Cumulative Residual Entropy: A New Measure of Information

Murali Rao, Yunmei Chen, Baba C. Vemuri, *Fellow, IEEE*, and Fei Wang

***Abstract*—In this paper, we use the cumulative distribution of a random variable to define its information content and thereby develop an alternative measure of uncertainty that extends Shannon entropy to random variables with continuous distributions. We call this measure cumulative residual entropy (CRE). The salient features of CRE are as follows: 1) it is more general than the Shannon entropy in that its definition is valid in the continuous and discrete domains, 2) it possesses more general mathematical properties than the Shannon entropy, and 3) it can be easily computed from sample data and these computations aymptotically converge to the true values. The properties of CRE and a precise formula relating CRE and Shannon entropy are given in the paper. Finally, we present some applications of CRE to reliability engineering and computer vision.**

***Index Terms*—Distribution, entropy, information measurement.**

I. Introduction

In [15], Shannon proposed a measure of uncertainty in a discrete distribution based on the Boltzmann entropy of classical statistical mechanics. He called it the entropy. The Shannon entropy of a discrete distribution F is defined by

$$H(F) = -\sum_i p_i \log p_i \tag{1}$$

where p_i's are the probabilities computed from the distribution F.

With this, he opened up a new branch of mathematics with far-reaching applications in many areas such as Financial Analysis [16], Data Compression [14], Statistics [9], and Information Theory [4].

This measure of uncertainty has many important properties which agree with our intuitive notion of randomness. We mention three. 1) It is always positive. 2) It vanishes if and only if it is a certain event. 3) Entropy is increased by the addition of an independent component, and decreased by conditioning.

However, extension of this notion to continuous distribution poses some challenges. A straightforward extension of the discrete case to continuous distributions F with density f, called differential entropy, reads

$$H(F) = -\int f(x) \log f(x) dx. \tag{2}$$

This definition raises the following concerns.

1) It is only defined for distributions with densities. For example, there is no definition of entropy for a mixture density comprised of a combination of Guassians and delta functions.
2) The entropy of a discrete distribution is always positive, while the differential entropy of a continuous variable may take any value on the extended real line.
3) It is "inconsistent" in the sense that the differential entropy of a uniform distribution in an interval of length a is $\log a$, which is zero if $a = 1$, negative if $a < 1$, and positive if $a > 1$.
4) The entropy of a discrete distribution and the differential entropy of a continuous variable are decreased by conditioning. Moreover, if X and Y are discrete (continuous) random variables, and the conditional entropy (differential entropy) of X given Y equals the entropy (differential entropy) of X, then X and Y are independent. Also, the conditional entropy of the discrete variable X given Y is zero, if and only if X is a function of Y, but the vanishing of the conditional differential entropy of X given Y does not imply that X is a function of Y.
5) Use of empirical distributions in approximations is of great value in practical applications. However, it is impossible, in general, to approximate the differential entropy of a continuous variable using the entropy of empirical distributions.
6) Consider the following situation: Suppose X and Y are two discrete random variables, with X taking on values $\{1, 2, 3, 4, 5, 6\}$, each with a probability $1/6$ and Y taking on values $\{1, 2, 3, 4, 5, 10^6\}$ again each with probability $1/6$. The information content measured in these two random variables using Shannon entropy is the same, i.e., Shannon entropy does not bring out any differences between these two cases. However, if the two random variables represented distinct payoff schemes in a game of chance, the information content in the two random variables would be considered as being dramatically different. Nevertheless, Shannon entropy fails to make any distinction whatsoever between them.

For additional discussion on some of these issues the reader is referred to [6].

In this work, we propose an alternative measure of uncertainty in a random variable X and call it the cumulative residual entropy (CRE) of X. The main objective of our study is to extend Shannon entropy to random variables with continuous distributions. The concept we propose overcomes the problems men-

Manuscript received February 24, 2003; revised December 15, 2003. This work was supported in part by the National Institutes of Health under Grant NS42075.

M. Rao and Y. Chen are with the Department of Mathematics, University of Florida, Gainesville, FL 32611-8205 USA.

B. C. Vemuri and F. Wang are with the Department of Computer and Infomation Sciences and Engineering, University of Florida, Gainesville, FL 32611 USA.

Communicated by R. W. Yeung, Associate Editor for Shannon Theory.

Digital Object Identifier 10.1109/TIT.2004.828057

tioned above, while retaining many of the important properties of Shannon entropy. For instance, both are decreased by conditioning, while increased by independent addition. They both obey the data processing inequality, etc. However, the differential entropy does not have the following important properties of CRE.

1) CRE has consistent definitions in both the continuous and discrete domains;
2) CRE is always nonnegative;
3) CRE can be easily computed from sample data and these computations asymptotically converge to the true values.
4) The conditional CRE (defined in Section III) of X given Y is zero, if and only if X is a function of Y.

The basic idea in our definition is to replace the density function with the cumulative distribution in Shannon's definition (2). The distribution function is more regular than the density function, because the density is computed as the derivative of the distribution. Moreover, in practice what is of interest and/or measurable is the distribution function. For example, if the random variable is the life span of a machine, then the event of interest is not whether the life span equals t, but rather whether the life span exceeds t. Our definition also preserves the well-established principle that the logarithm of the probability of an event should represent the information content in the event. The discussions about the properties of CRE in the following sections, we trust, are convincing enough for further development of the concept of CRE.

The remainder of the paper is organized as follows: Section II contains the definition of CRE and a description of its properties in the form of several theorems. In Section III, we present the definition of the conditional CRE and its properties. This is followed by a discussion and derivation of the relationship between CRE and the Shannon (differential) entropy in the form of theorems, in Section IV. In Section V, we present a discussion of empirical CRE and its relation to the continuous case. Section VI contains a discussion on applications of CRE in reliability engineering and computer vision. Finally, in Section VII, we present conclusions.

II. Cumulative Residual Entropy: A New Measure of Information

In this section, we define an alternate measure of uncertainty in a random variable and then derive some properties about this new measurement.

Definition: Let X be a random vector in $\mathcal{R}^N$, we define the CRE of X by

$$\mathcal{E}(X) = -\int_{\mathcal{R}_+^N} P(|X| > \lambda) \log P(|X| > \lambda)\, \mathrm{d}\lambda \tag{3}$$

where $X = (X_1, X_2, \ldots, X_N)$, $\lambda = (\lambda_1, \ldots. \lambda_N)$, and $|X| > \lambda$ means $|X_i| > \lambda_i$ and

$$\mathcal{R}_+^N = \Big(x_i \in \mathcal{R}^N; x_i \geq 0\Big).$$

Now we give a few examples.

Example 1: (CRE of the uniform distribution) Consider a general uniform distribution with the density function

$$p(x) = \begin{cases} \frac{1}{a}, & 0 \leq x \leq a \\ 0, & o.w. \end{cases} \tag{4}$$

Then its CRE is computed as follows:

$$\begin{aligned}\mathcal{E}(X) &= -\int_0^a P(|X| > x) \log P(|X| > x)\, \mathrm{d}x \\ &= -\int_0^a \left(1 - \frac{x}{a}\right) \log\left(1 - \frac{x}{a}\right) \mathrm{d}x \\ &= \frac{1}{4} a.\end{aligned} \tag{5}$$

Example 2: (CRE of the exponential distribution)

The exponential distribution with mean $1/\lambda$ has the density function

$$p(x) = \lambda e^{-\lambda x}. \tag{6}$$

Correspondingly, the CRE of the exponential distribution is

$$\begin{aligned}\mathcal{E}(x) &= -\int_0^\infty e^{-\lambda x} \log e^{-\lambda x}\, \mathrm{d}t \\ &= \int_0^\infty \lambda t e^{-\lambda x}\, \mathrm{d}t \\ &= \frac{1}{\lambda}.\end{aligned} \tag{7}$$

Example 3: (CRE of the Gaussian distribution)

The Gaussian probability density function is

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-m)^2}{2\sigma^2}\right] \tag{8}$$

where m is the mean and σ^2 is the variance.

The cumulative distribution function is

$$F(x) = 1 - \mathrm{erfc}\left(\frac{x-m}{\sigma}\right) \tag{9}$$

where erfc is the error function

$$\mathrm{erfc}(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty \exp(-t^2/2)\, \mathrm{d}t.$$

Then the CRE of the Gaussian distribution is

$$\mathcal{E}(x) = -\int_0^\infty \mathrm{erfc}\left(\frac{x-m}{\sigma}\right) \log\left[\mathrm{erfc}\left(\frac{x-m}{\sigma}\right)\right] \mathrm{d}x. \tag{10}$$

Next we will prove some properties of CRE.

Using the convexity of $x \log x$, it is easy to see CRE is a concave function of distribution.

Theorem 1: $\mathcal{E}(X) < \infty$ if for all i and some $p > N$, $E[|X_i|^p] < \infty$.

Proof: For simplicity we write the proof in steps.

Step 1) Using Hölder's inequality ([1, p. 125])

$$E\left[\prod_{i=1}^N f_i\right] \leq \prod_{i=1}^N E[f_i^N]^{1/N}$$

we see that for sets $A_1, \dots, A_n$

$$P[A_1 \cap A_2 \cap \dots \cap A_N] \le E\left[\prod_{i=1}^{N} 1_{A_i}\right] \le \prod_{i=1}^{N} P(A_i)^{1/N}. \quad (11)$$

Step 2) It is not difficult to see that for each $0 \le \alpha \le 1$

$$|x \log x| \le e^{-1}\frac{x^\alpha}{1-\alpha}, \qquad 0 \le x \le 1. \quad (12)$$

Step 3) From Step 2 and Step1, for any $0 < \alpha < 1$ and any i

$$\begin{aligned} P\big[|X_i| > x_i, 1 \le i \le N\big]\Big|\log P\big[|X_i| > x_i, 1 \le i \le N\big]\Big| &\le \frac{e^{-1}}{1-\alpha} P\big[|X_i| > x_i, 1 \le i \le N\big]^{\alpha} \\ &\le \frac{e^{-1}}{1-\alpha} \prod_{i=1}^{N} P[|X_i| > x_i]^{\alpha/N}. \quad (13) \end{aligned}$$

Integrating both sides of (13), on $\mathcal{R}_+^N$, we get

$$\begin{aligned} \mathcal{E}(X) &\le \frac{e^{-1}}{1-\alpha} \int_{\mathcal{R}_+^N} \prod_{i=1}^{N} P[|X_i| > x_i]^{\alpha/N}\, \mathrm{d}x \\ &= \frac{e^{-1}}{1-\alpha} \prod_{i=1}^{N} \left\{ \int_0^\infty P[|X_i| > x_i]^{\alpha/N}\, \mathrm{d}x_i \right\}. \quad (14) \end{aligned}$$

Step 4) We have for any positive random variable Y

$$\begin{aligned} \int_0^\infty P[Y > y]^{\alpha/N} dy &= \int_0^1 P[Y > y]^{\alpha/N}\, \mathrm{d}y + \int_1^\infty P[Y > y]^{\alpha/N} \mathrm{d}y \\ &\le 1 + \int_1^\infty \{\frac{1}{y^p} E[Y^p]\}^{\alpha/N}\, \mathrm{d}y. \quad (15) \end{aligned}$$

The second inequality on the right-hand side (RHS) above follows from the Markov inequality. The last integral is finite if $\frac{p\alpha}{N} > 1$, i.e., if $p < \frac{N}{\alpha}$. For $p > N$, we can choose $\alpha < 1$ to satisfy $p > \frac{N}{\alpha}$. Then the conclusion of the theorem follows from (14) and (15). □

The traditional Shannon entropy of a sum of independent variables is larger than that of either. We have analogously the following theorem.

Theorem 2: For any nonnegative and independent variables X and Y

$$\max\big(\mathcal{E}(X), \mathcal{E}(Y)\big) \le \mathcal{E}(X+Y).$$

Proof: Since X and Y are independent

$$P[X+Y > t] = \int \mathrm{d}F_Y(a) P[X > t-a]$$

where F_Y is the cumulative distribution function of Y. Using Jensen's inequality

$$P[X+Y > t] \log P[X+Y > t] \le \int \mathrm{d}F_Y(a) P[X > t-a] \log P[X > t-a].$$

Integrating both sides with respect to t from 0 to ∞

$$\begin{aligned} -\mathcal{E}(X+Y) &\le \int \mathrm{d}F_Y(a) \int_0^\infty P[X > t-a] \log P[X > t-a]\, \mathrm{d}t \\ &= \int \mathrm{d}F_Y(a) \int_a^\infty P[X > t-a] \log P[X > t-a]\, \mathrm{d}t \\ &= -\int \mathrm{d}F_Y(a) \mathcal{E}(X) = -\mathcal{E}(X) \end{aligned}$$

where in the first equality we used that for $t \le a$ $P[X > t-a] = 1$, and in the second one we change variables in the inner integral. □

The next theorem shows one of the salient features of CRE. In the discrete case, Shannon entropy is always nonnegative, and equals zero if and only if the random variable is a certain event. However, this is not valid for the Shannon entropy in the continuous case as defined in (1). In contrast, in this regard CRE does not differentiate between discrete and continuous cases, as shown by the following theorem.

Theorem 3: $\mathcal{E}(X) \ge 0$ and equality holds if and only if $P[|X| = \lambda] = 1$ for some vector λ, i.e.. $|X_i| = \lambda_i$ with probability 1.

Proof: Now $x \log x = 0$ if and only if $x = 0$ or 1. Thus, $\mathcal{E}(X) = 0$ implies $P[|X| > \lambda] = 0$ or 1 for almost all λ. If for all λ, $P[|X| > \lambda] = 0$, then $P[|X| = 0] = 1$. Now we consider the case that for some λ, $P[|X| > \lambda] = 1$. Note that if λ and μ satisfy $P[|X| > \lambda] = P[|X| > \mu] = 1$, then also $P[|X| > \lambda \vee \mu] = 1$, where $\lambda \vee \mu$ denotes the vector whose coordinates are maxima of coordinates of λ and μ. Then

$$\lambda_c = \max_{\lambda \in \Lambda} \lambda$$

satisfies $p[|X| = \lambda_c] = 1$, where $\Lambda = \{\lambda | P[|X| > \lambda] = 1\}$, and the maximum of $\lambda \in \Lambda$ is a vector whose coordinates are the maxima of the coordinates of all $\lambda \in \Lambda$. □

Theorem 4: If X_i are independent, then

$$\mathcal{E}(X) = \sum_i \left(\prod_{i \ne j} E(|X_j|) \right) \mathcal{E}(X_i).$$

We omit the proof.

No analog of the property that follows is valid for the Shannon entropy. Weak convergence [11] is a fundamentally important notion in probability theory.

Theorem 5: (Weak Convergence). Let the random vectors X_k converge in distribution to the random vector X

$$\lim_{k \to \infty} E[\varphi(X_k)] = E[\varphi(X)] \quad (16)$$

for all bounded continuous functions φ on $\mathcal{R}^N$. If all the X_k are bound in L^p for some $p > N$, then

$$\lim \mathcal{E}(X_k) = \mathcal{E}(X). \quad (17)$$

Proof: If X_k converges to X in distribution, it is known that for almost all $x \in \mathcal{R}_+^N$ [10]

$$\lim_k P[|X_k| > x] = P[|X| > x].$$

In particular, for almost all x

$$\lim_k P[|X_k| > x] \log P[|X_k| > x] = P[|X| > x] \log P[|X| > x] \quad (18)$$

and using (13)

$$\Big| P[|X_k| > x] \log P[|X_k| > x] \Big| \leq \frac{e^{-1}}{1-\alpha} \prod_i P[|X_{k_i}| > x_i]^{\alpha/N}$$

and for each i and k

$$P[|X_{k_i}| > x_i] \leq \mathbf{1}_{[0,1]}(x_i) + x_i^{-p} \mathbf{1}_{[1,\infty)}(x_i) E(|X|^p_{k_i}). \quad (19)$$

Thus, if $\frac{p\alpha}{N} > 1$, $|P[|X_k| > x] \log P[|X_k| > x]|$ is bounded by an integrable function. The use of dominated convergence theorem completes the proof. □

III. Conditional CRE

In this section, we will recall formal definitions of certain concepts from probability theory [5] that will be used subsequently.

Definition: A sigma field $\mathcal{F}$ is a class of subsets containing the empty set and closed under compliments and countable unions.

Given a sigma field $\mathcal{F}$ and a random variable X with finite expectation, we can define a random variable Y which is measurable with respect to $\mathcal{F}$ and is called the conditional expectation of X given $\mathcal{F}$ [5], denoted by $E(X/\mathcal{F})$. It should be noted that the measure functions of X and Y are the same.

Definition: Given a random $\mathcal{R}^N$ vector X and a σ-field $\mathcal{F}$, we define the conditional CRE: $\mathcal{E}(X|\mathcal{F})$ by

$$\mathcal{E}(X|\mathcal{F}) = -\int_{\mathcal{R}^N_+} P(|X| > x|\mathcal{F}) \log P(|X| > x|\mathcal{F})\, \mathrm{d}x \quad (20)$$

where $P(|X| > x|\mathcal{F})$ denotes the conditional expectation of the indicator function, namely, $E(\mathbf{1}_{(|X|>x)}/\mathcal{F})$. Note that $\mathcal{E}(X|\mathcal{F})$ is a random variable measurable with respect to $\mathcal{F}$. For example, if $\mathcal{F}$ is the σ-field generated by a random variable Y then, $\mathcal{E}(X/\mathcal{F}) = g(Y)$ where

$$g(Y) = -\int P(|X| > x/Y = y) \log(P(|X| > x/Y = y))\, \mathrm{d}x.$$

When $\mathcal{F}$ is the trivial field

$$\mathcal{E}(X|\mathcal{F}) = \mathcal{E}(X).$$

The following proposition says that the conditional CRE has the "super-martingale property."

Proposition 1: Let $X \in L^p$ for some $p > N$, then for σ-fields $\mathcal{G} \subset \mathcal{F}$

$$E[\mathcal{E}(X|\mathcal{F})|\mathcal{G}] \leq \mathcal{E}(X|\mathcal{G}). \quad (21)$$

Proof: The proof follows by applying Jensen's inequality [13] for the convex function $x \log x$.

In words, Proposition 1 states that conditioning decreases CRE. The same result holds for the Shannon entropy. In particular

$$H(Z|X,Y) \leq H(Z|X)$$

for any random variables X, Y, and Z. A simple consequence is the data processing inequality.

If $X \to Y \to Z$ is a Markov chain, i.e., if the conditional distribution of Z given (X, Y) equals that given Y, then $H(Z|Y) \leq H(Z|X)$. This is so because $H(Z|X,Y) = H(Z|Y)$ from Markov property.

For the same reason we have the data processing inequality for CRE. If $X \to Y \to Z$ is Markov

$$E[\mathcal{E}(Z|Y)] \leq E[\mathcal{E}(Z|X)].$$

Indeed, $\mathcal{E}(Z|Y) = \mathcal{E}(Z|X,Y)$, and from (21)

$$E[\mathcal{E}(Z|X,Y)|X] \leq \mathcal{E}(Z|X).$$

Theorem 6: Let $X \in L^p$ for some $p > N_\tau$ and $\mathcal{F}$ a σ-field, then

$$E[\mathcal{E}(X|\mathcal{F})] = 0 \quad \text{iff } X \text{ is } \mathcal{F}\text{-measurable}. \quad (22)$$

More generally, if

$$A = \Big\{\mathcal{E}(X|\mathcal{F}) = 0\Big\} \quad (23)$$

then $\mathbf{1}_A|X|$ is $\mathcal{F}$-measurable, where 1_A is the indicator function of the set A, and $\mathbf{1}_A|X| = |X|$ on the set A and zero elsewhere. In particular, for random vectors X and Y

$$\mathcal{E}(X|Y) = 0 \quad \text{iff } |X| \text{ is a function f } Y.$$

Note that $E[\mathcal{E}(X|\mathcal{F})]$ is a scalar whereas $\mathcal{E}(X/Y)$ is a function of Y. Also, if Y is a generator of $\mathcal{F}$ then $\mathcal{E}(X/\mathcal{F}) = \mathcal{E}(X/Y)$.

Proof: Let $\Phi(x) = x \log x$, note that for any positive random variable Z and any set A, $\mathbf{1}_A \Phi(Z) = \Phi(\mathbf{1}_A Z)$. In particular, if $A \in \mathcal{F}$

$$\mathbf{1}_A \Phi(E(Z|\mathcal{F})) = \Phi(\mathbf{1}_A E(Z|\mathcal{F})).$$

Therefore, proving (22) implies the more general statement (23). To prove (22) we will use the following easily verified fact:

$$\left.\begin{array}{l}\textbf{For any set } A,\ A \textbf{ is } \mathcal{F}\textbf{-measurable}\\ \text{iff } E(\mathbf{1}_A|\mathcal{F}) \textbf{ is } 0\text{–}1 \textbf{ valued.}\end{array}\right\} \quad (24)$$

Now suppose $|X|$ is $\mathcal{F}$-measurable. Then, $P(|X| > x|\mathcal{F}) = \mathbf{1}_A$, where $A = \{|X| > x\}$. Moreover, $\Phi(\mathbf{1}_A) = 0$, since $\Phi(\mathbf{1}_B) = 0$ for all sets B. Conversely, if $\mathcal{E}(X|\mathcal{F}) = 0$, then for almost all x

$$\Phi(P(|X| > x|\mathcal{F})) = 0$$

i.e., $P(|X| > x|\mathcal{F})$ is $0 - 1$ valued. Using (18) we get that for almost all $x \in \mathcal{R}^N_+$ the set $(|X| > x)$ belongs to $\mathcal{F}$. Otherwise stated, $|X|$ is $\mathcal{F}$-measurable. □

Theorem 7: For any X and σ-field $\mathcal{F}$

$$E[\mathcal{E}(X|\mathcal{F})] \leq \mathcal{E}(X). \quad (25)$$

Equality holds iff X is independent of $\mathcal{F}$. Where X is said to be independent of the σ-field $\mathcal{F}$ means X is independent of every random variable measurable with respect to $\mathcal{F}$.

Proof: The inequality (25) follows from (21) by taking $\mathcal{G}$ to be the trivial field. Now we prove the necessary and sufficient condition for the equality. First, it is clear from the definition that if $|X|$ is independent of $\mathcal{F}$ the equality in (21) holds. Conversely, suppose that there is equality in (25). By Jensen's inequality for conditional expectations

$$E\big[P[|X|>x|\mathcal{F}]\log P[|X|>x|\mathcal{F}]\big] \geq P\big[|X|>x\big]\log P\big[|X|>x\big] \quad (26)$$

for all x. Integrating both sides of (26) with respect to x and using (25), we see that equality holds in (26) for almost all x. Note the following fact, which can be proved using Taylor with remainder.

Fact: For any random variable X and strictly convex φ (i.e., $\varphi'' > 0$)

$$E[\varphi(X)] = \varphi(E[X])$$

implies $X = E(X)$ almost surely.

Using this fact and strict convexity of $x\log x$, we get from (26)

$$P\big[|X|>x|\mathcal{F}\big] = P\big[|X|>x\big]$$

for a.e. x, i.e., $|X|$ is independent of $\mathcal{F}$. □

IV. CRE and Differential Entropy

We show below that CRE dominates the differential entropy (which may exist when X has density).

Definition: The differential entropy $\mathcal{H}(X)$ of a random variable X with density f is defined as

$$\mathcal{H}(X) = -E[\log f] = -\int f(x)\log f(x)\mathrm{d}x.$$

Theorem 8: Let $X \geq 0$ have density f, Then

$$\mathcal{E}(X) \geq C\exp\{\mathcal{H}(X)\} \quad (27)$$

where

$$C = \exp\left\{\int_0^1 \log(x|\log x|)\,\mathrm{d}x\right\} \cong 0.2065.$$

Proof: Let $G(x) = P[X > x] = \int_x^\infty f(u)\,\mathrm{d}u$ using the log-sum inequality

$$\int_0^\infty f(x)\log\frac{f(x)}{G(x)|\log G(x)|}\mathrm{d}x \geq \log\frac{1}{\int_0^\infty G(x)|\log G(x)|\mathrm{d}x} = \log\frac{1}{\mathcal{E}(x)}. \quad (28)$$

If $\int G(x)\log G(x)$ is infinite then the proof trivially follows. The left-hand side (LHS) in (28) equals

$$-\mathcal{H}(X) - \int_0^\infty f(x)\log(G(x)|\log G(x)|)\mathrm{d}x$$

so that

$$\mathcal{H}(X) + \int_0^\infty f(x)\log(G(x)|\log G(x)|)\,\mathrm{d}x \leq \log\mathcal{E}(X). \quad (29)$$

Finally, a change of variable gives

$$\int_0^\infty f(x)\log\Big(G(x)|\log G(x)|\Big)\,\mathrm{d}x = \int_0^1 \log\Big(x|\log x|\Big)\,\mathrm{d}x.$$

Therefore, we have from (29) the following equation:

$$\mathcal{H}(X) + \int_0^1 \log\Big(x|\log x|\Big)\mathrm{d}x \leq \log\mathcal{E}(X). \quad (30)$$

Exponentiating both sides of (30), we get (27). □

More generally, we have the following proposition.

Proposition 2: Let $X \geq 0$ be a random variable, $\mathcal{F}$ a σ-field, and assume X has conditional density relative to $\mathcal{F}$, i.e., X has a conditional probability density function given $Y = y$. Then

$$\mathcal{E}(X|\mathcal{F}) \geq C\exp\{\mathcal{H}(X|\mathcal{F})\}$$

where

$$C = \exp\left\{\int_0^1 \log(x|\log x|)\,\mathrm{d}x\right\} \cong 0.2065$$

and $\mathcal{H}(X|\mathcal{F})$ is the conditional Shannon entropy defined as

$$-\mathcal{H}(X|\mathcal{F}) = \int f(x|\mathcal{F})\log f(x|\mathcal{F})\,\mathrm{d}x$$

with $f(x|\mathcal{F})$ the conditional density

$$P[X > t|\mathcal{F}] = \int_t^\infty f(x|\mathcal{F})\,\mathrm{d}x.$$

We omit the proof which is almost exactly the same as that of the previous theorem.

Definition: The joint entropy $\mathcal{H}(X,Y)$ of two random variable X and Y in Shannon's sense is defined as

$$\mathcal{H}(X,Y) = \mathcal{H}(X) + \mathcal{H}(Y/X). \quad (31)$$

Analogously, we define the **Cross CRE** (CCRE) by

$$\mathcal{E}(X,Y) = \mathcal{E}(X) + E[\mathcal{E}(Y/X)]. \quad (32)$$

Note $\mathcal{H}(X,Y)$ is symmetric, $\mathcal{E}(X,Y)$ need not be. If we want a symmetric measure, we can define **CCRE** as $1/2(\mathcal{E}(X,Y) + \mathcal{E}(Y,X))$.

Now using the last proposition we have the following proposition.

Proposition 3:

$$\mathcal{E}(X,Y) \geq 2C\exp\left(\frac{\mathcal{H}(X,Y)}{2}\right). \quad (33)$$

Proof: Using Proposition 2, the convexity of e^x, and Jensen inequality

$$\begin{aligned}\mathcal{E}(X,Y) &= \mathcal{E}(X) + E[\mathcal{E}(Y/X)]\\ &\geq C\exp\{\mathcal{H}(X)\} + C\exp\{E[\mathcal{H}(X/Y)]\}\\ &\geq 2C\exp\left\{\frac{\mathcal{H}(X)+E[\mathcal{H}(X/Y)]}{2}\right\}\\ &= 2C\exp\left\{\frac{\mathcal{H}(X,Y)}{2}\right\}.\end{aligned}$$

In the second inequality we used $2\exp\{\frac{x+y}{2}\} \leq e^x + e^y$. □

A. Analytical Form Relating CRE and Differential Entropy

We now present a theorem which provides the exact formula relating CRE and the differential entropy. Before stating our theorem, we first make some remarks. If F is a continuous distribution and U is uniformly distributed, then $F^{-1}(U)$ has the distribution F. Here $F^{-1}(U)$ is defined as the largest y such that $F(y) = x$ $(0 \leq x < 1)$, and $F^{-1}(1)$ is the smallest y such that $F(y) = 1$. Further, if X is a random variable with continuous distribution function F_X, then $F_X(X)$ is uniformly distributed. Thus, if X is any random variable with continuous distribution function F_X and if F is any continuous distribution function, then $F^{-1}(F_X(X))$ has the distribution F.

Proposition 4: Suppose X is a nonnegative random variable with continuous distribution. Then, there exists a function $\phi = \phi_X$ (a formula for ϕ is given in the proof) such that the Shannon entropy $\mathcal{H}(Y)$ of $Y = \phi(X)$ is related to $\mathcal{E}(X)$ by

$$\mathcal{H}(Y) = \frac{\mathcal{E}(X)}{E(X)} - \frac{1}{E(X)} \log\left(\frac{1}{E(X)}\right).$$

Proof: Let F be the distribution with density

$$\frac{P(X > t)}{E(X)}.$$

By the remarks preceding the proposition, simply choose $\phi = F^{-1}(F_X(x))$ or, equivalently, $Y = F^{-1}(F_X(X))$. Now use definitions of $\mathcal{E}(X)$ and $\mathcal{H}(Y)$. □

V. CRE and Empirical CRE

Let $X_1, X_2, \ldots, X_n$ be positive and independent and identically distributed (i.i.d.) with distribution F. Let F_n be the empirical distribution of the n-sample $X_1, \ldots, X_n$: put $\frac{1}{n}$ at each of the sample points $X_1, \ldots, X_n$, then F_n is the cumulative distribution function (CDF) of this mass distribution. Writing $G_n(x) = 1 - F_n(x)$, the CRE of the empirical distribution F_n is

$$-\mathcal{E}(F_n) = \int_0^\infty G_n(x) \log G_n(x)\,\mathrm{d}x. \qquad (34)$$

This is a random variable. A well-known theorem of Glivento–Cantelli (see [11, p. 18]) asserts that

$$\sup_x |F_n(x) - F(x)| = \sup_x |G_n(x) - G(x)| \to 0 \qquad (35)$$

almost surely as $n \to \infty$. Using this result, we prove the following theorem, which provides a method of computing the CRE of appropriate random variables using the CRE of empirical data.

Theorem 9: For any random X in L^p for some $p > 1$, the empirical CRE converges to the CRE of X, i.e.,

$$\mathcal{E}(F_n) \to \mathcal{E}(F) \text{ almost surely.}$$

Proof: We will use dominated convergence theorem to prove this. By dominated convergence theorem, the integral of $G_n(x) \log G_n(x)$ on any finite interval converges to that of $G(x) \log G(x)$. Therefore, we need only show that as $n \to \infty$

$$\left| \int_1^\infty G_n(x) \log G_n(x) \mathrm{d}x - \int_1^\infty G(x) \log G(x) \mathrm{d}x \right| \to 0$$

almost surely.

Recall that

$$G_n(x) = P_n(Z > x)$$

where P_n is the probability distribution on $\mathcal{R}_+$ assigning mass $\frac{1}{n}$ to each of the sample points $X_1, \ldots, X_n$, and Z is the random variable with $Z(y) = y$, $y \in \mathcal{R}_+$

It follows that

$$x^p G_n(x) \leq E_n[Z^p] = \frac{1}{n} \sum_1^n X_i^p \qquad (36)$$

where E_n is expectation relative to P_n. By the strong law [11]

$$\frac{1}{n} \sum_1^n X_i^p \to E[X_1^p] \qquad \text{almost surely.} \qquad (37)$$

In particular

$$\sup_n \frac{1}{n} \sum_1^n X_i^p < \infty \qquad \text{almost surely.}$$

The combination of (36) and (37) leads to

$$G_n(x) \leq x^{-p} \left(\sup_n \frac{1}{n} \sum_1^n X_i^p \right), \qquad \text{in } [1, \infty).$$

Now by applying the dominated convergence theorem and using (35), we proved the theorem.

VI. Applications

In this section, we will present some applications of our new measure of information, the CRE, introduced in earlier sections. We start with the characterization of distributions, in particular, the exponential distribution which is the only distribution with the "memoryless" property and is widely prevalent and fundamental in applications of queuing theory [8] and in reliability engineering [7]. This is followed by applications of CRE in computer vision, specifically to the image alignment/registration problem.

A. Application to Reliability Engineering

The exponential distribution is very widely used in reliability applications. The exponential distribution is used to model data with a constant failure rate (also described by the hazard function). We will not dwell on the various uses of the exponential distribution in reliability engineering here but present a way of characterizing this distribution using CRE.

There are many characterizations of the exponential distribution in literature. The one we give below is based on CRE. Since the exponential distribution has no "memory," intuition suggests that it should have maximum CRE. This is borne out by the following theorem.

Theorem 10: Let $X > 0$ be a random variable. Then

$$\text{CRE}(X) \leq \text{CRE}(X(\lambda))$$

where $X(\lambda)$ is the exponentially distributed random variable with mean $\lambda = E(X^2)/2E(X)$.

Proof: By log-sum inequality

$$\int_0^\infty P(X > t) \log(e^{t/\lambda} P(X > t)) dt \geq E(X) \log(E(X)/\lambda). \qquad (38)$$

Expanding the LHS of (38) we get

$$\int_0^\infty P(X > t)\log P(X > t)\,\mathrm{d}t + \int_0^\infty P(X > t)t/\lambda\,\mathrm{d}t \geq E(X)\log(E(X)/\lambda).$$

Since $\int_0^\infty P(X > t)t\,\mathrm{d}t = E(X^2)/2$, we get from above

$$\int_0^\infty P(X > t)\log P(X > t)\,\mathrm{d}t \geq -E(X^2)/2\lambda + E(X)\log(E(X)/\lambda). \quad (39)$$

This is valid for all λ positive. The maximum of the RHS of (39) is attained when $\lambda = E(X^2)/2E(X)$.

Substituting this value of λ into (39) we get

$$\begin{aligned}\int_0^\infty P(X > t)\log P(X > t)\,\mathrm{d}t &\geq -E(X) + E(X)\log 2E(X)^2/E(X^2)\\ &\geq -E(X) + E(X)[1 - E(X^2)/2E(X)^2]\\ &= -E(X^2)/2E(X). \qquad (40)\end{aligned}$$

Here in the second inequality we used $\log x \geq 1 - 1/x$. Finally, we note that the CRE of an exponential variable with mean μ is μ.

Thus, one way to characterize the exponential distribution is as the distribution that maximizes the CRE given the coefficient of variation of the distribution.

B. Computer Vision Applications

Matching two or more images under varying conditions—illumination, pose, acquisition parameters, etc.—is ubiquitous in computer vision, medical imaging, geographical information systems, etc. In the past several years, information-theoretic measures have been very widely used in defining cost functions to be optimized in achieving a match. An example problem common to all the aforementioned areas is the image registration problem. This problem may be defined as follows: Given two images $I_1(\boldsymbol{x})$ and $I_2(\boldsymbol{x}')$, with $\boldsymbol{x} = (x, y)^t$ and $\boldsymbol{x}' = (x', y')^t$ being the coordinates of each of the two images, respectively. Let $\boldsymbol{x}' = T \circ \boldsymbol{x}$, where T is an unknown transformation/registration between the two coordinate systems that needs to be estimated. The matching/registration problem is then posed as maximize/minimize the cost function $M(I_1(T(\boldsymbol{x})), I_2(\boldsymbol{x}))$ over all the appropriate (rigid, affine, or nonlinear nonparametric) class of transformations T.

The most popular cost function presented in literature has been the mutual information (MI), $\mathcal{I}(I_1(T(\boldsymbol{x})), I_2(\boldsymbol{x}))$, between the two given images which is maximized over the class of transformations mentioned earlier [18], [3]. The key strength of the MI-based methods is that the intensity functions in the two images being matched need not be related in any known way, e.g., the source and target images could be the images of the same scene under different lighting conditions; e.g., view of a room under artificial lighting and another under natural lighting. Since MI does not compare intensity values at corresponding points in the two images being registered, it is well suited for registering images of the same scene taken from two different sensing devices, e.g., video and infrared images. The MI between the source and the target images that are to be aligned was maximized using a stochastic analog of the gradient descent method in [18] and other optimization methods such as the Powells method in [3] and a multiresolution scheme in [17]. In recent times, most of the effort on the MI-based methods has been focussed on coping with nonrigid deformations between the source and target multimodal image data sets [12], [2]. Many variants of this technique have been published in literature and we refer the reader to a recent survey [20] for these.

In order to use the concept of CRE for the image alignment problem, we defined a quantity called the CCRE earlier in Section IV which is recalled here for convenience (also see [19])

$$\mathcal{C}(X, Y) = \mathcal{E}(X) - E[\mathcal{E}(Y/X)]. \quad (41)$$

Note that $\mathcal{I}(X, Y)$ is symmetric but $\mathcal{C}(X, Y)$ need not be. We can define a symmetrized version of CCRE by adding $\mathcal{E}(Y) - E[\mathcal{E}(X/Y)$ to $\mathcal{C}(X, Y)$ and pre-multiplying it by a factor of $\frac{1}{2}$. It is easy to show that the symmetrized CCRE is nonnegative. In all our image alignment experiments to date (see [19]), we have used the nonsymmetric CCRE as it was sufficient to yield the desired image registration results. The key experimental contribution in this context is that we empirically show superior performance of CCRE under low signal-to-noise ratio (SNR) and also depict its larger capture range with regards to the convergence to the optimal parameterized transformation [19]. We will now present one example depicting the superior performance of CCRE (in the context of higher amount of tolerance to noise in the data) over the MI-based registration that is widely used in computer vision and medical imaging fields. We refer the interested reader to [19] for a more detailed account of the image alignment applications.

The first experiment involves registering a magnetic resonance (MR) image of a human brain against a computerized tomographic (CT) image of the same brain. We choose the MR image as the source, the target image in this case is generated by applying a known misalignment (rigid transformation) to the CT image. The source and target image pair along with the result of estimated transformation using CCRE applied to the source with an overlay of the target edge map are shown in Fig. 1. As evident, the registration is quite accurate from a visual inspection. For quantitative assessment of accuracy of the estimated registration, we refer the reader to [19]. The summary result presented in [19], in this context, is that CCRE is at least as accurate as the competing methods, namely, the MI and normalized MI (NMI) based methods.

Using an aerial image of a city as our source image (see Fig. 2(a)), we generated the target image (Fig. 2(b)) by applying a fixed rigid motion. We conduct this experiment by varying the amount of Gaussian noise added and then for each instance of the added noise, we register the two images using CCRE, MI, and NMI. We expect that all the schemes are going to fail at some level of noise. By comparing the noise magnitude of the failure point, we can show the degree to which these methods are tolerant. We choose the fixed motion to be 10° rotation, and 5-pixel translation in both x and y direction. The numerical scheme we used to achieve the optimization of the cost functions (CCRE, MI, and NMI) to determine the estimated registrations is the sequential quadratic programming (SQP) technique. Table I shows the registration results obtained using optimization of the three cost functions. From the table, we observe that

TABLE I
COMPARISON OF THE REGISTRATION RESULTS BETWEEN CCRE-, MI-, AND NMI-BASED ALGORITHMS FOR A FIXED SYNTHETIC MOTION UNDER VARYING NOISE CONDITIONS. THE TRUE MOTION IS $(10°, 5, 5)$. FAILED INDICATES DIVERGENCE OF THE NUMERICAL OPTIMIZER WHEN USING THE SAME INITIAL GUESS AND PARAMETER SETTINGS

σ	CCRE	MI	NMI
10	9.998 5.016 4.996	9.993 4.999 5.007	10.002 5.256 5.235
15	9.998 5.077 5.005	0 6.003 − 3.000	10.132 5.046 5.998
19	9.998 5.006 5.001	FAILED	0 − 15.890 19.222
30	9.998 5.256 5.235		FAILED
59	10.027 5.124 4.995		
60	0 − 3.003 0		
61	FAILED		

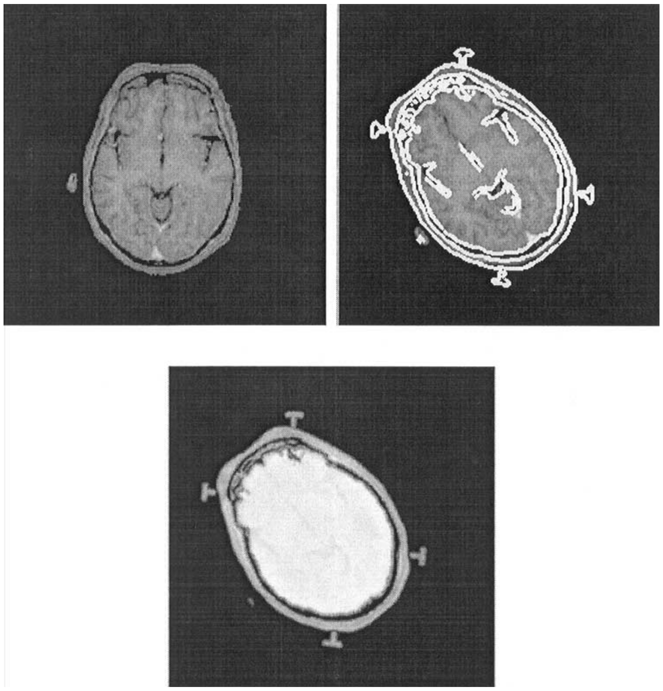

Fig. 1. Rigid motion example. Left: the MR (source) image; right: synthetically transformed (with a rigid motion) CT (source) image. Middle: overlay of the target edge map on the transformed source image obtained by applying the CCRE estimate of the rigid motion.

(a) (b)

Fig. 2. (a) An aerial image of a city. (b) Rotated and translated version of (a).

the MI fails when the standard deviation of the noise is increased to 15. It is slightly better for NMI, which fails at 19, while CCRE is tolerant until 60, a significant difference when compared to the traditional MI and NMI methods. This experiment depicts that CCRE has more noise immunity than both MI and the normalized MI. The *key idea* in our definition of CRE is to use the cumulative distribution in place of the density function as was done in Shannon's definition of differential entropy. The distribution function is more regular because it is defined in an integral form unlike the density function, which is computed as the derivative of the distribution. This is the reason why CCRE is more robust than MI or NMI in the presence of noise in the input image data sets being registered.

VII. SUMMARY

In this paper, we presented a novel measure of information which is based on the cumulative distribution of a random variable and is more general than the well-known Shannon's entropy. This definition is consistent in that it is valid in both discrete as well as continuous domains. We call this measure of information the cumulative residual entropy (CRE). We presented several theorems and propositions for CRE, some of which parallel those for the Shannon's entropy and others that are more general. The key advantages of CRE over Shannon's entropy were outlined as well. We also presented some applications of CRE, specifically, to reliability engineering and computer vision. We believe that this is just a scratch on the surface and envision that there will be many others in the near future. The results presented here are by no means comprehensive but hopefully will pave the way for re-examining the popular notion of entropy in a different, more general and rigorous setting.

REFERENCES

[1] D. P. Bertsekas, *Nonlinear Programming*. Nashua, NH: Athena Scientific.

[2] C. Chefd'Hotel, G. Hermosillo, and O. Faugeras, "A variational approach to multi-modal image matching," in *Proc. IEEE Workshop on Very Large Scale Integration* , Vancouver, BC, Canada, 2001, pp. 21–28.

[3] A. Collignon *et al.*, "Automated multimodality image registration using information theory," in *Proc. Int. Conf. Information Processing in Medical Imaging*, 1995, pp. 263–274.

[4] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[5] K. L. Chung, *A Course in Probability Theory*, 2nd ed. New York: Academic, 1974.

[6] G. Jumarie, *Relative Information*. New York: Springer-Verlag, 1990, pp. 55–57.

[7] D. Kececioglu, *Rliability Engineering Handbook*. Englewood Cliffs, NJ: Prentice-Hall, 1991, vol. 1.

[8] L. Kleinrock, *Queueing Systems, Volume I: Theory*. New York: Wiley-Interscience, 1975.

[9] S. Kullback, *Information Theory and Statistics*. New York: Wiely, 1959.

[10] D. Pollard, *Convergence of Stochastic Processes*. New York: Springer-Verlag, 2000, p. 46.

[11] B. L. S. P. Rao, *Asymptotic Theory of Statistical Inference*. New York: Wiley, 1987, p. 14.

[12] D. Ruckert, C. Hayes, C. Studholme, M. Leacha, and D. Hawkes, "Non-rigid registration of breast MRI using MI," in *Proc. Conf. Medical Image Computing and Computer Assisted Intervention*, 1998.

[13] W. Rudin, *Real and Complex Analysis*. New York: McGraw-Hill, 1987, p. 62.

[14] D. Salomon, *Data Compression*. New York: Springer-Verlag, 1998.

[15] C. E. Shannon, "The mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, 1948.

[16] W. F. Sharpe, *Investments*. Englewood Cliffs, NJ: Prentice-Hall, 1985.

[17] C. Studholme *et al.*, "An overlap invariant entropy measure of 3D medical image alignment," *Pattern Recogn.*, vol. 32, pp. 71–86, 1999.

[18] P. A. Viola and W. M. Wells, "Alignment by maximization of mutual information," in *Proc. 5th Int. Conf. Computer Vision*, 1995, pp. 16–23.

[19] F. E. Wang, B. C. Vemuri, M. Rao, and Y. Chen, "A new and robust information theoretic measure and its application to image alignment," in *Proc. Int. Conf. Information Processing in Medical Imaging*, Ambleside, U.K., 2003, pp. 388–400.

[20] B. Zitova and J. Flusser, "Image registration, a survey," *Image and Vision Computing*, vol. 21, pp. 977–1000, 2003.