ERGODICITY CONDITIONS FOR CONTROLLED STOCHASTIC NON-LINEAR SYSTEMS UNDER INFORMATION CONSTRAINTS∗

NICOLÁS GARCIA†, CHRISTOPH KAWAN‡, AND SERDAR YÜKSEL§

Abstract. Consider a stochastic nonlinear system controlled over a possibly noisy communication channel. An important problem is to characterize the largest class of channels for which there exist coding and control policies so that the closed-loop system is stochastically stable. In this paper we consider the stability notion of (asymptotic) ergodicity. We prove lower bounds on the channel capacity necessary to achieve the stability criterion. Under mild technical assumptions, we obtain that the necessary channel capacity is lower bounded by the log-determinant of the linearization, double-averaged over the state and noise space. We prove this bound by introducing a modified version of invariance entropy, and utilizing the almost sure convergence of sample paths guaranteed by the pointwise ergodic theorem. The fundamental bounds obtained generalize well-known formulas for linear systems, and are in some cases more refined than those obtained for nonlinear systems via information-theoretic methods.

Key words. Stochastic stabilization; control under communication constraints; asymptotic mean stationarity; ergodicity; invariance entropy

AMS subject classifications. 93E15, 93C10, 37A35

1. Introduction. In this paper, we consider a stochastic nonlinear system controlled over a possibly noisy communication channel. We consider the problem of determining necessary conditions on channel capacity required for the existence of coding and control policies so that the closed-loop system is stochastically stable. The stability criterion considered is asymptotic ergodicity, by which we mean the existence of an asymptotically mean stationary measure which is also ergodic. Our analysis considers systems of the form

\[ x_{t+1} = f(x_t, w_t) + u_t \]

where \( x_t \) and \( u_t \) take values in \( \mathbb{R}^N \) and \( w_t \) takes values in a probability space. The variables represent the state, noise, and control action at time \( t \), respectively.

In the case of a deterministic system, the notion of invariance entropy has been used to study a related problem, namely stabilization in the sense of set-invariance [3]. The invariance entropy of a compact subset \( Q \) of the state space is defined as

\[ h_{\text{inv}}(Q) := \lim_{\tau \to \infty} \frac{1}{\tau} \log r(\tau, Q), \]

where \( r(\tau, Q) \) is the minimum number of control inputs required to make \( Q \) invariant on the time interval \([0, \tau]\) for arbitrary initial states in \( Q \). The invariance entropy measures the smallest average rate of information that must be transmitted to a controller to render \( Q \) invariant.

The motivation for the above definition arises by observing that with \( n \) bits of

∗Submitted to the editors DATE.
†Is with the Department of Mathematics and Statistics at Queen’s University, Kingston ON, Canada. Email: 11ng31@queensu.ca
‡Is with the Institute of Informatics at Ludwig Maximilian University of Munich, Germany. Email: christoph.kawan@lmu.de
§Is with the Department of Mathematics and Statistics at Queen’s University, Kingston ON, Canada. Email: yuksel@queensu.ca

1 arXiv:1912.06351v1 [math.OC] 13 Dec 2019
information available at the controller side, at most $2^n$ different states can be distinguished, and therefore at most $2^n$ different control inputs can be generated. To ensure that the state “remains stable”, it is natural to require that certain subsets of the state space be made invariant. In the case of stochastic systems, asking for a compact subset of the state space to be invariant is too restrictive to be a useful notion of stability. For example, if the system is subject to unbounded noise, the state process may leave a given compact set regardless of the control decision. As such, we consider instead notions of stochastic stability such as asymptotic mean stationarity (AMS).

In [12], the notion of invariance entropy was generalized for use in the stability analysis of discrete-time stochastic systems controlled over finite-capacity channels. The introduced quantity, called stabilization entropy, was inspired by both invariance entropy and measure-theoretic entropy of dynamical systems, in particular by a characterization of the latter due to Katok [10] and a generalization thereof developed in Ren et al. [20].

Roughly speaking, the stabilization entropy of a subset $Q$ of the state space looks at the exponential growth rate of the number of length-$n$ control sequences required to render $Q$ invariant with a certain positive probability for a certain fraction of times $t$ between 0 and $n$. This concept is closely related to other criteria used in the literature, such as stability in probability [16] and finite $m$-th moment stability.

In the paper at hand, we provide an important refinement for this problem, where our stability criterion is stochastic in nature, but deterministic in its sample path limits; ergodicity of the AMS measure. This stronger notion of stability guarantees the almost sure convergence of sample paths which asymptotically visit each subset of the state space at a frequency given by the AMS measure of the subset. We further generalize the notion of stabilization entropy to a finite partition of a set, and prove stronger results using the pointwise ergodic theorem.

The paper is organized as follows. Section 2 provides a brief literature review. Some fundamental definitions and technical tools are introduced in Section 3. The main results are presented and discussed in Section 4, while their proofs are given in Section 5.

2. Literature review. The problem of determining necessary and sufficient conditions for stochastic stability of Markov chains, in the form of the existence of a stationary measure and positive Harris recurrence, has been studied using Lyapunov methods and we refer the reader to [17] for a comprehensive treatment.

To implement stabilizing control policies however, full feedback is often required, a condition which may be too restrictive in certain cases. For example, the controller may have access only to an estimate of the state encoded in $n$ bits at each time step, in which case the typically uncountable state space must be quantized using a finite ($2^n$) number of symbols. As such, the assumption that the controller has full state access, at arbitrary levels of precision, is no longer valid. In particular, this is the case in large-scale networked systems, where communication resources have to be distributed among many agents, and in underwater applications, where communication is naturally constrained due to the physical properties of the environment. The emergence of such problems has motivated the study of control problems subject to information constraints, and the development of the general theory of information-based control.

In the case of linear systems, formulas have been obtained for the smallest channel capacity above which stabilization is possible. Under certain stability notions, the capacity of the channel must not be smaller than the logarithm of the unstable open-
loop determinant. The earliest contributions can be found in [22] and [1]. These formulas, known as data-rate theorems, were further generalized in [9], [18] and [21]. For a more complete discussion of related results, see [16] and [7].

For nonlinear systems, most of the results in the literature have been obtained for deterministic systems controlled over noiseless channels. To this end, the notion of topological feedback entropy was introduced in [19] for the study of discrete-time systems. A related result, by the same authors, is a characterization of the smallest data rate required for stabilization to an equilibrium point as the log-sum of the unstable eigenvalues of the linearization. For the case of continuous-time systems, the notion of invariance entropy was introduced in [3]. Both topological feedback entropy and invariance entropy capture the smallest average rate of information required to keep the system state inside a compact set. When adapted to the discrete-time setting, the two notions are equivalent, as was shown in [4]. A comprehensive review of these concepts is provided in [11]. A recent related development was the introduction of metric invariance entropy in [2], a notion based on conditionally invariant measures.

Other studies on control of nonlinear systems over communication channels have focused on constructive schemes (and not on converse theorems), primarily for noise-free systems and channels, cf. [15, 5, 14]. Recently, necessary conditions in the form of lower bounds on the channel capacity for a certain class of stochastic nonlinear systems over both noiseless and noisy channels were established in [12] and [26], where the stability notion considered in the first paper is AMS, and the notions considered in the second paper are AMS, ergodicity, and positive Harris recurrence. In [12], the notion of stabilization entropy is used, while [26] relies on information-theoretic techniques, where the different approaches arrive at complementary results.

For a class of nonlinear systems controlled over noiseless channels [26], and for linear systems over Gaussian, discrete noiseless, erasure and discrete noisy channels [27], as well as [23, 24, 25] establish the ergodicity property under information constraints. Therefore, the goal of ergodicity is attainable even for systems with additive unbounded noise.

It is important to note that for linear systems, any local dynamical or control-theoretic property is also a global property. As such, the problems of local stabilization (stabilization to a point), semi-global stabilization (set invariance) and global stabilization (stochastic stability) can all be handled with similar methods, leading to the aforementioned data-rate theorem in each case. For nonlinear systems, however, the three stability problems are fundamentally different and require distinct approaches. For example, linearization techniques work well for local problems, for semi-global problems only under specific assumptions, and almost never for global problems. In addition, the presence of (possibly unbounded and additive) noise requires an approach fundamentally different from the machinery utilized for local stabilization problems.

3. Preliminaries.

3.1. Notation. Throughout the paper, $\mathbb{N}$ denotes the strictly positive integers, $\mathbb{Z}_{\geq 0}$ denotes $\mathbb{N} \cup \{0\}$ and $\mathbb{R}_{>0}$ the strictly positive real numbers. We write $[a:b]$ for a discrete interval, i.e., $[a:b] = \{a, a+1, \ldots, b\}$ for any $a \leq b$ in $\mathbb{Z}$. We write $\mathcal{B}(X)$ for the Borel $\sigma$-algebra of a polish space $X$. Furthermore, $\Sigma$ denotes the space of sequences in a polish space $X$, i.e., $\Sigma = X^{\mathbb{Z}_+}$, and $\mathcal{B}(\Sigma)$ the Borel $\sigma$-algebra of $\Sigma$, which is generated by the cylinder sets. If $x \in X^{\mathbb{Z}_+}$, we write $x_{[0,t]} = (x_0, x_1, \ldots, x_t)$ for any $t \in \mathbb{Z}_+$. By $m$ we denote the Lebesgue measure on $\mathbb{R}^N$ for any $N \in \mathbb{N}$. All logarithms are taken to the base 2.
3.2. Stochastic stability and ergodic properties. In this section, we provide some basic definitions, and characterize the stability criterion considered in this paper: asymptotic mean stationarity with the associated AMS measure resulting in an ergodic state process.

First, recall some basic facts from ergodic theory: A measurable map $T: \Omega \to \Omega$ on a probability space $(\Omega, \mathcal{F}, P)$ is called measure-preserving if $P(T^{-1}(A)) = P(A)$ for all $A \in \mathcal{F}$. An event $A \in \mathcal{F}$ is $T$-invariant if $A = T^{-1}(A)$ (up to a set of measure zero). We denote by $\mathcal{F}_{\text{inv}(T)}$ the set of all $T$-invariant measurable sets, which is a $\sigma$-algebra. A measure-preserving map $T$ is called ergodic if $P(A) \in \{0, 1\}$ for all $A \in \mathcal{F}_{\text{inv}(T)}$.

A measure-preserving map $T$ is called ergodic if $P(A) \in \{0, 1\}$ for all $A \in \mathcal{F}_{\text{inv}(T)}$. Note that ergodicity is a property of a system $(\Omega, \mathcal{F}, P, T)$, but sometimes we also say that “$T$ is ergodic”, or occasionally “$P$ is ergodic”, when the other components of the system are clear from the context.

A fundamental result in ergodic theory is the following pointwise ergodic theorem.

**Theorem 3.1.** Let $(\Omega, \mathcal{F}, P)$ be a probability space and $T: \Omega \to \Omega$ a measure-preserving map. Then for any $f \in L^1(\Omega, \mathcal{F}, P)$ we have

$$
\frac{1}{N} \sum_{k=0}^{N-1} f \circ T^k \xrightarrow{a.s. \quad N \to \infty} \varphi
$$

for some $\varphi \in L^1(\Omega, \mathcal{F}_{\text{inv}(T)}, P|_{\mathcal{F}_{\text{inv}(T)}})$ satisfying $\int \varphi \, dP = \int f \, dP$. If, in addition, $T$ is ergodic, then $\varphi$ is almost everywhere constant and thus

$$
\frac{1}{N} \sum_{k=0}^{N-1} f \circ T^k \xrightarrow{a.s. \quad N \to \infty} \int f \, dP.
$$

In the following, we fix a polish space $X$ and the associated sequence space $\Sigma = X^\mathbb{Z}_+$. The shift map on $\Sigma$ is defined by

$$
\theta: \Sigma \to \Sigma, \quad (\theta x)_t := x_{t+1}, \quad \forall x = (x_t)_{t \in \mathbb{Z}_+} \in \Sigma.
$$

A measure $\mu$ on $(\Sigma, \mathcal{B}(\Sigma))$ is called stationary if $\mu(B) = \mu(\theta^{-1}(B))$ for all $B \in \mathcal{B}(\Sigma)$, i.e., if $(\Sigma, \mathcal{B}(\Sigma), \mu, \theta)$ is a measure-preserving system.

A stochastic process $x = (x_t)_{t \geq 0}$ taking values in $X$ (with underlying probability space $(\Omega, \mathcal{F}, P)$) is called

- stationary if its process measure is stationary, i.e., $P(\{\omega \in \Omega : x(\omega) \in B\}) = P(\{\omega \in \Omega : (\theta x)(\omega) \in B\})$ for all $B \in \mathcal{B}(\Sigma)$.
- asymptotically mean stationary (AMS) if there exists a measure $\tilde{Q}$ on $(\Sigma, \mathcal{B}(\Sigma))$ such that

$$
\lim_{N \to \infty} \frac{1}{N} \sum_{k=0}^{N-1} P(\theta^{-k}(B)) = \tilde{Q}(B) \quad \text{for all } B \in \mathcal{B}(\Sigma).
$$

It can easily be shown that the measure $\tilde{Q}$ on the measurable space $(\Sigma, \mathcal{B}(\Sigma))$ is stationary. We can also obtain a measure $Q$ on $(X, \mathcal{B}(X))$ by projecting $\tilde{Q}$ down to any of its coordinates. It follows that for any $B \in \mathcal{B}(X)$

$$
\lim_{N \to \infty} \frac{1}{N} \sum_{t=0}^{N-1} P(x_t \in B) = Q(B).
$$
Let $\mu$ denote the process measure on $(\Sigma, B(\Sigma))$ and suppose that the system $(\Sigma, B(\Sigma), \mu, \theta)$ is ergodic. Observe that for a set $B \in B(\Sigma)$, the indicator function $1_{x_0 \in B} : \Sigma \mapsto \{0, 1\}$ belongs to $L^1(\Sigma, B(\Sigma), \mu)$. As such, by the pointwise ergodic theorem, we have

$$\mu\left( \left\{ x \in \Sigma : \lim_{N \to \infty} \frac{1}{N} \sum_{k=0}^{N-1} 1_{x_k \in B}(\theta^k(x)) = \int 1_{x_0 \in B}(x) \, d\mu \right\} \right) = 1$$

which we can (using the notation $\mu$ also for the projection of $\mu$ to $\Sigma$) rewrite as

$$\mu\left( \left\{ x \in \Sigma : \lim_{N \to \infty} \frac{1}{N} \sum_{k=0}^{N-1} 1_{x_k \in B}(x) = \mu(B) \right\} \right) = 1. \quad (3.1)$$

Thus, if the stochastic process is ergodic, then the set of sample paths which visit each Borel set $B$ with frequency $\mu(B)$ is of full measure. This is a strong notion of stability, and is a key ingredient in the proofs of the theorems in this paper. Asking ergodicity of the state process in many cases is an overly restrictive condition, however, and it turns out that (3.1) holds for a larger class of processes, which we now explore.

**Definition 3.2.** Consider a stochastic process which is AMS with asymptotic mean $Q$. If $Q$ is ergodic, we call the process AMS ergodic.

**Proposition 3.3.** An AMS ergodic process satisfies an equation similar to (3.1). Namely, for any $B \in B(\Sigma)$ it holds that

$$\mu\left( \left\{ x \in \Sigma : \lim_{N \to \infty} \frac{1}{N} \sum_{k=0}^{N-1} 1_{x_k \in B}(x) = Q(B) \right\} \right) = 1. \quad (3.2)$$

**Proof.** Let us fix a $B \in B(\Sigma)$. By stationarity, we can project $Q$ to the space $\Sigma$. By a slight abuse of notation, we also denote the projected measure by $Q$. We define

$$F := \left\{ x \in \Sigma : \lim_{N \to \infty} \frac{1}{N} \sum_{k=0}^{N-1} 1_{x_k \in B}(\theta_k(x)) = Q(B) \right\}.$$ 

From the ergodicity assumption on $Q$, it follows that $Q(F) = 1$. Also, $F$ is invariant under $\theta$ from which we obtain that $\mu(F) = 1$ (see [8, Lem. 6.3.1 and Eq. (6.22)]).

As such, we see that we do not need an ergodicity assumption on the state process to obtain almost sure convergence of time averages to a constant.

**4. Bounds on channel capacity.** We now state the main contributions of this paper. Proofs can be found in the next section. Consider the system

$$x_{t+1} = f(x_t, w_t) + u_t \quad (4.1)$$

where $x_t$ and $u_t$ are $\mathbb{R}^N$-valued for some $N \in \mathbb{N}$ and $w_t$ takes values in a standard probability space $\mathbb{W}$. For a fixed $w \in \mathbb{W}$, let us denote the map $x \mapsto f(x, w)$ by $f_w$. Suppose also that the following holds:

(i) The map $f : \mathbb{R}^N \times \mathbb{W} \to \mathbb{R}^N$ is Borel measurable.

(ii) The noise process $(w_t)_{t \in \mathbb{Z}_+}$ is i.i.d. By abuse of notation, $v$ denotes both the law of any individual $w_t$, as well as the process measure.

(iii) The map $f_w : \mathbb{R}^N \to \mathbb{R}^N$ is $C^1$ and injective for any $w \in \mathbb{W}$.
(iv) The initial state \( x_0 \) is random, distributed according to the measure \( \pi_0 \).
(v) The measure \( \pi_0 \) is absolutely continuous with respect to the \( N \)-dimensional Lebesgue measure \( m \), and its density (which exists by the Radon-Nikodym theorem) is bounded.
(vi) There exists a constant \( c > 0 \) with \( |\det Df_w(x)| > c \) for all \( x \in \mathbb{R}^N \) and \( w \in W \).

We write \( (\Omega, \mathcal{F}, P) \) for the probability space on which both \( x_0 \) and \( w_t \) are modeled.

We assume that the system is controlled over a possibly noisy communication channel as depicted in Fig. 1. The channel has a finite input alphabet \( M \) and a finite output alphabet \( M' \). The channel input \( q_t \) at time \( t \) is generated by a function \( \gamma_t^e \) so that \( q_t = \gamma_t^e(x_{[0,t]}, q_{[0,t-1]}) \). The channel maps \( q_t \) to \( q_t' \) in a stochastic fashion so that \( P(q_t' \in \cdot | q_t, q_{[0,t-1]}, q_{[0,t-1]}') = P(q_t' \in \cdot | q_t) \) is a conditional probability measure on \( M' \) for all \( t \in \mathbb{Z}_+ \), for every realization \( q_t, q_{[0,t-1]}, q_{[0,t-1]}' \). The controller, upon receiving the information from the channel, generates its decision at time \( t \), also causally: \( u_t = \gamma_t^c(q_{[0,t]}') \). Any coding and control policy of this kind is called causal.

If the channel is noiseless, we have \( M = M' \) and the channel capacity reduces to \( C = \log |M| \).

![Control of a system over a noisy channel with feedback](image)

**Fig. 1.** Control of a system over a noisy channel with feedback

**Theorem 4.1.** Consider system (4.1) satisfying assumptions (i)–(vi). Suppose the system is controlled over a discrete noiseless channel of capacity \( C \) and a coding and control policy achieves that the state process is AMS ergodic. Then the capacity must satisfy

\[
\int \int \log |\det Df_w(x)| \ dQ(x) \ dv(w) \leq C.
\]

Our second main theorem relaxes the condition of the channel being noiseless. On the other hand, the class of nonlinear systems considered will be slightly more restrictive.

**Theorem 4.2.** Consider the scalar additive-noise system

\[
x_{t+1} = f(x_t) + w_t + u_t
\]

satisfying assumptions (i)–(v) hold. Additionally, suppose that the following holds:
(i) \( |f'(x)| \geq 1 \) for every \( x \in \mathbb{R} \).
(ii) The support of \( \pi_0 \) is a compact interval \( K \subseteq \mathbb{R} \).
(iii) The essential infimum and supremum of the density of \( \pi_0 \), denoted by \( \rho_{\text{min}} \) and \( \rho_{\text{max}} \), satisfy \( 0 < \rho_{\text{min}} \leq \rho_{\text{max}} < \infty \).

Suppose that the system is controlled over a discrete memoryless channel with feedback of capacity \( C \) and a causal coding and control policy results in the state process being AMS ergodic. Then the channel capacity must satisfy

\[
\int \log |f'(x)| dQ(x) \leq C.
\]
The first theorem above is a counterpart to [12, Thm. 5.1], where it was shown for systems of the form $x_{t+1} = f(x_t) + w_t + u_t$, without the ergodicity assumption on the AMS measure, that for any Borel set $B$ of finite Lebesgue measure

$$Q(B) \inf_{x \in B} \log |\det Df(x)| \leq C$$

must be satisfied. The second theorem above is a counterpart to [12, Thm. 7.1] without the ergodicity assumption on the AMS measure.

To prove Theorem 4.1 and Theorem 4.2, the stabilization entropy introduced in [12] must be generalized and a technical lemma proven. This is carried out in the next section. Before doing this, we provide a discussion of the theorems.

Observe that our lower bound on channel capacity is $\leq 0$ (and thus vacuous) if $|\det Df_w(x)| \leq 1$ for all $(x, w)$. Recall that the determinant of a square matrix represents the “volume” of the unit cube after it is acted on by the matrix. As such, Theorem 4.1 is only interesting if the system is volume-expansive on some regions of the state space. This is intuitive, since if $f$ is nowhere volume-expansive, it may be possible for the uncontrolled system to have desirable stability properties.

The results obtained here are consistent with those obtained using information-theoretic techniques in [26], but are in fact a strict refinement. A similar converse result on channel capacity was obtained in [26] under the stronger stability criterion of positive Harris recurrence of the closed-loop stochastic process. It reads as follows:

**Theorem 4.3.** ([26, Thm. 4.2]) Consider the system

$$x_{t+1} = f(x_t, w_t) + Bu_t$$

and suppose that the following assumptions hold:

(i) For any fixed $w$, the function $f_w := f(\cdot, w) : \mathbb{R}^N \to \mathbb{R}^N$ is a $C^1$-diffeomorphism.

(ii) There exist $L, M \in \mathbb{R}$ such that $L \leq \log |\det Df_w(x)| \leq M$ for all $x, w \in \mathbb{R}^N$.

Suppose that a stationary coding and control policy (see [26] for a precise definition) is adopted so that under these policies

(i) the Markovian system state and encoder state is positive Harris recurrent (which implies the existence of a unique invariant measure).

(ii) $\limsup_{t \to \infty} h(x_t)/t \leq 0$.

Then the channel capacity must satisfy

$$\int \int \log |\det Df_w(x)| d\nu(w) dQ(x) \leq C.$$

Let us now compare Theorem 4.1 and Theorem 4.3.

- Theorem 4.1 is more general in the sense that it applies to arbitrary causal coding and control policies, not just Markov ones. Moreover, it does not require the assumption of sublinear growth of the differential entropy of the state process. Theorem 4.3 assumes that the state process is positive Harris recurrent which implies unique ergodicity, while Theorem 4.1 only assumes ergodicity of the AMS measure.

- On the other hand, Theorem 4.3 considers a more general class of channels (involving memory), more general class of system models as well as systems taking values in higher dimensions.

5. **Proofs.** In this section, we introduce and develop some machinery required to prove our two main theorems. We begin by generalizing the notion of stabilization entropy and proving a technical lemma.
5.1. Generalizing stabilization entropy. Consider system (4.1) without fixing a control policy and observe that given a control sequence \( u := (u_t)_{t \geq 0} \), a noise realization \( w := (w_t)_{t \geq 0} \), and an initial state \( x_0 \in \mathbb{R}^N \), the trajectory \( x := (x_t)_{t \geq 0} \in (\mathbb{R}^N)^{\mathbb{Z}_+} \) of the state is uniquely determined. Let us denote this trajectory by \( \phi(\cdot, x_0, u, w) \) so that for any \( t \in \mathbb{Z}_+ \), \( x_t = \phi(t, x_0, u, w) \).

We want to find a subset of control sequences that allow to render certain subsets of the state space invariant in a probabilistic sense. This leads us to the next definitions of spanning sets and stabilization entropy for finite collections of subsets of \( \mathbb{R}^N \) and \( W \), respectively, which generalize similar notions in [12] where a single set was considered.

**Definition 5.1.** Let \( B \in \mathcal{B}(\mathbb{R}^N) \) and \( D \in \mathcal{B}(W) \) be finite disjoint unions of Borel sets \( B_1, \ldots, B_n \) and \( D_1, \ldots, D_m \), respectively. Let also \( R \) denote a collection of numbers \( r_{k,l} \in [0, 1] \) for \( k \in \{1, \ldots, n\} \) and \( l \in \{1, \ldots, m\} \) satisfying

\[
1 - r := \sum_{k=1}^{n} \sum_{l=1}^{m} (1 - r_{k,l}) \in [0, 1].
\]

Fix \( T \in \mathbb{Z}_{\geq 0} \) and \( \rho \in [0, 1] \). A set of control sequences \( S \subseteq (\mathbb{R}^N)^T \) is called \((T, B, D, \rho, R)\)-spanning if there exists \( \tilde{\Omega} \in \mathcal{F} \) such that the following conditions hold:

- \( P(\tilde{\Omega}) \geq 1 - \rho \).
- For each \( \omega \in \tilde{\Omega} \), there exists a control sequence \( u \in S \) such that

\[
\frac{1}{T} |\{t \in [0; T - 1] : (\phi(t, x_0(\omega), u, w(\omega)), w_t(\omega)) \in B_k \times D_l\}| \geq 1 - r_{k,l}
\]

holds for all \( k \) and \( l \).

Note that we abuse notation by calling a set \((T, B, D, \rho, R)\)-spanning instead of \((T, (B_k)_{k=1}^{n}, (D_l)_{l=1}^{m}, \rho, R)\)-spanning. When doing so, there is the underlying assumption that some partitions of \( B \) and \( D \) are fixed. No confusion should arise, since we explicitly define the partitions whenever we use the definition.

In the above definition, the fact that all random variables are modeled on a common probability space ensures that given \( \omega \), the initial state and the noise sequence of length \( T \) are deterministic. Intuitively speaking, a subset of control sequences of length \( T \) is \((T, B, D, \rho, R)\)-spanning if the probability that, for all \( k, l \), we can maintain the state variable in \( B_k \) and the noise variable in \( D_l \) for at least \( 1 - r_{k,l} \) percent of the time, is at least \( 1 - \rho \). We want to use the size of spanning sets to quantify the difficulty of a control task, which leads to the next definition.

**Definition 5.2.** For the system (4.1), we define the \((B, D, \rho, R)\)-stabilization entropy by

\[
h(B, D, \rho, R) := \limsup_{T \to \infty} \frac{1}{T} \log s(T, B, D, \rho, R),
\]

where \( s(T, B, D, \rho, R) \) denotes the smallest cardinality of a \((T, B, D, \rho, R)\)-spanning set. We define this quantity to be \( \infty \) if no or no finite spanning set exists.

It is obvious that finite \((T, B, D, \rho, R)\)-spanning sets need not exist. As we will see however, they do exist in desired scenarios.

The following lemma is instrumental to prove Theorem 4.1.
Lemma 5.3. Consider system (4.1) with the assumptions of Theorem 4.1 (i.e., a coding and control policy exists over a noiseless channel of capacity $C = \log |\mathcal{M}|$ which makes the state process AMS ergodic). Let now

- $B := \bigcup_{k=1}^{n} B_k \in \mathcal{B}(\mathbb{R}^N)$ and $D := \bigcup_{l=1}^{m} D_l \in \mathcal{B}(\mathbb{W})$ be finite disjoint unions of Borel sets,
- $\rho \in (0, 1)$ be arbitrary.

Next, define the sequence of numbers $R_{\epsilon} := (r_{k,l})_{1 \leq k \leq n, 1 \leq l \leq m}$, where

$$
  r_{k,l} := \begin{cases} 
  (1 + \epsilon)(1 - Q(B_k)v(D_l)) & \text{if } Q(B_k)v(D_l) \in (0, 1) \\
  1 & \text{if } Q(B_k)v(D_l) = 0 \\
  \epsilon & \text{if } Q(B_k)v(D_l) = 1 
  \end{cases}
$$

and observe that for $\epsilon > 0$ small enough, the following conditions are satisfied:

(i) $1 - r := \sum_{k=1}^{n} \sum_{l=1}^{m} (1 - r_{k,l}) \in [0, 1]$.
(ii) $1 - (1 + \epsilon)(1 - Q(B_k)v(D_l)) \in (0, 1)$ for all $k, l$ with $Q(B_k)v(D_l) \in (0, 1)$.

Thus for such a small $\epsilon$, the generalized stabilization entropy $h(B, D, \rho, R_{\epsilon})$ is well-defined. (Of course, $r$ and the $r_{k,l}$’s are $\epsilon$-dependent, but we drop this from the notation.) Then for all $\epsilon > 0$ further small enough the capacity must satisfy

$$
  h(B, D, \rho, R_{\epsilon}) \leq C. 
$$

Proof. We distinguish two cases.

Case 1: We can remove the trivial sets with zero measure from the collections $\{B_k\}$ and $\{D_l\}$ and thus assume that $Q(B_k)v(D_l) > 0$ for all $(k, l)$. Indeed, if a spanning set can be found for the new collections, it is still spanning for the original ones. If $Q(B_k)v(D_l) = 1$ for some $(k, l)$, all the other Cartesian products have measure zero and we can remove them from the collection. Hence, this case reduces to the analysis of a single set as worked out in [12].

Case 2: We continue by considering the case where $Q(B_k)v(D_l) \in (0, 1)$ for all $k, l$. Let $\epsilon > 0$ be small enough such that conditions (i) and (ii) are satisfied and $\epsilon < \rho$. We will show that for any such $\epsilon$, the claim holds.

Let us denote the process measure by $\mu$, and note that by assumption it is AMS. Let $Q$ denote the asymptotic mean, which is by assumption ergodic. As $Q$ is stationary, we can project it unambiguously to a measure on $(\mathbb{R}^N, \mathcal{B}(\mathbb{R}^N))$. By a slight abuse of notation, we denote by $Q$ both the AMS measure and its projection. Let us consider some Borel set $C \subset \mathbb{R}^N$ and let $f : (\mathbb{R}^N)^{\mathbb{Z}_{\geq 0}} \to \mathbb{R}$ be defined by $f((x_t)_{t \geq 0}) := 1_{C}(x_0)$.

It is obvious that this function is in $L^1((\mathbb{R}^N)^{\mathbb{Z}_{\geq 0}})$ (with either $Q$ or $\mu$ as the measure). Recalling our ergodicity assumption, the pointwise ergodic theorem tells us that

$$
  \frac{1}{N} \sum_{j=0}^{N-1} f \circ \theta^j \frac{Q^{a.s.}}{N \to \infty} \int f \, dQ = \int 1_{C}(x) \, dQ(x) = Q(C).
$$

Crucially however, the above convergence also happens $\mu$-almost surely (see (3.2) or [8, Thm. 7.4.1]). Now, for any $B \in \mathcal{B}(\mathbb{W})$, it is clear by the i.i.d. property that

$$
  P\left(\left\{ \omega \in \Omega : \lim_{N \to \infty} \frac{1}{N} \sum_{t=0}^{N-1} 1_{V(w_t(\omega)) = v(V)} \right\} \right) = 1.
$$
As such, noting that $x_t$ and $w_t$ are independent at each time step $t$, it follows that

$$P\left(\left\{ \omega \in \Omega : \lim_{N \to \infty} \frac{1}{N} \sum_{t=0}^{N-1} \mathds{1}_{B_k}(x_t(\omega)) \cdot \mathds{1}_{D_l}(w_t(\omega)) = Q(B_k)v(D_l), \ \forall k, l \right\} \right) = 1.$$

Let us denote the full measure set, where this convergence happens, by $\hat{\Omega}$.

We continue by defining the events

$$E_i^j := \left\{ \omega \in \Omega : \left| \frac{1}{N} \sum_{t=0}^{N-1} \mathds{1}_{B_k}(x_t(\omega)) \cdot \mathds{1}_{D_l}(w_t(\omega)) - Q(B_k)v(D_l) \right| < \frac{1}{i} \right\},$$

$$E := \bigcap_{i=1}^{\infty} \bigcup_{j=1}^{\infty} E_i^j.$$

It is not hard to see that $\hat{\Omega} \subseteq E$, hence $P(E) = 1$. Furthermore, observe that $E$ is an infinite intersection of “decreasing” sets (in the containment sense). Hence,

$$P\left(\bigcap_{j=1}^{\infty} E_i^j\right) = 1 \quad \text{for all } i \in \mathbb{Z}_{>0}.$$

Let now $I_0$ be large enough such that

$$\frac{1}{I_0} \leq \epsilon(1 - Q(B_k)v(D_l)) \quad \text{for all } k \in \{1, \ldots, n\}, \ l \in \{1, \ldots, m\}$$

and observe that $E_{I_0}^1 \subseteq E_{I_0}^2 \subseteq E_{I_0}^3 \subseteq \cdots$. By continuity of probability, we have

$$\lim_{j \to \infty} P(E_{I_0}^j) = P\left(\bigcup_{j=1}^{\infty} E_{I_0}^j\right) = 1,$$

and thus there exists $J_0$ such that $P(E_{I_0}^j) \geq 1 - \epsilon$ for all $j \geq J_0$. For an arbitrary $T \geq J_0$, we define the set of control sequences

$$S_T := \{u_{[0;T-1]}(\omega) : \omega \in E_{I_0}^T\}.$$

We claim that this set is $(T, B, D, \rho, R_v)$-spanning. We use the set $\hat{\Omega}_T := E_{I_0}^T \in \mathcal{F}$ to show this, where we note that it satisfies $P(\hat{\Omega}_T) \geq 1 - \epsilon > 1 - \rho$, as required. For every $\omega \in \hat{\Omega}_T$ and all $k, l$ the control sequence $u_{[0;T-1]}(\omega)$ results in the joint state-noise process satisfying

$$P\left| \frac{1}{T} \sum_{t=0}^{T-1} \mathds{1}_{B_k}(x_t(\omega)) \cdot \mathds{1}_{D_l}(w_t(\omega)) - Q(B_k)v(D_l) \right| \leq \frac{1}{I_0} \leq \epsilon(1 - Q(B_k)v(D_l)).$$

To prove the claim, it now suffices to show that for all $\omega \in \hat{\Omega}_T$ and $k, l$ we have

$$\frac{1}{T}\{|t \in [0; T - 1] : (\varphi(t, x_0(\omega), u_{[0;T-1]}(\omega), \omega), \omega(t)) \in B_k \times D_l\} \geq 1 - (1 + \epsilon)(1 - Q(B_k)v(D_l)) = (1 + \epsilon)Q(B_k)v(D_l) - \epsilon.$$
This, however, follows directly from (5.2). Also, since the coding and control policy can generate at most \(|M|^T\) distinct control sequences by time \(T\), it follows that 
\[ |S_T| \leq |M|^T, \]
therefore \(s(T, B, \rho, R_c) \leq |M|^T\). Recalling that \(T \geq J_0\) was arbitrary, we find that
\[
\log s(T, B, R, \rho, R_c) \leq T \log |M| = TC
\]
and therefore dividing by \(T\) and letting \(T \to \infty\) yields the desired capacity bound (5.1), which completes the proof.

5.2. Proof of Theorem 4.1.

Proof. Let \(c \in (0, 1)\) be such that \(c < |\det Df_w(x)|\) for all \(x \in \mathbb{R}^N\) and \(w \in W\). Let also \(\delta > 0\) and \(\rho \in (0, 1)\) be arbitrary. Next, fix a partition of a Borel set \(B \subset \mathbb{R}^N\) and let \(D = W\), respectively; let \((B_k)_{k=1}^n\) be a partition of \(B\) and \((D_l)_{l=1}^m\) a partition of \(D\). Suppose that \(B\) has finite Lebesgue measure and
\[
Q(B) > 1 - \frac{\delta}{2|\log c|},
\]
where \(Q\) denotes the asymptotic mean of the state process. Let \(\epsilon > 0\) be small enough such that Lemma 5.3 holds, resulting in
\[
h(B, D, \rho, R_c) \leq C,
\]
where \(R_c\) is the associated collection of \(r_{k,l}\)'s as defined in Lemma 5.3. Let also \(1 - r := \sum (1 - r_{k,l})\). It is easy to see that
\[
r = 1 - (1 + \epsilon)Q(B) + n\epsilon
\]
(or \(r = \epsilon\) if one of the \(B_k \times D_l\) has full \(Q \times \nu\)-measure) thus we see that for every sufficiently small \(\epsilon\), we have
\[
2r < \frac{\delta}{|\log c|}.
\]
Now fix a sufficiently large \(T \in \mathbb{N}\) and let \(S\) be a finite \((T, B, D, \rho, R_c)\)-spanning set (whose existence is guaranteed by the proof of Lemma 5.3) with \(\hat{\Omega} \in \mathcal{F}, P(\hat{\Omega}) \geq 1 - \rho\), the associated subset of \(\Omega\). Also let
\[
A := \{(w(\omega), x_0(\omega)) : \omega \in \hat{\Omega}\},
\]
\[
A(u) := \{(w, x) \in W^{Z_+} \times \mathbb{R}^N : \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{1}_{B_k \times D_l}(\varphi(t, x, u, w), w_t) \geq 1 - r_{k,l}, \forall k,l\}
\]
\[
A(u, w) := \{x \in \mathbb{R}^N : (w, x) \in A(u)\}
\]
and observe that
\[
A \subseteq \bigcup_{u \in S} A(u).
\]
By the theorem of Fubini-Tonelli, we have
\[
(v \times m)(A(u)) = \int m(A(u, w)) \, dv(w).
\]
Let us now define a set consisting of disjoint collections of subsets of \( \{0, \ldots, T-1\} \):

\[
A := \{ \Lambda = \{ \Lambda_{k,l} \} \}_{k,l} : \bigcup_{k=1}^{n} \bigcup_{l=1}^{m} \Lambda_{k,l} \subseteq \{0, \ldots, T-1\}, |\Lambda_{k,l}| \geq (1 - r_{k,l})T, \forall k = 1, \ldots, n, l = 1, \ldots, m \}
\]

and note that as a consequence of the definition, \( |\bigcup_{k=1}^{n} \bigcup_{l=1}^{m} \Lambda_{k,l}| \geq (1 - r)T \) for all \( \Lambda \in A \). For such a \( \Lambda \), define the set

\[
A(u, w, \Lambda) := \{ x \in \mathbb{R}^N : (\varphi(t, x, u, w), w) \in B_k \times D_l \leftrightarrow t \in \Lambda_{k,l} \ \text{for all} \ k, l \}
\]

and also (writing \( \varphi(t, u, w)(\cdot) := \varphi(t, \cdot, u, w) \))

\[
A_t(u, w, \Lambda) = \varphi_{t,u,w}(A(u, w, \Lambda)), \quad t = 0, 1, \ldots, T-1.
\]

It is not hard to see that \( A(u, w) = \bigcup_{\Lambda \in A} A(u, w, \Lambda) \) is a disjoint union, implying

\[
\text{(5.6)} \quad m(A(u, w)) = \sum_{\Lambda \in A} m(A(u, w, \Lambda)).
\]

If \( M > 0 \) is an upper bound for the density of \( \pi_0 \), it follows that

\[
\text{(5.7)} \quad 1 - \rho \leq (v \times \pi_0)(A) \leq M \cdot (v \times m)(A)
\]

and

\[
A_t(u, w, \Lambda) \subseteq B_k \quad \text{whenever} \quad t \in \Lambda_{k,l}, \ \forall \ k \in \{1, \ldots, n\}, \ l \in \{1, \ldots, m\}.
\]

Next, we define the following numbers:

\[
c_{k,l} := \inf_{(x, w) \in B_k \times D_l} |\text{det} \ Df_w(x)|.
\]

Recalling the fact that \( f_w \) is injective and \( C^1 \), for all \( (k, l) \) we have

\[
m(A_{t+1}(u, w, \Lambda)) \geq c_{k,l} \cdot m(A_t(u, w, \Lambda)) \quad \text{whenever} \quad t \in \Lambda_{k,l},
\]

\[
m(A_{t+1}(u, w, \Lambda)) \geq c \cdot m(A_t(u, w, \Lambda)) \quad \text{whenever} \quad t \notin \bigcup \Lambda_{k,l}.
\]

Letting \( t^*_k(\Lambda) := \max \Lambda_{k,l}, t^*(\Lambda) := \max_{k,l} t^*_k \), and applying the above inequalities repeatedly, it is not hard to see that

\[
m(A(u, w, \Lambda)) \left( \prod_{k=1}^{n} \prod_{l=1}^{m} |\Lambda_{k,l}|^{-1} \right) c^{T+n m} \leq m(A_{t^*(\Lambda)}(u, w, \Lambda)).
\]

Recall that \( c \leq c_{k,l} \). Now in principle, all the exponents of the \( c_{k,l} \)'s should be \( |\Lambda_{k,l}| \), except for possibly one which should be \( |\Lambda_{k,l}| - 1 \). We do not know which one though, so we write the weaker inequality as above. Combining this with (5.4), (5.5), (5.6) and (5.7), we obtain

\[
\frac{1}{M} (1 - \rho) \leq (v \times m)(A)
\]

\[
\leq |S| \max_{u \in S} (v \times m)(A(u))
\]
In the last inequality we used that

\[ \text{This inequality holds for every } T \]

Moreover, all of these sets are contained in \( B \). Hence, all of these sets are contained in \( B \). Hence,

\[ \sum_{\Lambda \in A : t^*(\Lambda_k,l) = t_k,l \forall k,l} m(A_{t^*(\Lambda)}(u,w,\Lambda)) \leq m(B), \]

which, together with the above chain of inequalities, implies

\[ \frac{1}{M}(1 - \rho) \leq |S| \cdot m(B) \cdot c^{-(2rT + nm)} \cdot \prod_{k=1}^{n} \prod_{l=1}^{m} c_{k,l}^{-(1 - r_{k,l}) T - 1} \prod_{k=1}^{n} \prod_{l=1}^{m} (r_{k,l} T + 1). \]

Since this inequality holds for every \( T \) sufficiently large, we can take logarithms on both sides, divide by \( T \) and let \( T \to \infty \). This results in

\[ 0 \leq h(B, D, \rho, R) - 2r \log c - \sum_{k=1}^{n} \sum_{l=1}^{m} (1 - r_{k,l}) \log c_{k,l}. \]
Recalling the definition of $r_{k,l}$, the fact that $\epsilon$ can be chosen arbitrarily small and (5.3), this leads to the estimate

$$C + \delta \geq \sum_{k=1}^{n} \sum_{l=1}^{n} Q(B_k) v(D_l) \inf_{(x,w) \in B_k \times D_l} \log |\det Df_w(x)|.$$ 

Considering the supremum of the right-hand side over all finite measurable partitions of $B$ and $W$ leads to

$$C + \delta \geq \int \int 1_B(x) \log |\det Df_w(x)| \, dQ(x) \, dv(w),$$

where we use that the integrand is uniformly bounded below by $\log c$ (and hence, we can assume that it is non-negative). Considering now an increasing sequence of sets $B_k \subset \mathbb{R}^N$ whose union is $\mathbb{R}^N$, we can invoke the theorem of monotone convergence to obtain the desired estimate, observing that $\delta$ can be made arbitrarily small as $B_k$ becomes arbitrarily large.

5.3. Proof of Theorem 4.2.

Proof. Suppose for a contradiction that a causal coding and control policy is such that the state process is AMS ergodic, but that the converse of inequality (4.2) holds. Let $r > 0$ be small enough so that

$$C < (1 - 3r) \int \log |f'(x)| \, dQ(x).$$

Since we can approximate the integral by the integral over associated step functions, for any $b \in \mathbb{N}$ large enough there exists a disjoint collection of intervals $B_1, \ldots, B_{2^{b+1}}$ such that $B := [-b,b] = \bigsqcup_{k=1}^{2^{b+1}} B_k$, and

$$C < (1 - 3r) \sum_{k=1}^{2^{b+1}} Q(B_k) \log c_k,$$

where $c_k := \inf_{x \in B_k} |f'(x)|$. Put $n := 2^{b+1} + 1$, and fix a $b$ (and the associated collection $(B_k)_{k=1}^{n-1}$ of intervals) further large enough such that

$$Q([-b,b])(1 - r) > 1 - \frac{2.5}{2^r},$$

which is possible by continuity of probability. Finally, let $B_n$ denote the set of remaining real numbers not covered by the intervals so that $\mathbb{R} = \bigsqcup_{k=1}^{n} B_k$. Next, we define the following sets in a slightly different manner than in the previous proofs:

$$A_T(u,w) := \{x \in \mathbb{R} : \forall k \in \{1, \ldots, n\} \text{ and } \forall N \in \{T(1 - 3r), \ldots, T\}, \frac{1}{N} |\{t \in [0; N-1] : \varphi(t,x,u,w) \in B_k\}| \geq Q(B_k)(1 - r)\}.$$ 

Let also

$$\overline{A}_T(u,w) := [\inf A_T(u,w), \sup A_T(u,w)]$$

and let $x_0(T,u,w)$ denote the midpoint of the above interval (as we will see, this interval is bounded so the midpoint is well defined). We claim that for given $u$ and $w$, 


a $T$ larger than some threshold $M_1(r)$ (which as we will see does not depend on the choice of $u$ and $w$), and any $x_1, x_2 \in A_T(u, w)$, there exists a $t^*$ with $[(1 - 2.5r)T] \leq t^* \leq T - 1$ satisfying

$$\varphi(t^*, x_i, u, w) \in B \text{ for } i \in \{1, 2\}.$$ 

To see this, suppose otherwise. Then for any $i \in \{1, 2\}$ we have

$$|\{t \in [0; T - 1] : \varphi(t, x_i, u, w) \in B\}| \leq [(1 - 2.5r)T] + \frac{1}{2}(T - [(1 - 2.5r)T])$$

$$\leq (1 - 2.5r)T + 1 + \frac{1}{2}(T - ((1 - 2.5r)T + 1))$$

$$= \frac{1}{2}((1 - 2.5r)T + 1) + \frac{1}{2}T$$

$$= \frac{1}{2} + (1 + (1 - 2.5r))\frac{1}{2}T$$

$$= \frac{1}{2} + \left(1 - \frac{2.5}{2}r\right)T < (1 - r)Q(B)T$$

where the last inequality holds for $T$ large enough from the assumption (5.8) on $Q(B)$.

This is a contradiction to $x_i \in A_T(u, w)$, which follows by recalling the definition of $A_T(u, w)$. Let now $\epsilon > 0$ and $\delta > 0$ be given. By the pointwise ergodic theorem (see the construction in the proof of Lemma 5.3), there exists an $M_2 := M_2(\epsilon, \delta) \in \mathbb{N}$ such that for all $T \geq M_2$

$$P(\{\omega \in \Omega : \forall k \in \{1, \ldots, n\}, \forall N \geq (1 - 2.5r)T, \frac{1}{N} \sum_{t=0}^{N-1} \mathbb{1}_{B_k}(x_t(\omega)) \geq Q(B_k)(1 - \delta)\}) > 1 - \epsilon.$$ 

We denote by $\hat{\Omega}(\epsilon, \delta, M_2)$ the set of $\omega$’s for which the event within the brackets of the above expression occurs.

Letting $u, w$ and $x_1, x_2 \in A_T(u, w)$ be arbitrary, we have

$$(5.9) \quad |x_1 - x_2| \leq \frac{2b}{\prod_{k=1}^{n} c_k^{Q(B_k)(1 - \delta)T}} \leq \frac{2b}{\prod_{k=1}^{n} c_k^{Q(B_k)(1 - \delta)T(1 - 2.5r)}}$$

which follows by noting that

$$\prod_{k=1}^{n} c_k^{(1 - \delta)Q(B_k)(1 - 2.5r)T} |x_1 - x_2|$$

$$\leq \prod_{k=1}^{n} c_k^{(1 - \delta)Q(B_k)(1 - 2.5r)T} |x_1| + \prod_{k=1}^{n} c_k^{(1 - \delta)Q(B_k)(1 - 2.5r)T} |x_2|$$

$$\leq \prod_{k=1}^{n} c_k^{(1 - \delta)Q(B_k)t^*} |x_1| + \prod_{k=1}^{n} c_k^{(1 - \delta)Q(B_k)t^*} |x_2|$$

$$\leq |\varphi(t^*, x_1, u, w)| + |\varphi(t^*, x_2, u, w)| \leq 2b.$$ 

A moment’s thought makes it clear that if $\omega \in \hat{\Omega}(\epsilon, \delta, M_2)$, then $x_0(\omega) \in A_T(u(\omega), w(\omega))$ for all $T \geq M_2$. As such and combining with (5.9), we conclude
that
\[ |x_0(\omega) - x_0(T, u(\omega), w(\omega))| \leq \frac{b}{\prod_{k=1}^{n} c_k^{(1-\delta)Q(B_k)(1-2.5r)T}} \]

for every \( T \geq M_2(\epsilon, \delta) \) and every \( \omega \in \hat{\Omega}(\epsilon, \delta, M_2) \). Letting \( \delta \) be small enough so that \( (1 - 3r) \leq (1 - 2.5r)(1 - \delta) \), we conclude that
\[ \liminf_{T \to \infty} P\left( \left\{ \omega \in \Omega : |x_0(\omega) - x_0(T, u(\omega), w(\omega))| \leq \frac{b}{\prod_{k=1}^{n} c_k^{Q(B_k)(1-3r)T}} \right\} \right) \geq 1 - \epsilon. \]

and since \( \epsilon > 0 \) was also arbitrary, it follows that
\[ \limsup_{T \to \infty} P\left( \left\{ \omega \in \Omega : |x_0(\omega) - x_0(T, u(\omega), w(\omega))| > \frac{b}{\prod_{k=1}^{n} c_k^{Q(B_k)(1-3r)T}} \right\} \right) = 0. \]

We will see that our initial hypothesis leads to a contradiction with the above equation. To this effect, let us choose \( \alpha \in (0, 1/2) \) small enough so that
\[ \limsup_{L \to \infty} \left( 1 - \frac{\rho_{\min} \cdot (1 - \alpha)}{2 \cdot \rho_{\max}} + \frac{\rho_{\max}^2}{2L \rho_{\min}^2} + \frac{2 \cdot \rho_{\max} \cdot \alpha}{\rho_{\min} - 1 - \alpha} \right) < 1. \]

Let also \( \hat{\Omega} \in F \) be such that \( P(\hat{\Omega}) > 1 - \alpha \), and such that for all \( T \) large enough (say, bigger than \( C(\alpha) \)),
\[ |x_0(\omega) - x_0(T, u(\omega), w(\omega))| \leq \frac{b}{\prod_{k=1}^{n} c_k^{Q(B_k)(1-3r)T}} \]

for all \( \omega \in \hat{\Omega} \). The idea from here on is to treat \( \hat{\Omega} \) as “the universe”, since conditioning on this set gives the above deterministic bound. We proceed by defining
\[ U_T := \{ (\gamma_0(q_{00}'), \ldots, \gamma_{T-1}(q_{0T-1}')) \in U^T : q_{0T-1}' \in (\mathcal{M}')^T \}, \]
\[ \hat{U}_T := \{ (\gamma_0(q_{00}(\omega)), \ldots, \gamma_{T-1}(q_{0T-1}(\omega))) \in U^T : \omega \in \hat{\Omega} \} \]

and also
\[ R := \limsup_{T \to \infty} \frac{1}{T} \log |U_T|, \quad \hat{R} := \limsup_{T \to \infty} \frac{1}{T} \log |\hat{U}_T|. \]

We now treat two distinct cases.

**Case 1:** Let us suppose that
\[ \hat{R} < (1 - 3r) \sum_{k=1}^{n} Q(B_k) \log c_k. \]

Let \( \epsilon > 0 \) be small enough so that \( \hat{R} + 2\epsilon < (1 - 3r) \sum_{k=1}^{n} Q(B_k) \log c_k \) and observe that for all \( T \) large enough,
\[ |\hat{U}_T| \leq 2^{(\hat{R}+\epsilon)T}. \]
Recall also that $\tilde{\Omega}$ is such that for all $T$ large enough,

$$|x_0(\omega) - x_0(T, u(\omega), w(\omega))| \leq \frac{b}{\prod_{k=1}^{n} Q(B_k)(1 - 3r)T}$$

for all $\omega \in \tilde{\Omega}$. We now fix a noise realization $w$. For all $T$ large enough so that (5.13) holds,

$$m\left(\bigcup_{u \in U_T} \mathcal{A}_T(u, w)\right) \leq \frac{2b \cdot 2^{(R+\epsilon)T}}{\prod_{k=1}^{n} Q(B_k)(1 - 3r)T} \leq \frac{2b \cdot 2^{((1 - 3r) \sum_{k=1}^{n} Q(B_k) \log c_k) - \epsilon)T}}{\prod_{k=1}^{n} Q(B_k)(1 - 3r)T} \leq \frac{2b \cdot 2^{-T \sum_{k=1}^{n} Q(B_k) \log c_k}}{\prod_{k=1}^{n} Q(B_k)(1 - 3r)T} = \frac{2b}{2^{2T}},$$

where the inequalities follow by applying the union bound, and from (5.13) and (5.12).

The above yields

$$\lim_{T \to \infty} m\left(\bigcup_{u \in U_T} \mathcal{A}_T(u, w)\right) = 0,$$

and thus by the absolute continuity and boundedness assumptions of $\pi_0$, we have

$$\lim_{T \to \infty} \pi_0\left(\bigcup_{u \in U_T} \mathcal{A}_T(u, w)\right) = 0.$$

On the other hand, let us define $J := \{w \in \mathbb{R}^n : P(\omega \in \tilde{\Omega} | w(\omega) = w) > 0\}$. We note that $J$ is the projection of $\tilde{\Omega}$ onto $\mathbb{R}^n$ from which the set $\{w : P(\omega \in \tilde{\Omega} | w(\omega) = w) = 0\}$ is taken out: these ensure that $J$ is a universally measurable set since the image of a Borel set under a measurable map is universally measurable [6].

We can therefore write

$$\limsup_{T \to \infty} P\{\omega \in \Omega : |x_0(\omega) - x_0(T, u(\omega), w(\omega))| \leq \frac{b}{\prod_{k=1}^{n} Q(B_k)(1 - 3r)T} | \omega \in \tilde{\Omega}\} = \limsup_{T \to \infty} \left( P\{\omega \in \Omega : |x_0(\omega) - x_0(T, u(\omega), w(\omega))| \leq \frac{b}{\prod_{k=1}^{n} Q(B_k)(1 - 3r)T} | \omega \in \tilde{\Omega}, w(\omega) \in J\} \cdot P(J) \right)$$

$$+ P\{\omega \in \Omega : |x_0(\omega) - x_0(T, u(\omega), w(\omega))| \leq \frac{b}{\prod_{k=1}^{n} Q(B_k)(1 - 3r)T} | \omega \in \tilde{\Omega}, w(\omega) \in J^c\} \cdot P(J^c) \right)$$

Now, noting that $P(\tilde{\Omega}) > 1 - \alpha$ implies that $P(J^c) \leq \alpha$ we can further write

$$\leq \limsup_{T \to \infty} P\{\omega \in \Omega : |x_0(\omega) - x_0(T, u(\omega), w(\omega))| \leq \frac{b}{\prod_{k=1}^{n} Q(B_k)(1 - 3r)T} | \omega \in \tilde{\Omega}, w(\omega) \in J\} \cdot P(J) + \alpha$$

Now, observe that for a noise realization $w \in J$, we have

$$\limsup_{T \to \infty} P\{\omega \in \Omega : |x_0(\omega) - x_0(T, u(\omega), w(\omega))| \leq \frac{b}{\prod_{k=1}^{n} Q(B_k)(1 - 3r)T} | \omega \in \tilde{\Omega}, w(\omega) \in J\} \cdot P(J) + \alpha$$

$$\leq \limsup_{T \to \infty} P\{\omega \in \Omega : |x_0(\omega) - x_0(T, u(\omega), w(\omega))| \leq \frac{b}{\prod_{k=1}^{n} Q(B_k)(1 - 3r)T} | \omega \in \tilde{\Omega}, w(\omega) \in J\} \cdot P(J) + \alpha$$

$$\leq \limsup_{T \to \infty} P\{\omega \in \Omega : |x_0(\omega) - x_0(T, u(\omega), w(\omega))| \leq \frac{b}{\prod_{k=1}^{n} Q(B_k)(1 - 3r)T} | \omega \in \tilde{\Omega}, w(\omega) \in J^c\} \cdot P(J^c) \right)$$

$$+ P\{\omega \in \Omega : |x_0(\omega) - x_0(T, u(\omega), w(\omega))| \leq \frac{b}{\prod_{k=1}^{n} Q(B_k)(1 - 3r)T} | \omega \in \tilde{\Omega}, w(\omega) \in J^c\} \cdot P(J^c) \right)$$

$$\leq \limsup_{T \to \infty} P\{\omega \in \Omega : |x_0(\omega) - x_0(T, u(\omega), w(\omega))| \leq \frac{b}{\prod_{k=1}^{n} Q(B_k)(1 - 3r)T} | \omega \in \tilde{\Omega}, w(\omega) \in J\} \cdot P(J) + \alpha$$
Recall that by definition of \( \tilde{\Omega} \), we have that for any theorem for discrete memoryless channels with feedback (see [13] and Theorem 6.8), this case, we will obtain a contradiction to a generalized version of the strong converse thus by assumption we also have \( \tilde{\Omega} \) holds for any \( \omega \) on it. As such, the proof for Case 1 is complete.

Therefore, the first inequality can be justified by noting that \( |x_0(\omega) - x_0(T, u(\omega), w(\omega))| \leq \Pi_{k=1}^n \frac{b}{Q(B_k)(1-3\tau)T} |\omega \in \tilde{\Omega}, w(\omega) = w| \)

\[
\leq \limsup_{T \to \infty} P\{\omega \in \tilde{\Omega} : x_0(\omega) \in \bigcup_{u \in \tilde{\mathcal{U}}_T} \mathcal{A}_T(u, w) | \omega \in \tilde{\Omega}, w(\omega) = w\}
\leq \frac{1}{P(\omega \in \tilde{\Omega}|w(\omega) = w)} \limsup_{T \to \infty} P\{\omega \in \tilde{\Omega} : x_0(\omega) \in \bigcup_{u \in \tilde{\mathcal{U}}_T} \mathcal{A}_T(u, w) | w(\omega) = w\}
\]

\[
= \frac{1}{P(\omega \in \tilde{\Omega}|w(\omega) = w)} \limsup_{T \to \infty} \pi_0 \{ \bigcup_{u \in \tilde{\mathcal{U}}_T} \mathcal{A}_T(u, w) \} = 0,
\]

where the first inequality can be justified by noting that \( |x_0(\omega) - x_0(T, u(\omega), w(\omega))| \leq \Pi_{k=1}^n \frac{b}{Q(B_k)(1-3\tau)T} \) implies \( x_0(\omega) \in \mathcal{A}_T(u(\omega), w) \) for all \( T \) sufficiently large (see (5.14)) and the last inequality follows by independence of noise and initial state. We thus have a uniform upper bound on the limsup when conditioned on \( w \in J \), thus we have

\[
\limsup_{T \to \infty} P\left( \{\omega \in \tilde{\Omega} : |x_0(\omega) - x_0(T, u(\omega), w(\omega))| \leq \frac{b}{\Pi_{k=1}^n Q(B_k)(1-3\tau)T} | \omega \in \tilde{\Omega}, w(\omega) \in J \} \right) = 0.
\]

Therefore,

\[
\limsup_{T \to \infty} P\{\omega \in \tilde{\Omega} : |x_0(\omega) - x_0(T, u(\omega), w(\omega))| \leq \frac{b}{\Pi_{k=1}^n Q(B_k)(1-3\tau)T} | \omega \in \tilde{\Omega} \} \leq \alpha,
\]

which is a contradiction to (5.10) since \( \alpha < 1/2 \) in addition to all the other constraints on it. As such, the proof for Case 1 is complete.

**Case 2:** Now we suppose that

\[
\tilde{R} \geq (1 - 3\tau) \sum_{k=1}^n Q(B_k) \log c_k,
\]

thus by assumption we also have \( \tilde{R} > C \). Recall that the proof is by contradiction. In this case, we will obtain a contradiction to a generalized version of the strong converse theorem for discrete memoryless channels with feedback (see [13] and Theorem 6.8).

Recall that by definition of \( \tilde{\Omega} \), we have that for any \( T \) sufficiently large, the inequality

\[
|x_0(\omega) - x_0(T, u(\omega), w(\omega))| \leq \frac{b}{\Pi_{k=1}^n Q(B_k)(1-3\tau)T} | \omega \in \tilde{\Omega}\]

holds for any \( \omega \in \tilde{\Omega} \). Also recall that \( P(\tilde{\Omega}) > 1 - \alpha \) for \( \alpha \) satisfying the important assumption (5.11). As such, there must exist some noise realization \( w \) such that \( P(\{\Omega|w(\omega) = w\}) > 1 - \alpha \), which yields

\[
\liminf_{T \to \infty} P\left( \{\omega \in \tilde{\Omega} : |x_0(\omega) - x_0(T, u(\omega), w| \leq \frac{b}{\Pi_{k=1}^n Q(B_k)(1-3\tau)T} | w(\omega) = w \} \right) > 1 - \alpha.
\]

In the remainder of this section, we condition on the occurrence of the noise realization \( w \). We follow an almost identical approach as in the proof from [12]; we will construct
a sequence of codes to transmit a uniform random variable which contradicts a version of the strong converse result for DMCs. This is accomplished in four steps.

**Step 1 (Construction of bins):** For every \( T \geq 1 \), define \( S_T := \{ x_0(T, u, w) : u \in \hat{U}_T \} \) and enumerate the elements of this set so that

\[
(5.16) \quad S_T := \{ x_1(T), \ldots, x_{n_1(T)}(T) \}.
\]

Since \( |\hat{U}_T| \leq |U_T| \) for each \( T \), we have

\[
\hat{R} = \limsup_{T \to \infty} \frac{1}{T} \log n_1(T) \leq R.
\]

We continue by defining the not necessarily disjoint collection of bins

\[
B_i^T := \{ x \in \mathbb{R} : |x - x_i(T)| \leq \frac{b}{\prod_{k=1}^n Q(B_k)(1-3\alpha)^T} \}
\]

for each \( i \in \{ 1, \ldots, n_1(T) \} \). Note that for a fixed \( T \), each bin has the same Lebesgue measure which we denote by \( \rho_T := \frac{2b}{\prod_{k=1}^n Q(B_k)(1-3\alpha)^T} \). Recalling that \( P(\{ \omega \mid w(\omega) = w \}) \), it follows that

\[
1 - \alpha < \liminf_{T \to \infty} P \left( \{ \omega \in \Omega : x_0(\omega) \in \bigcup_{i=1}^{n_1(T)} B_i^T \mid w(\omega) = w \} \right),
\]

from which by independence of noise and initial state, we obtain

\[
1 - \alpha < \liminf_{T \to \infty} \pi_0 \left( \bigcup_{i=0}^{n_1(T)} B_i^T \right).
\]

We will disregard the bins that are only partially contained in \( K \). Since \( \rho_T \to 0 \) as \( T \to \infty \) and the union of the measure of bins that are partially inside of \( K \) can have at most a Lebesgue measure of \( 2\rho_T \), they will contribute negligible measure as \( T \) gets large. Also, let us suppose without loss of generality that the ordering of the bins in (5.16) is such that the last \( n(T) \) are the ones not contained in \( K \). Observing that

\[
\liminf_{T \to \infty} \pi_0 \left( \bigcup_{i=0}^{n_1(T)} B_i^T \right) = \liminf_{T \to \infty} \pi_0 (K \cap \bigcup_{i=0}^{n_1(T)} B_i^T) = \liminf_{T \to \infty} \pi_0 \left( \bigcup_{i=0}^{n_1(T) - n(T)} B_i^T \right)
\]

\[
\leq \liminf_{T \to \infty} \frac{n_1(T) - n(T)}{\rho_{\max}} \cdot m \left( \bigcup_{i=0}^{n_1(T) - n(T)} B_i^T \right) \leq \liminf_{T \to \infty} \left( \frac{(n_1(T) - n(T))}{\prod_{k=1}^n Q(B_k)(1-3\alpha)^T} \right),
\]

we obtain

\[
1 - \alpha \leq \frac{1}{2b \cdot \rho_{\max}} \liminf_{T \to \infty} \left( \frac{(n_1(T) - n(T))}{\prod_{k=1}^n Q(B_k)(1-3\alpha)^T} \right),
\]

from which we conclude that the number of bins \( n_1(T) - n(T) \) which are entirely contained in \( K \) must grow at an exponential rate of at least \( \sum_{k=1}^n c_k Q(B_k)(1-3\alpha) \) with \( T \), just as \( n_1(T) \) does. As such, since we will be concerned only with the number
of bins entirely contained in \( K \), we may as well assume without loss of generality that all are entirely in \( K \) (or alternatively, relabel \( n_1(T) - n(T) \) to be \( n_1(T) \)).

We continue by extracting a sub-collection of disjoint bins \( (C_i^T)_{i=1}^{n_2(T)} \) as described in [12, App. A]. This new sub-collection has the property that

\[
\frac{1}{2} m \left( \bigcup_{i=1}^{n_1(T)} B_i^T \right) \leq m \left( \bigcup_{i=1}^{n_2(T)} C_i^T \right).
\]

Also, it is clear that for any given \( T \), \( \frac{1}{2} n_1(T) \leq n_2(T) \). Hence, we also have the exponential growth condition of

\[
\limsup_{T \to \infty} \frac{1}{T} \log n_2(T) \geq (1 - 3\epsilon) \sum_{k=1}^{n} Q(B_k) \log c_k.
\]

Analogously to [12], define the collection \( (D_i^T)_{i=1}^{n_2(T)} \) and observe that \( m(D_i^T \setminus C_i^T) \leq \rho_T \) for each \( i \).

Finally, for a fixed \( L \in \mathbb{N} \) we join \( L \) successive \( D_i^T \) blocks (see [12, p. 27] for an exact formulation) to get a collection \( (E_i^T)_{i=1}^{n_3(T)} \), where \( n_3(T) = \lfloor \frac{2n_2(T)}{T} \rfloor + 1 \), possibly adding some empty sets in the last block. Again, the following inequalities hold:

\[
\limsup_{T \to \infty} \frac{1}{T} \log n_3(T) \geq (1 - 3\epsilon) \sum_{k=1}^{n} Q(B_k) \log c_k, \quad m(E_i^T) \geq L \rho_T.
\]

We also define

\[
M_T := \bigcup_{i=1}^{n_1(T)} B_i^T, \quad M_T := \bigcup_{i=1}^{n_3(T)} E_i^T \setminus (D_i^T \setminus C_i^T)
\]

and observe that \( m(M_T) \leq 2n_2(T)\rho_T \leq 2n_3(T)L\rho_T \).

**Step 2 (Auxiliary Coding Scheme):** We now construct a sequence of codes to transmit information over the channel. For a fixed \( L \) and for each \( T \), we will construct a code. Note that we are considering a channel with feedback, which can be used by the encoding function. For a given \( T \), the encoding and decoding processes are specified as follows.

**Encoder:** We give to the decoder the noise realization \( w \) that we have conditioned on throughout, the function \( f \) corresponding to the system dynamics, and the fixed causal coding and control policy. In the classical notion of a code, the encoding function is a deterministic map \( g : \{1, \ldots, n_3(T)\} \to M^T \). In our case, for \( i \in \{1, \ldots, n_3(T)\} \), \( g(i) \) will be a random variable, determined as follows. The first symbol of the sequence \( g(i) \) is just the random variable \( q_0 = \gamma_0(x_0) \). Now, because the channel has feedback, the encoder can determine \( u_0 \) by applying the decoding function of the fixed causal coding and control policy. As such, using the fixed and known noise realization \( w \), \( x_1 \) can be computed. Then, \( q_1 \) is computed again using the causal coding and control policy, and so on until \( q_{T-1} \) is determined (note that the encoder makes use of the channel feedback from the channel, and thus we use the generalized version of the strong converse theorem for channel capacity to obtain a contradiction).

**Decoder:** At time \( T \), the decoder will receive \( T \) symbols from the channel, which are used to compute the control decisions \( u_0, \ldots, u_{T-1} \) according to the fixed causal coding and control policy. The decoder also has knowledge of the noise sequence \( w \).
and uses it to compute the point \( x_0(T, u, w) \). Our goal is to use the received channel output and control sequence to reconstruct the index \( Y \) of the bin \( E_Y^T \) containing \( x_0 \). We do this by looking at the point \( x_0(T, u, w) \) for the observed control sequence \( u \). Note that \( w \) can be thought of as deterministic since we are conditioning on its occurrence. Recall also that \( x_0(T, u, w) \) is the “midpoint” of the set \( A_Y^T(u, w) \), and can be computed without knowledge of the initial state \( x_0 \). We simply decide on our guess \( \tilde{Y} \) of the index as follows.

- If \( x_0(T, u, w) \in M_T^T \), take the index \( i \) of the set \( E_Y^T \) that \( x_0(T, u, w) \) belongs to.
- If \( x_0(T, u, w) \notin M_T^T \), then decide randomly between \( i \) and \( i+1 \), where \( i \) is the index of the set \( E_Y^T \) that \( x_0(T, u, w) \) belongs to.

### Analysis of probability of error for the code.

To study the probability of error, let \( Y \) be a random variable on the indices \( \{1, \ldots, n_3(T)\} \), where \( P(Y = i) = \pi_0(E_Y^T) \). We analyze \( P(\tilde{Y} \neq Y) \).

First, by construction of the bins and estimation scheme, we have

\[
P(\tilde{Y} \neq Y | x_0 \in M_T^T, |x_0 - x_0(T, u, w)| \leq b \prod_{k=1}^{n} c_k^{Q(B_j^T)} (1-3r)^T) = 0.
\]

and

\[
P(\tilde{Y} \neq Y | x_0 \in M_T \setminus M_T^T, |x_0 - x_0(T, u, w)| \leq b \prod_{k=1}^{n} c_k^{Q(B_j^T)} (1-3r)^T) \leq \frac{1}{2}.
\]

As such, from (5.15), it is not hard to see that for every \( T \) sufficiently large,

\[
P(Y \neq \tilde{Y}) \leq \frac{1}{2} \pi_0(M_T \setminus M_T^T) + \alpha.
\]

By an analysis exactly as in [12], we have

\[
\pi_0(M_T \setminus M_T^T) \leq \frac{1}{L} \rho_{\max} \pi_0(M_T).
\]

Combining the above two equations, we obtain

\[
\sum_{i=1}^{n_3(T)} P(Y = i) P(\tilde{Y} \neq Y | Y = i) \leq \frac{1}{2L} \frac{\rho_{\max}}{\rho_{\min}} \pi_0(M_T) + \alpha.
\]

**Step 3 (Introduction of an auxiliary uniform random variable):** In order to obtain a contradiction with the strong converse theorem for DMCs, we will need to transmit a random variable uniformly distributed on the indices \( 1, \ldots, n_3(T) \). Let us call this random variable \( W := W_T \). Of course, at any time step, \( W \) must be conditionally independent from the channel output, given the channel input. To obtain the desired contradiction, we must show that \( \lim_{T \to \infty} P(W \neq \tilde{Y}) < 1 \). Before considering this quantity, note that by following exactly the same steps as in [12], we obtain

\[
\pi_0(M_T) \leq \rho_{\max} \cdot m(M_T) \leq 2n_3(T) \rho_{\max} \cdot L \rho_T
\]
and also
\[ \sum_{i=1}^{n_3(T)} \frac{1}{n_3(T)} P(Y = i) P(Y = i) \leq \alpha + \frac{\rho_{\max} \pi_0(M_T)}{4L \rho_{\min}}. \]

Again as in [12] we have
\[(5.17) \quad P(W \neq \tilde{Y}) = \sum_{i=1}^{n_3(T)} P(W = i) P(Y = i) \leq P(Y \neq W) + \frac{\alpha + \rho_{\max} \pi_0(M_T)}{4L \rho_{\min}}. \]

**Step 4 (Application of optimal transport):** Recall the independence condition mentioned above that \( W \) must satisfy. To achieve this, one could adjoin \( W \) to the common probability space using the product measure, thus keeping \( W \) independent from all other random variables. Observe however, that the random variable \( x_0 \) satisfies the independence condition that we require \( W \) to satisfy. As such, we are free to choose any possible coupling between \( W \) and \( x_0 \) while ensuring that \( W \) will remain independent from the channel output given the channel input (in particular, \( x_0 \) and \( W \) need not be independent). We will take advantage of this observation in this section.

Consider (5.17) and note that if the limit as \( T \rightarrow \infty \) of the right-hand side is strictly less than 1, then we will have the desired contradiction with the strong converse. As such, we proceed by finding a coupling between \( W \) and \( x_0 \) which makes \( P(Y \neq W) \) small enough so that the limit is less than 1.

We continue by letting \( \mu \) denote the law of \( Y \). That is, for some index \( i \in 1, \ldots, n_3(T) \), \( \mu(i) = \pi_0(E_i^T) \). Let also \( \nu \) represent the law of \( W \), i.e., a uniform measure on the set \( \{1, \ldots, n_3(T)\} \). We now invoke Lemma 6.3, which guarantees the existence of a coupling \( (Y, W) : (\Omega, \mathcal{F}, P) \rightarrow \{1, \ldots, n_3(T)\}^2 \) such that
\[ P(Y \neq W) = \frac{1}{2} \sum_{i=1}^{n_3(T)} |\mu(i) - \nu(i)|. \]

Let now \( A = \{i \in 1, \ldots, n_3(T) : \mu(i) \geq \nu(i)\} \) and observe that
\[ 1 - \sum_{i=1}^{n_3(T)} \min(\mu(i), \nu(i)) = \frac{1}{2} \sum_{i=1}^{n_3(T)} \mu(i) + \frac{1}{2} \sum_{i=1}^{n_3(T)} \nu(i) - \sum_{i \in A} \nu(i) - \sum_{i \in A^c} \mu(i) \]
\[ = \frac{1}{2} \sum_{i \in A} \mu(i) - \sum_{i \in A} \nu(i) - \frac{1}{2} \sum_{i \in A^c} \mu(i) - \sum_{i \in A^c} \nu(i) \]
\[ = \frac{1}{2} \left( \sum_{i \in A} \mu(i) - \nu(i) \right) + \frac{1}{2} \left( \sum_{i \in A^c} \nu(i) - \mu(i) \right) = \frac{1}{2} \sum_{i=1}^{n_3(T)} |\mu(i) - \nu(i)|. \]

thus we can write
\[ P(Y \neq W) = \frac{1}{2} \sum_{i=1}^{n_3(T)} |\mu(i) - \nu(i)| = 1 - \sum_{i=1}^{n_3(T)} \min(\mu(i), \nu(i)). \]

To get an upper bound for the right-hand side of the above, note that
\[ \mu(i) = \pi_0(E_i^T) \geq \rho_{\min} \cdot m(E_i^T) = \frac{n_3(T)}{n_3(T)} \cdot m(E_i^T) \cdot \rho_{\min}. \]
\[ \geq \frac{n_2(T) \cdot \rho_T \cdot \rho_{\min}}{n_3(T)} \geq \frac{m(M_T) \cdot \rho_{\min}}{2 \cdot \rho_{\max} \cdot n_3(T)} \geq \frac{\rho_{\min} \cdot (1 - \alpha)}{2 \cdot \rho_{\max} \cdot n_3(T)}. \]

Recalling that \( \nu(i) = \frac{1}{n_3(T)} \) for each \( i \), we have \( \min(\mu(i), \nu(i)) \geq \frac{\rho_{\min} \cdot (1 - \alpha)}{2 \cdot \rho_{\max} \cdot n_3(T)} \) for all \( i \) and therefore

\[ P(Y \neq W) \leq 1 - \frac{\rho_{\min} \cdot (1 - \alpha)}{2 \cdot \rho_{\max}}. \]

Combining with the (5.17), we obtain

\[ P(W \neq \tilde{Y}) \leq 1 - \frac{\rho_{\min} \cdot (1 - \alpha)}{2 \cdot \rho_{\max}} + \frac{\alpha + \frac{\rho_{\max} \pi_0(M_T)}{4L\rho_{\min}}}{\frac{\rho_{\min} \pi_0(M_T)}{2\rho_{\max}}} \]

which holds for all \( T \) sufficiently large. We now evaluate the right-hand side to determine its behavior as \( T \) tends to infinity. We have

\[
\limsup_{T \to \infty} \left( 1 - \frac{\rho_{\min} \cdot (1 - \alpha)}{2 \cdot \rho_{\max}} + \frac{\alpha + \frac{\rho_{\max} \pi_0(M_T)}{4L\rho_{\min}}}{\frac{\rho_{\min} \pi_0(M_T)}{2\rho_{\max}}} \right)
\leq 1 - \frac{\rho_{\min} \cdot (1 - \alpha)}{2 \cdot \rho_{\max}} + \frac{\rho_{\max}^2}{2L\rho_{\min}^2} + \frac{2 \cdot \rho_{\min} \cdot \rho_{\max}}{\rho_{\max} \cdot 1 - \alpha}. \]

Recall now that throughout, \( L \in \mathbb{N} \) was fixed but arbitrary. Taking \( L \) large enough so that (5.11) holds, and writing \( T \) subscripts to emphasize \( T \)-dependence, we obtain

\[ \limsup_{T \to \infty} P(W_T \neq \tilde{Y}_T) < 1, \]

which is a contradiction, since it negates the strong converse theorem for DMCs with feedback. Thus in Case 2, we have also obtained a contradiction, thus combining with Case 1, the proof is complete.

6. Appendix. In this section, we state a few results required in the paper.

6.1. A result from optimal transport. In the proof of Theorem 4.2, a basic result from optimal transport is used, which we state here.

Definition 6.1. Let \( \mu \) and \( \nu \) be Borel probability measures on a metric spaces \((S,d)\). A coupling of \( \mu \) and \( \nu \) is a pair of random variables \( X, Y \) defined on some probability space \((\Omega, \mathcal{F}, P)\) such that the law of the random variable \((X,Y)\) on \( S^2 \) admits \( \mu \) and \( \nu \) as its marginals.

The notion of coupling can easily be generalized for the case where the measures \( \mu \) and \( \nu \) are on distinct spaces, however we do not require that level of generality. The total variation distance between probability measures on the same measurable space serves as a measure for how distinct they are. The definition reads as follows.

Definition 6.2. Let \( \mu \) and \( \nu \) be probability measures on a measurable space \((\Omega, \mathcal{F})\). We define the total variation distance as

\[ \|\mu - \nu\|_{TV} := 2 \sup_{A \in \mathcal{F}} |\mu(A) - \nu(A)|. \]
Lemma 6.3. Let \((X, Y) : (\Omega, \mathcal{F}, P) \rightarrow S^2\) be a coupling of the probability measures \(\mu\) and \(\nu\) on the metric space \((S, d)\). Then
\[
\|\mu - \nu\|_{TV} \leq P(\{\omega \in \Omega : X(\omega) \neq Y(\omega)\}).
\]
If in addition, \(S\) is a finite set, then a coupling \((X, Y)\) exists which achieves the above bound. Note also that if \(S\) is finite, then we have the identity
\[
\|\mu - \nu\|_{TV} = \frac{1}{2} \sum_{x \in S} |\nu(x) - \mu(x)|.
\]

6.2. Channel coding theorem. When considering a system controlled over a noisy channel, we make use of the strong converse of the noisy channel coding theorem. We state the necessary definitions and theorems here without proof.

Definition 6.4. Consider a memoryless finite alphabet channel with input alphabet \(X\), output alphabet \(Y\) and a given transition probability measure. The capacity of the channel is defined by \(C := \sup_{p(x)} I(X; Y)\) where the sup is taken over all possible probability measures on the input alphabet \(X\). We call such a channel a Discrete Memoryless Channel (DMC). A DMC with feedback is as above, but with the additional property that the encoder has knowledge of the channel output. It is well-known that feedback does not increase channel capacity.

Next, we provide the definition of a code.

Definition 6.5. An \((M, n)\)-code consists of an encoding function \(x^n : \{1, \ldots, M\} \rightarrow X^n\) and a decoding function \(g : Y^n \rightarrow \{1, \ldots, M\}\). We define the rate of an \((M, n)\)-code by \(R := \frac{\log(M)}{n}\).

For a code as above, we call \(x^n(1), x^n(2), \ldots, x^n(M)\) the codewords. Because the channel distorts the codewords, we must consider the probability that we can decode correctly. This leads to the following definition.

Definition 6.6. The maximal error of a \((M, n)\)-code is given by
\[
\lambda(n) := \max_{i \in \{1, \ldots, M\}} P(g(Y^n) \neq i | X^n = x^n(i)).
\]

Definition 6.7. A rate \(R\) is called achievable if there exists a sequence of \((\lceil 2^nR \rceil, n)\) codes with the property that \(\lambda(n) \rightarrow 0\) as \(n \rightarrow \infty\).

The following is the strong converse of the noisy channel coding theorem in information theory.

Theorem 6.8. Consider a DMC \((X, p(\cdot|\cdot), Y)\) of capacity \(C\). Let \(R > C\) and consider an arbitrary sequence of \((\lceil 2^{nR} \rceil, n)\)-codes, used to transmit the uniform random variables \(W_n\), uniformly distributed on the set \(\{1, \ldots, 2^{nR}\}\), respectively. Then
\[
P(W_n \neq g_n(Y^n)) \rightarrow 1 \text{ as } n \rightarrow \infty.
\]

The above theorem also holds for DMCs with feedback (see [13] for a proof). In the proof of Theorem 4.2, the encoding functions require that the channel have feedback, hence the need for this assumption in the theorem statement.

REFERENCES

[1] J. Baillieu, Feedback designs for controlling device arrays with communication channel bandwidth constraints, in ARO workshop on smart structures, University Park PA, 1999, pp. 16–18.
[2] F. Colonius, Metric invariance entropy and conditionally invariant measures, Ergodic Theory and Dynamical Systems, 38 (2018), pp. 921–939.
[3] F. Colonius and C. Kawan, Invariance entropy for control systems, SIAM Journal on Control and Optimization, 48 (2009), pp. 1701–1721.
[4] F. Colonius, C. Kawan, and G. Nair, A note on topological feedback entropy and invariance entropy, Systems & Control Letters, 62 (2013), pp. 377–381.
[5] C. De Persis and A. Isidori, Stabilizability by state feedback implies stabilizability by encoded state feedback, Systems & Control Letters, 53 (2004), pp. 249–258.
[6] E. B. Dynkin and A. A. Yushkevich, Controlled Markov processes, vol. 235, Springer, 1979.
[7] M. Franceschetti and P. Minero, Elements of information theory for networked control systems, in Information and Control in Networks, Springer, 2014, pp. 3–37.
[8] R. M. Gray and R. Gray, Probability, random processes, and ergodic properties, Springer, 1988.
[9] J. 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, Citeseer, 2002, p. 1.
[10] A. Katok, Lyapunov exponents, entropy and periodic orbits for diffeomorphisms, Publications Mathématiques de l’IHES, 51 (1980), pp. 137–173.
[11] C. Kawan, Invariance entropy for deterministic control systems, Lecture notes in mathematics, 2089 (2013).
[12] C. Kawan and S. Yüksel, Invariance properties of nonlinear stochastic dynamical systems under information constraints, 2019, https://arxiv.org/abs/1901.02825.
[13] J. Kemperman, Strong converses for a general memoryless channel with feedback, in Proceedings 6th Prague Conference on Information Theory, Statistical Decision Functions, and Random Processes, 1971, pp. 375–409.
[14] K. Li and J. Bailleul, Data-rate requirements for nonlinear feedback control, IFAC Proceedings Volumes, 37 (2004), pp. 997–1002.
[15] D. Liberzon and J. P. Hespanha, Stabilization of nonlinear systems with limited information feedback, IEEE Transactions on Automatic Control, 50 (2005), pp. 910–915.
[16] A. S. Matveev and A. V. Savkin, Estimation and control over communication networks, Springer Science & Business Media, 2009.
[17] S. P. Meyn and R. L. Tweedie, Markov chains and stochastic stability, Springer Science & Business Media, 2012.
[18] G. N. Nair and R. J. Evans, Stabilizability of stochastic linear systems with finite feedback data rates, SIAM Journal on Control and Optimization, 43 (2004), pp. 413–436.
[19] G. N. Nair, R. J. Evans, I. M. Mareels, and W. Moran, Topological feedback entropy and nonlinear stabilization, IEEE Transactions on Automatic Control, 49 (2004), pp. 1585–1597.
[20] Y. Ren, L. He, J. Lu, and G. Zheng, Topological r-entropy and measure-theoretic r-entropy of a continuous map, Science China Mathematics, 54 (2011), pp. 1197–1205.
[21] S. Tatikonda and S. Mitter, Control under communication constraints, IEEE Transactions on Automatic control, 49 (2004), pp. 1056–1068.
[22] W. S. Wong and R. W. Brockett, Systems with finite communication bandwidth constraints. ii. stabilization with limited information feedback, IEEE Transactions on Automatic Control, 44 (1999), pp. 1049–1053.
[23] S. Yüksel, Stochastic stabilization of noisy linear systems with fixed-rate limited feedback, IEEE Transactions on Automatic Control, 55 (2010), pp. 2847–2853.
[24] S. Yüksel, Characterization of information channels for asymptotic mean stationarity and stochastic stability of non-stationary/unstable linear systems, IEEE Transactions on Information Theory, 58 (2012), pp. 6332–6354.
[25] 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, 58 (2013), pp. 47 – 59.
[26] S. Yüksel, Stationary and ergodic properties of stochastic nonlinear systems controlled over communication channels, SIAM Journal on Control and Optimization, 54 (2016), pp. 2844–2871.
[27] S. Yüksel and T. Başar, Stochastic networked control systems: Stabilization and optimization under information constraints, Springer Science & Business Media, 2013.