Towards a Non-Stochastic Information Theory

Anshuka Rangi  
University of California, San Diego  
Email: arangi@ucsd.edu

Massimo Franceschetti  
University of California, San Diego  
Email: massimo@ece.ucsd.edu

Abstract—The δ-mutual information between uncertain variables is introduced as a generalization of Nair’s non-stochastic information functional. Several properties of this new quantity are illustrated, and used to prove a channel coding theorem in a non-stochastic setting. Namely, it is shown that the largest δ-mutual information between a metric space and its δ-packing equals the (ε, δ)-capacity of the space. This notion of capacity generalizes the Kolmogorov ε-capacity to packing sets of overlap at most δ, and is a variation of a previous definition proposed by one of the authors. These results provide a framework for developing a non-stochastic information theory motivated by potential applications in control and learning theories. Compared to previous non-stochastic approaches, the theory admits the possibility of decoding errors as in Shannon’s probabilistic setting, while retaining its worst-case non-stochastic character.

Index Terms—Shannon capacity, Kolmogorov capacity, zero-error capacity, ε-capacity, (ε, δ)-capacity, mutual information, coding theorem.

I. INTRODUCTION

When Shannon laid the mathematical foundations of communication theory he embraced a probabilistic approach [1]. A tangible consequence of this choice is that in today’s communication systems performance is guaranteed in an average sense, or with high probability. Occasional violations from a specification are permitted, and cannot be avoided. This approach is well suited for consumer-oriented digital communication devices, where occasional loss of data packets is not critical. In contrast, in the context of control of safety-critical systems, error bounds must often be guaranteed at any time, not only on average. In this case, at each time step of the evolution of a dynamical system, sensor measurements are used by a controller to generate the next input, which is then fed back into the system. When this feedback loop is closed over a communication channel, occasional decoding errors can quickly drive the system out of control and lead to catastrophic failures. The emerging paradigm of cyber-physical systems (CPS), integrating computation, communication, and control on a single networked platform, makes these considerations particularly relevant [2]. These systems have a variety of applications, including health care, intelligent farming, transportation, security, and robotics. They typically employ a distributed network of sensors and actuators that interact with each other, and are monitored and controlled by a supervisory control and data acquisition (SCADA) system [3] via wireless communication links. In this setting, classical information theory has little role in providing non-stochastic guarantees of meeting the control objectives. On the other hand, information in some sense must be flowing across the network, and this observation motivates the need for a meaningful theory of information in a non-stochastic setting.

Another motivating example for developing a non-stochastic information theory arises in the context of learning theory, while studying the performance of classifiers with a rejection option [4]. For example, consider a medical doctor that makes a diagnosis based on some observed symptoms. The diagnosis is inherently uncertain, but this uncertainty cannot be described in a probabilistic framework. The doctor can shrink the uncertainty by rejecting the task of identifying the disease and ordering more tests. This reduces the margin of error of any diagnosis, but leads to higher costs. The doctor’s objective is then to correctly identify the disease from an ensemble, while keeping the cost contained within a certain level. The (non-stochastic) capacity of this system corresponds to the largest number of distinct diseases that can be reliably diagnosed by the doctor with a given margin of error, and at a given cost.

A third motivation for our work comes from the observation that most engineering systems operate in a regime well described by classical physics, and non-stochastic approaches can be better suited to describe information from first physical principles in this classical setting [5]. Information-theoretic results in a quantum setting can then be derived starting with a deterministic wave description and imposing quantum resolution constraints in a deterministic setting [6].

Coming back to control of CPS, there is a long tradition in control theory of treating noise disturbances as non-random perturbations with bounded magnitude, energy, or power. The main reason is that systems usually have mechanical or chemical components, in addition to electrical, whose dominant disturbances may not be governed by known (e.g. Gaussian) probability distributions. The typical approach is then to provide worst-case performance guarantees assuming a non-stochastic, bounded noise model. In this paper, we embrace this approach and cast it into an information-theoretic setting, considering worst-case communication errors in a non-stochastic setting.

The idea of adopting a non-stochastic approach to information theory is not new. A few years after introducing the notion of capacity of a communication system [1], Shannon introduced the zero-error capacity [7]. While the first notion corresponds to the largest rate of communication such that the probability of decoding error tends to zero, the second corresponds to the largest rate of communication such that the probability of decoding error equals zero. Both definitions of
capacity satisfy a coding theorem: Shannon’s channel coding theorem states that the capacity is the supremum of the mutual information between the input and the output of the channel [1]. In the context of control and estimation of dynamical systems, Nair introduced a non-stochastic mutual information functional and established an analogous coding theorem for the zero-error capacity in a non-stochastic setting [8]. Motivated by Shannon’s results, Kolmogorov introduced the deterministic notions of $\epsilon$-entropy and $\epsilon$-capacity in the context of functional spaces [9]. The $\epsilon$-capacity is defined as the logarithm base two of the packing number of the space, namely the logarithm of the maximum number of balls of radius $\epsilon$ that can be placed in the space without overlap. Determining this number is analogous to designing a codebook such that the distance between any two codewords is at least $2\epsilon$. In this way, any transmitted codeword that is subject to a perturbation of at most $\epsilon$ can be recovered at the receiver without error. The $\epsilon$-capacity also corresponds to the zero-error capacity of an additive channel having arbitrary, but bounded noise of support at most $\epsilon$. Lim and Franceschetti extended this concept introducing the $(\epsilon, \delta)$ capacity [10], defined as the logarithm base two of the largest number of balls of radius $\epsilon$ that can be placed in the space with average codeword overlap of at most $\delta$. In this setting, $\delta$ is a measure of the amount of error that can be tolerated when designing a codebook in a non-stochastic setting. Both the Kolmogorov capacity and its $(\epsilon, \delta)$ generalization rely on operational definitions, and neither of them has a corresponding information-theoretic characterization in terms of mutual information and associated coding theorem.

Our contributions are as follows. Following [10], we introduce a new notion of $(\epsilon, \delta)$-capacity, which is defined as the logarithm base two of the largest number of balls of radius $\epsilon$ that can be placed in the space such that the overlap between any two balls is at most $\delta$. A key point here is that instead of bounding the average overlap among all the balls as in [10], our definition requires to bound the overlap between every pair of balls. This guarantees a certain margin of error for every single transmitted codeword. Following [8], we then introduce a new notion of $\delta$-mutual information between uncertain variables. In contrast to [8], our definition considers the information revealed by one variable regarding the other with a given level of confidence, and relies on the notion of non-stochastic partial association between variables. We show that the largest $\delta$-mutual information, corresponding to the largest amount of non-stochastic information at level $\delta$ between a transmitted codeword and its received version, corrupted with noise of level at most $\epsilon$, is the $(\epsilon, \delta)$-capacity; thus establishing a channel coding theorem for functional spaces in a non-stochastic setting. For $\delta = 0$ our capacity definition reduces to the Kolmogorov capacity, viz. the zero-error capacity of an additive bounded noise channel, and our mutual information functional reduces to Nair’s one, thus providing the relevant mathematical definition of information in these settings.

The rest of the paper is organized as follows: Section II introduces uncertain variables; Section III introduces $\delta$-mutual information; Section IV establishes the relationship between $\delta$-mutual information and the $(\epsilon, \delta)$-capacity. Due to space constraints, proofs of all results are available in the open-access on-line repository ArXiv [11].

II. Uncertain Variables

We start by briefly reviewing the mathematical framework used in [8] to describe non-stochastic uncertain variables (UVs). An UV $X$ is a mapping from a sample space $\Omega$ to a set $\mathcal{X}$, i.e. for all $\omega \in \Omega$, we have $x = X(\omega) \in \mathcal{X}$. Given an UV $X$, the marginal range of $X$ is

$$\{X(\omega) : \omega \in \Omega\}. \quad (1)$$

The joint range of two UVs $X$ and $Y$ is

$$\{X(\omega), Y(\omega) : \omega \in \Omega\}. \quad (2)$$

Given an UV $Y$, the conditional range of $X$ given $Y = y$ is

$$\{X(\omega) : Y(\omega) = y, \omega \in \Omega\}, \quad (3)$$

and the conditional range of $X$ given $Y$ is

$$\{X|Y\} = \{X|y : y \in \{Y\}\}. \quad (4)$$

Thus, $\{X|Y\}$ denotes the uncertainty in $X$ given the realization of $Y$ and $\{X, Y\}$ represents the total joint uncertainty contributed by both $X$ and $Y$, i.e.

$$\{X, Y\} = \bigcup_{y \in \{Y\}} \{X|y\} \times \{y\}. \quad (5)$$

Finally, two UVs $X$ and $Y$ are independent if for all $x \in \{X\}$

$$\{Y|x\} = \{Y\}. \quad (6)$$

III. $\delta$-Mutual Information

A. Association and dissociation between UVs

We now introduce our notion of association and dissociation between UVs $X$ and $Y$. In the following definitions, we let $m_{\mathcal{X}}(\cdot)$ and $m_{\mathcal{Y}}(\cdot)$ denote the measures over the uncertainty sets $\mathcal{X}$ and $\mathcal{Y}$ respectively, and use the notation $\mathcal{A} \triangleright \delta$ to indicate that for all $a \in A$ we have $a \geq \delta$. We also assume that $y_i \neq y_j$ and $x_i \neq x_j$ whenever $i \neq j$.

**Definition 1.** The sets of association for UVs $X$ and $Y$ are

$$\mathcal{A}(X, Y) = \left\{ \frac{m_{\mathcal{X}}([X|y_1] \cap [X|y_2])}{m_{\mathcal{X}}([X])} : y_1, y_2 \in \{Y\} \text{ and } y_1 \neq y_2 \right\} \setminus \{0\}. \quad (7)$$

$$\mathcal{A}(Y, X) = \left\{ \frac{m_{\mathcal{Y}}([Y|x_1] \cap [Y|x_2])}{m_{\mathcal{Y}}([Y])} : x_1, x_2 \in \{X\} \text{ and } x_1 \neq x_2 \right\} \setminus \{0\}. \quad (8)$$

**Definition 2.** For any $\delta_1, \delta_2 \in [0, 1)$, UVs $X$ and $Y$ are disassociated at levels $(\delta_1, \delta_2)$ if the following inequalities hold:

$$\mathcal{A}(X, Y) \not\succ \delta_1, \quad (9)$$
\(\mathcal{A}(Y, X) \succ \delta_2,\) \hspace{1cm} (10)

and this case we write \((X, Y) \not\sim (\delta_1, \delta_2).\)

Intuitively, the levels of disassociation between two UVs represent lower bounds on the amount of residual uncertainty in each variable when the other is known. If \(X\) and \(Y\) are independent, then \(\mathcal{A}(X, Y)\) and \(\mathcal{A}(Y, X)\) contain only the element 1, and the variables are maximally disassociated. In this case, knowledge of \(X\) does not reduce the uncertainty of \(Y\), and vice versa. On the other hand, when no conditional ranges intersect, then \(\mathcal{A}(X, Y)\) and \(\mathcal{A}(Y, X)\) are empty. In the first case, \(X\) and \(Y\) are minimally disassociated and there is no residual uncertainty in \(Y\) given knowledge of \(X\). In the second case, there is no residual uncertainty in \(X\) given knowledge of \(Y\). Finally, in the case of partial disassociation, if there exist points \(y_1, y_2 \in [Y]\) such that \(m_X([X|y_1] \cap [X|y_2]) \neq 0\), then the measure of this intersection is at least a \(\delta_1\)-fraction of the total measure of the uncertainty set of \([X]\), indicating that the measure of uncertainty in \(X\) induces uncertainty in \(Y\) at least \(\delta_1\). The same reasoning holds for points \(x_1, x_2 \in [X]\) and level \(\delta_2\).

An analogous definition of association is given to provide upper bounds on the residual uncertainty of one random variable when the other is known.

**Definition 3.** For any \(\delta_1, \delta_2 \in [0, 1]\), we say that UVs \(X\) and \(Y\) are associated at levels \((\delta_1, \delta_2)\) if the following inequalities hold:

\[ \mathcal{A}(X, Y) \preceq \delta_1, \quad \mathcal{A}(Y, X) \preceq \delta_2, \] \hspace{1cm} (11)\hspace{1cm} (12)

and in this case we write \((X, Y) \not\mathcal{A} (\delta_1, \delta_2)\).

The following lemma provides necessary and sufficient conditions for association at a given level to hold. These conditions are stated for all points in \([Y]\) and \([X]\) rather than for points in \([Y|X]\) and \([X|Y]\). Similar global conditions do not hold for disassociation.

**Lemma 1.** \((X, Y) \not\mathcal{A} (\delta_1, \delta_2)\) if and only if for all \(y_1, y_2 \in [Y]\), we have

\[ \frac{m_X([X|y_1] \cap [X|y_2])}{m_X([X])} \leq \delta_1, \] \hspace{1cm} (13)

and for all \(x_1, x_2 \in [X]\), we have

\[ \frac{m_Y([Y|x_1] \cap [Y|x_2])}{m_Y([Y])} \leq \delta_2. \] \hspace{1cm} (14)

The following lemma provides a way to determine the levels of association and disassociation.

**Lemma 2.** \((X, Y) \not\mathcal{A} (\delta_1, \delta_2)\) if and only if the following inequalities hold:

\[ \min \mathcal{A}(X, Y) > \delta_1, \] \hspace{1cm} (15)

\[ \min \mathcal{A}(Y, X) > \delta_2. \] \hspace{1cm} (16)

Similarly, \((X, Y) \mathcal{A} (\delta_1, \delta_2)\) if and only if the following inequalities hold:

\[ \max \mathcal{A}(X, Y) \leq \delta_1, \] \hspace{1cm} (17)

\[ \max \mathcal{A}(Y, X) \leq \delta_2. \] \hspace{1cm} (18)

An immediate, yet important consequence the above lemma is that both association and disassociation at given levels \((\delta_1, \delta_2)\) cannot hold simultaneously.

**B. \(\delta\)-mutual information**

In order to introduce the notion of \(\delta\)-mutual information we need some additional definitions.

**Definition 4.** \(\delta\)-Connectedness and \(\delta\)-isolation.

- Points \(x_1, x_2 \in [X]\) are \(\delta\)-connected via \([X|Y]\) and are denoted by \(x_1 \leftrightarrow^\delta x_2\), if there exists a finite sequence \([X|y_i]\) such that \(x_1 \in [X|y_i], x_2 \in [X|y_N]\) and for all \(1 \leq i \leq N\), we have

\[ \frac{m_X([X|y_i] \cap [X|y_{i+1}])}{m_X([X])} > \delta. \] \hspace{1cm} (19)

If \(x_1 \leftrightarrow^\delta x_2\) and \(N = 1\), then we say that \(x_1\) and \(x_2\) are singly \(\delta\)-connected, i.e. there exists a \(y\) such that \(x_1, x_2 \in [X|y]\).

- A set \(\mathcal{A} \subseteq [X]\) is (singly) \(\delta\)-connected via \([X|Y]\) if every pair of points in the set is (singly) \(\delta\)-connected via \([X|Y]\).

- Two sets \(\mathcal{A}_1, \mathcal{A}_2 \subseteq [X]\) are \(\delta\)-isolated via \([X|Y]\) if no point in \(\mathcal{A}_1\) is \(\delta\)-connected to any point in \(\mathcal{A}_2\).

**Definition 5.** \(\delta\)-isolated partition and \(\delta\)-overlap family.

- An \([X|Y]\) \(\delta\)-isolated partition of \([X]\), denoted by \([X|Y]_\delta\), is a partition of \([X]\) such that any two sets in the partition are \(\delta\)-isolated via \([X|Y]\).

- An \([X|Y]\) \(\delta\)-overlap family of \([X]\), denoted by \([X|Y]_\delta^\ast\), is a largest family of sets covering \([X]\) such that each set in the family is \(\delta\)-connected and contains a \(\delta\)-connected set of the form \([X|y]\), there exists a set containing any two singly \(\delta\)-connected points, and the measure of overlap between any two sets in the family is at most \(\delta \cdot m_X([X])\).

The following theorem establishes some key properties of the \(\delta\)-overlap family \([X|Y]_\delta^\ast\).

**Theorem 3.** Suppose that \((X, Y) \not\mathcal{A} (\delta, \delta_2)\). Then, the following statements hold:

(i) There is a unique \(\delta\)-overlap family \([X|Y]_\delta^\ast\); (ii) The \(\delta\)-overlap family is the \(\delta\)-isolated partition with highest cardinality, namely for any \([X|Y]_\delta\), we have

\[ |X|Y|_\delta^\ast \leq |X|Y|_\delta | , \] \hspace{1cm} (20)

where the equality holds if and only if \([X|Y]_\delta = [X|Y]_\delta^\ast\).

**Theorem 4.** If \((X, Y) \mathcal{A} (\delta, \delta_2)\), then there exists a \(\delta\)-overlap family \([X|Y]_\delta^\ast\).
The stage is now set to define the $\delta$-mutual information in terms of the $\delta$-overlap family. The definition is analogous to the one in [8], extended here to a $\delta$-overlap family.

**Definition 6.** The $\delta$-mutual information between two UVs $X$ and $Y$ is

$$I_\delta(X,Y) = \log |[X|Y]|_\delta^\ast.$$  

(21)

Since in general we have $[X|Y] \neq [Y|X]$, one may reasonably suspect the definition of mutual information to be asymmetric in its arguments. We show in the next section that this is not the case and symmetry is retained, provided that when swapping $X$ with $Y$ one also rescales $\delta$ appropriately.

C. Taxicab symmetry of the mutual information

**Definition 7.** $(\delta_1, \delta_2)$-taxicab connectedness and $(\delta_1, \delta_2)$-taxicab isolation.

- Points $(x, y), (x', y') \in [X,Y]$ are $(\delta_1, \delta_2)$-taxicab connected, and are denoted by $(x, y) \overset{\delta_1, \delta_2}{\leftrightarrow} (x', y')$, if there exists a finite sequence $\{(x_i, y_i)\}_{i=1}^{N}$ of points in $[X,Y]$ such that $(x, y) = (x_1, y_1), (x', y') = (x_N, y_N)$ and for all $2 < i \leq N$, at least one of the following holds:

$$A_1 = \{x_i = x_{i-1} \text{ and } \frac{m_X([X|y_i] \cap [X|y_{i-1}])}{m_X([X])} > \delta_1\},$$

$$A_2 = \{y_i = y_{i-1} \text{ and } \frac{m_Y([X|x_i] \cap [Y|x_{i-1}])}{m_Y([Y])} > \delta_2\}.$$

- A set $\mathcal{S} \subseteq [X,Y]$ is (singly) $(\delta_1, \delta_2)$-taxicab connected if every pair of points in the set is (singly) $(\delta_1, \delta_2)$-taxicab connected in $[X,Y]$.

- Two sets $\mathcal{S}_1, \mathcal{S}_2 \subseteq [X,Y]$ are $(\delta_1, \delta_2)$-taxicab isolated if no point in $\mathcal{S}_1$ is $(\delta_1, \delta_2)$-taxicab connected to any point in $\mathcal{S}_2$.

**Theorem 5.** Suppose that $(X,Y) \overset{\delta_1}{\leftrightarrow} (\delta_1, \delta_2)$. Then, the following statements hold:

- Points $(x, y)$ and $(x', y')$ in $[X,Y]$ are $(\delta_1, \delta_2)$-taxicab connected if and only if $x$ and $x'$ are $\delta_1$-connected via $[X|Y]$. 
- A set $\mathcal{S} \subseteq [X,Y]$ is $(\delta_1, \delta_2)$-taxicab connected if and only if $\mathcal{S}^+ \subseteq [X]$ is $\delta_1$-connected via $[X|Y]$ where $\mathcal{S}^+$ is a projection of the set $\mathcal{S}$ on the x-axis.
- Two sets $\mathcal{S}_1, \mathcal{S}_2 \subseteq [X,Y]$ are $(\delta_1, \delta_2)$-taxicab isolated if and only if $\mathcal{S}_1^+, \mathcal{S}_2^+ \subseteq [X]$ are $\delta_1$-isolated via $[X|Y]$.

The following theorem establishes some key properties of the $(\delta_1, \delta_2)$-taxicab family $[X,Y]^{\ast}_{\delta_1, \delta_2}$.

**Theorem 6.** Suppose that $(X,Y) \overset{\delta_1}{\leftrightarrow} (\delta_1, \delta_2)$. Then, the following statements hold:

- There is a unique $(\delta_1, \delta_2)$-taxicab family $[X,Y]^{\ast}_{\delta_1, \delta_2}$.
- The $(\delta_1, \delta_2)$-taxicab family is the $(\delta_1, \delta_2)$-taxicab isolated partition with highest cardinality, namely for any $(\delta_1, \delta_2)$-taxicab isolated partition $[X,Y]^{\ast}_{\delta_1, \delta_2}$, we have

$$|[X,Y]^{\ast}_{\delta_1, \delta_2}| \leq |[X,Y]^{\ast}_{\delta_1, \delta_2}|,$$

(22)

where the equality holds if and only if $\|[X,Y]^{\ast}_{\delta_1, \delta_2} = |[X,Y]^{\ast}_{\delta_1, \delta_2}|$.

**Theorem 7.** Suppose that $(X,Y) \overset{\delta_1}{\leftrightarrow} (\delta_1, \delta_2)$. Then, the following statements hold:

- There exists a $(\delta_1, \delta_2)$-taxicab family $[X,Y]^{\ast}_{\delta_1, \delta_2}$.
- The cardinality of the $(\delta_1, \delta_2)$-taxicab family is the same as the one of the $\delta_1$-overlap family, namely

$$|[X,Y]^{\ast}_{\delta_1, \delta_2}| = |[X,Y]^{\ast}_{\delta_1}|.$$

Combining the results in Theorem 5, Theorem 6 and Theorem 7 with Definition 6, we have the following important corollary, showing the symmetry in the mutual information.

**Corollary 7.1.**

$$\log |[X,Y]^{\ast}_{\delta_1, \delta_2}| = I_{\delta_1}(X,Y) = I_{\delta_2}(Y,X).$$

(23)

IV. $(\epsilon, \delta)$-CAPACITY

Let $X$ be a normed metric space such that for all $x \in X$ we have $\|x\| \leq 1$. This normalization is for convenience of notation and all results can be easily be extended to metric spaces of bounded norm. Let $Y \subseteq X$ be a discrete set of points in the space. Any point $y \in Y$ represents a codeword that can be selected at the transmitter, sent over the channel, and received with perturbation at most $\epsilon$. Namely, for any $y \in Y$ we receive a point $x \in X$ that is within the set

$$S^X_{\epsilon}(y) = \{x \in X : \|x - y\| \leq \epsilon\}.$$

(24)

It should be clear that transmitted codewords can be decoded correctly as long as the corresponding uncertainty sets do not overlap. This can be done by simply associating the received codeword to the point in the codebook that is closest to it. For any $y_1, y_2 \in Y$, we now introduce the error measure

$$e_{\epsilon}(y_1, y_2) = \frac{m_X(S^X_{\epsilon}(y_1) \cap S^X_{\epsilon}(y_2))}{m_X(X)},$$

(25)
where $m_X(.)$ is a finite measure over the metric space $X$. We also let $V_X^X$ be the ball of radius $\epsilon$ in the metric space $X$.

**Definition 9.** For any $0 < \epsilon \leq 1$, $0 \leq \delta < m_X(V_X^X)$, a codebook $Y \subseteq X$ is $(\epsilon, \delta)$-distinguishable if for all $y_1, y_2 \in Y$, we have $e_\epsilon(y_1, y_2) \leq \delta$.

**Definition 10.** For any $0 < \epsilon \leq 1$, $0 \leq \delta < m_X(V_X^X)$ and normed metric space $X$, the $(\epsilon, \delta)$-capacity of $X$ is

$$C^\delta_\epsilon = \sup_{Y \in \mathcal{Y}^\delta} \log |Y|,$$

(26)

where $\mathcal{Y}^\delta$ is the set of all possible $(\epsilon, \delta)$-distinguishable codebooks of $X$.

We point out that in the above definitions we restrict $\delta < m_X(V_X^X)$ to rule out the trivial case when the decoding error is greater than the error introduced by the channel. Clearly, when $\delta \geq m_X(V_X^X)$ the $(\epsilon, \delta)$-capacity would be infinity.

Points in our metric space can be considered the image through an UV map with sample space $\Omega = X$, set $\mathcal{X} = X$ and one-to-one mapping from $\Omega$ to $\mathcal{X}$. Since the codebook $Y \subseteq X$, then the elements of $Y$ can also be considered the image of an UV map. For our $\epsilon$-perturbation channel, these UVs are such that for all $y \in Y$ and $x \in X$, we have

$$[X|y] = \{x \in X : ||x - y|| \leq \epsilon\} = S_X^X(y),$$

(27)

$$[Y|x] = \{y \in Y : ||x - y|| \leq \epsilon\} = S_Y^X(x).$$

(28)

To measure the level of association and disassociation between UVs $X$ and $Y$, we use the measure $m_X(.)$ defined over the uncertainty set $\mathcal{X}$ and let $m_X(.)$ be the cardinality over $\mathcal{X}$, and we introduce the feasible set

$$\mathcal{F}_\delta = \{Y \subseteq X : (X, Y) \not\leftrightarrow (\delta, 0) \text{ or } (X, Y) \not\leftrightarrow (\delta, 1)\},$$

(29)

representing the set of codebooks $Y$ that can achieve $(\delta, 0)$ levels of disassociation or $(\delta, 1)$ levels of association with $X$.

In our channel model, this feasible set also depends on the $\epsilon$-perturbation through (27) and (28). We can now state the nonstochastic channel coding theorem for our $\epsilon$-perturbed channel.

**Theorem 8.** For any normed metric space $X$, $0 < \epsilon \leq 1$, $0 \leq \delta < m_X(V_X^X)$, codebook $Y \subseteq X$, and $\epsilon$-perturbation channel satisfying (27) and (28), we have

$$C^\delta_\epsilon = \sup_{Y \in \mathcal{F}_\delta} I_\delta(X, Y).$$

(30)

**Proof.** The outline of the proof is briefly described here. First, we show that there exists a $\tilde{\delta} \leq \delta$, such that the set $\mathcal{F}_{\tilde{\delta}}$ is not empty, so that the supremum is well defined. Second, we show that

$$\sup_{Y \in \mathcal{F}_{\tilde{\delta}}} I_\delta(X, Y) \leq C^\delta_\epsilon,$$

(31)

Finally, we show the existence of a $\tilde{\delta} \leq \delta$ and a codebook $Y \in \mathcal{F}_{\tilde{\delta}}$, such that $I_{\tilde{\delta}}(X, Y) = C^\delta_\epsilon$.

Theorem 8 characterizes the capacity as the supremum of the mutual information over all codebooks in the feasible set. The following corollary shows that the same characterization is obtained if we optimize over all codebooks in the space.

**Corollary 8.1.** The $(\epsilon, \delta)$-capacity in (30) can also be written as

$$C^\delta_\epsilon = \sup_{Y \subseteq X : \epsilon, \delta} I_\delta(X, Y).$$

(32)

**V. Final Remarks**

According to Theorem 8, rather than optimizing over all codebooks as stated in Corollary 8.1, a capacity achieving codebook can also be found within the smaller class $F_{\tilde{\delta}} : \delta \leq \delta$, representing all feasible sets with error at most $\delta$. In Shannon’s formulation, a capacity achieving codebook has vanishing probability of error, and thus the mutual information does not depend on the probability of error. In our setting, the decoding error $\delta$ is a parameter of the model, and our information functional depends on it. The Shannon capacity depends on the type of channel, namely it is a function of the probability distribution over the output given the channel input. Likewise, in our setting the $(\epsilon, \delta)$-capacity is a function of the worst-case uncertainty $\epsilon$ that characterizes the channel. Finally, we make some considerations with respect to previous results in the literature. For $\delta = 0$, Theorem 8 recovers Nair’s coding theorem for the zero-error capacity [8, Theorem 4.1] in the case of an additive $\epsilon$-noise channel. The $(\epsilon, \delta)$-capacity considered in [10] defines the set of $(\epsilon, \delta)$-distinguishable codewords such that the average overlap among all codewords is at most $\delta$. In contrast, with our definition we have that the overlap for each pair of codewords is at most $\delta$. In the full paper, we prove the upper bound:

$$C^\delta_\epsilon \leq \tilde{C}^\delta_\epsilon / 2,$$

(33)

where $\tilde{C}^\delta_\epsilon$ is the capacity considered in [10], and the constant $\alpha = m_X(V_X^X)/m_X([X])$.

**References**

[1] C. E. Shannon, “A mathematical theory of communication,” Bell system technical journal, vol. 27(3), pp. 379–423, 1948.

[2] K.-D. Kim and P. R. Kumar, “Cyber-physical systems: A perspective at the centennial,” Proceedings of the IEEE, vol. 100, no. Centennial Issue, pp. 1287–1308, 2012.

[3] S. A. Boyer, SCADA: supervisory control and data acquisition. International Society of Automation, 2009.

[4] C. Cortes, G. DeSalvo, and M. Mohri, “Learning with rejection,” in International Conference on Algorithmic Learning Theory. Springer, 2016, pp. 67–82.

[5] M. Franceschetti, Wave Theory of Information. Cambridge University Press, 2018.

[6] ———, “Quantum limits on the entropy of bandlimited radiation,” Journal of Statistical Physics, vol. 196(2), pp. 374–394, 2017.

[7] C. Shannon, “The zero error capacity of a noisy channel,” IRE Transactions on Information Theory, vol. 2(3), pp. 8–19, 1956.

[8] G. N. Nair, “A nonstochastic information theory for communication and state estimation,” IEEE Transactions on automatic control, vol. 58(6), pp. 1497–1510, 2013.

[9] V. M. Tkhomov and A. N. Kolmogorov, “c-entropy and c-capacity of sets in functional spaces,” Uspekhi Matematicheskikh Nauk, vol. 14(2), pp. 3–86, 1959.

[10] T. J. Lim and M. Franceschetti, “Information without rolling dice,” IEEE Transactions on Information Theory, vol. 63(3), pp. 1349–1363, 2017.

[11] A. Rangi and M. Franceschetti, “Towards a non-stochastic information theory.” arXiv preprint arXiv:1904.11632, 2019.