

# 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  $\alpha$ -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 that  $M = \lfloor a \rfloor + 1$  is necessary and sufficient for  $\beta$ -moment stability, for any  $\beta < \alpha$ . Our achievable scheme is a uniform quantizer of zoom-in / zoom-out type whose performance is analyzed using probabilistic arguments. The matching converse is shown using information-theoretic techniques. The analysis generalizes to vector systems, to systems with dependent Gaussian noise, and to the scenario in which a small fraction of transmitted messages is lost.

## 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  $\alpha$ -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  $g_n: [M]^n \mapsto \mathbb{R}$  is the decoding (controlling) function, and  $[M] \triangleq \{1, 2, \dots, M\}$ . At time  $i$ , the controller outputs

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

The fundamental operational limit of quantized control of interest in this paper is the minimum number of quantization points to achieve  $\beta$ -moment stability:

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

V. Kostina ([vkostina@caltech.edu](mailto:vkostina@caltech.edu)) is with California Institute of Technology, Pasadena, CA. Y. Peres ([peres@microsoft.com](mailto:peres@microsoft.com)) and G. Ranade ([gireeja@eecs.berkeley.edu](mailto:gireeja@eecs.berkeley.edu)) are with Microsoft Research, Redmond, WA. M. Sellke ([msellke@gmail.com](mailto:msellke@gmail.com)) is with University of Cambridge, Cambridge, UK. 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.

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

The main result of the paper is the following theorem.

**Theorem 1.** *Let  $X_1, Z_n$  in (1) be independent random variables with bounded  $\alpha$ -moments. Then for any  $0 < \beta < \alpha$ , the minimum number of quantization points to achieve  $\beta$ -moment stability is*

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

The result of Theorem 1 is tight, as the following converse shows.

**Theorem 2.** *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  $\beta > 0$ ,*

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

In the special case of unstable scalar systems with bounded disturbances, i.e.  $|Z_n| \leq B$  a.s., the results of Theorem 1 and Theorem 2 are well known since [1], [2], where it was shown that a simple uniform quantizer with the number of quantization bins in (4) stabilizes such systems. That corresponds to the special case  $\alpha = \beta = \infty$  in Theorem 1.

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  $\alpha > 2$ .

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 originally proposed by Brockett and Liberzon [4]. 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. They are known to achieve input-to-state stability for linear systems with bounded disturbances [5]. Nair and Evans [3] proposed a stabilizing quantization scheme in which the number of quantization levels is finite at each step but varies with time, and showed that it suffices to use  $\log_2 a$  bits on average to achieve second moment stability, as long as system noise has bounded  $2 + \epsilon$  moment, for some  $\epsilon > 0$ . In this

paper, we do not allow the communication rate to vary with time: our communication channel noiselessly transmits one of  $M$  messages at each time step.

The stabilizing performance of fixed-rate quantizer-controller pairs that fit the setting of this paper was studied by Yüksel [6], who proved that for Gaussian system noises,

$$M_2^* \leq \lfloor a \rfloor + 2. \quad (6)$$

Yüksel's result leaves a gap of 1 between the upper and lower bounds. The gap might seem insignificant, especially if the gain  $a$  is large, but the gain of many realistic systems is in  $[1, 2)$ . The state of the art thus leaves open the question of whether such systems are stabilizable with a single-bit quantizer.

This paper resolves that question in the affirmative. We construct a controller that stabilizes linear systems with  $a \in [1, 2)$  while using only 1 bit per sample to choose its control action. We show that  $\beta$ -moment stability is achievable as long as system noise has bounded  $\alpha$  moment, for some  $\alpha > \beta$ . The scheme and its analysis extend naturally to higher  $a$ 's.

Note that both schemes [3], [6] rely on the special treatment of the overflow bins of the quantizer, which are its unbounded leftmost and rightmost bins. Once the quantizer overflows, the controllers of [3], [6] enter their zoom-out modes. Such controller strategies cannot be used with single bit quantizers, because single bit quantizers are always in overflow. Furthermore, as Yüksel [6] discusses, the special treatment of the overflow bin is what causes the extra 1 in (6).

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.

## 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

$$C_2 \triangleq (a/2)C_1 + B \leq C_1, \quad (7)$$

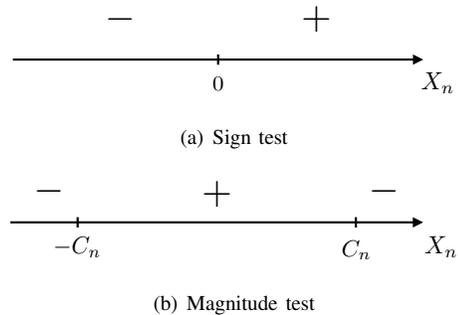


Fig. 1. The binary quantizer uses two kinds of tests on a schedule determined by the previous  $\pm$ 's to produce the next  $+$  or  $-$ .

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

$$U_n = (a/2)C_n \operatorname{sgn}(X_n) \quad (8)$$

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

$$|X_{n+1}| \leq (a/2)C_n + B = C_{n+1}, \quad (9)$$

and

$$\limsup_{n \rightarrow \infty} |X_n| \leq \frac{B}{1-a/2}. \quad (10)$$

Actually, this is the best achievable bound on the uncertainty about the location of  $X_n$ , as a simple volume-division argument shows [7].

When  $Z_n$  merely have bounded  $\alpha$ -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 runs the above procedure, but occasionally, on a schedule, the controller performs a *magnitude test* and sends a bit whose sole purpose is to test whether the  $X_n$  is staying within desired bounds. If the answer is affirmative, the controller reverts back to 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 allowed region tends to decrease exponentially. The zoom-in mode operates almost exactly as in the bounded case, except that we restrict the constants  $C_n$  from below by some (sufficiently large) constant  $C$ . When the size of the allowed region reaches  $C$ , the uncertainty interval does not shrink anymore, which prevents non-extremal values of

$Z_n$  from triggering emergencies. 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  $\alpha$ -moments. We find  $U_n$  - a function only of the sequence bits received from the quantizer - that achieves  $\beta$ -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

$$(a/2)^{k-1}a \leq 1 - 3\delta. \quad (11)$$

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 (8), 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

$$C_{m+1} = aC_m + B, \quad (12)$$

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

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

$$= (a/2)^{i-1}C_{m+1} + \frac{1 - (a/2)^{i-1}}{1 - a/2}B. \quad (14)$$

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 (8) 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

$$C_{m+k+j} = P C_{m+k+j-1} = P^j C_{m+k} \quad j \geq 0 \quad (15)$$

until the first time  $j$  that the magnitude test is passed, i.e.  $|X_{m+k+j}| \leq C_{m+k+j}$ . We then return to Step 1.

### 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, it 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 (11), (12) and (13), we bound the state at the end of the round as

$$|X_{m+k}| \leq C_{m+k} \quad (18)$$

$$\leq (1 - 3\delta) C_m + \frac{B}{1 - a/2}, \quad (19)$$

which means that  $C_{m+k} \leq C_m$ , provided that  $C_m \geq \frac{B}{3\delta(1-a/2)}$ . Thus, even after using 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  $\alpha$ -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 prevent a catastrophe. Recall that in emergency mode our procedure will take exponentially growing  $C_n$  (see (15)) 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  $\alpha$ -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  $\alpha$ -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  $\alpha$ -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  $\beta$ -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  $\alpha$  moments. Then the random variables*

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

*also have uniformly bounded absolute  $\alpha$ -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 weaken some conditions on our system in Section IV-C below.

**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  $m + k + 1 + \tau$ . 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  $\delta$  in (11).

- 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  $m + k + 1 + \tau$ , with stopping time  $\tau = 0$  usually.

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

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

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  $P, k, \delta$  (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  $\sigma$ -algebra generated by the sequences  $Z_1, Z_2, \dots, Z_{n-1}$  and  $\tilde{X}_1, \tilde{X}_2, \dots, \tilde{X}_n$ . Note that knowledge of  $\tilde{X}_n$  involves a small 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 end of the round:

(a) For sufficiently large  $k$  and  $P$  in (11) and (29), respectively, it holds that<sup>1</sup>

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

(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  $\Delta$ . Furthermore, since the right side of (38) is independent of  $\mathcal{F}_m$ , the  $\beta + \Delta$ -moment of  $Y_n$  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

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

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

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

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  $P, k, \delta$ ),

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

Rewriting (43) 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).

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

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 (8), (12), (13) 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

$$t(1 - \delta)^{k+j-1} \left( \tilde{X}_m + \frac{B}{(1-a/2)(1-3\delta)} \right) \quad (50)$$

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

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 (12), (14), and (15). 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

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

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

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

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  $\lfloor a+1 \rfloor$ -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  $\lfloor a+1 \rfloor$  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 the values of  $w_i$ . However, this is easy; in the obvious generalization of our algorithm to higher  $a$ , the estimate  $C_n$  of the state magnitude will still be shared knowledge at all times; the (uniform) quantizer simply breaks up the interval  $[-C_n, C_n]$  into  $\lfloor a+1 \rfloor$  equal parts as in the  $1 \leq a < 2$  case.

In the case  $a < 1$ , the controller does nothing, which by Lemma 1 achieves  $\beta$ -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], [8]. 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

$$N(X|U) \triangleq \frac{1}{2\pi e} \exp(2h(X|U)) \quad (64)$$

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. [9, Appendix 2]):

$$N(X) \leq \kappa_\beta \mathbb{E}[|X|^\beta]^{\frac{2}{\beta}} \quad (65)$$

$$\kappa_\beta \triangleq \frac{2}{\pi e} \left( e^{\frac{1}{\beta}} \Gamma\left(1 + \frac{1}{\beta}\right) \beta^{\frac{1}{\beta}} \right)^2, \quad (66)$$

Thus,

$$\kappa_\beta \mathbb{E}[|X_n|^\beta]^{\frac{2}{\beta}} \geq N(X_n) \quad (67)$$

$$\geq N(X_n|U^{n-1}), \quad (68)$$

where (68) holds because conditioning reduces entropy. Next, we show a recursion on  $N(X_n|U^{n-1})$ :

$$N(X_n|U^{n-1}) = N(\mathbf{A}X_{n-1} + Z_{n-1}|U^{n-1}) \quad (69)$$

$$\geq a^2 N(X_{n-1}|U^{n-1}) + N(Z_{n-1}) \quad (70)$$

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

where (70) is due to the conditional entropy power inequality:<sup>2</sup>

$$N(X+Y|U) \geq N(X|U) + N(Y|U), \quad (72)$$

<sup>2</sup>Conditional EPI follows by convexity from the unconditional EPI first stated by Shannon [10] and proved by Stam [11].

which holds as long as  $X$  and  $Y$  are conditionally independent given  $U$ , and (71) 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

$$\min_{P_{U|X}: I(X;U) \leq r} h(X|U) \geq h(X) - r. \quad (73)$$

It follows from (71) that  $r > \log a$  is necessary to keep  $N(X_n | U^{n-1})$  bounded. Due to (68), it is also necessary to keep  $\beta$ -th moment of  $X_n$  bounded.  $\square$

Consider the following notion of stability.

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

$$\limsup_{n \rightarrow \infty} \mathbb{P}[X_n \in \mathcal{I}] > 0. \quad (74)$$

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,  $M \geq \lceil a \rceil$  is necessary.*

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

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

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  $\tilde{U}_n \triangleq \sum_{i=0}^n a^{n-i} U_i$  is the combined effect of  $t$  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  $\mathcal{I}$ ,

$$\mathbb{P}[X_{n+1} \in \mathcal{I}] = \mathbb{P}[a^n X_1 + \tilde{U}_n \in \mathcal{I}] \quad (76)$$

$$= \sum_{m=1}^{M^n} \mathbb{P}[a^n X_1 + u(m) \in \mathcal{I}, X_1 \in \mathcal{I}_m] \quad (77)$$

$$\leq M^n a^{-n} f_{\max} |\mathcal{I}|, \quad (78)$$

and (75) follows for any  $M < a$ , confirming the necessity of  $M \geq \lceil a \rceil$  to achieve weak stability.

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

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

$$\mathbb{P}[X_1 \in \mathcal{S}_b] = \int_{\mathbb{R}} f_{X_1}(x) 1\{f_{X_1}(x) > b\} dx \rightarrow 0 \text{ as } b \rightarrow \infty. \quad (79)$$

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 (75) holds for bounded  $f_{X_1}$ , we conclude

$$\limsup_{n \rightarrow \infty} \mathbb{P}[X_n \in \mathcal{I}] \leq \epsilon + \limsup_{n \rightarrow \infty} \mathbb{P}[X_n \in \mathcal{I} | X_1 \in \mathcal{S}_b] = \epsilon, \quad (80)$$

which implies that (75) 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  $M_\beta^*$  would be 3 if  $X_1$  had a density. Beyond distributions with densities, we conjecture that  $M_\beta^*$  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.

$$U_n = g_n(f_1(X_1), f_2(X^2), \dots, f_{n-\ell}(X^{n-\ell})), \quad (81)$$

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  $\alpha$ -moments. Assume that  $h(X_1) > -\infty$ .*

The minimum number of quantization points to achieve  $\beta$ -moment stability, for any  $0 < \beta < \alpha$  and with any constant delay  $\ell$  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  $\ell$  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

$$X_n = a^\ell \tilde{X}_n + a^{\ell-1} Z_{n+1} + \dots + Z_{n+\ell}. \quad (82)$$

By Theorem 1, we can keep  $\mathbb{E}[|\tilde{X}_n|^\beta]$  bounded for any  $\beta < \alpha$ . But each  $Z_i$  has bounded  $\beta$  moment, and so by Lemma 1 their sum will have bounded  $\beta$ -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,

$$U_n = \mathbf{g}_n(\{f_n(X^n) : n \in \mathcal{T}\}). \quad (83)$$

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

$$\frac{|n+i : n+i \in \mathcal{T}, i=0, \dots, N-1|}{N} > p. \quad (84)$$

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  $\alpha$ -moments. Assume that  $h(X_1) > -\infty$ . The minimum number of quantization points to achieve  $\beta$ -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

$$(\lfloor a \rfloor + 1)^p > a. \quad (85)$$

*Proof.* The requirement (85) ensures that the bounded case works; indeed, it is equivalent to

$$\left( \frac{a}{\lfloor a \rfloor + 1} \right)^p a^{1-p} < 1, \quad (86)$$

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  $\lambda$ . 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  $\lambda^2$ .

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

If the rows of  $M$  have  $\ell^1$  norm at most  $\lambda$ , the assumption of Proposition 2 that  $M$  has spectrum bounded by  $\lambda$  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  $\lambda$ : 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

$$\begin{pmatrix} |x| & -x \\ -x & |x| \end{pmatrix} \quad (87)$$

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

$$X_{n+1} = AX_n + Z_n - BU_n, \quad (88)$$

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

$$[B, AB, A^2B, \dots, A^{d-1}B] \quad (89)$$

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 (99), 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,

$$\begin{bmatrix} X_{n+1}^u \\ X_{n+1}^s \end{bmatrix} = \begin{bmatrix} A^u & A^s \\ 0 & A^s \end{bmatrix} \begin{bmatrix} X_n^u \\ X_n^s \end{bmatrix} + Z_n - \begin{bmatrix} B^u \\ 0 \end{bmatrix} U_n, \quad (90)$$

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 (88) with (A, B) stabilizable. Let  $X_1, Z_n$  be independent random  $\mathbb{R}^d$ -valued random vectors with bounded  $\alpha$ -moments. Assume that  $h(X_1) > -\infty$ . Let  $(\lambda_1, \dots, \lambda_d)$  be the eigenvalues of A, and set*

$$a \triangleq \prod_{j=1}^d \max(1, |\lambda_j|). \quad (91)$$

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

$$M_\beta^* = \lceil a \rceil + 1. \quad (92)$$

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

$$X_{n+1} = AX_n + Z_n - U_n, \quad (93)$$

and then explain how to deal with the general stabilizable system in (88).

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

$$A = \bigoplus_j A_j \quad (94)$$

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

$$\bigoplus_j \mathcal{V}_j = \mathbb{R}^d \quad (95)$$

such that:

- 1) The spectrum of each  $A_j$  is either a single real  $\lambda_j$  (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

$$\sup_n \mathbb{E}[|X_{n,j}|^\beta] < \infty \quad (96)$$

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 (96) 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

$$(\lfloor a \rfloor + 1)^{p_j} > |\lambda_j|^{\dim(\mathcal{V}_j)}. \quad (97)$$

The assumption (91) 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  $\epsilon > 0$ )

$$C' = (|\lambda_j| + \epsilon)^{\left(\frac{kp_j}{\dim \mathcal{V}_j}\right)} C \quad (98)$$

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  $O\left(\left(\lfloor a \rfloor + 1 - \epsilon\right)^{kp_j}\right)$  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  $\beta$ -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 [8], 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 (88), 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 (88), we will reduce the problem to the delayed version of (93). 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 (90). We will use the spanning set of matrices to give an arbitrary control with a delay of  $\ell$  steps, where  $\ell \leq d - 1$  is such that the range of

$$[B, AB, A^2B, \dots, A^\ell B] \quad (99)$$

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

$$v = Bv_0 + ABv_1 + \dots + A^\ell Bv_\ell, \quad (100)$$

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  $\ell$ , 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

$$\hat{X}_{n+1} = A\hat{X}_n + Z_n - \hat{U}_{n-\ell} \quad (101)$$

has bounded moments. We assume that the noises  $Z_n$  are common to (88) and (101), so that  $\{X_n\}$  and  $\{\hat{X}_n\}$  are coupled together rather than independent. Per (100) we can write

$$\hat{U}_n = BU_{0,n} + ABU_{1,n} + A^2BU_{2,n} + \dots + A^\ell BU_{\ell,n}. \quad (102)$$

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  $\hat{U}_n = BU_n$  applied at time  $n$  will consist of contributions from all  $\hat{U}_{n-\ell}, \dots, \hat{U}_n$ :

$$U_n = \sum_{i=0}^{\ell} U_{i,n-\ell+i}, \quad (103)$$

and the actual accumulated control by time  $n + \ell$  is larger than  $\hat{U}_n$ . Therefore, by applying (103) to the state of the original system  $X_n$  (88) 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 (8), in terms of the same majorizing sequence  $\tilde{X}_n$  ((30), Lemma 2) that we used to bound the  $\beta$ -moment of  $\hat{X}_n$ . Since  $\tilde{X}_n$  has bounded  $\beta$ -moment according to (35) (although we haven't emphasized it,  $\tilde{X}_n$  is also bounded in  $\beta$ -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

We studied control of stochastic linear systems over noiseless, time-invariant, rate-constrained communication links, and we showed that conveying  $\lceil a \rceil + 1$  distinct values is necessary and sufficient to achieve  $\beta$ -moment stability, where  $a$  is defined in (91), provided that the independent additive noises have bounded  $\alpha$  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 [12], 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 [3], [8] 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.

This paper resolves the question of the minimum number of bits necessary and sufficient for stability, when fixed-rate quantizers are used. While we picked the constants to guarantee a bounded  $\beta$ -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 [8], the tradeoff between rate and the attainable  $\beta$ -moment. It will be interesting to see whether our scheme can approach the lower bound in [8], and to compare its performance with that of the Lloyd-Max quantizer, explored in the context of control in [13].

## 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] 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.

- [5] Y. Sharon and D. Liberzon, "Input to state stabilizing controller for systems with coarse quantization," *IEEE Transactions on Automatic Control*, vol. 57, no. 4, pp. 830–844, 2012.
- [6] 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.
- [7] 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.
- [8] V. Kostina and B. Hassibi, "Rate-cost tradeoffs in control," *ArXiv preprint*, Oct. 2016.
- [9] 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.
- [10] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, July and October 1948.
- [11] 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.
- [12] V. Kostina, Y. Peres, G. Ranade, and M. Sellke, "Stabilizing a system with an unbounded random gain using only a finite number of bits," *ArXiv preprint arXiv:1805.05535*, May 2018.
- [13] 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.

## 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  $m+i$ :

$$X_{m+k} = a^{k-i} \left( X_{m+i} + \sum_{\ell=0}^{k-i-1} a^{-\ell-1} U_{m+i+\ell} + \sum_{\ell=0}^{k-i-1} a^{-\ell-1} Z_{m+i+\ell} \right). \quad (104)$$

Applying (8), (12) and (14), we can crudely bound the cumulative effect of controls on  $X_{m+k}$  as

$$\left| \sum_{\ell=0}^{k-1} a^{-\ell-1} U_{m+i+\ell} \right| \leq (a/2) \sum_{\ell=1}^{\infty} a^{-\ell-1} \left( (a/2)^{\ell} C_{m+i} + \frac{1 - (a/2)^{\ell}}{1 - a/2} B \right) \quad (105)$$

$$= C_{m+i} + \frac{B}{a-1} \quad (106)$$

$$\leq (a/2)^{-k} C_m + \frac{aB}{(2-a)(a-1)} \quad (107)$$

Unifying (104) and (107), we get

$$|X_{m+i}| \leq |X_{m+k}| + (a/2)^{-k} C_m + \frac{aB}{(2-a)(a-1)} + \sum_{\ell=0}^{k-i-1} a^{-\ell-1} |Z_{m+i+\ell}| \quad (108)$$

By Lemma 1, the sum of random variables on the right side of (108) has uniformly bounded  $\alpha$ -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  $i = 1, 2, \dots, k + \tau$ ,

$$|X_{m+i}| \leq a^i \left( 2C_m + \frac{B}{a-1} + \sum_{\ell=0}^{i-1} a^{-\ell-1} |Z_{m+\ell}| \right), \quad (109)$$

Maximizing the right side of (109) over  $1 \leq i \leq k+j$  and using (109), we conclude

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

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

$$C_{m+k+j} \leq \text{right side of (110)} + \frac{a^{k+j} B}{1 - a/2}. \quad (111)$$

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

$$C_{m+k+j} < P|X_{m+k+j-1}|. \quad (112)$$

Combining (110), (111) and (112) yields (27).