This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2021.3126679, IEEE Transactions on Automatic Control

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 the performance of our scheme 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

$$X_{n+1} = aX_n + Z_n - U_n, \tag{1}$$

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 $\{\mathsf{f}_n, \mathsf{g}_n\}_{n=1}^{\infty}$, where $\mathsf{f}_n \colon \mathbb{R}^n \mapsto [M]$ is the encoding (quantizing) function, and $\mathsf{g}_n \colon [M] \mapsto \mathbb{R}^n$ is the decoding (controlling) function, and $[M] \triangleq \{1, 2, \dots, M\}$. At time i, the controller outputs

$$U_n = \mathsf{g}_n(\mathsf{f}_1(X_1), \mathsf{f}_2(X^2), \dots, \mathsf{f}_n(X^n)). \tag{2}$$

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:

$$M_\beta^\star \triangleq \inf \Big\{ M \colon \exists\, M\text{-bin causal quantizer-controller} \quad \text{s.t.} \quad \limsup_n \mathbb{E}\left[|X_n|^\beta\right] < \infty \Big\}, \tag{3}$$

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 $0 < \beta < \alpha$*

$$M_\beta^\star \leq \lfloor a \rfloor + 1. \tag{4}$$

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 $\beta > 0$,*

$$M_\beta^\star \geq \lfloor a \rfloor + 1. \tag{5}$$

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 [5] showed that time-invariant fixed-rate quantizers are unable to attain bounded cost if the noise is unbounded [5], 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 [**?**], [**?**], [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 [5] 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 [9] (see [10], [11] 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 [5], [9] attain (4) for a large enough k. See also [**?**], 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 [5], 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 [3], for noiseless vector systems under different stability criteria [4], and for packet-drop channels [**?**].

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 [5], [9] 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 controller-quantizer enters the zoom-out mode, which is essentially the same as in [9]: 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.

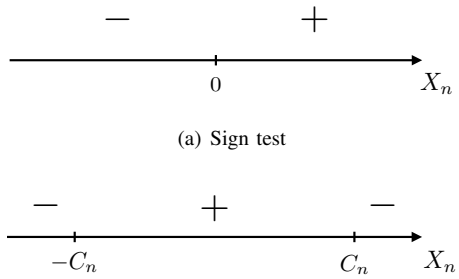
We also present a short proof of the converse result in Theorem 2 that uses information-theoretic arguments.

In Section II, we describe our achievable scheme and give its analysis. In Section III, we present a short proof of the converse in Theorem 2. Section IV presents an extension of Theorem 1 to vector systems. The results in this paper were partially presented at CDC [12]. The achievability proof for vector systems, an elementary converse proof for stabilizability in probability that is tight for non-integer a, and extensions to packet drop channels and dependent system noise are presented in the extended version [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.



(a) Sign test

(b) Magnitude test

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

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

$$C_2 \triangleq (a/2)C_1 + B \leq C_1, \tag{6}$$

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

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

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

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

and

$$\limsup_{n\to\infty} |X_n| \leq \frac{B}{1-a/2}. \tag{9}$$

Actually, this is the best achievable bound on the uncertainty about the location of X_n, as a simple volume-division argument shows [3], [13].

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,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2021.3126679, IEEE Transactions on Automatic Control

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

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

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

$$C_{m+1} = aC_m + B, \tag{11}$$

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

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

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

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 $X_m, \ldots, X_{m+k-1}$. 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

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

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

$$\tau \triangleq \inf \{j \geq 0 \colon |X_{m+k+j}| \leq C_{m+k+j}\}. \tag{15}$$

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 $i \in \{0, 1, \ldots, k\}$, we have

$$|Z_{m+i}| \leq B. \tag{16}$$

and thus

$$|X_{m+i}| \leq C_{m+i}. \tag{17}$$

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

$$|X_{m+k}| \leq C_{m+k} \tag{18}$$

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

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

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

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

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

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*

$$t^\alpha \mathbb{P}[|Z| > t] \tag{22}$$

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. □

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*

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

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

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

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

$$(1+y)^\alpha \le (1+\varepsilon) y^\alpha. \tag{25}$$

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

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

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

Remark 1. The mild assumptions of Lemma 1 make it easy to generalize our results to dependent noise in [17, Sec. IV].

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 $0 \le j \le \tau$, it holds that*

$$\max\left\{|X_{m+1}|, \dots, |X_{m+k+j}|, C_{m+k+j}\right\} \tag{27}$$
$$\le P a^{k+j} \left(2C_m + \frac{a\,B}{(2-a)(a-1)} + \sum_{\ell=0}^{k+j-1} a^{-\ell-1} |Z_{m+\ell}| \right),$$

Proof. Appendix. □

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 $a' \le a$. 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
$$\beta = \alpha - 3\Delta. \tag{28}$$
- Fix P large enough so that
$$P/a \ge \max\left\{ \left(\frac{a}{1-\delta}\right)^{\alpha-\Delta},\ 2^k,\ \frac{a^{k+1}}{2(a-1)} \right\}. \tag{29}$$

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 sequence[1] $\tilde{X}_n$ through, for $1 \le i \le k + \tau$,

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

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

$$|X_{m+k+j}| \le \tilde{X}_{m+k+j} \quad 0 \le j \le \tau. \tag{31}$$

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

$$\mathbb{E}\left[|X_{m+i}|^\beta\right] \le K_1\, \mathbb{E}\left[\tilde{X}_{m+k}^\beta\right] + K_2\, \mathbb{E}\left[\tilde{X}_m^\beta\right] + K_3. \tag{32}$$

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

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

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

[1] $\tilde{X}_n$ serves as a Lyapunov function that stochastically controls the growth of the state process X_n. See [11, Th. 2.1], [14] for a general approach to proving stability using Lyapunov functions for general Markov chains.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2021.3126679, IEEE Transactions on Automatic Control

By definition (30),

$$\tilde{X}_{m+i} \le \tilde{X}_{m+1} \quad i = 2, \dots, k+\tau, \tag{34}$$

with equality for $i \le k$.

We will show that

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

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

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

Define

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

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 that [2]

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

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

$$\mathbb{P}\left[Y_m \le 1-3\delta \mid \mathcal{F}_m\right] \to 1. \tag{39}$$

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 $\beta p \le \beta + \Delta$, and let q satisfy $\frac{1}{p} + \frac{1}{q} = 1$. Write

$$\begin{aligned}
&\mathbb{E}\left[Y_m^{\beta} \mid \mathcal{F}_m\right] \\
&\le (1-3\delta)^{\beta} + \mathbb{E}\left[Y_m^{\beta} 1\left\{Y_m > 1-3\delta\right\} \mid \mathcal{F}_m\right] && (40)\\
&\le (1-3\delta)^{\beta} + \left(\mathbb{E}\left[Y_m^{\beta p} \mid \mathcal{F}_m\right]\right)^{\frac{1}{p}} \left(\mathbb{P}\left[Y_m > 1-3\delta \mid \mathcal{F}_m\right]\right)^{\frac{1}{q}} && (41)\\
&\to (1-3\delta)^{\beta}, \quad B \to \infty, && (42)
\end{aligned}$$

where (41) is by Hölder's inequality, and the second term in (41) vanishes as $B \to \infty$ due to (39) and uniform boundedness of the $\beta+\Delta$ - moment of $\{Y_m \mid \mathcal{F}_m\}$. 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),

$$\mathbb{E}\left[Y_m^{\beta} \mid \mathcal{F}_m\right] \le (1-2\delta)^{\beta}. \tag{43}$$

Rewriting (43) using (37) yields

$$\begin{aligned}
\mathbb{E}[\tilde{X}_{m+1}^{\beta} \mid \mathcal{F}_m] &\le (1-2\delta)^{\beta}\left(\tilde{X}_m + \frac{B}{(1-a/2)(1-3\delta)}\right)^{\beta} && (44)\\
&\le (1-\delta)^{\beta}\tilde{X}_m^{\beta} + K, && (45)
\end{aligned}$$

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

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

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 $\mathbb{P}[Y_m \ge t, \tau = j | \mathcal{F}_m]$ 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:

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

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

$$\begin{aligned}
a^i\left|\sum_{\ell=0}^{k-1} a^{-\ell-1}U_{m+\ell}\right| &\le a^i\,(a/2)\sum_{\ell=1}^{\infty} a^{-\ell-1} && (48)\\
&\quad \left((a/2)^{\ell-1}C_{m+1} + \frac{1-(a/2)^{\ell-1}}{1-a/2}B\right)\\
&= a^i\left(C_m + \frac{B}{a-1}\right). && (49)
\end{aligned}$$

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

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

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

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

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

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

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

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,

$$|X_{m+k+j-1}| > C_{m+k+j-1} \tag{53}$$
$$= P^j\left(2\left(a/2\right)^k C_m + \frac{B}{1-a/2}\right), \tag{54}$$

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

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

the event equivalent to

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

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]

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

[3]Similar exponential bounds to the event $\mathbb{P}\left[\tau \geq j\right]$ are provided in [11, Lem. 5.2] and in [15, Lem. 5.2].

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

$$\mathbb{P}\left[\tau = \infty\right] \leq \lim_{j\to\infty} \mathbb{P}\left[\tau \geq j\right] = 0, \tag{58}$$

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

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

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

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

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 $B \to \infty$, all terms $Z_m, \ldots, Z_{m+k}$ are within $[-B, B]$, and $\tau = 0$. In such a case, applying (19) and recalling (30), we get

$$\tilde{X}_{m+1} = \max\{|X_{m+k}|, C_{m+k}\} \tag{61}$$
$$\leq (1-3\delta)\,\tilde{X}_m + \frac{B}{1-a/2}, \tag{62}$$

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

□

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

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

and the controller receives the index $b_n \in \{0, 1, \ldots, \lfloor a \rfloor\}$ 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 (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

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

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 [5], [16]. See [17, Th. 3] for an alternative converse result that uses elementary probability, which implies Theorem 2 unless a is an integer.

Proof of Theorem 2. Conditional entropy power is defined as

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

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. [18, Appendix 2]):

$$N(X) \le \kappa_\beta \mathbb{E}\left[|X|^\beta\right]^{\frac{2}{\beta}} \tag{66}$$

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

Thus,

$$\kappa_\beta \mathbb{E}\left[|X_n|^\beta\right]^{\frac{2}{\beta}} \ge N\left(X_n\right) \tag{68}$$

$$\ge N\left(X_n|U^{n-1}\right), \tag{69}$$

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

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

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

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

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

$$N(X+Y|U) \ge N(X|U) + N(Y|U), \tag{73}$$

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

$$\min_{P_{U|X}:\, I(X;U)\le r} h(X|U) \ge h(X) - r. \tag{74}$$

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

[4]Conditional EPI follows by convexity from the unconditional EPI first stated by Shannon [19] and proved by Stam [20].

IV. Vector systems

The results generalize to higher dimensional systems

$$X_{n+1} = \mathsf{A}X_n + Z_n - \mathsf{B}U_n, \tag{75}$$

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.

Theorem 3. *Consider the stochastic vector linear system in (75) 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 $(\lambda_1, ..., \lambda_d)$ be the eigenvalues of A, and set*

$$a \triangleq \prod_{j=1}^{d} \max(1, |\lambda_j|). \tag{76}$$

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

$$M_\beta^\star = \lfloor a \rfloor + 1. \tag{77}$$

The proof of Theorem 3 relies on an idea previously explored in e.g. [5] that one can decompose $\mathbb{R}^d$ into eigenspaces of A and rotate attention between these parts. Full proof is contained in [17, Sec. IV].

V. Conclusion

This paper studies the minimum number of bits necessary and sufficient for stability, when fixed-rate quantizers are used. Extensions of the results of this paper to constant-length time delays, to control over communication channels that drop a small fraction of packets, and to systems with dependent Gaussian noise are discussed in [17, Sec. IV].

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 [16], the tradeoff between rate and the attainable β-moment. It will be interesting to see whether our scheme can approach the lower bound in [16], and to compare its performance with that of the Lloyd-Max quantizer, explored in the context of control in [21].

Appendix

A. Proof of (32)

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

$$X_{m+k} = a^{k-i}\left(X_{m+i} + \sum_{\ell=0}^{k-i-1} a^{-\ell-1} U_{m+i+\ell} + \sum_{\ell=0}^{k-i-1} a^{-\ell-1} Z_{m+i+\ell}\right). \tag{78}$$

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

$$\left|\sum_{\ell=0}^{k-1} a^{-\ell-1} U_{m+i+\ell}\right| \le (a/2)\sum_{\ell=1}^{\infty} a^{-\ell-1} \left((a/2)^{\ell} C_{m+i} + \frac{1-(a/2)^{\ell}}{1-a/2} B\right) \quad (79)$$

$$= C_{m+i} + \frac{B}{a-1} \quad (80)$$

$$\le (a/2)^{-k} C_m + \frac{aB}{(2-a)(a-1)} \quad (81)$$

Unifying (78) and (81), we get

$$|X_{m+i}| \le |X_{m+k}| + (a/2)^{-k} C_m + \frac{aB}{(2-a)(a-1)} + \sum_{\ell=0}^{k-i-1} a^{-\ell-1} |Z_{m+i+\ell}| \quad (82)$$

By Lemma 1, the sum of random variables on the right ride of (82) has uniformly bounded α-moments, and since by definition of $\tilde{X}_n$ in (30), $\tilde{X}_m \le C_m$ and $|X_{m+k}| \le \tilde{X}_{m+k}$, (32) follows by the means of (24).

B. Proof of Lemma 2

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

$$|X_{m+i}| \le a^i \left(2C_m + \frac{B}{a-1} + \sum_{\ell=0}^{i-1} a^{-\ell-1} |Z_{m+\ell}|\right), \quad (83)$$

Maximizing the right side of (83) over $1 \le i \le k+j$ and using (83), we conclude

$$\max_{1\le i\le k+j} |X_{m+i}| \le a^{k+j}\left(2C_m + \frac{B}{a-1} + \sum_{\ell=0}^{k+j-1} a^{-\ell-1}|Z_{m+\ell}|\right), \quad (84)$$

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

$$C_{m+k+j} \le \text{right side of (84)} + \frac{a^{k+j}B}{1-a/2}. \quad (85)$$

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

$$C_{m+k+j} < P|X_{m+k+j-1}|. \quad (86)$$

Combining (84), (85) and (86) 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] 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.

[13] 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.

[14] 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.

[15] ——, "Exact minimum number of bits to stabilize a linear system," *ArXiv preprint arXiv:1807.07686*, July 2018.

[16] 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.

[17] S. Yüksel and T. Başar, *Stochastic networked control systems: Stabilization and optimization under information constraints*. Springer Science & Business Media, 2013.

[18] 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.

[19] V. Kostina and B. Hassibi, "Rate-cost tradeoffs in control," *IEEE Transactions on Automatic Control*, vol. 64, no. 11, pp. 4525–4540, Apr. 2019.

[20] 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.

[21] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, July and October 1948.

[22] 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.

[23] 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.