Source coding and channel requirements for unstable processes

Anant Sahai, Sanjoy Mitter

sahai@eecs.berkeley.edu, mitter@mit.edu

Abstract

Our understanding of information in systems has been based on the foundation of memoryless processes. Extensions to stable Markov and auto-regressive processes are classical. Berger proved a source coding theorem for the marginally unstable Wiener process, but the infinite-horizon exponentially unstable case has been open since Gray's 1970 paper. There were also no theorems showing what is needed to communicate such processes across noisy channels.

In this work, we give a fixed-rate source-coding theorem for the infinite-horizon problem of coding an exponentially unstable Markov process. The encoding naturally results in two distinct bitstreams that have qualitatively different QoS requirements for communicating over a noisy medium. The first stream captures the information that is accumulating within the nonstationary process and requires sufficient anytime reliability from the channel used to communicate the process. The second stream captures the historical information that dissipates within the process and is essentially classical. This historical information can also be identified with a natural stable counterpart to the unstable process. A converse demonstrating the fundamentally layered nature of unstable sources is given by means of information-embedding ideas.

Index Terms

Nonstationary processes, rate-distortion, anytime reliability, information embedding

Department of Electrical Engineering and Computer Science at the University of California at Berkeley. A few of these results were presented at ISIT 2004 and a primitive form of others appeared at ISIT 2000 and in his doctoral dissertation.

Department of Electrical Engineering and Computer Science at the Massachusetts Institute of Technology. Support for S.K. Mitter was provided by the Army Research Office under the MURI Grant: Data Fusion in Large Arrays of Microsensors DAAD19-00-1-0466 and the Department of Defense MURI Grant: Complex Adaptive Networks for Cooperative Control Subaward #03-132 and the National Science Foundation Grant CCR-0325774.

Source coding and channel requirements for unstable processes

I. Introduction

The source and channel models studied in information theory are not just interesting in their own right, but also provide insights into the architecture of reliable communication systems. Since Shannon's work, memoryless sources and channels have always been at the base of our understanding. They have provided the key insight of separating source and channel coding with the bit rate alone appearing at the interface [1], [2]. The basic story has been extended to many different sources and channels with memory for point-to-point communication [3].

However, there are still many issues for which information theoretic understanding eludes us. Networking in particular has a whole host of such issues, leading Ephremides and Hajek to entitle their survey article "Information Theory and Communication Networks: An Unconsummated Union!" [4]. They comment:

> The interaction of source coding with network-induced delay cuts across the classical network layers and has to be better understood. The interplay between the distortion of the source output and the delay distortion induced on the queue that this source output feeds into may hold the secret of a deeper connection between information theory. Again, feedback and delay considerations are important.

Real communication networks and networked applications are quite complicated. To move toward a quantitative and qualitative of understanding of the issues, tractable models that exhibit at least some of the right qualitative behavior are essential. In [5], [6], the problem of stabilization of unstable plants across a noisy feedback link is considered. There, delay and feedback considerations become intertwined and the notion of feedback anytime capacity is introduced. To stabilize an otherwise unstable plant over a noisy channel, not only is it necessary to have a channel capable of supporting a certain minimal rate, but the channel when used with noiseless feedback must also support a high enough error-exponent (called the anytime reliability) with fixed delay in a delay-universal fashion. This turns out to be a sufficient condition as well, thereby establishing a separation theorem for stabilization. In [7], upper bounds are given for the fixed-delay reliability functions of DMCs with and without feedback, and these bounds are shown to be tight for certain classes of channels. Moreover, the fixed-delay reliability functions with feedback are shown to be fundamentally better than the traditional fixed-block-length reliability functions.

While the stabilization problem does provide certain important insights into interactive applications, the separation theorem for stabilization given in [5], [6] is coarse — it only addresses performance as a binary valued entity: stabilized or not stabilized. All that matters is the tail-behavior of the closed-loop process. To get a more refined view in terms of steady-state performance, this paper instead considers the corresponding open-loop estimation problem. This is the seemingly classical question of lossy source coding for an *unstable* scalar Markov processes — mapping the source into bits and then seeing what is required to communicate such bits using a point-to-point communication system.

A. Communication of Markov processes

Coding theorems for stable Markov and auto-regressive processes under mean-squared-error distortion are now well established in the literature [8], [9]. We consider real-valued Markov processes, modeled as

$$X_{t+1} = \lambda X_t + W_t \tag{1}$$

where $\{W_t\}_{t\geq 0}$ are white and X_0 is an independent initial condition uniformly distributed on $[-\frac{\Omega_0}{2}, +\frac{\Omega_0}{2}]$ where $\Omega_0 > 0$ is small. The essence of the problem is depicted in Fig. 1: to minimize the rate of the encoding while maintaining an adequate fidelity of reconstruction. Once the source has been compressed, the resulting bitstreams can presumably be reliably communicated across a wide variety of noisy channels.

The infinite-horizon source-coding problem is to design a source code minimizing the rate R used to encode the process while keeping the reconstruction close to the original source in an average sense

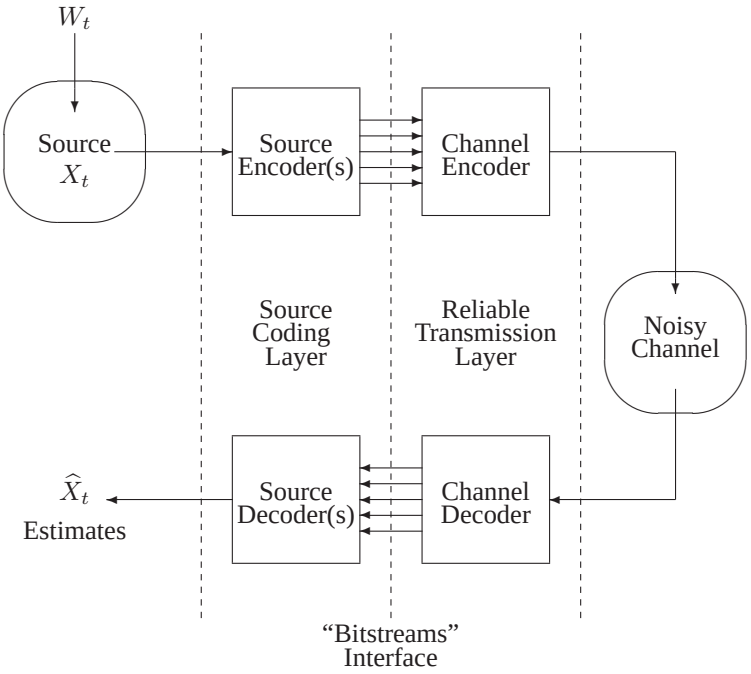

Fig. 1. The point-to-point communication problem considered here. The goal is to minimize end-to-end average distortion $\rho(X_t, \widehat{X}_t)$. Finite, but possible large, end-to-end delay will be permitted. One of the key issues explored is what must be made available at the source/channel interface.

$\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} E[|X_t - \widehat{X}_t|^\eta]$. *The key issue is that any given encoder/decoder system must have a bounded delay when used over a fixed-rate noiseless channel.* The encoder is not permitted to look into the entire infinite future before committing to an encoding for $\widehat{X}_t$. To allow the laws of large numbers to work, a finite, but potentially large, end-to-end delay is allowed between when the encoder observes X_t and when the decoder emits $\widehat{X}_t$. However, this delay must remain bounded and not grow with t.

For the stable cases $|\lambda| < 1$, standard block-coding arguments work since long blocks separated by an intervening block look relatively independent of each other and are in their stationary distributions. The ability to encode blocks in an independent way also tells us that Shannon's classical sense of ϵ-reliability also suffices for communicating the encoded bits across a noisy channel. The study of unstable cases $|\lambda| \geq 1$ is substantially more difficult since they are neither ergodic nor stationary and furthermore their variance grows unboundedly with time. As a result, Gray was able to prove only finite horizon results for such nonstationary processes and the general infinite-horizon unstable case has remained essentially open since Gray's 1970 paper [9]. As he put it:

> It should be emphasized that when the source is non-stationary, the above theorem is not as powerful as one would like. Specifically, it does not show that one can code a long sequence by breaking it up into smaller blocks of length n and use the same code to encode each block. The theorem is strictly a "one-shot" theorem unless the source is stationary, simply because the blocks $[(k-1)n, kn]$ do not have the same distribution for unequal k when the source is not stationary.

On the computational side, Hashimoto and Arimoto gave a parametric form for computing the $R(d)$ function for unstable auto-regressive Gaussian processes [10] and mean-square distortion. Toby Berger gave an explicit coding theorem for an important sub-case, the marginally unstable Wiener process with $\lambda = 1$, by introducing an ingenious parallel stream methodology. He noticed that although the Wiener process is nonstationary, it does have stationary and independent increments [11]. However, Berger's source-coding theorem said nothing about what is required from a noisy channel. In his own words:[12]

> It is worth stressing that we have proved only a source coding theorem for the Wiener process, not an information transmission theorem. If uncorrected channel errors were to occur, even in extremely rare instances, the

> user would eventually lose track of the Wiener process completely. It appears (although it has never been proved) that, even if a *noisy* feedback link were provided, it still would not be possible to achieve a finite [mean squared error] per letter as $t \to \infty$.

In an earlier conference work [13] and the first author's dissertation [14], we gave a variable rate coding theorem that showed the $R(d)$ bound is achievable in the infinite-horizon case if variable-rate codes are allowed. The question of whether or not fixed-rate and finite-delay codes could be made to work was left open, and is resolved here along with a full information transmission theorem.

B. Asymptotic equivalences and direct reductions

Beyond the technical issue of fixed or variable rate lies a deeper question regarding the nature of "information" in such processes. [15] contains an analysis of the traditional Kalman-Bucy filter in which certain entropic expressions are identified with the accumulation and dissipation of information within a filter. No explicit source or channel coding is involved, but the idea of different kinds of information flows is raised through the interpretation of certain mutual information quantities. In the stabilization problem of [5], it is hard to see if any qualitatively distinct kinds of information are present since to an external observer, the closed-loop process is stable.

Similarly, the variable-rate code given earlier in [13], [14] does not distinguish between kinds of information since the same high QoS requirements were imposed on all bits. However, it was clear that all the bits *do not require* the same treatment since there are examples in which access to an additional lower reliability medium can be used to improve end-to-end performance [16], [14]. The true nature of the information within the unstable process was left open and while exponentially unstable processes certainly appeared to be accumulating information, there was no way to make this interpretation precise and quantify the amount of accumulation.

In order to understand the nature of information, this paper builds upon the "asymptotic communication problem equivalence" perspective introduced at the end of [5]. This approach associates communication problems (e.g. communicating bits reliably at rate R or communicating iid Gaussian random variables to average distortion $\leq d$) with the set of channels that are good enough to solve that problem (e.g. noisy channels with capacity $C > R$). This parallels the "asymptotic computational problem equivalence" perspective in computational complexity theory [17] except that the critical resource shifts from computational operations to noisy channel uses. The heart of the approach is the use of "reductions" that show that a system made to solve one communication problem can be used as a black box to solve another communication problem. Two problems are asymptotically equivalent if they can be reduced to each other.

The equivalence perspective is closely related to the traditional source/channel separation theorems. The main difference is that traditional separation theorems give a privileged position to one communication problem — reliable bit-transport in the Shannon sense — and use reductions in only one direction: from the source to bits. The "converse" direction is usually proved using properties of mutual information. In [18], [19], we give a direct proof of the "converse" for classical problems by showing the existence of randomized codes that embed iid message bits into iid seeming source symbols at rate R. The embedding is done so that the bits can be recovered with high probability from distorted reconstructions as long as the average distortion on long blocks stays below the distortion-rate function $D(R)$. Similar results are obtained for the conditional distortion-rate function. This equivalence approach to separation theorems considers the privileged position of reliable bit-transport to be purely a pedagogical matter.

This paper uses the results from [18], [19] to extend the results of [5] from the control context to the estimation context. We demonstrate that the problem of communicating an unstable Markov process to within average distortion d is asymptotically equivalent to a pair of communication problems: classical reliable bit-transport at a rate $\approx R(d) - \log_2 |\lambda|$ and anytime-reliable bit-transport at a rate $\approx \log_2 |\lambda|$. This gives a precise interpretation to the nature of information flows in such processes.

C. Outline

Section II states the main results of this paper. A brief numerical example for the Gaussian case is given to illustrate the behavior of such unstable processes. The proofs follow in the subsequent sections.

Section III considers lossy source coding for unstable Markov processes with the driving disturbance W_t constrained to have bounded support. A fixed-rate code at a rate arbitrarily close to $R(d)$ is constructed by encoding process into two simultaneous fixed-rate message streams. The first stream has a bit-rate arbitrarily close to $\log_2|\lambda|$ and encodes what is needed from the past to understand the future. It captures the information that is accumulating within the unstable process. The other stream encodes those aspects of the past that are not relevant to the future and so captures the purely historical aspects of the unstable process in a way that meets the average distortion constraint. This second stream can be made to have a rate arbitrarily close to $R(d) - \log_2|\lambda|$.

Section IV then examines this historical information more carefully by looking at the process formally going backward in time. The $R(d)$ curve for the unstable process is shown to have a shape that is the stable historical part translated by $\log_2|\lambda|$ to account for the unstable accumulation of information.

Section V first reviews the fact that random codes exist achieving anytime reliability over noisy channels even without any feedback. Then, for η-difference distortion measures, an anytime reliability $> \eta\log_2|\lambda|$ is shown to be sufficient to encode the first bitstream of the code of Section III across a noisy channel. The second bitstream is shown to only require classical Shannon ϵ-reliability. This completes the reduction of the lossy-estimation problem to a two-tiered reliable bit-transportation problem and resolves the conjecture posed by Berger regarding an information transmission theorem for unstable processes.

Section VI tackles the other direction. The problem of anytime-reliable bit-transport is directly reduced to the problem of lossy-estimation for a decimated version of the unstable process. This is done using the ideas in [5], reinterpreted as information-embedding and shows that the higher QoS requirements for the first stream are unavoidable for these processes. A second stream of messages is then embedded into the historical segments of the unstable process and this stream is recovered in the classical Shannon ϵ-reliable sense. *Exponentially unstable Markov processes are thus the first nontrivial examples of stochastic processes that naturally generate two qualitatively distinct kinds of information.*

In Section VII, the results are then extended to cover the Gauss-Markov case with the usual squared-error distortion. Although the proofs are given in terms of Gaussian processes and squared error, the results actually generalize to any η-distortion as well as driving noise distributions W that have at least an exponentially decaying tail.

This paper focuses throughout on scalar Markov processes for clarity. It is possible to extend all the arguments to cover the general autoregressive moving average (ARMA) case. The techniques used to cover the ARMA case are discussed in the control context in [6] where a state-space formulation is used. A brief discussion of how to apply those techniques is present here in Section VIII.

II. Main results

A. Performance bound in the limit of large delays

To define $R(d)$ for unstable Markov processes, the infinite-horizon problem is viewed as the limit of a sequence of finite-horizon problems:

Definition 2.1: Given the scalar Markov source given by (1), the *finite n-horizon* version of the source is defined to be the random variables $X_0^{n-1} = (X_0, X_1, \ldots, X_{n-1})$.

Definition 2.2: The *$\eta-$distortion* measure is $\rho(X_i, \widehat{X}_i) = |X_i - \widehat{X}|^\eta$. It is an additive distortion measure when applied to blocks.

The standard information-theoretic rate-distortion function for the finite-horizon problem using η-difference distortion is:

$$R_n^X(d) = \inf_{\{\mathcal{P}(Y_0^{n-1}|X_0^{n-1}):\frac{1}{n}\sum_{i=0}^{n-1}E[|X_i-Y_i|^\eta]\le d\}} \frac{1}{n} I(X_0^{n-1}; Y_0^{n-1}) \tag{2}$$

We can consider the block X_1^n as a single vector-valued random variable $\vec{X}$. The $R_n^X(d)$ defined by (2) is related to $R_1^{\vec{X}}(d)$ by $R_n^X(d) = \frac{1}{n} R_1^{\vec{X}}(nd)$ with the distortion measure on $\vec{X}$ given by $\rho(\vec{X}, \widehat{\vec{X}}) = \sum_{i=0}^{n-1} |X_i - \widehat{X}|^\eta$.

The infinite-horizon case is then defined as a limit:

$$R_\infty^X(d) = \liminf_{n\to\infty} R_n^X(d) \tag{3}$$

The distortion-rate function $D_\infty^X(R)$ is also defined in the same manner, except that the mutual-information is fixed and the distortion is what is infimized.

B. The stable counterpart to the unstable process

It is insightful to consider what the stable counterpart to this unstable process would be. There is a natural choice, just formally turn the recurrence relationship around and flip the order of time. This gives the "backwards in time process" governed by the recursion

$$\overleftarrow{X}_t = \lambda^{-1} \overleftarrow{X}_{t+1} - \lambda^{-1} W_t. \tag{4}$$

This is purely a formal reversal. In place of an initial condition X_0, it is natural to consider a $\overleftarrow{X}_n$ for some time n and then consider time going backwards from there. Since $|\lambda^{-1}| < 1$, this is a stable Markov process and falls under the classical theorems of [9].

C. Encoders and decoders

For notational convenience, time is synchronized between the source and the channel. Thus both delay and rate can be measured against either source symbols or channel uses.

Definition 2.3: A *discrete time channel* is a probabilistic system with an input. At every time step t, it takes an input $a_t \in \mathcal{A}$ and produces an output $c_t \in \mathcal{C}$ with probability $p(C_t | a_1^t, c_1^{t-1})$ where the notation a_1^t is shorthand for the sequence $(a_1, a_2, \ldots, a_t)$. In general, the current channel output is allowed to depend on all inputs so far as well as on past outputs.

The channel is *memoryless* if conditioned on a_t, the random variable C_t is independent of any other random variable in the system that occurs at time t or earlier. So all that needs to be specified is $p_t(C_t|a_t)$. The channel is memoryless and stationary if $p_t(C_t|a_t) = p(C_t|a_t)$ for all times t.

Definition 2.4: A *rate R source-encoder $\mathcal{E}_s$* is a sequence of maps $\{\mathcal{E}_{s,i}\}$. The range of each map is a single bit $b_i \in \{0, 1\}$ if it is a pure source encoder and is from the channel input alphabet $\mathcal{A}$ if it is a joint source-channel encoder. The i-th map takes as input the available source symbols $X_1^{\lfloor \frac{i}{R} \rfloor}$.

Similarly, a *rate R channel-encoder $\mathcal{E}_c$ without feedback* is a sequence of maps $\{\mathcal{E}_{c,t}\}$. The range of each map is the channel input alphabet $\mathcal{A}$. The t-th map takes as input the available message bits $B_1^{\lfloor Rt \rfloor}$.

Randomized encoders also have access to random variables denoting the common randomness available in the system. This common randomness is independent of the source and channel.

Definition 2.5: A *delay ϕ rate R source-decoder* is a sequence of maps $\{\mathcal{D}_{s,t}\}$. The range of each map is just an estimate $\widehat{X}_t$ for the t-th source symbol. For pure source decoders, the t-th map takes as input the available message bits $B_1^{\lfloor (t+\phi)R \rfloor}$. For joint source-channel decoders, it takes as input the available channel outputs $C_1^{t+\phi}$. Either way, it can see ϕ time units beyond the time when the desired source symbol first had a chance to impact its inputs.

Similarly, a *delay ϕ rate R channel-decoder* is a sequence of maps $\{\mathcal{D}_{c,i}\}$. The range of each map is just an estimate $\widehat{B}_i$ for the i-th bit taken from $\{0, 1\}$. The i-th map takes as input the available channel outputs $C_1^{\lceil \frac{i}{R} \rceil + \phi}$ which means that it can see ϕ time units beyond the time when the desired message bit first had a chance to impact the channel inputs.

Randomized decoders also have access to the random variables denoting common randomness.

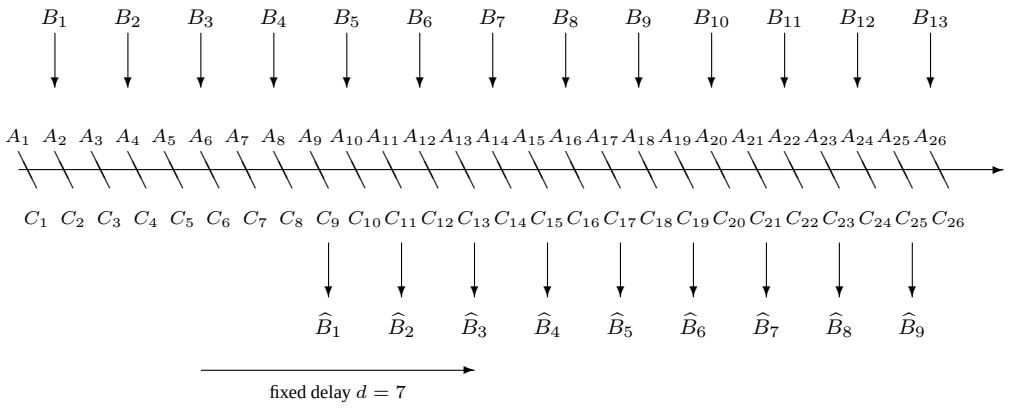

Fig. 2. The timeline in a rate $\frac{1}{2}$ delay 7 channel code. Both the encoder and decoder must be causal so A_i and $\widehat{B}_i$ are functions only of quantities to the left of them on the timeline. If noiseless feedback is available, the A_i can also have an explicit functional dependence on the C_1^{i-1} that lie to the left on the timeline.

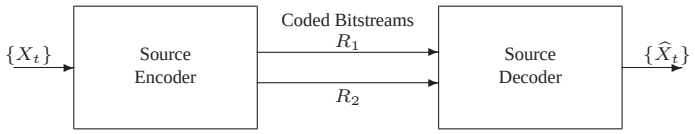

Fig. 3. The source-coding problem of translating the source into two simultaneous bitstreams of fixed rates R_1 and R_2. End-to-end delay is permitted but must remain bounded for all time. The goal is to get $R_1 \approx \log_2 |\lambda|$ and $R_2 \approx R(d) - \log_2 |\lambda|$.

The timeline is illustrated in Fig. 2 for channel coding and a similar timeline holds for either pure source coding or joint source-channel coding.

For a specific channel, the maximum rate achievable for a given sense of reliable communication is called the associated capacity. Shannon's classical ϵ-reliability requires that for a suitably large end-to-end delay ϕ, the probability of error on each bit is below a specified ϵ.

Definition 2.6: A rate R *anytime communication system* over a noisy channel is a single channel encoder $\mathcal{E}_c$ and decoder $\mathcal{D}_c^{\phi}$ *family* for all end-to-end delays ϕ.

A rate R communication system achieves *anytime reliability* α if there exists a constant K such that:

$$\mathcal{P}(\widehat{B}_1^i(t) \neq B_1^i) \leq K 2^{-\alpha(t-\frac{i}{R})} \tag{5}$$

holds for every i. The probability is taken over the channel noise, the message bits B, and all of the common randomness available in the system. If (5) holds for every possible realization of the message bits B, then we say that the system achieves *uniform anytime reliability* α.

Communication systems that achieve *anytime reliability* are called *anytime codes* and similarly for *uniform anytime codes*.

The important thing to understand about anytime reliability is that it is not considered to be a proxy used to study encoder/decoder complexity as traditional reliability functions often are [8]. Instead, the anytime reliability parameter α indexes a sense of reliable transmission for a bitstream in which the probability of bit error tends to zero exponentially as time goes on.

D. Main results

The first result concerns the source coding problem illustrated in Fig. 3 for unstable Markov processes with bounded-support driving noise.

Theorem 2.1: Assume both the source encoder and source decoder can be randomized. Given the unstable ($|\lambda| > 1$) scalar Markov process from (1) driven by independent noise $\{W_t\}_{t\geq 0}$ with bounded

support, it is possible to encode the process to average fidelity $E[|X_i - \widehat{X}_i|^\eta]$ arbitrarily close to d using two fixed-rate bitstreams and a suitably high end-to-end delay ϕ.

The first stream (called the *checkpoint stream*) can be made to have rate R_1 arbitrarily close to $\log_2|\lambda|$ while the second (called the *historical stream*) can have rate R_2 arbitrarily close to $R^X_\infty(d) - \log_2|\lambda|$.
Proof: See Section III.

In a very real sense, the first stream in Theorem 2.1 represents an initial description of the process to some fidelity, while the second represents a refinement of the description [20]. These two descriptions turn out to be qualitatively different when it comes to communicating them across a noisy channel.

Theorem 2.2: Suppose that a communication system provides uniform anytime reliability $\alpha > \eta \log_2|\lambda|$ for the checkpoint message stream at bit-rate R_1. Then given sufficient end-to-end delay ϕ, it is possible to reconstruct the checkpoints to arbitrarily high fidelity in the η-distortion sense.
Proof: See Section V-B.

Theorem 2.3: Suppose that a communication system can reliably communicate message bits meeting any bit-error probability ϵ given a long enough delay. Then, that communication system can be used to reliably communicate the historical information message stream generated by the fixed-rate source code of Theorem 2.1 in that the expected end-to-end distortion can be made arbitrarily close to the distortion achieved by the code over a noiseless channel.
Proof: See Section V-C.

The Gauss-Markov case with mean squared error is covered by corollaries:

Corollary 2.1: Assume both the encoder and decoder are randomized and the finite end-to-end delay ϕ can be chosen to be arbitrarily large. Given an unstable ($|\lambda| > 1$) scalar Markov process (1) driven by iid Gaussian noise $\{W_t\}_{t\geq 0}$ with zero mean and variance σ^2, it is possible to encode the process to average fidelity $E[|X_i - \widehat{X}_i|^2]$ arbitrarily close to d using two fixed-rate bitstreams.

The checkpoint stream can be made to have rate R_1 arbitrarily close to $\log_2|\lambda|$ while the historical stream can have rate R_2 arbitrarily close to $R^X_\infty(d) - \log_2|\lambda|$.
Proof: See Section VII-A.

Corollary 2.2: Suppose that a communication system provides us with the ability to carry two message streams. One at rate $R_1 > \log_2|\lambda|$ with uniform anytime reliability $\alpha > 2\log_2|\lambda|$, and another with classical Shannon reliability at rate $R_2 > R^X_\infty(d) - \log_2|\lambda|$ where $R^X_\infty(d)$ is the rate-distortion function for an unstable Gauss-Markov process with unstable gain $|\lambda| \geq 1$ and squared-error distortion.

Then it is possible to successfully transport the two-stream code of Corollary 2.1 using this communication system by picking a sufficiently large end-to-end delay ϕ. The mean squared error of the resulting system will be as close to d as desired.
Proof: See Section VII-B.

Theorems 2.2 and 2.3 together with the source code of Theorem 2.1 combine to establish a reduction of the d-lossy joint source/channel coding problem to the problem of communicating message bits at rate $R(d)$ over the same channel, wherein a substream of message bits at rate $\approx \log_2|\lambda|$ is given an anytime reliability of at least $\eta\log_2|\lambda|$. This reduction is in the sense of Section VII of [5]: any channel that is good enough to solve the second pair of problems is good enough to solve the first problem.

The asymptotic relationship between the forward and backward rate-distortion functions is captured in the following theorem.

Theorem 2.4: Let X be the unstable Markov process of (1) with $|\lambda| > 1$ and let the stable backwards-in-time version from (4) be denoted $\overleftarrow{X}$. Assume that the iid driving noise W has a Riemann-integrable density f_W and there exists a constant K so that $E[|\sum_{i=1}^t \lambda^{-i} W_i|^\eta] \leq K$ for all $t \geq 1$. Furthermore for the purpose of calculating the rate-distortion functions below, assume that for the backwards-in-time version is initialized with $\overleftarrow{X}_n = 0$. Let Q_Δ be the uniform quantizer that maps its input to the nearest neighbor of the form $k\Delta$ for integer k.

$$R_{\infty}^{\overleftarrow{X}}(d) =_{(a)} \lim_{\Delta\to 0}\lim_{n\to\infty} R_n^{\overleftarrow{X}|Q_\Delta(\overleftarrow{X}_0)}(d) =_{(b)} \lim_{n\to\infty} R_n^{\overleftarrow{X}|\overleftarrow{X}_0}(d) =_{(c)} R_{\infty}^{X}(d) - \log_2|\lambda|. \tag{6}$$

or expressed in terms of distortion-rate functions for $R > \log_2|\lambda|$:

$$D_{\infty}^{X}(R) = D_{\infty}^{\overleftarrow{X}}(R - \log_2|\lambda|).$$

This implies that the process generally undergoes a phase transition from infinite to bounded average distortion at the critical rate $\log_2|\lambda|$.
Proof: See Section IV-B.

Notice that there are no explicitly infinite distortions in the original setup of the problem. Consequently, the appearance of infinite distortions is interesting as is the abrupt transition from infinite to finite distortions around the critical rate of $\log_2|\lambda|$. This abrupt transition gives a further indication that there is something fundamentally nonclassical about the rate $\log_2|\lambda|$ information inside the process.

To make this precise, a converse is needed. Classical rate-distortion results only point out that the mutual information across the communication system must be at least $R(d)$ on average. However, as [21] points out, having enough mutual information is not enough to guarantee a reliable-transport capacity since the situation here is not stationary and ergodic. The following theorem gives the converse, but adds an intuitively required additional condition that the probability of excess average distortion over any long enough segment can be made as small as desired.

Theorem 2.5: Consider the unstable process given by (1) with the iid driving noise W having a Riemann-integrable density f_W satisfying the conditions of Theorem 2.4.

Suppose there exists a family (indexed by window size n) of joint source-channel codes $(\mathcal{E}_s, \mathcal{D}_s)$) so that the n-th member of the family has reconstructions that satisfy

$$E[|X_{kn} - \widehat{X}_{kn}|^\eta] \le d \tag{7}$$

for every positive integer k. Furthermore, assume the family collectively also satisfies

$$\lim_{n\to\infty}\sup_{\tau\ge 0}\mathcal{P}(\frac{1}{n}\sum_{i=\tau}^{\tau+n-1}|\widehat{X}_i - X_i|^\eta > d) = 0 \tag{8}$$

so that the probability of excess distortion can be made arbitrarily small on long enough blocks.

Then for any $R_1 < \log_2|\lambda|, \alpha < \eta\log_2|\lambda|, R_2 < R_{\infty}^{X}(d) - \log_2|\lambda|, P_e > 0$, the channel must support the simultaneous carrying of a bit-rate R_1 priority message stream with anytime reliability α along with a second message stream of bit-rate R_2 with a probability of bit error $\le P_e$ for some end-to-end delay ϕ.
Proof: See Section VI-A.

Note that a Gaussian disturbance W is covered by Theorems 2.4 and 2.5, even if the difference distortion measure is not mean squared-error.

E. An example and comparison to the sequential rate distortion problem

In the case of Gauss-Markov processes with squared-error distortion, Hashimoto and Arimoto in [10] give an explicit way of calculating $R(d)$. Tatikonda in [22], [23] gives a similar explicit lower bound to the rate required when the reconstruction $\widehat{X}_t$ is forced to be causal in that it can only depend on X_j observations for $j \le t$.

Assuming unit variance for the driving noise W and $\lambda > 1$, Hashimoto's formula is parametric in terms of the water-filling parameter κ and for the Gauss-Markov case considered here simplifies to:

$$\begin{aligned} D(\kappa) &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\min\left[\kappa, \frac{1}{1-2\lambda\cos(\omega)+\lambda^2}\right]d\omega, \\ R(\kappa) &= \log_2\lambda + \frac{1}{2\pi}\int_{-\pi}^{\pi}\max\left[0, \frac{1}{2}\log_2\frac{1}{\kappa(1-2\lambda\cos(\omega)+\lambda^2)}\right]d\omega. \end{aligned} \tag{9}$$

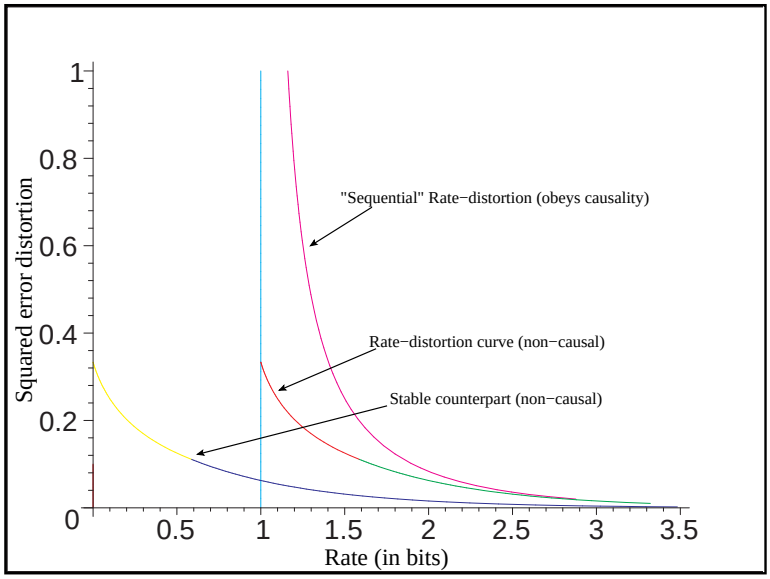

Fig. 4. The distortion-rate curves for an unstable Gauss-Markov process with $\lambda = 2$ and its stable backwards-version. The stable and unstable $D(R)$ curves are related by a simple translation by 1 bit per symbol.

The rate-distortion function for the stable counterpart given in (4) has a water-filling solution that is identical to 9, except without the $\log_2 \lambda$ term in the $R(\kappa)$! Thus, in the Gaussian case with squared error distortion direct calculation verifies the claim

$$R_\infty^X(d) = \log_2 \lambda + R_\infty^{\overleftarrow{X}}(d)$$

from Theorem 2.4.

For the unstable process, Tatikonda's formula for causal reconstructions is given by

$$R_{\mathsf{seq}}(d) = \frac{1}{2}\log_2(\lambda^2 + \frac{1}{d}). \tag{10}$$

Fig. 4 shows the distortion-rate frontier for both the original unstable process and backwards stable process. It is easy to see that the forward and backward process curves are translations of each other. In addition, the sequential rate-distortion curve for the forward process is qualitatively distinct. $D_{\mathsf{seq}}(R)$ goes to infinity as $R \downarrow \log_2 \lambda$ while $D(R)$ approaches a finite limit.

The results in this paper show that the lower curve for the regular distortion-rate frontier can be approached arbitrarily closely by increasing the acceptable (finite) end-to-end delay. This suggests that it takes some time for the randomness entering the unstable process through W to sort itself into the two categories of fundamental accumulation and transient history. The difference in the resulting distortion is not that significant at high rates, but becomes unboundedly large as the rate approaches $\log_2 \lambda$. It is open whether similar information-embedding theorems similar to Theorem 2.5 exist that give an operational meaning to the gap between $R_{seq}(d)$ and $R(d)$. If a communication system can be used to satisfy distortion d in a causal way, does that mean the underlying communication resources also must be able to support messages at this higher rate $R_{seq}(d)$?

III. Two stream source encoding: approaching $R(d)$

This section proves Theorem 2.1.

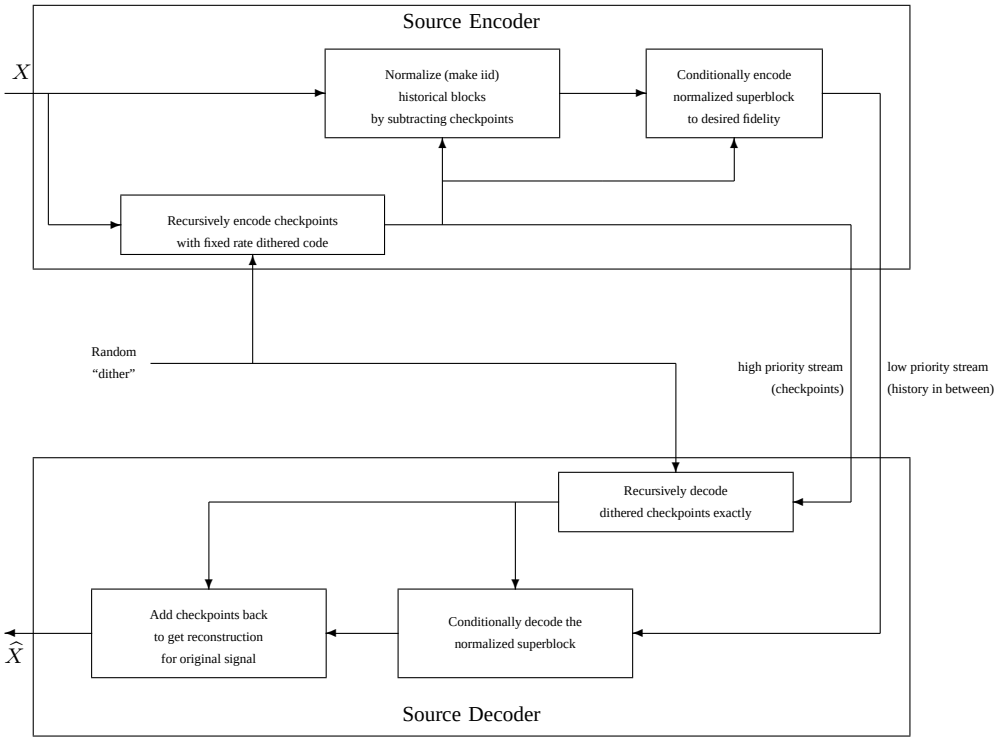

Fig. 5. A flowchart showing how to do fixed-rate source coding for Markov sources using two streams and how the streams are decoded.

A. Proof strategy

The code for proving Theorem 2.1 is illustrated in Fig. 5. Without loss of generality, assume $\lambda = |\lambda| > 1$ to avoid the notational complication of keeping track of the sign.

- Look at time in blocks of size n and encode the values of endpoints (X_{kn-1}, X_{kn}) recursively to very high precision using rate $n(\log_2 \lambda + \epsilon_1)$ per pair. Each block $X_{kn}, X_{kn+1}, \ldots, X_{(k+1)n-1}$ will have encoded checkpoints $(\check{X}_{kn}, \check{X}_{kn+n-1})$ at both ends.
- Use the encoded checkpoints $\{\check{X}_{kn}\}$ at the start of the blocks to transform the process segments in between (the history) so that they look like an iid sequence of finite horizon problems $\vec{X}$.
- Use the checkpoints $\{\check{X}_{kn+n-1}\}$ at the end of the blocks as side-information to encode the history to fidelity d at a rate of $n(R^X_\infty(d) - \log_2 \lambda + \epsilon_2 + o(1))$ per block.
- "Stationarize" the encoding by choosing a random starting offset so that no times t are *a priori* more vulnerable to distortion.

The source decoding proceeds in the same manner and first recovers the checkpoints, and then uses them as known side-information to decode the history. The two are then recombined to give a reconstruction of the original source to the desired fidelity.

The above strategy follows the spirit of Berger's encoding[11]. In Berger's code for the Wiener process, the first stream's rate is negligible relative to that of the second stream. In our case, the first stream's rate is significant and cannot be averaged away by using large blocks n.

The detailed constructions and proof for this theorem are in the next few subsections, with some technical aspects relegated to the appendices.

B. Recursively encoding checkpoints

This section relies on the assumption of bounded support for the driving noise $|W_t| \leq \frac{\Omega}{2}$, but does not care about any other property of the $\{W_t\}_{t\geq 0}$ like independence or stationarity. The details of the distortion measure are also not important for this section.

Proposition 3.1: Given the unstable ($\lambda > 1$) scalar Markov process of (1) driven by noise $\{W_t\}_{t\geq 0}$ with bounded support, and any $\Delta > 0$, it is possible to causally and recursively encode checkpoints spaced by n so that $|\check{X}_{kn} - X_{kn}| \leq \frac{\Delta}{2}$. For any $R_1 > \log_2 \lambda$, this can be done with rate nR_1 bits per checkpoint by choosing n large enough. Furthermore, if an iid sequence of independent pairs of continuous uniform random variables $\{\Theta_i, \Theta_i'\}_{i\geq 0}$ is available to both the encoder and decoder for dithering, the errors $(\check{X}_{kn-1} - X_{kn-1}, \check{X}_{kn} - X_{kn})$ can be made an iid sequence of pairs of independent uniform random variables on $[-\frac{\Delta}{2}, +\frac{\Delta}{2}]$.

Proof: First, consider the initial condition at X_0. It can be quantized to be within an interval of size Δ by using $\log_2\lceil\frac{\Omega_0}{\Delta}\rceil$ bits.

With a block length of n, the successive endpoints are related by:

$$X_{(k+1)n} = \lambda^n X_{kn} + [\lambda^{n-1}\sum_{i=0}^{n-1}\lambda^{-i}W_{kn+i}] \tag{11}$$

The second term $[\cdots]$ on the left of (11) can be denoted $\widetilde{W}_k$ and bounded using

$$|\widetilde{W}_k| = |\lambda^{n-1}\sum_{i=0}^{n-1}\lambda^{-i}W_{kn+i}| \leq |\lambda^{n-1}|\sum_{i=0}^{n-1}\lambda^{-i}\frac{\Omega}{2} < \lambda^n\frac{\Omega}{2(\lambda-1)}. \tag{12}$$

Proceed by induction. Assume that $\check{X}_{kn}$ satisfies $|X_{kn} - \check{X}_{kn}| \leq \frac{\Delta}{2}$. This clearly holds for $k = 0$. Without any further information, it is known that $X_{(k+1)n}$ must lie within an interval of size $\lambda^n\Delta + \lambda^n\frac{\Omega}{\lambda-1}$. By using nR_1' bits (where R_1' is chosen to guarantee an integer nR_1') to encode where the true value lies, the uncertainty is cut by a factor of $2^{nR_1'}$. To have the resulting interval of size Δ or smaller, we must have:

$$\Delta \geq 2^{-nR_1'}\lambda^n(\Delta + \frac{\Omega}{(\lambda-1)}).$$

Dividing through by $\Delta 2^{-nR_1'}\lambda^n$ and taking logarithms gives

$$n(R_1' - \log_2\lambda) \geq \log_2(1 + \frac{\Omega}{\Delta(\lambda-1)}).$$

Encoding $\check{X}_{kn-1}$ given $\check{X}_{kn}$ requires very little additional rate since $|X_{kn-1} - \lambda^{-1}\check{X}_{kn}| < \Omega + \Delta$ and so $\log_2\lceil\frac{\Omega}{\Delta}+1\rceil < \log_2(2+\frac{\Omega}{\Delta})$ additional bits are good enough to encode both checkpoints. Putting everything together in terms of the original R_1 gives

$$R_1 \geq \max\left(\log_2\lambda + \frac{\log_2(1+\frac{\Omega}{\Delta(\lambda-1)}) + \log_2(2+\frac{\Omega}{\Delta})}{n}, \frac{\log_2\lceil\frac{\Omega_0}{\Delta}\rceil}{n}\right). \tag{13}$$

It is clear from (13) that no matter how small a Δ we choose, by picking an n large enough the rate R_1 can get as close to $\log_2\lambda$ as desired. In particular, picking $n = K(\log_2\frac{1}{\Delta})^2$ works with large K and small Δ.

To get the uniform nature of the final error $\check{X}_{kn} - X_{kn}$, subtractive dithering can be used [24]. This is accomplished by adding a small iid random variable Θ_k, uniform on $[-\frac{\Delta}{2}, +\frac{\Delta}{2}]$, to the X_{kn}, and only then quantizing $(X_{kn} + \Theta_k)$ to resolution Δ. At the decoder, Θ_k is subtracted from the result to get $\check{X}_{kn}$. Similarly for $\check{X}_{kn-1}$. This results in the checkpoint error sequence $(X_{kn-1} - \check{X}_{kn-1}, X_{kn} - \check{X}_{kn})$ being iid uniform pairs over $[-\frac{\Delta}{2}, +\frac{\Delta}{2}]$. These pairs are also independent of all the W_t and initial condition X_0. □

In what follows, we always assume that Δ is chosen to be of high fidelity relative to the target distortion d (e.g. For squared-error distortion, this means that $\Delta^2 \ll d$.) as well as small relative to the the initial condition so $\Delta \ll \Omega_0$.

C. Transforming and encoding the history

Having dealt with the endpoints, focus attention on the historical information between them. Here the bounded support assumption is not needed for the $\{W_t\}$, but the iid assumption is important. First, the encoded checkpoints are used to transform the historical information so that each historical segment looks iid. Then, it is shown that these segments can be encoded to the appropriate fidelity and rate when the decoder has access to the encoded checkpoints as side information.

1) Forward transformation: The simplest transformation is to effectively restart the process at every checkpoint and view time going forward. This can be considered normalizing each of the historical segments $X_{kn}^{(k+1)n-1}$ to $(\widetilde{X}_{(k,i)}, 0 \le i \le n-1)$ for $k = 0, 1, 2, \ldots$.

$$\widetilde{X}_{(k,i)} = X_{kn+i} - \lambda^i \check{X}_{kn} \tag{14}$$

For each k, the block $\widetilde{X}_k = \{\widetilde{X}_{(k,i)}\}_{0 \le i \le n-1}$ satisfies $\widetilde{X}_{(k,i+1)} = \lambda \widetilde{X}_{(k,i)} + W_{(k,i)}$. By dithered quantization, the initial condition ($i = 0$) of each block is a uniform random variable of support Δ that is independent of all the other random variables in the system. The initial conditions are iid across the different k. Thus, except for the initial condition, the blocks $\widetilde{X}_k$ are identically distributed to the finite horizon versions of the problem.

Since $\Delta < \Omega_0$, each $\widetilde{X}_k$ block starts with a tighter initial condition than the original X process did. Since the initial condition is uniform, this can be viewed as a genie-aided version of the original problem where a genie reveals a few bits of information about the initial condition. Since the initial condition enters the process dynamics in a linear way and the distortion measure ρ depends only on the difference, this implies that the new process with the smaller initial condition requires no more bits per symbol to achieve a distortion d than did the original process. Thus:

$$R_n^X(d) - \frac{1 + \log_2 \frac{\Omega_0}{\Delta}}{n} \le R_n^{\widetilde{X}}(d) \le R_n^X(d)$$

for all n and d. So in the limit of large n

$$R_\infty^{\widetilde{X}}(d) = R_\infty^X(d). \tag{15}$$

In simple terms, the normalized history behaves like the finite horizon version of the problem when n is large.

2) Conditional encoding: The idea is to encode the normalized history between two checkpoints conditioned on the ending checkpoint. The decoder has access to the exact values of these checkpoints through the first bitstream.

For a given k, shift the encoded ending checkpoint $\check{X}_{(k+1)n-1}$ to

$$Z_k^q = \check{X}_{(k+1)n-1} - \lambda^{n-1} \check{X}_{kn}. \tag{16}$$

Z_k^q is clearly available at both the encoder and the decoder since it only depends on the encoded checkpoints. Furthermore, it is clear that

$$\widetilde{X}_{(k,n-1)} - Z_k^q = (X_{(k+1)n-1} - \lambda^{n-1}\check{X}_{kn}) - (\check{X}_{(k+1)n-1} - \lambda^{n-1}\check{X}_{kn}) = X_{(k+1)n-1} - \check{X}_{(k+1)n-1}$$

which is a uniform random variable on $[-\frac{\Delta}{2}, +\frac{\Delta}{2}]$. Thus Z_k^q is just a dithered quantization to Δ precision of the endpoint $\widetilde{X}_{(k,n-1)}$.

Define the conditional rate-distortion function $R_\infty^{X|Z^q,\Theta}(d)$ for the limit of long historical blocks $\widetilde{X}_{k,0}^{n-1}$ conditioned on their quantized endpoint as

$$R_\infty^{X|Z^q,\Theta}(d) = \liminf_{n\to\infty} \frac{1}{n} \inf_{\{\mathcal{P}(Y_0^{n-1}|\widetilde{X}_0^{n-1},Z^q,\Theta):\frac{1}{n}\sum_{i=0}^{n-1} E\left[|\widetilde{X}_i - Y_i|^\eta\right] \le d\}} \frac{1}{n} I(\widetilde{X}_0^{n-1}; Y_0^{n-1} | Z^q, \Theta). \tag{17}$$

Proposition 3.2: Given an unstable ($\lambda > 1$) scalar Markov process $\{\widetilde{X}_{k,t}\}$ obeying (1) and whose driving noise satisfies $E[|\sum_{i=1}^{t} \lambda^{-i} W_i|^{\eta}] \leq K$ for all $t \geq 1$ for some constant K, together with its encoded endpoint Z_k^q obtained by Θ-dithered quantization to within a uniform random variable with small support Δ, the limiting conditional rate-distortion function

$$R_{\infty}^{X|Z^q,\Theta}(d) = R_{\infty}^{X}(d) - \log_2 \lambda. \tag{18}$$

Proof: See Appendix C.

The case of driving noise with bounded support clearly satisfies the conditions of this proposition since geometric sums converge. The conditional rate-distortion function in Proposition 3.2 has a corresponding coding theorem:

Proposition 3.3: Given an unstable ($\lambda > 1$) scalar Markov process $\{X_t\}$ given by (1) together with its n-spaced pairs of encoded checkpoints $\{\check{X}\}$ obtained by dithered quantization to within iid uniform random variables with small support Δ, for every $\epsilon_4 > 0$ there exists an M large enough so that a conditional source-code exists that maps a length M superblock of the historical information $\{\widetilde{X}_k\}_{0 \leq k < M}$ into a superblock $\{T_k\}_{0 \leq k < M}$ satisfying

$$\frac{1}{M} \sum_{k=0}^{M-1} \frac{1}{n} \sum_{j=1}^{n} E[\rho(\widetilde{X}_{(k,j)}, T_{(k,j)})] \leq d + \epsilon_4. \tag{19}$$

By choosing n large enough, the rate of the superblock code can be made as close as desired to $R_{\infty}^{X}(d) - \log_2 \lambda$ if the decoder is also assumed to have access to the encoded checkpoints $\check{X}_{kn}$.

Proof: M of the $\widetilde{X}_k$ blocks are encoded together using conditioning on the encoded checkpoints at the end of each block. The pair $(\widetilde{X}_k, Z_k^q)$ have a joint distribution, but are iid across k by the independence properties of the subtractive dither and the driving noise $W_{(k,i)}$. Furthermore, the $\widetilde{X}_{(k,i)}$ are bounded and as a result, the all zero reconstruction results in a bounded distortion on the $\widetilde{X}$ vector that depends on n. Even without the bounded support assumption, Theorem 2.4 reveals that there is a reconstruction based on the Z_k^q alone that has bounded average distortion where the bound does not even depend on n.

Since the side information Z_k^q is available at both encoder and decoder, the classical conditional rate-distortion coding theorems of [25] tell us that there exists a block-length $M(n)$ so that codes exist satisfying (19). The rate can be made arbitrarily close to $R_n^{X|Z^q,\Theta}(d)$. By letting n get large, Proposition 3.2 reveals that this rate can be made as close as desired to $R_{\infty}^{X}(d) - \log_2 \lambda$. □

D. Putting history together with checkpoints

The next step is to show how the decoder can combine the two streams to get the desired rate/distortion performance.

The rate side is immediately obvious since there is $\log_2 \lambda$ from Proposition 3.1 and $R_{\infty}^{X}(d) - \log_2 \lambda$ from Proposition 3.3. The sum is as close to $R_{\infty}^{X}(d)$ as desired. On the distortion side, the decoder runs (14) in reverse to get reconstructions. Suppose that $T_{(k,i)}$ are the encoded transformed source symbols from the code in Proposition 3.3. Then $\widehat{X}_{kn+i} = T_{(k,i)} + \lambda^i \check{X}_{kn}$ and so $X_{kn+i} - \widehat{X}_{kn+i} = \widetilde{X}_{(k,i)} - T_{(k,i)}$. Since the differences are the same, so is the distortion.

E. "Stationarizing" the code

The underlying X_t process is non-stationary so there is no hope to make the encoding truly stationary. However, as it stands, only the average distortion across each of the Mn length superblocks is close to d in expectation giving the resulting code a potentially "cyclostationary" character. Nothing guarantees that source symbols at every time will have the same level of expected fidelity. To fix this, a standard trick can be applied by making the encoding have two phases:

- An initialization phase that lasts for a random T time-steps. T is a random integer chosen uniformly from $0, 1, \ldots Mn-1$ based on common randomness available to the encoder and decoder. During the first phase, all source symbols are encoded to fidelity Δ recursively using the code of Proposition 3.1 with $n = 1$.
- A main phase that applies the two-part code described above but starts at time $T + 1$.

The extra rate required in the first phase is negligible on average since it is a one-time cost. This takes a finite amount of time to drain out through the rate R_1 message stream. This time can be considered an additional delay that must be suffered for everything in the second phase. Thus it adds to the delay of n required by the causal recursive code for the checkpoints. The rest of the end-to-end delay is determined by the total length Mn of the superblock chosen inside Proposition 3.3.

Let d_i be such that the original super-block code gives expected distortion d_i at position i ranging from 0 to $Mn - 1$. It is known from Proposition 3.3 that $\frac{1}{Mn}\sum_{i=0}^{Mn-1} d_i \leq d + \epsilon_4$. Because the first phase is guaranteed to be high fidelity and all other time positions are randomly and uniformly assigned positions within the superblock of size Mn, the expected distortion $E[|X_i - \widehat{X}_i|^\eta] \leq d + \epsilon_4$ for every bit position i.

The code actually does better than that since the probability of excess average distortion over a long block is also guaranteed to go to zero. This property is inherited from the repeated use of independent conditional rate-distortion codes in the second stream [25].

This completes the proof of Theorem 2.1.

IV. Time-reversal and the essential phase transition

It is interesting to note that the distortion of the code in the previous section turns out to be entirely based on the conditional rate-distortion performance for the historical segments. The checkpoints merely contribute a $\log_2 \lambda$ term in the rate.

The nature of historical information in the unstable Markov process described by (1) can be explored more fully by transforming the historical blocks going locally backward in time. The informational distinction between the process going forward and the purely historical information parallels the concepts of information production and dissipation explored in the context of the Kalman Filter [15].

First, the original problem is formally decomposed into forward and backward parts. Then, Theorem 2.4 is proved.

A. Endpoints and history

It is useful to think of the original problem as being broken down into two analog sub-problems:

1) The n-endpoint problem: This is the communication of the process $\{X_{kn}\}$ where each sample arrives every n time steps and the samples are related to each other through (11) with $\widetilde{W}_k$ being iid and having the same distribution as $\lambda^{n-1}\sum_{i=0}^{n-1}\lambda^{-i}W_i$.

This process must be communicated so that $E[|X_{kn} - \widehat{X}_{kn}|^\eta] \leq K$ for some performance K. This is essentially a decimated version of the original problem.

2) The conditional history problem: The stable $\overleftarrow{X}$ process defined in (4) can be viewed in blocks of length n. The conditional history problem is thus the problem of communicating an iid sequence of n-vectors $\vec{X}_k^- = (\overleftarrow{X}_{k,1}, \ldots, \overleftarrow{X}_{k,n-1})$ conditioned on iid Z_k that are known perfectly at the encoder and decoder. The joint distribution of $\vec{X}^-, Z$ are given by:

$$\begin{aligned} Z &= -\sum_{i=0}^{n-1}\lambda^{-i}W_i \\ \overleftarrow{X}_{n-1} &= -\lambda^{-1}W_{n-1} \\ \overleftarrow{X}_t &= \lambda^{-1}\overleftarrow{X}_{t+1} - \lambda^{-1}W_t \end{aligned}$$

where the underlying $\{W_t\}$ are iid. Unrolling the recursion gives $\overleftarrow{X}_t = -\sum_{i=0}^{n-1-t} \lambda^{-i-1} W_{t+i}$. The Z is thus effectively the endpoint $Z = \overleftarrow{X}_0$. The vectors $\vec{X}_k^-$ are made available to the encoder every n time units along with their corresponding side-information Z_k. The goal is to communicate these to a receiver that has access to the side-information Z_k so that $\frac{1}{n}\sum_{i=1}^{n-1} E[\rho(\overleftarrow{X}_{k,i}, \widehat{X}_{k,i}^-)] \leq d$ for all k.

The relevant rate distortion function for the above problem is the conditional rate-distortion function $R_n^{\overleftarrow{X}|Z}(d)$. The proof of Theorem 2.1 in the previous section involves a slightly modified version of the above where the side-information Z is known only to some quantization precision Δ. The quantized side-information is $Z^q = Q_{(\Delta,\Theta)}(Z)$. The relevant conditional rate-distortion function is $R_n^{\overleftarrow{X}|Z^q,\Theta}(d)$.

3) Reductions back to the original problem: It is obvious how to put these two problems together to construct an unstable $\{X_t\}$ stream: the endpoints problem provides the skeleton and the conditional history interpolates in between. To reduce the endpoints problem to the original unstable source communication problem, just use randomness at the transmitter to sample from the interpolating distribution and fill in the history.

To reduce the conditional history problem to the original unstable source communication problem, just use the iid Z_k to simulate the endpoints problem and use the interpolating $\vec{X}$ history to fill out $\{X_t\}$. Because the distortion measure is a difference distortion measure, the perfectly known endpoint process allows us to translate everything so that the same average distortion is attained.

B. Rate-distortion relationships proved

Theorem 2.4 tells us that the unstable $|\lambda| > 1$ Markov processes are nonclassical only as they evolve into the future. The historical information is a stable Markov process that fleshes out the unstable skeleton of the nonstationary process. This fact also allows a simplification in the code depicted in Fig. 5. Since the side-information does not impact the rate-distortion curve for the stable historical process, the encoding of the historical information can be done unconditionally and on a block-by-block basis. There is no need for superblocks.

The remainder of this section proves Theorem 2.4.

Proof:

1) (a): It is easy to see that $R_\infty^{\overleftarrow{X}}(d) = \lim_{n\to\infty} R_n^{\overleftarrow{X}|Q_\Delta(\overleftarrow{X}_0)}(d)$ since the endpoint $\overleftarrow{X}_0$ is distributed like $-\sum_{i=1}^{t} \lambda^{-i} W_i$ and has a finite η-th moment by assumption. By Lemma B.1 (in the Appendix), the entropy of $Q_\Delta(\overleftarrow{X}_0)$ is bounded below a constant that depends only on the precision Δ. This finite number is then amortized away as $n \to \infty$.

2) (b): Next, we show

$$\lim_{\Delta\to 0} R_\infty^{\overleftarrow{X}|Q_\Delta(\overleftarrow{X}_0)}(d) = \lim_{n\to\infty} R_n^{\overleftarrow{X}|\overleftarrow{X}_0}(d). \tag{20}$$

For notational convenience, let $Z^q = Q_\Delta(\overleftarrow{X}_0)$. First, $R_n^{\overleftarrow{X}|\overleftarrow{X}_0}(d)$ is immediately bounded above by $R_n^{\overleftarrow{X}|Z^q}(d)$ since knowledge of $\overleftarrow{X}_0$ exactly is better than knowledge of only the quantized Z^q. To get a lower bound, imagine a hypothetical problem that is one time-step longer and consider the choice between knowing $\overleftarrow{X}_0$ to fine precision Δ or knowing $\overleftarrow{X}_{-1}$ exactly.

$$\begin{aligned} R^{\overleftarrow{X}_0^{n-1}|\overleftarrow{X}_{-1}}(d) &\geq_{(i)} R^{\overleftarrow{X}_0^{n-1}|\overleftarrow{X}_{-1},Z^q}(d) \\ &\geq_{(ii)} R^{\overleftarrow{X}_0^{n-1}|\overleftarrow{X}_{-1},Z^q,C_\gamma,G_\delta,W''_\delta}(d) \end{aligned}$$

where (i) and (ii) above hold since added conditioning can only reduce the conditional rate-distortion function, and $C_\gamma, G_\delta, W''_\delta$ are from the following lemma applied to the hypothesized W_{-1} driving noise.

Lemma 4.1: Given a random variable W with density f_W, arbitrary $1 > \gamma > 0$, there exists a $\delta > 0$ so that it is possible to realize W as

$$W = (1 - C_\gamma)(G_\delta + U_\delta) + C_\gamma W''_\delta \tag{21}$$

where

- C_γ is a Bernoulli random variable with probability γ of being 1.
- U_δ is a continuous uniform random variable on $[-\frac{\delta}{2}, +\frac{\delta}{2}]$.
- G_δ and W''_δ are some random variables whose distributions depend on f_W, δ, γ.
- $C_\gamma, U_\delta, G_\delta, W''_\delta$ are all independent of each other.

Proof: See Appendix A.

Pick γ small and then choose $\Delta \ll \delta$. Notice that $\overleftarrow{X}_{-1} = \lambda^{-1}\overleftarrow{X}_0 - \lambda^{-1}(1-C_\gamma)(G_\delta + U_\delta) + \lambda^{-1}C_\gamma W''_\delta$ where $C_\gamma, U_\delta, G_\delta, W''_\delta$ are independent of each other as well as the entire vector $\overleftarrow{X}_0^{n-1}$. Because the $\{\overleftarrow{X}_t\}$ process is Markov, the impact of the observations $\overleftarrow{X}_{-1}, Z^q, C_\gamma, G_\delta, W''_\delta$ on the conditional rate-distortion function is factored entirely through the posterior distribution for $\overleftarrow{X}_0$.

There are two cases:

- $C_\gamma = 1$ The value for $\overleftarrow{X}_0$ is entirely revealed by the observations. The posterior is a Dirac delta.
- $C_\gamma = 0$ There are two independent measurements of $\overleftarrow{X}_0$. The first is the quantization Z^q. The second is $\lambda\overleftarrow{X}_{-1} + G_\delta = \overleftarrow{X}_0 - U_\delta$. This is just $\overleftarrow{X}_0$ blurred by uniform noise.
  It is useful to view them as coming one after the other. After seeing $Z^q = Q_\Delta(\overleftarrow{X}_0) = z_1$, the posterior distribution $\mathcal{P}(\overleftarrow{X}_0 | Z^q = z_1)$ has support only within $[z_1 - \frac{\Delta}{2}, z_1 + \frac{\Delta}{2}]$.
  The distribution $\mathcal{P}(Z_2 | Z^q = z_1)$ for the second observation $Z_2 = \overleftarrow{X}_0 - U_\delta$ conditioned on the first observation has a pair of interesting properties. First, it has support only on $[z_1 - \frac{\delta+\Delta}{2}, z_1 + \frac{\delta+\Delta}{2}]$. Second, the distribution is uniform over the interval $(z_1 - \frac{\delta-\Delta}{2}, z_1 + \frac{\delta-\Delta}{2})$ since the $\mathcal{P}(\overleftarrow{X}_0 | Z^q = z_1)$ has support with total span $\Delta \ll \delta$.
  Consider the posterior $\mathcal{P}(\overleftarrow{X}_0 | Z^q = z_1, Z_2 = z_2)$ for $z_2 \in (z_1 - \frac{\delta-\Delta}{2}, z_1 + \frac{\delta-\Delta}{2})$ and apply Bayes rule:

$$\begin{aligned}
\mathcal{P}(\overleftarrow{X}_0 \le x | Z^q = z_1, Z_2 = z_2) &= \frac{\mathcal{P}(\overleftarrow{X}_0 \le x, Z_2 = z_2 | Z^q = z_1)}{\mathcal{P}(Z_2 = z_2 | Z^q = z_1)} \\
&= \delta \mathcal{P}(\overleftarrow{X}_0 \le x, Z_2 = z_2 | Z^q = z_1) \\
&= \left(\delta \mathcal{P}(Z_2 = z_2 | Z^q = z_1, \overleftarrow{X}_0 \le x)\right) \mathcal{P}(\overleftarrow{X}_0 \le x | Z^q = z_1) \\
&= \mathcal{P}(\overleftarrow{X}_0 \le x | Z^q = z_1).
\end{aligned}$$

  So if it lands in this region, the second observation is useless. Notice that $U_\delta \in (\frac{\delta-2\Delta}{2}, \frac{\delta-2\Delta}{2})$ forces the second observation to be inside this region. Thus the second observation is useless with probability at least $(1-\gamma)\frac{\delta-2\Delta}{\delta}$ regardless of what the actual $\overleftarrow{X}_0^{n-1}$ are.

Define a new hypothetical observation Z' that with probability $(1-\gamma)\frac{\delta-2\Delta}{\delta}$ is just equal to Z^q and is equal to $\overleftarrow{X}_0$ otherwise. The above tells us that this is a more powerful observation than than the original $(\overleftarrow{X}_{-1}, Z^q, C_\gamma, G_\delta, W''_\delta)$. Thus

$$\begin{aligned}
R^{\overleftarrow{X}_0^{n-1} | \overleftarrow{X}_{-1}}(d) &\ge R^{\overleftarrow{X}_0^{n-1} | Z'}(d) \\
&= (1-\gamma)\frac{\delta-2\Delta}{\delta} R^{\overleftarrow{X}_0^{n-1} | Z^q}(d) + \left(1 - (1-\gamma)\frac{\delta-2\Delta}{\delta}\right) R^{\overleftarrow{X}_0^{n-1} | \overleftarrow{X}_0}(d) \\
&\ge (1-\gamma)\frac{\delta-2\Delta}{\delta} R^{\overleftarrow{X}_0^{n-1} | Z^q}(d).
\end{aligned}$$

Simple algebra then reveals that

$$\begin{aligned} R_n^{\overleftarrow{X}|Z^q} &= \frac{1}{n} R^{\overleftarrow{X}_0^{n-1}|Z^q}(d) \\ &\leq \frac{\delta R^{\overleftarrow{X}_0^{n-1}|\overleftarrow{X}_{-1}}(d)}{n(1-\gamma)\delta - 2\Delta} \\ &= \frac{\delta(n+1)}{n(1-\gamma)\delta - 2\Delta} R_{n+1}^{\overleftarrow{X}|Z}(d). \end{aligned}$$

Taking the limits of $n \to \infty, \frac{\Delta}{\delta} \to 0, \delta \to 0, \gamma \to 0$ establishes the desired result.

Notice that an identical argument works to show that

$$\lim_{\Delta \to 0} \lim_{n \to \infty} R_n^{X|Z^q}(d) = \lim_{n \to \infty} R_n^{X|X_n}(d)$$

for the forward unstable process. It does not matter if it is conditioned on the exact endpoint or a finely quantized version of it. Notice also that the argument is unchanged if the quantization was dithered rather than undithered.

3) (c): This follows almost immediately from (18) from Proposition 3.2. The only remaining task is to show that

$$R_\infty^{\overleftarrow{X}|Z}(d) = R_\infty^{X|Z}(d).$$

It is clear that the iid $\{Z_k\}$ in the "conditional history" problem are just scaled-down (by a factor of $\lambda^{-(n-1)}$) versions of the $\{\widetilde{W}_k\}$ from the "endpoints" problem. The forward $\vec{X}_k = (X_{k,1}, \ldots, X_{k,n-1})$ can be recovered using a simple translation of $\vec{X}_k^-$ by the vector $(Z_k, \lambda Z_k, \ldots, \lambda^{n-1} Z_k)$ since

$$\begin{aligned} X_t &= \sum_{i=0}^{t-1} \lambda^{t-i-1} W_i \\ &= \sum_{i=0}^{n-1} \lambda^{t-i-1} W_i - \sum_{i=t}^{n-1} \lambda^{t-i-1} W_i \\ &= \lambda^{t-1} \sum_{i=0}^{n-1} \lambda^{-i} W_i - \sum_{i=0}^{n-1-t} \lambda^{-i-1} W_{t+i} \\ &= \lambda^{t-1} Z + \overleftarrow{X}_t. \end{aligned}$$

Similarly, the conditional history problem can be recovered from the forward one by another simple translation of $\vec{X}_k$ by the vector $(-\lambda^{-(n-1)} Z_k, \ldots, -\lambda^{-1} Z_k, -Z_k)$.

Thus, the problem of encoding the conditional history to distortion d conditioned on its endpoints is the same whether we are considering the unstable forward or stable backwards processes.

4) Phase transition: At rates strictly less than $\log_2 \lambda$, the distortion for the original X process is necessarily infinite. This is shown in Lemma 6.2 where finite distortion implies the ability to carry $\approx \log_2 \lambda$ bits through the communication medium. □

V. Quality of service requirements for communicating unstable processes: sufficiency

In Section V-A, the sense of anytime reliability is reviewed from [5] and related to classical results on sequential coding for noisy channels. Then in Section V-B, anytime reliable communication is shown to be sufficient for protecting the encoding of the checkpoint process, thereby proving Theorem 2.2. Finally in Section V-C, it is shown that it is sufficient to communicate the historical information using traditional Shannon ϵ-reliability, thereby proving Theorem 2.3.

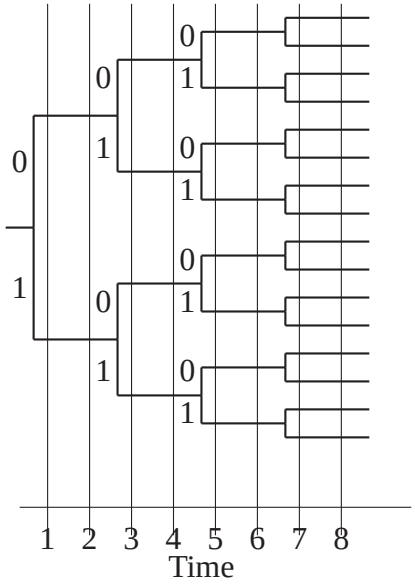

Fig. 6. A channel encoder viewed as a tree. At every integer time, each path of the tree has a channel input symbol. The path taken down the tree is determined by the message bits to be sent. Infinite trees have no intrinsic target delay and bit/path estimates can get better as time goes on.

A. Anytime reliability

It should be clear that the encoding given for the checkpoint process in Section III-B is very sensitive to bit errors since it is decoded recursively in a way that propagates errors in an unbounded fashion. To block this propagation of errors, the channel code must guarantee not only that every bit eventually is received correctly, but that this happens fast enough. This is what motivates the definition of anytime reliability given in Definition 2.6. The relationship of anytime reliability to classical concepts of error exponents as well as bounds are given in [7], [5].

Here, the focus is on the case where there is no explicit feedback of channel outputs. Consider maximum-likelihood decoding [26] or sequential-decoding [28] as applied to an infinite tree code like the one illustrated in Fig. 6. The estimates $\widehat{B}_i(t)$ describe the current estimate for the most likely path through the tree based on the channel outputs received so far. Because of the possibility of "backing up," in principle the estimate for $\widehat{B}_i$ could change at any point in time. The theory of both ML and sequential decoding tells us that generically, the probability of bit error on bit i approaches zero exponentially with increasing delay.

In traditional analysis, random ensembles of infinite tree codes are viewed as idealizations used to study the asymptotic behavior of finite sequential encoding schemes such as convolutional codes. We can instead interpret the traditional analysis as telling us that random infinite tree codes achieve anytime reliability. In particular, we know from the analysis of [26] that at rate R bits per channel use, we can achieve anytime reliability α equal to the block random coding error exponent. Pinsker's argument in [29] as generalized in [7] tells us also that we cannot hope to do any better, at least in the high-rate regime for symmetric channels. We summarize this interpretation in the following theorem:

Theorem 5.1: Random anytime codes exist for all DMCs For a stationary discrete memoryless channel (DMC) with capacity C, randomized anytime codes exist without feedback at all rates $R < C$ and have anytime reliability $\alpha = E_r(R)$ where $E_r(R)$ is the random coding error exponent as calculated in base 2.

Proof: See Appendix D.

B. Sufficiency for the checkpoint process

The effect of any bit error in the checkpoint encoding of Section III-B will be to throw us into a wrong bin of size Δ. This bin can be at most $\lambda^n \frac{\Omega}{\lambda-1}$ away from the true bin. The error will then propagate and grow by a factor λ^n as we move from checkpoint to checkpoint.

If we are interested in the $\eta-$difference distortion, then the distortion is growing by a factor of $\lambda^{n\eta}$ per checkpoint, or a factor of λ^η per unit of time. As long as the probability of error on the message bits goes down faster than that, the expected distortion will be small. This parallels Theorem 4.1 in [5] and results in this proof for Theorem 2.2.

Proof: Let $\check{X}'_{kn}(\phi)$ be the best estimate of the checkpoint $\check{X}_{kn}$ at time $kn+\phi$. By the anytime reliability property, grouping the message bits into groups of nR_1 at a time, and the nature of exponentials, it is easy to see that there exists a constant K' so that:

$$\begin{aligned} E[|\check{X}'_{kn}(\phi) - \check{X}_{kn}|^\eta] &\leq \sum_{j=0}^{k} K' 2^{-\alpha(\phi+nj)} \lambda^{jn\eta} \frac{\Omega}{\lambda-1} \\ &= K'' 2^{-\alpha\phi} \sum_{j=0}^{k} 2^{-jn(\alpha-\eta\log_2\lambda)} \\ &\leq K'' 2^{-\alpha\phi} \sum_{j=0}^{\infty} 2^{-jn(\alpha-\eta\log_2\lambda)} \\ &= K''' 2^{-\alpha\phi} \end{aligned}$$

where K''' is a constant that depends on the anytime code, rate R_1, support Ω, and unstable λ. Thus by making sure $\alpha > \eta\log_2\lambda$ and choosing ϕ large enough, $2^{-\alpha\phi}$ will become small enough so that $K'''2^{-\alpha\phi}$ is as small as we like and the checkpoints will be reconstructed to arbitrarily high fidelity. □

Theorem 2.2 applies even in the case that $\lambda = 1$ and hence answers the question posed by Berger in [12] regarding the ability to track an unstable process over a noisy channel without perfect feedback. Theorem 5.1 tells us that it is in principle possible to get anytime reliability without any feedback at all, and thus also with only noisy feedback.

This idea of tracking an unstable process using an anytime code is useful beyond the source-coding context. In [30], [31], [32], anytime codes are used over a noisy feedback link to study the reliability functions for communication using ARQ schemes and expected delay. The sequence numbers of blocks are considered to be an unstable process that needs to be tracked at the encoder. The random requests for retransmissions make it behave like a random walk with a forward drift, but that can stop and wait from time to time.

C. Sufficiency for the history process

It is easy to see that the history information for the two stream code does not propagate errors from superblock to superblock and so does not require any special QoS beyond what one would need for an iid or stationary-ergodic process. This is the basis for proving Theorem 2.3.

Proof: Since the impact of a bit error is felt only within the superblock, no propagation of errors needs to be considered. Theorem 2.4 tells us that there is a maximum possible distortion on the historical component. Thus the standard achievability argument [8] for $D(R)$ tells us that as long as the probability of block error can be made arbitrarily small ϵ with increasing block-length, then the additional expected distortion induced by decoding errors will also be arbitrarily small. The desired probability of bit error can then be set to be ϵ divided by the superblock length. □

The curious fact here is that the QoS requirements of the second stream of messages only need to hold on a superblock-by-superblock basis. To achieve a small ensemble average distortion, there is no need

to have a secondary bitstream available with error probability that gets arbitrarily small with increased delay! The secondary channel could be nonergodic and go into outage for the entire semi-infinite length of time as long as that outage event occurs sufficiently rarely so that the average on each superblock is kept small. Thus the second stream of messages is compatible with the approach put forth in [33].

VI. Quality of service requirements for communicating unstable processes: necessity

The goal is to prove Theorem 2.5 by showing that unstable Markov processes require communication channels capable of supporting two-tiered service: a high priority core of rate $\log_2 \lambda$ with anytime-reliability of at least $\eta \log_2 \lambda$, and the rest with Shannon reliable bit-transport. To do this, this section proceeds in stages and follows the asymptotic equivalence approach of [5].

This section builds on Section IV-A where the pair of communication problems (the endpoint communication problem and conditional history communication problem) were introduced. In Section VI-A, it is shown that the anytime-reliable bit-transport problem reduces to the first problem (endpoint communication) in the pair. Then Section VI-B finishes the necessity argument by showing how traditional Shannon-reliable bit-transport reduces to the second problem and that the two of them can be put together. This reduces a pair of data-communication problems — anytime-reliable bit transport and Shannon-reliable bit-transport — to the original problem of communicating a single unstable process to the desired fidelity.

The proof construction is illustrated in Fig. 7. Two message streams need to be embedded — a priority stream that requires anytime reliability and a remaining stream for which Shannon-reliability is good enough. The priority stream is used to generate the endpoints while the the history part is filled in with the appropriate conditional distribution. This simulated process is then run through the joint source-channel encoder $\mathcal{E}_s$ to generate channel inputs. The channel outputs are given to the joint source-channel decoder $\mathcal{D}_s$ which produces, after some delay ϕ, a fidelity d reconstruction of the simulated unstable process. By looking at the reconstructions corresponding to the endpoints, it is possible to recover the priority message bits in an anytime reliable fashion. With these in hand, the remaining stream can also be extracted from the historical reconstructions.

A. Necessity of anytime reliability

We follow the spirit of information embedding[34] except that we have no *a-priori* covertext. Instead we use a simulated unstable process that uses common randomness and without loss of generality, message bits assumed to be from iid coin tosses. If the message bits were not fair coin tosses to begin with, XOR them with a one-time pad using common randomness before embedding them. This section parallels the necessity story in [5], except that in this context, there is the additional complication of having a specified distribution for the $\{W_t\}$, not just a bound on the allowed $|W_t|$.

The result is proved in stages. First, we assume that the density of W is a continuous uniform random variable plus something independent. After that, this assumption is relaxed to having a Riemann-integrable density f_W.

1) Uniform driving noise:

Lemma 6.1: Assume the driving noise $W = G + U_\delta$ where G, U_δ are independent random variables with U_δ being a uniform random variable on the interval $[-\frac{\delta}{2}, +\frac{\delta}{2}]$ for some $\delta > 0$.

If a joint source-channel encoder/decoder pair exists for the endpoint process given by (11) that achieves (7) for every position kn, then for every rational rate $R = \frac{nR}{n} < \log_2 \lambda$, there exists a randomized anytime code for the channel that achieves an anytime reliability of $\alpha = \eta \log_2 \lambda$.

Proof: The goal is to simulate the the endpoint process using the message bits and then to recover the message bits from the reconstructions of the endpoints. Pick the initial condition X_0 using common randomness so it can be ignored in what follows.

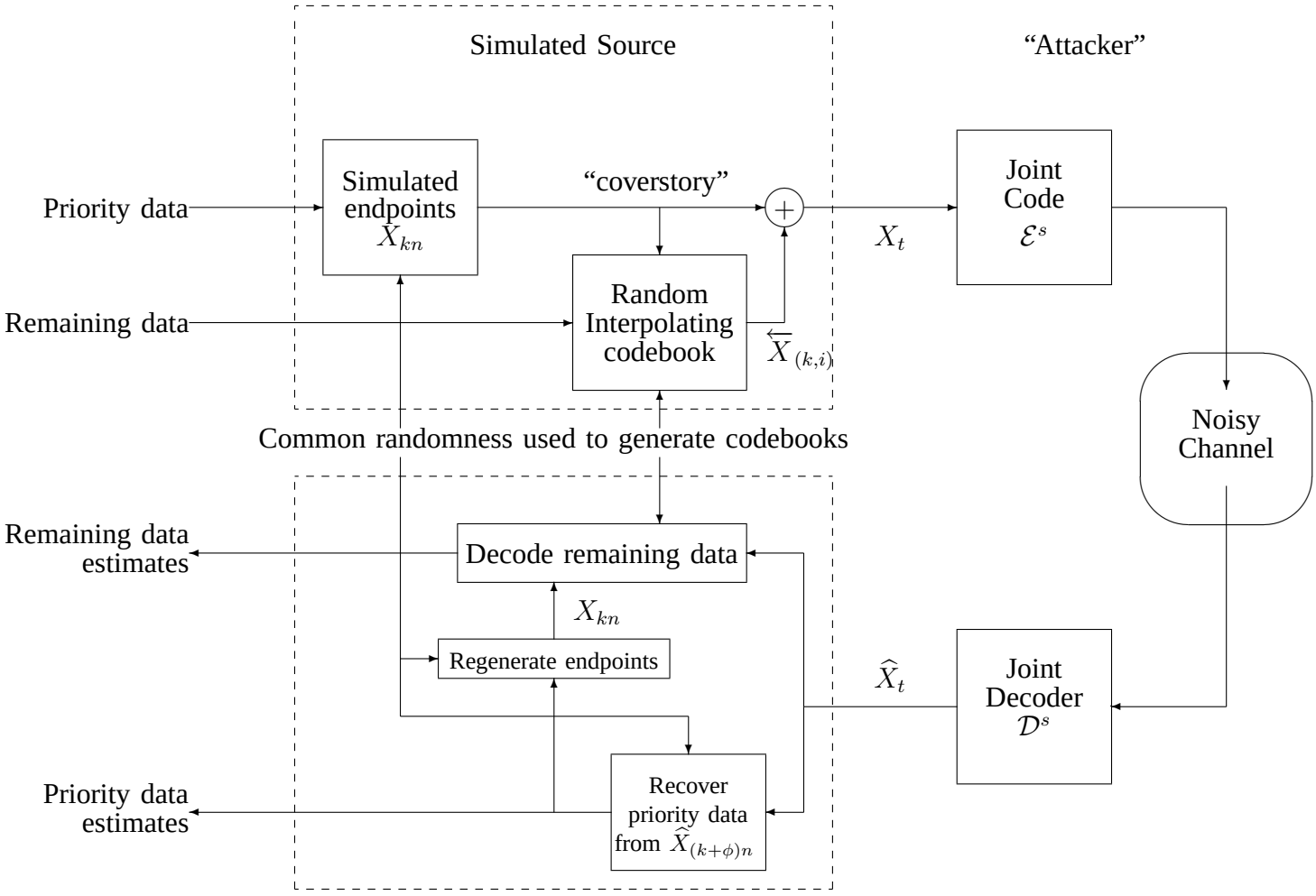

Fig. 7. Turning a joint-source-channel code into a two-stream code using information embedding. The good joint-source-channel code is like an attacker that will not impose too much distortion. Our goal is to simulate a source that carries our messages so that they can be recovered from the attacker's output.

At the encoder, the goal is to simulate

$$\begin{aligned}
\widetilde{W}_k &= \lambda^{n-1}\sum_{i=0}^{n-1}\lambda^{-i}W_{k,i} \\
&= \lambda^{n-1}W_{k,0} + \lambda^{n-1}\sum_{i=0}^{n-1}\lambda^{-i}W_{k,i} \\
&= \lambda^{n-1}U_{\delta,k} + \lambda^{n-1}(G_k + \sum_{i=0}^{n-1}\lambda^{-i}W_{k,i}) \\
&= U_{\lambda^{n-1}\delta,k} + [\lambda^{n-1}(G_k + \sum_{i=0}^{n-1}\lambda^{-i}W_{k,i})]
\end{aligned}$$

The $[\lambda^{n-1}(G_k+\sum_{i=0}^{n-1}\lambda^{-i}W_{k,i})]$ term is simulated entirely using common randomness and is hence known to both the transmitter and receiver. The $U_{\lambda^{n-1}\delta,k}$ term is a uniform random variable on $[-\frac{\lambda^{n-1}\delta}{2}, +\frac{\lambda^{n-1}\delta}{2}]$ and is simulated using a combination of common randomness and the fair coin tosses coming from the message bits.

Fig. 8. The priority message bits are used to refine a point on a Cantor set. The natural tree structure of the Cantor set construction allows us to encode bits sequentially. The Cantor set also has finite gaps between all points corresponding to bit sequences that first differ in a particular bit position. These gaps allow us to reliably extract bit values from noisy observations of the Cantor set point regardless of which point it is.

Since a uniform random variable has a binary expansion that is fair coin tosses, we can write $U_{\lambda^{n-1}\delta,k} = \frac{\lambda^{n-1}\delta}{2}\sum_{\ell=1}^{\infty}(\frac{1}{2})^{\ell}S_{k,\ell}$ where the $S_{k,\ell}$ are iid random variables taking on values ± 1 each with probability $\frac{1}{2}$.

The idea is to embed the iid nR message bits into positions $\ell = 1, 2, \ldots, nR$ while letting the rest — a uniform random variable $U'_{\delta 2^{nR},k}$ representing the semi-infinite sequence of bits $(S_{k,nR+1}, S_{k,nR+2}, \ldots)$ — be chosen using common randomness. The result is:

$$\widetilde{W}_k = \lambda^{n-1}\frac{\delta}{2}M_k + [\lambda^{n-1}(U'_{\delta 2^{nR},k}G_k + \sum_{i=0}^{n-1}\lambda^{-i}W_{k,i})] \tag{22}$$

where M_k is the nR bits of the message as represented by 2^{nR} equally likely points in the interval $[-1, +1]$ spaced apart by 2^{1-nR}, and the rest of the terms $[\cdots]$ are chosen using common randomness known at both the transmitter and receiver side.

Since the simulated endpoints process is a linear function of the $\{\widetilde{W}_k\}$ and the distortion measure is a difference distortion, it suffices to just consider the $\{X'_{kn}\}$ process representing the response to the discrete messages $\{M_k\}$ alone. This has a zero initial condition and evolves like

$$X'_{(k+1)n} = \lambda^n X'_{kn} + \beta M_k \tag{23}$$

where $\beta = \lambda^{n-1}\frac{\delta}{2}$. Expanding this recursion out as a sum gives

$$X'_{(k+1)n} = (\lambda^n)^k \beta \sum_{i=0}^{k}\lambda^{-ni}M_{k-i}. \tag{24}$$

This looks like a generalized binary expansion in base λ^n and therefore implies that the X' process takes values on a growing Cantor set (illustrated in Fig. 8 for $nR = 1$)

The key property is that there are gaps in the Cantor set:

Property 6.1: If the rate $R < \log_2\lambda + \frac{\log_2(1-\lambda^{-n})}{n}$ and the message-streams M and $\bar{M}$ first differ at position j (message $M_j \neq \bar{M}_j$), then at time $k > j$, the encoded X'_{kn} and $\bar{X}'_{kn}$ corresponding to M_1^{k-1} and $\bar{M}_1^{k-1}$ respectively differ by at least:

$$|X'_{kn} - \bar{X}'_{kn}| \leq K\lambda^{n(k-j)} \tag{25}$$

for some constant $K > 0$ that does not depend on the values of the message bits, k, or j.

Proof: See Appendix E.

In coding theory terms, Property 6.1 can be interpreted as an infinite Euclidean free-distance for the code with the added information that the distance increases exponentially as $\lambda^{n(k-j)}$. Thus, a bit error can only happen if the received "codeword" is more than half the minimum distance away.

At the decoder, the common randomness means that the estimation error $X_{kn} - \widehat{X}_{kn}$ is the error in estimating X'_{kn}. By applying Markov's inequality to this using (7), we immediately get a bound on the

probability of an error on the prefix M_0^i for $i < k$:

$$\begin{aligned}
\mathcal{P}(\widehat{M}_1^i(kn) \neq M_1^i) &\leq \mathcal{P}(|\widehat{X}'_{kn} - X'_{kn}| > \frac{K}{2}\lambda^{n(k-i)}) \\
&= \mathcal{P}(|\widehat{X}_{kn} - X_{kn}| > \frac{K}{2}\lambda^{n(k-i)}) \\
&= \mathcal{P}(|\widehat{X}_{kn} - X_{kn}|^\eta > (\frac{K}{2})^\eta(\lambda^{n(k-i)})^\eta \\
&\leq d(\frac{K}{2})^{-\eta}(\lambda^{n(k-i)})^\eta \\
&= K'2^{-(\eta\log_2\lambda)n(k-i)}.
\end{aligned}$$

But $n(k-i)$ is the delay that is experienced at the nR-bit message level. If bits have to be buffered-up to form messages, then the delay at the bit level includes another constant n. This only increases the constant K' further but does not change the exponent with large delays. Thus, the desired anytime reliability is obtained. □

2) General driving noise: Lemma 6.1 can have the technical smoothness condition weakened to simply requiring a Riemann-integrable density for the white W driving process.

Lemma 6.2: Assume the driving noise W has a Riemann-integrable density f_W. If there exists a family of joint source-channel encoder/decoder pairs for a sequence of increasing n-endpoint problems given by (11) that achieve (7) for every position kn, then for every rate $R < \log_2\lambda$ and anytime reliability $\alpha < \eta\log_2\lambda$, there exists a randomized anytime code for the underlying channel.

Proof: Since the density is Riemann-integrable, Lemma 4.1 applies. Choose δ such that $\gamma < \lambda^{-2\eta n}$. When simulating $W_{k,0}$ in the endpoint process, use common randomness for C_γ and W''_δ, and follow the procedure from the proof of Lemma 6.1 for G_δ and U_δ.

We can thus interpret a "heads" for C_γ as an "erasure" with probability γ since no message can be encoded in that time period. From the point of view of Lemma 6.1, this can be considered a known null message.

Since the outcome of these coin tosses come from common randomness, the position of these erasures are known to both the transmitter and the receiver. In this way, it behaves like a packet erasure channel with feedback. This problem is studied in Theorem 3.3 of [7], and the delay-optimal coding strategy relative to the erasure channel is to place incoming packets into a FIFO queue awaiting a non-erased opportunity for transmission. The following lemma summarizes the results needed from [7].

Lemma 6.3: Suppose packets arrive deterministically at a rate of R packets per unit time and enter a FIFO queue drained at constant rate 1 per unit time.

- Suppose $\gamma < \frac{1}{16}$. If each packet has a size distribution that is bounded below a geometric$(1-\gamma)$ (i.e. $\mathcal{P}(\text{Size} > s) \leq \gamma^s$ for all non-negative integers s), then the random delay ϕ experienced by any individual packet from arrival to departure from the queue satisfies $\mathcal{P}(\phi > s) \leq K2^{-\alpha s}$ for all non-negative s and some constant K that does not depend on s. Furthermore, if $R < \frac{1}{1+2r}$ for some $r > 0$, then $\alpha \geq -\log_2\gamma - 2\gamma^r$.
- Assume the rate $R = \frac{1}{n}$ and each packet has a size distribution that is bounded by: $\mathcal{P}(\text{Size} > n(1-\epsilon) + s) \leq \gamma^s$ for all non-negative integers s. Then the delay ϕ experienced by any individual packet has a tail distribution bounded in the same way as for $R' = \frac{1}{n^\epsilon}$ and packets with geometric$(1-\gamma)$ size. That is $\mathcal{P}(\phi > s) \leq K2^{-\alpha s}$ where $\alpha \geq -\log_2\gamma - 2\gamma^{\frac{n\epsilon-1}{2}}$.

Proof: See Theorem 3.3 and Corollary 6.1 of [7].

For our problem, the message bits are arriving deterministically at bit-rate $R < \log_2\lambda$ per unit time to the transmitter. Pick $r > 0$ small enough so that $R' = (1+3r)R < \log_2\lambda$. Group message bits into packets of size nR'. These packets arrive deterministically at rate $\frac{1}{1+3r} < \frac{1}{1+2r}$ packets per n time units.

Thus, Lemma 6.3 applies and the delay (in n units) experienced by a packet in the queue has a delay error exponent α of least

$$\begin{aligned} -\log_2 \gamma - 2\gamma^r &\geq -\log_2 \lambda^{-2\eta n} - 2\lambda^{-2\eta n r} \\ &= n2\eta \log_2 \lambda - 2\lambda^{-2\eta n r} \end{aligned}$$

per n time steps or $2\eta \log_2 \lambda - \frac{2\lambda^{-2\eta n r}}{n}$ per unit time step. When n is large, this exponent is much faster than the delay exponent of $\eta \log_2 \lambda$ obtained in the proof of Lemma 6.1. The two delays experienced by a bit are independent by construction. Thus, the dominant delay-exponent remains $\eta \log_2 \lambda$ as desired. □

Notice that the simulated endpoint process depends only on common randomness and the message packets. Since the common randomness is known perfectly at the receiver by assumption and the message packets are known with a probability that tends to 1 with delay, the endpoint process is also known with zero distortion with a probability tending to 1 as the delay increases.

B. Embedding classical bits

All that remains is to embed the classical message bits into the historical process. The overall construction is described in Fig. 7. First, n is chosen to be large enough so that the R_1 stream can be successfully embedded in the endpoint process by Lemma 6.2.

Now, n is further increased so that $R_2 < R_n^{\overleftarrow{X}|\overleftarrow{X}_0}(d)$ the conditional rate-distortion function for the history given the endpoint. This can be done since $\lim_{n\to\infty} R_n^{\overleftarrow{X}|\overleftarrow{X}_0}(d) = R_\infty^X(d) - \log_2 \lambda$ by Theorem 2.4.

By choosing an appropriate additional delay, Lemma 6.2 assures us that the receiver will know all the past high-priority messages and hence simulated endpoints correctly with an arbitrarily small probability of error ϵ. As described in Section IV-A, this means we now have a family of systems (indexed by m) that solve the conditional history problem. The condition (8) translates into

$$\lim_{m\to\infty} \sup_{\tau\geq 0} \mathcal{P}\left(\frac{1}{m}\sum_{k=\tau}^{\tau+m-1} \frac{1}{n}\sum_{i=1}^{n-1} |\overleftarrow{X}_{(k,i)} - \widehat{X}^-_{(k,i)}|^\eta > d\right) = 0. \tag{26}$$

It tells us that by picking m large enough, the probability of having excess distortion can be made as small as desired.

The simulated $\{Z_k\}$ containing the high-priority messages are interpreted as the "coverstory" that must be respected when embedding messages into the $\{\vec{X}_k^-\}$ process. The $\{Z_k\}$ are iid by construction and hence Theorem 3 from [18] (full proofs in [19]) applies and tells us that a length $m' > m$ random code with $\vec{X}_k^-$ drawn independently of each other, but conditional on the iid Z_k, can be used to embed information at any rate $nR_2 < nR_n^{\overleftarrow{X}|Z}(d+\epsilon) = nR_n^{X|X_n}(d+\epsilon)$ per vector symbol with arbitrarily low probability of error. □

The "weak law of large numbers"-like condition (8), or something like it, is required for the theorem to hold since there are joint source-channel codes for which mutual information cannot be turned into the reliable communication of bits at arbitrarily low probabilities of error. Consider the following contrived example. Suppose there are two different joint source-channel codes available: one has a target distortion of d_1 and the other has a target distortion of $d_2 = 10d_1$. The actual joint code, which is presumed to have access to common randomness, could decide with probability $\frac{1}{1000}$ to use the second code rather than the first. In such a case, the ensemble average mutual information is close to $R(d_1) - \log_2 \lambda$ bits, but with non-vanishing probability $\frac{1}{1000}$ we might not be able sustain such a rate over the virtual channel.

We conjecture that for DMCs, if any joint source-channel code exists that hits the target distortion on average, then one should also exist that meets (8) and it should be possible to simultaneously communicate two streams of messages reliably with anytime reliability on the first stream and enough residual rate on the second.

VII. Unstable Gauss-Markov processes with squared-error distortion

A. Source-coding for Gaussian processes

The goal here is to prove Corollary 2.1. The strategy is essentially as before. One simplification is that we can make full use of Theorem 2.4 and rely on $R_\infty^{\overleftarrow{X}}(d) = R_\infty^X(d) - \log_2 \lambda$. There is thus no rate loss in encoding the historical segments on a block-by-block basis rather than using superblocks and conditional encodings. The only issue that remains is dealing with the unbounded support when encoding the checkpoints.

The overall approach is: (key differences *italicized*)

(a) Look at time in blocks of size n and encode the values of checkpoints X_{kn} recursively to very high precision using a prefix-free *variable-length* code with rate $n(\log_2 \lambda + \epsilon_1) + L_k$ bits per value, where the L_k are iid random variables with appropriately nice properties.

(b) *Smooth out the variable-length code by running it through a FIFO queue drained at constant rate $R_1 = \log_2 \lambda + \epsilon_1 + \epsilon_q$. Make sure that the delay exponent in the queue is high enough.*

(c) Use the *exact value* for the ending checkpoint $X_{(k+1)n}$ (*instead of the quantized* $\check{X}$) to transform the segment immediately before it so that it looks exactly like a stable backwards Gaussian process of length n with initial condition 0. Encode each block of the backwards history process to average-fidelity d using a fixed-rate rate-distortion code for the backwards process that operates at rate $R_\infty^{\overleftarrow{X}}(d) + \epsilon_s$.

(d) At the decoder, wait ϕ time units and attempt to decode the checkpoints to high fidelity. *If the FIFO queue is running too far behind, then extrapolate a reconstruction based on the last fully decoded checkpoint.*

(e) Decode the history process to average-fidelity d and combine it with the recursively quantized checkpoints to get the reconstruction.

a) Encoding the checkpoints: (11) remains valid, but the term $\widetilde{W}_k = \lambda^{n-1} \sum_{i=0}^{n-1} \lambda^{-i} W_{kn+i}$ is not bounded since the W_i are iid Gaussians. The $\widetilde{W}_k$ are instead Gaussian with variance

$$\begin{aligned} \widetilde{\sigma}^2 &= \lambda^{2(n-1)} \sum_{i=0}^{n-1} \lambda^{-2i} \sigma^2 \\ &\leq \lambda^{2(n-1)} \sigma^2 \sum_{i=0}^{\infty} \lambda^{-2i} \\ &= \lambda^{2n} \frac{\sigma^2}{\lambda^2 - 1}. \end{aligned}$$

The standard deviation $\widetilde{\sigma}$ is therefore $\lambda^n \frac{\sigma}{\sqrt{\lambda^2-1}}$. Pick $l = 2^{\frac{\epsilon_1}{3} n}$ and essentially pretend that this random variable $\widetilde{W}_k$ has bounded support of $l\widetilde{\sigma}$ during the encoding process. By comparing (12) to the above, the effective Ω is simply $l\sigma \frac{2(\lambda-1)}{\sqrt{\lambda^2-1}} = 2^{\frac{\epsilon_1}{3} n} \sigma \sqrt{\frac{\lambda-1}{\lambda+1}}$. Define $\widetilde{\Omega} = \sigma \sqrt{\frac{\lambda-1}{\lambda+1}}$ so that the effective $\Omega = 2^{\frac{\epsilon_1}{3} n} \widetilde{\Omega}$.

Encode the checkpoint increments recursively as before, only add an additional variable-length code for the value of $\lfloor \frac{\widetilde{W}}{l\widetilde{\sigma}} + \frac{1}{2} \rfloor$ while treating the remainder using the fixed-rate code as before. The variable length code is a unary encoding that counts how many $l\widetilde{\sigma}$ away from the center the $\widetilde{W}_k$ actually is. (Fig. 9 illustrates the unary code.) Let L_k be the length of the k-th unary codeword. This is bounded above by

$$P(L_k \geq 3 + j) = P(|\widetilde{W}| > jl\widetilde{\sigma}).$$

Let N be a standard Gaussian random variable and rewrite this as

$$P(L_k \geq 3 + j) = P(|N| > j 2^{\frac{\epsilon_1}{3} n}) \leq \exp(-\frac{1}{2} j^2 2^{\frac{2\epsilon_1}{3} n}) \tag{27}$$

and so L_k is very likely indeed to be small and certainly has a finite expectation $\bar{L} < 4$ if n is large.

Offset	Codeword
0	100
+1	1110
-1	1100
+2	11110
-2	11010
+3	111110
-3	110110
+4	1111110
-4	1101110
⋮	⋮

Fig. 9. Unary encoding of integer offsets to deal with the unbounded support. The first bit denotes start while the nest two bits reflect the sign. The length of the rest reflects the magnitude of the offset with a zero termination. The encoding is prefix-free and hence uniquely decodable. The length of the encoding of integer S is bounded by $3 + |S|$

The fixed-rate part of the checkpoint encoding has a rate that is the same as that given by (13), except that Ω is now mildly a function of n. Plugging in $2^{\frac{\epsilon_1}{3}n}\widetilde{\Omega}$ for Ω in (13) gives

$$\begin{aligned} R_{1,f} &\geq \max\left(\log_2\lambda + \frac{\log_2(1+\frac{\Omega}{\Delta(\lambda-1)}) + \log_2(2+\frac{\Omega}{\Delta})}{n}, \frac{\log_2\lceil\frac{\Omega_0}{\Delta}\rceil}{n}\right) \\ &= \max\left(\log_2\lambda + \frac{\log_2(1+\frac{2^{\frac{\epsilon_1}{3}n}\widetilde{\Omega}}{\Delta(\lambda-1)}) + \log_2(2+\frac{2^{\frac{\epsilon_1}{3}n}\widetilde{\Omega}}{\Delta})}{n}, \frac{\log_2\lceil\frac{\Omega_0}{\Delta}\rceil}{n}\right) \\ &= \max\left(\log_2\lambda + \frac{2}{3}\epsilon_1 + \frac{\log_2(2^{-\frac{\epsilon_1}{3}n}+\frac{\widetilde{\Omega}}{\Delta(\lambda-1)}) + \log_2(2^{1-\frac{\epsilon_1}{3}n}+\frac{\widetilde{\Omega}}{\Delta})}{n}, \frac{\log_2\lceil\frac{\Omega_0}{\Delta}\rceil}{n}\right). \end{aligned}$$

Essentially, the required rate $R_{1,f}$ for the fixed-rate part has only increased by a small constant $\frac{2}{3}\epsilon_1$. Holding Δ fixed and assuming n is large enough, we can see that

$$R_{1,f} = \log_2\lambda + \epsilon_1 \tag{28}$$

is sufficient.

b) Smoothing out the flow: The code so far is variable-rate and to turn this into a fixed-rate $R_1 = \log_2\lambda + \epsilon_1 + \epsilon_q$ bitstream, it is smoothed by going through a FIFO queue. First, encode the offset using the variable-length code and then recursively encode the increment as was done in the finite support case. All such codes will begin with a 1 and thus we can use zeros to pad the end of a codeword whenever the FIFO is empty. When n is large, the average input rate to the FIFO is smaller than the output rate and hence it will be empty infinitely often.

c) Getting history and encoding it: Section IV explains why such a transformation is possible by subtracting off a scaled version of the endpoint. The result is a stable Gaussian process and so [9] reveals that it can be encoded arbitrarily close to its rate-distortion bound $R_\infty^{\overleftarrow{X}}(d) = R_\infty^X(d) - \log_2\lambda$ if n is large enough.

d) Decoding the checkpoints: The decoder can wait long enough so that the checkpoint we are interested in is very likely to have made it through the FIFO queue by now. The ideas here are similar to [7], [35] in that a FIFO queue is used to smooth out the rate variation with good large deviations performance. There is $n\epsilon_q$ slack that has to accommodate L_k bits. Because n can be made large, the error exponent with delay here can be made as large as needed.

More precisely, a packet of size $n(\epsilon_1+\log_2\lambda)+L_k$ bits arrives every n time units where the L_k are iid. This is drained at rate $R_1=\epsilon_q+\epsilon_1+\log_2\lambda$. An alternative view is therefore that a point packet arrives deterministically every n time units and it has a random service time T_k given by $n\frac{\epsilon_1+\log_2\lambda}{\epsilon_q+\epsilon_1+\log_2\lambda}+\frac{L_k}{\epsilon_q+\epsilon_1+\log_2\lambda}$. Define $(1-\epsilon'_q)=\frac{\epsilon_1+\log_2\lambda}{\epsilon_q+\epsilon_1+\log_2\lambda}$. Then the random service time $T_k=(1-\epsilon'_q)n+\frac{L_k}{\epsilon_q+\epsilon_1+\log_2\lambda}$ when measured in time units or $T_k^b=(1-\epsilon'_q)nR_1+L_k$ when measured in bit-units.

This can be analyzed using large-deviations techniques or by applying standard results in queuing. The important thing is a bound on the length L_k which is provided by (27). It is clear that[1],

$$\begin{aligned} P(L_k\geq 3+j) &\leq \exp(-\frac{1}{2}j^2 2^{\frac{2\epsilon_1}{3}n}) \\ &\leq \exp(-2^{\frac{2\epsilon_1}{3}n-1})j). \end{aligned}$$

Since an exponential eventually dominates all constants, we know that for any $\beta>0$, there exists a sufficiently large n so that:

$$P(L_k-3>j)\leq 2^{-\beta j}. \tag{29}$$

Thus, the delay (in bits) experienced by a block in the queue will behave no worse than that of point messages arriving every nR_1 bits where each requires at least $nR_1(1-\epsilon'_q)+3=nR_1(1-\epsilon''_q)$ bits plus an iid geometric$(1-p)$ number of bits with $p=2^{-\beta}$.

Lemma 6.3 applies to this queuing problem and the second part of that lemma tells us that the delay performance is exactly the same as that of a system with point messages arriving every $n\epsilon''_q$ bits requiring only an iid geometric number of bits. Since $\frac{1}{n\epsilon''_q}$ is small, the first part of Lemma 6.3 applies. Set $r=\frac{n\epsilon''_q}{3}-1$, then the bit-delay exponent α_b is at least

$$\begin{aligned} \alpha_b &\geq -\log_2 2^{-\beta}-2^{-\beta r} \\ &= \beta-2^{-\beta(\frac{n\epsilon''_q}{3}-1)} \end{aligned}$$

which is at least $\beta-1$ when $n\epsilon''_q\geq 3$. Converting between bit-delay and time-delay is essentially just a factor of $\log_2\lambda$ and so the time-delay exponent is at least $\frac{\beta-1}{\log_2\lambda}$. But β can be made as large as we want by choosing n large enough.

e) Getting the final reconstruction: The history process is added to the recovered checkpoint. This differs from the original process by only the error in the history plus the impact of the error in the checkpoint. The checkpoint reconstruction-error's impact dies exponentially since the history process is stable. So the target distortion is achieved if the checkpoint has arrived by the time reconstruction is attempted. By choosing a large enough end-to-end delay ϕ, the probability of this can be made as high as we like.

However, the goal is not just to meet the target distortion level d with high probability, it is also to hit the target in expectation. Thus, we must bound the impact of not having the checkpoint available in time. When this happens, the un-interpretable history information is ignored and the most recent checkpoint is simply extrapolated forward to the current time. The expected squared errors grow as $\lambda^{2\psi}$ where ψ is the delay in time-units. The arguments here exactly parallel those of Theorem 2.2, where the FIFO queue is acting like an anytime code. Since the delay-exponent of the queue is as large as we want, it can be made larger than $2\log_2\lambda$. Thus, the expected distortion coming from such "overflow" situations is as small as desired. This completes the proof of Corollary 2.1. □

[1]While this proof is written for the Gaussian case, the arguments here readily generalize to any driving distribution W that has at least an exponential tail probability. To accommodate W with power-law tail distributions would require the use of logarithmic encodings as described in [36], [37]. This does not work for our case because the unary nature of the encoding is important when we consider transporting such bitstreams across a noisy channel.

B. Channel sufficiency for communicating Gaussian processes

This section shows why Corollary 2.2 is true. The story in the Gaussian case is mostly unchanged since the historical information is as classical as ever. The only issue is with the checkpoint stream. An error in a bit ψ steps ago can do more than propagate through the usual pathway. It could also damage the bits corresponding to the variable-length offset.

Because of the unary encoding and the $2^{\frac{\epsilon_1}{3}n}$ expansion in the effective Ω, an uncorrected bit-stream error ψ time-steps ago can only impact the error in the checkpoint reconstruction by $4\psi(\log_2\lambda)2^{\frac{\epsilon_1}{3}n}$ since the worst error is clearly to flip the sign bit and keep the unary codeword from terminating thereby making it at most $2\psi\log_2\lambda$ bits long. The current reconstruction is therefore incorrect by an $O(\psi 2^{\frac{\epsilon_1}{3}n}\lambda^{\psi})$ change in its value. As far as η-distortion is concerned, the distortion grows by a factor $O(\psi^{\eta}2^{\eta\frac{\epsilon_1}{3}n}\lambda^{\eta\psi})$ from what it would be with correct reconstruction. Asymptotically, the delay ψ is much larger than the block-length n and so the polynomial term in front is insignificant relative to the exponential in ψ. If the code has anytime reliability $\alpha > \eta\log_2\lambda$, then the same argument as Theorem 2.2 applies and the Corollary holds. □

VIII. Extensions to the vector case

With the scalar case explored, it is natural to consider what happens for general finite-dimensional linear models where λ is replaced with a matrix A and X is a vector. In the Gaussian process case, these will correspond to cases with formally rational power-spectral densities. Though the details are left to the reader, the story is sketched here. No fundamentally new phenomena arise in the vector case, except that different anytime reliabilities can be required on different streams arising from the same source as is seen in the control context [6].

The source-coding results here naturally extend to the fully observed vector case with generic driving noise distributions. Instead of two message streams, there is one special stream for each unstable eigenvalue λ_i of A and a singe final stream capturing the residual information across all dimensions. All the sufficiency results also generalize in a straightforward manner — each of the unstable streams requires a corresponding anytime reliability depending on the distortion function's η and the magnitude of the eigenvalue. The multiple priority-stream necessity results also follow generically.[2] This is a straightforward application of a system diagonalization[3] argument followed by an eigenvalue by eigenvalue analysis. The necessity result for the residual rate follows the same proof as here based on inverse-conditional rate-distortion with the endpoints in all dimensions used as side-information.

The case of partially observed vector Markov processes where the observations $C_y\vec{X}$ are linear in the system state requires one more trick. We need to invoke the observability[4] of the system state. Instead of a single checkpoint pair, use an appropriate number[5] of consecutive values for the observation and encode them together to high fidelity Δ. This can be done by transforming coordinates linearly so that the system is diagonal, though driven by correlated noise, from checkpoint-block to the next checkpoint-block. The initial condition is governed by the self-noise that is unavoidable while trying to observe the state. Each unstable eigenvalue will contribute its own $\log_2\lambda_i$ term to the first stream rate and will require the appropriate anytime reliability. The overhead continues to be sublinear in n and the residual information continues to be classical in nature by the same arguments given here.

[2]The required condition is that the the driving noise distribution W should not have support isolated to an invariant subspace of A. If that were to happen, there would be modes of the process that are never excited.

[3]The case of non-diagonal Jordan blocks is only a challenge for the necessity part regarding anytime reliability. It is covered in [6] in the control context. The same argument holds here with a Riemann-integrable joint-density assumption on the driving noise.

[4]The linear observation should not be restricted to a single invariant subspace. If it were, we could drop the other subspaces from the model as irrelevant to the observed process under consideration.

[5]The appropriate number is twice the number of observations required before all of the unstable subspaces show up in the observation. This number is bounded above by twice the dimensionality of the vector state space. The factor of two is to allow each block to have its own beginning and end.

The partially observed necessity story is essentially unchanged on the information embedding side, except that every long block should be followed by a miniblock of length equal to the dimensionality k during which no message is embedded and only common-randomness is used to generate the driving noise. This will allow the decoder to easily use observability to get noisy access to the unstable state itself.

In [6], these techniques are applied in the context of control rather than estimation. The interested reader is referred there for the details. Some simplifications to the general story might be possible in the case of SISO autoregressive processes, but we have not explored them in detail.

IX. Conclusions

We have characterized the nature of information in an unstable Markov process. On the source coding side, this was done by giving the fixed-rate coding Theorem 2.1. This theorem's code construction naturally produces two streams — one that captures the essential unstable nature of the process and requires a rate of at least $\log_2 \lambda$, and another that captures the essentially classical nature of the information left over. The quantitative distortion is dominated by the encoding of the second stream, while the first stream serves to ensure its finiteness as time goes on. The essentially stable nature of the second stream's information is then made precise by Theorem 2.4 which relates the forward $D(R)$ curve to the "backwards" one corresponding to a stable process.

At the intersection of source and channel coding, the notion of anytime reliability was reviewed and Theorem 5.1 shows that it is nonzero for DMCs at rates below capacity. Theorem 2.2 and Lemma 6.2 then shows that the first stream requires a high-enough anytime reliability from a communication system rather than merely enough rate. In contrast, Theorems 2.3 and 2.5 show that the second stream requires only sufficient rate. Together, all these results establish the relevant separation principle for such unstable Markov processes.

This work brings exponentially unstable processes firmly into the fold of information theory. More fundamentally, it shows that reliability functions are not a matter purely internal to channel coding. In the case of unstable processes, the demand for appropriate reliability arises at the source-channel interface. Thus unstable processes have the potential to be useful models while taking an information-theoretic look at QoS issues in communication systems. The success of the "reductions and equivalences" paradigm of [5], [19] here suggests that this approach might also be useful in understanding other situations in which classical approaches to separation theorems break down.

Appendix A
Riemann-integrable densities as mixtures

It is often conceptually useful to think of generic random variables with Riemann-integrable densities as being mixtures of a blurred uniform random variable along with something else. This appendix proves Lemma 4.1.

Since the density is Riemann-integrable,

$$\int_{-\infty}^{+\infty} f_W(w)dw \;=\; \lim_{\delta\to 0}\sum_{i=-\infty}^{+\infty}\delta \min_{x\in[i\delta-\frac{\delta}{2},i\delta+\frac{\delta}{2}]} f_W(x)$$

Thus, f_W can be expressed as a non-negative piecewise constant function f'_W that only changes every δ units plus a non-negative function f''_W representing the "error" in Riemann-integration from below. By choosing δ small enough, the total mass in f''_W can be made as small as desired since the Riemann sums above converge.

Choose δ such that the total mass in f''_W is γ. So

$$f_W = (1-\gamma)(\frac{f'_W}{1-\gamma}) + \gamma(\frac{f''_W}{\gamma}) \tag{30}$$

and thus W can thus be simulated in the following way:

1) Flip an independent biased coin C_γ with probability of heads γ.
2) If heads, independently draw the value of W from the density $\frac{f''_W}{\gamma}$ corresponding to a random variable W''.
3) If tails, independently draw the value of W from the random variable W'' with piecewise constant density $\frac{f'_W}{1-\gamma}$. This can clearly be done by using a discrete random variable G_δ plus an independent uniform random variable U_δ so that $W'' = G_\delta + U_\delta$ has density $\frac{f'_W}{1-\gamma}$.

This proves the result. □

APPENDIX B
ENTROPY BOUND FOR QUANTIZED RANDOM VARIABLES WITH BOUNDED MOMENTS

Lemma B.1: Consider a random variable Z that is quantized to precision Δ so $Z^q = Q_\Delta(Z)$. Further suppose that $E[|Z|^\eta] \leq K$ where $K > \Delta^\eta$. Then

$$H(S) < 7 + \frac{\log_2 K}{\eta} + 2\log_2 \frac{\log_2 K}{\eta} + \log_2 \frac{1}{\Delta} + 2\log_2 \log_2 \frac{1}{\Delta} + \frac{5+\ln 2}{\eta \ln 2}. \tag{31}$$

Proof: Let $Z^q = S\Delta$ where S is an integer. Then $|S| \leq 1 + \frac{|Z|}{\Delta}$ and so

$$\begin{aligned}
E[|S|^\eta] &\leq E[(1+\frac{|Z|}{\Delta})^\eta] \\
&\leq E[(2\max\left(1, \frac{|Z|}{\Delta}\right))^\eta] \\
&= E[2^\eta \max\left(1^\eta, (\frac{|Z|}{\Delta})^\eta\right)] \\
&\leq E[2^\eta + 2^\eta \frac{|Z|^\eta}{\Delta^\eta}] \\
&= 2^\eta + \frac{2^\eta}{\Delta^\eta} E[|Z|^\eta] \\
&\leq 2^\eta + \frac{2^\eta K}{\Delta^\eta} \\
&< 2^{\eta+1} \frac{K}{\Delta^\eta}.
\end{aligned}$$

Applying the Markov inequality gives

$$\mathcal{P}(|S| \geq s) \leq \min(1, \frac{2^{\eta+1}K}{\Delta^\eta} s^{-\eta}). \tag{32}$$

The integer S can be encoded into bits using a self-punctuated code using less than $4 + \log_2(|S|) + 2\log_2(1+\log_2(|S|+1))$ bits to encode $S \neq 0$ [38]. First encode the sign of S using a single bit. There are at most $1+\log_2(|S|+1)$ digits in the natural binary expansion of $|S|$. This length can be encoded using at most $2 + 2\log_2(1+\log_2(|S|+1))$ bits by giving its binary expansion with each digit followed by a 0 if it is not the last digit, and a 1 if it is the last digit. Finally, $|S|$ itself can be encoded using at most $1 + \log_2 |S|$ bits.

Since the entropy must be less than the expected code-length for any code,

$$\begin{aligned}
H(S) &\leq 4 + E[\log_2(|S|)] + 2E[\log_2(1+\log_2(|S|+1))] \\
&= 4 + \int_0^\infty \mathcal{P}(\log_2(|S|) > l) dl + 2\int_0^\infty \mathcal{P}(\log_2(1+\log_2(|S|+1)) > l) dl.
\end{aligned}$$

First, we deal with the dominant term

$$\begin{aligned}
& \int_0^\infty \mathcal{P}(\log_2(|S|) > l) dl \\
= & \int_0^\infty \mathcal{P}(|S| > 2^l) dl \\
\leq & \int_0^\infty \min(1, \frac{2^{\eta+1}K}{\Delta^\eta} 2^{-\eta l}) dl \\
= & \frac{1}{\eta} \log_2(\frac{2^{\eta+1}K}{\Delta^\eta}) + \int_0^\infty 2^{-\eta u} dl \\
= & 1 + \frac{\log_2 K}{\eta} + \log_2 \frac{1}{\Delta} + \frac{1+\ln 2}{\eta \ln 2}
\end{aligned}$$

Next, consider the smaller term

$$\begin{aligned}
& 2\int_0^\infty \mathcal{P}(\log_2(1+\log_2(|S|+1)) > l) dl \\
= & 2\int_0^\infty \mathcal{P}(\log_2(|S|+1) > 2^l - 1) dl \\
= & 2\int_0^\infty \mathcal{P}(|S|+1 > \frac{2^{2^l}}{2}) dl \\
\leq & 2(1 + \int_0^\infty \mathcal{P}(|S| > 2^{2^l}) dl) \\
\leq & 2 + 2\int_0^\infty \min(1, \frac{2^{\eta+1}K}{\Delta^\eta} 2^{-\eta 2^l}) dl \\
= & 2 + 2\log_2 \frac{\log_2 \frac{2^{\eta+1}K}{\Delta^\eta}}{\eta} + 2\int_0^\infty 2^{-\eta 2^l} dl \\
< & 2 + 2\log_2 \frac{\log_2 K}{\eta} + 2\log_2(1+\frac{1}{\eta}) + 2\log_2 \log_2 \frac{1}{\Delta} + \frac{2}{\eta \ln 2} \\
\leq & 2 + 2\log_2 \frac{\log_2 K}{\eta} + 2\log_2 \log_2 \frac{1}{\Delta} + \frac{4}{\eta \ln 2}
\end{aligned}$$

where the final inequalities come from the concave $\cap$ nature of $\log_2$ and lower bounding 2^l with just l. Putting everything together gives the desired result. □

APPENDIX C
PROOF OF PROPOSITION 3.2

From (3) and (2), we know for every $\epsilon_2 > 0$, if Δ is small enough and n is large enough, that there exists a random vector Y_0^{n-1} so that $\frac{1}{n}\sum_{i=0}^{n-1} \rho(\widetilde{X}_i, Y_i) = d + \epsilon_3$ and that even the best such vector must satisfy

$$n(R_\infty^X(d) - \epsilon_2) \leq I(\widetilde{X}_0^{n-1}; Y_0^{n-1}) \leq n(R_\infty^X(d) + \epsilon_2).$$

Decompose the relevant mutual information as

$$I(\widetilde{X}_0^{n-1}; Y_0^{n-1} | Z^q, \Theta) = -I(\widetilde{X}_0^{n-1}; Z^q | \Theta) + I(\widetilde{X}_0^{n-1}; Y_0^{n-1}, Z^q | \Theta). \tag{33}$$

To get the desired result of asymptotic equality, this conditional mutual information has to be both upper and lower bounded. To upper bound the conditional mutual information, we lower bound $I(\widetilde{X}_0^{n-1}; Z^q|\Theta)$ and upper bound $I(\widetilde{X}_0^{n-1}; Y_0^{n-1}, Z^q|\Theta)$. Vice-versa to get the lower bound.

A. Lower bounding $I(\widetilde{X}_0^{n-1}; Z^q|\Theta)$

The first term is easily lower bounded for Δ small enough since

$$\begin{aligned} I(\widetilde{X}_0^{n-1}; Z^q|\Theta) &= H(Z^q|\Theta) - H(Z^q|\widetilde{X}_1^n, \Theta) \\ &= H(Z^q|\Theta) \\ &\geq H(Z^q|\Theta, W_0^{n-2}) \\ &\geq \lfloor \log_2 \lambda^{n-1} \rfloor \\ &= \lfloor (n-1)\log_2 \lambda \rfloor. \end{aligned} \tag{34}$$

This holds since conditioned on the final dither Θ, the quantized endpoint is a discrete random variable that is a deterministic function of $\widetilde{X}_{n-1}$ and conditioning reduces entropy. But Z^q conditioned on the driving noise W_0^{n-2} is just the Δ-precision quantization of λ^{n-1} times a uniform random variable of width Δ and hence has discrete entropy $\geq \log_2 \lambda^{n-1}$.

B. Upper bounding $I(\widetilde{X}_0^{n-1}; Z^q|\Theta)$

To upper-bound this term, Lemma B.1 can be used to see

$$\begin{aligned} I(\widetilde{X}_0^{n-1}; Z^q|\Theta) &= H(Z^q|\Theta) \\ &< 7 + \frac{\log_2 K'}{\eta} + 2\log_2 \frac{\log_2 K'}{\eta} + \log_2 \frac{1}{\Delta} + 2\log_2 \log_2 \frac{1}{\Delta} + \frac{5+\ln 2}{\eta \ln 2} \end{aligned} \tag{35}$$

where K' is an upper-bound to $E[|Z|^\eta]$. Such an upper-bound is readily available since

$$\begin{aligned} E[|Z|^\eta] &= E[|\widetilde{X}_n|^\eta] \\ &\leq E[(\frac{\Delta}{2} + |\sum_{i=0} n-1 \lambda^{n-1-i} W_i|)^\eta] \\ &= E[(\frac{\Delta}{2} + \lambda^{n-1}|\sum_{i=0} n-1 \lambda^{-i} W_i|)^\eta] \\ &= \lambda^{\eta(n-1)} E[(\frac{\Delta}{2\lambda^{n-1}} + |\sum_{i=0} n-1 \lambda^{-i} W_i|)^\eta] \\ &\leq \lambda^{\eta(n-1)} E[(2\max(\frac{\Delta}{2\lambda^{n-1}}, |\sum_{i=0} n-1 \lambda^{-i} W_i|)^\eta] \\ &< \lambda^{\eta(n-1)} (\frac{\Delta^\eta}{\lambda^{\eta(n-1)}} + 2^\eta K). \end{aligned}$$

Using this for K' and taking logs shows

$$\begin{aligned} \frac{\log_2 K'}{\eta} &= \frac{\log_2(\lambda^{\eta(n-1)}(\frac{\Delta^\eta}{\lambda^{\eta(n-1)}} + 2^\eta K))}{\eta} \\ &= 1 + (n-1)\log_2 \lambda + \frac{\log_2(\frac{\Delta^\eta}{\lambda^{\eta(n-1)}} + 2^\eta K)}{\eta} \end{aligned}$$

Substituting this in gives the desired bound

$$\begin{aligned} & I(\widetilde{X}_0^{n-1}; Z^q|\Theta) \\ < \; & 8 + (n-1)\log_2 \lambda + 2\log_2(1 + (n-1)\log_2 \lambda + \frac{\log_2(\frac{\Delta^\eta}{\lambda^{\eta(n-1)}} + 2^\eta K)}{\eta}) + \log_2 \frac{1}{\Delta} + 2\log_2 \log_2 \frac{1}{\Delta} + \frac{5+\ln 2}{\eta \ln 2} \end{aligned} \tag{36}$$

There is only a single $O(n)$ term above, and it is $(n-1)\log_2 \lambda$. Everything else is $o(n)$.

C. Lower bounding $I(\widetilde{X}_0^{n-1}; Y_0^{n-1}, Z^q|\Theta)$

We need to establish

$$I(\widetilde{X}_0^{n-1}; Y_0^{n-1}, Z^q|\Theta) \geq n(R_\infty^X(d) - \epsilon_2). \tag{37}$$

This is immediately obvious from

$$\begin{aligned} I(\widetilde{X}_0^{n-1}; Y_0^{n-1}, Z^q|\Theta) &= I(\widetilde{X}_0^{n-1}; Y_0^{n-1}|\Theta) + H(Z^q|\Theta, Y_0^{n-1}) - H(Z^q|\Theta, Y_0^{n-1}, \widetilde{X}_0^{n-1}) \\ &= I(\widetilde{X}_0^{n-1}; Y_0^{n-1}|\Theta) + H(Z^q|\Theta, Y_0^{n-1}) \\ &\geq n(R_\infty^X(d) - \epsilon_2). \end{aligned}$$

The first equality is just expanding the mutual information and recognizing the fact that Z^q is discrete once conditioned on the dither Θ and so H is the regular discrete entropy here. Let $Q_{(\Delta,\Theta)}$ denote the dithered scalar quantizer used to generate the encoded checkpoints, just appropriately translated so it can apply to the $\widetilde{X}$ giving $Z^q = Q_{(\Delta,\Theta)}(\widetilde{X}_{n-1})$. The next equality is a consequence of this deterministic relationship. Finally, the discrete entropy is always positive and can be dropped to give a lower bound.

D. Upper bounding $I(\widetilde{X}_0^{n-1}; Y_0^{n-1}, Z^q|\Theta)$

The second term of (33) is upper bounded in a way similar to the first term. We need to establish

$$I(\widetilde{X}_0^{n-1}; Y_0^{n-1}, Z^q|\Theta) \leq n(R_\infty^X(d) + \epsilon_2) + o(n). \tag{38}$$

Expand the mutual information as before

$$\begin{aligned} I(\widetilde{X}_0^{n-1}; Y_0^{n-1}, Z^q|\Theta) &= I(\widetilde{X}_0^{n-1}; Y_0^{n-1}|\Theta) + H(Z^q|\Theta, Y_0^{n-1}) \\ &\leq I(\widetilde{X}_0^{n-1}; Y_0^{n-1}|\Theta) + H(Z^q|\Theta, Y_{n-1}) \\ &= n(R_\infty^X(d) + \epsilon_2) + H(Z^q - Q_{(\Delta,\Theta)}(Y_{n-1})|\Theta, Y_{n-1}) \\ &\leq n(R_\infty^X(d) + \epsilon_2) + \log_2 3 + H(Q_{(\Delta,\Theta)}(\widetilde{X}_{n-1} - Y_{n-1})|\Theta). \end{aligned}$$

The first inequality comes from dropping conditioning. After that, the quantizer $Q_{(\Delta,\Theta)}$ can be applied to Y_{n-1} so that $Z^q - Q_{(\Delta,\Theta)}(Y_{n-1}) = S\Delta$ where S is an integer-valued random variable representing how many steps up or down the Δ-quantization ladder are needed to get from $Q_{(\Delta,\Theta)}(Y_{n-1})$ to Z^q. The difference of two quantized numbers differs by at most 1 quantization bin from the quantization of the difference. This slack of up to 1 bin in either direction can be encoded using $\log_2 3$ bits.

At this point, Lemma B.1 applies using the trivial upper bound $n(d + \epsilon_3)$ for the η-th moment of $\widetilde{X}_{n-1} - Y_{n-1}$, since the worst case is for the entire distortion to fall on the last component of the vector.

$$\begin{aligned} & H(Q_{(\Delta,\Theta)}(\widetilde{X}_{n-1} - Y_{n-1})|\Theta) \\ < \; & 7 + \frac{\log_2 n(d+\epsilon_3)}{\eta} + 2\log_2 \frac{\log_2 n(d+\epsilon_2)}{\eta} + \log_2 \frac{1}{\Delta} + 2\log_2 \log_2 \frac{1}{\Delta} + \frac{5 + \ln 2}{\eta \ln 2} \end{aligned}$$

The $\log_2 n$ term is certainly $o(n)$. The only other term that might raise concern is $\log_2 \frac{1}{\Delta}$, but that is $o(n)$ since (13) tells us that we are already required to choose n much larger than that to have R_1 close to $\log_2 \lambda$ in the first stream. The order of limits is to always let n go to infinity before Δ goes to zero.

E. Putting pieces together

With (38) established, it can be applied along with (34) to (33) and gives

$$I(\widetilde{X}_0^{n-1}; Y_0^{n-1}|Z^q, \Theta) \leq n(R_\infty^X(d) - \log_2 \lambda + \epsilon_2) + o(n). \tag{39}$$

Taking n to ∞ and dividing through by n establishes the desired result on the upper bound.

Similarly putting together (36) and (37) gives

$$I(\widetilde{X}_0^{n-1}; Y_0^{n-1}|Z^q, \Theta) \geq n(R_\infty^X(d) - \log_2 \lambda - \epsilon_2) - o(n). \tag{40}$$

Taking n to ∞ and dividing through by n establishes the desired result on the lower bound.

But ϵ_2 was arbitrary and this establishes the desired result. □

APPENDIX D
PROOF OF THEOREM 5.1

Interpret the random ensemble of infinite tree codes as a single code with both encoder and decoder having access to the common-randomness used to generate the code-tree. Populate the tree with iid channel inputs drawn from the distribution that achieves $E_r(R)$ for block codes. Theorem 7 in [26] tells us that the code achieves anytime reliability $\alpha = E_r(R)$ since the analysis uses the same infinite ensemble for all i and delays.

Alternatively, this can be seen from first principles for ML decoding by observing that any false path $\widetilde{B}_1^i$ can be divided into a true prefix B_1^{j-1} and a false suffix $\widetilde{B}_j^i$. The iid nature of the channel inputs on the code tree tells us that the true code-suffix corresponding to the received channel outputs from time $\frac{j}{R}$ to t is independent of any false code-suffix. Since there are $\leq 2^{R(t-\frac{j}{R})}$ such false code-suffixes (ignoring integer effects) at depth j, Gallager's random block-coding analysis from [8] applies since all that it requires is pairwise independence between true and false codewords.

$$\begin{aligned}
& \mathcal{P}(\widehat{B}_j(t) \neq B_j | B_1^{j-1} \text{ already known}) \\
\leq\ & \mathcal{P}(\text{error on random code with } 2^{R(t-\frac{j}{R})} \text{ words and block length } t - \lceil \frac{j}{R} \rceil) \\
\leq\ & 2^{-(t-\lceil \frac{j}{R} \rceil)E_r(R)} \\
\leq\ & 2^{-(t-\frac{j}{R}-1)E_r(R)}
\end{aligned}$$

The probability of error on B_1^i can be bounded by the union bound over $j = 1 \ldots i$.

$$\begin{aligned}
\mathcal{P}(\widehat{B}_1^i(t) \neq B_1^i) &\leq \sum_{j=1}^{i} \mathcal{P}(\widehat{B}_j(t) \neq B_j | B_1^{j-1} \text{ already known}) \\
&\leq \sum_{j=1}^{i} 2^{-(t-\frac{j}{R}-1)E_r(R)} \\
&< \sum_{j=0}^{\infty} 2^{-(t-\frac{i}{R}-j-1)E_r(R)} \\
&= K 2^{-(t-\frac{i}{R})E_r(R)}
\end{aligned}$$

The exponent for the probability of error is dominated by the shortest codeword length in the union bound, and this corresponds to $t - \frac{i}{R}$. □

APPENDIX E
PROOF OF PROPERTY 6.1

$$\begin{aligned}
|X'_{kn} - \bar{X}'_{kn}| &\geq \lambda^{n(k-j)}\beta(|M_j - \bar{M}_j| - \left|\sum_{i=j+1}^{\infty} \lambda^{-n(i-j)}(M_i - \bar{M}_i)\right|) \\
&\geq \lambda^{n(k-j)}\beta(|M_j - \bar{M}_j| - 2\lambda^{-n}\sum_{i=0}^{\infty}\lambda^{-ni}) \\
&\geq \lambda^{n(k-j)}\beta(2^{1-nR} - 2\frac{\lambda^{-n}}{1-\lambda^{-n}}) \\
&= \lambda^{n(k-j)}2\beta(2^{-nR} - \frac{1}{\lambda^n - 1})
\end{aligned}$$

which is positive as long as $2^{-nR} > \frac{1}{\lambda^n - 1}$ or $nR < \log_2(\lambda^n - 1)$. We can thus use $K = 2\beta(2^{-nR} - \frac{1}{\lambda^n - 1}) = \frac{\delta}{\lambda}(2^{n(\log_2 \lambda - R)} - \frac{\lambda^n}{\lambda^n - 1})$ and the property is proved. □

ACKNOWLEDGMENTS

The authors would like to give special thanks to Mukul Agarwal for the discussions regarding this paper and the resulting contributions in [19] where one of the key results required is proved. We also thank Nigel Newton, Nicola Elia, and Sekhar Tatikonda for several constructive discussions about this general subject over a long period of time which have influenced this work in important ways.

REFERENCES

[1] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, Jul./Oct. 1948.
[2] ——, "Coding theorems for a discrete source with a fidelity criterion," *IRE National Convention Record*, vol. 7, no. 4, pp. 142–163, 1959.
[3] S. Vembu, S. Verdu, and Y. Steinberg, "The source-channel separation theorem revisited," *IEEE Trans. Inf. Theory*, vol. 41, no. 1, pp. 44–54, Jan. 1995.
[4] A. Ephremides and B. Hajek, "Information theory and communication networks: An unconsummated union," *IEEE Trans. Inf. Theory*, vol. 44, pp. 2416–2434, Oct. 1998.
[5] A. Sahai and S. K. 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. Inf. Theory*, vol. 52, no. 8, pp. 3369–3395, Aug. 2006.
[6] ——, "The necessity and sufficiency of anytime capacity for stabilization of a linear system over a noisy communication link. part II: vector systems," *IEEE Trans. Inf. Theory*, submitted for publication. [Online]. Available: http://www.eecs.berkeley.edu/˜sahai/Papers/control-part-II.pdf
[7] A. Sahai, "Why block-length and delay behave differently if feedback is present," *IEEE Trans. Inf. Theory*, Submitted. [Online]. Available: http://www.eecs.berkeley.edu/˜sahai/Papers/FocusingBound.pdf
[8] R. G. Gallager, *Information Theory and Reliable Communication*. New York, NY: John Wiley, 1971.
[9] R. Gray, "Information rates of autoregressive processes," *IEEE Trans. Inf. Theory*, vol. 16, no. 4, pp. 412–421, Jul. 1970.
[10] T. Hashimoto and S. Arimoto, "On the rate-distortion function for the nonstationary autoregressive process," *IEEE Trans. Inf. Theory*, vol. 26, no. 4, pp. 478–480, Apr. 1980.
[11] T. Berger, "Information rates of Wiener processes," *IEEE Trans. Inf. Theory*, vol. 16, no. 2, pp. 134–139, Mar. 1970.
[12] ——, *Rate Distortion Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1971.
[13] A. Sahai, "A variable rate source-coding theorem for unstable scalar markov processes," in *Proceedings of the 2001 IEEE Symposium on Information Theory*, Washington, DC, Jun. 2001.
[14] ——, "Any-time information theory," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, 2001.
[15] S. K. Mitter and N. J. Newton, "Information flow and entropy production in the Kalman-Bucy filter," *Journal of Statistical Physics*, vol. 118, no. 1, pp. 145–175, Jan. 2005.
[16] A. Sahai and S. K. Mitter, "A fundamental need for differentiated 'quality of service' over communication links: An information theoretic approach," in *Proceedings of the Allerton Conference on Communication, Control, and Computing*, Monticello, IL, Oct. 2000.
[17] J. E. Hopcroft and J. D. Ullman, *Introduction to Automata Theory, Languages, and Computation*. Reading, MA: Addison-Wesley, 1979.
[18] M. Agarwal, A. Sahai, and S. K. Mitter, "Coding into a source: A direct inverse rate-distortion theorem," in *Proceedings of the Allerton Conference on Communication, Control, and Computing*, Monticello, IL, Sep. 2006.
[19] ——, "A direct equivalence perspective on the separation theorem," *IEEE Trans. Inf. Theory*, In preparation.
[20] B. Rimoldi, "Successive refinement of information: Characterization of the achievable rates," *IEEE Trans. Inf. Theory*, vol. 40, no. 1, pp. 253–259, Jan. 1994.
[21] S. Verdu and T. S. Han, "A general formula for channel capacity," *IEEE Trans. Inf. Theory*, vol. 40, no. 4, pp. 1147–1157, Jul. 1994.
[22] S. Tatikonda, "Control under communication constraints," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, 2000.
[23] S. Tatikonda, A. Sahai, and S. K. Mitter, "Stochastic linear control over a communication channel," *IEEE Trans. Autom. Control*, vol. 49, no. 9, pp. 1549–1561, Sep. 2004.
[24] R. Zamir and M. Feder, "On universal quantization by randomized uniform/lattice quantizers," *IEEE Trans. Inf. Theory*, vol. 38, no. 2, pp. 428–436, Mar. 1992.
[25] B. M. Leiner and R. M. Gray, "Rate-distortion theory for ergodic sources with side information," *IEEE Trans. Inf. Theory*, vol. 20, no. 5, pp. 672–675, Sep. 1974.
[26] G. D. Forney, "Convolutional codes II. maximum-likelihood decoding," *Information and Control*, vol. 25, no. 3, pp. 222–266, Jul. 1974.
[27] ——, "Convolutional codes III. sequential decoding," *Information and Control*, vol. 25, no. 3, pp. 267–297, Jul. 1974.
[28] F. Jelinek, "Upper bounds on sequential decoding performance parameters," *IEEE Trans. Inf. Theory*, vol. 20, no. 2, pp. 227–239, Mar. 1974.
[29] M. S. Pinsker, "Bounds on the probability and of the number of correctable errors for nonblock codes," *Problemy Peredachi Informatsii*, vol. 3, no. 4, pp. 44–55, Oct./Dec. 1967.
[30] A. Sahai and T. Simsek, "On the variable-delay reliability function of discrete memoryless channels with access to noisy feedback," in *Proceedings of the IEEE Workshop on Information Theory*, San Antonio, TX, Nov. 2004.
[31] S. Draper and A. Sahai, "Noisy feedback improves communication reliability," in *Proc. IEEE International Symposium on Information Theory*, Jul. 2006. [Online]. Available: http://www.eecs.berkeley.edu/˜sahai/Papers/DraperSahaiISIT06.pdf
[32] ——, "Variable-length coding with noisy feedback," *European Transactions on Telecommunications*, Submitted.

[33] Y. Liang, A. Goldsmith, and M. Effros, "Distortion metrics of composite channels with receiver side information," in *Proceedings of the 2007 Information Theory Workshop*, Sep. 2007, pp. 559–564.
[34] P. Moulin and J. A. O'Sullivan, "Information-theoretic analysis of information hiding," *IEEE Trans. Inf. Theory*, vol. 49, no. 3, pp. 563–593, Mar. 2003.
[35] A. Sahai and S. K. Mitter, "Source coding and channel requirements for unstable processes," *IEEE Trans. Inf. Theory*, Submitted, 2006. [Online]. Available: http://www.eecs.berkeley.edu/˜sahai/Papers/anytime.pdf
[36] G. N. Nair and R. J. Evans, "Communication-limited stabilization of linear systems," in *Proceedings of the 39th IEEE Conference on Decision and control*, Sydney, Australia, Dec. 2000, pp. 1005–1010.
[37] ——, "Stabilizability of stochastic linear systems with finite feedback data rates," *SIAM Journal on Control and Optimization*, vol. 43, no. 2, pp. 413–436, Jul. 2004.
[38] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.