


Metric and Topological Entropy Bounds for Optimal Coding of Stochastic Dynamical Systems

Christoph Kawan  and Serdar Yüksel , *Member, IEEE*

Abstract—We consider the problem of optimal zero-delay coding and estimation of a stochastic dynamical system over a noisy communication channel under three estimation criteria concerned with the low-distortion regime. The criteria considered are (i) a strong and (ii) a weak form of almost sure stability of the estimation error as well as (ii) asymptotic quadratic stability in expectation. For all three objectives, we derive lower bounds on the smallest channel capacity C_0 above which the objective can be achieved with an arbitrarily small error. We first obtain bounds through a dynamical systems approach by constructing an infinite-dimensional dynamical system and relating the capacity with the topological and the metric entropy of this system. We also consider information-theoretic and probability-theoretic approaches to address the different criteria. Finally, we prove that a memoryless noisy channel in general constitutes no obstruction to asymptotic almost sure state estimation with arbitrarily small errors, when there is no noise in the system. The results provide new solution methods for the criteria introduced (e.g., standard information-theoretic bounds cannot be applied for some of the criteria) and establish further connections between dynamical systems, networked control, and information theory, especially in the context of nonlinear stochastic systems.

Index Terms—Dynamical systems, information theory, metric entropy, nonlinear systems, state estimation, topological entropy.

I. INTRODUCTION

IN THIS article, we study the problem of optimal coding and estimation of a stochastic dynamical system over a noisy communication channel. This is a fundamental and classical problem in stochastic and networked control, information and communication theory, and estimation theory. Accordingly, this problem has an extensive history and literature which we present

Manuscript received November 16, 2018; revised May 2, 2019; accepted July 8, 2019. Date of publication August 26, 2019; date of current version May 28, 2020. This research was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada. Recommended by Associate Editor W. X. Zheng. Some results of this article appeared in part at the 2017 IEEE International Symposium on Information Theory. (*Corresponding author: Serdar Yüksel.*)

C. Kawan is with the Institute for Informatics at the Ludwig-Maximilians-Universität Munich, 80538 Munich, Germany (e-mail: kawanchr123@gmail.com).

S. Yüksel is with the Department of Mathematics and Statistics, Queen's University, Kingston, ON K7L 3N6, Canada (e-mail: yuksel@mast.queensu.ca).

Digital Object Identifier 10.1109/TAC.2019.2937732

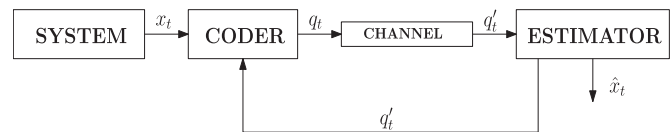


Fig. 1. Coding and state estimation over a noisy channel with feedback.

further below once we have added more specificity to the setup considered in the article and have stated the problem.

In this article, we consider nonlinear stochastic systems given by an equation of the form

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

Here x_t is the state at time t and $(w_t)_{t \in \mathbb{Z}_+}$ is an independent and identically distributed (i.i.d.) sequence of random variables with common distribution $w_t \sim \nu$, modeling the noise. In general, we assume that

$$f : X \times W \rightarrow X$$

is a Borel measurable map, where (X, d) is a complete metric space and W a measurable space, so that for any $w \in W$, the map $f(\cdot, w)$ is a homeomorphism of X . We further assume that x_0 is a random variable on X with an associated probability measure π_0 , stochastically independent of $(w_t)_{t \in \mathbb{Z}_+}$. We use the notations

$$f_w(x) = f(x, w), \quad f^x(w) = f(x, w)$$

so that $f_w : X \rightarrow X$ and $f^x : W \rightarrow X$.

System (1) is connected over a possibly noisy channel with finite capacity to an estimator, as shown in Fig. 1. The estimator has access to the information it has received through the channel. A source coder maps the source symbols (i.e., state values) to corresponding channel inputs. These inputs are transmitted through the channel, which we assume to be discrete with input alphabet \mathcal{M} and output alphabet \mathcal{M}' .

We refer by a *coding policy* Π , to a sequence of functions $(\gamma_t^e)_{t \in \mathbb{Z}_+}$ which are causal such that the channel input at time t , $q_t \in \mathcal{M}$, under Π is generated by a function of its local information, i.e.,

$$q_t = \gamma_t^e(\mathcal{I}_t^e)$$

where $\mathcal{I}_t^e = \{x_{[0,t]}, q'_{[0,t-1]}\}$ and $q_t \in \mathcal{M}$, the channel input alphabet given by $\mathcal{M} = \{1, 2, \dots, M\}$, for $0 \leq t \leq T-1$. Here, we use the notation $x_{[0,t-1]} = \{x_s : 0 \leq s \leq t-1\}$ for $t \geq 1$.

The channel maps q_t to q'_t in a stochastic fashion so that $P(q'_t|q_t, q_{[0,t-1]}, q'_{[0,t-1]})$ is a conditional probability measure on \mathcal{M}' for all $t \in \mathbb{Z}_+$. If this expression is equal to $P(q'_t|q_t)$, the channel is said to be memoryless, i.e., the past variables do not affect the channel output q'_t given the current channel input q_t .

The receiver, upon receiving the information from the channel, generates an estimate \hat{x}_t at time t , also causally: An admissible causal estimation policy is a sequence of functions $(\gamma_t^d)_{t \in \mathbb{Z}_+}$ such that $\hat{x}_t = \gamma_t^d(q'_{[0,t]})$ with

$$\gamma_t^d : (\mathcal{M}')^{t+1} \rightarrow X, \quad t \geq 0.$$

We often classify such coding and estimation policies as *zero-delay* policies, since the encoder and the estimator operate instantaneously and do not wait for future data to arrive prior to selecting their outputs. Such policies are crucial in delay-sensitive applications, such as networked control systems. This is unlike much of the information theory literature in the context of Shannon theory. We provide a detailed literature review below on zero-delay coding.

For a given $\varepsilon > 0$, we denote by C_ε the smallest channel capacity above which there exist an encoder and an estimator so that one of the following estimation objectives is achieved:

- E1) Eventual almost sure stability of the estimation error:
There exists $T(\varepsilon) \geq 0$ so that

$$\sup_{t \geq T(\varepsilon)} d(x_t, \hat{x}_t) \leq \varepsilon \quad \text{a.s.}$$

- E2) Asymptotic almost sure stability of the estimation error:

$$P(\limsup_{t \rightarrow \infty} d(x_t, \hat{x}_t) \leq \varepsilon) = 1.$$

- E3) Asymptotic quadratic stability of the estimation error in expectation:

$$\limsup_{t \rightarrow \infty} E[d(x_t, \hat{x}_t)^2] \leq \varepsilon.$$

We are interested in characterizations of C_ε and, in particular $C_0 := \lim_{\varepsilon \downarrow 0} C_\varepsilon$.

A. Literature Review and Contributions

In a recent work [20], we investigated the same problem for the special case involving only deterministic systems and discrete noiseless channels. In this article, we will provide further connections between the ergodic theory of dynamical systems and information theory by answering the problems posed in the previous section and relating the answers to the concepts of either metric or topological entropy. Our findings complement and generalize our results in [20] since here we consider stochasticity in the system dynamics and/or the communication channels.

As we note in [20], optimal coding of stochastic processes is a problem that has been studied extensively: in information theory in the context of per-symbol cost minimization, in dynamical systems in the context of identifying representational and equivalence properties between dynamical systems, and in networked control in the context of identifying information transmission

requirements for stochastic stability or cost minimization. As such, for the criteria laid out in (E1)–(E3) above, the results in our article are related to the efforts in the literature in the following three general areas.

Dynamical Systems and Ergodic Theory: Historically, there has been a symbiotic relation between the ergodic theory of dynamical systems and information theory (see, e.g., [7] and [43] for comprehensive reviews). Information-theoretic tools have been foundational in the study of dynamical systems, for example, the metric (also known as Kolmogorov–Sinai or measure-theoretic) entropy is crucial in the celebrated Shannon–McMillan–Breiman theorem as well as two important representation theorems: Ornstein’s (isomorphism) theorem and the Krieger’s generator theorem [7] and [12], [17], [36], [37]. The concept of sliding block encoding [11] is a stationary encoding of a dynamical system defined by the shift process, leading to fundamental results on the existence of stationary codes which perform as good as the limit performance of a sequence of optimal block codes. For topological dynamical systems, the theory of entropy structures and symbolic extensions answers the question to which extent a system can be represented by a symbolic system (under preservation of some topological structure), cf. [7] for an overview of this theory. Entropy concepts have extensive operational practical usage in identifying limits on source and channel coding for a large class of sources [11], [13], [43]. We also refer the reader to [10] for a more general overview of the concept of entropy and its applications in various branches of applied mathematics, mathematical physics, and engineering.

Networked Control and Stochastic Stability Under Information Constraints: In networked control, there has been a recurrent interest in identifying limitations on state estimation and control under information constraints. The results in this area have typically involved linear systems, and in the nonlinear case, the studies have only been on deterministic systems estimated/controlled over deterministic channels, with few exceptions. For linear systems, data-rate theorem type results have been presented in [28], [31], [33], [44], and [49].

The papers [23], [24], [29], [30], [39] studied state estimation for nonlinear deterministic systems and noise-free channels. In [23] and [24], Liberzon and Mitra characterized the critical data rate C_0 for exponential state estimation with a given exponent $\alpha \geq 0$ for a continuous-time system on a compact subset K of its state-space. As a measure for C_0 , they introduced a quantity called estimation entropy $h_{\text{est}}(\alpha, K)$, which equals the topological entropy on K in case $\alpha = 0$, but for $\alpha > 0$ is no longer a purely topological quantity. The paper [19] provided a lower bound on $h_{\text{est}}(\alpha, K)$ in terms of Lyapunov exponents under the assumption that the system preserves a smooth measure. In [29], [30], Matveev and Pogromsky studied three estimation objectives of increasing strength for discrete-time nonlinear systems. For the weakest one, the smallest bit rate was shown to be equal to the topological entropy. For the other ones, general upper and lower bounds were obtained which can be computed directly in terms of the linearized right-hand side of the equation generating the system.

A further closely related paper is due to Savkin [41], which uses topological entropy to study state estimation for a class of nonlinear systems over noise-free digital channels. In fact, our results can be seen as stochastic analogs of some of the results presented in [41], which show that for sufficiently perturbed deterministic systems, state estimation with arbitrarily small error is not possible over finite-capacity channels.

A related problem is the control of nonlinear systems over communication channels. This problem has been studied in few publications, and mainly for deterministic systems and/or deterministic channels. Recently, [53] studied stochastic stability properties for a more general class of stochastic nonlinear systems building on information-theoretic bounds and Markov-chain-theoretic constructions. However, these bounds do not distinguish between the unstable and stable components of the tangent space associated with a dynamical nonlinear system, while the entropy bounds established in this article make such a distinction, but only for estimation problems and in the low-distortion regime.

Zero-Delay Coding Over Communication Channels: In our setup, we have causality as a restriction in coding and decoding. Zero-delay coding is an increasingly important research area of significant practical relevance, as we review in [20]. Notable papers include the classical works by Witsenhausen [48], Walrand and Varaiya [47], and Teneketzis [46]. The findings of [47] have been generalized to continuous sources in [52] (see also [25] and [3], where the latter imposes a structure a priori); and the structural results on optimal fixed-rate coding in [48] and [47] have been shown to be applicable to setups when one also allows for variable-length source coding in [15]. Structural results on coding over noisy channels have been studied in [27], [46], [47], [50] among others. Related work also includes [1], [14], [27], [50] which have primarily considered the coding of discrete sources. A few works [1], [3], [14], [25], [50] have considered infinite horizon problems and, in particular, [50] has established the optimality of stationary and deterministic policies for finite aperiodic and irreducible Markov sources. A related lossy coding procedure was introduced by Neuhoff and Gilbert [35], called *causal source coding*, which has a different operational definition, since delays in coding and decoding are allowed so that efficiencies through entropy coding can be utilized. Further discussions on the literature are available in [1], [15], [25], and [34]. Among those that are most relevant to our article is [26], where causal coding under a high rate assumption for stationary sources and individual sequences was studied, though only in a source coding context.

The setup with Gaussian channels is a special case studied extensively for the coding of linear systems. We will not consider such a setup in this article, though we note that explicit results have been obtained for a variety of criteria in the literature.

Contributions: In view of this literature review, we make the following contributions. We establish that for (E1), the topological entropy of a properly defined infinite-dimensional dynamical system defining the stochastic evolution of the process provides lower bounds; for (E2), a lower bound is provided by the metric entropy; and for (E3), the metric entropy of this system also provides a lower bound under a restriction on the class of encoders

considered. Through a novel analysis, we also provide achievability results for the case where the only stochasticity is in the communication channel and the initial state. We also establish impossibility results when the system is sufficiently mixing due to the noise. We show that our results reduce to those reported in [20] for deterministic systems. An implication is that the rate bounds may not depend continuously on the noise, i.e., an arbitrarily small noisy perturbation of the system dynamics may lead to a discontinuous change in the rate requirements for each of the criteria. Throughout the analysis, we provide further connections between information theory and dynamical systems by identifying the operational usage of entropy concepts for the three different estimation criteria.

II. PRELIMINARIES

Notation: All logarithms in this article are taken to the base 2. We write $|E|$ for the cardinality of a set E . By \mathbb{N} , we denote the set of positive integers. We write \mathbb{Z} for the set of all integers and $\mathbb{Z}_+ = \mathbb{N} \cup \{0\}$. By $\mathbf{1}_A$, we denote the characteristic function of a set A . We write $B_\varepsilon(x)$ for the open ball of radius $\varepsilon > 0$ centered at $x \in \mathbb{R}^N$. If $f : X \rightarrow Y$ is a measurable map between measurable spaces (X, \mathcal{F}) and (Y, \mathcal{G}) , we write f_* for the push-forward operator associated with f on the space of measures on (X, \mathcal{F}) , i.e., for any measure μ on (X, \mathcal{F}) , $f_*\mu$ is the measure on (Y, \mathcal{G}) defined by $(f_*\mu)(G) := \mu(f^{-1}(G))$ for all $G \in \mathcal{G}$.

A. Entropy Notions for Dynamical Systems

In the following, we explain the notions of topological and metric entropy for dynamical systems. Metric entropy was first introduced by Kolmogorov and Sinai in the 1950s as a measure-theoretic invariant of dynamical systems that preserve a probability measure on their state-space. In information-theoretic terms, metric entropy measures the expected average (over time) growth of uncertainty about the initial state of the system as time tends to infinity. In the 1960s, an analogous notion of topological entropy was introduced by Adler, Konheim, and McAndrew for dynamical systems in the topological category. While it has no direct information-theoretic interpretation, topological entropy turns out to be the supremal metric entropy, when all probability measures preserved by the system are considered at once. For further details about these concepts and their operational meaning, we refer the reader to the excellent survey [17] and the monograph [7].

Let $f : X \rightarrow X$ be a continuous map on a metric space (X, d) and write f^i for its iterates, i.e., $f^0 = \text{id}_X$ and $f^{i+1} = f \circ f^i$ for $i \geq 0$. For a compact set $K \subset X$, we say that $E \subset K$ is (n, ε, f) -separated for some $n \in \mathbb{N}$ and $\varepsilon > 0$ if for all $x, y \in E$ with $x \neq y$, $d(f^i(x), f^i(y)) > \varepsilon$ for some $i \in \{0, 1, \dots, n-1\}$. We write $r_{\text{sep}}(n, \varepsilon, K; f)$ for the maximal cardinality of an (n, ε, f) -separated subset of K and define the *topological entropy* $h_{\text{top}}(f, K)$ of f on K by

$$h_{\text{sep}}(f, \varepsilon, K) := \limsup_{n \rightarrow \infty} \frac{1}{n} \log r_{\text{sep}}(n, \varepsilon, K; f)$$

$$h_{\text{top}}(f, K) := \lim_{\varepsilon \downarrow 0} h_{\text{sep}}(f, \varepsilon, K).$$

If X is compact and $K = X$, we omit the argument K and call $h_{\text{top}}(f)$ the *topological entropy of f* . Alternatively, one can define $h_{\text{top}}(f, K)$ using (n, ε) -spanning sets. A set $F \subset X$ (n, ε) -spans another set $K \subset X$ if for each $x \in K$ there is $y \in F$ with $d(f^i(x), f^i(y)) \leq \varepsilon$ for $i = 0, 1, \dots, n-1$. Letting $r_{\text{span}}(n, \varepsilon, K; f)$ (or $r_{\text{span}}(n, \varepsilon, K)$ if the map f is clear from the context) denote the minimal cardinality of a set which (n, ε) -spans K , the topological entropy of f on K satisfies

$$h_{\text{top}}(f, K) = \lim_{\varepsilon \downarrow 0} \limsup_{n \rightarrow \infty} \frac{1}{n} \log r_{\text{span}}(n, \varepsilon, K; f).$$

If $f : X \rightarrow X$ is a measure-preserving map on a probability space $(\Omega, \mathcal{F}, \mu)$, i.e., $f_*\mu = \mu$, its *metric entropy* $h_\mu(f)$ is defined as follows. Let \mathcal{A} be a finite measurable partition of X . Then, the entropy of f with respect to \mathcal{A} is defined by

$$h_\mu(f; \mathcal{A}) := \lim_{n \rightarrow \infty} \frac{1}{n} H_\mu \left(\bigvee_{i=0}^{n-1} f^{-i} \mathcal{A} \right). \quad (2)$$

Here \bigvee denotes the join operation, i.e., $\bigvee_{i=0}^{n-1} f^{-i} \mathcal{A}$ is the partition of X consisting of all intersections of the form $A_0 \cap f^{-1}(A_1) \cap \dots \cap f^{-n+1}(A_{n-1})$ with $A_i \in \mathcal{A}$. For any partition \mathcal{B} of X , $H_\mu(\mathcal{B}) = -\sum_{B \in \mathcal{B}} \mu(B) \log \mu(B)$ is the Shannon entropy of \mathcal{B} . The existence of the limit in (2) follows from a subadditivity argument. The metric entropy of f is then defined by

$$h_\mu(f) := \sup_{\mathcal{A}} h_\mu(f; \mathcal{A})$$

the supremum taken over all finite measurable partitions \mathcal{A} of X . If f is continuous, X is compact metric, and μ is ergodic, there is an alternative characterization of $h_\mu(f)$ due to Katok [16]:

For any $n \in \mathbb{N}$, $\varepsilon > 0$ and $\delta \in (0, 1)$ put

$$r_{\text{span}}(n, \varepsilon, \delta) := \min \{ r_{\text{span}}(n, \varepsilon; A) : A \subset X \text{ Borel, } \mu(A) \geq 1 - \delta \}.$$

Then, for every $\delta \in (0, 1)$, it holds that

$$h_\mu(f) = \lim_{\varepsilon \downarrow 0} \limsup_{n \rightarrow \infty} \frac{1}{n} \log r_{\text{span}}(n, \varepsilon, \delta).$$

Topological and metric entropy are related to each other via a variational principle [32]: for a continuous map $f : X \rightarrow X$ on a compact metric space X ,

$$h_{\text{top}}(f) = \sup_{\mu} h_\mu(f)$$

the supremum taken over all f -invariant Borel probability measures μ , i.e., such with $f_*\mu = \mu$.

If two maps $f : X \rightarrow X$ and $g : Y \rightarrow Y$ on compact metric spaces X and Y satisfy $h \circ f = g \circ h$ with a homeomorphism $h : X \rightarrow Y$, they are called *topologically conjugate* and h is called a *topological conjugacy*. In this case, $h_{\text{top}}(f) = h_{\text{top}}(g)$. If h is only a continuous surjection from X to Y , then g is called a *topological factor* of f and $h_{\text{top}}(g) \leq h_{\text{top}}(f)$.

B. Entropy Notions for Random Variables

The (*Shannon*) *entropy* of a random variable X taking values in a finite or countable set \mathbb{X} is defined by

$$H(X) := - \sum_{x \in \mathbb{X}} P(X = x) \log P(X = x)$$

where by convention $0 \cdot \log 0 = 0$. $H(X)$ is a measure for the uncertainty of the outcome of a random experiment described by X . If X and Y are two discrete random variables with values in \mathbb{X} and \mathbb{Y} , respectively, the *conditional entropy of X given $Y = y_0$* is defined by

$$H(X|Y = y_0) := - \sum_{x \in \mathbb{X}} P(X = x|Y = y_0) \log P(X = x|Y = y_0).$$

The *conditional entropy of X given Y* is defined by

$$H(X|Y) := \sum_{y \in \mathbb{Y}} P(Y = y) H(X|Y = y).$$

Observe that $H(X|Y) \leq H(X)$, i.e., additional knowledge can only decrease the uncertainty about X . Moreover, $H(X|Y) = H(X)$ if and only if X and Y are independent.

The *mutual information* of two discrete random variables X and Y is defined by

$$I(X; Y) := H(X) - H(X|Y).$$

The number $I(X; Y)$ is a measure for how much of the uncertainty of X is removed by knowing Y . It holds that $I(X; Y) = I(Y; X)$. Moreover, $I(X; Y) = 0$ if and only if X and Y are independent. There also exists a conditional version of mutual information: if X, Y , and Z are three discrete random variables, the *conditional mutual information of X and Y given Z* is defined by

$$I(X; Y|Z) := H(X|Z) - H(X|Y, Z).$$

If X is an \mathbb{R}^N -valued random variable whose associated probability measure has a density p_X , its *differential entropy* is defined by

$$h(X) := - \int_{\mathbb{R}^N} p_X(x) \log p_X(x) dx$$

if the integral exists. Similarly to the discrete case, also a conditional version of differential entropy, denoted by $h(X|Y)$, can be defined. Moreover, we can introduce the mutual information $I(X; Y)$ for two \mathbb{R}^N -valued random variables with densities in an analogous way as in the discrete case.

Finally, we can also define the conditional entropy and the mutual information for random variables of mixed type, i.e., we can talk about $h(X|Y)$, when X is continuous- and Y is discrete-valued, for instance. In general, if X and Y are any two random variables with values in \mathbb{X} and \mathbb{Y} , respectively, we can define their mutual information by

$$I(X; Y) := \sup_{Q_1, Q_2} I(Q_1(X), Q_2(Y))$$

where the supremum is taken over all quantizers Q_1, Q_2 (i.e., finite-valued measurable maps) on \mathbb{X} and \mathbb{Y} , respectively.

III. DYNAMICAL SYSTEMS APPROACH

In order to use the concepts of topological and metric entropy, defined for deterministic maps, we associate a shift map with the given stochastic system (1). More precisely, we consider the space $X^{\mathbb{Z}_+}$ of all sequences in X , equipped with the product topology. We write $\bar{x} = (x_0, x_1, x_2, \dots)$ for the elements of $X^{\mathbb{Z}_+}$ and we fix the product metric

$$D(\bar{x}, \bar{y}) := \sum_{t=0}^{\infty} \frac{1}{2^t} \frac{d(x_t, y_t)}{1 + d(x_t, y_t)} \quad (3)$$

where $d(\cdot, \cdot)$ is the given metric on X . A natural dynamical system on $X^{\mathbb{Z}_+}$ is the shift map $\theta : X^{\mathbb{Z}_+} \rightarrow X^{\mathbb{Z}_+}$, $(\theta\bar{x})_t \equiv x_{t+1}$, which is continuous with respect to the product topology. An analogous shift map is defined on $W^{\mathbb{Z}_+}$ and denoted by ϑ .

In the following, we assume that the channel is noiseless. In particular, its capacity is given by $C = \log |\mathcal{M}|$, where $\mathcal{M} = \mathcal{M}'$ is the coding alphabet. We will derive lower bounds on C_0 for the objectives (E1)–(E3).

Observing that the sequence of random variables $(x_t)_{t \in \mathbb{Z}_+}$ forms a Markov chain, when x_0 is fixed, the following lemma shows how a stationary measure of this Markov chain defines an invariant measure for θ . Its proof is given in the Appendix.

Lemma 3.1: Let π be a stationary measure of the Markov chain $(x_t)_{t \in \mathbb{Z}_+}$. Then, an invariant Borel probability measure Θ for θ is defined on cylinder sets by

$$\begin{aligned} \Theta(B_0 \times B_1 \times \dots \times B_n \times X^{[n+1, \infty)}) \\ := \int_{B_0 \times B_1 \times \dots \times B_n} \pi(dx_0) P(dx_1|x_0) \dots P(dx_n|x_{n-1}) \end{aligned}$$

where B_0, B_1, \dots, B_n are arbitrary Borel sets in X . Here

$$\begin{aligned} P(x_{n+1} \in B | x_n = x) &= P(f(x_n, w) \in B | x_n = x) \\ &= \nu((f^x)^{-1}(B)). \end{aligned}$$

The support of Θ is contained in the closure of the set of all trajectories, i.e.,

$$\text{supp}\Theta \subset \text{cl}\mathcal{T}$$

with

$$\mathcal{T} := \left\{ \bar{x} \in X^{\mathbb{Z}_+} : \exists w_t \in W \text{ with } x_{t+1} \equiv f(x_t, w_t), t \in \mathbb{Z}_+ \right\}.$$

We will also need the following characterization of topological entropy.

Lemma 3.2: Let $f : X \rightarrow X$ be a homeomorphism on a compact metric space (X, d) . Fix $\varepsilon > 0$ and $n_0 \in \mathbb{N}$. For $n > n_0$ we say that a set $E \subset X$ is $(n, \varepsilon; n_0)$ -separated if $d(f^i(x), f^i(y)) > \varepsilon$ for some $i \in \{n_0, n_0 + 1, \dots, n - 1\}$, whenever $x, y \in E$ with $x \neq y$. We write $r_{\text{sep}}(n, \varepsilon; n_0, f)$ for the maximal cardinality of an $(n, \varepsilon; n_0)$ -separated set. Then, for any choice of $n_0(\varepsilon) \in \mathbb{N}$, $\varepsilon > 0$, we have

$$h_{\text{top}}(f) = \lim_{\varepsilon \downarrow 0} \limsup_{n_0(\varepsilon) < n \rightarrow \infty} \frac{1}{n} \log r_{\text{sep}}(n, \varepsilon; n_0(\varepsilon)). \quad (4)$$

The proof of the above lemma can also be found in the Appendix.

Theorem 3.3: Consider the estimation objective (E1):

$$\sup_{t \geq T(\varepsilon)} d(x_t, \hat{x}_t) \leq \varepsilon \quad \text{a.s.} \quad (5)$$

Let C_0 denote the smallest channel capacity above which this objective can be achieved for every $\varepsilon > 0$ for a fixed initial measure π_0 which is stationary under the Markov chain $(x_t)_{t \in \mathbb{Z}_+}$. Let $\Theta = \Theta(\pi_0)$ be the measure introduced in Lemma 3.1. Then, the following holds: If $\text{supp}\Theta$ is not compact, we have $C_0 = \infty$. Otherwise we have

$$C_0 \geq h_{\text{top}}(\theta|_{\text{supp}\Theta}).$$

Proof: Assume that for some $\varepsilon > 0$, the objective (5) is achieved by a coder-estimator pair via a noiseless channel of capacity $C = \log |\mathcal{M}|$. Then, for every $k \in \mathbb{N}$, we define the set

$$\mathcal{E}_k := \{(\hat{x}_0, \hat{x}_1, \dots, \hat{x}_{k-1}) : q_t \in \mathcal{M}, 0 \leq t \leq k-1\}$$

of all possible estimation sequences of length k the estimator can generate in the time interval $[0, k-1]$.

Assume to the contrary that there exists a measurable set $A \subset X^{\mathbb{Z}_+}$ of positive measure $\alpha := \Theta(A) > 0$ so that for every $\bar{x} = (x_t)_{t \in \mathbb{Z}_+} \in A$ there is $t \geq T(\varepsilon)$ with $d(x_t, \hat{x}_t) > \varepsilon$ in case the sequence (x_t) is realized as a trajectory of the system. If $G : X \times W^{\mathbb{Z}_+} \rightarrow X^{\mathbb{Z}_+}$ is the map from the proof of Lemma 3.1, mapping a pair (x_0, \bar{w}) to the corresponding trajectory, then the preimage $G^{-1}(A)$ is measurable in $X \times W^{\mathbb{Z}_+}$ with $\pi_0 \times \nu^{\mathbb{Z}_+}$ -measure $\alpha > 0$. This contradicts the assumption that the almost sure estimation objective (5) is achieved. Hence, the set

$$\tilde{\mathcal{T}} := \{ \bar{x} \in X^{\mathbb{Z}_+} : d(x_t, \hat{x}_t) \leq \varepsilon \text{ for all } t \geq T(\varepsilon) \}$$

has measure one and consequently is dense in $\text{supp}\Theta$.

Choose $\tau = \tau(\varepsilon)$ large enough so that

$$\sum_{t=\tau}^{\infty} \frac{1}{2^t} \leq \varepsilon.$$

Let $E \subset \text{supp}\Theta$ be a finite $(k, 5\varepsilon; T(\varepsilon))$ -separated set for some $k > T(\varepsilon)$. Since $\tilde{\mathcal{T}}$ is dense in $\text{supp}\Theta$, a small perturbation of E yields a $(k, 5\varepsilon; T(\varepsilon))$ -separated set in $\tilde{\mathcal{T}}$ with the same cardinality as E (using that θ is continuous). Hence, we may assume $E \subset \tilde{\mathcal{T}}$. We define a map $\alpha : E \rightarrow \mathcal{E}_{k+\tau}$ by assigning to $(x_t)_{t \in \mathbb{Z}_+} \in E$ the estimation sequence generated by the estimator when it receives the signals $q_t = q_t(x_0, \dots, x_t)$ for $t = 0, 1, \dots, k + \tau - 1$.

Assuming $\alpha(\bar{x}) = \alpha(\bar{y})$ for some $\bar{x}, \bar{y} \in E$, we find for $T(\varepsilon) \leq t \leq k$ that

$$\begin{aligned} &D(\theta^t(\bar{x}), \theta^t(\bar{y})) \\ &\leq \sum_{s=0}^{\tau-1} \frac{1}{2^s} \frac{d(x_{t+s}, y_{t+s})}{1 + d(x_{t+s}, y_{t+s})} + \sum_{s=\tau}^{\infty} \frac{1}{2^s} \\ &\leq \sum_{s=0}^{\tau-1} \frac{1}{2^s} d(x_{t+s}, \hat{x}_{t+s}) + \sum_{s=0}^{\tau-1} \frac{1}{2^s} d(\hat{y}_{t+s}, y_{t+s}) + \varepsilon \\ &\leq 2\varepsilon + 2\varepsilon + \varepsilon = 5\varepsilon \end{aligned}$$

implying $\bar{x} = \bar{y}$, since E is $(k, 5\varepsilon; T(\varepsilon))$ -separated. Hence, the map α is injective.

The set $\text{supp}\Theta$ is a closed subset of the complete metric space $(X^{\mathbb{Z}_+}, D)$. Hence, it is also a complete metric space. If we assume that $\text{supp}\Theta$ is not compact, it thus follows that $\text{supp}\Theta$ is not totally bounded, implying that $(k, 5\varepsilon; T(\varepsilon))$ -separated subsets of $\text{supp}\Theta$ of arbitrarily large (finite) cardinality exist, when ε is sufficiently small. Hence, $\mathcal{E}_{k+\tau}$ must be infinite, leading to the contradiction $|\mathcal{M}| = \infty$. Consequently, in this case, the estimation problem cannot be solved via a channel of finite capacity.

Now assume that $\text{supp}\Theta$ is compact. Choosing a maximal $(k, 5\varepsilon; T(\varepsilon))$ -separated set E for the dynamical system $\theta_{|\text{supp}\Theta} : \text{supp}\Theta \rightarrow \text{supp}\Theta$, we obtain the inequality

$$r_{\text{sep}}(k, 5\varepsilon; T(\varepsilon)) \leq |\mathcal{E}_{k+\tau}| \leq |\mathcal{M}|^{k+\tau}.$$

This implies

$$\limsup_{k \rightarrow \infty} \frac{1}{k} \log r_{\text{sep}}(k, 5\varepsilon; T(\varepsilon)) \leq \log |\mathcal{M}| = C.$$

Using Lemma 3.2, the result follows by letting $C \rightarrow C_\varepsilon$ and $\varepsilon \downarrow 0$. ■

Remark 3.4: To make the statement of the theorem clearer, let us consider the two extreme cases when there is no noise and when there is only noise:

- i) If the system is deterministic, i.e., $x_{t+1} = f(x_t)$ for a homeomorphism $f : X \rightarrow X$ of a compact metric space X , then π_0 is an invariant measure of f . Moreover, $P(x_t \in B | x_{t-1} = x) = 1$ if $f(x) \in B$ and 0 otherwise, implying

$$\begin{aligned} \Theta(B_0 \times B_1 \times \dots \times B_n \times X^{[n+1, \infty)}) \\ = \pi_0(B_0 \cap f^{-1}(B_1) \cap f^{-2}(B_2) \cap \dots \cap f^{-n}(B_n)). \end{aligned}$$

From this expression, we see that the support of Θ is contained in the set \mathcal{T} of all trajectories of f (which in this case coincides with its closure), as already proved in Lemma 3.1. The map $h : \mathcal{T} \rightarrow X$ defined by $h(\bar{x}) := x_0$, is easily seen to be a homeomorphism, which conjugates $\theta_{|\mathcal{T}}$ and f . That is, the following diagram commutes:

$$\begin{array}{ccc} \mathcal{T} & \xrightarrow{\theta} & \mathcal{T} \\ h \downarrow & & \downarrow h \\ X & \xrightarrow{f} & X \end{array}$$

Since $h_*\Theta = \pi_0$ and conjugate systems have the same entropy, our theorem implies

$$C_0 \geq h_{\text{top}}(f, \text{supp}\pi_0). \tag{6}$$

The right-hand side of this inequality is finite under mild assumptions, e.g., if f is Lipschitz continuous on $\text{supp}\pi_0$ and $\text{supp}\pi_0$ has finite lower box dimension (see [2, Thm. 6.1.2]). These conditions are in particular satisfied when f is a diffeomorphism on a finite-dimensional manifold. However, one should be aware that even on a compact interval, there exist continuous maps with infinite topological entropy on the support of an invariant

measure. The lower bound (6) has already been derived in [20, Thm. 3.1], and in fact for the deterministic case considered here the bound was shown to be tight.

- ii) Assume that $X = W$ is compact and the system is given by $x_{t+1} = w_t$, i.e., the trajectories are only determined by the noise. In this case, with $\pi_0 := \nu$, the measure Θ is the product measure $\nu^{\mathbb{Z}_+}$. Hence, C_0 is bounded below by the topological entropy of the shift on $W^{\mathbb{Z}_+}$ restricted to $\text{supp}\nu^{\mathbb{Z}_+} = (\text{supp}\nu)^{\mathbb{Z}_+}$. This number is finite if and only if $\text{supp}\nu$ is finite and in this case is given by $\log |\text{supp}\nu|$.

If the system is not deterministic, then usually $C_0 = \infty$. In fact, this is always the case if the estimator is able to recover the noise to a sufficiently large extent. The following corollary treats the case, when the noise can be recovered completely from the state trajectory.

Corollary 3.5: Additionally to the assumptions in Theorem 3.3, suppose that W and X are compact and $f^x : W \rightarrow X$ is invertible for every $x \in X$ so that $(x, y) \mapsto (f^x)^{-1}(y)$ is continuous. Then, for (E1),

$$C_0 \geq h_{\text{top}}(\Phi_{|\text{supp}(\pi_0 \times \nu^{\mathbb{Z}_+})}) \geq h_{\text{top}}(\vartheta_{|\text{supp}\nu^{\mathbb{Z}_+}}) \tag{7}$$

where $\Phi : X \times W^{\mathbb{Z}_+} \rightarrow X \times W^{\mathbb{Z}_+}$ is the skew-product map $(x, \bar{w}) \mapsto (f_{w_0}(x), \vartheta\bar{w})$. As a consequence, $C_0 = \infty$ whenever $\text{supp}\nu$ contains infinitely many elements.

Proof: We consider the map $h : X^{\mathbb{Z}_+} \rightarrow W^{\mathbb{Z}_+}$, $\bar{x} \mapsto \bar{w} = (w_t)_{t \in \mathbb{Z}_+}$ with

$$w_t = (f^{x_t})^{-1}(x_{t+1}).$$

If we equip $W^{\mathbb{Z}_+}$ with the product topology, h becomes continuous. Indeed, if the distance of two points $\bar{x}^1, \bar{x}^2 \in X^{\mathbb{Z}_+}$ is small, then the distances $d_X(\bar{x}_t^1, \bar{x}_t^2)$ are small for finitely many values of t . Hence, by the uniform continuity of $(x, y) \mapsto f_x^{-1}(y)$ on the compact space $X \times X$, also the distances $d_W(h(\bar{x}^1)_t, h(\bar{x}^2)_t)$ can be made small for sufficiently many values of t , guaranteeing that $D(h(\bar{x}^1), h(\bar{x}^2))$ becomes small, where D is a product metric on $W^{\mathbb{Z}_+}$.

The map $G : X \times W^{\mathbb{Z}_+} \rightarrow X^{\mathbb{Z}_+}$, used in the proof of Lemma 3.1, satisfies

$$h(G(x_0, \bar{w})) = \bar{w} \quad \text{for all } (x_0, \bar{x}) \in X \times W^{\mathbb{Z}_+}$$

because we can write

$$G(x_0, \bar{w}) = (x_0, f^{x_0}(w_0), f^{x_1}(w_1), f^{x_2}(w_2), \dots).$$

Consequently, G —as a map from $X \times W^{\mathbb{Z}_+}$ to the space \mathcal{T} of trajectories—is invertible with

$$G^{-1}(\bar{x}) = (x_0, h(\bar{x})).$$

From the assumptions, it follows that G is continuous, hence G is a homeomorphism and \mathcal{T} is compact. By the proof of Lemma 3.1, we have $\theta \circ G = G \circ \Phi$, where Φ is the skew-product map $\Phi(x, \bar{w}) = (f(x, w_0), \vartheta\bar{w})$ and $G_*(\pi_0 \times \nu^{\mathbb{Z}_+}) = \Theta$. Hence, G is a topological conjugacy between $\theta_{|\text{supp}\Theta}$ and $\Phi_{|\text{supp}(\pi_0 \times \nu^{\mathbb{Z}_+})}$, implying

$$C_0 \geq h_{\text{top}}(\theta_{|\text{supp}\Theta}) = h_{\text{top}}(\Phi_{|\text{supp}(\pi_0 \times \nu^{\mathbb{Z}_+})}).$$

Since the projection map $\pi : (x, \bar{w}) \mapsto \bar{w}$ exhibits ϑ as a topological factor of Φ and $\pi_*(\pi_0 \times \nu^{\mathbb{Z}_+}) = \nu^{\mathbb{Z}_+}$, the second inequality in (7) follows. ■

Example 3.6: Let $X = W = S^1 = \mathbb{R}/\mathbb{Z}$. Let $f(x, w) = x + w \bmod 1$ and let $\pi_0 = \nu$ be the normalized Lebesgue measure on S^1 . In this case, the map $f^x : S^1 \rightarrow S^1, w \mapsto x + w$, is obviously invertible and $(x, y) \mapsto (f^x)^{-1}(y) = x - y$ is continuous. Hence, $C_0 = \infty$ for the estimation objective (E1).

Theorem 3.7: Consider the estimation objective (E2)

$$P\left(\limsup_{t \rightarrow \infty} d(x_t, \hat{x}_t) \leq \varepsilon\right) = 1. \quad (8)$$

Let C_0 denote the smallest channel capacity above which this objective can be achieved for every $\varepsilon > 0$ for a fixed initial measure π_0 which is stationary and ergodic under the Markov chain $(x_t)_{t \in \mathbb{Z}_+}$. Let $\Theta = \Theta(\pi_0)$ be the measure introduced in Lemma 3.1. Then, if $\text{supp}\Theta$ is compact, we have

$$C_0 \geq h_\Theta(\theta).$$

Proof: First observe that the ergodicity of π_0 implies the ergodicity of Θ . Indeed, it is well known that the product measure $\pi_0 \times \nu^{\mathbb{Z}_+}$ is ergodic for the skew-product Φ if π_0 is ergodic (cf. [22]). Since $G_*(\pi_0 \times \nu^{\mathbb{Z}_+}) = \Theta$ and $\theta \circ G = G \circ \Phi$, this implies the ergodicity of Θ . Now consider a noiseless channel with input alphabet \mathcal{M} and a pair of coder and decoder/estimator which solves the estimation problem (E2) for some $\varepsilon > 0$. For every $\bar{x} \in X^{\mathbb{Z}_+}$ and $\delta > \varepsilon$, let

$$T(\bar{x}, \delta) := \inf \left\{ k \in \mathbb{N} : \sup_{t \geq k} d(x_t, \hat{x}_t) \leq \delta \right\}$$

where the infimum is defined as $+\infty$ if the corresponding set is empty. Note that $T(\bar{x}, \delta)$ depends measurably on \bar{x} . Define

$$B^K(\delta) := \{\bar{x} \in \text{supp}\Theta : T(\bar{x}, \delta) \leq K\} \forall \delta > \varepsilon, K \in \mathbb{N}$$

and observe that these sets are measurable. From (8), it follows that for every $\delta > \varepsilon$

$$\lim_{K \rightarrow \infty} \Theta(B^K(\delta)) = \Theta\left(\bigcup_{K \in \mathbb{N}} B^K(\delta)\right) = 1.$$

Fixing a K large enough so that $\Theta(B^K(\delta)) > 0$, Katok's characterization of metric entropy yields the assertion, which is proved with the same arguments as in the proof of Theorem 3.3, using the simple fact a maximal (n, ε) -separated set contained in some set K also (n, ε) -spans K . ■

In the following, we consider (E3). To obtain a lower bound, we restrict the encoder to have finite memory and be periodic.

Theorem 3.8: Consider the estimation objective (E3)

$$\limsup_{t \rightarrow \infty} E[d(x_t, \hat{x}_t)^2] \leq \varepsilon \quad (9)$$

for an initial measure π_0 which is stationary and ergodic under the Markov chain $(x_t)_{t \in \mathbb{Z}_+}$. Additionally, assume that there exists $\tau > 0$ so that the coder map δ_t is of the form

$$q_t = \delta_t(x_{[t-\tau+1, t]}) \quad (10)$$

and is periodic so that $\delta_{t+\tau} \equiv \delta_t$. Further assume that the estimator map is of the form

$$\hat{x}_t = \gamma_t(q_{[t-\tau+1, t]})$$

and also $\gamma_{t+\tau} \equiv \gamma_t$. Then, if the support of the measure $\Theta = \Theta(\pi_0)$, introduced in Lemma 3.1, is compact, the smallest channel capacity C_0 above which (E3) can be achieved for every $\varepsilon > 0$ satisfies

$$C_0 \geq h_\Theta(\theta).$$

Proof: First note that we would obtain a lower bound on C_0 if we allowed the periodic encoders to be of the form

$$q_t = \delta_t(x_{[t-\tau+1, \infty)}) \quad (11)$$

i.e., if we allow the encoder to have noncausal access to the realizations of x_t . Note that every encoder policy of the form (10) would be of the form (11). We keep the structure of the decoder as is.

The criterion (E3) considered in this article implies (E3) considered in [20] with the distortion metric d being the product metric D introduced in (3) for the dynamical system θ and $p = 2$: This follows since $\limsup_{t \rightarrow \infty} E[(d(x_t, \hat{x}_t))^2] \leq \varepsilon$ implies that with $\bar{x}_t = (x_t, x_{t+1}, \dots)$ and $\hat{\bar{x}}_t = (\hat{x}_t, \hat{x}_{t+1}, \dots)$, we have

$$\begin{aligned} & \limsup_{t \rightarrow \infty} E[D(\bar{x}_t, \hat{\bar{x}}_t)^2] \\ &= \limsup_{t \rightarrow \infty} E \left[\left(\sum_{i=0}^{\infty} 2^{-i} \frac{d(x_{t+i}, \hat{x}_{t+i})}{1 + d(x_{t+i}, \hat{x}_{t+i})} \right)^2 \right] \\ &\leq \limsup_{t \rightarrow \infty} E \left[4 \left(\sum_{i=0}^{\infty} 2^{-(i+1)} \frac{d(x_{t+i}, \hat{x}_{t+i})}{1 + d(x_{t+i}, \hat{x}_{t+i})} \right)^2 \right] \\ &\leq 4 \limsup_{t \rightarrow \infty} \sum_{i=0}^{\infty} 2^{-(i+1)} E \left[\left(\frac{d(x_{t+i}, \hat{x}_{t+i})}{1 + d(x_{t+i}, \hat{x}_{t+i})} \right)^2 \right] \quad (12) \\ &\leq 4 \sum_{i=0}^{\infty} 2^{-(i+1)} \limsup_{t \rightarrow \infty} E \left[\left(\frac{d(x_{t+i}, \hat{x}_{t+i})}{1 + d(x_{t+i}, \hat{x}_{t+i})} \right)^2 \right] \\ &\leq 4 \sum_{i=0}^{\infty} 2^{-(i+1)} \limsup_{t \rightarrow \infty} E \left[\left(\frac{d(x_{t+i}, \hat{x}_{t+i})^2}{1} \right) \right] \\ &= 4 \sum_{i=0}^{\infty} 2^{-(i+1)} \limsup_{t \rightarrow \infty} E[d(x_{t+i}, \hat{x}_{t+i})^2] \\ &\leq 4 \sum_{i=0}^{\infty} 2^{-(i+1)} \varepsilon \\ &= 4\varepsilon =: \bar{\varepsilon}. \quad (13) \end{aligned}$$

In particular, $\bar{\varepsilon} \rightarrow 0$ as $\varepsilon \rightarrow 0$. Thus, if (E3) can be achieved for every $\varepsilon > 0$, (E3) considered in [20] can also be achieved for every $\varepsilon > 0$. Here, we apply Jensen's inequality in (12).

Thus, if the encoder is of the form (11), the problem can be viewed as an instance of [20, Thm. 5.2] for the dynamical system θ .

Under the stated periodicity assumption and (11), [20, Thm. 5.2] directly implies that $C_0 \geq h_\Theta(\theta)$. ■

Three remarks are in order.

Remark 3.9: It is worth noting here that for a deterministic dynamical system, the property of being ergodic is typically very restrictive; however, for a stochastic system, ergodicity is often very simple to satisfy: the presence of noise often leads to strong mixing conditions which directly leads to ergodicity.

Remark 3.10: The discussion in Theorem 3.8 leads to an interesting relation between the classical information-theoretic problem of optimally encoding (noncausally) sequences of random variables and metric entropy of an infinite-dimensional dynamical system defined via the shift operator. A close look at the proof of [20, Thm. 5.1] reveals that under a stationarity and ergodicity assumption, when the channel is noise-free, the lower bound presented in Theorem 3.8 is essentially achievable, provided that the encoder has noncausal access to the source realizations and in particular a large enough *look-ahead* is sufficient for an approximately optimal performance. Note though that the decoder is still restricted to be zero-delay.

Remark 3.11: Besides our results in [20], a related study to the approach of this section is due to Savkin [41]. This article is concerned with nonlinear systems of the form

$$x(t+1) = F(x(t), \omega(t))$$

where $\omega(t)$ is interpreted as an uncertainty input or a disturbance. However, no statistical structure is imposed on ω so that the system can be regarded as deterministic (thus, the formulation is distribution-free). A characterization of the smallest channel capacity C_0 above which the state $x(t)$ can be estimated with arbitrary precision and for every initial state $x(0)$ in a specified compact set via a noiseless channel is given by [41, Thm. 3.1]. A close inspection of this result shows that it characterizes C_0 precisely as the topological entropy of the associated shift operator acting on system trajectories. Moreover, [41, Thm. 3.2] shows that under mild assumptions on the system, the entropy of this operator is infinite, and hence observation via a finite-capacity channel is not possible.

IV. INFORMATION-THEORETIC AND PROBABILITY-THEORETIC APPROACH

In this section, we consider a much broader class of channels, the so-called *Class A type channels* (see [51, Def. 8.5.1]). We restrict ourselves to systems with state-space $X = \mathbb{R}^N$ and provide lower bounds of the channel capacity for the objectives (E2) and (E3), using information-theoretic methods.

Definition 4.1: A channel is said to be of Class A type, if

1) it satisfies the following Markov chain condition:

$$q'_t \leftrightarrow q_t, q_{[0,t-1]}, q'_{[0,t-1]} \leftrightarrow \{x_0, w_s : s \geq 0\},$$

i.e., almost surely, for all Borel sets B ,

$$\begin{aligned} P(q'_t \in B | q_t, q_{[0,t-1]}, q'_{[0,t-1]}, x_0, \{w_s\}_{s \geq 0}) \\ = P(q'_t \in B | q_t, q_{[0,t-1]}, q'_{[0,t-1]}) \end{aligned}$$

for all $t \geq 0$, and

2) its capacity with feedback is given by

$$\begin{aligned} C &= \lim_{T \rightarrow \infty} \max_{\{P(q_t | q_{[0,t-1]}, q'_{[0,t-1]}), 0 \leq t \leq T-1\}} \\ &\frac{1}{T} I(q_{[0,T-1]} \rightarrow q'_{[0,T-1]}) \end{aligned}$$

where the directed mutual information is defined by

$$\begin{aligned} I(q_{[0,T-1]} \rightarrow q'_{[0,T-1]}) &:= \sum_{t=1}^{T-1} I(q_{[0,t]}; q'_t | q'_{[0,t-1]}) \\ &+ I(q_0; q'_0) \end{aligned}$$

Discrete noiseless channels and memoryless channels belong to this class; for such channels, feedback does not increase the capacity [5]. Class A type channels also include finite-state stationary Markov channels which are indecomposable [38] and non-Markov channels which satisfy certain symmetry properties [42]. Further examples can be found in [6] and [45].

Theorem 4.2: Consider system (1) with state-space $X = \mathbb{R}^N$. Suppose that

$$\limsup_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^{T-1} h(x_t | x_{t-1}) > -\infty$$

and $h(x_t) < \infty$ for all $t \in \mathbb{Z}_+$. Then, the smallest channel capacity C_ε above which (E3) can be achieved over a Class A type channel satisfies

$$C_\varepsilon \geq \left(\limsup_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^{T-1} h(x_t | x_{t-1}) \right) - \frac{N}{2} \log(2\pi\varepsilon).$$

This, in particular, implies $C_0 = \infty$ for the smallest channel capacity C_0 above which (E3) can be achieved for every $\varepsilon > 0$. That is, (E3) cannot be achieved for every $\varepsilon > 0$ over a finite-capacity Class A type channel.

Proof: Let $(\varepsilon_t)_{t \in \mathbb{Z}_+}$ be a sequence of nonnegative real numbers so that $E[\|x_t - \hat{x}_t\|^2] \leq \varepsilon_t$ for all $t \in \mathbb{Z}_+$ and $\limsup_{t \rightarrow \infty} \varepsilon_t \leq \varepsilon$. Observe that for every $t \geq 1$, we have

$$\begin{aligned} I(q'_t; q_{[0,t]} | q'_{[0,t-1]}) \\ &= H(q'_t | q'_{[0,t-1]}) - H(q'_t | q_{[0,t]}, q'_{[0,t-1]}) \\ &= H(q'_t | q'_{[0,t-1]}) - H(q'_t | q_{[0,t]}, x_t, q'_{[0,t-1]}) \\ &\geq H(q'_t | q'_{[0,t-1]}) - H(q'_t | x_t, q'_{[0,t-1]}) \\ &= I(x_t; q'_t | q'_{[0,t-1]}). \end{aligned} \tag{14}$$

Here, (14) follows from the assumption that the channel is of Class A type. Define

$$R_T := \max_{\substack{P(q_t | q_{[0,t-1]}, q'_{[0,t-1]}), \\ 0 \leq t \leq T-1}} \frac{1}{T} \sum_{t=0}^{T-1} I(q'_t; q_{[0,t]} | q'_{[0,t-1]}).$$

Now consider the following identities and inequalities:

$$\begin{aligned}
& \lim_{T \rightarrow \infty} R_T \\
& \geq \limsup_{T \rightarrow \infty} \frac{1}{T} \left(\sum_{t=1}^{T-1} I(x_t; q'_t | q'_{[0,t-1]}) + I(x_0; q'_0) \right) \\
& = \limsup_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^{T-1} \left(h(x_t | q'_{[0,t-1]}) - h(x_t | q'_{[0,t]}) \right) \\
& \geq \limsup_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^{T-1} \left(h(x_t | x_{[0,t-1]}, q'_{[0,t-1]}) - h(x_t | q'_{[0,t]}) \right) \\
& = \limsup_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^{T-1} h(x_t | x_{[0,t-1]}, q'_{[0,t-1]}) - h(x_t - \hat{x}_t | q'_{[0,t]}) \\
& \geq \limsup_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^{T-1} \left(h(x_t | x_{[0,t-1]}, q'_{[0,t-1]}) - h(x_t - \hat{x}_t) \right) \\
& \geq \limsup_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^{T-1} \left(h(x_t | x_{[0,t-1]}, q'_{[0,t-1]}) - \frac{N}{2} \log(2\pi e \varepsilon_t) \right) \\
& = \limsup_{T \rightarrow \infty} \frac{1}{T} \left(\sum_{t=1}^{T-1} h(x_t | x_{t-1}) \right) - \frac{N}{2} \log(2\pi e \varepsilon). \quad (15)
\end{aligned}$$

Here, the second inequality uses the property that entropy decreases under conditioning on more information. The second equality follows from the fact that \hat{x}_t is a function of $q'_{[0,t]}$, and the last inequality follows from that fact that among all real random variables X that satisfy a given second moment constraint $E[X^2] \leq \varepsilon$, a Gaussian maximizes the entropy and the differential entropy in this case is given by $\frac{1}{2} \log(2\pi e \varepsilon)$. Using the fact that for an N -dimensional vector $X = (X_1, \dots, X_N)^T$, $h(X) = h(X_1) + \sum_{i=2}^N h(X_i | X_{[1,i-1]}) \leq \sum_{i=1}^N h(X_i)$, it follows with $E[\|x_t - \hat{x}_t\|^2] \leq \varepsilon_t$ that $h(x_t - \hat{x}_t) \leq \frac{N}{2} \log(2\pi e \varepsilon_t)$. The final equality then follows from the fact that conditioned on x_{t-1} , x_t and $q'_{[0,t-1]}$ are independent. For the final result, in (15), taking the limit as $\varepsilon \rightarrow 0$, $\log(\varepsilon) \rightarrow -\infty$, and $C_0 = \infty$ follows. ■

Theorem 4.3: Suppose that $X \subset \mathbb{R}^N$ and the system given by $x_{t+1} = f(x_t, w_t)$ satisfies

$$P(x_{t+1} \in B | x_t = x) \leq K\lambda(B)$$

for all Borel sets $B \subset \mathbb{R}^N$, where λ denotes the Lebesgue measure, $K \in \mathbb{R}_+$, and w_t is an i.i.d. noise process. Then, the smallest channel capacity C_0 above which (E2) can be achieved for every $\varepsilon > 0$ satisfies $C_0 = \infty$. That is, (E2) cannot be achieved for every $\varepsilon > 0$ over a finite-capacity Class A type channel.

A special case for the above is a system of the form

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

where $w_t \sim \nu$ with the noise measure ν admitting a bounded density function.

Proof: Given a finite alphabet channel with $|\mathcal{M}'| < \infty$, for a given time $t > 0$ under any encoding and decoding policy, there exists a finite partition of the state-space X for encoding x_t

leading to \hat{x}_t . Thus, there exists $\bar{\varepsilon} > 0$ so that for all $\varepsilon \in (0, \bar{\varepsilon})$, for each set

$$A_t(q'_0, \dots, q'_t) := \{x \in \mathbb{R}^N : d(x, \hat{x}_t(q'_0, \dots, q'_{t-1}, q'_t)) \geq \varepsilon\}$$

where $q'_0, \dots, q'_t \in \mathcal{M}'$, we find that

$$\begin{aligned}
& P\left(x_t \in A_t(q'_0, \dots, q'_t) | x_{[0,t-1]}, q'_{[0,t-1]}\right) \\
& = \sum_{q' \in \mathcal{M}'} P\left(q'_t = q' | x_{[0,t-1]}, q'_{[0,t-1]}\right) \\
& \quad \times P\left(x_t \in A_t(q'_0, \dots, q'_t) | x_{[0,t-1]}, q'_{[0,t-1]}, q'_t = q'\right) \\
& \geq 1 - \sum_{q' \in \mathcal{M}'} P\left(q'_t = q' | x_{[0,t-1]}, q'_{[0,t-1]}\right) \\
& \quad \times P\left(d(x_t, \hat{x}_t(q'_0, \dots, q'_{t-1}, q')) < \varepsilon | x_{[0,t-1]}, q'_{[0,t-1]}, q'_t = q'\right) \\
& \geq 1 - |\mathcal{M}'| K \lambda(B_\varepsilon(0)) > 0.
\end{aligned}$$

This implies that

$$P\left(x_t \in A_t(q'_0, \dots, q'_t) \mid x_{[0,t-1]}, q'_{[0,t-1]}\right) > 0 \quad (16)$$

uniform over all realizations of $x_{[0,t-1]}, q'_{[0,t-1]}$.

Let

$$\eta := \sum_{t=1}^{\infty} \mathbb{1}_{\{x_t \in A_t(q'_0, \dots, q'_t)\}}.$$

Our goal is to show that $\eta = \infty$ almost surely, leading to the desired conclusion. Let

$$\tau(1) = \min\{t > 0 : x_t \in A_t(q'_0, \dots, q'_t)\}$$

and for $z > 1, z \in \mathbb{N}$, let

$$\tau(z) = \min\{t > \tau(z-1) : x_t \in A_t(q'_0, \dots, q'_t)\}.$$

It follows that $P(\tau(1) < \infty) = 1$ by a repeated use of (16), since the event $\tau(1) = \infty$ would imply that the event (whose probability is lower bounded by (16)) would be avoided infinitely many times, leading to a zero measure. Thus, $P(\eta \geq 1) = 1$. By a repeated use of (16) and induction if $P(\eta \geq k-1) = 1$, we find that

$$\begin{aligned}
P(\eta \geq k) & = P(\eta \geq k, \eta \geq k-1) \\
& = P(\tau(1) < \infty | \mathcal{F}_{\tau(k-1)}) P(\eta \geq k-1) = 1
\end{aligned}$$

where $\mathcal{F}_{\tau(k-1)}$ is the σ -field generated by $\{x_s, q'_s\}$ up to time $\tau(k-1)$. Thus, for every $k \in \mathbb{N}$, $P(\eta \geq k) = 1$, and it follows by continuity in probability that $P(\eta = \infty) = \lim_{k \rightarrow \infty} P(\eta \geq k) = 1$. Hence, for any finite communication rate, almost sure boundedness is not possible for arbitrarily small $\varepsilon > 0$. ■

V. ACHIEVABILITY BOUNDS

A. Coding of Deterministic Dynamical Systems Over Noisy Communication Channels

In this section, we show that for a noise-free system, a discrete memoryless noisy communication channel is no obstruction for

achieving the objectives (E2) and (E3) with finite capacity. More precisely, we prove the following theorem.

Theorem 5.1: Consider a nonlinear deterministic system $x_{t+1} = f(x_t)$ given by a continuous map $f : X \rightarrow X$ on a compact metric space X , estimated via a discrete memoryless channel (DMC). Then, for the asymptotic estimation objectives (E2) and (E3), the smallest channel capacity C_0 above which these objectives can be achieved for every $\varepsilon > 0$ satisfies

$$C_0 \leq h_{\text{top}}(f).$$

Proof: It suffices to prove the result for (E2), since for a compact metric space, (E2) implies (E3); therefore, the construction below also applies for the objective (E3).

Without loss of generality, we may assume that $h_{\text{top}}(f) < \infty$, since otherwise the statement trivially holds. Then, it suffices to show that for any $\varepsilon > 0$, the estimation objective can be achieved whenever the channel capacity satisfies $C > h_{\text{top}}(f)$. Since the capacity of a DMC can take any positive value, it follows that $C_\varepsilon \leq h_{\text{top}}(f)$ for every $\varepsilon > 0$ and thus $C_0 \leq h_{\text{top}}(f)$.

Now, consider a channel with capacity $C > h_{\text{top}}(f)$ and fix $\varepsilon > 0$. Recall that the input alphabet is denoted by \mathcal{M} and the output alphabet by \mathcal{M}' . By the random coding construction of Shannon [9], we can achieve a rate R satisfying

$$h_{\text{top}}(f) < R < C \quad (17)$$

with a sequence of increasing sets $\{1, \dots, M_n\}$ of input messages so that for all n

$$2^{nR} \leq M_n \quad \text{and} \quad \lim_{n \rightarrow \infty} \frac{1}{n} \log M_n = C. \quad (18)$$

Furthermore, there exists a sequence of encoders $E^n : \{1, \dots, M_n\} \rightarrow \mathcal{M}^n$, yielding codewords $x^n(1), \dots, x^n(M_n)$, and a sequence of decoders $D^n : (\mathcal{M}')^n \rightarrow \{1, \dots, M_n\}$ so that

$$P(D^n(q'_{[0,n-1]}) \neq c | q_{[0,n-1]} = x^n(c)) \leq e^{-nE(R)+o(n)}$$

uniformly for all $c \in \{1, \dots, M_n\}$. Here, $\frac{o(n)}{n} \rightarrow 0$ as $n \rightarrow \infty$ and $E(R) > 0$. In particular, we observe that with $c_n \in \{1, \dots, M_n\}$ being the message transmitted and $D^n(q'_{[0,n-1]})$ the decoder output

$$\begin{aligned} & P(D^n(q'_{[0,n-1]}) \neq c_n) \\ &= \sum_{c \in \{1, \dots, M_n\}} P(D^n(q'_{[0,n-1]}) \neq c | q_{[0,n-1]} = x^n(c)) \\ & \quad \times P(q_{[0,n-1]} = x^n(c)) \\ & \leq e^{-nE(R)+o(n)}. \end{aligned}$$

This also implies that the bound holds even when the messages to be transmitted are not uniformly distributed. Thus, for the sequence of encoders and decoders constructed above, we have

$$\sum_n P(D^n(q'_{[0,n-1]}) \neq c_n) \leq \sum_n e^{-nE(R)+o(n)} < \infty.$$

The Borel–Cantelli Lemma then implies

$$P\left(\left\{D^n(q'_{[0,n-1]}) \neq c_n \text{ infinitely often}\right\}\right) = 0. \quad (19)$$

Now we choose $\delta \in (0, \varepsilon)$ so that, by uniform continuity,

$$d(x, y) < \delta \quad \Rightarrow \quad d(f(x), f(y)) < \varepsilon \quad \text{for all } x, y \in X. \quad (20)$$

Furthermore, we choose N sufficiently large so that

$$r_{\text{span}}(n, \delta) \leq M_n \quad \text{for all } n \geq N. \quad (21)$$

This is possible, because by (17) and (18), for every $n \in \mathbb{N}$, we have

$$\begin{aligned} \limsup_{n \rightarrow \infty} \frac{1}{n} \log r_{\text{span}}(n, \delta) & \leq h_{\text{top}}(f) \\ & < R = \frac{1}{n} \log 2^{nR} \leq \frac{1}{n} \log M_n. \end{aligned}$$

Let S_j be a (j, ε) -spanning set of cardinality $r_{\text{span}}(j, \delta)$ and fix injective functions

$$\iota_j : S_j \rightarrow \{1, \dots, M_j\}.$$

In fact, by possibly enlarging the set S_j , we can assume that ι_j is bijective. For any $a \in X$, let $x_j^*(a)$ denote a fixed element of S_j satisfying $d(f^t(x^*(a)), f^t(a)) \leq \delta$ for $0 \leq t \leq j-1$.

Define sampling times by

$$\tau_0 := 0 \quad \text{and} \quad \tau_{j+1} := \tau_j + j + 1 \quad \text{for } j \geq 0.$$

In the following, we specify the coding scheme. In this coding scheme, the encoder from τ_j to $\tau_{j+1} - 1$ encodes the information regarding the orbit of the state from τ_{j+1} to $\tau_{j+2} - 1$. For all $j \geq N$, at time τ_j , use the input $\iota_{j+1}(x_{j+1}^*(f^{j+1}(x_{\tau_j})))$ for the encoder, where x_{τ_j} is the state at time τ_j . Then, $x^{j+1}(\iota_{j+1}(x_{j+1}^*(f^{j+1}(x_{\tau_j}))))$ is sent during the next $j+1$ units of time. This is possible by (21). For $j < N$, it is not important what we transmit.

Let the estimator apply $x_{j+1}^* \circ \iota_{j+1}^{-1}$ to the output of the decoder, obtaining an element $y_{j+1} \in S_{j+1}$, and use $y_{j+1}, f(y_{j+1}), \dots, f^j(y_{j+1}), f^{j+1}(y_{j+1})$ as the estimates during the forthcoming time interval of length $\tau_{j+2} - \tau_{j+1} = \tau_{j+1} - \tau_j + 1 = j + 2$. Then, $\delta < \varepsilon$ (20), and the fact that S_{j+1} is $(j+1, \delta)$ -spanning implies that the desired estimation accuracy is achieved, provided that there was no error in the transmission. Now (19) implies that after a finite random time, there are no more errors in the transmission. By the analysis above, the errors will be uniformly bounded by ε . Hence, the objective (E2) is achieved. ■

The proof above crucially depends on the fact that the system is deterministic. The theorem is essentially a possibility result and we note that the proof does not make use of the fact that the encoder has access to the realizations of the channel output. For linear systems, a constructive proof is given in [31, Thm. 6.4.1]. We present this result for completeness. It provides a positive answer to the question whether the estimation objective (E2) can be achieved, when no noise is present in the system.

Theorem 5.2: Consider the noiseless linear system

$$x_{t+1} = Ax_t \quad (22)$$

with $A \in \mathbb{R}^{N \times N}$, estimated over a memoryless erasure channel with finite capacity. Then, the smallest channel capacity C_0

above which (E2) can be achieved for every $\varepsilon > 0$ satisfies

$$C_0 \leq \sum_{i=1}^N \max\{0, \log\lceil |\lambda_i| \rceil\} \quad (23)$$

where $\lambda_1, \dots, \lambda_N$ are the eigenvalues of A .

For completeness, we also note that [40, Cor. 5.3 and Thm. 4.3] shows that for a discrete memoryless channel, it suffices that $C > \sum_{|\lambda_i| \geq 1} \log |\lambda_i|$ for the existence of encoder and controller policies leading to almost sure stability.

Remark 5.3: An implication on the achievability for noncausal codes over discrete memoryless channels: In Section III, we utilized the fact that one can view a stochastic dynamical system as a deterministic one using the shift operator. Building on a similar argument as that in the proof of Theorem 3.8, it can be shown that (E2) considered in this article implies and is implied by (E2) considered in [20] with the distortion metric d being the product metric D introduced in (3) for the dynamical system θ . Therefore, provided that the encoder has access to future realizations of the state sequence, the proof of Theorem 5.1 implies an achievability result: If the encoder has non-causal access to the source realizations, for (E2), we have $C_0 \leq h_{\text{top}}(\theta_{|\text{supp}\mu})$, and this can be achieved through the construction in the proof of Theorem 5.1 using an encoder which has non-causal access to the future state realizations. Note though that the decoder is still restricted to be zero-delay. We note that in traditional Shannon theory, block codes are allowed to be noncausal.

VI. EXAMPLES

Example 6.1: Consider the diffeomorphism $f_A : \mathbb{T}^2 \rightarrow \mathbb{T}^2$ on the 2-torus $\mathbb{T}^2 = \mathbb{R}^2/\mathbb{Z}^2$, induced by the linear map

$$A = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \quad (24)$$

i.e., $f_A(x + \mathbb{Z}^2) = Ax + \mathbb{Z}^2$. Note that the inverse of f_A is given by $f_{A^{-1}}$, which is well defined, since $\det A = 1$. The map f_A is known as *Arnold's Cat Map*, and is one of the simplest examples of an Anosov diffeomorphism.

Since $\det Df_A(x) \equiv \det A \equiv 1$, the map f_A is area-preserving. The eigenvalues of the matrix A are given by

$$\gamma_1 = \frac{3}{2} - \frac{1}{2}\sqrt{5} \quad \text{and} \quad \gamma_2 = \frac{3}{2} + \frac{1}{2}\sqrt{5}$$

and satisfy $|\gamma_2| > 1 > |\gamma_1|$. It is well known that both the topological entropy and the metric entropy of f_A with respect to Lebesgue measure are given by $\log |\gamma_2| > 0$. Hence, Theorem 5.1 yields

$$C_0 \leq \log \left| \frac{3}{2} + \frac{1}{2}\sqrt{5} \right| \approx 1.3885$$

for (E2) to be achieved over a DMC.

Now, suppose we have additive noise for the Cat Map so that $x_{t+1} = f(x_t, w_t)$ with $f(x, w) = Ax + w \pmod{\mathbb{Z}^2}$, with $w \sim \nu$ which admits a density supported on \mathbb{T}^2 . In this case, the map $f^x : \mathbb{T}^2 \rightarrow \mathbb{T}^2$, $w \mapsto Ax + w$, is invertible and $(x, y) \mapsto (f^x)^{-1}(y) = y - Ax$ is continuous. By Corollary 3.5, $C_0 = \infty$ for the estimation objective (E1), under a stationary initial

measure. For the objective (E2), it can be shown that, under corresponding initial measure conditions, Theorem 3.7 leads to $C_0 = \infty$.

In the following, we consider a system without noise for which an explicit estimate for the metric entropy is available.

Example 6.2: We consider the map

$$f(x, y) = (5 - 0.3y - x^2, x), \quad f : \mathbb{R}^2 \rightarrow \mathbb{R}^2.$$

There exists a natural (physical) invariant measure π_0 on the nonwandering set of f , i.e., the set of all points (x, y) so that for every neighborhood U of (x, y) , there is $n \geq 1$ with $f^n(U) \cap U \neq \emptyset$. A numerical approximation of the metric entropy $h_{\pi_0}(f)$ is given in [8, Ex. 6.4], namely

$$h_{\pi_0}(f) \approx 0.655.$$

According to Theorems 3.7 and 3.8, $h_{\pi_0}(f) \leq C_0$ for the estimation objectives (E2) and (E3).

VII. DISCUSSION AND CONCLUDING REMARKS

In this article, we considered three estimation objectives for stochastic nonlinear systems $x_{t+1} = f(x_t, w_t)$ with i.i.d. noise (w_t) , assuming that the estimator receives state information via a noisy channel of finite capacity.

- 1) For noiseless channels, assuming that the initial measure π_0 is stationary, we proved that C_0 is bounded below by either the topological or the metric entropy of a shift dynamical system on the space of trajectories (Theorems 3.3, 3.7, and 3.8).
- 2) For systems on Euclidean space and noisy channels, we provided information-theoretic and probability-theoretic conditions enforcing $C_0 = \infty$. In particular, Theorem 4.2 shows that $C_0 = \infty$ for the quadratic stability objective, whenever

$$\limsup_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^{T-1} h(x_t | x_{t-1}) > -\infty. \quad (25)$$

We have a corresponding negative result under (E2) in Theorem 4.3 for noisy systems which are sufficiently *irreducible*. Since $h(x_t | x_{t-1})$ is a measure for the uncertainty of x_t given x_{t-1} , condition (25) means that the noise on the long run (in average) influences the state process in a substantial way. Similarly to the results in Section III, this means that the noise makes the space of relevant trajectories too large (or too complicated) to estimate the state with arbitrarily small error over a finite-capacity channel.

- 3) Compared with our earlier work [20], our results reveal that the rate requirements are not robust with respect to the presence of noise: That is, even an arbitrarily small noise level may lead to drastic effects in the rate requirements. However, the metric or topological entropy bounds are always present and our lower bounds reduce to those established in [20]. We also note that the metric entropy definition for random dynamical systems [22] in the ergodic theory literature is not the answer to the operational questions we proposed in this article, unlike the

one for the deterministic case which precisely answered the operational question related to (E2).

- 4) In Section V-A, we assumed that the system is deterministic with a compact state-space, but the channel is noisy. We proved that in this case, C_0 is bounded from above by the topological entropy of the system for the asymptotic almost sure objective (thus, leading to an achievability result). Our result strictly generalizes the previously known results in the literature which have considered only linear systems, to our knowledge.
- 5) Some of the ideas used in this article can also be applied to the study of stochastic stabilization over digital channels for noisy nonlinear systems of the form

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

with u_t being a control variable. The preprint [21] introduces a notion of stabilization entropy that is tailored to derive lower bounds on the necessary channel capacity for the control objective of generating an asymptotically mean stationary state process. This notion of entropy combines the idea of invariance entropy [4], [18] with Katok's characterization of metric entropy used in Section III.

APPENDIX

This Appendix contains the proofs of Lemmas 3.1 and 3.2, used in Section III.

Proof: (of Lemma 3.1) We consider the map

$$G : X \times W^{\mathbb{Z}_+} \rightarrow X^{\mathbb{Z}_+}$$

which maps a pair (x_0, \bar{w}) with $\bar{w} = (w_t)_{t \in \mathbb{Z}_+}$ to the trajectory $(x_t)_{t \in \mathbb{Z}_+}$ obtained by $x_{t+1} := f(x_t, w_t)$. We claim that this map is measurable and its associated push-forward operator on measures maps $\pi \times \nu^{\mathbb{Z}_+}$ to Θ . To prove that G is measurable, consider a cylinder set $A = B_0 \times \dots \times B_n \times X^{[n+1, \infty)}$ in $X^{\mathbb{Z}_+}$. Then

$$\begin{aligned} G^{-1}(A) &= \{(x_0, \bar{w}) : x_0 \in B_0, G(x_0, \bar{w})_1 \in B_1, \dots, G(x_0, \bar{w})_n \in B_n\}. \end{aligned}$$

Hence, $G^{-1}(A)$ can be expressed as the preimage of $B_0 \times \dots \times B_n \subset X^{n+1}$ under the map

$$\begin{aligned} (x_0, \bar{w}) &\mapsto (x_0, f_{w_0}(x_0), f_{w_1} \circ f_{w_0}(x_0), \dots, f_{w_{n-1}} \\ &\quad \circ \dots \circ f_{w_0}(x_0)). \end{aligned}$$

To show that this map is measurable, it suffices to show that each component is a measurable map. This follows from the fact that the projection $W^{\mathbb{Z}_+} \rightarrow W^{n+1}$ to the first $n+1$ components is measurable and f is measurable. Hence, we have proved that G is measurable. To see that $G_* (\pi \times \nu^{\mathbb{Z}_+}) = \Theta$, observe that for

a set of the form $A = B_0 \times B_1 \times X^{[2, \infty)}$, we have

$$\begin{aligned} &\pi \times \nu^{\mathbb{Z}_+} (\{(x_0, \bar{w}) : x_0 \in B_0, f_{w_0}(x_0) \in B_1\}) \\ &= \int_X \int_{W^{\mathbb{Z}_+}} \nu^{\mathbb{Z}_+} (d\bar{w}) \pi(dx_0) \mathbb{1}_{B_0}(x_0) \mathbb{1}_{B_1}(f_{w_0}(x_0)) \\ &= \int_{B_0} \pi(dx_0) \int_W \nu(dw) \mathbb{1}_{B_1}(f_w(x_0)) \\ &= \int_{B_0} \pi(dx_0) \nu(\{w \in W : f_w(x_0) \in B_1\}) \\ &= \Theta(B_0 \times B_1 \times X^{[2, \infty)}). \end{aligned} \tag{26}$$

For more general cylinder sets, the claim follows inductively. The fact that $\text{supp } \Theta$ is contained in $\text{cl } \mathcal{T}$ follows from:

$$\begin{aligned} \Theta(\text{cl } \mathcal{T}) &= G_* [\pi \times \nu^{\mathbb{Z}_+}] (\text{cl } \mathcal{T}) = \pi \times \nu^{\mathbb{Z}_+} (G^{-1}(\text{cl } \mathcal{T})) \\ &\geq \pi \times \nu^{\mathbb{Z}_+} (G^{-1}(\mathcal{T})) \\ &= \pi \times \nu^{\mathbb{Z}_+} (G^{-1}(G(X \times W^{\mathbb{Z}_+}))) \\ &\quad \pi \times \nu^{\mathbb{Z}_+} (X \times W^{\mathbb{Z}_+}) = 1. \end{aligned}$$

Finally, we show that Θ is θ -invariant. To this end, note that the map $\Phi : X \times W^{\mathbb{Z}_+} \rightarrow X \times W^{\mathbb{Z}_+}$, $(x, \bar{w}) \mapsto (f(x, w_0), \vartheta \bar{w})$, satisfies $\theta \circ G = G \circ \Phi$. Using that

$$\begin{aligned} &\pi \times \nu^{\mathbb{Z}_+} (\Phi^{-1}(A \times B)) \\ &= \pi \times \nu^{\mathbb{Z}_+} (\{(x_0, \bar{w}) : f_{w_0}(x_0) \in A, \vartheta \bar{w} \in B\}) \\ &= \pi \times \nu^{\mathbb{Z}_+} \left(\bigcup_{x_0 \in X} \{x_0\} \times ((f^{x_0})^{-1}(A) \times B) \right) \\ &= \int_X \pi(dx_0) \nu(\{w : f(x_0, w) \in A\}) \nu^{\mathbb{Z}_+}(B) \\ &= \nu^{\mathbb{Z}_+}(B) \int_X \pi(dx) P(x, A) \\ &= \pi(A) \nu^{\mathbb{Z}_+}(B) = \pi \times \nu^{\mathbb{Z}_+}(A \times B) \end{aligned}$$

i.e., $\Phi_* (\pi \times \nu^{\mathbb{Z}_+}) = \pi \times \nu^{\mathbb{Z}_+}$, we find that

$$\begin{aligned} \theta_* \Theta &= \theta_* G_* (\pi \times \nu^{\mathbb{Z}_+}) = G_* \Phi_* (\pi \times \nu^{\mathbb{Z}_+}) \\ &= G_* (\pi \times \nu^{\mathbb{Z}_+}) = \Theta \end{aligned}$$

completing the proof. \blacksquare

Proof: (of Lemma 3.2) Any $(n, \varepsilon; n_0(\varepsilon))$ -separated set is trivially (n, ε) -separated, hence $r_{\text{sep}}(n, \varepsilon) \geq r_{\text{sep}}(n, \varepsilon; n_0(\varepsilon))$, implying the inequality “ \geq ” in (4). Conversely, assume that E is (n, ε) -separated and put $E' := f^{-n_0(\varepsilon)}(E)$. Then, $|E'| = |E|$ and E' is $(n_0(\varepsilon) + n, \varepsilon; n_0(\varepsilon))$ -separated. This implies

$$\begin{aligned} &\frac{n + n_0(\varepsilon)}{n} \frac{1}{n + n_0(\varepsilon)} \log r_{\text{sep}}(n_0(\varepsilon) + n, \varepsilon; n_0(\varepsilon)) \\ &\geq \frac{1}{n} \log r_{\text{sep}}(n, \varepsilon). \end{aligned}$$

Letting $n \rightarrow \infty$ on both sides, we find that

$$\limsup_{n_0(\varepsilon) < n \rightarrow \infty} \frac{1}{n} \log r_{\text{sep}}(n, \varepsilon; n_0(\varepsilon)) \geq h_{\text{sep}}(f, \varepsilon).$$

Finally, letting $\varepsilon \downarrow 0$, the desired inequality follows. ■

REFERENCES

- [1] H. Asnani and T. Weissman, "Real-time coding with limited lookahead," *IEEE Trans. Inform. Theory*, vol. 59, no. 6, pp. 3582–3606, Jun. 2013.
- [2] V. A. Boichenko, G. A. Leonov, and V. Reitmann, *Dimension Theory for Ordinary Differential Equations*. Stuttgart, Germany: Teubner, 2005.
- [3] V. S. Borkar, S. K. Mitter, and S. Tatikonda, "Optimal sequential vector quantization of Markov sources," *SIAM J. Control Optim.*, vol. 40, no. 1, pp. 135–148, 2001.
- [4] F. Colonius and C. Kawan, "Invariance entropy for control systems," *SIAM J. Control Optim.*, vol. 48, no. 3, pp. 1701–1721, 2009.
- [5] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley, 1991.
- [6] R. Dabora and A. Goldsmith, "On the capacity of indecomposable finite-state channels with feedback," in *Proc. Allerton Conf. Commun. Control Comput.*, Sep. 2008, pp. 1045–1052.
- [7] T. Downarowicz, *Entropy in Dynamical Systems*. Cambridge, U.K.: Cambridge University Press, 2011.
- [8] G. Froyland, "Using Ulam's method to calculate entropy and other dynamical invariants," *Nonlinearity*, vol. 12, no. 1, pp. 79–101, 1999.
- [9] R. G. Gallager, "A simple derivation of the coding theorem and some applications," *IEEE Trans. Inform. Theory*, vol. 11, no. 1, pp. 3–18, Jan. 1965.
- [10] H. O. Georgii, "Probabilistic aspects of entropy," in *Entropy*, A. Greven, G. Keller, and G. Warnecke, Eds., Princeton, NJ, USA: Princeton University Press, 2003, pp. 37–54.
- [11] R. M. Gray, *Entropy and Information Theory*. New York, NY, USA: Springer, 2011.
- [12] R. M. Gray, *Probability, Random Processes, and Ergodic Properties*. 2nd ed. Dordrecht, Springer, 2009.
- [13] R. M. Gray and D. L. Neuhoff, "Quantization: 1948–1998," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2325–2383, Oct. 1998.
- [14] T. Javidi and A. Goldsmith, "Dynamic joint source-channel coding with feedback," in *Proc. IEEE Int. Symp. Inf. Theory*, IEEE, 2013, pp. 16–20.
- [15] Y. Kaspi and N. Merhav, "Structure theorems for real-time variable-rate coding with and without side information," *IEEE Trans. Inform. Theory*, vol. 58, no. 12, pp. 7135–7153, Dec. 2012.
- [16] A. Katok, "Lyapunov exponents, entropy and periodic orbits for diffeomorphisms," *Inst. Hautes Études Sci. Publ. Math.*, vol. 51, pp. 137–173, 1980.
- [17] A. Katok, "Fifty years of entropy in dynamics: 1958–2007," *J. Modern Dyn.*, vol. 1, no. 4, pp. 545–596, 2007.
- [18] C. Kawan, "Invariance entropy for deterministic control systems. An introduction," in *Lecture Notes Math.*, vol. 2089. Berlin Germany: Springer, 2013.
- [19] C. Kawan, "Exponential state estimation, entropy and Lyapunov exponents," *Syst. Control Lett.*, vol. 113, pp. 78–85, 2018.
- [20] C. Kawan and S. Yüksel, "On optimal coding of non-linear dynamical systems," *IEEE Trans. Inf. Theory*, vol. 64, no. 10, pp. 6816–6829, Oct. 2018.
- [21] C. Kawan and S. Yüksel, "Invariance properties of nonlinear stochastic dynamical systems under information constraints," submitted for publication, *arXiv: 1901.02825*.
- [22] F. Ledrappier and L.-S. Young, "Entropy formula for random transformations," *Probability Theory Related Fields*, vol. 80, pp. 217–240, 1988.
- [23] D. Liberzon and S. Mitra, "Entropy and minimal data rates for state estimation and model detection," in *Proc. 19th Int. Conf. Hybrid Syst.: Comput. Control*, ACM, 2016, pp. 247–256.
- [24] D. Liberzon and S. Mitra, "Entropy and minimal bit rates for state estimation and model detection," *IEEE Trans. Autom. Control*, vol. 63, no. 10, pp. 3330–3344, Oct. 2018.
- [25] T. Linder and S. Yüksel, "On optimal zero-delay quantization of vector Markov sources," *IEEE Trans. Inform. Theory*, vol. 60, no. 10, pp. 5975–5991, Oct. 2014.
- [26] T. Linder and R. Zamir, "Causal coding of stationary sources and individual sequences with high resolution," *IEEE Trans. Inform. Theory*, vol. 52, no. 2, pp. 662–680, Feb. 2006.
- [27] A. Mahajan and D. Teneketzis, "Optimal design of sequential real-time communication systems," *IEEE Trans. Inform. Theory*, vol. 55, no. 11, pp. 5317–5338, Nov. 2009.
- [28] A. S. Matveev, "State estimation via limited capacity noisy communication channels," *Math. Control Signals Syst.*, vol. 20, no. 1, pp. 1–35, 2008.
- [29] A. Matveev and A. Pogromsky, "Observation of nonlinear systems via finite capacity channels: Constructive data rate limits," *Automatica*, vol. 70, pp. 217–229, 2016.
- [30] A. Matveev and A. Pogromsky, "Observation of nonlinear systems via finite capacity channels. Part II: Restoration entropy and its estimates," *Automatica*, vol. 70, pp. 217–229, 2016.
- [31] A. S. Matveev and A. V. Savkin, *Estimation and Control Over Communication Networks*. Control Engineering. Boston, MA, USA: Birkhäuser Boston, Inc., 2009.
- [32] M. Misiurewicz, *Topological entropy and metric entropy*. Ergodic theory (Sem., Les Plans-sur-Bex, 1980) (in French), pp. 61–66, Monograph. Enseign. Math., vol. 29, Univ. Genève, Geneva, 1981.
- [33] G. N. Nair and R. J. Evans, "Stabilizability of stochastic linear systems with finite feedback data rates," *SIAM J. Control Optim.*, vol. 43, pp. 413–436, 2004.
- [34] A. Nayyar and D. Teneketzis, "On the structure of real-time encoders and decoders in a multi-terminal communication system," *IEEE Trans. Inf. Theory*, vol. 57, no. 9, pp. 6196–6214, Sep. 2011.
- [35] D. L. Neuhoff and R. K. Gilbert, "Causal source codes," *IEEE Trans. Inf. Theory*, vol. 28, no. 5, pp. 701–713, Sep. 1982.
- [36] D. Ornstein, "Bernoulli shifts with the same entropy are isomorphic," *Advances Math.*, vol. 4, pp. 337–352, 1970.
- [37] D. Ornstein, "An application of ergodic theory to probability theory," *Ann. Probability*, vol. 1, no. 1, pp. 43–65, 1973.
- [38] H. H. Permuter, T. Weissman, and A. J. Goldsmith, "Finite state channels with time-invariant deterministic feedback," *IEEE Trans. Inform. Theory*, vol. 55, pp. 644–662, Feb. 2009.
- [39] A. Pogromsky and A. Matveev, "A topological entropy approach for observation via channels with limited data rate," *IFAC Proc.*, vol. 44, no. 1, pp. 14416–14421, 2011.
- [40] 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," *IEEE Trans. Inf. Theory*, vol. 52, no. 8, pp. 3369–3395, Aug. 2006.
- [41] A. V. Savkin, "Analysis and synthesis of networked control systems: Topological entropy, observability, robustness and optimal control," *Automatica*, vol. 42, no. 1, pp. 51–62, 2006.
- [42] N. Şen, F. Alajaji, and S. Yüksel, "Feedback capacity of a class of symmetric finite-state Markov channels," *IEEE Trans. Inf. Theory*, vol. 57, no. 7, pp. 4110–4122, Jul. 2011.
- [43] P. C. Shields, "The interactions between ergodic theory and information theory," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2079–2093, Oct. 1998.
- [44] S. Tatikonda, "Control under communication constraints," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, USA, 2000.
- [45] S. Tatikonda and S. Mitter, "The capacity of channels with feedback," *IEEE Trans. Inf. Theory*, vol. 55, no. 1, pp. 323–349, Jan. 2009.
- [46] D. Teneketzis, "On the structure of optimal real-time encoders and decoders in noisy communication," *IEEE Trans. Inform. Theory*, vol. 52, no. 9, pp. 4017–4035, Sep. 2006.
- [47] J. C. Walrand and P. Varaiya, "Optimal causal coding-decoding problems," *IEEE Trans. Inform. Theory*, vol. 29, no. 6, pp. 814–820, Nov. 1983.
- [48] H. S. Witsenhausen, "On the structure of real-time source coders," *Bell Syst. Tech. J.*, vol. 58, pp. 1437–1451, Jul./Aug. 1979.
- [49] W. S. Wong and R. W. Brockett, "Systems with finite communication bandwidth constraints - Part II: Stabilization with limited information feedback," *IEEE Trans. Autom. Control*, vol. 44, no. 5, pp. 1294–1299, May 1999.
- [50] R. G. Wood, T. Linder, and S. Yüksel, "Optimal zero delay coding of Markov sources: Stationary and finite memory codes," *IEEE Trans. Inform. Theory*, vol. 63, no. 9, pp. 5968–5980, Sep. 2017.
- [51] S. Yüksel and T. Başar, *Stochastic Networked Control Systems: Stabilization and Optimization Under Information Constraints*. New York, NY, USA: Birkhäuser, 2013.
- [52] S. Yüksel, "On optimal causal coding of partially observed Markov sources in single and multi-terminal settings," *IEEE Trans. Inf. Theory*, vol. 59, no. 1, pp. 424–437, Jan. 2013.
- [53] S. Yüksel, "Stationary and ergodic properties of stochastic nonlinear systems controlled over communication channels," *SIAM J. Control Optim.*, vol. 54, no. 5, pp. 2844–2871, 2016.



Christoph Kawan received the diploma and doctoral degrees in mathematics (both under supervision of Prof. Fritz Colonius) from the University of Augsburg, Augsburg, Germany, in 2006 and 2009, respectively.

He was a Research Scholar for four months with the State University of Campinas, Brazil, in 2011, and nine months with the Courant Institute of Mathematical Sciences, New York University, NY, USA, in 2014. He currently works as a Researcher with the Ludwig-Maximilians-

Universität, Munich, Germany. He is the author of the book "Invariance Entropy for Deterministic Control Systems - An Introduction." His research interests include networked and information-based control and nonautonomous dynamical systems.



Serdar Yüksel (S'02–M'11) received the B.Sc. degree in electrical and electronics engineering from the Bilkent University, Ankara, Turkey, in 2001, the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Illinois at Urbana-Champaign, Urbana-Champaign, IL, USA, in 2003 and 2006, respectively.

He was a Postdoctoral Researcher with the Yale University before joining Queen's University, Kingston, ON, Canada, as an Assistant Professor in the Department of Mathematics and Statistics, where he is now an Associate Professor. His research interests include stochastic control, decentralized control, information theory, and probability.

Dr. Yüksel has received the 2013 CAIMS/PIMS Early Career Award in Applied Mathematics. He has been an Associate Editor for the IEEE TRANSACTIONS ON AUTOMATIC CONTROL, *Automatica*, and *Systems and Control Letters*.