error with these two schemes. In the solid curves, the prior probabilities of the hypotheses were chosen to be the uniform distribution $P[H=h_1]=P[H=h_2]=0.5$, which essentially reflects no prior knowledge of the hypotheses. The dashed curves are with $P[H=h_1]=0.95$ and $P[H=h_2]=0.05$. As expected, the actual asymptotic behavior is independent of the prior probabilities (provided neither is zero). From the figure, we clearly see that the Bayesian constant signal selection scheme (which employs no feedback) outperforms the scheme proposed by Chernoff [3] (which does employ feedback). This, however, is to be expected as Chernoff's scheme was not explicitly designed to be optimal in the Bayesian sense for fixed-length tests. Based on Theorems 5 and 7, we expect our scheme to have an asymptotic exponent of 0.0411 decades/input which compares well with an estimated exponent of 0.0416 decades/sample based on a window of samples from $t=100$ to $t=160$. By comparison, we measure an exponent of 0.0275 over the same window for Chernoff's scheme.

VI. CONCLUSION

We have derived the error exponent for open-loop, finite hypothesis channel identification. This result is related to the error exponent for Bayesian hypothesis testing. A somewhat surprising result has been shown for the case of detection between two hypotheses—in particular, it was shown that there is no advantage to mixing inputs: the input sequence distribution that maximizes the error exponent is achieved by consistently repeating the same input. Also, a simple numerical approach to calculating the error exponent for a given input distribution has been presented.

In the second (feedback) case, we have considered time-invariant Markov feedback policies. We have shown that regardless of the order k of the feedback policy, in the case of two hypotheses, the exponent is upper-bounded by an open-loop exponent $E_t(Q)$. Furthermore, for an optimal feedback policy, the bound is tight and equals that of an optimal open-loop exponent $E_t(Q)$.

In light of this, we suggested a simple signal selection strategy in a binary hypothesis scenario. The scheme was shown to be optimal in two different ways: optimal time-varying open loop and optimal time-invariant Markov of any order. Numerical results also show that thescheme outperforms an approach suggested by Chernoff, which is not unexpected since Chernoff's scheme was not designed to be optimal in the Bayesian setting employed here.

APPENDIX

The appendix contains technical lemmas.

Lemma 2: Let $\mathcal{U}$ and $\mathcal{V}$ be two compact metric spaces and $f:\mathcal{U}\times\mathcal{V}\to\mathbb{R}$ a continuous function. Let $u_1,u_2,\ldots$ be a sequence in $\mathcal{U}$ which converges to u. Then $\lim_{n\to\infty}\min_{v\in\mathcal{V}}f(u_n,v)=\min_{v\in\mathcal{V}}f(u,v)$. □

Proof: Since $\mathcal{U}\times\mathcal{V}$ is compact, f is uniformly continuous. Hence, for any $\epsilon>0$, $u\in\mathcal{U}$ and $v\in\mathcal{V}$, one can find $\delta(\epsilon)$ such that $|f(u,v)-f(u',v')|<\epsilon$ whenever $d(u,u')+d(v,v')<\delta(\epsilon)$.

It will suffice to show that $g(u)=\min_{v\in\mathcal{V}}f(u,v)$ is continuous. Now, fix u and u' such that $d(u,u')<\delta(\epsilon)$ and let v^* be a minimizer of $f(u,b)$, i.e., $v^*=\arg\min_v f(u,v)$. However,

$$g(u')=\min_v f(u',v)\le f(u',v^*)<f(u,v^*)+\epsilon=g(u)+\epsilon.$$

Hence, $g(u')-g(u)<\epsilon$. By interchanging the role of u and u', one may obtain the complementary relation $g(u)-g(u')<\epsilon$. ■

Lemma 3: For any sequence of types $Q_n\in\mathcal{Q}_X^n$ such that $Q_n\to Q\in\mathcal{E}_X$ and any $Q_{Y\mid X}\in\mathcal{E}_{Y\mid X}$, there exists a sequence of joint types $Q_{XY}^n\in\mathcal{Q}_{XY}^n$ with $Q_{XY}^n\to Q_{XY}\in\mathcal{E}_{XY}$ such that the X marginal of Q_{XY}^n is Q_n and $Q_{XY}[x^{(i)},y^{(j)}]=Q_{Y\mid X}[y^{(j)}\mid x^{(i)}]Q[x^{(i)}]$. □

Proof: We will show this by construction. First, pick a sequence x_1^n of type Q_n. Let k_n^i be the number of occurrences of $x^{(i)}$ in x_1^n. If $Q[x^{(i)}]>0$, then we must have $\lim_{n\to\infty}k_n^i=\infty$. Hence, asymptotically, one can pair $y^{(j)}$'s with the $x^{(i)}$'s such that the relative frequency of $y^{(j)}$ with respect to $x^{(i)}$ approaches any number between 0 and 1. Therefore, $Q_{XY}[x^{(i)},y^{(j)}]=Q_{Y\mid X}[y^{(j)}\mid x^{(i)}]Q[x^{(i)}]$.

Now, assume that $Q[x^{(i)}]=0$. Since by design, the type Q_{XY}^n has X marginal Q_n, we have that $Q_{XY}^n[x^{(i)},y^{(j)}]\le Q_n[x^{(i)}]$. Hence, it follows that $Q_{XY}[x^{(i)},y^{(j)}]=0$ since $\lim_{n\to\infty}Q_n[x^{(i)}]=0$. ■

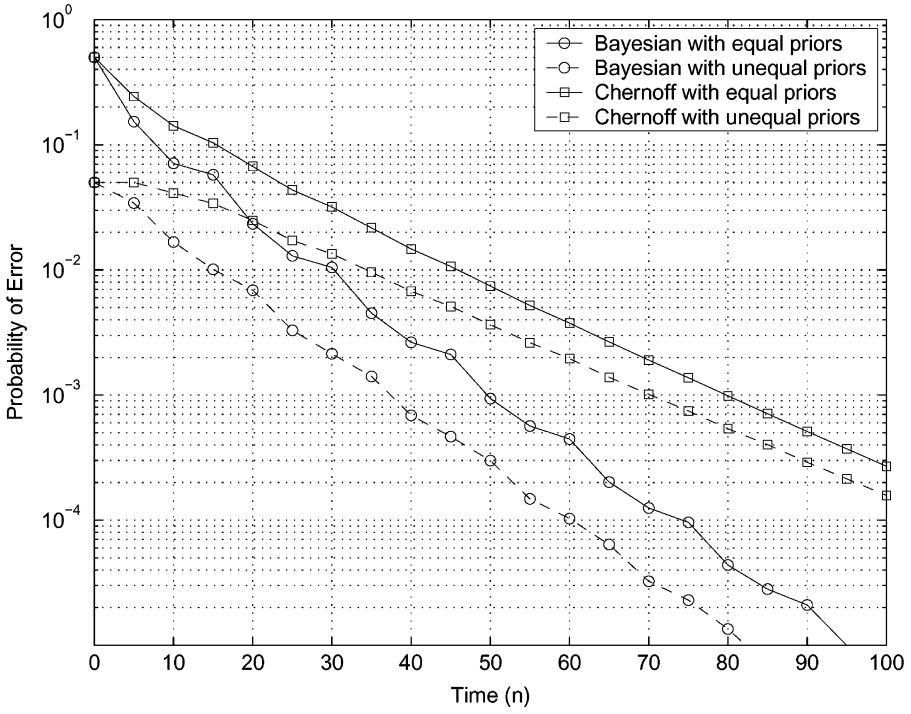

Fig. 2. Comparison of average probability of error using Bayesian approach and Chernoff's approach for binary channels in Fig. 1. We observe that the prior probabilities do not affect the error exponent, i.e., the asymptotic slopes of the curves are equal under both the equal and unequal prior scenarios.

Lemma 4: Let $Q_V \in \mathcal{E}_V$ and $Q_{U\,|\,V}, P^1_{U\,|\,V}, P^2_{U\,|\,V} \in \mathcal{E}_{U\,|\,V}$ with $P^1_{U\,|\,V} > 0$ and $P^2_{U\,|\,V} > 0$. If

$$D(Q_{U\,|\,V}Q_V \,\|\, P^1_{U\,|\,V}) > D(Q_{U\,|\,V}Q_V \,\|\, P^2_{U\,|\,V})$$

then there exists a $Q^*_{U\,|\,V}$ such that

$$D(Q_{U\,|\,V}Q_V \,\|\, P^1_{U\,|\,V}) > D(Q^*_{U\,|\,V}Q_V \,\|\, P^1_{U\,|\,V})$$

and

$$D(Q_{U\,|\,V}Q_V \,\|\, P^1_{U\,|\,V}) > D(Q^*_{U\,|\,V}Q_V \,\|\, P^2_{U\,|\,V}) \qquad \square$$

Proof: Consider the mixture

$$Q_{U\,|\,V}(\lambda) = (1-\lambda)Q_{U\,|\,V} + \lambda P^1_{U\,|\,V}$$

$\lambda \in [0.0, 1.0]$. Then both $D_1(\lambda) \triangleq D(Q_{U\,|\,V}(\lambda)Q_V \,\|\, P^1_{U\,|\,V})$ and $D_2(\lambda) \triangleq D(Q_{U\,|\,V}(\lambda)Q_V \,\|\, P^2_{U\,|\,V})$ are continuous on $[0.0, 1.0]$ and convex in λ as Kullback–Leibler divergence $D(Q \,\|\, P)$ is convex in Q. Furthermore, since $D_1(0) = D(Q_{U\,|\,V}Q_V \,\|\, P^1_{U\,|\,V}) > 0$ and $D_1(1) = 0$, we must have that $D_1(\lambda)$ is strictly decreasing for λ in some neighborhood of 0. Therefore, by choosing $\lambda > 0$ sufficiently small and by continuity of D_1 and D_2, the desired $Q^*_{U\,|\,V}$ may be found. ■

Lemma 5: Let $Q_{Z_0^k} \in \mathcal{F}_{Z_0^k}$ and consider two ergodic chains $P^1_{Z_k\,|\,Z_0^{k-1}}$ and $P^2_{Z_k\,|\,Z_0^{k-1}}$ with the same set of allowable state transitions. If

$$D(Q_{Z_0^k} \,\|\, P^1_{Z_k\,|\,Z_0^{k-1}}) > D(Q_{Z_0^k} \,\|\, P^2_{Z_k\,|\,Z_0^{k-1}})$$

then there exists a $Q^*_{Z_0^k}$ such that

$$D(Q_{Z_0^k} \,\|\, P^1_{Z_k\,|\,Z_0^{k-1}}) > D(Q^*_{Z_0^k} \,\|\, P^1_{Z_k\,|\,Z_0^{k-1}})$$

and

$$D(Q_{Z_0^k} \,\|\, P^1_{Z_k\,|\,Z_0^{k-1}}) > D(Q^*_{Z_0^k} \,\|\, P^2_{Z_k\,|\,Z_0^{k-1}}). \qquad \square$$

Proof: The proof is similar to that of Lemma 4. Because $P^1_{Z_k\,|\,Z_0^{k-1}}$ is ergodic, there is a unique invariant distribution $P^1_{Z_0^{k-1}}$ such that

$$P^1_{Z_0^k} = P^1_{Z_k\,|\,Z_0^{k-1}} P^1_{Z_0^{k-1}} \in \mathcal{F}_{Z_0^k}.$$

Furthermore, consider the mixture $Q_{Z_0^k}(\lambda) = (1-\lambda)Q_{Z_0^k} + \lambda P^1_{Z_0^k}$ for $\lambda \in [0,1]$. Then clearly, for all such λ, $Q_{Z_0^k}(\lambda) \in \mathcal{F}_{Z_0^k}$. Similar to the proof of Lemma 4, we have that $D_1(\lambda) = D(Q_{Z_0^k}(\lambda) \,\|\, P^1_{Z_k\,|\,Z_0^{k-1}})$ is continuous and convex on $[0,1]$ with

$$D_1(0) > D_2(0) = D(Q_{Z_0^k}(\lambda) \,\|\, P^2_{Z_k\,|\,Z_0^{k-1}}) \geq 0$$

and $D(1) = 0$. Hence, $D_1(\lambda)$ is strictly decreasing in some neighborhood of 0 and choosing λ sufficiently small will produce the desired $Q^*_{Z_0^k}$. ■

REFERENCES

[1] S. Arimoto and H. Kimura, "Optimum input test signals for systems identification—An information-theoretic approach," *Int. J. Syst. Sci.*, vol. 1, no. 3, pp. 279–290, 1971.

[2] P. Brémaud, *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*. New York: Springer-Verlag, 1998.

[3] H. Chernoff, *Sequential Analysis and Optimal Design*. Philadelphia, PA: Soc. Ind. Appl. Math, 1979.

[4] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[5] I. Csiszár, "The method of types," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2505–2523, Oct. 1998.

[6] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic, 1981.

[7] L. Davisson, G. Longo, and A. Sgarro, "The error exponent for the noiseless encoding of finite ergodic Markov sources," *IEEE Trans. Inf. Theory*, vol. IT-27, no. 4, pp. 431–438, Jul. 1981.

[8] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*. New York: Springer-Verlag, 1998.

[9] K. Itô, Ed., *Encyclopedic Dictionary of Mathematics*, 2nd ed. Cambridge, MA: MIT Press, 1987.

[10] S. Natarajan, "Large deviations, hypothesis testing, and source coding for finite Markov chains," *IEEE Trans. Inf. Theory*, vol. IT-31, no. 3, pp. 360–365, May 1985.
[11] O. Shalvi and E. Weinstein, "New criteria for blind deconvolution of minimum phase systems," *IEEE Trans. Inf. Theory*, vol. 36, no. 2, pp. 312–321, Mar. 1990.
[12] S. M. Sowelam and A. H. Tewfik, "Waveform selection in radar target classification," *IEEE Trans. Inf. Theory*, vol. 46, no. 3, pp. 1014–1029, May 2000.
[13] L. Tong, G. Xu, and T. Kailath, "Blind identification and equalization based on second-order statistics: A time domain approach," *IEEE Trans. Inf. Theory*, vol. 40, no. 2, pp. 340–349, Mar. 1994.
[14] H. L. Van Trees, *Detection, Estimation, and Modulation Theory*. New York: Wiley, 1968, pt. 1.

On the Exponential Convergence of Matching Pursuits in Quasi-Incoherent Dictionaries

Rémi Gribonval, *Member, IEEE,* and
Pierre Vandergheynst, *Member, IEEE*

***Abstract*—The purpose of this correspondence is to extend results by Villemoes and Temlyakov about exponential convergence of Matching Pursuit (MP) with some structured dictionaries for "simple" functions in finite or infinite dimension. The results are based on an extension of Tropp's results about Orthogonal Matching Pursuit (OMP) in finite dimension, with the observation that it does not only work for OMP but also for MP. The main contribution is a detailed analysis of the approximation and stability properties of MP with quasi-incoherent dictionaries, and a bound on the number of steps sufficient to reach an error no larger than a penalization factor times the best m-term approximation error.**

***Index Terms*—Dictionary, greedy algorithm, matching pursuit (MP), nonlinear approximation, sparse representation.**

I. Introduction

In a Hilbert space $\mathcal{H}$ of finite or infinite dimension, we consider the problem of getting m-term approximants of a function f from a possibly redundant *dictionary* $\mathcal{D} = \{g_k, k \in \mathbb{Z}\}$ of unit norm basis functions also called *atoms*. It will often be convenient to see a dictionary as a synthesis operator (or, in finite dimension, as a matrix)

$$\mathbf{D} : \boldsymbol{c} = (c_k) \mapsto \mathbf{D}\boldsymbol{c} = \sum_k c_k g_k$$

that maps sequences to vectors in $\mathcal{H}$. A special class of dictionaries that is widely used in signal and image processing is the family of frames: a dictionary $\mathcal{D}$ is a frame for $\mathcal{H}$ if, and only if, $\mathbf{D}$ is a bounded operator from ℓ^2 onto $\mathcal{H}$ [2]. However, in this correspondence, we consider dictionaries that may not be frames, hence $\mathbf{D}$ shall be defined essentially on sequences $\boldsymbol{c}$ with a finite number of nonzero entries. For any index set I (not necessarily finite) we will also consider the restricted synthesis operator

$$\mathbf{D}_I : \boldsymbol{c} \mapsto \mathbf{D}_I\boldsymbol{c} = \sum_{k \in I} c_k g_k$$

that corresponds to the subset $\mathcal{D}_I = \{g_k, k \in I\}$ of the full dictionary.

When $\mathcal{D}$ is an orthonormal basis for $\mathcal{H}$, it is well known how to get the best m-term approximant to any f: the solution is to keep the m atoms of the basis which have the largest inner products $|\langle f, g_k\rangle|$ with f. However, for arbitrary redundant dictionaries, the problem becomes NP-hard [3]. In the recent years, much effort has been made to understand what structure should be imposed on f (for a given dictionary) or on the dictionary itself so that good approximants can be obtained with computationally feasible algorithms.

One of the first algorithms that appeared in the signal processing community for approximating signals from a redundant dictionary was the Matching Pursuit (MP) algorithm of Mallat and Zhang [25], which iteratively decomposes the analyzed function f into an m-term approximant $f_m = \sum_{n=1}^{m} \alpha_n g_{k_n}$ and a residual $r_m = f - f_m$. MP is also known as Projection Pursuit in the statistics community [10], [22] and as a Pure Greedy Algorithm [27] in the approximation community. In finite dimension, MP is known to converge exponentially, i.e., for some $0 < \beta < 1$

$$\|r_m\|^2 = \|f_m - f\|^2 \leq \beta^m \cdot \|f\|^2, \qquad m \geq 1.$$

In infinite-dimensional Hilbert spaces, Jones [24] proved that MP is still convergent, i.e., $\|f_m - f\| \to 0$, but gave no estimate of the speed of convergence. DeVore and Temlyakov [4] exhibited a "bad" dictionary $\mathcal{D}$ where there exists a "simple" function (sum of two dictionary elements) for which MP gives "bad" approximations (i.e., with a slow convergence $\|f_m - f\| \geq Cm^{-1/2}$). On the positive side, Villemoes [30] showed that for Walsh wavelet packets, MP on "simple" functions ($f = c_i g_i + c_j g_j$ any sum of any two wavelet packets) was exponentially convergent (just as MP in finite dimension) with $\|f_m - f\|^2 \leq (3/4)^m \|f\|^2$. Temlyakov obtained similar results [26]: in particular, for f a function on the erval $[0, 1)$ taking constant values on a partition of $[0, 1)$ into n disjoint intervals, and $\mathcal{D}$ a highly redundant dictionary containing all (normalized) characteristic functions of intervals $I \subset [0, 1)$

$$\|f_m - f\|_2 \leq (1 - 1/n)^{m/2} \|f\|_2.$$

In this correspondence, we extend Villemoes and Temlyakov results about MP to more general dictionaries and "simple functions," as stated in the following featured theorem.

Featured Theorem 1: Let $\mathcal{D}$ be a dictionary in a finite- or infinite-dimensional Hilbert space and I an index set such that the stability condition (SC)

$$\eta(I) := \sup_{k \notin I} \|(\mathbf{D}_I)^\dagger g_k\|_1 < 1 \tag{1}$$

is met, where $(\cdot)^\dagger$ denotes pseudoinversion.[1] Then, for any $f = \sum_{k \in I} c_k g_k \in \operatorname{span}(g_k,\ k \in I)$, MP

1) picks up only correct atoms at each step: $(\forall n, k_n \in I)$;
2) if I is a finite set, then the residual r_m converges exponentially to zero.

The stability condition (1) may look fairly abstract, but for so-called *quasi-incoherent* dictionaries, one can obtain more explicit sufficient conditions [28]. For such dictionaries, we derive estimates of the rate of

[1]Basic reminders on pseudoinversion are given in Section II-E

Manuscript received April 9, 2004; revised October 13, 2005.

R. Gribonval is with IRISA-INRIA, Campus de Beaulieu, 35042 Rennes Cedex, France (e-mail: remi.gribonval@irisa.fr).

P. Vandergheynst is with the Signal Processing Institute, the Swiss Federal Institute of Technology (EPFL), CH-1015 Lausanne, Switzerland (e-mail: pierre.vandergheynst@epfl.ch).

Communicated by G. Battail, Associate Editor At Large.

Digital Object Identifier 10.1109/TIT.2005.860474