Control with Minimum Communication Cost per Symbol

Technical Report

Justin Pearson, João P. Hespanha, Daniel Liberzon

May 11, 2014

Abstract

We address the problem of stabilizing a continuous-time linear time-invariant process under communication constraints. We assume that the sensor that measures the state is connected to the actuator through a finite capacity communication channel over which an encoder at the sensor sends symbols from a finite alphabet to a decoder at the actuator. We consider a situation where one symbol from the alphabet consumes no communication resources, whereas each of the others consumes one unit of communication resources to transmit. This paper explores how the imposition of limits on an encoder's bit-rate and average resource consumption affect the encoder/decoder/controller's ability to keep the process bounded. The main result is a necessary and sufficient condition for a bounding encoder/decoder/controller which depends on the encoder's average bit rate, its average resource consumption, and the unstable eigenvalues of the process.

1 Introduction

This paper addresses the problem of stabilizing a continuous-time linear time-invariant process under communication constraints. As in [1–7], we assume that the sensor that measures the state is connected to the actuator through a finite capacity communication channel. At each sampling time, an encoder sends a symbol through the channel. The problem of determining whether or not it is possible to bound the state of the process under this type of encoding scheme is not new; it was established in [2–4] that a necessary and sufficient condition for stability can be expressed as a simple relationship between the unstable eigenvalues of A and the average communication bit-rate.

We expand upon this result by considering the notion that encoders can effectively save communication resources by not transmitting information, while noting that the absence of an explicit transmission nevertheless conveys information. To capture this, we suppose that one symbol from the alphabet consumes no communication resources to transmit, whereas each of the others consumes one unit of communication resources. We then proceed to define the *average cost per symbol* of an encoder, which is essentially the

average fraction of non-free symbols emitted. This paper's main technical contribution is a necessary and sufficient condition for the existence of an encoder/decoder/controller that bounds the state of the process. This condition depends on the channel's average bit rate, the encoder's average cost per symbol, and the unstable eigenvalues of A.

Our result extends [2] in the sense that as the constraint on the average cost per symbol is allowed to increase (becomes looser), our necessary and sufficient condition becomes the condition from [2]. As with [2–4], our result is constructive in the sense that we describe a family of encoder/decoder pairs that bound the process when our condition holds. A counterintuitive corollary to our main result shows that if the process may be bounded with average bit-rate r, then there exists a bounding encoder/decoder/controller with average bit-rate r which uses no more than 50% non-free symbols in its symbol-stream.

The remainder of this paper is organized as follows. Section 3 contains the main negative result of the paper, namely that boundedness is not possible when our condition does not hold. To prove this result we actually show that it is not possible to bound the process with a large class of encoders — which we call M-of-N encoders — that includes all the encoders with average cost per symbol not exceeding a given threshold. Section 4 contains the positive result of the paper, showing that when our condition *does* hold, there is an encoder/decoder pair that can bound the process; we provide the encoding scheme.

2 Problem Statement

Consider a stabilizable linear time-invariant process

$$\dot{x} = Ax + Bu, \qquad x \in \mathbb{R}^n, u \in \mathbb{R}^m, \tag{1}$$

for which it is known that $x(0)$ belongs to some bounded set $\mathcal{X}_0 \subset \mathbb{R}^n$. A sensor that measures the state $x(t)$ is connected to the actuator through a finite-data-rate, error-free, and delay-free communication channel. An *encoder* collocated with the sensor samples the state once every T time units, and from this sequence of measurements $\{x(kT) : k \in \mathbb{N}_{>0}\}$ causally constructs a sequence of symbols $\{s_k \in \mathcal{A} : k \in \mathbb{N}_{>0}\}$ from a finite alphabet $\mathcal{A}$. The encoder sends this symbol sequence through the channel at a rate of 1 symbol every T time units to a *decoder/controller* collocated with the actuator, which causally constructs the control signal $u(t)$, $t \geqslant 0$ from the sequence of symbols $\{s_k \in \mathcal{A} : k \in \mathbb{N}_{>0}\}$ that arrive at the decoder. We assume the channel faithfully transmits each symbol without error.

The positive time T between successive samplings is called the *sampling period* and has units of time units. The *average bit-rate* of an encoder/decoder pair is the amount of information that the encoder transmits in units of bits per time unit. For an encoder/decoder pair whose encoder transmits one symbol from an alphabet $\mathcal{A}$ every T time units, the pair's average bit-rate is given by

$$r \coloneqq \frac{\log_2 |\mathcal{A}|}{T}. \tag{2}$$

We consider encoder/decoder pairs whose alphabets each contain one special symbol $0 \in \mathcal{A}$ that can be transmitted without consuming any communication resources, and S additional symbols that each require one unit of communication resources per transmission. One can think of the "free" symbol 0 as the absence of an explicit transmission. The "communication resources" at stake may be energy, time, or any other resource that may be consumed in the course of the communication process. In order to capture the average rate that an encoder consumes communication resources, we define the *average cost per symbol* of an encoder as follows: We say an encoder has *average cost per symbol not exceeding* $\gamma_{\max}$ if for every symbol sequence $\{s_k\}$ the encoder may generate, we have

$$\frac{1}{N-M+1}\sum_{k=M}^{N} I_{s_k\neq 0} \leqslant \gamma_{\max} + O\left(\frac{1}{N-M+1}\right) \tag{3}$$

for all integers N, M satisfying $N \geqslant M \geqslant 0$, where $I_{s_k\neq 0} \coloneqq 1$ if the kth symbol is not the free symbol, and 0 if it is. The summation in (3) captures the total resources spent transmitting symbols $s_M, s_{M+1}, \ldots, s_N$. Motivating this definition of average cost per symbol is the observation that the lefthand side has the intuitive interpretation as the average cost per transmitted symbol between symbols s_M and s_N. As $N - M \to \infty$, the rightmost "big-oh" term vanishes, leaving $\gamma_{\max}$ as an upper bound on the average long-term cost per symbol of the symbol sequence. Note that the average cost per symbol of any encoder never exceeds 1 and does not depend on the sampling period T.

Whereas an encoder/decoder pair's average bit-rate r only depends on its symbol alphabet $\mathcal{A}$ and sampling period T, its average cost per symbol depends on every possible symbol sequence it may generate, and therefore depends on the encoder/decoder pair, the controller, the process (1), and the initial condition $x(0)$.

The specific question considered in this paper is: under what conditions on the bit-rate and average cost per symbol does there exist a controller and encoder/decoder pair that keep the state of process (1) bounded?

3 Necessary condition for boundedness with limited-communication encoders

It is well known from [2–4] that it is possible to construct a controller and encoder/decoder pair that bounds process (1) with average bit-rate r only if

$$r \ln 2 \geqslant \sum_{i:\Re\lambda_i[A]\geqslant 0} \lambda_i[A], \tag{4}$$

where ln denotes the base-e logarithm, and the summation is over all eigenvalues of A with nonnegative real part. The result that follows shows that a larger bit-rate may be needed when one poses constraints on the encoder's average cost per symbol $\gamma_{\max}$. Specifically, when $\gamma_{\max} \geqslant S/(S+1)$ the (necessary) stability condition reduces to (4), but when $\gamma_{\max} < S/(S+1)$ a bit-rate larger than (4) is necessary to compensate for the "dilution" of information content in the transmitted symbols due to the constraint on the average cost per symbol.

Theorem 1. *Consider an encoder/decoder pair with sampling period T, an alphabet with S nonfree symbols and one free symbol, and an average cost per symbol not exceeding $\gamma_{\max}$. If this pair keeps the state of process (1) bounded for every initial condition $x_0 \in \mathcal{X}_0$, then we must have*

$$r\, f(\gamma_{\max}, S) \ln 2 \geqslant \sum_{i:\Re\lambda_i[A]\geqslant 0} \lambda_i[A], \tag{5}$$

where the bit-rate r is related to S and T via Equation (2), and the function $f : [0,1] \times [0,\infty) \to [0,\infty)$ is defined as

$$f(\gamma, S) \coloneqq \begin{cases} \frac{H(\gamma)+\gamma \log_2 S}{\log_2(S+1)} & 0 \leqslant \gamma \leqslant \frac{S}{S+1} \\ 1 & \frac{S}{S+1} < \gamma \leqslant 1, \end{cases} \tag{6}$$

and $H(p) \coloneqq -p\log_2(p) - (1-p)\log_2(1-p)$ is the base-2 entropy of a Bernoulli random variable with parameter p. □

Remark 1. The function f is plotted in Figure 1 for several values of S. It is worth making two observations regarding f. The first observation is that the function $f(\gamma, S)$ is monotone nonincreasing in S for any fixed $\gamma \in [0,1]$, which implies that smaller alphabets are preferable to large ones when trying to satisfy (5) with a given fixed bit-rate. The second observation is that the average cost per time unit, which is γ/T, can be made arbitrarily small in two different ways.

1. One could pick T very large, then leveraging the fact that $rf(\gamma, S)$ is monotone increasing in S, pick S large enough to satisfy (5). This approach has two downsides: First, with a large sampling period, the state, although remaining bounded, can grow quite large between transmissions. Second, large S means that the encoder/decoder pair must store and process a large symbol library, adding complexity to the pair's implementation.

2. Alternatively, one can make γ/T arbitrary small by observing that the sequences

$$\gamma_k \coloneqq e^{-k}, \qquad T_k \coloneqq e^{-k}\sqrt{k}, \qquad k \in \mathbb{N}_{>0}$$

have the property that $\gamma_k \to 0$, $T_k \to 0$, and $\gamma_k/T_k \to 0$, but $H(\gamma_k)/T_k \to \infty$, and hence $r_k f(\gamma_k, S)\ln 2 \to \infty$ (where $r_k \coloneqq \log_2(S+1)/T_k$). This means that one can find $k \in \mathbb{N}_{>0}$ to make the average cost per time unit γ_k/T_k arbitrarily small, and also satisfy the necessary condition (5). The drawback of this approach is that to achieve a small sampling period T in practice requires an encoder/decoder pair with a very precise clock.

Remark 2. The addition of the "free" symbol effectively increases the data rate without increasing the rate of resource consumption, as seen by the following two observations:

- Without the free symbols, the size of the alphabet would be S and the bit-rate would be $\log_2(S)/T$. It could happen that this bit-rate is too small to bound the plant, yet after the introduction of the free symbol, the condition (5) is satisfied.

- Since $\gamma_{\max}$ is the fraction of non-free symbols, then the quantity $r\ \gamma_{\max}$ is the number of bits per time unit spent transmitting non-free symbols. But since $f(\gamma, S) \geqslant \gamma$, again we see that the free symbols help satisfy (5).

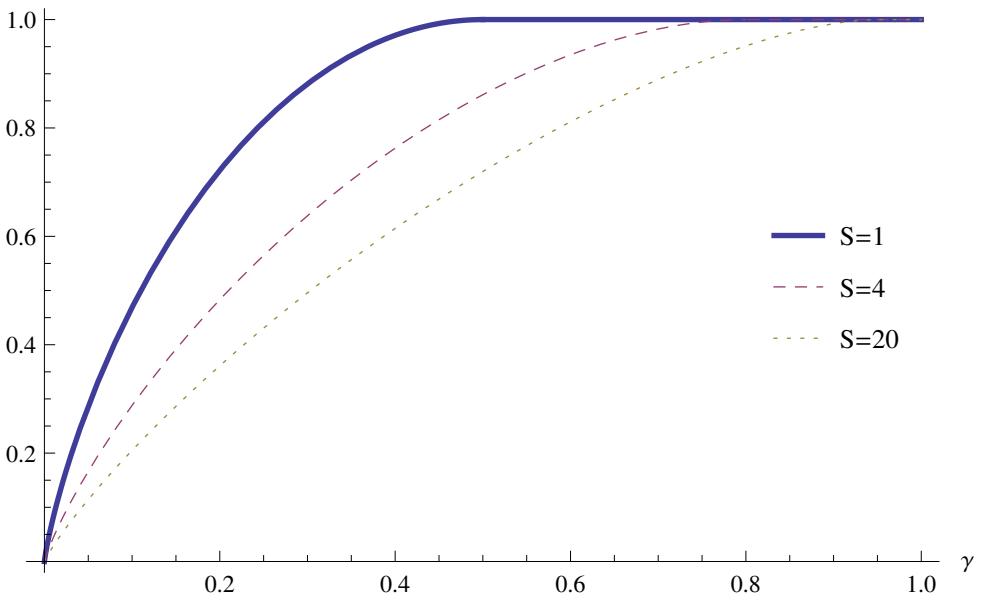

Figure 1: A plot of $f(\gamma, S)$ for $S = 1, 4, 20$.

3.1 Theorem 1 Proof Setup

We lead up to the proof of Theorem 1 by first establishing three lemmas centered around a restricted large class of encoders called M-of-N encoders. We first define M-of-N encoders, which essentially partition their symbol sequences into N-length *codewords*, each with M or fewer non-free symbols. Lemma 1 demonstrates that every encoder with a limited average cost per symbol is an M-of-N encoder for appropriate N and M. Next, in Lemma 2 we establish a relationship between the number of codewords available to an M-of-N encoder and the function f as defined in (6). Then, in Lemma 3 we leverage previous work to establish a necessary condition for an M-of-N encoder to bound the state of the process. Finally, the proof of Theorem 1 leverages these three results.

To introduce the class of M-of-N encoders, we define an *N-symbol codeword* to be a sequence

$$\{s_{\ell N+1}, s_{\ell N+2}, \ldots, s_{\ell N+N}\}$$

of N consecutive symbols starting at an index $k = \ell N + 1$, with $\ell \in \mathbb{N}_{\geqslant 0}$. An *$M$-of-$N$ encoder* is an encoder for which every N-symbol codeword has M or fewer non-free symbols, i.e.,

$$\sum_{k=\ell N+1}^{\ell N+N} I_{s_k \neq 0} \leqslant M, \qquad \forall \ell \in \mathbb{N}_{\geqslant 0}. \tag{7}$$

The total number of distinct N-symbol codewords available to an M-of-N encoder is thus given by

$$L(N, M, S) \coloneqq \sum_{i=0}^{\lfloor M \rfloor} \binom{N}{i} S^i, \tag{8}$$

where the ith term in the summation counts the number of N-symbol codewords with exactly i non-free symbols.

Note that in keeping with the problem setup, the M-of-N encoders considered here each draw their symbols from the symbol library $\mathcal{A} \coloneqq \{0, 1, \ldots, S\}$ and transmit symbols with sampling period T.

An intuitive property of M-of-N encoders is that an M-of-N encoder has average cost per symbol not exceeding M/N. To see this, suppose M and N are fixed and let $N_1, N_2 \in \mathbb{N}_{\geqslant 0}$ be arbitrary with $N_2 \geqslant N_1 \geqslant 0$. The width of the interval $[N_1, N_2]$ may not be an integer multiple of N, but nevertheless it is between two integer multiples of N: we have $N(k-2) \leqslant N_2 - N_1 \leqslant \hat{N}_2 - \hat{N}_1 = kN$ for some non-negative integer $k \in \mathbb{N}_{\geqslant 0}$. From the first inequality we obtain

$$k \leqslant \frac{N_2 - N_1 + 1}{N} + 2 - \frac{1}{N},$$

and from the second inequality we have

$$\sum_{k=N_1}^{N_2} I_{s_k \neq 0} \leqslant kM,$$

because in each of the k N-length intervals containing $[N_1, N_2]$ there are at most M non-free symbols. Combining these yields

$$\sum_{k=N_1}^{N_2} I_{s_k \neq 0} \leqslant \frac{M}{N}(N_2 - N_1 + 1) + N_0, \tag{9}$$

where N_0 is some constant. Dividing both sides by $(N_2 - N_1 + 1)$ and rearranging, (3) follows.

The fact that an M-of-N encoder refrains from sending "expensive" codewords effectively reduces its ability to transmit information. Indeed, since an M-of-N encoder takes NT time units to transmit one of $L(N, M, S)$ codewords, the encoder transmits merely $\frac{\log_2 L(N,M,S)}{NT}$ bits per time unit. For $M < N$, this is strictly less than the encoder's average bit-rate r.

The first lemma, proved in the appendix, shows that the set of M-of-N encoders is "complete" in the sense that every encoder with average cost per symbol not exceeding a finite threshold $\gamma_{\max}$ is actually an M-of-N encoder for N sufficiently large and $M \approx \gamma_{\max} N$.

Lemma 1. *For every encoder/decoder pair with average bit-rate r and average cost per symbol not exceeding $\gamma_{\max} \in [0, 1]$, and every constant $\epsilon > 0$, there exist positive integers M and N with $M < N\gamma_{\max}(1 + \epsilon)$ such that the encoder is an M-of-N encoder.* □

The next lemma establishes a relationship between the number of codewords $L(N, M, S)$ available to an M-of-N encoder and the function f defined in (6).

Lemma 2. *For any $N, S \in \mathbb{N}_{>0}$ and $\gamma \in [0, 1]$, the function L defined in (8) and the function f defined in (6) satisfy*

$$\frac{\log_2 L(N, N\gamma, S)}{N} \leqslant \log_2(S + 1) f(\gamma, S), \tag{10}$$

with equality holding only when $\gamma = 0$ or $\gamma = 1$. Moreover, we have asymptotic equality in the sense that

$$\lim_{N\to\infty} \frac{\log_2 L(N, N\gamma, S)}{N} = \log_2(S+1) f(\gamma, S). \tag{11}$$

□

The left and right sides of (10) are plotted in Figure 2.

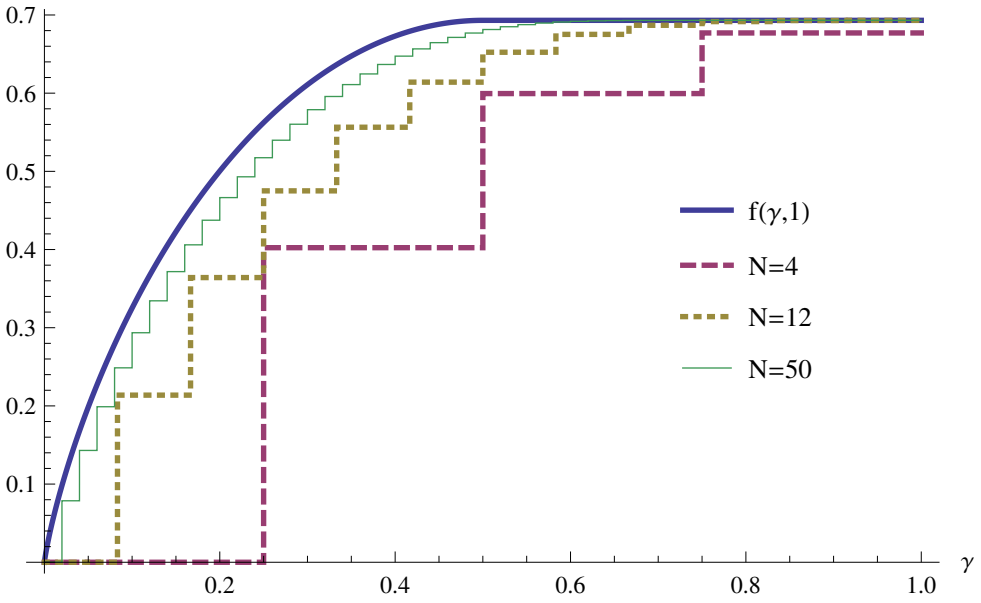

Figure 2: A plot of $\log_2(S+1)f(\gamma, S)$ and $\log_2 L(N, N\gamma, S)/N$ for $S = 1$ and $N = 4, 12, 50$.

Proof of Lemma 2. Let N and S be arbitrary positive integers. First we prove (10) for $\gamma \in (0, \frac{S}{S+1}]$. Applying the Binomial Theorem to the identity $1 = (\gamma + (1-\gamma))^N$, we obtain

$$1 = \sum_{i=0}^{N} \binom{N}{i} \gamma^i (1-\gamma)^{N-i} \qquad \forall \gamma > 0, \quad \forall N \in \mathbb{N}_{>0}.$$

Since each term in the summation is positive, keeping only the first $\lfloor N\gamma \rfloor$ terms yields the inequality

$$1 > \sum_{i=0}^{\lfloor N\gamma \rfloor} \binom{N}{i} \gamma^i (1-\gamma)^{N-i} \qquad \forall \gamma > 0, \quad \forall N \in \mathbb{N}_{>0}. \tag{12}$$

Next, a calculation presented as Lemma 5 in the appendix reveals that

$$\gamma^i (1-\gamma)^{N-i} \geqslant 2^{-N\,H(\gamma)} \frac{S^i}{S^{N\gamma}} \tag{13}$$

for all $N, S \in \mathbb{N}_{>0}$, $\gamma \in (0, S/(S+1)]$, and $i \in [0, N\gamma]$. Using this in (12) and simplifying yields

$$\frac{\log_2 L(N, N\gamma, S)}{N} < H(\gamma) + \gamma \log_2 S \qquad \forall N, S \in \mathbb{N}_{>0}, \quad \forall \gamma \in \left(0, \frac{S}{S+1}\right]. \tag{14}$$

Since $\log_2(S+1) f(\gamma, S) = H(\gamma) + \gamma \log_2 S$ when $\gamma \in \left[0, \frac{S}{S+1}\right]$, (14) proves (10) for $\gamma \in (0, \frac{S}{S+1}]$. Next, suppose $\gamma \in (\frac{S}{S+1}, 1)$ and observe from (8) that $L(N, M, S)$ is a sum of positive terms whose index reaches $\lfloor M \rfloor$, hence $L(N, N\gamma, S)$ is strictly less than $L(N, N, S)$ for any $\gamma < 1$. We conclude that

$$\frac{\log_2 L(N, N\gamma, S)}{N} < \frac{\log_2 L(N, N, S)}{N} = \log_2(S+1) \qquad \forall N, S \in \mathbb{N}_{>0}, \quad \forall \gamma \in \left(\frac{S}{S+1}, 1\right). \tag{15}$$

Since $\log_2(S+1) f(\gamma, S) = \log_2(S+1)$ for $\gamma \in (\frac{S}{S+1}, 1)$, this concludes the proof of (10) for $\gamma \in (0, 1)$. The proof of (10) for $\gamma = 0$, follows merely from inspection of (10), and the $\gamma = 1$ case follows from the equality in (15).

Next we prove the asymptotic result (11) using information-theoretic methods. First we prove (11) for $\gamma \in [0, \frac{S}{S+1})$. Consider a random variable X parameterized by $S \in \mathbb{N}_{>0}$ and $\gamma \in [0, \frac{S}{S+1})$ which takes values in $\mathcal{X} \coloneqq \{0, 1, \ldots, S\}$ with probabilities given by

$$\begin{aligned} \mathbb{P}(X = 0) &\coloneqq (1 - \gamma) \\ \mathbb{P}(X = i) &\coloneqq \gamma / S \qquad i = \{1, 2, \ldots, S\}. \end{aligned}$$

Following our convention, we call 0 the "free" symbol and $1, \ldots, S$ the "nonfree" symbols. To lighten notation we write $p(x) \coloneqq \mathbb{P}(X = x), x \in \mathcal{X}$. The entropy of the random variable X is

$$H(X) \coloneqq -\sum_{i=0}^{S} p(i) \log_2 p(i) = H(\gamma) + \gamma \log_2 S, \tag{16}$$

where we have overloaded the symbol H so that $H(\gamma) \coloneqq -\gamma \log_2 \gamma - (1-\gamma) \log_2 (1-\gamma)$ is the entropy of a Bernoulli random variable with parameter γ.

Next, for some arbitrary $N \in \mathbb{N}_{>0}$, we consider N-length sequences of i.i.d. copies of X. Let $\mathcal{X}^N \coloneqq \{(x_1, \ldots, x_N) : x_i \in \mathcal{X}\}$. We use the symbol x^N as shorthand for $(x_1, \ldots, x_N)$, and we use $p(x^N)$ as shorthand for $\mathbb{P}\Big((X_1, X_2, \ldots, X_N) = (x_1, x_2, \ldots, x_N)\Big)$. Given an N-length sequence $x^N \in \mathcal{X}^N$, the probability that the N i.i.d. random variables $(X_1, \ldots, X_N)$ take on the values in the N-tuple x^N is given by

$$p(x^N) = (1-\gamma)^{N - \sum_{i=1}^N I_{x_i \neq 0}} \frac{\gamma^{\sum_{i=1}^N I_{x_i \neq 0}}}{S^{\sum_{i=1}^N I_{x_i \neq 0}}}. \tag{17}$$

The summation $\sum_{i=1}^N I_{x_i \neq 0}$ is the number of nonfree symbols in the N-tuple x^N.

For arbitrary $\epsilon > 0$, define the set $A_\epsilon^{(N)} \subseteq \mathcal{X}^N$ as

$$A_\epsilon^{(N)} \coloneqq \left\{ x^N \in \mathcal{X}^N \;\middle|\; N(\gamma - \delta_\epsilon) \leqslant \sum_{i=1}^N I_{x_i \neq 0} \leqslant N(\gamma + \delta_\epsilon) \right\}, \tag{18}$$

where $\delta_\epsilon \coloneqq \epsilon / \log_2 \frac{(1-\gamma)S}{\gamma}$. That is, $A_\epsilon^{(N)}$ is the set of all N-length sequences with "roughly" $N\gamma$ nonfree symbols. Using (17) and standard algebraic manipulations, we express the inequalities in (18) as

$$A_\epsilon^{(N)} = \left\{ x^N \in \mathcal{X}^N \;\middle|\; 2^{-N(H(X)+\epsilon)} \leqslant p(x^N) \leqslant 2^{-N(H(X)-\epsilon)} \right\}. \tag{19}$$

Here we relied on the fact that $\log_2 \frac{(1-\gamma)S}{\gamma} > 0$ for $S \in \mathbb{N}_{>0}, \gamma \in [0, \frac{S}{S+1})$. In the form of (19), we recognize $A_\epsilon^{(N)}$ as the so-called "typical set" of N-length sequences of i.i.d. copies of X as defined in [11]. Theorem 3.1.2 of [11] uses the Asymptotic Equipartition Property of sequences of i.i.d. random variables to prove that for any $\epsilon > 0$, we have

$$(1-\epsilon)2^{N(H(X)-\epsilon)} \leqslant \left|A_\epsilon^{(N)}\right| \tag{20}$$

for $N \in \mathbb{N}_{>0}$ large enough.
Next, we observe that

$$\left|A_\epsilon^{(N)}\right| \leqslant L\left(N, N\left(\gamma+\delta_\epsilon\right), S\right), \tag{21}$$

because $\left|A_\epsilon^{(N)}\right|$ is the number of N-length sequences with a number of nonfrees in the interval $[N(\gamma-\delta_\epsilon), N(\gamma+\delta_\epsilon)]$, whereas the right-hand side counts sequences with a number of nonfrees in the larger interval $[0, N(\gamma+\delta_\epsilon)]$. Combining (20) and (21), we obtain that for any $\epsilon > 0$,

$$\frac{1}{N}\log_2(1-\epsilon) + H(\gamma) + \gamma \log_2 S - \epsilon \leqslant \frac{1}{N}\log_2 L\left(N, N\left(\gamma+\delta_\epsilon\right), S\right) \tag{22}$$

for N large enough. Moreover, by (10) we have

$$\frac{1}{N}\log_2 L\left(N, N\left(\gamma+\delta_\epsilon\right), S\right) \leqslant H\left(\gamma+\delta_\epsilon\right) + \left(\gamma+\delta_\epsilon\right)\log_2 S \tag{23}$$

for any $\gamma \in [0, \frac{S}{S+1})$, $N, S \in \mathbb{N}_{>0}$, and $\epsilon > 0$. Combining these two observations and letting $\epsilon \to 0$, we conclude (11) for $\gamma \in [0, \frac{S}{S+1})$.
Next we prove (11) for $\gamma \in [\frac{S}{S+1}, 1]$. In (22) and (23), L is bounded by functions which are continuous in γ on $\gamma \in [0, \frac{S}{S+1})$ and which become equal as $\epsilon \to 0$. We conclude that $\lim_{N\to\infty} L(N, N\gamma, S)$ must be continuous on $\gamma \in [0, \frac{S}{S+1})$, and hence its closure. In other words, we are saying that

$$\lim_{N\to\infty} L(N, N\gamma, S)\Big|_{\gamma=\frac{S}{S+1}} = \lim_{\gamma\to\frac{S}{S+1}^-} \lim_{N\to\infty} L(N, N\gamma, S) = \log_2(S+1) \qquad \forall S \in \mathbb{N}_{>0}.$$

Since L is monotonically nondecreasing in its second argument, and $L(N, N\gamma, S) < \log_2(S+1)$ for $N, S \in \mathbb{N}_{>0}$, $\gamma \in [\frac{S}{S+1}, 1]$ by (10), we conclude that the limit of L has reached its maximum value when $\gamma = \frac{S}{S+1}$, and so $\lim_{N\to\infty} L(N, N\gamma, S)$ is constant in γ for $\gamma \in [\frac{S}{S+1}, 1]$, with value $\log_2(S+1)$.
This concludes the proof of Lemma 2. ∎

The following lemma provides a necessary condition for an M-of-N encoder to bound process (1).

Lemma 3. *Consider an M-of-N encoder/decoder pair using symbols $\{0,\dots,S\}$ with sampling period T. If the pair keeps the state of (1) bounded for every initial condition, then we must have*

$$\frac{\ln L(N,M,S)}{NT} > \sum_{i:\Re\lambda_i[A]\geqslant 0} \lambda_i[A]. \tag{24}$$

□

Proof of Lemma 3. We proceed with a proof by contradiction inspired by [2] which considers the rate at which the initial state volume $\mathcal{X}_0 \subset \mathbb{R}^n$ grows through process (1) and shrinks upon the receipt of information from the encoder. Consider an encoder/decoder/controller triple whose encoder is an M-of-N encoder using symbols $\{0,\dots,S\}$ and sampling period T. For the sake of contradiction, suppose the encoder/decoder/controller triple keeps the state of process (1) bounded for every initial condition, but that

$$\epsilon \coloneqq \sum_{i:\Re\lambda_i[A]\geqslant 0} \lambda_i[A] - \frac{\ln L(N,M,S)}{NT} > 0. \tag{25}$$

We will show that a contradiction arises.

Consider two instants of time t, $t+T$ at which consecutive codewords were received, and let $\mu(\mathcal{X}_t) := \int_{x\in\mathcal{X}_t} dx$ be the volume of the set for which $x(t)$ is known to belong. By the variation of constants formula,

$$x(t+NT) = e^{ANT}x(t) + u_t, \tag{26}$$

where $u_t := \int_t^{t+NT} e^{ANT(t+\tau)}Bu(\tau)d\tau$. Therefore, before the symbol s_{t+NT} is received, $x(t+NT)$ is only known to be in the set $\mathcal{X}^-_{t+NT} := e^A\mathcal{X}_t + u_t$. We are assuming here that the decoder knows the input u that it applied during this period. The volume of $\mathcal{X}^-_{t+NT}$ is then given by

$$\mu(\mathcal{X}^-_{t+NT}) := \int_{x\in e^{ANT}\mathcal{X}_t+u_t} dx = \int_{z\in\mathcal{X}_t} |\det e^{ANT}|\,\mu(\mathcal{X}_t)\,dz = |\det e^{ANT}|\;\mu(\mathcal{X}_t). \tag{27}$$

Here, we made the change of integration variable $z := e^{ANT}(x-u_t)$. Using the fact that $\det e^A = e^{\operatorname{trace} A} = e^{\sum_{i=1}^n \lambda_i[A]}$ for any $n\times n$ matrix A with eigenvalues $\lambda_1[A],\lambda_2[A],...,\lambda_n[A]$, we conclude that

$$\mu(\mathcal{X}^-_{t+NT}) = e^{NT\sum_{i=1}^n \lambda_i[A]}\mu(\mathcal{X}_t). \tag{28}$$

Suppose now that the coding is optimal in the specific sense that it would minimize the volume of the set $\mathcal{X}_{t+NT}$ to which $x(t+NT)$ will be known to belong. To maximize the volume reduction provided by the codeword $\{s_{t+1},\dots,s_{t+NT}\}$, this symbol should allow us to locate $x(t+NT)$ in one of $L(N,M,S)$ disjoint sub-volumes of $\mathcal{X}^-_{t+NT}$. Thus,

$$\mu(\mathcal{X}_{t+NT}) \geqslant \frac{1}{L(N,M,S)}\mu(\mathcal{X}^-_{t+NT}) = \frac{1}{L(N,M,S)}e^{NT\sum_{i=1}^n \lambda_i[A]}\;\mu(\mathcal{X}_t). \tag{29}$$

with equality achievable when $\mathcal{X}^-_{t+NT}$ is evenly divided into $L(N,M,S)$ parts. In other words, the receipt of an N-symbol word from a set of $L(N,M,S)$ possible words can reduce the volume $\mu(\mathcal{X}^-_{t+NT})$ by a factor of no more than $1/L(N,M,S)$.

Iterating this formula from 0 to t, we conclude that

$$\mu(\mathcal{X}_t) \geqslant e^{t\left(\sum_{i=1}^n \lambda_i[A] - \frac{1}{NT}\ln L(N,M,S)\right)}\ \mu(\mathcal{X}_0) = e^{t\epsilon}\mu(\mathcal{X}_0).$$

This means that the volume of sets $\{\mathcal{X}_t\}$ grows to infinity as $t \to \infty$, which in turn means that we can find points in these sets arbitrarily far apart for sufficiently large t. In this case, stability is not possible because the state could be arbitrarily far from the origin regardless of the control signal used. Moreover, any other encoding would also necessarily lead to unbounded volumes and therefore instability because it could do no better than the optimal volume-reduction coding considered. Hence, for each time t, one can find two initial conditions x_0^t and $\bar{x}_0^t$ such that

$$\lim_{t\to\infty} \|\varphi(t;x_0^t) - \varphi(t;\bar{x}_0^t)\| = \infty,$$

where $\varphi(t;x_0)$ denotes the solution to the closed-loop that starts at $x(0) = x_0$. We thus conclude that the encoder-decoder-controller triple does not stabilize the process. ■

Now we are ready to prove Theorem 1.

Proof of Theorem 1. By Lemma 1, for any $\epsilon > 0$ there exist $M, N \in \mathbb{N}_{>0}$ with $M < N\gamma_{\max}(1+\epsilon)$ for which the encoder/decoder is an M-of-N encoder. Since the state of the process is kept bounded, by Lemma 3 we have

$$\sum_{i:\Re\lambda_i[A]\geqslant 0} \lambda_i[A] < \frac{\log_2 L(N,M,S)}{NT}\ \ln 2. \tag{30}$$

Since L is monotonically nondecreasing in its second argument and $M < N\gamma_{\max}(1+\epsilon)$, we have

$$\frac{\log_2 L(N,M,S)}{NT} \leqslant \frac{\log_2 L(N,N\gamma_{\max}(1+\epsilon),S)}{NT}. \tag{31}$$

Lemma 2 implies that

$$\frac{\log_2 L(N,N\gamma_{\max}(1+\epsilon),S)}{NT} \leqslant rf(\gamma_{\max}(1+\epsilon),S). \tag{32}$$

Combining these and letting $\epsilon \to 0$, we obtain (5). This completes the proof of Theorem 1. ■

4 Sufficient condition for boundedness with limited-communication encoders

The previous section established a necessary condition (5) on the average bit-rate and average cost per symbol of an encoder/decoder pair in order to bound the state of (1). In this section, we show that that condition is also sufficient for a bounding encoder/decoder to exist. The proof is constructive in that we provide the encoder/decoder.

Theorem 2. *Assume that A is diagonalizable. For every $S \in \mathbb{N}_{>0}$, $T > 0$, and $\gamma_{\max} \in [0,1]$ satisfying*

$$r f(\gamma_{\max}, S) \ln 2 > \sum_{i:\Re\lambda_i[A]\geqslant 0} \lambda_i[A], \tag{33}$$

where r is defined in (2) and the function f is defined in (6), there exists a controller and an encoder/decoder pair using S nonfree symbols, sampling period T, and average cost per symbol not exceeding $\gamma_{\max}$, that keeps the state of the process (1) bounded for every initial condition $x_0 \in \mathcal{X}_0$. □

The proof of Theorem 2 relies on the following lemma, which provides a sufficient condition for the existence of an M-of-N encoder to bound the state of process (1).

Lemma 4. *Assume that A is diagonalizable. For every $T > 0$ and $N, M, S \in \mathbb{N}_{>0}$ with $N \geqslant M \geqslant 0$ satisfying*

$$\frac{\ln L(N,M,S)}{NT} > \sum_{i:\Re\lambda_i[A]\geqslant 0} \lambda_i[A] \tag{34}$$

there exists an M-of-N encoder on alphabet $\{0,\ldots,S\}$ with sampling period T that keeps the state of the process (1) bounded for every initial condition. □

Proof of Lemma 4. As with Lemma 3, this proof is inspired by [2]. We define an encoder/decoder/controller triple whose encoder is an M-of-N encoder using symbols $\{0,\ldots,S\}$ and sampling period T. The encoder generates its symbol-stream in the following way. It samples the state $x(t)$, places a bounding hyper-rectangle $\mathcal{R}_t^-$ around $\mathcal{X}_t^-$, then partitions this hyper-rectangle into $L(N,M,S)$ sub-rectangles $\mathcal{R}_t^1, ..., \mathcal{R}_t^{L(N,M,S)}$ along the largest axis of $\mathcal{R}_t^-$. The encoder and decoder both contain a bijective mapping between the $L(N,M,S)$ sub-rectangles and the $L(N,M,S)$ codewords of N symbols from alphabet $\{0,\ldots,S\}$ with M or fewer non-free symbols. The encoder then transmits the N-length codeword corresponding to the sub-rectangle that contains $x(t)$. The encoder starts this process at $t = 0$ and repeats it every NT time units.

We show next that all the axes of these hyper-rectangles converge to zero as $t \to \infty$. To this effect, for each $i \in \{1, 2, ..., n\}$ let $a_{i,t}$ denote the size of the ith axis of the hyper-rectangle $\mathcal{R}_t$ and $r_i \in [0,1]$ the average rate (per codeword sent) at which the ith axis is divided into $L(N,M,S)$ pieces.

At time 0, the original bounding hyper-rectangle around $\mathcal{X}_0$ is $\mathcal{R}_0$. Consider how the axes of $\mathcal{R}_0$ change in size over time. At time t, the ith axis of $\mathcal{R}_0$ has expanded by a factor of $e^{\lambda_i t}$ merely from process (1). However, for each N-length codeword (with M or fewer nonfree symbols) sent by the encoder which acts on the ith axis, the ith axis shrinks by a factor of $1/L(N,M,S)$. At time t, the encoder has sent $\lfloor t/NT \rfloor$ of these codewords so the volume of $\mathcal{R}_t$ has shrunk by a factor $1/L(N,M,S)^{\lfloor t/NT \rfloor}$ from the original volume of $\mathcal{R}_0$. But we are interested in the ith axis in particular. The encoder shrinks the ith axis a fraction r_i of the time, so the average rate that the ith

axis shrinks is $1/L(N,M,S)^{r_i\lfloor t/NT\rfloor}$. Hence, the average size of the ith axis of $\mathcal{R}_t$ at time t is

$$e^{\lambda_i t}\ \frac{1}{L(N,M,S)^{r_i\lfloor t/NT\rfloor}}\ a_{i,0} = e^{\lambda_i t - r_i\lfloor \frac{t}{NT}\rfloor \ln L(N,M,S)} a_{i,0}. \tag{35}$$

As $t\to\infty$, the ith axis' average size matches the actual one a_t^i:

$$\limsup_{t\to\infty} e^{\lambda_i t - r_i\lfloor \frac{t}{NT}\rfloor \ln L(N,M,S)} a_{i,0} = \limsup_{t\to\infty} a_{i,t} \qquad i\in\{1,2,...,n\}.$$

We next need to prove that the size of each axis of the bounding hyper-rectangle converges to zero. We prove this by contradiction. Suppose that m of the $a_{i,t}$ axes are not converging to zero (the first m for simplicity) and that the remaining are. This means that, for each $i\in\{1,2,...,m\}$,

$$\limsup_{t\to\infty} e^{\lambda_i t - r_i\lfloor \frac{t}{NT}\rfloor \ln L(N,M,S)} > 0 \tag{36}$$

$$\Leftrightarrow \qquad \limsup_{t\to\infty} t\left(\lambda_i - \frac{r_i}{NT}\ln L(N,M,S)\right) > -\infty \tag{37}$$

$$\Rightarrow \quad \lambda_i - \frac{r_i}{NT}\ln L(N,M,S) \geqslant 0, \tag{38}$$

where the last implication follows because otherwise the limit would be unbounded below. Moreover, $\sum_{i=1}^m r_i = 1$ because these must be the largest axes after some finite time. Therefore, summing over (38), we obtain

$$\frac{\ln L(N,M,S)}{NT} \leqslant \sum_{i=1}^m \lambda_i \leqslant \sum_{i:\Re\lambda_i[A]\geqslant 0} \lambda_i[A], \tag{39}$$

where the last inequality derives from the fact that if axis i is not converging to zero, then λ_i must be unstable or marginally stable. Equation (39) contradicts (34), establishing that each axis of the bounding hyper-rectangle converges to zero. Since $x(t)$ is known to be inside $\mathcal{R}_t$, this means that as $t\to\infty$ the decoder/controller can determine the precise value of the state and drive it to the origin with any stabilizing linear feedback controller. ■

Now we are ready to prove Theorem 2.

Proof of Theorem 2. Assume that S, T, and $\gamma_{\max}$ satisfy (33), so that

$$\epsilon \coloneqq rf(\gamma_{\max},S)\ln 2 - \sum_{i:\Re\lambda_i[A]\geqslant 0} \lambda_i[A] > 0. \tag{40}$$

Equation (11) establishes that $\frac{\ln L(N,N\gamma_{\max},S)}{NT}$ gets arbitrarily close to $rf(\gamma_{\max},S)$ as we increase N, so we pick N sufficiently large to satisfy

$$rf(\gamma_{\max},S)\ln 2 - \frac{\ln L(N,N\gamma_{\max},S)}{NT} < \epsilon/2. \tag{41}$$

By (8), we have $L(N,\lfloor N\gamma_{\max}\rfloor,S) = L(N,N\gamma_{\max},S)$ for every N, $\gamma_{\max}$, and S. Setting $M\coloneqq\lfloor N\gamma_{\max}\rfloor$, then by (40) and (41) we have found N and M satisfying $\frac{\ln L(N,M,S)}{NT} >$

$\sum_{i:\Re\lambda_i[A]\geqslant 0}\lambda_i[A]$. Hence by Lemma 4, there exists an M-of-N encoder/decoder which bounds the state of (1). Since all M-of-N encoder/decoders have average cost per symbol not exceeding M/N, and this encoder/decoder satisfies $M/N \leqslant \gamma_{\max}$, we conclude that this encoder has an average cost per symbol not exceeding $\gamma_{\max}$. This concludes the proof, since we have found the desired encoder. ■

An unexpected consequence of Theorems 1 and 2 is that when it is possible to keep the state of a process bounded with a given bit-rate $r \coloneqq \log_2(S+1)/T$, one can always find M-of-N encoders that bound it for (essentially) the same bit-rate and average cost per symbol not exceeding $S/(S+1)$, i.e., approximately a fraction $1/(S+1)$ of the symbols will not consume communication resources. In the most advantageous case, the encoder/decoder use the alphabet $\{0,1\}$ and the encoder's symbol stream consumes no more than 50% of the communication resources. The price paid for using an encoder/decoder with average cost per symbol near $S/(S+1)$ is that it may require prohibitively long codewords (large N) as compared to an encoder with higher average cost per symbol. This is due to the fact that $\ln L(N, N\gamma_{\max}, S)/NT)$ is monotonically nondecreasing in $\gamma_{\max}$, so that a smaller $\gamma_{\max}$ generally requires a larger N to satisfy (41).

Corollary 1. *If the process (1) can be bounded with an encoder/decoder pair with symbol alphabet $\{0,1,\ldots,S\}$ and sampling period T, then for any $\epsilon > 0$ with $T > \epsilon$, there exists an M-of-N encoder with alphabet $\{0,1,\ldots,S\}$, sampling period $T-\epsilon$, and average cost per symbol not exceeding $S/(S+1)$ that bounds its state.* □

Proof of Corollary 1. Let ϵ be arbitrary in $(0,T)$. Since the controller and encoder/decoder pair bound the system and the average cost per symbol never exceeds 1, by Theorem 1 we have

$$rf(1,S)\ln 2 \geqslant \sum_{i:\Re\lambda_i[A]\geqslant 0}\lambda_i[A], \tag{42}$$

where $r \coloneqq \log_2(S+1)/T$. Since $f(S/(S+1),S) = f(1,S)$ for all $S \in \mathbb{N}_{>0}$, we have

$$\frac{\log_2(S+1)}{T-\epsilon}\, f\left(\frac{S}{S+1},S\right) > r\, f\left(\frac{S}{S+1},S\right) = rf(1,S)\ln 2 \geqslant \sum_{i:\Re\lambda_i[A]\geqslant 0}\lambda_i[A]. \tag{43}$$

By Theorem 2, there exists an encoder/decoder pair with symbol alphabet $\{0,\ldots,S\}$, sampling period $T-\epsilon$, and average cost per symbol not exceeding $S/(S+1)$ which bounds the state of (1). By Lemma 1, this encoder is an M-of-N encoder for appropriately chosen M and N. ■

5 Conclusion and Future Work

In this paper we considered the problem of bounding the state of a continuous-time linear process under communication constraints. We considered constraints on both the average bit-rate and the average cost per symbol of an encoding scheme. Our main contribution was a necessary and sufficient condition on the process and constraints

for which a bounding encoder/decoder/controller exists. In the absence of a limit on the average cost per symbol, the conditions recovered previous work. A surprising corollary to our main result was the observation that one may impose a constraint on the average cost per symbol without necessarily needing to loosen the bit-rate constraint. Specifically, we proved that if a process may bounded with a particular bit-rate, then there exists a (possibly very complex) encoder/decoder that can bound it using no more than 50% non-free symbols on average, yet obeying the *same* bit-rate. This was surprising because one would expect the prohibition of some codewords would require that the encoder always compensate by transmitting at a higher bit-rate.

We observed in Remark 1 that smaller alphabets incur a smaller penalty on the conditions for boundedness in Theorems 1 and 2. This suggests that encoding schemes with small alphabets may be able to bound the state of the process with bit-rates and average costs not far above the minimum theoretical bounds as established in Theorems 1 and 2. *Event-based* control strategies comprise one such class of encoders; they use a small number of non-free symbols to notify the decoder/controller about certain state-dependent events. We have preliminary results showing that event-based encoders can indeed be used to produce encoder/decoder pairs that almost achieve the minimum achievable average cost per symbol that appears in Theorems 1 and 2.

Finally, our problem setup considered merely wherether there exists a bounding encoder/decoder/controller triple. It seems natural to extend this setup to finding stabilizing triples.

A Appendix

Proof of Lemma 1. By the definition of average cost per symbol not exceeding $\gamma_{\max}$ in (3) there exists an integer $N_0 \in \mathbb{N}_{>0}$ such that for any symbol sequence $\{s_k\}$ that the encoder generates, we have

$$\sum_{i=1}^{N} I_{s_i \neq 0} < N_0 + N\gamma_{\max}, \quad \forall N \in \mathbb{N}_{>0}. \tag{44}$$

Pick $N \in \mathbb{N}_{>0}$ large enough to satisfy $N_0 + 2 < \epsilon N \gamma_{\max}$ and pick $M \coloneqq \lfloor N_0 + 2 + N\gamma_{\max} \rfloor$. Combining these with (44), we obtain

$$\sum_{i=1}^{N} I_{s_i \neq 0} < N_0 + N\gamma_{\max} < M \leqslant N_0 + 2 + N\gamma_{\max} < N\gamma_{\max}(1+\epsilon),$$

which establishes that $M < N\gamma_{\max}(1+\epsilon)$. This completes the proof. ∎

Lemma 5. *The following inequality holds for all $N, S \in \mathbb{N}_{>0}$, $q \in (0, S/(S+1)]$, and $i \in [0, Nq]$:*

$$q^i (1-q)^{N-i} \geqslant 2^{-N\,H(q)} \, \frac{S^i}{S^{Nq}} \tag{45}$$

where $H(q) \coloneqq -q \log_2 q - (1-q)\log_2(1-q)$ is the the base-2 entropy of a Bernoulli random variable with parameter q.

Proof of Lemma 5. Let N, S, q, and i take arbitrary values from the sets described in the lemma's statement. Since $\log_2$ is a monotone increasing function, $\log_2(q/(1-q))$ for $q > 0$ is maximized at the right endpoint value, $q = S/(S+1)$, where it equals $\log_2 S$. This leads to

$$\log_2 q - \log_2(1-q) \leqslant \log_2 S \tag{46}$$

for all $S \in \mathbb{N}_{>0}$ and $q \in (0, S/(S+1)]$. Next, since $i \in [0, Nq]$ by assumption, then $i - Nq \leqslant 0$. Multiplying (46) by $i - Nq$, we obtain

$$\begin{aligned}
& (i-Nq)\log_2 q - (i-Nq)\log_2(1-q) \geqslant (i-Nq)\log_2 S \\
\Leftrightarrow \quad & (i-Nq)\log_2 q + (N-i-N+Nq)\log_2(1-q) \geqslant (i-Nq)\log_2 S \\
\Leftrightarrow \quad & i\log_2 q + (N-i)\log_2(1-q) \geqslant Nq\log_2 q + N(1-q)\log_2(1-q) \\
& \qquad + (i-Nq)\log_2 S \\
\Leftrightarrow \quad & i\log_2 q + (N-i)\log_2(1-q) \geqslant -N\left(-q\log_2 q - (1-q)\log_2(1-q)\right) \\
& \qquad + (i-Nq)\log_2 S \\
\Leftrightarrow \quad & i\log_2 q + (N-i)\log_2(1-q) \geqslant -NH(q) + (i-Nq)\log_2 S \\
\Leftrightarrow \quad & q^i(1-q)^{N-i} \geqslant 2^{-NH(q)}\frac{S^i}{S^{Nq}}
\end{aligned}$$

■

References

[1] R. Brockett and D. Liberzon, "Quantized feedback stabilization of linear systems," *Automatic Control, IEEE Transactions on*, vol. 45, no. 7, pp. 1279–1289, Jul 2000. (cited in p. 1)

[2] J. P. Hespanha, A. Ortega, and L. Vasudevan, "Towards the control of linear systems with minimum bit-rate," in *Proc. of the Int. Symp. on the Mathematical Theory of Networks and Syst.*, Aug. 2002. (cited in p. 1, 2, 3, 10, 12)

[3] G. Nair and R. Evans, "Communication-limited stabilization of linear systems," in *Decision and Control, 2000. Proceedings of the 39th IEEE Conference on*, vol. 1, 2000, pp. 1005 –1010 vol.1. (cited in p.)

[4] S. Tatikonda and S. Mitter, "Control under communication constraints," *Automatic Control, IEEE Transactions on*, vol. 49, no. 7, pp. 1056 – 1068, july 2004. (cited in p. 1, 2, 3)

[5] G. N. Nair and R. J. Evans, "Exponential stabilisability of finite-dimensional linear systems with limited data rates," *Automatica*, vol. 39, no. 4, pp. 585 – 593, 2003. (cited in p.)

[6] A. Sahai and S. Mitter, "The necessity and sufficiency of anytime capacity for stabilization of a linear system over a noisy communication link — part i: Scalar systems," *Information Theory, IEEE Transactions on*, vol. 52, no. 8, pp. 3369–3395, Aug 2006. (cited in p.)

[7] A. Matveev and A. Savkin, "Multirate stabilization of linear multiple sensor systems via limited capacity communication channels," *SIAM Journal on Control and Optimization*, vol. 44, no. 2, pp. 584–617, 2005. (cited in p. 1)

[8] K. Astrom and B. Bernhardsson, "Comparison of riemann and lebesgue sampling for first order stochastic systems," in *Decision and Control, 2002, Proceedings of the 41st IEEE Conference on*, vol. 2, dec. 2002, pp. 2011 – 2016 vol.2. (cited in p.)

[9] K. J. Åström, "Event based control," in *Analysis and Design of Nonlinear Control Systems: In Honor of Alberto Isidori*. Springer Verlag, 2007. (cited in p.)

[10] P. Tabuada, "Event-triggered real-time scheduling of stabilizing control tasks," *Automatic Control, IEEE Transactions on*, vol. 52, no. 9, pp. 1680 –1685, sept. 2007. (cited in p.)

[11] T. Cover and J. Thomas, *Elements of Information Theory*. Wiley, 2012. (cited in p. 9)