

Universal Anytime Coding

Stark C. Draper and Anant Sahai

Dept. of Electrical Engineering and Computer Sciences
University of California at Berkeley
Berkeley CA 94720 USA

Email: sdraper@eecs.berkeley.edu, sahai@eecs.berkeley.edu

Abstract—We show the existence of random codes without feedback for anytime coding over compound channels — i.e. codes that operate without knowledge of the transition probabilities of the communication channel. The anytime error exponent achieved is the same as if the decoder were given the probabilities governing the realized channel and allowed to do maximum likelihood decoding. Because of the equivalence between system stabilization over noisy feedback channels and communication over noisy channels, the results immediately give sufficient conditions for stabilization over compound channels.

I. INTRODUCTION

A fundamental control problem is the tracking and stabilization of an unstable process. When the observers and controllers are not co-located, the details of how sensor data is transmitted from observer to controller can severely impact overall control performance. In many cases the classic communication system structures and paradigms posed by Shannon do not suffice. Anytime codes have proved a basic tool in the solution of this problem [1]. Work in this area has usually assumed that there is a precise and perfectly known model for both the plant and the communication constraint.

However, in many communication contexts there is uncertainty about the probabilistic model for the channel law. Such uncertainty can have a substantial impact on the performance of block-coded communication systems. In a control application where delay and error requirements can be very tight, the degradation incurred may be intolerable.

These considerations motivate us to study an extension of anytime coding to so called “compound channels.” In this setting, the channel law is modeled as being known only to within a (possibly non-finite) set of possibilities. The code must be chosen without knowledge of which channel law will be realized, and the decoder must operate universally — without knowledge of the channel law selected by nature. The difference between universal anytime-coding and universal block-coding (see, e.g., [2] and [3] for an extensive review) is that in the anytime setting data bits are not all realized prior to transmission; rather they are realized in a on-line fashion as they would be at the observer of a control system.

In control applications the “anytime reliability” of the code proves to be crucial. This is closely connected to the code’s error exponent performance. We bound the error exponent achieved by our scheme. We provide the same performance guarantee as is demanded of a universal decoder for a compound channel in the the block coding context: the error

exponent performance of the code for whatever channel is realized must (asymptotically) match that achieved if the decoder were to be told the channel law (in which case it could perform maximum likelihood decoding). Naturally, if the channel law were known at design time one might chose to use a code matched to the channel. But, given that the code design must be fixed knowing only the set of possible channel laws, meeting the reliability function of maximum likelihood decoding is the most one can ask for.

Applying this result to the context of stabilization over a communication channel described in [1], this means that the observer and controller induced by the approach here will achieve a sense of stabilization in terms of η -moments stabilized that will be depend on the channel that is actually selected by nature. If nature picks a channel that has a higher random coding error exponent at the appropriate rate, then the closed-loop system will be stable for higher η -moments.

II. NOTATION

We use serifed-fonts x to indicate sample values, and sans-serif x to indicate random variables. Bolded fonts are reserved to indicate length- n sample or random vectors, $\mathbf{x} = x^n$ and $\mathbf{x} = x^n$. A subsequence such as x_l, x_{l+1}, \dots, x_n is denoted as x_l^n . Distributions are indicated with lower-case p , e.g., x is distributed according to $p_x(x)$. Sets and their elements are denoted as, e.g., $x \in \mathcal{X}$, and their cardinality by $|\mathcal{X}|$. We use standard notation for types, see, e.g., [2]. If a sequence \mathbf{x} is of type P is we write $\mathbf{x} \in \mathcal{T}_P$ and if \mathbf{y} is in the V -shell of \mathbf{x} we write $\mathbf{y} \in \mathcal{T}_V(\mathbf{x})$. We write the joint distribution of this (\mathbf{x}, \mathbf{y}) as $P \times V$.

III. MAIN RESULT

In this section we state the main contributions of this paper. We define an anytime code and its delay-dependent error exponent. We define a compound channel and state a lower bound on the achievable exponent of this class of codes for compound channels.

The techniques we use are adapted from those developed for universal block codes in [2]. Familiarity with those results is useful, but not assumed. We first developed ideas of universal streaming decoders in the closely related context of sequential Slepian-Wolf coding [4], [5].

Definition 1: A randomized anytime code for a channel with input set \mathcal{X} and output set \mathcal{Y} is a pair of sequence of

mappings $(\{\mathcal{E}_j\}, \{\mathcal{D}_j\})$ where $j = 1, 2, \dots$

$$\mathcal{E}_j : \mathcal{B}^j \rightarrow \mathcal{X}, \quad (1)$$

$$\mathcal{D}_j : \mathcal{Y}^j \rightarrow \mathcal{B}^j, \quad (2)$$

where \mathcal{B} is the data set and common randomness, shared between encoder and decoder, is assumed. The latter allows us to randomize the mapping independently of the data and channel sequences.

In this paper we study the delay-dependent reliability of anytime codes. We use b^n to denote the input data sequence known by the encoder at time n , and $\hat{b}^{n-\Delta}(y^n)$ to denote the delay- Δ decoder estimate. For compactness of notation we suppress the explicit dependence on y^n and simply write $\hat{b}^{n-\Delta}$.

An error exponent $E > 0$ is said to be achievable at rate R for a given channel if there exists a rate- R anytime code such that

$$\lim_{\Delta \rightarrow \infty} \lim_{n \rightarrow \infty} -\frac{1}{\Delta} \log \Pr[\hat{b}_{n-\Delta} \neq b_{n-\Delta}] \geq E. \quad (3)$$

Note that the order of limits is important since the total transmission time n is allowed to go to infinity faster than the Δ else we would be asking for decisions about bits before they are available at encoder.

For simplicity of exposition, in this paper we assume a constant $|\mathcal{B}| = 2^R$ and further that R is integer. Neither assumption is fundamental. Non-integer and non-rational rates can be accommodated by allowing a time-varying \mathcal{B}_j satisfying $|\mathcal{B}^n| = 2^{\lfloor nR \rfloor}$ for all n . As long as the decoding delay $\Delta \gg 1/R$, the results derived closely approximate those of the non-integer case. Since this condition is always true in the large delay limit, the lower bound on the error exponent that we derive applies to both integer and noninteger rates R .

Definition 2: (Following [2]) A compound channel with input set \mathcal{X} and output set \mathcal{Y} is a (not necessarily finite) family \mathcal{W} of channels $W : \mathcal{X} \rightarrow \mathcal{Y}$.

Definition 3: A compound discrete memoryless channel with input set \mathcal{X} and output set \mathcal{Y} is a sequence of compound channels $\{W^n : W \in \mathcal{W}\}$ where \mathcal{W} is some given (not necessarily finite) set of stochastic matrices $W : \mathcal{X} \rightarrow \mathcal{Y}$. The same matrix W is used for all time.

Our main result is an achievability theorem for anytime coding over compound channels.

Theorem 1: Given a rate $R > 0$ and compound discrete memoryless channel specified by \mathcal{W} there exists an anytime code such that for all $E < E_{\text{any,univ}}(R)$ there is a constant $K > 0$ such that $\Pr[\hat{b}_{n-\Delta} \neq b_{n-\Delta}] \leq K2^{-\Delta E}$ for all $n, \Delta \geq 0$ where

$$E_{\text{any,univ}}(R) = \sup_Q \inf_{W \in \mathcal{W}} \inf_{P, V} D(P \times V \| Q \times W) + |I(P, V) - R|^+, \quad (4)$$

and where $|\cdot|^+ = \max\{0, \cdot\}$.

IV. PROOF

The proof of our result proceeds in three main steps. First, we break the error events down into their delay-dependence. Then we introduce the decoding rule and use the characteristics of the codebook generation process to simplify the expression of error probability. Finally we show how the resulting expression can be bounded by an exponential that decays in Δ .

$$\Pr[\hat{b}_{n-\Delta} \neq b_{n-\Delta}] = \sum_{\mathbf{b} \in \mathcal{B}^n} p_{\mathbf{b}}(\mathbf{b}) \Pr[\hat{b}_{n-\Delta} \neq b_{n-\Delta} | \mathbf{b}] \quad (5)$$

$$\leq \sum_{\mathbf{b} \in \mathcal{B}^n} \Pr[\hat{b}^{n-\Delta} \neq b^{n-\Delta} | \mathbf{b}] p_{\mathbf{b}}(\mathbf{b})$$

$$= \sum_{\mathbf{b} \in \mathcal{B}^n} \Pr \left[\bigcup_{l=1}^{n-\Delta} (\hat{b}^{l-1} = b^{l-1}) \cap (\hat{b}_l \neq b_l) \middle| \mathbf{b} \right] p_{\mathbf{b}}(\mathbf{b}) \quad (6)$$

$$= \sum_{\mathbf{b} \in \mathcal{B}^n} \sum_{l=1}^{n-\Delta} \Pr \left[(\hat{b}^{l-1} = b^{l-1}) \cap (\hat{b}_l \neq b_l) \middle| \mathbf{b} \right] p_{\mathbf{b}}(\mathbf{b}) \quad (7)$$

$$= \sum_{l=1}^{n-\Delta} \sum_{\mathcal{C}} \sum_{\mathbf{b} \in \mathcal{B}^n} \sum_{\mathbf{y} \in \mathcal{Y}^n} \Pr \left[(\hat{b}^{l-1} = b^{l-1}) \cap (\hat{b}_l \neq b_l) \middle| \mathbf{b}, \mathcal{C}, \mathbf{y} \right] p_{\mathbf{b}}(\mathbf{b}) \Pr[\mathcal{C}] \Pr[\mathbf{y} = \mathbf{y} | \mathbf{b}, \mathcal{C}] \quad (8)$$

$$= \sum_{l=1}^{n-\Delta} \sum_{\mathcal{C}} \sum_{\mathbf{y} \in \mathcal{Y}^n} \Pr \left[(\hat{b}^{l-1} = 0^{(l-1)R}) \cap (\hat{b}_l \neq 0^R) \middle| 0^{nR}, \mathcal{C}, \mathbf{y} \right] \Pr[\mathcal{C}] \Pr[\mathbf{y} = \mathbf{y} | 0^{nR}, \mathcal{C}] \quad (9)$$

In the conditioning on the data sequence in (5), \mathcal{B} is the data alphabet. In (6)–(7) we decompose the error event into a number of mutually exclusive events by partitioning all possible data estimates $\hat{\mathbf{b}}$ into disjoint sets defined by the time l of the first sample in which they differ from the true data sequence \mathbf{b} . In (8) we condition on the randomly generated codebook \mathcal{C} and the observed sequence \mathbf{y} . The codebook is independent of the data sequence while the observation depends on both the codebook and the data sequence. Note that the sum over \mathcal{C} in (8) is a shorthand that denotes the marginalization over all codeword realizations. We choose the codewords to be mutually independent and identically distributed. The error probability therefore does not depend on the particular codeword transmitted and so in (9) we can assume that the realized data sequence is the all-zeros sequence. (Recall that for simplicity of presentation we assume that R is integer.)

We now introduce our universal decoding rule. The decoding rule creates its estimate of the information symbols in a symbol-by-symbol manner. The l th symbol of the estimate \hat{b}_l at time $n, l \in \{1, 2, \dots, n\}$, is defined as

$$\hat{b}_l = w_l[l] \quad \text{where} \\ w^n[l] = \arg \max_{\tilde{b}^n \in \mathcal{B}^n \text{ s.t. } \tilde{b}^{l-1} = \hat{b}^{l-1}} I(x_l^n(\tilde{\mathbf{b}}); y_l^n), \quad (10)$$

and $I(x_l^n(\tilde{\mathbf{b}}); y_l^n)$ denotes the empirical mutual information between the last $n - l + 1$ symbols of the codeword \mathbf{x} corresponding to information sequence $\tilde{\mathbf{b}}$ and the last $n - l + 1$ symbols of the observation \mathbf{y} .

This decoding rule is decision-directed. At each decoding opportunity (i.e., at each n) estimates are made of each information bit, starting from the earliest. Earlier bits are fixed to their estimate when estimating later bits. While sequential hard decisions are made for each information bit at each decoding opportunity, as soon as the next channel observation is received all estimates are recalculated.

Each suffix comparison in (10) uses a maximum mutual information decoding rule (operating just on the suffix). The error exponent of maximum mutual information and maximum likelihood block decoding have the same reliability function, see, e.g., [2]. The overall reliability will be determined by the shortest suffix comparison, the one of length Δ . Therefore we expect reliability to decay exponentially in Δ .

The reason we use this ‘‘maximum suffix-mutual-information’’ decoder rather than the standard maximum mutual information decoder [2] is that each suffix comparison has a polynomial term in the suffix length $n - l + 1$ that results from summing over the type class of possible channel laws. This term multiplies an exponential decay in $n - l + 1$. If we were to use straight block maximum mutual information decoding the polynomial term would be in n which would dominate the decay in $n - l + 1$ for large n and all finite l . The decay of the probability of decoding error on messages with delay would thus differ from bit-position to bit-position. This is not acceptable for anytime coding and is incompatible with any sense of ‘‘steady state’’ behavior in a control system.

We now continue our derivation by indicating the events that lead to errors in the decoding rule (10). Errors are only made if two events occur. First, some erroneous information sequence matches the true information sequence up to a certain point. And second, after that point the suffix of the corresponding erroneous codeword and observation have an empirical mutual information greater than that between the suffix of the transmitted codeword and the observation.

$$\begin{aligned} & \Pr[\hat{b}_{n-\Delta} \neq b_{n-\Delta}] \\ & \leq \sum_{l=1}^{n-\Delta} \sum_{\mathcal{C}} \sum_{\mathbf{y} \in \mathcal{Y}^n} \Pr \left[\exists \mathbf{x}(\tilde{\mathbf{b}}) \in \mathcal{C} \text{ s.t.} \right. \\ & \quad \left(\tilde{b}^{l-1} = 0^{(l-1)R} \right) \cap \left(\tilde{b}_l \neq 0^R \right) \cap \\ & \quad \left. \left(I(x_l^n(\tilde{\mathbf{b}}); y_l^n) > I(x_l^n(0^{nR}); y_l^n) \right) \middle| 0^{nR}, \mathcal{C}, \mathbf{y} \right] \\ & \Pr[\mathcal{C}] \Pr[\mathbf{y} = \mathbf{y} | 0^{nR}, \mathcal{C}] \end{aligned} \quad (11)$$

$$\begin{aligned} & = \sum_{l=1}^{n-\Delta} \sum_{\mathbf{x}(1) \in \mathcal{X}^n} p_{\mathbf{x}}(\mathbf{x}(1)) \sum_{\mathbf{y} \in \mathcal{Y}^n} \\ & \quad \left[\sum_{\mathcal{C}/\mathbf{x}(1)} \Pr[\mathbf{x}(2), \mathbf{x}(3), \dots, \mathbf{x}(2^{nR})] \right. \\ & \quad \Pr \left[\exists \mathbf{x}(\tilde{\mathbf{b}}) \in \mathcal{C} \text{ s.t.} \left(\tilde{b}^{l-1} = 0^{(l-1)R} \right) \cap \left(\tilde{b}_l \neq 0^R \right) \right. \\ & \quad \left. \left. \cap \left(I(x_l^n(\tilde{\mathbf{b}}); y_l^n) > I(x_l^n(0^{nR}); y_l^n) \right) \middle| 0^{nR}, \mathcal{C}, \mathbf{y} \right] \right] \\ & \quad \Pr[\mathbf{y} = \mathbf{y} | \mathbf{x}(1)] \end{aligned} \quad (12)$$

$$\begin{aligned} & \leq \sum_{l=1}^{n-\Delta} \sum_{\mathbf{x}(1) \in \mathcal{X}^n} p_{\mathbf{x}}(\mathbf{x}(1)) \sum_{\mathbf{y} \in \mathcal{Y}^n} \Pr[\mathbf{y} = \mathbf{y} | \mathbf{x}(1)] \\ & \quad \min \left\{ 1, \sum_{\mathcal{C}/\mathbf{x}(1)} \prod_{i=2}^{2^{nR}} p_{\mathbf{x}}(\mathbf{x}(i)) \sum_{\substack{\tilde{\mathbf{b}} \text{ s.t. } \tilde{b}_l \neq 0^R, \\ \tilde{b}^{l-1} = 0^{(l-1)R}}} \right. \\ & \quad \left. 1 \left[I(x_l^n(\tilde{\mathbf{b}}); y_l^n) > I(x_l^n(0^{nR}); y_l^n) \middle| 0^{nR}, \mathcal{C}, \mathbf{y} \right] \right\} \end{aligned} \quad (13)$$

$$\begin{aligned} & \leq \sum_{l=1}^{n-\Delta} \sum_{\mathbf{x}(1) \in \mathcal{X}^n} p_{\mathbf{x}}(\mathbf{x}(1)) \sum_{\mathbf{y} \in \mathcal{Y}^n} \Pr[\mathbf{y} = \mathbf{y} | \mathbf{x}(1)] \\ & \quad \min \left\{ 1, \sum_{\substack{\tilde{\mathbf{b}} \text{ s.t. } \tilde{b}_l \neq 0^R, \\ \tilde{b}^{l-1} = 0^{(l-1)R}}} \sum_{\mathbf{x}(\tilde{\mathbf{b}}) \in \mathcal{X}^n} p_{\mathbf{x}}(\mathbf{x}(\tilde{\mathbf{b}})) \right. \\ & \quad \left. 1 \left[I(x_l^n(\tilde{\mathbf{b}}); y_l^n) > I(x_l^n(0^{nR}); y_l^n) \middle| 0^{nR}, \mathcal{C}, \mathbf{y} \right] \right\} \end{aligned} \quad (14)$$

$$\begin{aligned} & \leq \sum_{l=1}^{n-\Delta} \sum_{\mathbf{x}(1) \in \mathcal{X}^n} p_{\mathbf{x}}(\mathbf{x}(1)) \sum_{\mathbf{y} \in \mathcal{Y}^n} \Pr[\mathbf{y} = \mathbf{y} | \mathbf{x}(1)] \\ & \quad \min \left\{ 1, \sum_{\substack{\tilde{\mathbf{b}} \text{ s.t. } \tilde{b}_l \neq 0^R, \\ \tilde{b}^{l-1} = 0^{(l-1)R}}} \sum_{\tilde{\mathbf{x}}^{n-l+1} \in \mathcal{X}^{n-l+1}} p_{\mathbf{x}_l^n}(\tilde{\mathbf{x}}^{n-l+1}) \right. \\ & \quad \left. 1 \left[I(\tilde{\mathbf{x}}^{n-l+1}; y_l^n) > I(x_l^n(0^{nR}); y_l^n) \middle| 0^{nR}, \mathcal{C}, \mathbf{y} \right] \right\} \end{aligned} \quad (15)$$

$$\begin{aligned} & \leq \sum_{l=1}^{n-\Delta} \sum_{x_l^n(1) \in \mathcal{X}^{n-l+1}} p_{x_l^n}(x_l^n(1)) \sum_{y_l^n \in \mathcal{Y}^{n-l+1}} \Pr[y_l^n = y_l^n | x_l^n(1)] \\ & \quad \min \left\{ 1, 2^{(n-l+1)R} \sum_{\tilde{\mathbf{x}}^{n-l+1} \in \mathcal{X}^{n-l+1}} p_{x_l^n}(\tilde{\mathbf{x}}^{n-l+1}) \right. \\ & \quad \left. 1 \left[I(\tilde{\mathbf{x}}^{n-l+1}; y_l^n) > I(x_l^n(0^{nR}); y_l^n) \middle| 0^{nR}, \mathcal{C}, \mathbf{y} \right] \right\} \end{aligned} \quad (16)$$

$$\begin{aligned} & = \sum_{l=1}^{n-\Delta} \sum_P \sum_{x_l^n \in \mathcal{T}_P} \Pr[x_l^n] \sum_V \sum_{y_l^n \in \mathcal{T}_V(x_l^n)} \Pr[y_l^n | x_l^n] \\ & \quad \min \left\{ 1, \sum_{\bar{P}} \sum_{\substack{\bar{V} \text{ s.t.} \\ I(\bar{P}, \bar{V}) \geq I(P, V)}} \sum_{\tilde{\mathbf{x}}^{n-l+1} \in \mathcal{T}_{\bar{P}}} \right. \\ & \quad \left. \Pr[\tilde{\mathbf{x}}^{n-l+1}] 1[y_l^n \in \mathcal{T}_{\bar{V}}(\tilde{\mathbf{x}}^{n-l+1})] 2^{(n-l+1)R} \right\} \end{aligned} \quad (17)$$

Since the observation \mathbf{y} is only dependent on the transmitted codeword $\mathbf{x}(\mathbf{b})$, in (12) we can interchange the averaging over the non-transmitted codeword realizations $\mathcal{C}/\mathbf{x}(1)$ with the sum over \mathbf{y} . In (12) we introduce a slight abuse of notation in our codeword indexing. We allow a codeword to be indexed either by its information sequence, e.g., $\mathbf{x}(\mathbf{b})$,

or by a scalar index, e.g., $\mathbf{x}(j)$, where $j \in \{1, 2, \dots, 2^{nR}\}$. We identify message $j = 1$ with the all-zeros data sequence, i.e., $\mathbf{x}(1) = \mathbf{x}(0^{nR})$. Our usage will be clear from the context.

In (13) we use the union bound and the fact that the codewords are identically distributed and mutually independent. We use the notation $1[\cdot]$ for the indicator function. Because the decoding rule is a pairwise comparison, in (14) we need only consider the particular codeword under consideration, determined by $\tilde{\mathbf{b}}$ and marginalize out the rest. We generate the symbols in each codeword in an independent and identically distributed manner. Therefore, in (15) we can marginalize out the prefix of the non-transmitted codeword under consideration. Similarly we marginalize out the prefix of the true codeword and the observation. In (16) there are $2^{(n-l+1)R}$ data sequences that satisfy the conditions given in (15). In (17) we define P to be the type of the length- $(n-l+1)$ suffix of the transmitted codeword x_l^n , and V to be the conditional type relating x_l^n to y_l^n . \tilde{P} and \tilde{V} are similarly defined by \tilde{x}_l^n and y_l^n .

We now bound the term $\sum_{\tilde{\mathbf{x}}^{n-l+1}} \Pr[\tilde{x}^{n-l+1}] \Pr[y_l^n \in \mathcal{T}_{\tilde{V}}(\tilde{x}^{n-l+1})]$ in (17). For simplicity we drop the explicit time indexes of (17), assume the vectors are of length N , and recall the the input distribution is Q .

$$\begin{aligned} & \sum_{\tilde{\mathbf{x}} \in \mathcal{T}_{\tilde{P}}} \Pr[\tilde{\mathbf{x}}] 1[\mathbf{y} \in \mathcal{T}_{\tilde{V}}(\tilde{\mathbf{x}})] \\ &= \sum_{\tilde{\mathbf{x}} \in \mathcal{T}_{\tilde{P}}} 2^{-N[D(\tilde{P}\|Q)+H(\tilde{P})]} 1[\mathbf{y} \in \mathcal{T}_{\tilde{V}}(\tilde{\mathbf{x}})] \\ &\leq 2^{-ND(\tilde{P}\|Q)} \sum_{\tilde{\mathbf{x}} \in \mathcal{T}_{\tilde{P}}} \frac{1[\mathbf{y} \in \mathcal{T}_{\tilde{V}}(\tilde{\mathbf{x}})]}{|\mathcal{T}_{\tilde{P}}|} \\ &= 2^{-ND(\tilde{P}\|Q)} \frac{|\{\tilde{\mathbf{x}} : \tilde{\mathbf{x}} \in \mathcal{T}_{\tilde{P}}, \mathbf{y} \in \mathcal{T}_{\tilde{V}}(\tilde{\mathbf{x}})\}|}{|\mathcal{T}_{\tilde{P}}|} \\ &\leq 2^{-ND(\tilde{P}\|Q)} (N+1)^{|\mathcal{X}|} 2^{-NI(\tilde{P}, \tilde{V})} \end{aligned}$$

The last inequality, bounding the probability that a fixed \mathbf{y} falls in the \tilde{V} shell of any $\tilde{\mathbf{x}}$ picked uniformly at random from type class $\mathcal{T}_{\tilde{P}}$ is proved in [2, p. 164].

Using this result in (17) with $N = n-l+1$ gives

$$\begin{aligned} & \Pr[\hat{b}_{n-\Delta} \neq b_{n-\Delta}] \\ &\leq \sum_{l=1}^{n-\Delta} \sum_P \sum_{x_l^n \in \mathcal{T}_P} \Pr[x_l^n] \sum_V \sum_{y_l^n \in \mathcal{T}_V(x_l^n)} \Pr[y_l^n | x_l^n] \\ &\quad \min \left\{ 1, \sum_{\tilde{P}} (n-l+2)^{|\mathcal{X}|} 2^{-(n-l+1)D(\tilde{P}\|Q)} \right. \\ &\quad \left. \sum_{\substack{\tilde{V} \text{ s.t.} \\ I(\tilde{P}, \tilde{V}) \geq I(P, V)}} 2^{-(n-l+1)[I(\tilde{P}, \tilde{V})-R]} \right\} \quad (18) \\ &\leq \sum_{l=1}^{n-\Delta} \sum_P \sum_{x_l^n \in \mathcal{T}_P} \Pr[x_l^n] \sum_V \sum_{y_l^n \in \mathcal{T}_V(x_l^n)} \Pr[y_l^n | x_l^n] \\ &\quad \min \left\{ 1, (n-l+2)^{2|\mathcal{X}|} \sum_{\tilde{V}} 2^{-(n-l+1)[I(P, V)-R]} \right\} \end{aligned}$$

$$\begin{aligned} & \leq \sum_{l=1}^{n-\Delta} (n-l+2)^{2|\mathcal{X}|} \sum_P 2^{-(n-l+1)D(P\|Q)} \\ & \quad \sum_V 2^{-(n-l+1)D(V\|W|P)} \sum_{\tilde{V}} 2^{-(n-l+1)|I(P, V)-R|^+} \quad (19) \end{aligned}$$

$$\begin{aligned} & \leq \sum_{l=1}^{n-\Delta} (n-l+2)^{4|\mathcal{X}||\mathcal{Y}|} \\ & \quad 2^{-(n-l+1) \min_{P, V} [D(V \times P \| W \times Q) + |I(P, V) - R|^+]} \quad (20) \end{aligned}$$

$$\begin{aligned} & \leq \sum_{l=1}^{n-\Delta} K 2^{-(n-l+1)[\min_{P, V} [D(V \times P \| W \times Q) + |I(P, V) - R|^+] - \gamma]} \quad (21) \end{aligned}$$

$$\leq K' 2^{-\Delta[\min_{P, V} [D(V \times P \| W \times Q) + |I(P, V) - R|^+] - \gamma]} \quad (22)$$

In (19) we use the fact that \mathbf{x} is generated in an i.i.d. manner from the distribution Q and the true channel law is W . In (21) we incorporate the polynomial into the exponent. Namely, for all $a > 0$, $b > 0$, there exists a K such that $z^a \leq K 2^{b(z-1)}$ for all $z \geq 1$. Therefore in (21) the $\gamma > 0$ can be set arbitrarily small. Pulling out the exponent in Δ , the remaining summation is a sum over decaying exponentials, and can be bounded by a constant. Combining this constant with K gives the constant K' in (22).

Since the channel law W is not used in the decoding rule (10), (22) holds for all W . Therefore, choosing Q as in Thm. 1 and taking the limits defined in (3) means that the exponent defined in Thm. 1 is universally attained.

V. APPLICATION TO STABILIZATION: A NEW SUFFICIENT CONDITION

The sufficiency results of [1] say that the η -moment of the closed-loop state of a scalar system with unstable open-loop gain λ can be made bounded if a fixed channel is available whose random coding error exponent $E_r(\log \lambda) > \eta \log \lambda$. Viewed in a different way, it says that the η -moments of the closed-loop system will be finite as long as $\eta < \frac{E_r(\log \lambda)}{\log \lambda}$. Moreover, [1] shows that this can be done using a very simple “nearly memoryless” encoder structure based on random binning. The result here implies that if the same basic encoder structure is used based on the input distribution Q , then the closed-loop η -moments up to

$$\eta < \frac{\inf_{P, V} D(P \times V \| Q \times W) + |I(P, V) - \log \lambda|^+}{\log \lambda} \quad (23)$$

will be finite when the decoding rule given here is used.

While the results in (4) give no easily computable bound on this, the solution (given in [6]) to exercise 5.23 in [7] tells us that the random-coding error exponent is lower-bounded by a simple function $\frac{(C-R)^2}{\frac{8}{e^2} + 4(\ln |\mathcal{Y}|)^2}$ (when both rate and exponents are expressed in natural ln units rather than the log units used here.). This depends only on the output-alphabet size and the square of the *gap* to capacity.

Thus, the results of this paper are easily adapted to give rise to a *universal sufficient condition* for η -stabilization of a system with a compound channel on the feedback link. By

simple algebra, the η -moment of the closed-loop system can be made finite if a compound channel with output alphabet \mathcal{Y} has a capacity (in nats):

$$C(\mathcal{W}) > \ln \lambda + (2 \ln |\mathcal{Y}|) \sqrt{(\eta \ln \lambda) \left(1 + \frac{2}{e^{2(\ln |\mathcal{Y}|)^2}}\right)} \quad (24)$$

Converting to traditional \log_2 units (bits) and viewing¹ this in terms of the excess capacity required for performance beyond the $\log \lambda$ needed for mere stabilization gives:

$$C(\mathcal{W}) - \log \lambda > (2 \log |\mathcal{Y}|) \sqrt{\frac{\eta \log \lambda}{\log e} \left(1 + \frac{2(\log e)^2}{e^{2(\log |\mathcal{Y}|)^2}}\right)}. \quad (25)$$

This bound is quite conservative, but has the advantage of giving an easy to evaluate sufficient condition for the compound channel capacity $C(\mathcal{W})$ only in terms of the size $|\mathcal{Y}|$ of the channel output alphabet, the unstable gain λ , and the desired moment η . The interesting fact is that the rate overhead to get the η -moment stabilized seems to grow proportional to $\sqrt{\eta \log \lambda}$. Normalizing by $\log |\mathcal{Y}|$, the condition (25) becomes:

$$\frac{C(\mathcal{W})}{\log |\mathcal{Y}|} - \frac{\log \lambda}{\log |\mathcal{Y}|} > \sqrt{\frac{4\eta \log \lambda}{\log e} \left(1 + \frac{2(\log e)^2}{e^{2(\log |\mathcal{Y}|)^2}}\right)}. \quad (26)$$

Given that $\frac{C(\mathcal{W})}{\log |\mathcal{Y}|} < 1$ measures the factor deviation from an ideal channel with the same output alphabet, this result implies that even as the output alphabet size goes to infinity this approach requires $\frac{4\eta \log \lambda}{\log e} < 1$ or $\lambda < \exp(\frac{1}{4\eta})$. This upper bound on λ is curious and it would be interesting to see if this is merely an artifact of the bounding techniques used here or whether some upper bound like this is essential whenever the channel is only specified in terms of the compound channel capacity. Such a bound would reflect a fundamental performance penalty due to ignorance of the channel.

VI. CONCLUSION

In this paper we have developed universal anytime codes for compound channels. The decoding rule for these codes is different from that of universal channel codes. In particular we must use a suffix-decoding rule. The resulting exponent achieved is identical to the random coding exponent of block coding for compound channels.

The main challenge of the presented algorithm is that as $n \rightarrow \infty$, the decoding complexity grows without bound since the decoder is required to revisit all decoding estimates with every channel output. Initial work in complexity-limited anytime decoders (for a fixed and precisely known channel model) is presented in [8].

Furthermore, as applied to control with communication constraints, the results here suffer from the same limitation as [1]. They only address closed-loop performance obliquely through the question of which η -moments are guaranteed to be finite. As far as more standard notions of closed-loop performance

go, it would be interesting to see how the results here interact with Tatikonda's sequential rate distortion framework of [9], [10], [11]. In particular, it would be interesting to see if there are ever situations where $D_{seq}(C(\mathcal{W}))$ can be attained in closed-loop.

REFERENCES

- [1] A. Sahai and S. Mitter, "The necessity and sufficiency of anytime capacity for stabilization of a linear system over a noisy communication link—Part I: Scalar systems," *IEEE Trans. Inform. Theory*, vol. 52, pp. 3369–3395, Aug. 2006.
- [2] I. Csiszár and J. Körner, *Information Theory, Coding Theorems for Discrete Memoryless Systems*. Akadémiai Kiadó, 1981.
- [3] A. Lapidoth and P. Narayan, "Reliable communication under channel uncertainty," *IEEE Trans. Inform. Theory*, vol. 44, pp. 2148–2177, Oct. 1998.
- [4] S. C. Draper, C. Chang, and A. Sahai, "Sequential random binning for streaming distributed source coding," in *Proc. Int. Symp. Inform. Theory*, Adelaide, Australia, 2005.
- [5] C. Chang, S. C. Draper, and A. Sahai, "Lossless coding for distributed streaming sources," *IEEE Trans. Inform. Theory*, submitted.
- [6] C. Chang and A. Sahai, "Solution to gallager's exercise 5.23," http://www.eecs.berkeley.edu/~cchang/5_23.pdf, 2006.
- [7] R. G. Gallager, *Information Theory and Reliable Communication*. John Wiley and Sons, 1968.
- [8] A. Sahai and H. Palaiyanur, "A simple encoding and decoding strategy for stabilization over discrete memoryless channels," in *Proc. 43rd Allerton Conf. on Communication, Control and Computing*, Allerton House, Monticello, IL, Sept. 2005.
- [9] S. Tatikonda, A. Sahai, and S. K. Mitter, "Stochastic linear control over a communication channel," vol. 49, no. 9, pp. 1549–1561, Sept. 2004.
- [10] S. Tatikonda and S. K. Mitter, "Control under communication constraints," vol. 49, no. 7, pp. 1056–1068, July 2004.
- [11] S. Tatikonda, "Control under communication constraints," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, 2000.

¹The form derived in (25) is also helpful in terms of thinking of units. The quantity within the square-root is clearly unitless.