

Exact minimum number of bits to stabilize a linear system

Victoria Kostina, Yuval Peres, Gireeja Ranade, Mark Sellke

Abstract—We consider an unstable scalar linear stochastic system, $X_{n+1} = aX_n + Z_n - U_n$, where $a \geq 1$ is the system gain, Z_n 's are independent random variables with bounded α -th moments, and U_n 's are the control actions that are chosen by a controller who receives a single element of a finite set $\{1, \dots, M\}$ as its only information about system state X_i . We show new proofs that $M > a$ is necessary and sufficient for β -moment stability, for any $\beta < \alpha$. Our achievable scheme is a uniform quantizer of the zoom-in / zoom-out type that codes over multiple time instants for data rate efficiency; the controller uses its memory of the past to correctly interpret the received bits. We analyze its performance using probabilistic arguments. We show a simple proof of a matching converse using information-theoretic techniques. Our results generalize to vector systems, to systems with dependent Gaussian noise, and to the scenario in which a small fraction of transmitted messages is lost.

Index Terms—Linear stochastic control, source coding, data rate theorem.

I. INTRODUCTION

We study the tradeoff between stabilizability of a linear stochastic system and the coarseness of the quantizer used to represent the state. The evolution of the system is described by

$$X_{n+1} = aX_n + Z_n - U_n, \quad (1)$$

where constant $a \geq 1$; X_1 and Z_1, Z_2, \dots are independent random variables with bounded α -th moments, and U_n is the control action chosen based on the history of quantized observations. More precisely, an M -bin causal quantizer-controller for X_1, X_2, \dots is a sequence $\{f_n, g_n\}_{n=1}^{\infty}$, where $f_n: \mathbb{R}^n \mapsto [M]$ is the encoding (quantizing) function, and $g_n: [M] \mapsto \mathbb{R}^n$ is the decoding (controlling) function, and $[M] \triangleq \{1, 2, \dots, M\}$. At time i , the controller outputs

$$U_n = g_n(f_1(X_1), f_2(X^2), \dots, f_n(X^n)). \quad (2)$$

V. Kostina (vkostina@caltech.edu) is with California Institute of Technology, Pasadena, CA. Y. Peres (yuval@yuvalperes.com) is an independent researcher. G. Ranade (ranade@eecs.berkeley.edu) is with the University of California, Berkeley, CA. M. Sellke (msellke@stanford.edu) is with Stanford University, CA. This work was supported in part by the National Science Foundation (NSF) under Grant CCF-1751356, and by the Simons Institute for the Theory of Computing. Research of Y. Peres was partially supported by NSF grant DMS-1900008. G. Ranade acknowledges the Siebel Energy Institute Seed Funding.

The fundamental operational limit of quantized control of interest in this paper is the minimum number of quantization bins to achieve β -moment stability:

$$M_\beta^* \triangleq \min \left\{ M: \exists M\text{-bin causal quantizer-controller} \right. \\ \left. \text{s.t. } \limsup_n \mathbb{E} [|X_n|^\beta] < \infty \right\}, \quad (3)$$

where $0 < \beta < \alpha$ is fixed.

The main results of the paper are new proofs of the following achievability and converse theorems, whose various special cases have been previously shown in literature.

Theorem 1 (achievability). *Let X_1, Z_n in (1) be independent random variables with bounded α -moments. Then for any $0 < \beta < \alpha$*

$$M_\beta^* \leq \lfloor a \rfloor + 1. \quad (4)$$

Theorem 2 (converse). *Let X_1, Z_n in (1) be independent random variables. Let $h(X_1) > -\infty$, where $h(X) \triangleq -\int_{\mathbb{R}} f_X(x) \log f_X(x) dx$ is the differential entropy. Then, for all $\beta > 0$,*

$$M_\beta^* \geq \lfloor a \rfloor + 1. \quad (5)$$

The first achievability results [1], [2] focused on unstable scalar systems with bounded disturbances, i.e. $|Z_n| \leq B$ a.s., and showed that a simple uniform quantizer with the number of quantization bins in (4) stabilizes such systems. That corresponds to the special case $\alpha = \beta = \infty$. Nair and Evans [3] showed that time-invariant fixed-rate quantizers are unable to attain bounded cost if the noise is unbounded [3], regardless of their rate. The reason is that since the noise is unbounded, over time, a large magnitude noise realization will inevitably be encountered, and the dynamic range of the quantizer will be exceeded by a large margin, not permitting recovery. This necessitates the use of adaptive quantizers of zooming type [4]–[6]. Such quantizers “zoom out” (i.e. expand their quantization intervals) when the system is far from the target and “zoom in” when the system is close to the target. Nair and Evans [3] constructed such an adaptive fixed-length quantizer with nonuniform quantization levels and showed second-moment stability via a recursive bound on its mean-squared error, under the assumption that the system noise has bounded $2 + \epsilon$ moment, for some $\epsilon > 0$. Under the same assumption, Yüksel [7] (see [8], [9] for generalizations to vector systems and to $\beta = 1, 2, \dots$) showed second-moment stabilizability using a uniform scalar quantizer that enters its

zoom-out mode whenever its input falls outside its dynamic range. When applied to encode each k -th system state over the following k time instances, the schemes in [3], [7] attain (4) for a large enough k . See also [10], which explores the use of constrained quantizers to encode the overflow event over multiple time instances.

The converse in the special case of $\beta = 2$ was proved in [3], where it was shown that it is impossible to achieve second moment stability in the system in (1) using a quantizer-controller with the number of bins $< \lfloor a \rfloor + 1$. This implies the validity of Theorem 2 for $\beta \geq 2$. Variants of the necessity result in Theorem 2 are known for vector systems with bounded disturbances [11] and model uncertainty [12]; for noiseless vector systems under different stability criteria [13]; for vector linear stochastic systems under second moment constraint stabilized over rate-constrained noiseless [3] and packet-drop [14] channels; for vector linear stochastic systems stabilized in probability over noisy channels [15]; and for nonlinear systems with additive noise stabilized in probability over noisy channels [16, Th. 3.1].

In this paper, we construct a new zoom-in zoom-out scheme that most of the time operates as if the noise were bounded, and relies on a periodic magnitude test to determine whether the state has left the quantized region. Similar to an application of the schemes in [3], [7] to an undersampled system with the transmission of a codeword over multiple time slots mentioned above, our strategy uses coding over multiple time instants, and the controller uses its memory of the past to correctly interpret the received bits. While the controller in the above modification of the known schemes is almost always silent, producing a large signal once in k time instances, our controller is almost always active, producing a control signal optimized for bounded noise. Thus it introduces less delay. If the periodic magnitude test is failed, the quantizer-controller enters the zoom-out mode, which is essentially the same as in [7]: the controller looks for the X_n in exponentially larger intervals until it is located, at which point it returns to the zoom-in mode. We provide an elementary analysis of our scheme with an explicit bound on k leading to Theorem 1.

We also present a short proof of the converse result in Theorem 2 that uses information-theoretic arguments. We also provide an elementary converse proof for stabilizability in probability that is tight for non-integer a .

In Section II, we describe our achievable scheme and give its analysis. In Section III, we give a proof of the converse in Theorem 2. Our results generalize to constant-length time delays, to control over communication channels that drop a small fraction of packets, to systems with dependent Gaussian noise, and to vector systems. These extensions are presented in Section IV. This is a full version of the conference paper [17]. This paper presents full proofs (the proofs in [17] are either omitted or replaced with proof outlines), and a much more comprehensive discussion of our results and their relationship to past and future research. In addition, Theorem 3, which presents a converse using an elementary

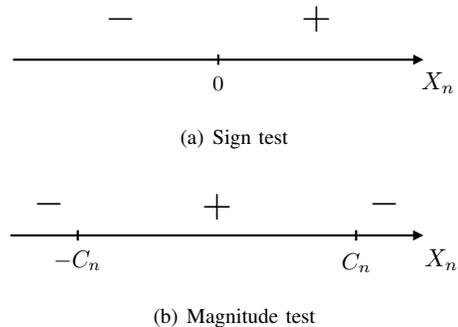


Fig. 1. The binary quantizer uses two kinds of tests on a schedule determined by the previous \pm 's to produce the next $+$ or $-$.

probabilistic argument, is not contained in [17].

II. ACHIEVABLE SCHEME

A. The idea

Here we explain the idea of our achievable scheme. For readability we focus on the case $a \in [1, 2)$ and show that the system can be controlled with 1 bit. In this case we will be able to restrict to two types of tests, a *sign test* and a *magnitude test* (see Fig. 1), which simplifies our procedure. The straightforward extension to an arbitrary $a \geq 1$, in which the sign test is replaced by a uniform quantizer, is found in Section II-E below.

In the case of bounded noise a uniform time-invariant quantizer deterministically keeps X_n bounded [1], [2]. Indeed, when $|Z_n| \leq B$, $n = 1, 2, \dots$ and $|X_1| \leq C_1$, if $C_1 \geq \frac{B}{1-a/2}$ one can put

$$C_2 \triangleq (a/2)C_1 + B \leq C_1, \quad (6)$$

and putting further $C_{n+1} \triangleq (a/2)C_n + B$, we obtain a monotonically decreasing to $\frac{B}{1-a/2}$ sequence numbers $\{C_n\}_{n=1}^{\infty}$. Setting

$$U_n = (a/2)C_n \operatorname{sgn}(X_n) \quad (7)$$

requires only 1 bit of knowledge about X_n (i.e., its sign). If $|X_n| \leq C_n$ then

$$|X_{n+1}| \leq (a/2)C_n + B = C_{n+1}, \quad (8)$$

and

$$\limsup_{n \rightarrow \infty} |X_n| \leq \frac{B}{1-a/2}. \quad (9)$$

Actually, this is the best achievable bound on the uncertainty about the location of X_n , as a simple volume-division argument shows [11], [18].

When Z_n merely have bounded α -moments the above does not work because a single large value of Z_n will cause the system to explode. However we can use the idea of the bounded case with the following modification. Most of the time, in *normal*, or *zoom-in*, mode, the controller assumes the X_n are bounded by constants C_n and forms the control

actions according to the above procedure, but occasionally, on a schedule, the quantizer performs a *magnitude test* and sends a bit whose sole purpose is to inform the controller whether the X_n is staying within desired bounds. If the test is passed, the controller continues in the normal mode, and otherwise, it enters the *emergency*, or *zoom-out*, mode, whose purpose is to look for the X_n in exponentially larger intervals until it is located, at which point it returns to the zoom-in mode while still occasionally checking for anomalies. We will show that all this can be accomplished with only 1 bit per controller action.

The intuition behind our scheme is the following. At any given time, with high probability X_n is not too large. Thus, the emergencies are rare, and when they do occur, the size of the uncertainty region tends to decrease exponentially. The zoom-in mode operates almost exactly as in the bounded case, except that we choose B large enough to diminish the probability that the noise exceeds it. We now proceed to making these intuitions precise in Section II-B.

B. The Algorithm

Here we describe the algorithm precisely and then prove that it works. Specifically, we consider the setting of Theorem 1 with $a \in [1, 2)$ and Z_n with bounded α -moments. We find U_n - a function only of the sequence bits received from the quantizer - that achieves β -moment stability, for $0 < \beta < \alpha$.

First we prepare some constants. We fix $B \geq 1$ large enough. We set the *probing factor* $P = P(\alpha, \beta)$ - a large positive constant (how large will be explained below, but roughly P blows up as $\beta \uparrow \alpha$). Fix a small $\delta > 0$ and a large enough $k = k(a)$ so that

$$(a/2)^{k-1}a \leq 1 - 3\delta. \quad (10)$$

We proceed in “rounds” of at least $k + 1$ moves, k moves in normal (zoom-in) mode and $k + 1$ ’th move to test whether X_n escaped the desired bounds. If that magnitude test comes back normal, the round ends; otherwise the controller enters the emergency (zoom-out) mode, whose duration is variable and which ends once the controller learns a new (larger) bound on X_n . In normal mode, we use the update rule in (7), where $C_n \geq B$ is positive. In the emergency mode, $U_n \equiv 0$ while C_n grows exponentially. A precise description of the operation of the algorithm is given below.

- 1) At the start of a round at time-step m , $|X_m| \leq C_m$, the controller is silent, $U_m = 0$, and $X_{m+1} = aX_m + Z_m$. Set

$$C_{m+1} = aC_m + B, \quad (11)$$

and for each $i \in \{2, \dots, k\}$,

$$C_{m+i} = \frac{a}{2}C_{m+i-1} + B \quad (12)$$

$$= (a/2)^{i-1}C_{m+1} + \frac{1 - (a/2)^{i-1}}{1 - a/2}B. \quad (13)$$

In this normal mode operation, the quantizer sends a sequence of signs of X_n (see Fig. 1(a)), while the controller applies the controls (7) successively to X_m, \dots, X_{m+k-1} . This normal mode operation will keep X_{m+i} bounded by C_{m+i} unless some Z_{m+i} is atypically large.

- 2) The quantizer applies the magnitude test to check whether $|X_{m+k}| \leq C_{m+k}$ (see Fig. 1(b)). If $|X_{m+k}| \leq C_{m+k}$, we return to step 1. If $|X_{m+k}| > C_{m+k}$, this means some Z_{m+i} was abnormally large; the system has blown up and we must do damage control. In this case we enter *emergency (zoom-out) mode* in Step 3 below.
- 3) In emergency mode, we repeatedly perform silent ($U_{m+k+j} \equiv 0$) magnitude tests via

$$C_{m+k+j} = P C_{m+k+j-1} = P^j C_{m+k} \quad j \geq 0 \quad (14)$$

until the first time τ that the magnitude test is passed, i.e.

$$\tau \triangleq \inf \{j \geq 0: |X_{m+k+j}| \leq C_{m+k+j}\}. \quad (15)$$

We then set $m \leftarrow m + k + \tau$ and return to Step 1.

The controller is silent at the start of a round because it does not know the sign of X_m . Each round thus includes one silent step at the start, and $\tau \geq 0$ silent steps of the emergency mode.

C. Overview of the Analysis

We analyze the result of each round. At the start of each round m we know that X_m is contained within interval $[-C_m, C_m]$. We will show that when C_m is large, the uncertainty interval tends to decrease by a constant factor each round.

At the start of the round, $|X_m| \leq C_m$. Assume that for each $i \in \{0, 1, \dots, k\}$, we have

$$|Z_{m+i}| \leq B. \quad (16)$$

and thus

$$|X_{m+i}| \leq C_{m+i}. \quad (17)$$

In particular, applying (10), (11) and (12), we bound the state at the end of the round as

$$\begin{aligned} |X_{m+k}| &\leq C_{m+k} \\ &\leq (1 - 3\delta)C_m + \frac{B}{1 - a/2}, \end{aligned} \quad (18)$$

which means that $C_{m+k} \leq C_m$, provided that $C_m \geq \frac{B}{3\delta(1-a/2)}$. Thus, even starting with the silent step we have successfully decreased C_m , provided that it was large enough.

What if (16) fails to hold? Because the Z_i have bounded α -moments, by the union bound and Markov’s inequality, the chance (16) fails is at most

$$\mathbb{P} \left[\bigcup_{i=0}^k \{|Z_{m+i}| > B\} \right] \leq (k+1) \mathbb{E}[|Z|^\alpha] B^{-\alpha}. \quad (20)$$

In this case, we show that we can control the blow-up to avert a catastrophe. Recall that in emergency mode our

procedure will take exponentially growing C_n (see (14)) so that we will soon observe that $|X_n| \leq C_n$. The controller then exits emergency mode and returns to the normal mode, starting a new round at time step n . Using boundedness of α -moments of Z_i , we will show in Section II-D below that the chance that on step $n = m + k + j$ this *fails* is exponentially small in j . We will see that in each round starting at $X_m \in [-C_m, C_m]$, there is a high chance to shrink the magnitude of the state and a small chance to grow larger. In the next section we explain how to obtain precise moment control.

D. Precise Analysis

Here we give details of the analysis outlined in Section II-C, demonstrating that when the Z_n are i.i.d. with bounded α -moments, our strategy in Section II-B yields

$$\limsup_n \mathbb{E}[|X_n|^\beta] < \infty \quad (21)$$

for all $0 < \beta < \alpha$.

The following tools will be instrumental in controlling the tails of the accumulated noise.

Proposition 1. *If the random variable Z has finite α -moment, then*

$$t^\alpha \mathbb{P}[|Z| > t] \quad (22)$$

are bounded in t . Conversely, if (22) are bounded in t then Z has a finite β -moment for any $0 < \beta < \alpha$.

Proof. The first part is the Markov inequality. The second is a standard use of the tail-sum formula. \square

Lemma 1. *Suppose $a > 1$ is fixed and Z_i are (arbitrarily coupled) random variables with uniformly bounded absolute α moments. Then the random variables*

$$\tilde{Z}_j \triangleq \sum_{i=0}^j a^{-i} Z_i \quad (23)$$

also have uniformly bounded absolute α -moments.

Proof. It is easy to see that for any $\alpha > 0$, $\varepsilon > 0$ there is $c = c_{\alpha, \varepsilon}$ such that for all

$$(x + y)^\alpha \leq c_{\alpha, \varepsilon} x^\alpha + (1 + \varepsilon) y^\alpha \quad (24)$$

holds for all $x, y \geq 0$. Indeed, to see this, assume without loss of generality that $x = 1$, and note that when y is sufficiently large we already have

$$(1 + y)^\alpha \leq (1 + \varepsilon) y^\alpha. \quad (25)$$

The set of y for which (25) does not hold is bounded, hence so is the value of $(1 + y)^\alpha$; take c to be an upper bound for this expression. The equation will now hold for any value of y .

Applying (24) repeatedly yields

$$|\tilde{Z}_k|^\alpha \leq c|Z_0|^\alpha + c \sum_{i=1}^{k-1} (1 + \varepsilon)^i a^{-\alpha i} |Z_i|^\alpha + (1 + \varepsilon)^k a^{-\alpha k} |Z_k|^\alpha. \quad (26)$$

Since $\mathbb{E}[|Z_i|^\alpha]$ are uniformly bounded and for $1 + \varepsilon < a^\alpha$ the geometric series $\sum_{i=1}^{j-1} (1 + \varepsilon)^i a^{-\alpha i}$ converges, $\mathbb{E}[|\tilde{Z}_j|^\alpha]$ is bounded uniformly in j , as desired. \square

Remark 1. The mild assumptions of Lemma 1 will make it easy to generalize our results to dependent noise in Section IV-C below.

The bound in Lemma 2 below considers the evolution of the system over $k + 1 + \tau$ steps, where τ (15) determines the end of the round. Note that τ is a stopping time of the filtration generated by $\{X_n\}$.

Lemma 2. *Fix $B, P > 0$ and consider our algorithm described in Section II-B with these parameters. Suppose that time-step m is the start of a round, so that the round ends on time-step $m + k + \tau$. For all $1 < a < 2$ and for all $0 \leq j \leq \tau$, it holds that*

$$\begin{aligned} & \max\{|X_{m+1}|, \dots, |X_{m+k+j}|, C_{m+k+j}\} \\ & \leq Pa^{k+j} \left(2C_m + \frac{aB}{(2-a)(a-1)} + \sum_{\ell=0}^{k+j-1} a^{-\ell-1} |Z_{m+\ell}| \right), \end{aligned} \quad (27)$$

Proof. Appendix. \square

Proof of Theorem 1 for the case $a \in [1, 2)$. To avoid a special treatment of the case $a = 1$, we assume that $a > 1$. This is without loss because showing stability for a implies stability for all $a' \leq a$. First we prepare some constants. Recall the choices of k and δ in (10).

- Fix $\Delta < \alpha - \beta$ an arbitrary fixed constant, e.g. $\Delta = \frac{\alpha - \beta}{3}$, so that

$$\beta = \alpha - 3\Delta. \quad (28)$$

- Fix P large enough so that

$$P/a \geq \max \left\{ \left(\frac{a}{1 - \delta} \right)^{\alpha - \Delta}, 2^k, \frac{a^{k+1}}{2(a-1)} \right\}. \quad (29)$$

Suppose that time-step m is the start of a round, so that the round ends on time-step $m + k + \tau$, with stopping time $\tau = 0$ usually.

We define a modified sequence¹ \tilde{X}_n through, for $1 \leq i \leq k + \tau$,

$$\begin{aligned} \tilde{X}_{m+i} & \triangleq \left(\frac{1}{1 - \delta} \right)^{\tau - |i - k|_+} \\ & \max\{|X_{m+k}|, \dots, |X_{m+k+\tau}|, C_{m+k+\tau}\}, \end{aligned} \quad (30)$$

¹ \tilde{X}_n serves as a Lyapunov function that stochastically controls the growth of the state process X_n . See [9, Th. 2.1], [19] for a general approach to proving stability using Lyapunov functions for general Markov chains.

where $|\cdot|_+ \triangleq \max\{0, \cdot\}$. Clearly this definition ensures that

$$|X_{m+k+j}| \leq \tilde{X}_{m+k+j} \quad 0 \leq j \leq \tau. \quad (31)$$

Furthermore, for all $1 \leq i \leq k-1$, there exists universal constants K_1, K_2, K_3 that depend on a, k and B such that (Appendix A)

$$\mathbb{E}[|X_{m+i}|^\beta] \leq K_1 \mathbb{E}[\tilde{X}_{m+k}^\beta] + K_2 \mathbb{E}[\tilde{X}_m^\beta] + K_3. \quad (32)$$

Inequalities (31) and (32) together mean that to establish (21), it is sufficient to prove

$$\limsup_n \mathbb{E}[\tilde{X}_n^\beta] < \infty. \quad (33)$$

The rest of the proof is focused in establishing (33).

By definition (30),

$$\tilde{X}_{m+i} \leq \tilde{X}_{m+1} \quad i = 2, \dots, k + \tau, \quad (34)$$

with equality for $i \leq k$.

We will show that

$$\mathbb{E}[\tilde{X}_{m+1}^\beta] \leq (1-\delta)^\beta \mathbb{E}[\tilde{X}_m^\beta] + K, \quad (35)$$

where $K = K(P, k, \delta)$ is a constant that may depend on P, k, δ (but is independent of m). Together, inequalities (34) and (35) ensure that $\limsup_n \mathbb{E}[\tilde{X}_n^\beta]$ is bounded above by $\frac{K}{1-(1-\delta)^\beta}$.

The intuition behind the definition for \tilde{X}_n is as follows. We want to construct a dominating sequence \tilde{X}_n with the expected decrease property in (35). During emergency mode, the original sequence X_n may increase on average during rounds. The sequence \tilde{X}_n in (30) takes the potential increase during each round up front, achieving the desired expected decrease property. We will see that P in (29) is chosen so that the constant-factor decrease of the system is preserved when switching between rounds.

To show (35), we define the filtration \mathcal{F}_n as follows: \mathcal{F}_n is the σ -algebra generated by the sequences Z_1, Z_2, \dots, Z_{n-1} and $\tilde{X}_1, \tilde{X}_2, \dots, \tilde{X}_n$. Unless n is the end of a round, knowledge of \tilde{X}_n involves a peek into the future, so \mathcal{F}_n encompasses slightly more information than the naive notion of ‘‘information up to time n ’’. The inequality we will show, clearly stronger than (35), is

$$\mathbb{E}[\tilde{X}_{m+1}^\beta | \mathcal{F}_m] \leq (1-\delta)^\beta \mathbb{E}[\tilde{X}_m^\beta | \mathcal{F}_m] + K. \quad (36)$$

Define

$$Y_n \triangleq \frac{\tilde{X}_{n+1}}{\tilde{X}_n + \frac{B}{(1-a/2)(1-3\delta)}}. \quad (37)$$

We will show (36) by the means of the following two statements, where m is the transition between rounds:

(a) For sufficiently large k and P in (10) and (29), respectively, it holds that²

$$\mathbb{P}[Y_m \geq t | \mathcal{F}_m] = O\left(t^{-(\alpha-\Delta)}\right), \quad (38)$$

²Throughout this section, the implicit constants $O(\cdot)$ may depend on P, k, δ (but are independent of n and $B \geq 1$).

(b) As $B \rightarrow \infty$,

$$\mathbb{P}[Y_m \leq 1 - 3\delta | \mathcal{F}_m] \rightarrow 1. \quad (39)$$

We use (38) and (39) to show (36) as follows. First, observe that by (38) and Proposition 1, $\{Y_m | \mathcal{F}_m\}$ has bounded $\beta + \Delta$ -moment since we assumed (28) when choosing Δ . Furthermore, since the right side of (38) is independent of \mathcal{F}_m , the $\beta + \Delta$ -moment of Y_m is bounded uniformly in m . Now, pick $p > 1$ so that $\beta p \leq \beta + \Delta$, and let q satisfy $\frac{1}{p} + \frac{1}{q} = 1$. Write

$$\mathbb{E}[Y_m^\beta | \mathcal{F}_m] \leq (1-3\delta)^\beta + \mathbb{E}[Y_m^\beta \mathbf{1}\{Y_m > 1-3\delta\} | \mathcal{F}_m] \quad (40)$$

$$\leq (1-3\delta)^\beta + (\mathbb{E}[Y_m^{\beta p} | \mathcal{F}_m])^{\frac{1}{p}} (\mathbb{P}[Y_m > 1-3\delta | \mathcal{F}_m])^{\frac{1}{q}} \quad (41)$$

$$\rightarrow (1-3\delta)^\beta, \quad B \rightarrow \infty, \quad (42)$$

where (41) is by Hölder’s inequality, and the second term in (41) vanishes as $B \rightarrow \infty$ due to (39) and uniform boundedness of the $\beta + \Delta$ -moment of $\{Y_m | \mathcal{F}_m\}$. Note that convergence in (42) is uniform in m . It follows that for a large enough B (how large depends on the values of P, k, δ),

$$\mathbb{E}[Y_m^\beta | \mathcal{F}_m] \leq (1-2\delta)^\beta. \quad (43)$$

Rewriting (43) using (37) yields

$$\mathbb{E}[\tilde{X}_{m+1}^\beta | \mathcal{F}_m] \leq (1-2\delta)^\beta \left(\tilde{X}_m + \frac{B}{(1-a/2)(1-3\delta)} \right)^\beta \quad (44)$$

$$\leq (1-\delta)^\beta \tilde{X}_m^\beta + K, \quad (45)$$

where to write (45) we used (24). This establishes the inequality (36).

To complete the proof of Theorem 1, it remains to establish (38) and (39).

To show (38), recall that the round ends at stopping time $m+k+\tau$. Since the events $\{\tau=j\}$ are disjoint, we have

$$\begin{aligned} \mathbb{P}[Y_m \geq t | \mathcal{F}_m] &= \sum_{j=0}^{\infty} \mathbb{P}[Y_m \geq t, \tau=j | \mathcal{F}_m] \\ &\quad + \mathbb{P}[Y_m \geq t, \tau=\infty | \mathcal{F}_m] \end{aligned} \quad (46)$$

Note that since m is the end of the previous round, \mathcal{F}_m does not contain any information about the future.

We estimate the probability of the event in $\mathbb{P}[Y_m \geq t, \tau=j | \mathcal{F}_m]$ in two ways, and use the better estimate on each term individually.

We express the system state at time $m+i$ in terms of the system state at time m :

$$X_{m+i} = a^i \left(X_m + \sum_{\ell=0}^{i-1} a^{-\ell-1} U_{m+\ell} + \sum_{\ell=0}^{i-1} a^{-\ell-1} Z_{m+\ell} \right). \quad (47)$$

Using (7), (11), (12) and recalling that $U_m = 0$, we can crudely bound the cumulative effect of controls on X_{m+i} as

$$a^i \left| \sum_{\ell=0}^{k-1} a^{-\ell-1} U_{m+\ell} \right| \leq a^i (a/2) \sum_{\ell=1}^{\infty} a^{-\ell-1} \quad (48)$$

$$\begin{aligned} & \left((a/2)^{\ell-1} C_{m+1} + \frac{1 - (a/2)^{\ell-1}}{1 - a/2} B \right) \\ &= a^i \left(C_m + \frac{B}{a-1} \right). \end{aligned} \quad (49)$$

Recalling the definitions of \tilde{X}_n, Y_n in (30), (37), respectively, and invoking Lemma 2, we see that if $\{Y_m \geq t, \tau = j\}$ holds, then

$$\begin{aligned} & t(1-\delta)^{k+j-1} \left(\tilde{X}_m + \frac{B}{(1-a/2)(1-3\delta)} \right) \\ & \leq P a^{k+j} \left(2C_m + \frac{aB}{(2-a)(a-1)} + \sum_{\ell=0}^{k+j-1} a^{-\ell-1} |Z_{m+\ell}| \right). \end{aligned} \quad (50)$$

Noting that both C_m and $\frac{aB}{2-a}$ are dominated by $\tilde{X}_m + \frac{B}{(1-a/2)(1-3\delta)} \geq 1$, we can weaken (50) as

$$t(1-\delta)^{k+j-1} \leq P a^{k+j} \left(2 + \frac{1}{a-1} + \sum_{\ell=0}^{k+j-1} a^{-\ell-1} |Z_{m+\ell}| \right). \quad (51)$$

Applying Lemma 1 and Proposition 1, we deduce that the probability of the event in (51) is

$$O \left(\left(\frac{a}{1-\delta} \right)^{\alpha j} t^{-\alpha} \right). \quad (52)$$

The bound in (52) works well for small j / large t . For large j / small t , we observe that $\{Y_m \geq t, \tau = j\} \subseteq \{\tau \geq j\}$ and apply the following reasoning. The event $\{\tau \geq j\}$ means that the emergency did not end at time j ; in other words,

$$|X_{m+k+j-1}| > C_{m+k+j-1} \quad (53)$$

$$= P^j \left(2(a/2)^k C_m + \frac{B}{1-a/2} \right), \quad (54)$$

where to write (54) we used (11), (13), and (14). Substituting $i \leftarrow k+j$ into (47) and recalling (49) and $|X_m| \leq C_m$, we weaken (53)–(54) as

$$\begin{aligned} & a^{k+j} \left(2C_m + \frac{aB}{(2-a)(a-1)} + \sum_{\ell=0}^{k+j-1} a^{-\ell-1} |Z_{m+\ell}| \right) \\ & > P^j \left(2(a/2)^k C_m + \frac{B}{1-a/2} \right), \end{aligned} \quad (55)$$

the event equivalent to

$$\begin{aligned} & (a/P)^j \sum_{\ell=0}^{k+j-1} a^{-\ell-1} |Z_{m+\ell}| \geq 2 \left((1/2)^k - (a/P)^j \right) C_m \\ & + \left((1/a)^k - \frac{a(a/P)^j}{2(a-1)} \right) \frac{B}{1-a/2}. \end{aligned} \quad (56)$$

Due to the choice of P in (29), the coefficients in front of C_m and B in the right side of (56) are nonnegative for all $j \geq 1$. Bounding the probability of the event in (56) using Lemma 1 and Proposition 1, we conclude that ³

$$\mathbb{P}[\tau \geq j] = O \left((P/a)^{-j\alpha} \right). \quad (57)$$

Furthermore, (57) means that $\mathbb{P}[\tau = \infty] = 0$. Indeed, $1\{\tau = \infty\} = \prod_{j=0}^{\infty} 1\{\tau \geq j\} = \lim_{j \rightarrow \infty} 1\{\tau \geq j\}$ and by Fatou's lemma,

$$\mathbb{P}[\tau = \infty] \leq \lim_{j \rightarrow \infty} \mathbb{P}[\tau \geq j] = 0, \quad (58)$$

thus the corresponding term can be eliminated from (46).

Juxtaposing (52) and (57), we conclude that the probability $\mathbb{P}[Y_m \geq t, \tau = j | \mathcal{F}_m]$ is bounded by

$$O \left(\min \left\{ \left(\frac{a}{1-\delta} \right)^{\alpha j} t^{-\alpha}, (P/a)^{-j\alpha} \right\} \right). \quad (59)$$

Since (29) ensures that $(P/a)^\Delta \geq \left(\frac{a}{1-\delta} \right)^\alpha$, we weaken (59) as

$$O \left((P/a)^{j\Delta} \min \left\{ t^{-\alpha}, (P/a)^{-j\alpha} \right\} \right). \quad (60)$$

Recall that we have fixed t and are varying j ; this upper bound peaks at j such that $(P/a)^j = t$ at the value $t^{-(\alpha-\Delta)}$ and decays geometrically on each side at rates $(P/a)^\Delta$ and $(P/a)^{\alpha-\Delta}$. Hence the sum of all $\mathbb{P}[Y_m \geq t, \tau = j | \mathcal{F}_m]$ terms in (46) is bounded by the maximum up to a constant factor and therefore (38) holds.

To complete the proof of Theorem 1, it remains to establish (39). By Markov's inequality (20), with probability converging to 1 as $B \rightarrow \infty$, all terms Z_m, \dots, Z_{m+k} are within $[-B, B]$, and $\tau = 0$. In such a case, applying (19) and recalling (30), we get

$$\tilde{X}_{m+1} = \max\{|X_{m+k}|, C_{m+k}\} \quad (61)$$

$$\leq (1-3\delta) \tilde{X}_m + \frac{B}{1-a/2}, \quad (62)$$

which implies that $Y_m \leq 1-3\delta$, establishing (39). \square

E. Finer Quantization

For $a \geq 2$, the controller receives an element of an $\lfloor a \rfloor + 1$ -element set instead of a single bit. In this case we restrict our attention to *order-statistic* tests, meaning that we split the real line into $\lfloor a \rfloor + 1$ intervals

$$(-\infty, w_{1,n}), [w_{1,n}, w_{2,n}), \dots, [w_{\lfloor a \rfloor, n}, \infty), \quad (63)$$

and the controller receives the index $b_n \in \{0, 1, \dots, \lfloor a \rfloor\}$ of the interval containing X_n . The only real issue is for the quantizer and the controller to agree upon a rule for updating the values of w_i . However, this is easy; in the obvious generalization of our algorithm to higher a , the

³Similar exponential bounds to the event $\mathbb{P}[\tau \geq j]$ are provided in [9, Lem. 5.2] and in [16, Lem. 5.2].

(uniform) quantizer simply breaks up the interval $[-C_n, C_n]$ into $\lfloor a \rfloor + 1$ equal parts, where C_n is the same bound on the state magnitude as before. Both quantizer and controller follow the rules in (12) (with $a/2$ replaced by $a/(\lfloor a \rfloor + 1)$ and in (14) to update C_n . During the normal mode, the controller applies the control

$$U_n = -C_n + C_n \frac{2b_n + 1}{\lfloor a \rfloor + 1} \quad (64)$$

which reduces to (7) when $\lfloor a \rfloor = 1$.

In the case $a < 1$, the controller does nothing, which by Lemma 1 achieves β -moment stability.

III. CONVERSE

In this section, we prove the converse result in Theorem 2 using information-theoretic arguments similar to those employed in [3], [20]. Then, we use elementary probability to show an alternative converse result, which implies Theorem 2 unless a is an integer.

Proof of Theorem 2. Conditional entropy power is defined as

$$N(X|U) \triangleq \frac{1}{2\pi e} \exp(2h(X|U)) \quad (65)$$

where $h(X|U) = -\int_{\mathbb{R}} f_{X,U}(x, u) \log f_{X|U=u}(x) dx$ is the conditional differential entropy of X .

Conditional entropy power is bounded above in terms of moments (e.g. [21, Appendix 2]):

$$N(X) \leq \kappa_\beta \mathbb{E} [|X|^\beta]^{\frac{2}{\beta}} \quad (66)$$

$$\kappa_\beta \triangleq \frac{2}{\pi e} \left(e^{\frac{1}{\beta}} \Gamma \left(1 + \frac{1}{\beta} \right) \beta^{\frac{1}{\beta}} \right)^2, \quad (67)$$

Thus,

$$\kappa_\beta \mathbb{E} [|X_n|^\beta]^{\frac{2}{\beta}} \geq N(X_n) \quad (68)$$

$$\geq N(X_n|U^{n-1}), \quad (69)$$

where (69) holds because conditioning reduces entropy. Next, we show a recursion on $N(X_n|U^{n-1})$:

$$N(X_n|U^{n-1}) = N(A X_{n-1} + Z_{n-1}|U^{n-1}) \quad (70)$$

$$\geq a^2 N(X_{n-1}|U^{n-1}) + N(Z_{n-1}) \quad (71)$$

$$\geq a^2 N(X_{n-1}|U^{n-2}) \exp(-2r) + N(Z_{n-1}), \quad (72)$$

where (71) is due to the conditional entropy power inequality:⁴

$$N(X + Y|U) \geq N(X|U) + N(Y|U), \quad (73)$$

which holds as long as X and Y are conditionally independent given U , and (72) is obtained by weakening the constraint $|U_{n-1}| \leq M$ to a mutual information constraint $I(X_{n-1}; U_{n-1}|U^{n-2}) \leq \log M = r$ and observing that

$$\min_{P_{U|X}: I(X;U) \leq r} h(X|U) \geq h(X) - r. \quad (74)$$

⁴Conditional EPI follows by convexity from the unconditional EPI first stated by Shannon [22] and proved by Stam [23].

It follows from (72) that $r > \log a$ is necessary to keep $N(X_n|U^{n-1})$ bounded. Due to (69), it is also necessary to keep β -th moment of X_n bounded. \square

Consider the following notion of stability.⁵

Definition 1. *The system is stabilizable in probability if there exists a control strategy such that for some bounded interval \mathcal{I} ,*

$$\limsup_{n \rightarrow \infty} \mathbb{P}[X_n \in \mathcal{I}] > 0. \quad (75)$$

As a simple consequence of Markov's inequality, if the system is moment-stable, it is also stable in probability. Therefore the following converse for stability in probability implies a converse for moment stability.

Theorem 3. *Assume that X_1 has a density. To achieve stability in probability, $M \geq \lfloor a \rfloor$ is necessary.*

Proof. We want to show that for any bounded interval \mathcal{I} , if $r < \log a$ then

$$\limsup_{n \rightarrow \infty} \mathbb{P}[X_n \in \mathcal{I}] = 0. \quad (76)$$

At first we assume that the density of X_1 is bounded, that is, $f_{X_1}(x) \leq f_{\max}$ and that X_1 is supported on a finite interval, i.e. $|X_1| \leq x_{\max}$, for some constants f_{\max}, x_{\max} .

Since we are showing a converse (impossibility) result, we may relax the operational constraints by revealing the noises Z_n , $n = 1, 2, \dots$ noncausally to both encoder and decoder. Since then the controller can simply subtract the effect of the noise, we may put $Z_n \equiv 0$ in (1). Then, $X_{n+1} = a^n X_1 + \tilde{U}_n$, where $\tilde{U}_n \triangleq \sum_{i=0}^{n-1} a^{n-i} U_i$ is the combined effect of n controls, which can take one of M^n values, i.e. $U_n = u(m)$ if $X_1 \in \mathcal{I}_m$, $m = 1, \dots, M^n$. Regardless of the particular choice of control actions $u(m)$ and quantization intervals \mathcal{I}_m , for any bounded interval \mathcal{I} ,

$$\mathbb{P}[X_{n+1} \in \mathcal{I}] = \mathbb{P}[a^n X_1 + \tilde{U}_n \in \mathcal{I}] \quad (77)$$

$$= \sum_{m=1}^{M^n} \mathbb{P}[a^n X_1 + u(m) \in \mathcal{I}, X_1 \in \mathcal{I}_m] \quad (78)$$

$$\leq M^n a^{-n} f_{\max} |\mathcal{I}|, \quad (79)$$

and (76) follows for any $M < a$, confirming the necessity of $M \geq \lfloor a \rfloor$ to achieve weak stability.⁶

Finally, if the density of X_1 is unbounded, consider the set $\mathcal{S}_b \triangleq \{x \in \mathbb{R}: f_{X_1}(x) \leq b\}$ and notice that since

⁵A more stringent to Definition 1 notion of stability in probability, in which > 0 in the right side of (75) is replaced by $= 1$, was considered in [15, Def. 2.1], and in [16, Th. 3.1].

⁶An argument similar to (77)–(79) showing that the Lebesgue measure of \mathcal{I} cannot be sustained is the key to the data-rate theorems for invariance entropy [24]–[27].

$1\{f_{X_1}(x) > b\} \rightarrow 0$ pointwise as $b \rightarrow \infty$, by dominated convergence theorem,

$$\mathbb{P}[X_1 \in \mathcal{S}_b] = \int_{\mathbb{R}} f_{X_1}(x) 1\{f_{X_1}(x) > b\} dx \rightarrow 0 \text{ as } b \rightarrow \infty. \quad (80)$$

Therefore for any $\epsilon > 0$, one can pick $b > 0$ such that $\mathbb{P}[X_1 \in \mathcal{S}_b] \leq \epsilon$. Then, since we already proved that (76) holds for bounded f_{X_1} , we conclude

$$\limsup_{n \rightarrow \infty} \mathbb{P}[X_n \in \mathcal{I}] \leq \epsilon + \limsup_{n \rightarrow \infty} \mathbb{P}[X_n \in \mathcal{I} | X_1 \in \mathcal{S}_b] = \epsilon, \quad (81)$$

which implies that (76) continues to hold for unbounded f_{X_1} . \square

The quantities $\lceil a \rceil$ and $\lfloor a \rfloor + 1$ coincide unless a is an integer, thus Theorem 3 shows that for non-integer a , the converse (impossibility) part of Theorem 1 continues to hold in the sense of weak stability. Note that the proof of Theorem 3 relaxes the causality requirement. We conjecture that $\lceil a \rceil$ can be replaced by $\lfloor a \rfloor + 1$ in its statement, but proving that will require bringing causality back in the picture, and the simple argument in the proof of Theorem 3 will not work.

We conclude Section III with a technical remark.

Remark 2. The assumptions in Theorem 3 are weaker than those in Theorem 2, because the differential entropy of X_1 not being $-\infty$ implies that X_1 must have a density. The assumption that X_1 must have a density is not superficial. For example, consider $Z_i \equiv 0$ and X_1 uniformly distributed on the Cantor set, and $a = 2.9$. Clearly this system can be stabilized with 1 bit, by telling the controller at each step the undeleted third of the interval the state is at. This is lower than the result of Theorem 1, which states that M_β^* would be 3 if X_1 had a density. Beyond distributions with densities, we conjecture that M_β^* will depend on the Hausdorff dimension of the probability measure of X_1 .

IV. GENERALIZATIONS

In this section, we generalize our results in several directions. In most cases we only outline the mild differences in the proof.

A. Constant-Length Time Delays

Many systems have a finite delay in feedback. To model this, we can force U_n to depend on only the feedback up to round $n - \ell$, i.e.

$$U_n = g_n(f_1(X_1), f_2(X^2), \dots, f_{n-\ell}(X^{n-\ell})), \quad (82)$$

where $f_n(X^n)$ is the quantizer's output at time n , as before.

We argue here that this makes no difference in terms of the minimum number of bits required for stability. We state the modified result next.

Theorem 4. *Let X_1, Z_n in (1) be independent random variables with bounded α -moments. Assume that $h(X_1) > -\infty$.*

The minimum number of quantization points to achieve β -moment stability, for any $0 < \beta < \alpha$ and with any constant delay ℓ is given by $\lfloor a \rfloor + 1$.

Proof. The problem here is that the encoder sees the system before the controller can act on it. However, if we also delay the encoder seeing the system by ℓ time steps, then we can directly use the algorithm we have already constructed. Specifically, if our artificially delayed sequence of system states is $\{\tilde{X}_n\}$, then the real sequence is given by

$$X_n = a^\ell \tilde{X}_n + a^{\ell-1} Z_{n+1} + \dots + Z_{n+\ell}. \quad (83)$$

By Theorem 1, we can keep $\mathbb{E}[|\tilde{X}_n|^\beta]$ bounded for any $\beta < \alpha$ by applying to \tilde{X}_n the control in (64) in normal mode and $U_n = 0$ in the emergency mode. Since with delay, the controller acts on X_n rather than on \tilde{X}_n , we multiply the control action (64) by a^ℓ to achieve the same effect on \tilde{X}_n as without delay. Furthermore, each Z_i has bounded β moment, so by Lemma 1 their sum will have bounded β -moment, as desired.

The converse is obvious as even with $\ell = 0$, Theorem 2 asserts that the system cannot be stabilized with fewer than $\lfloor a \rfloor + 1$ bits. \square

B. Packet drops

Suppose that the encoder cannot send information to the controller at all time-steps. Instead, the encoder can only send information at a deterministic set $\mathcal{T} \subseteq \mathbb{N}$ of times. Formally,

$$U_n = g_n(\{f_n(X^n) : n \in \mathcal{T}\}). \quad (84)$$

As long as the density of \mathcal{T} is high enough on all large, constant-sized scales, the same results go through.

Definition 2. *A set $\mathcal{T} \subseteq \mathbb{N}$ is strongly p -dense if there exists N such that for all n we have*

$$\frac{|n+i : n+i \in \mathcal{T}, i=0, \dots, N-1|}{N} > p. \quad (85)$$

Note that the constant delay scenario in Section IV-A amounts to control on a strongly p -dense set, with $p \in [0, 1]$ as close to 1 as desired.

Theorem 5. *Let X_1, Z_n in (1) be independent random variables with bounded α -moments. Assume that $h(X_1) > -\infty$. The minimum number of quantization points to achieve β -moment stability is $\lfloor a \rfloor + 1$, for any $0 < \beta < \alpha$ and on any strongly p -dense set with some $p \in [0, 1]$ large enough so that*

$$(\lfloor a \rfloor + 1)^p > a. \quad (86)$$

Proof. The requirement (86) ensures that the bounded case works; indeed, it is equivalent to

$$\left(\frac{a}{\lfloor a \rfloor + 1}\right)^p a^{1-p} < 1, \quad (87)$$

which means that the logarithm of the range of X_n decreases on average each time-step.

In the unbounded noise case, we perform the same basic algorithm, but ensuring that the normal mode has enough times in \mathcal{T} , so the duration of the normal mode gets longer if N is large. Likewise, in the emergency mode, it will take longer to catch blow-ups. However, a much weaker condition on \mathcal{T} suffices for the emergency mode to end: even if \mathcal{T} contains only 1 element out of every N , we make the probe factor P large enough depending on N . The difference between probing every N time steps vs. every time step at most a factor of a^N which is a constant. \square

C. Dependent Noise

Here we address a modification in which the noise is correlated rather than independent.

Proposition 2. *Suppose $\{Z_n\}$ is a Gaussian process whose covariance matrix M (for any number of samples) has spectrum bounded by λ . Then there is an independent Gaussian process $\{Z'_n\}$ such that the random variables $\{Z_n + Z'_n\}$ are i.i.d. Gaussians with variances λ^2 .*

Proof. Just make the covariances matrices of $\{Z_n\}$ and $\{Z'_n\}$ add to λI ; the assumption means both are positive semidefinite, hence define Gaussian processes. \square

If the rows of M have ℓ^1 norm at most λ , the assumption of Proposition 2 that M has spectrum bounded by λ will be satisfied. Indeed, we can add a positive semidefinite matrix to such an M to obtain a diagonal matrix with each entry at most λ : if M has an entry of x at positions (i, j) and (j, i) then we add $|x|$ to the (i, i) and (j, j) entries; it is easy to see that the symmetric matrix

$$\begin{pmatrix} |x| & -x \\ -x & |x| \end{pmatrix} \quad (88)$$

is always positive semidefinite. Therefore doing this for all non-diagonal entries adds a positive semidefinite matrix to M and still results in a spectrum contained in $[0, \lambda]$, meaning M also had spectrum contained in $[0, \lambda]$.

Theorem 6. *The results in Theorems 1, 4, 5 extend to the case when $\{Z_n\}$ is correlated Gaussian noise whose covariance matrix has bounded spectrum.*

Proof. As a result of Proposition 2, for any Gaussian noise with known covariance matrix M of bounded spectrum, the controller can simply add extra noise to the system via U_n to effectively make the noise i.i.d. Gaussian, reducing this scenario to the i.i.d. case. \square

D. Vector systems

The results generalize to higher dimensional systems

$$X_{n+1} = AX_n + Z_n - BU_n, \quad (89)$$

where A is a $d \times d$ matrix and Z_n, U_n are vectors. The dimensionality of control signals U_n can be less than d , in which case B is a tall matrix.

For the controls to potentially span the whole space \mathbb{R}^d when combined with the multiplication-by- A amplification, the range of

$$[B, AB, A^2B, \dots, A^{d-1}B] \quad (90)$$

needs to span \mathbb{R}^d . Such a pair (A, B) of matrices is commonly referred to as *controllable*. (The $0, \dots, d-1$ powers of A are sufficient in (90), because by the Cayley-Hamilton theorem any higher power of A is a linear combination of those lower powers). For our results to hold, a weaker condition suffices, namely, we need *stabilizability* of (A, B) , which is to say that only unstable modes need be controllable. More precisely, in the *canonical* representation of a linear system,

$$\begin{bmatrix} X_{n+1}^u \\ X_{n+1}^s \end{bmatrix} = \begin{bmatrix} A^u & A^s \\ 0 & A^s \end{bmatrix} \begin{bmatrix} X_n^u \\ X_n^s \end{bmatrix} + Z_n - \begin{bmatrix} B^u \\ 0 \end{bmatrix} U_n, \quad (91)$$

where the matrix A^s has all stable eigenvalues, the state coordinates X_{n+1}^s cannot be reached by the control U_n . Stabilizability means that the pair (A^u, B^u) is controllable, which ensures that unstable modes can be controlled.

The idea behind our generalization to the vector case, previously explored in e.g. [3], is that we can decompose \mathbb{R}^d into eigenspaces of A and rotate attention between these parts.

Theorem 7. *Consider the stochastic vector linear system in (89) with (A, B) stabilizable. Let X_1, Z_n be independent random \mathbb{R}^d -valued random vectors with bounded α -moments. Assume that $h(X_1) > -\infty$. Let $(\lambda_1, \dots, \lambda_d)$ be the eigenvalues of A , and set*

$$a \triangleq \prod_{j=1}^d \max(1, |\lambda_j|). \quad (92)$$

Then for any $0 < \beta < \alpha$, the minimum number of quantization points to achieve β -moment stability is

$$M_\beta^* = \lfloor a \rfloor + 1. \quad (93)$$

Proof. We first consider the case $U_n \in \mathbb{R}^d, B = I$:

$$X_{n+1} = AX_n + Z_n - U_n, \quad (94)$$

and then explain how to deal with the general stabilizable system in (89).

By using a real Jordan decomposition, we can block-diagonalize A into

$$A = \bigoplus_j A_j \quad (95)$$

where $A_j: \mathcal{V}_j \rightarrow \mathcal{V}_j$ with

$$\bigoplus_j \mathcal{V}_j = \mathbb{R}^d \quad (96)$$

such that:

- 1) The spectrum of each A_j is either a single real λ_j (possibly with multiplicity) or a pair of complex numbers $\lambda_j, \bar{\lambda}_j$ (with equal multiplicity).
- 2) The spectral norm of A_j^k is $|\lambda_j|^k k^{O(1)}$.

This decomposition splits any vector X_n into a sum $X_n = \sum_j X_{n,j}$. Each $X_{n,j}$ individually satisfies a control equation with matrix A_j , and we will control these separately. Indeed, if

$$\sup_n \mathbb{E}[|X_{n,j}|^\beta] < \infty \quad (97)$$

for all j , then we get the desired result. (Note that we do *not* need to assume that the noise Z_n behaves independently on each subspace; getting from separate moment bounds on $X_{n,j}$ to a moment bound on X_n does not require any sort of independence.)

If $|\lambda_j| < 1$, we can leave that subspace alone; we will have (97) without doing anything (by Lemma 1). If $|\lambda_j| \geq 1$ we will act on this subspace at times in $\mathcal{T}_j \subseteq \mathbb{N}$ where \mathcal{T}_j is strongly p_j -dense for some p_j (Definition 2) with

$$(\lfloor a \rfloor + 1)^{p_j} > |\lambda_j|^{\dim(\mathcal{V}_j)}. \quad (98)$$

The assumption (92) precisely means that we can pick such p_j with $\sum_j p_j < 1$. Generating a partition of \mathbb{N} into strongly p_j -dense sets \mathcal{T}_j is simple if this constraint holds.

Now we are left to explain how to handle the problem on each \mathcal{V}_j separately. If $\dim(\mathcal{V}_j) = 1$, then we have done this before. The key point is the second property above on the growth of the spectral norm of A_j^k ; for fixed, large enough k , this growth is slow enough that we can use the same procedure, and the non-trivial Jordan blocks won't matter.

Now we proceed as before in rounds of k steps, except that the encoder sends everything at the end of a round rather than bit-by-bit (we can do this by introducing a constant amount of delay). At the end of k steps, assuming that $X_{n,j}$ started in some ball $B_C(0)$ at the start of the round for C large, the ending value $X_{n+k,j}$ will with high probability be contained in a ball $B_{C'}(0)$ with C' given by (for some $\epsilon > 0$)

$$C' = (|\lambda_j| + \epsilon)^{\left(\frac{k p_j}{\dim \mathcal{V}_j}\right)} C \quad (99)$$

assuming that no noise term was very large (the ‘‘high probability’’ is independent of j by another use of Lemma 1 - note that the high-dimensionality doesn't matter since the Euclidean norm is subadditive).

We also recall that any set in \mathcal{S}_j of diameter D can be covered by at most $O\left(\left(\frac{D}{r}\right)^{\dim \mathcal{S}_j}\right)$ balls of radius r . Hence for large enough k , we can cover $B_{C'}(0)$ with $O\left(\left(\lfloor a \rfloor + 1 - \epsilon\right)^{k p_j}\right)$ balls of radius $\left(\frac{(|\lambda_j| + \epsilon)C}{\lfloor a \rfloor + 1 - \epsilon}\right)$.

The upshot of this is that at the end of a round of length k , assuming no blow-up happened, the encoder has enough bandwidth to point to one of many balls of smaller radius than the starting ball and assert that $X_{n,j}$ is now inside that ball. Hence, typical behavior of the system will reduce the radius of $X_{n,j}$ by a constant factor each round.

Emergency mode proceeds in the same way as before, using balls of larger and larger size. The effect is still that

the β -moment of the radius decreases in expectation when large, hence is bounded.

To prove the converse, we can project out stable eigenmodes of A as done in [20], and then apply a straightforward generalization of the reasoning in Section III to the resulting vector system. This converse will apply to the system with low-dimensional controls in (89), because we can always augment the matrix B to make it full rank and extend the dimension of the control signal accordingly.

To show an achievable scheme for (89), we will reduce the problem to the delayed version of (94). Although we only addressed delays in the 1-dimensional setting in Section IV-A, the exact same argument shows that delays change absolutely nothing in all dimensions. We will focus on the case of controllable (A, B) , because if (A, B) is merely stabilizable we can always ignore the uncontrollable stable part as per the canonical representation (91). We will use the spanning set of matrices to give an arbitrary control with a delay of ℓ steps, where $\ell \leq d - 1$ is such that the range of

$$[B, AB, A^2B, \dots, A^\ell B] \quad (100)$$

spans \mathbb{R}^d . Then any vector $v \in \mathbb{R}^d$ can be written as

$$v = Bv_0 + ABv_1 + \dots + A^\ell Bv_\ell, \quad (101)$$

where $v_i \in \text{ran}(A^i B)$ for each i , and $|v_i| = O(|v|)$.

Now, suppose that the sequence $\{\hat{U}_n\}_{n \in \mathbb{Z}^+}$ solves the control problem with delay ℓ , meaning that the control \hat{U}_n is chosen at time n but kicks in at time $n + \ell$. That is, $\{\hat{U}_n\}$ is chosen so that the sequence \hat{X}_n given by

$$\hat{X}_{n+1} = A\hat{X}_n + Z_n - \hat{U}_{n-\ell} \quad (102)$$

has bounded moments. We assume that the noises Z_n are common to (89) and (102), so that $\{X_n\}$ and $\{\hat{X}_n\}$ are coupled together rather than independent. Per (101) we can write

$$\hat{U}_n = B U_{0,n} + AB U_{1,n} + A^2 B U_{2,n} + \dots + A^\ell B U_{\ell,n}. \quad (103)$$

To realize control \hat{U}_n that takes full effect at time $n + \ell$, we can have \hat{U}_n contribute $B U_{i,n}$ to the control at time $n + \ell - i$. If we do this for every n , however, we see that the control $\tilde{U}_n = B U_n$ applied at time n will consist of contributions from all $\hat{U}_{n-\ell}, \dots, \hat{U}_n$:

$$U_n = \sum_{i=0}^{\ell} U_{i,n-\ell+i}, \quad (104)$$

and the actual accumulated control by time $n + \ell$ is larger than \hat{U}_n . Therefore, by applying (104) to the state of the original system X_n (89) at time n , we will not get exactly \hat{X}_n . However, the difference $\hat{X}_n - X_n$ is a finite sum of terms of type $A^{j_1} U_{j_2, \ell}$ that are bounded, according to (7), in terms of the same majorizing sequence \tilde{X}_n ((30), Lemma 2) that we used to bound the β -moment of \hat{X}_n . Since \tilde{X}_n has bounded β -moment according to (35) (although we haven't emphasized it, \tilde{X}_n is also bounded in β -moment in higher dimensions just as in the 1-dimensional case), we conclude

that the finite number of extra controls has no bearing on stabilizability. \square

V. CONCLUSION

This paper studies the minimum number of bits necessary and sufficient for stability, when fixed-rate quantizers are used, and proves that conveying $\lceil a \rceil + 1$ distinct values is necessary and sufficient to achieve β -moment stability, where a is defined in (92), provided that the independent additive noises have bounded α moments, for some $\alpha > \beta$. Theorem 1, which is the main technical result of the paper, proposes and analyses a time-varying strategy to achieve stability of a scalar system under this minimum communication requirement. We use probabilistic arguments to show this result. Theorem 2 shows a matching converse (impossibility) result, attesting that no strategy can achieve stability with a lower amount of communication. We use information-theoretic arguments to show this result. Theorem 3 relaxes the assumptions of Theorem 2, and shows, using a purely probabilistic argument, that $\lceil a \rceil$ distinct messages are necessary for stability even in the absence of additive noise. Generalizations to constant-length time delays, communication channels with packet drops, dependent noise, and vector systems are presented in Theorem 4, Theorem 5, Theorem 6, Theorem 7, respectively.

In [28], we applied a similar strategy to stabilize a system with random gain a (which is constant in the present paper) using finitely many bits at each time step.

An advantage of the scheme presented in this paper compared to [20] is that it uses a fixed number of bits at each time step, and thus is directly compatible with standard block error-correcting codes used for the transmission over noisy channels. Analyzing how our strategy can be applied together with an appropriate error-correcting code to control over noisy channels and whether fundamental limits can be attained that way is an important future research direction.

While we picked the constants to guarantee a bounded β -moment, we did not try to optimize them in order to minimize it. A natural future research direction, then, is to study, in the spirit of [20], the tradeoff between rate and the attainable β -moment. It will be interesting to see whether our scheme can approach the lower bound in [20], and to compare its performance with that of the Lloyd-Max quantizer, explored in the context of control in [29].

APPENDIX

A. Proof of (32)

For $1 \leq i \leq k$, we express the system state at time $m+k$ in terms of the system state at time $m+i$:

$$X_{m+k} = a^{k-i} \left(X_{m+i} + \sum_{\ell=0}^{k-i-1} a^{-\ell-1} U_{m+i+\ell} + \sum_{\ell=0}^{k-i-1} a^{-\ell-1} Z_{m+i+\ell} \right). \quad (105)$$

Applying (7), (11) and (13), we can crudely bound the cumulative effect of controls on X_{m+k} as

$$\left| \sum_{\ell=0}^{k-1} a^{-\ell-1} U_{m+i+\ell} \right| \leq (a/2) \sum_{\ell=1}^{\infty} a^{-\ell-1} \left((a/2)^\ell C_{m+i} + \frac{1 - (a/2)^\ell}{1 - a/2} B \right) \quad (106)$$

$$= C_{m+i} + \frac{B}{a-1} \quad (107)$$

$$\leq (a/2)^{-k} C_m + \frac{aB}{(2-a)(a-1)} \quad (108)$$

Unifying (105) and (108), we get

$$|X_{m+i}| \leq |X_{m+k}| + (a/2)^{-k} C_m + \frac{aB}{(2-a)(a-1)} + \sum_{\ell=0}^{k-i-1} a^{-\ell-1} |Z_{m+i+\ell}| \quad (109)$$

By Lemma 1, the sum of random variables on the right side of (109) has uniformly bounded α -moments, and since by definition of \tilde{X}_n in (30), $\tilde{X}_m \leq C_m$ and $|X_{m+k}| \leq \tilde{X}_{m+k}$, (32) follows by the means of (24).

B. Proof of Lemma 2

Combining (47), (49) and $|X_m| \leq C_m$ yields for $i = 1, 2, \dots, k + \tau$,

$$|X_{m+i}| \leq a^i \left(2C_m + \frac{B}{a-1} + \sum_{\ell=0}^{i-1} a^{-\ell-1} |Z_{m+\ell}| \right), \quad (110)$$

Maximizing the right side of (110) over $1 \leq i \leq k+j$ and using (110), we conclude

$$\max_{1 \leq i \leq k+j} |X_{m+i}| \leq a^{k+j} \left(2C_m + \frac{B}{a-1} + \sum_{\ell=0}^{k+j-1} a^{-\ell-1} |Z_{m+\ell}| \right), \quad (111)$$

It remains to bound C_{m+k+j} . If $j = 0$, we may simply apply (19), which means, crudely,

$$C_{m+k+j} \leq \text{right side of (111)} + \frac{a^{k+j} B}{1 - a/2}. \quad (112)$$

If $j > 0$, since the round did not end on step $m+k+j-1$, we have $C_{m+k+j-1} < |X_{m+k+j-1}|$, which means that

$$C_{m+k+j} < P |X_{m+k+j-1}|. \quad (113)$$

Combining (111), (112) and (113) yields (27).

REFERENCES

- [1] J. Baillieul, "Feedback designs for controlling device arrays with communication channel bandwidth constraints," in *ARO Workshop on Smart Structures, Pennsylvania State Univ.*, 1999, pp. 16–18.
- [2] W. S. Wong and R. W. Brockett, "Systems with finite communication bandwidth constraints. II. Stabilization with limited information feedback," *IEEE Transactions on Automatic Control*, vol. 44, no. 5, pp. 1049–1053, 1999.

- [3] G. N. Nair and R. J. Evans, "Stabilizability of stochastic linear systems with finite feedback data rates," *SIAM Journal on Control and Optimization*, vol. 43, no. 2, pp. 413–436, 2004.
- [4] A. Gersho and D. Goodman, "A training mode adaptive quantizer," *IEEE Transactions on Information Theory*, vol. 20, no. 6, pp. 746–749, 1974.
- [5] J. Kieffer and J. Dunham, "On a type of stochastic stability for a class of encoding schemes," *IEEE Transactions on Information Theory*, vol. 29, no. 6, pp. 793–797, 1983.
- [6] R. W. Brockett and D. Liberzon, "Quantized feedback stabilization of linear systems," *IEEE transactions on Automatic Control*, vol. 45, no. 7, pp. 1279–1289, 2000.
- [7] S. Yüksel, "Stochastic stabilization of noisy linear systems with fixed-rate limited feedback," *IEEE Transactions on Automatic Control*, vol. 55, no. 12, pp. 2847–2853, 2010.
- [8] A. P. Johnston and S. Yüksel, "Stochastic stabilization of partially observed and multi-sensor systems driven by unbounded noise under fixed-rate information constraints," *IEEE Transactions on Automatic Control*, vol. 59, no. 3, pp. 792–798, 2014.
- [9] S. Yüksel and S. P. Meyn, "Random-time, state-dependent stochastic drift for Markov chains and application to stochastic stabilization over erasure channels," *IEEE Transactions on Automatic Control*, vol. 58, no. 1, pp. 47–59, 2012.
- [10] O. Sabag, V. Kostina, and B. Hassibi, "Stabilizing dynamical systems with fixed-rate feedback using constrained quantizers," in *2020 IEEE International Symposium on Information Theory (ISIT)*, June 2020, pp. 2855–2860.
- [11] S. Tatikonda and S. Mitter, "Control under communication constraints," *IEEE Transactions on Automatic Control*, vol. 49, no. 7, pp. 1056–1068, 2004.
- [12] N. C. Martins, M. A. Dahleh, and N. Elia, "Feedback stabilization of uncertain systems in the presence of a direct link," *IEEE Transactions on Automatic Control*, vol. 51, no. 3, pp. 438–447, 2006.
- [13] S. Yüksel and T. Başar, "Minimum rate coding for LTI systems over noiseless channels," *IEEE Transactions on Automatic Control*, vol. 51, no. 12, pp. 1878–1887, 2006.
- [14] P. Minero, M. Franceschetti, S. Dey, and G. N. Nair, "Data rate theorem for stabilization over time-varying feedback channels," *IEEE Transactions on Automatic Control*, vol. 54, no. 2, pp. 243–255, 2009.
- [15] A. S. Matveev, "State estimation via limited capacity noisy communication channels," *Mathematics of Control, Signals, and Systems*, vol. 20, no. 1, pp. 1–35, 2008.
- [16] S. Yüksel, "Stationary and ergodic properties of stochastic nonlinear systems controlled over communication channels," *SIAM Journal on Control and Optimization*, vol. 54, no. 5, pp. 2844–2871, 2016.
- [17] V. Kostina, Y. Peres, G. Ranade, and M. Sellke, "Exact minimum number of bits to stabilize a linear system," in *Proceedings 57th IEEE Conference on Decision and Control*, Miami, FL, Dec. 2018, pp. 453–458.
- [18] B. G. N. Nair, F. Fagnani, S. Zampieri, and R. J. Evans, "Feedback control under data rate constraints: An overview," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 108–137, 2007.
- [19] S. Yüksel and T. Başar, *Stochastic networked control systems: Stabilization and optimization under information constraints*. Springer Science & Business Media, 2013.
- [20] V. Kostina and B. Hassibi, "Rate-cost tradeoffs in control," *IEEE Transactions on Automatic Control*, vol. 64, no. 11, pp. 4525–4540, Apr. 2019.
- [21] R. Zamir and M. Feder, "On universal quantization by randomized uniform/lattice quantizers," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 428–436, Mar. 1992.
- [22] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, July and October 1948.
- [23] A. J. Stam, "Some inequalities satisfied by the quantities of information of Fisher and Shannon," *Information and Control*, vol. 2, no. 2, pp. 101–112, 1959.
- [24] F. Colonius and C. Kawan, "Invariance entropy for control systems," *SIAM Journal on Control and Optimization*, vol. 48, no. 3, pp. 1701–1721, 2009.
- [25] F. Colonius, C. Kawan, and G. Nair, "A note on topological feedback entropy and invariance entropy," *Systems & Control Letters*, vol. 62, no. 5, pp. 377–381, 2013.
- [26] C. Kawan and J.-C. Delvenne, "Network entropy and data rates required for networked control," *IEEE Transactions on Control of Networked Systems*, vol. 3, no. 1, pp. 57–66, 2015.
- [27] A. Da Silva and C. Kawan, "Robustness of critical bit rates for practical stabilization of networked control systems," *Automatica*, vol. 93, pp. 397–406, 2018.
- [28] V. Kostina, Y. Peres, G. Ranade, and M. Sellke, "Stabilizing a system with an unbounded random gain using only a finite number of bits," *IEEE Transactions on Information Theory*, vol. 67, no. 4, pp. 2554–2561, Apr. 2021.
- [29] A. Khina, Y. Nakahira, Y. Su, and B. Hassibi, "Algorithms for optimal control with fixed-rate feedback," in *Proceedings 2017 IEEE Conference on Decision and Control*, Melbourne, Australia, Dec. 2017.



Victoria Kostina received the bachelor's degree from Moscow Institute of Physics and Technology (MIPT) in 2004, the master's degree from University of Ottawa in 2006, and the Ph.D. degree from Princeton University in 2013. During her studies at MIPT, she was affiliated with the Institute for Information Transmission Problems of the Russian Academy of Sciences. She is currently a Professor of electrical engineering and computing and mathematical sciences at California Institute of Technology. Her research interests include information theory, coding, control, learning, and communications. She received the Natural Sciences and Engineering Research Council of Canada postgraduate scholarship during 2009–2012, Princeton Electrical Engineering Best Dissertation Award in 2013, Simons-Berkeley research fellowship in 2015 and the NSF CAREER award in 2017.

formation theory, coding, control, learning, and communications. She received the Natural Sciences and Engineering Research Council of Canada postgraduate scholarship during 2009–2012, Princeton Electrical Engineering Best Dissertation Award in 2013, Simons-Berkeley research fellowship in 2015 and the NSF CAREER award in 2017.



Yuval Peres obtained his PhD in 1990 from the Hebrew University in Jerusalem. In 1993, he joined the faculty of the University of California at Berkeley, where he served as a professor in the mathematics and statistics departments until 2006. From 2006 to 2018, he was a Principal Researcher at Microsoft Research. He has also taught at Yale and at the Hebrew University, and is currently visiting Kent State University. Yuval Peres has published more than 300 papers with 200 co-authors and has mentored 21 PhD theses. His research encompasses most areas of probability theory, including random walks, Brownian motion, percolation, and random graphs. He has co-authored books on Markov chains, probability on graphs, game theory and Brownian motion. Dr. Peres is an IMS fellow and a recipient of the Rollo Davidson prize and the Loeve prize. In 2002, he was an invited speaker at the International Congress of Mathematicians in Beijing, and in 2016 he was elected to the National Academy of Sciences.

His research encompasses most areas of probability theory, including random walks, Brownian motion, percolation, and random graphs. He has co-authored books on Markov chains, probability on graphs, game theory and Brownian motion. Dr. Peres is an IMS fellow and a recipient of the Rollo Davidson prize and the Loeve prize. In 2002, he was an invited speaker at the International Congress of Mathematicians in Beijing, and in 2016 he was elected to the National Academy of Sciences.



Gireeja Ranade is an Assistant Teaching Professor at UC Berkeley. Before this she was a Researcher at Microsoft Research AI, Redmond. She received an MS and PhD in EECS from UC Berkeley and an SB in EECS from MIT. Her research interests have revolved around understanding stochastic systems and problems in control theory, information theory and wireless communications.



Mark Sellke is a PhD student in mathematics at Stanford advised by Sébastien Bubeck and Andrea Montanari. He graduated from MIT in 2017 with a B.S. in mathematics and from the University of Cambridge with a Masters in mathematics with distinction in 2018. He is the recipient of an NSF Graduate Fellowship and a Stanford Graduate Fellowship. Mark's primary research interests are in probability and theoretical machine learning. His work "Chasing Convex Bodies Optimally" won the best paper and best student paper awards at

SODA 2020.