Proc. 54th IEEE Conf. Decision and Control, Osaka, Japan, Dec. 2015, pp.45 -56.

Nonstochastic Information Concepts for Estimation and Control

Girish N. Nair

Abstract— **Entropy and information are crucial notions in stochastic communication systems. However, they have arguably not been as central in control theory, which has a rich tradition of non-random models and techniques. This tutorial session aims to describe the key elements of certain non-probabilistic entropy and information concepts for state estimation and control. In this paper, which comprises the first half of the session, the focus is on a recently developed theory of nonstochastic information. Motivated by worst-case estimation and control, this framework allows non-statistical analogues of mutual independence, Markovness, information, and directed information to be rigorously defined. This yields powerful information-theoretic tools for finding fundamental bounds in zero-error communication and worst-case control systems.**

In the second half of this session, notions of entropy for deterministic nonlinear control systems are described, based on dynamical systems theory. These notions lead to characterisations of minimal feedback data rates for set-invariance. Taken together, the concepts discussed in this session give deterministic control theorists a way to use information and entropy ideas, without having to adopt a stochastic formulation.

I. INTRODUCTION

Information is a difficult concept to generally define without vagueness. However, in 1948 the electrical engineer Claude Shannon [1] gave a mathematically precise definition for random variables (rv's) in a probability space. Shannon was interested in the problem of reliable communication over a noisy channel. Assuming that the data and noise sources in the system could be modelled as rv's, he proposed statistical indices of *entropy*, i.e. the *a priori* uncertainty in an rv, and the *mutual information* shared by two correlated rv's, in units of bits.

These indices exhibit properties that are natural to expect from any reasonable measure of uncertainty and information. However, it is the *operational* relevance of Shannon's concepts that has given them a huge impact in communications. If the goal in a communication system is to make the probability of a decoding error negligible, then the entropy rate of a stationary stochastic process coincides with the lowest possible compressed bit-rate, while the the maximum mutual information rate across a channel, in bits per sample, coincides with the highest block-coding rate. Prior to Shannon's work, the belief among engineers was that small decoding error probabilities could be achieved only with large signal-to-noise (SNR) ratios in the physical channel. By thinking in terms of bits and codes, Shannon showed that arbitrarily small error probabilities could be achieved, even at low SNR.

Information theory is now the basis of modern digital communications, and has also percolated into many other areas, including computer science [2], system identification [3], [4] and medicine [5]. In the context of control and state estimation, minimum entropy and maximum information have been proposed as design criteria for control and estimation [4], [6], [7]. In addition, information-theoretic methods have been exploited, under various models for the noise in the channel and plant, to find minimum communication rates for estimating or stabilising the states of a linear-time invariant (LTI) plant, [8], [9], [10] and to study fundamental limitations for disturbance rejection in control [11], [12], [13], [14]. Inspired by [1], in 1958 Kolmogorov and Sinai formulated an analogous, measure-theoretic definition of entropy rate for an open-loop dynamical system evolving on an invariant measure space. With some surprise, it was discovered that systems with purely deterministic dynamics could have nonzero entropy rates, similar to stationary random processes [15].

A. *Information in Control*

Despite these advances, information theory has not played as central a role in control as it has in communications. One reason for this is its probabilistic model of uncertainty. In communications, statistical models make sense for several good reasons. Firstly, most communication systems consist of some mix of electronic, electromagnetic and photonic devices, and the dominant disturbances in these domains - e.g. thermal and shot noise, fading - arise from physical laws that yield well-defined distributions. Furthermore, in everyday communications, each telephone call and data byte may not be crucially important or expensive, and so performance usually needs to be guaranteed only on average, over many uses of the system.

In contrast, automatic control is frequently used in safety- or mission-critical applications, where performance, safety or regulatory bounds must be guaranteed *every* time an often expensive plant is used, not just on average; e.g. in aircraft, automobiles and critical infrastructure. Moreover, as control systems frequently contain mechanical or chemical components, the dominant disturbances do not always arise from electronic or photonic circuit noise, and so may not obey probability distributions based on physics. Even if some circuit noise is present, control systems typically operate with bit- or sample periods T several orders of magnitude longer than modern communication systems. At these slower

This work was supported by the Australian Research Council under the Future Fellowships scheme (FT140100527).

Dept. Electrical & Electronic Engineering, University of Melbourne, VIC 3010, Australia, gnair@unimelb.edu.au

scales, the variance of thermal or shot noise ($\propto T$) after filtering is often negligible compared to the corresponding signal power ($\propto T^2$). For these reasons, control theory often treats uncertainties and disturbances as unknown variables and signals without any probabilistic structure, and measures performance in a *worst-case sense* over all admissible disturbances with specified bounds on magnitude, power or energy.

Another factor that has impeded the application of standard information theory to networked control is the presence of feedback. Feedback is the core idea in control theory, and its vital role in maintaining performance and stability in the face of disturbances has been well-understood for over a century. However, although feedback is a feature of many practical communication protocols, it has not been a major focus in information theory. This is partly because for *stationary memoryless channels*, the ordinary capacity C is completely insensitive to the availability of feedback from the receiver back to the transmitter [16], [17]. More importantly, the correct formulation of a stochastic channel with feedback was lacking until Massey's 1990 paper [17], in which he remarked *'it is hardly a wonder that information theory has had problems dealing with feedback'*.

B. Aim

Networked control combines both the disciplines of communications and control, and its rapid emergence in recent years raises the important question of how to define operationally meaningful analogues of concepts like independence, Markovnessand information for systems with worst-case objectives. These concepts are powerful aids in stochastic settings, and it would be useful to be able to apply them in some form to nonstochastic problems, without having to impose a probability space.

This tutorial paper provides an overview of a recent framework [18], [19] that gives nonstochastic analogues for mutual information and underlying concepts in terms of the ranges of the variables in question, rather than their probability distributions. Knowledge of classical information theory is not necessary to understand this construction. It turns out that many of the basic bricks in it already exist in various guises, but in discrete, stochastic settings – e.g. the *qualitative independence* of [20], the *common information elements* of [21], and the *ergodic decomposition* and *connected components* of [22], [23]. Importantly, these notions are essentially independent of the assumed probability measures, apart from their supports. By excising all their stochastic elements, cementing them together, and adapting them for continuous-valued variables, a useful information-theoretic framework is obtained for analysing noisy systems under worst-case or zero-error objectives. While other definitions of information without probability exist and also possess natural properties [24], [25], their operational relevance for finding performance limitations in such systems is unclear.

C. Structure

In the next section, the *uncertain variable* framework is described, followed by a description of nonstochastic information in sec. III. These ideas then lead to definitions of nonstochastic conditional and directed information in sec. IV. In secs. V and VI, the uses of nonstochastic information and directed information in analysing zero-error communications and state estimation or control via noisy channels are discussed. The paper then ends with a discussion of several open problems. Formal statements of results are avoided, but where possible sketches of proofs are provided.

Throughout this paper, set cardinality is denoted by $|\cdot|$, with the value ∞ permitted, and all logarithms are to the base 2. With a small abuse of terminology, the cardinality of the range of a variable X will just be called the cardinality of the variable X.

II. Uncertain Variables, Unrelatedness, and Markov Chains

In this section, the *uncertain variable (uv)* framework of [18] is described, and seen to yield nonstochastic analogues of probabilistic concepts such as independence, conditional independence, and Markovness. These analogues are weaker than their stochastic counterparts, since they depend only on the ranges of the variables in question. On the other hand, this suits situations where statistical models of uncertainty are unjustified or unavailable.

A. Uncertain Variables

The key idea in the uv framework is to keep the probability convention of regarding an unknown variable of interest as a mapping X from some underlying *sample space* Ω to a set $\mathbb{X}$ of interest. When an experiment is run, a specific *sample* $\omega \in \Omega$ is selected, yielding a realisation $X(\omega)$, denoted by lower-case $x \in \mathbb{X}$. For instance, in a dynamic system each possible $\omega \in \Omega$ may be identified with a particular combination of initial states and exogenous noise signals, and $x = X(\omega)$ may be the realised state or output. The sample ω itself may not be observed. As in probability theory, the dependence on ω will usually be suppressed for conciseness, so that a statement such as $X \in \mathbf{K}$ may be taken to mean $X(\omega) \in \mathbf{K}$, unless stated otherwise. Unlike in probability theory, no measure is imposed on Ω, nor is it necessary to assume some σ-algebra $\subset 2^{\Omega}$ of valid ω-sets.[1]

The mapping $X : \Omega \to \mathbb{X}$ is called an *uncertain variable (uv)*. Similar to the case of a random variable (rv), this mapping may not be known. However, its range is known, and some realisation $x = X(\omega)$ is seen. Given another uv Y taking values in $\mathbb{Y}$, write

$$\begin{aligned}
[\![X]\!] &:= \{X(\omega) : \omega \in \Omega\}, && (1)\\
[\![X,Y]\!] &:= \{(X(\omega),Y(\omega)) : \omega \in \Omega\} \subseteq [\![X]\!] \times [\![Y]\!], && (2)\\
[\![Y|x]\!] &:= \{Y(\omega) : X(\omega) = x, \omega \in \Omega\}. && (3)
\end{aligned}$$

Call $[\![X]\!]$ the *marginal range* of X, $[\![X,Y]\!]$, the *joint range* of X and Y, and $[\![Y|x]\!]$ the *conditional range* of Y given (or *range conditional on*) $X = x$, In the absence of a joint distribution,

[1] In a probability space with uncountable Ω, the restriction to a σ-algebra strictly smaller than the power set is required in order to avoid paradoxes when a σ-additive measure is imposed - see [26].

the joint range fully characterises the relationship between X and Y. In particular, (1) and (3) can each be produced from the two-dimensional set (2), by projection onto the horizontal x-axis, or intersection with a vertical line at x followed by projection onto the vertical y-axis, respectively.

Conversely, the marginal range (1) and conditional ranges (3) together fully determine the joint range (2), i.e.

$$[\![X,Y]\!] = \bigcup_{x\in[\![X]\!]} [\![Y|x]\!] \times \{x\}. \tag{4}$$

This is similar to the way that joint probability distributions are determined by the conditional and marginal ones.

B. Unrelatedness

The next step is to define a nonstochastic analogue of statistical independence. Call two uv's X and Y *mutually unrelated* if their joint range coincides with the Cartesian product of the marginal ones, i.e.

$$[\![X,Y]\!] = [\![X]\!] \times [\![Y]\!], \tag{5}$$

denoted $X \perp Y$; otherwise, call them *mutually related*. This directly parallels mutual independence in probability theory, with distributions replaced by ranges, and multiplication, by a Cartesian product. However, it is an essentially weaker notion. This is because the joint distribution of independent rv's always has support in the form of a Cartesian product, but conversely, a joint support that takes Cartesian product form does not entail a joint probability density or mass function in product form.[2]

In some situations, it is more convenient to think in terms of conditional ranges. It can be shown that unrelatedness[3] as defined in (5) is equivalent to the property that

$$[\![X|y]\!] = [\![X]\!], \quad \forall y \in [\![Y]\!]. \tag{6}$$

In other words, a realisation of one uv does not affect the range of values that the other can take. Note that when Ω is discrete, unrelatedness is equivalent to the mappings X and Y inducing *qualitatively independent* [20] partitions of Ω.

The discussion above easily extends to handle more than two uv's. A finite collection $X_1, X_2, \ldots, X_n$ of uv's is said to be unrelated if

$$[\![X_1,\ldots,X_n]\!] = [\![X_1]\!] \times \cdots \times [\![X_n]\!], \tag{7}$$

or equivalently if

$$[\![X_i|X_{1:i-1}]\!] = [\![X_i]\!], \quad \forall i \in [2:n]. \tag{8}$$

An infinite collection $X_1, X_2, \ldots$ of uv's is called unrelated if every finite subcollection $X_{i_1}, \ldots, X_{i_n}$ is mutually unrelated.

C. Markovness without Probability

In probability theory, a sequence of three or more rv's is said to be Markovian if the conditional distribution of any rv given all previous ones depends only on the previous rv. This notion can also be extended quite directly to uv's, by replacing distributions with ranges.

First consider a sequence X, Y, Z of three uv's. This is said to form a *Markov uncertainty chain*, denoted $X \leftrightarrow Y \leftrightarrow Z$, if

$$[\![X|y,z]\!] = [\![X|y]\!], \ \forall (y,z) \in [\![Y,Z]\!], \tag{9}$$

i.e. the conditional range of X given a realisation (y,z) depends only y. It can be shown that this is exactly equivalent to the condition

$$[\![X,Z|y]\!] = [\![X|y]\!] \times [\![Z|y]\!], \quad \forall y \in [\![Y]\!], \tag{10}$$

i.e. X, Z are *conditionally unrelated given* Y, denoted $X \perp Z|Y$. In this form it is clear that the Markovness is preserved if the sequence is reversed by swapping X and Z; hence the double-headed arrows.

Now consider a general sequence $X_1, \ldots, X_n$ of $n \geq 3$ uv's. This sequence is said to form a Markov uncertainty chain $X_1 \leftrightarrow X_2 \leftrightarrow \cdots \leftrightarrow X_n$ if

$$X_{1:k-1} \perp X_{k+1}|X_k, \quad \forall k \in [2:n-1]. \tag{11}$$

i.e. the conditional range of X_{k+1} given past realisations depends only on the most recent one. It can be shown that this is equivalent to the reversal-invariant property

$$X_{1:k-1} \perp X_{k+1:n}|X_k, \quad \forall k \in [2:n-1],$$

i.e. the future and past of the sequence are conditionally unrelated given the present. The definition (11) can easily be extended to semi-infinite sequences, i.e. with $n \to \infty$.

As a final definition, a sequence $X_1, X_2 \ldots$ of uv's is said to be a *conditional Markov uncertainty chain given* W if

$$X_{1:k-1} \perp X_{k+1}|(X_k, W), \quad k = 2, 3, \ldots.$$

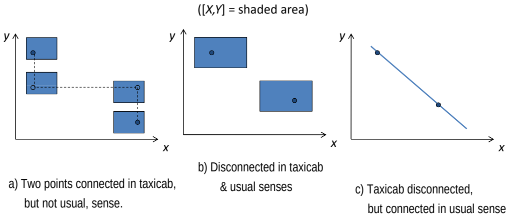
III. NONSTOCHASTIC INFORMATION

It is now shown that the framework described above can be used to measure the amount of information $\mathrm{I}_*[X;Y]$ common to two uncertain variables (uv's) X and Y. This construction allows information-theoretic tools to be used to analyse problems without statistical structure, as discussed later in this paper.

Before commencing, it is worth remarking that Shannon defines mutual information as the reduction in the entropy of one variable after the other variable is observed. In [24], [25], the same philosophy was followed to define some non-probabilistic indices of information, but with entropy replaced by the *Hartley entropy* (log-cardinality) for discrete variables [27], and the *0th order Rényi differential entropy* (log-Lebesgue-measure) for continuous ones [28]. The nonstochastic information proposed here does not follow this approach, but instead directly measures how much both variables share, as described below.

[2] Of course the range of an rv and the support of its distribution are not exactly identical, but coincide up to a zero-probability set.

[3] For conciseness the adjective 'mutual' will often be dropped.



a) Two points connected in taxicab, but not usual, sense.

b) Disconnected in taxicab & usual senses

c) Taxicab disconnected, but connected in usual sense

Fig. 1. Examples of taxicab (dis)connectedness

A. Taxicab Connectedness and Nonstochastic Information

Before nonstochastic information is defined, certain coordinate-geometrical properties of the joint range $\llbracket X,Y \rrbracket$ must be defined. Call two points (x,y) and $(x',y') \in \llbracket X,Y \rrbracket$ *taxicab connected* $(x,y) \leftrightsquigarrow (x',y')$ if $\exists$ a finite sequence $((x_i,y_i))_{i=1}^n$ of points in $\llbracket X,Y \rrbracket$, beginning in $(x_1,y_1) = (x,y)$ and ending in $(x_n,y_n) = (x',y')$, such that $\forall i \in [2:n]$, either $x_i = x_{i-1}$ or $y_i = y_{i-1}$, i.e each point in the sequence shares at least one coordinate with its predecessor.[4]

As $\leftrightsquigarrow$ is evidently transitive and symmetric, it is an equivalence relation on $\llbracket X,Y \rrbracket$, and hence induces *equivalence classes*. Call the collection of equivalence classes induced by $\leftrightsquigarrow$ the *taxicab partition* $\mathscr{T}[X;Y]$ of $\llbracket X,Y \rrbracket$. Nonstochastic information is then defined as

$$\mathrm{I}_*[X;Y] := \log|\mathscr{T}[X;Y]| \in [0,\infty]. \tag{12}$$

It is important to note that $\mathrm{I}_*[X;Y]$ is not generally the same as mutual information with a uniform probability distribution on $\llbracket X,Y \rrbracket$. Nor is it the same as mutual information infimised or supremised over all joint distributions with support on or in $\llbracket X,Y \rrbracket$ (which would in any case yield an infimum $= 0$, and a supremum $= \infty$).

As such, it may not be immediately obvious why (12) should be treated as a measure of information. In summary, it turns out that each label of a taxicab partition-set is a realisation of the maximal *common variable* [23] Z_*, or equivalently the *common information element* [21], for X and Y. In other words $\mathrm{I}_*[X;Y]$, which equals the log-cardinality of Z_*, directly enumerates the maximum number of bits on which X and Y can always agree. In contrast, it has long been known that mutual information does not correspond to an actual variable common to X and Y (see e.g. [22]), even though for pedagogical reasons it is often depicted as an intersection of two sets in a Venn diagram.

The notion of a (maximal) common variable is described next.

B. Maximal Common Variables

Suppose two agents observe realisations of X and Y separately and want to agree on some value, which is to be produced by each agent applying a function to its observation. That is, the first agent applies a function f to the realisation x, and the other agent a function g to y, so that $f(x) = g(y)$ for all possible (x,y) pairs. Then $Z := f(X) = g(Y)$ is a common variable on the value of which they are guaranteed to agree; if they cannot agree on anything unambigously, then Z is a constant.

It turns out that for any pair of variables X,Y, there exists a common variable $Z_* \equiv f_*(X) \equiv g_*(Y)$ that is *maximal* in the sense that any other common variable $Z = f(X) = g(Y)$ can be deduced from Z_*, i.e. there is a mapping $Z_* \mapsto Z$.[5] Thus Z_* has the maximum number - possibly infinite - of distinct values over all common variables.

However, the log-cardinality of the maximal common variable is not very attractive as a definition of information, since it involves a maximisation over auxiliary functions f and g. Fortunately, maximal common variables coincide (up to isomorphism) with the labels of the taxicab partition-sets, which are defined directly in terms of the range $\llbracket X,Y \rrbracket$. This equivalence was proved in [23], for discrete rv's and in terms of the connected components of a discrete bipartite graph rather than taxicab partitions. It is useful to go through the argument in terms of uv's and taxicab sequences.

1) Equivalence Between Taxicab Partition and Maximal Common UV's: Observe that any two points (x,y) and (x',y') in the same taxicab partition set must give the same realisation $f_*(x) \equiv g_*(y) = f_*(x') \equiv g_*(y')$ of the maximal common variable Z_*. To see this, first note that by definition there is a taxicab sequence $((x_i,y_i))_{i=1}^n$ connecting one point with the other, and let $z_* := f_*(x) \equiv g_*(y)$. Going from $(x_1,y_1) := (x,y)$ to (x_2,y_2), it must hold that either $x_2 = x_1$, implying that $g_*(y_2) \equiv f_*(x_2) = f_*(x_1) \equiv z_*$, or that $y_2 = y_1$, meaning that $f_*(x_2) \equiv g_*(y_2) = g_*(y_1) \equiv z_*$. Proceeding by induction until $(x_n,y_n) := (x',y')$, it follows that the common value $f_*(x') \equiv g_*(y')$ is also equal to z_*. In other words, the label of the taxicab partition set determines the value of the maximal common uv.

Now it is shown that the active taxicab partition set containing a realisation of (X,Y) can be determined from a realisation of X or Y alone. Call the taxicab partition sets $\mathbf{T}_i$, with label i running over some set[6], and let $\mathbf{T}_i^{\mathrm{x}}$ denote the projections of $\mathbf{T}_i$ onto the x-axis. Clearly $\{\mathbf{T}_i^{\mathrm{x}}\}_i$ covers $\llbracket X \rrbracket$. Moreover none of the projected sets overlap; if for some $i \neq j$ there is an $x_0 \in \mathbf{T}_i^{\mathrm{x}} \cap \mathbf{T}_j^{\mathrm{x}}$, then there would have to be a point in $\mathbf{T}_i$ and another in $\mathbf{T}_j$ with identical x-coordinate x_0, immediately implying a taxicab sequence from $\mathbf{T}_i$ to $\mathbf{T}_j$, which is impossible. In other words, the projection of the taxicab partition of the joint range $\llbracket X,Y \rrbracket$ onto the x-axis is itself a partition, of the marginal range $\llbracket X \rrbracket$, and in particular $x \in \mathbf{T}_z^{\mathrm{x}}$ iff $(x,y) \in \mathbf{T}_z$. Now define a uv Z by the rule $Z = z$

[4] The name arises because the sequence yields a path in $\llbracket X,Y \rrbracket$ with only vertical or horizontal segments, like a taxi in a grid of streets. This usage comes from metric analysis.

[5] This is because the space of variables forms a *lattice* [21].

[6] If the label set is uncountable, then each label is chosen as a representative point selected from the taxicab partition set.

iff $X \in \mathbf{T}_z^{\mathrm{x}}$. As this occurs iff $(X,Y) \in \mathbf{T}_z$, the label Z of the active taxicab partition-set is fully determined by X.

However, these arguments also hold for projections onto the y-axis, meaning that the label Z of the taxicab partition-set is also fully determined as a function of Y. In other words, Z is common to X and Y. Recall from above that the value of any common uv is fully determined by the label of the taxicab partition set. By definition, this label Z is then a maximal common uv, as desired.

C. *Equivalent View of* I_* *via Overlap Partitions*

Analogous to probability theory, it often easier to work with conditional rather than joint ranges, e.g. when dealing with Markov chains. Let

$$[\![X|Y]\!] := \{[\![X|y]\!] : y \in [\![Y]\!]\} \tag{13}$$

be the family of conditional ranges of X given Y. Note this is an unordered family, with no repeated elements. That is, given a set $\mathbf{B} \in [\![X|Y]\!]$, the number and specific values $y \in [\![Y]\!]$ that yield $[\![X|y]\!] = \mathbf{B}$ cannot be determined. Nonetheless, it turns out that knowing $[\![X|Y]\!]$ is enough to find $\mathrm{I}_*[X;Y]$. This contrasts sharply with mutual information $\mathrm{I}[X;Y]$, which cannot be determined from knowledge of each conditional distribution of rv X given $Y = y$.

As in subsection III-A, a notion of connectedness is first needed. Two points $x, x' \in [\![X]\!]$ are called $[\![X|Y]\!]$*-overlap-connected*, concisely denoted $x \sim x'$, if $\exists$ a finite sequence of successively overlapping sets in $[\![X|Y]\!]$ going from one to the other. In other words, $\exists \mathbf{B}_1, \ldots, \mathbf{B}_n \in [\![X|Y]\!]$ s.t.

- $x \in \mathbf{B}_1$ and $x' \in \mathbf{B}_n$
- $\mathbf{B}_{i-1} \cap \mathbf{B}_i \neq \emptyset$, $\forall i \in [2:n]$.

It is easy to see that $[\![X|Y]\!]$-overlap connectedness is symmetric and transitive, i.e. $\sim$ is an equivalence relation on $[\![X]\!]$, induced by $[\![X|Y]\!]$. The *overlap partition* $[\![X|Y]\!]_*$ of $[\![X]\!]$ is then defined as the family of equivalence classes.

It turns out that for any uv's X, Y,

$$\mathrm{I}_*[X;Y] = \log_2 |[\![X|Y]\!]_*| . \tag{14}$$

The proof of this equation may be found in [18].[7] In summary, the proof shows that for any two points (x,y) and $(x',y') \in [\![X,Y]\!]$,

$$x \sim x' \quad \Leftrightarrow \quad (x,y) \leftrightsquigarrow (x',y'). \tag{15}$$

This allows a bijection to be set up between the sets of the taxicab partition $\mathscr{T}[X;Y]$ and the sets of the overlap partition $[\![X|Y]\!]_*$, which then implies that they have equal cardinality. In terms of the maximal common uv $Z_* \equiv f_*(X) \equiv g_*(Y)$, the partition $[\![X|Y]\!]_*$ gives the level sets of f_*, while $[\![Y|X]\!]_*$ gives the level sets of g_*.

D. *History*

Some historical comments are appropriate to close this section. As mentioned above and in the Introduction, many of the notions discussed in this section have existed in the literature under different names, in the context of random variables. The notion of a discrete, maximal common rv [23] was first introduced in [21], where it was called a *common information element*. When constrained to discrete variables, the taxicab partition is the same as the *ergodic decomposition* of [22] and the *connected components* of [23]; in the last-mentioned paper the equivalence to maximal common rv's is proved. Furthermore, for rv's the sets of the overlap partition correspond to the *communicating classes* of [22]; however this earlier notion is specified not in terms of conditional support sets, but rather in terms of taxicab connectedness, i.e. by using (15) as a definition. For possibly continuous-valued rv's, there is also the related notion of *common knowledge* [29] from economics, which yields the smallest measurable set, in an intersected σ-algebra on Ω, that contains the (not generally measurable) pre-images of a taxicab partition-set.

In the first three of the articles mentioned above, joint probability mass functions were imposed on the variables in order to use Shannon's discrete entropy functional. For instance in [21], a lattice metric was defined in terms of marginal and joint entropies. The entropy of the labels of the ergodic decomposition of [22] defines the *Gács-Körner common information*[8], also called *zero-error information* [23].

In contrast, nonstochastic information in the uv framework corresponds to the *Hartley entropy* (log-cardinality) of the taxicab partition or ergodic decomposition, or equivalently of the maximal common uv. In the next subsection, some basic properties of I_* are discussed.

E. *Properties of* I_*

Mutual information obeys several key properties, such as nonnegativity, symmetry, *monotonicity*, and the *data processing inequality* for Markov chains, which are usually proved by decomposing it into a linear combination of joint and marginal entropies and then applying function inequalities. Similar properties can be proved rather more directly for the nonstochastic information $\mathrm{I}_*[X;Y]$, even though it does not enjoy a similar decomposition.

To begin, note that the nonnegativity of $\mathrm{I}_*[X;Y]$ is obvious, from the definition (12). Similarly, the symmetry between X and Y is also clear from the definition of taxicab connectedness, by which

$$(x,y) \leftrightsquigarrow (x',y') \in [\![X,Y]\!] \Longleftrightarrow (y,x) \leftrightsquigarrow (y',x') \in [\![Y,X]\!].$$

Monotonicity states that for any uv's W, X, Y,

$$\mathrm{I}_*[X;Y,W] \geq \mathrm{I}_*[X;Y]. \tag{16}$$

This is obviously a very desirable property for any information index, since it states that the information gained about X

[7] However, note that in [18] nonstochastic information was defined by (14), and then shown to be equivalent to (12).

[8] 'Common information' is a common phrase; in addition to the two usages above, there is a *Wyner common information*, which is somewhat different.

from Y cannot decrease if Y is augmented with extra data W. It can be proved very directly via the equivalence to maximal common uv's. The maximal common uv $Z_* = f_*(X) \equiv g_*(Y)$ is clearly also a common uv for X and (Y,W), but possibly submaximal since it ignores W. Thus the cardinality of Z_* cannot be greater than that of the maximal common uv for X and (Y,W), implying (16).

Finally, the uv version of the *data processing inequality* states that for any Markov uncertainty chain $W \leftrightarrow X \leftrightarrow Y$,

$$\mathrm{I}_*[W;Y] \leq \mathrm{I}_*[X;Y], \tag{17}$$

i.e. inner uv pairs in the chain share more information than outer ones. This is a natural property, since each link in the chain inserts unrelated noise, in a sense. It is proved by first using monotonicity and the overlap partition characterisation of I_* to write

$$\mathrm{I}_*[Y;W] \overset{(16)}{\leq} \mathrm{I}_*[Y;X,W] \overset{(14)}{=} \log|[\![Y|X,W]\!]_*|. \tag{18}$$

By Markovness (9), $[\![Y|x,w]\!] = [\![Y|x]\!]$, $\forall (x,w) \in [\![X,W]\!]$. Thus the conditional range family $[\![Y|X,W]\!] = [\![Y|X]\!]$, from which it follows that $[\![Y|X,W]\!]_* = [\![Y|X]\!]_*$. Substituting this into the RHS of (18) completes the proof.

IV. Nonstochastic Conditional and Directed Information

In information theory, conditional mutual information measures how much information two random variables (rv's) share on average, on top of a third rv that is available to both. It forms part of the *(Marko-Massey) directed information*, which has been shown to characterise the ordinary capacity of channels with perfect feedback, and has also been proposed as a quantifier of causality between two random processes [30], [31].

In this section, nonstochastic versions of conditional and directed information are described. These have applications in scenarios where joint probability distributions do not exist or are hard to estimate, but where ranges or support sets are known with high confidence.

To begin, for any uv's X,Y and any realisation w of a uv W, let $\mathscr{T}[X;Y|w]$ denote the taxicab partition (see subsection III-A) of the conditional joint range $[\![X,Y|w]\!]$, given $W = w$ (3). Then define the *nonstochastic conditional information* of X and Y given W to be the minimum log-cardinality of $\mathscr{T}[X;Y|w]$

$$\mathrm{I}_*[X;Y|W] := \min_{w \in [\![W]\!]} \log|\mathscr{T}[X;Y|w]|\,. \tag{19}$$

This functional has a simple interpretation: if two agents are given data W and then each observe uv's X and Y privately, then $2^{\mathrm{I}_*[X;Y|W]}$ is the cardinality of the maximal common uv Z that is *new* with respect to what they were both given. This is explained below.

A. Meaning of Nonstochastic Conditional Information

In mathematical terms, if $\mathscr{Z}$ denotes the set of all uv's $Z \perp W$ such that $Z \equiv f(X,W) \equiv g(Y,W)$, then

$$\max_{Z \in \mathscr{Z}} \log|[\![Z]\!]| = \mathrm{I}_*[X;Y|W] \tag{20}$$

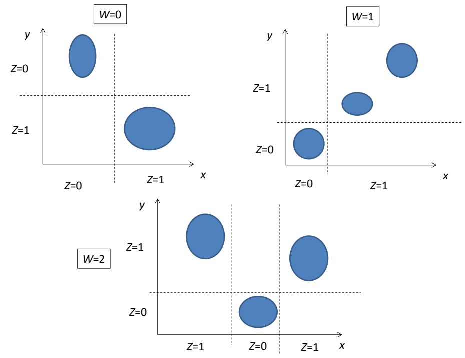


Fig. 2. Example of conditional joint ranges $[\![X,Y|w]\!]$, with conditional taxicab partitions and corresponding Z values as indicated.

The proof proceeds as follows.

Let $w^* \in \operatorname{Arg\,min}_{w \in [\![W]\!]} |\mathscr{T}[X;Y|w]|$, so that $\mathrm{I}_*[X;Y|W] = \log|\mathscr{T}[X;Y|W = w^*]|$, by (19). It is first shown that the left-hand side (LHS) of (20) can never exceed the right-hand side (RHS). The argument is by contradiction. Suppose that $\exists Z \in \mathscr{Z}$ with $\log|[\![Z]\!]| > \mathrm{I}_*[X;Y|W]$. As $Z \perp W$, it follows that

$$|[\![Z|w^*]\!]| = |[\![Z]\!]| > |\mathscr{T}[X;Y|w^*]|\,.$$

However the first term here, which gives the number of distinct values that the common variable $f(X,w^*) = g(Y,w^*)$ can take, must must be bounded above by the third term, which is the number of elements in the taxicab partition of the conditional joint range $[\![X,Y|w^*]\!]$. This is a contradiction, so we must have $\log|[\![Z]\!]| \leq \mathrm{I}_*[X;Y|W]$, $\forall$ uv's $Z \in \mathscr{Z}$.

It remains to construct $Z \in \mathscr{Z}$ that attains the RHS of (20). Let $n := \min_{w \in [\![W]\!]} |\mathscr{T}[X;Y|w]|$. For each $w \in [\![W]\!]$, let $Z_w = f'(X,w) = g'(Y,w)$ denote a maximal common variable between X and Y, given $W = w$. Note that this common variable takes $|\mathscr{T}[X;Y|w]| \geq n$ distinct values. Without loss of generality, suppose that Z_w takes integer values in $[1 : |\mathscr{T}[X;Y|w]|]$. Now define the uv $Z := \min\{Z_W, n\}$. Evidently

$$Z \equiv \min\{f'(X,W), n\} \equiv \min\{g'(X,W), n\}, \tag{21}$$

i.e. Z is a common uv for (X,W) and (Y,W). Furthermore $\{Z_w | w \in [\![W]\!]\}$ always contains the integers $1, \ldots, n \equiv \min_{w \in [\![W]\!]} |\mathscr{T}[X;Y|w]|$. Thus $\forall w \in [\![W]\!]$,

$$[\![Z|w]\!] = [1 : \min\{|[\![Z_w]\!]|, n\}] = [1 : n] = [\![Z]\!],$$

implying that Z is unrelated with W. Thus $Z \in \mathscr{Z}$ and from the second equality above has cardinality $n \equiv \min_{w \in [\![W]\!]} |\mathscr{T}[X;Y|w]|$, as required.

B. Properties of Nonstochastic Conditional Information

Nonstochastic conditional information shares four important properties with its stochastic analogue.

- *Nonnegativity:* $\mathrm{I}_*[X;Y|W] \geq 0$.
- *Symmetry:* $\mathrm{I}_*[X;Y|W] = \mathrm{I}_*[Y;X|W]$
- *Monotonicity:* $\mathrm{I}_*[X;Y|W] \leq \mathrm{I}_*[X;Y,Z|W]$.

- *Data Processing Inequality:* If $X \leftrightarrow Y \leftrightarrow Z|W$ is any conditional Markov uncertainty chain given W, then

$$\mathrm{I}_*[X;Z|W] \leq \mathrm{I}_*[X;Y|W].$$

The proofs of these properties are omitted, since they follow almost the same lines as for unconditional I_*.

C. Nonstochastic Directed Information

In a probabilistic setting, the Marko-Massey directed information [17] between two sequences of rv's is defined as

$$\mathrm{I}[X_{0:n} \to Y_{0:n}] := \sum_{k=0}^{n} \mathrm{I}[X_{0:k};Y_k|Y_{0:k-1}], \qquad (22)$$

where the conditional mutual information $I[X;Y|Z] = H[X|Z] - H[X|Y,Z]$. It is a relatively recent concept, which sprang from attempts to understand feedback in communication systems and quantify causality, and is not the same as the mutual information $\mathrm{I}[X_{0:n};Y_{0:n}]$.

With a notion of nonstochastic conditional information in place, it is possible to construct a parallel of (22). Let $X_{0:n}$ and $Y_{0:n}$ be two sequences of uv's, of the same length $n+1$. Then the *nonstochastic directed information* is defined as

$$\mathrm{I}_*[X_{0:n} \to Y_{0:n}] := \sum_{k=0}^{n} \mathrm{I}_*[X_{0:k};Y_k|Y_{0:k-1}]. \qquad (23)$$

It turns out that this concept precisely characterises the zero-error capacity of a channel in the presence of perfect, one-step delayed feedback - see subsection V-C. There is also a related interpretation as a measure of influence in causal systems. This is described as follows.

Suppose we treat $X_{0:n}$ and $Y_{0:n}$ as time sequences of uv's associated with some unknown causal system, and wish to quantify how well these sequences can be treated as 'input' and 'output' respectively. As the system structure is unknown, the putative output Y_k may generally depend on all past outputs $Y_{0:k-1}$, as well as all inputs $X_{0:k}$ up to present time.

The nonstochastic information $\mathrm{I}_*[X_{0:n};Y_{0:n}]$ is useless here, since it is symmetric, and insensitive to reorderings of time. However, recall that the nonstochastic conditional information $\mathrm{I}_*[X_{0:k};Y_k|Y_{0:k-1}]$ (19) corresponds to the maximal common uv Z_k that is common to $(X_{0:k},Y_{0:k-1})$ and $Y_{0:k}$ but unrelated with the past $Y_{0:k-1}$ (20). In other words, it measures how much new information can be stated about the outputs up to time k, by the inputs and the past outputs. As Z_k is unrelated with past $Y_{0:k-1}$, it is consequently unrelated with $Z_{0:k-1}$, which is a function of $Y_{0:k-1}$. Thus the sum on the RHS of (23) coincides precisely with the log-cardinality of $Z_{0:n}$, and captures how much can be stated about the output process as a causal function of the inputs and past outputs. It is maximised when Y_k is a deterministic, causal function of the inputs and past outputs.

V. NOISY COMMUNICATION CHANNELS

In this section, it is discussed how the uncertain variable (uv) framework and nonstochastic information concepts above give intrinsic characterisations of certain zero-error coding capacity concepts for communication channels. In particular, the *zero-error capacity* C_0 of a memoryless channel coincides with the highest rate of nonstochastic information (12) across it, while the *zero-error capacity with feedback* $C_{0\mathrm{f}}$ coincides with the highest rate of nonstochastic directed information (23).

Before these results can be given, signals and channels must be appropriately defined in the uv framework.

A. Signals and Channels

A little care is needed to define uncertain signals and systems that exist on the semi-infinite discrete-time axis $\mathbb{Z}_{\geq 0}$. Let $\mathbb{X}^\infty$ be the space of all $\mathbb{X}$-valued, discrete-time sequences $x = (x_k)_{k=0}^{\infty}$. Similar to the way that discrete-time random processes are defined, let an *uncertain signal* X be a mapping from the sample space Ω to $\mathbb{X}^\infty$.[9] Confining this mapping to any time $t \in \mathbb{Z}_{\geq 0}$ yields a uv X_t. As with uv's, the dependence on $\omega \in \Omega$ will not usually be indicated.Also note that $[\![X]\!]$ is a subset of the function space $\mathscr{X}$.

The next step is to define a suitable notion of an *uncertain channel*. In communications theory, discrete stochastic channels are often defined in terms of conditional probabilities of outputs given inputs, e.g. the binary symmetric channel, with the channel noise remaining implicit. On the other hand, for analog channels the usual custom is to indicate the channel noise explicitly, e.g. the additive white Gaussian noise channel. In [18], [19], the former approach was adapted to define uncertain channels in terms of conditional ranges. However, it turns out to be simpler to take the second approach, and treat channel noise as an explicit object. In particular, the mutual unrelatedness between messages and channels when feedback is present becomes easier to capture.

So let a *stationary memoryless uncertain channel (SMUC)* be defined by

a) an *unrelated, identically spread (uis)* noise signal $V = (V_k)_{k=0}^{\infty}$ on a space $\mathbb{V}$, i.e.

$$[\![V_k|v_{0:k-1}]\!] = [\![V_k]\!] = \mathbb{V}, \quad \forall v_{0:k-1} \in \mathbb{V}^k, k \in \mathbb{Z}_{\geq 0}; \quad (24)$$

b) input and output spaces $\mathbb{X}, \mathbb{Y}$, and a *transition function* $\tau : \mathbb{X} \times \mathbb{V} \to \mathbb{Y}$.

A family of pairs of input-output signals that are compatible with the channel also needs to be defined. This depends on the problem being studied. At one extreme, the transmitter may have one-step delayed perfect feedback from the channel receiver. For this case, define an associated family $\mathscr{G}_{\mathrm{f}}$ of all uncertain input-output signal pairs (X,Y) s.t. $\forall k \in \mathbb{Z}_{\geq 0}$,

- $Y_k = \tau(X_k, V_k)$,
- and $X_{0:k} \perp V_k$

This last condition expresses the fact that the current channel noise is unrelated to all channel inputs up to now, but may influence a future input through the feedback.

At the other extreme, the transmitter may have no feedback whatsoever. In this case, tighten the last condition so that $X \perp V$, i.e. observations of X and V at any finite number of

[9] A smaller range $\mathscr{X} \subset \mathbb{X}^\infty$ would be required if power, run-length or other dynamic constraints must be imposed on realisations of X.

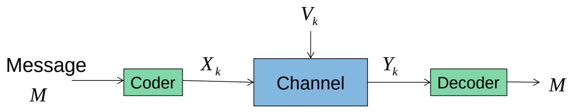


Fig. 3. Noisy channel without feedback

times (not necessarily identical times, or equal in number) are unrelated. This yields the smaller family $\mathscr{G}_{\text{nf}} \subset \mathscr{G}_{\text{f}}$.

B. Zero Error Communication without Feedback

The aim in a zero-error communication system is to transmit any message M that can take up to $\mu \geq 1$ distinct values as a sequence of n inputs, so that after receiving the n uncertain channel outputs, the message M can be reproduced exactly. Evidently, a code is required to convert M to a sequence of inputs. More formally, let a zero-error code *without feedback* be defined by

- a block length $n \in \mathbb{Z}_{\geq 1}$;
- a message cardinality $\mu \geq 1$;
- and an encoder mapping $\gamma : [1 : \mu] \to \mathbb{X}^n$, s.t. for any message $M \perp V$ taking μ distinct values $m^1, \ldots, m^\mu$ and a channel input sequence given by $X_{0:n-1} = \gamma(i)$ if $M = m^i$, it holds that

$$|[\![M|y_{0:n-1}]\!]| = 1, \quad \forall y_{0:n-1} \in [\![Y_{0:n-1}]\!].$$

Note that the last condition is equivalent to the existence of a decoder at the receiver that always takes $Y_{0:n-1} \mapsto M$, despite channel noise.

The rate of the code is defined as $(\log \mu)/n$. The *zero-error capacity* C_0 of the channel is then defined as the highest rate of all zero-error block codes,

$$C_0 := \sup_{n,\mu \in \mathbb{Z}_{\geq 1}, \gamma} \frac{\log \mu}{n} = \lim_{n \to \infty} \sup_{\mu \in \mathbb{Z}_{\geq 1}, \gamma} \frac{\log_2 \mu}{n}, \tag{25}$$

where the limit follows from subadditivity. Note that the zero-error rate is typically much smaller than the ordinary capacity C, which is defined by allowing a small probability of decoding error that approaches 0. Insisting on exactly no decoding errors seems like a small difference, but in fact introduces significant conservatism.

The definition (25) is *operational*, i.e. in terms of the highest zero-error code rate. However, by using nonstochastic information, it can be shown that C_0 has an intrinsic characterisation, as the highest nonstochastic information rate that is possible across a channel:

$$\begin{aligned} C_0 &= \sup_{n \geq 1, (X,Y) \in \mathscr{G}_{\text{nf}}} \frac{\mathrm{I}_*[X_{0:n-1}; Y_{0:n-1}]}{n} \\ &= \lim_{n \to \infty} \sup_{(X,Y) \in \mathscr{G}_{nf}} \frac{\mathrm{I}_*[X_{0:n-1}; Y_{0:n}]}{n}. \end{aligned} \tag{26}$$

This is a partial analogue of Shannon's *channel coding theorem*, which characterises the ordinary channel capacity C as the largest mutual information rate $\mathrm{I}[X_{0:n-1}; Y_{0:n-1}]/n$ across a stochastic memoryless channel [1], maximised over all input sequence distributions compatible with the channel. However, Shannon's expression can be further simplified to a 'single-letter' supremum of $\mathrm{I}[X;Y]$ over the distribution of a single input, thanks to certain convenient properties of mutual information. For discrete channels this reduces the computation of C to solving a finite-dimensional optimisation.

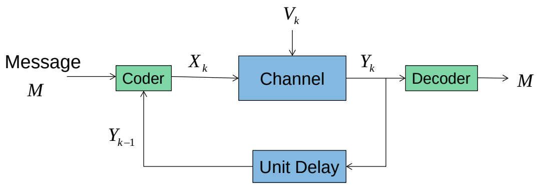


Fig. 4. Noisy channel with perfect feedback

Such a single-letter characterisation is not possible in (26). Nonetheless, it is a powerful theoretical aid in the analysis of problems such as uniform state estimation via noisy channels, as discussed in section VI. For discrete channels, the computation of C_0 as defined in (25) is a very difficult combinatorial problem, with formulas available in few cases [32]. There is a possibility that (26) may give alternative approaches for estimating or bounding C_0, but this is not explored here.

The reader is referred to [18] for a proof of (26) in terms of overlap partitions. It should also be noted that in [23], C_0 for a stochastic discrete memoryless channel was characterised as the largest Shannon entropy rate of the maximal common rv for $X_{0:n-1}$ and $Y_{0:n-1}$. The result above is similar, except that the setting is nonstochastic and the Shannon entropy is replaced with the Hartley entropy, i.e. log-cardinality, of the maximal common uv.

C. Zero Error Communication with Perfect Feedback

In this subsection, zero-error communication is studied under the assumption that the transmitter is told the exact channel outputs, with a one-step delay. As mentioned in the Introduction, the ordinary capacity C of a stochastic memoryless channel does not change when such feedback is allowed. However, this is not so for the zero-error capacity.

Similar to before, define a zero-error code with perfect feedback by

- a block length $n \in \mathbb{Z}_{\geq 1}$;
- a message cardinality $\mu \in \mathbb{Z}_{\geq 1}$;
- and a sequence $\gamma_{0:n-1}$ of encoder functions s.t. for any message $M \perp V$ taking values $m^1, \ldots, m^\mu$, and any input and output signals satisfying $X_k = \gamma_k(i, Y_{0:k-1})$ if $M = m^i$,

$$|[\![M|y_{0:n-1}]\!]| = 1, \quad \forall y_{0:n-1} \in [\![Y_{0:n-1}]\!].$$

As before, the code rate is defined by $(\log \mu)/n$. The zero-error feedback capacity $C_{0\text{f}}$ is defined as the highest feedback coding rate that yields no decoding errors, i.e.

$$C_{0\mathrm{f}} := \sup_{n,\mu\in\mathbb{Z}_{\geq 1},\gamma_{0:n-1}} \frac{\log_2 \mu}{n} = \lim_{n\to\infty} \sup_{\mu\in\mathbb{Z}_{\geq 1},\gamma_{0:n-1}} \frac{\log_2 \mu}{n}. \quad (27)$$

In other words, it is the growth rate of the maximum cardinality of sets of feedback coding functions that can be unambiguously determined from channel outputs [16].

The definition above is an operational one. However, it turns out to have an information-theoretic characterisation as the maximum nonstochastic directed information rate across the channel [19]:

$$\begin{aligned} C_{0\mathrm{f}} &= \sup_{n\in\mathbb{Z}_{\geq 1},(X,Y)\in\mathscr{G}} \frac{\mathrm{I}_*[X_{0:n-1} \to Y_{0:n-1}]}{n} \\ &= \lim_{n\to\infty} \sup_{(X,Y)\in\mathscr{G}} \frac{\mathrm{I}_*[X_{0:n-1} \to Y_{0:n-1}]}{n}, \end{aligned} \quad (28)$$

where $\mathrm{I}_*[\cdot \to \cdot]$ is the nonstochastic directed information (19). This parallels the characterisation in [33], [34] of the ordinary feedback capacity C_f of stochastic channels with memory as the maximum rate of Marko-Massey directed information (22) across the channel, Although the uncertain channel here is memoryless, it is possible to extend (28) to channels with memory; this will be reported elsewhere.

Note that unlike in the stochastic framework, for a memoryless channel $C_{0\mathrm{f}}$ may be strictly larger than the zero-error capacity without feedback C_0. Counter-intuitively, for discrete memoryless channels $C_{0\mathrm{f}}$ is an easier object to study than C_0, and can be obtained through an auxiliary, finite-dimensional optimisation [16]. However, this optimisation problem has no information-theoretic interpretation, and arises from a coding analysis. This coding analysis has recently been extended to find zero-error feedback capacity formulas for certain classes of channels with memory. The formula (28) may well allow $C_{0\mathrm{f}}$ to be determined for other channels, but its main use at present is in allowing $C_{0\mathrm{f}}$ to be thought of as an information-theoretic object. This lets information-theoretic tools to be applied to certain nonstochastic feedback problems, as described in the next section.

VI. LTI PLANTS AND NOISY CHANNELS

In this section, I_* is used to study the problems of uniformly estimating or stabilising the state of a linear time-invariant (LTI) plant via a stationary memoryless uncertain channel (SMUC). First, some related work is discussed.

In the case where the channel is an errorless digital bit-pipe, the 'data rate theorem' states that the estimation errors or states can be bounded or taken to zero iff $R > H$, where R is the average channel bit-rate and H is the sum of the log-magnitudes of the unstable eigenvalues of the plant dynamical matrix. This tight condition holds with no noise, stochastic plant noise, or bounded plant disturbances, and under different notions of convergence or stability, e.g. uniform, in rth moment or almost surely (a.s.) [35], [36], [37], [38], [39], [40], [9]. The stochastic formulations use differential entropy to prove necessity, while deterministic ones typically employ volume-partitioning arguments.

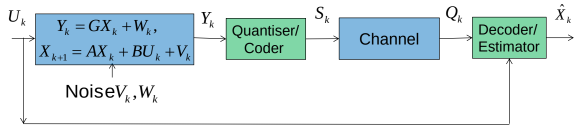


Fig. 5. State Estimation via a Noisy Channel ($U_k \equiv 0$)

However, with channel noise, the controller or estimator does not necessarily know what the encoder sent, and this unified criterion splits into multiple ones. Depending on the notion of stability on the states or estimation errors, the assumptions on initial states and plant noise, the channel model and the availability of explicit channel feedback, the data rate R in the inequality must replaced with either the ordinary capacity C [41], [10], [42], the *anytime capacity* C_any [43], or the zero-error capacities $C_{0\mathrm{f}}, C_0$ with and without feedback [44]. In other words, with noisy channels no single figure of merit is appropriate for all situations.

For the purposes of this section, the results of [44] are immediately relevant. In that article, the problem of almost surely (a.s.) uniformly estimating or stabilising the states of an LTI plant via a noisy memoryless channel was studied. The channel and initial plant state were modelled stochastically, but the disturbances entering the plant were treated as bounded unknowns without statistical structure. When the plant disturbances are nulled, it is known that $C > H$ is a tight condition to be able to ensure a.s. bounded errors or states [10]. However, when plant disturbances are allowed, then the criterion changes to become $C_0 > H$ for a.s. uniform state estimation and $C_{0\mathrm{f}} > H$ for a.s. uniform stabilisation, even when the disturbance bound is arbitrarily small As $C_0 \leq C_{0\mathrm{f}}$ are usually much smaller than C, these are typically more restrictive conditions. The reason is that nonstochastic disturbances do not average out in the long run. Therefore it becomes crucial for no decoding errors to occur in the channel, not just for their average probability to be arbitrarily small. These important results were proved using volume-partitioning arguments and a law of large numbers that exploited the random initial state and channel. No information theory was applied.

In this section, neither the initial state, plant noise or the noisy channel are modelled stochastically. As a consequence, probability and laws of large numbers cannot be employed in the analysis. Nonetheless, analogous bounds can be obtained with the aid of nonstochastic information theory.

A. Uniform State Estimation

Proofs of the results in this subsection can be found in [18]. First the problem of state estimation for a disturbance-free LTI plant is considered.

1) Disturbance-Free Plant: The components of the system are:

- A noiseless LTI plant with zero input and uncertain initial state:

$$X_{k+1} = AX_k, \quad Y_k = GX_k, \quad X_0 \text{ a uv}, \ \forall k \in \mathbb{Z}_{\geq 0}.$$

- A coder that maps $Y_{0:k}$ to S_k at each time k.
- A stationary memoryless uncertain channel with inputs S_k, outputs Q_k and channel noise terms Z_k.
- An estimator that maps the past symbols $Q_{0:k-1}$ to an estimate $\hat{X}_k$.

Consider the objective of *uniform ρ-exponential convergence* from an ℓ-ball. I.e. given $\rho, \ell > 0$, construct a coder-estimator s.t. for any uv X_0 with range $[\![X_0]\!] \subseteq \mathbf{B}_\ell(0)$,

$$\lim_{k\to\infty} \sup_{\omega\in\Omega} \rho^{-k} \|X_k - \hat{X}_k\| = 0.$$

Assume the following:

DF1: A has one or more eigenvalues with magnitude $> \rho$.
DF2: (G,A) is observable.[10]
DF3: $X_0 \perp Z$

Under these assumptions, it is shown in [18] that if uniform ρ-exponential convergence is achieved from some ℓ-ball, then

$$C_0 \geq \sum_{|\lambda_i|\geq\rho} \log\left(\frac{|\lambda_i|}{\rho}\right). \tag{29}$$

Conversely, if (29) holds strictly, then for any $\ell > 0$, a coder-estimator that achieves uniform ρ-exponential convergence from $\mathbf{B}_\ell(0)$ can be constructed.

The proof of sufficiency relies on constructing a coder-estimator, given a channel that satisfies (29) strictly. First, the plant is down-sampled, and then a zero-error code with sufficiently long block-length is applied, so as to convert the erroneous channel into an errorless one with average bit rate arbitrarily close to C_0. This reduces the proof of sufficiency to an application of the 'data rate theorem', without requiring any of the concepts presented here.

The proof of necessity is more complex. With no channel noise, the basic idea is to analyse the growth rate of state uncertainty volumes. In [10], [44], this technique is extended to discrete stochastic channels to prove a.s. convergence and boundedness. Importantly, treating the initial state and channel noise as random variables allows a strong law of large numbers to be applied.

Unfortunately, no laws of large numbers hold for the uncertain variables (uv's) of this paper. Nonetheless, it turns out that (29) can still be proved in the uv framework, by using the nonstochastic information-theoretic characterisation of C_0 in terms of I_* (26). The proof, which is relatively direct, considers uncertainty diameters not volumes, and exploits properties of I_* such as monotonicity and data processing. It is illuminating to go through a sketch of the necessity argument, restricted here to scalar plants; the reader is referred to [18] for the complete proof.

a) Necessity Argument - Scalar Case: First, pick arbitrarily large $t \in \mathbb{Z}_{\geq 1}$ and small $\varepsilon \in \left(0, 1 - \frac{\rho}{|\lambda|}\right)$. Divide $[-\ell, \ell]$ into

$$\kappa := \left\lfloor \left| \frac{(1-\varepsilon)\lambda}{\rho} \right|^t \right\rfloor \geq 1$$

equal intervals of length $2\ell/\kappa$. Now, inside each interval construct a centred subinterval $\mathbf{I}(s)$ of **shorter** length ℓ/κ. Define the subinterval family

$$\mathscr{H} := \{\mathbf{I}(s) : s = 1, \ldots, \kappa\}, \tag{30}$$

noting that subintervals $\in \mathscr{H}$ are separated by a gap $\geq \ell/\kappa$. Now, consider an initial state uv X_0 with range $[\![X_0]\!] = \bigcup_{\mathbf{H}\in\mathscr{H}} \mathbf{H} \subset [-\ell, \ell]$.

Let $E_k := X_k - \hat{X}_k$ denote the estimation error. By hypothesis, $\exists \phi > 0$ s.t.

$$\begin{aligned}
\phi\rho^k &\geq \sup[\![|E_k|]\!] \geq 0.5\mathrm{diam}[\![E_k]\!] && (31)\\
&\geq 0.5\mathrm{diam}[\![E_k|q_{0:k-1}]\!] && (32)\\
&= 0.5\mathrm{diam}\left[\!\!\left[\lambda^k X_0 - \eta_k(q_{0:k-1})\,|q_{0:k-1}\right]\!\!\right] &&\\
&= 0.5\mathrm{diam}[\![\lambda^k X_0|q_{0:k-1}]\!] && (33)\\
&= 0.5|\lambda|^k \mathrm{diam}[\![X_0|q_{0:k-1}]\!], && (34)
\end{aligned}$$

where (31) arises from the fact that the diameter of a real set is at most twice the maximum magnitude of an element, (32) from the fact that conditioning can only reduce a set, and (33) from the invariance of set-diameter to translations.

Next it is shown that for large t, no two sets in $\mathscr{H}$ (30) can be $[\![X_0|Q_{0:t-1}]\!]$-overlap-connected. Suppose in contradiction that $\exists \mathbf{H} \in \mathscr{H}$ that is $[\![X_0|Q_{0:t-1}]\!]$-overlap-connected with another set in $\mathscr{H}$. This would imply that there is a conditional range $[\![X_0|q_{0:t-1}]\!]$ containing both a point $u \in \mathbf{H}$ and a point v in some $\mathbf{H}' \in \mathscr{H} \setminus \{\mathbf{H}\}$. Consequently,

$$|u - v| \leq \mathrm{diam}[\![X_0|q_{0:t-1}]\!] \overset{(34)}{\leq} \frac{2\phi\rho^t}{|\lambda|^t}. \tag{35}$$

However, recall that any two sets $\in \mathscr{H}$ are separated by a distance of at least ℓ/κ. So

$$\begin{aligned}
|u - v| &\geq \frac{l}{\kappa} = \frac{\ell}{\lfloor ((1-\varepsilon)|\lambda|/\rho)^t \rfloor}\\
&\geq \frac{\ell}{((1-\varepsilon)|\lambda|/\rho)^t} = \frac{l\rho^t}{|(1-\varepsilon)\lambda|^t}.
\end{aligned}$$

The RHS of this would exceed the RHS of (35) when t is sufficiently large that $\left(\frac{1}{1-\varepsilon}\right)^t > 2\phi/\ell$, yielding a contradiction. So for large enough t, no two sets of $\mathscr{H}$ are $[\![X_0|Q_{0:t-1}]\!]$-overlap-connected. Consequently,

$$\begin{aligned}
2^{\mathrm{I}_*[X_0;Q_{0:t-1}]} \equiv |[\![X_0|Q_{0:t-1}]\!]_*| &\geq |\mathscr{H}|\\
&= \left\lfloor \left| \frac{(1-\varepsilon)\lambda}{\rho} \right|^t \right\rfloor\\
&\geq 0.5 \left| \frac{(1-\varepsilon)\lambda}{\rho} \right|^t, && (36)
\end{aligned}$$

since $\lfloor x \rfloor > x/2$, for every $x \geq 1$. However, $X_0 \leftrightarrow S_{0:t-1} \leftrightarrow Q_{0:t-1}$ is a Markov uncertainty chain, so

$$\mathrm{I}_*[X_0; Q_{0:t-1}] \leq \mathrm{I}_*[S_{0:t-1}; Q_{0:t-1}] \leq tC_0.$$

Substitute this last inequality into the LHS of (36), take logarithms and divide by t to get

$$C_0 \geq \log_2(1-\varepsilon) + \log_2|\lambda/\rho| - 1/t.$$

[10] This can be relaxed to requiring observability of the modes with eigenvalue magnitudes $\geq \rho$.

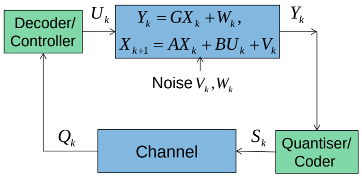


Fig. 6. Stabilisation via a Noisy Channel

Letting $t \to \infty$ yields

$$C_0 \geq \log_2(1-\varepsilon) + \log_2|\lambda/\rho|.$$

As ε can be made arbitrarily small, this completes the necessity argument for the scalar case.

2) Plant with Disturbances: Now suppose that process and measurement noise is present, i.e.

$$X_{k+1} = AX_k + V_k, \quad Y_k = GX_k + W_k,$$

where X_0 is a uv and V, W are uncertain signals. The coder, channel and estimator remain the same. With plant noise present convergence of the estimation errors to zero is impossible, so the objective is relaxed to attaining uniformly bounded estimation errors beginning from an ℓ-ball. I.e. given $\ell > 0$, construct a coder-estimator s.t. for any initial state X_0 with $[\![X_0]\!] \subseteq \mathbf{B}_\ell(0)$,

$$\sup_{k\in\mathbb{Z}_{\geq 0}, \omega\in\Omega} \|X_k - \hat{X}_k\| < \infty.$$

Impose the following assumptions:

D1: A has one or more eigenvalues with magnitude > 1.
D2: (G,A) is observable.[11]
D3: $[\![V_k]\!]$ and $[\![W_k]\!]$ are uniformly bounded over k.
D4: X_0, V, W and Z are mutually unrelated.
D5: The zero-noise sequence pair $(v,w) = (0,0)$ is valid, i.e. $(0,0) \in [\![V,W]\!]$.

Under these conditions,if uniformly bounded estimation errors are achieved from some ℓ-ball, then

$$C_0 \geq \sum_{|\lambda_i|\geq 1} \log_2|\lambda_i|. \tag{37}$$

Conversely, if (37) holds strictly, then for any $\ell > 0$, a coder-estimator that achieves uniformly bounded estimation errors from $\mathbf{B}_\ell(0)$ can be constructed.

B. Uniform Stabilisation via Noisy Channels

In this section, the analogous problem of controlling an LTI plant via a noisy channel is addressed. As before, the case of plants without disturbances is first addressed.

1) Disturbance-Free Plant: Now suppose that the plant is given by

$$X_{k+1} = AX_k + BU_k, \quad Y_k = GX_k,$$

where X_0 is a uv. The coder and channel are as before, but the estimator is replaced by a controller that maps the past channel output sequence $Q_{0:k-1}$ to a plant input U_k. The objective is *uniform ρ-exponential stability* on an ℓ-ball. I.e. given $\rho, \ell > 0$, the aim is to construct a coder-controller s.t. for any uv X_0 with range $[\![X_0]\!] \subseteq \mathbf{B}_\ell(0)$,

$$\lim_{k\to\infty} \sup_{\omega\in\Omega} \rho^{-k}\|X_k\| = 0.$$

Make the following assumptions:

DFC1: A has one or more eigenvalues with magnitude $> \rho$.
DFC2: (G,A) is observable and (A,B) is controllable.
DFC3: $X_0 \perp Z$

It can then be shown that if uniform ρ-exponential stability is achieved on some ℓ-ball, then

$$C_{0\mathrm{f}} \geq \sum_{|\lambda_i|\geq\rho} \log\left(\frac{|\lambda_i|}{\rho}\right). \tag{38}$$

Conversely, if (38) holds strictly, then for any $\ell > 0$, a coder-controller that achieves uniform ρ-exponential stability on $\mathbf{B}_\ell(0)$ can be constructed [19].

2) Plants with Disturbances: Now suppose the plant is given by

$$X_{k+1} = AX_k + BU_k + V_k, \quad Y_k = GX_k + W_k,$$

where X_0 is a uv and V, W are process and observation noise signals. The coder, channel and controller remain the same, but the objective is now to achieve uniformly bounded states beginning from an ℓ-ball. In other words, given $\ell > 0$, the aim is to construct a coder-controller s.t. for any initial state X_0 with $[\![X_0]\!] \subseteq \mathbf{B}_l(0)$,

$$\sup_{k\in\mathbb{Z}_{\geq 0}, \omega\in\Omega} \|X_k\| < \infty.$$

Impose the following assumptions

DC1: A has one or more eigenvalues with magnitude ≥ 1.
DC2: (G,A) is observable and (A,B) is controllable.
DC3: $[\![V_k]\!]$ and $[\![W_k]\!]$ are uniformly bounded over k.
DC4: X_0, V, W and Z are mutually unrelated.
DC5: The zero-noise sequence pair $(v,w) = (0,0)$ is valid, i.e. $(0,0) \in [\![V,W]\!]$.

It can be shown that if uniformly bounded estimation errors are achieved from some ℓ-ball, then

$$C_{0f} \geq \sum_{|\lambda_i|\geq 1} \log_2|\lambda_i|. \tag{39}$$

Conversely, if (39) holds strictly, then for any $\ell > 0$, a coder-controller that achieves uniformly bounded states from $\mathbf{B}_\ell(0)$ can be constructed.

[11] This can be relaxed to observability of the modes with eigenvalue magnitudes ≥ 1.

VII. FUTURE DIRECTIONS

This paper discussed a recent nonstochastic theory of information, and its application to the analysis of certain problems in zero-error communications and uniform estimation and control. This theory is far from mature, and there are numerous challenges and open problems. Three of the major ones are listed below.

- How can this framework be adapted to handle disturbances with bounds on energy or average power over time, rather than on magnitudes?
- The systems considered here consist of two agents, i.e. transmitter and receiver, coder and controller, etc. However, taxicab connectedness can be extended to three or more variables. Could a corresponding triple or n-tuple information help to analyse systems with three or more agents?
- Can the characterisations of zero-error capacity and zero-error feedback capacity in terms of I_* and directed I_* be exploited to estimate them for various channels of interest, perhaps using Monte Carlo methods?

REFERENCES

[1] C. E. Shannon, “A mathematical theory of communication,” *Bell Syst. Tech. Jour.*, vol. 27, pp. 379–423, 623–56, 1948, reprinted in ‘Claude Elwood Shannon Collected Papers’, IEEE Press, 1993.

[2] G. J. Chaitin, “A theory of program size formally identical to information theory,” *J. Assoc. Computing Machinery*, vol. 22, no. 3, pp. 329–40, 1975.

[3] J. P. Burg, “The relationship between maximum entropy spectra and maximum likelihood spectra,” *Geophysics*, vol. 37, no. 2, pp. 375–6, 1972.

[4] E. T. Jaynes, “On the rationale of maximum-entropy methods,” *Proc. IEEE*, vol. 70, no. 9, pp. 939–52, 1982.

[5] W. van Drongelen, S. Nayak, D. M. Frim, M. H. Kohrman, V. L. Towle, H. C. Lee, A. B. McGee, M. S. Chico, and K. E. Hecox, “Seizure anticipation in pediatric epilepsy: use of Kolmogorov entropy,” *Pediatric Neurology*, vol. 29, no. 3, pp. 208–13, 2003.

[6] X. Feng and K. A. Loparo, “Active probing for information in control systems with quantized state measurements: a minimum entropy approach,” *IEEE Trans. Autom. Contr.*, vol. 42, pp. 216–38, 1997.

[7] ——, “Optimal state estimation for stochastic systems: an information theoretic approach,” *IEEE Trans. Autom. Contr.*, vol. 42, pp. 771–85, 1997.

[8] S. Tatikonda, A. Sahai, and S. Mitter, “Stochastic linear control over a communication channel,” *IEEE Trans. Autom. Contr.*, vol. 49, no. 9, pp. 1549–61, Sep. 2004.

[9] G. N. Nair and R. J. Evans, “Stabilizability of stochastic linear systems with finite feedback data rates,” *SIAM J. Contr. Optim.*, vol. 43, no. 2, pp. 413–36, July 2004.

[10] A. S. Matveev and A. V. Savkin, “An analogue of Shannon information theory for detection and stabilization via noisy discrete communication channels,” *SIAM J. Contr. Optim.*, vol. 46, no. 4, pp. 1323–67, 2007.

[11] N. C. Martins and M. A. Dahleh, “Feedback control in the presence of noisy channels: Bode-like fundamental limitations of performance,” *IEEE Trans. Autom. Contr.*, vol. 53, pp. 1604–15, 2008.

[12] J. S. Freudenberg, R. H. Middleton, and V. Solo, “Stabilization and disturbance attenuation over a Gaussian communication channel,” *IEEE Trans. Autom. Contr.*, vol. 55, pp. 795–799, 2010.

[13] J. S. Freudenberg, R. H. Middleton, and J. H. Braslavsky, “Minimum variance control over a Gaussian communication channel,” *IEEE Trans. Autom. Contr.*, vol. 56, pp. 1751–1765, 2011.

[14] H. Ishii, K. Okano, and S. Hara, “Achievable sensitivity bounds for mimo control systems via an information theoretic approach,” *Sys. Contr. Lett.*, vol. 60, pp. 111–118, 2011.

[15] A. Katok, “Fifty years of entropy in dynamics: 1958–2007,” *J. Modern Dynamics*, vol. 1, no. 4, pp. 545–96, 2007.

[16] C. E. Shannon, “The zero-error capacity of a noisy channel,” *IRE Trans. Info. Theory*, vol. 2, pp. 8–19, 1956.

[17] J. L. Massey, “Causality, feedback and directed information,” in *Proc. Int. Symp. Inf. Theory App.*, Nov. 1990, pp. 1–6, full preprint downloaded from http://csc.ucdavis.edu/ rgjames/static/pdfs/.

[18] G. N. Nair, “A nonstochastic information theory for communication and state estimation,” *IEEE Trans. Autom. Contr.*, vol. 58, no. 6, pp. 1497–510, June 2013.

[19] ——, “A nonstochastic information theory for feedback,” in *Proc. IEEE Conf. Decision and Control*, Maui, USA, 2012, pp. 1343–8.

[20] A. Rényi, *Foundations of Probability*. Holden-Day, 1970.

[21] C. E. Shannon, “The lattice theory of information,” *Trans. IRE Prof. Group on Info. Theory*, vol. 1, no. 1, pp. 105–8, 1953.

[22] P. Gács and J. Körner, “Common information is far less than mutual information,” *Problems of Control and Information Theory*, vol. 2, no. 2, pp. 119–162, 1972.

[23] S. Wolf and J. Wullschleger, “Zero-error information and applications in cryptography,” in *Proc. IEEE Info. Theory Workshop*, San Antonio, USA, 2004, pp. 1–6.

[24] G. J. Klir, “Generalized information theory: aims, results, and open problems,” *Reliability Engineering and System Safety*, vol. 84, pp. 214–38, 2004, sec. 3.1.

[25] H. Shingin and Y. Ohta, “Disturbance rejection with information constraints: Performance limitations of a scalar system for bounded and Gaussian disturbances,” *Automatica*, vol. 48, no. 6, pp. 1111–6, 2012.

[26] P. Billingsley, *Probability and Measure*. Wiley, 1995.

[27] R. V. L. Hartley, “Transmission of information,” *Bell Syst. Tech. Jour.*, vol. 7, no. 3, pp. 535–63, 1928.

[28] A. Rényi, “On measures of entropy and information,” in *Proc. 4th Berkeley Symp. Maths., Stats. and Prob.*, Berkeley, USA, 1960, pp. 547–61.

[29] R. Aumann, “Agreeing to disagree,” *Annals of Statistics*, vol. 4, pp. 1236–1239, 1976.

[30] C. J. Quinn, T. P. Coleman, N. Kiyavash, and N. G. Hatsopoulos, “Estimating t he directed information to infer causal relationships in ensemble neural spike train recordings,” *J. Comput. Neuroscience*, vol. 30, no. 1, pp. 17–44, 2011.

[31] P. O. Amblard and O. J. J. Michel, “The relation between Granger causality and directed information theory: a review,” *Entropy*, vol. 15, no. 1, pp. 113–143, 2013.

[32] J. Korner and A. Orlitsky, “Zero-error information theory,” *IEEE Trans. Info. Theory*, vol. 44, pp. 2207–29, 1998.

[33] Y. H. Kim, “A coding theorem for a class of stationary channels with feedback,” *IEEE Trans. Info. Theory*, pp. 1488–99, 2008.

[34] S. Tatikonda and S. Mitter, “The capacity of channels with feedback,” *IEEE Trans. Info. Theory*, pp. 323–49, 2009.

[35] W. S. Wong and R. W. Brockett, “Systems with finite communication bandwidth constraints I,” *IEEE Trans. Autom. Contr.*, vol. 42, pp. 1294–9, 1997.

[36] ——, “Systems with finite communication bandwidth constraints II: stabilization with limited information feedback,” *IEEE Trans. Autom. Contr.*, vol. 44, pp. 1049–53, 1999.

[37] J. Hespanha, A. Ortega, and L. Vasudevan, “Towards the control of linear systems with minimum bit-rate,” in *Proc. 15th Int. Symp. Math. The. Netw. Sys. (MTNS)*, U. Notre Dame, USA, Aug 2002.

[38] J. Baillieul, “Feedback designs in information-based control,” in *Stochastic Theory and Control. Proceedings of a Workshop held in Lawrence, Kansas*, B. Pasik-Duncan, Ed. Springer, 2002, pp. 35–57.

[39] G. N. Nair and R. J. Evans, “Exponential stabilisability of finite-dimensional linear systems with limited data rates,” *Automatica*, vol. 39, pp. 585–93, Apr. 2003.

[40] S. Tatikonda and S. Mitter, “Control under communication constraints,” *IEEE Trans. Autom. Contr.*, vol. 49, no. 7, pp. 1056–68, July 2004.

[41] ——, “Control over noisy channels,” *IEEE Trans. Autom. Contr.*, vol. 49, no. 7, pp. 1196–201, July 2004.

[42] J. H. Braslavsky, R. H. Middleton, and J. S. Freudenberg, “Feedback stabilization over signal-to-noise ratio constrained channels,” *IEEE Trans. Autom. Contr.*, vol. 52, no. 8, pp. 1391–403, 2007.

[43] A. Sahai and S. Mitter, “The necessity and sufficiency of anytime capacity for stabilization of a linear system over a noisy communication link part 1: scalar systems,” *IEEE Trans. Info. Theory*, vol. 52, no. 8, pp. 3369–95, 2006.

[44] A. S. Matveev and A. V. Savkin, “Shannon zero error capacity in the problems of state estimation and stabilization via noisy communication channels,” *Int. Jour. Contr.*, vol. 80, pp. 241–55, 2007.