Entropy for Singular Distributions

Georg Pichler, Günther Koliander, Erwin Riegler, and Franz Hlawatsch
Institute of Telecommunications, Vienna University of Technology, Vienna, Austria

***Abstract*—Entropy and differential entropy are important quantities in information theory. A tractable extension to singular random variables (which are neither discrete nor continuous) has not been available so far. Here, we propose such an extension for the practically relevant class of singular probability measures that are supported on a lower-dimensional subset of Euclidean space. We show that our entropy transforms in a natural manner under Lipschitz functions and that it conveys useful expressions of the mutual information. Potential applications of the proposed entropy definition include capacity calculations for the vector interference channel, compressed sensing in a probabilistic setting, and capacity bounds for block-fading channel models.**

***Keywords*—mutual information, information entropy, singular measures, rectifiable sets, information measures.**

I. Introduction

The entropy of discrete random variables and the differential entropy of continuous random variables are basic and important quantities in information theory [1]. Applications include the calculation of channel capacity [1, Ch. 8] and of rate-distortion functions [1, Ch. 10]. Unfortunately, a tractable entropy definition for *singular* random variables, which are neither discrete nor continuous, does not appear to be available. In particular, the entropy definition proposed in [2] is quite general and also applies to singular random variables; however, this definition is difficult to evaluate and manipulate and thus has not been used as a tool in information-theoretic problems.

A tractable definition of entropy for singular random variables is desirable due to a number of information-theoretic problems involving singular random variables. Examples include the following:

- For the vector interference channel, a singular input distribution has to be used to achieve full degrees of freedom [3].
- In a probabilistic formulation of compressed sensing, the underlying source distribution is singular [4]. While in [4] the information dimension [2] of the source was considered, an entropy term might lead to a more detailed analysis.
- In block-fading channel models, two different kinds of singular distributions arise: the optimal input distribution is singular in some settings [5, Ch. 6], and the noiseless output distribution is singular except for special cases [6].

In this paper, we propose a tractable definition of entropy for the broad class of singular random variables that are supported on lower-dimensional subsets of Euclidean space. More specifically, we extend the definition of entropy to random variables (or vectors) that are distributed according to a *rectifiable probability measure*, which means that the random variable takes its values on a lower-dimensional subset with probability one. Our definition encompasses as special cases the entropy of a discrete random variable and the differential entropy of a continuous random variable. Furthermore, we prove that our entropy transforms similarly to differential entropy under one-to-one Lipschitz mappings. Finally, we define joint and conditional entropy and show that the mutual information of two random variables that are distributed according to a rectifiable measure can be expressed in terms of (joint, conditional) entropy in the usual ways.

The rest of this paper is organized as follows. Rectifiable sets and measures are defined in Section II. In Section III, we define entropy for rectifiable measures, discuss connections to the entropy of discrete random variables and differential entropy, and derive a transformation property under Lipschitz mappings. In Section IV, we introduce joint and conditional entropy and demonstrate relations with mutual information.

This work was supported by the WWTF under grants ICT10-066 (NOWIRE) and ICT12-054 (TINCOIN) and by the FWF under grant S10603 (Statistical Inference) within the National Research Network SISE.

II. Rectifiable Sets and Measures

The notion of rectifiable sets [7, §3.2.14] is convenient for defining the subsets of Euclidean space on which the considered random variables take values with probability one. Because rectifiability is not defined consistently in the literature, we provide the definition we will use.

Definition 1: Let $M \in \mathbb{N}$. For $m \in \{1, \ldots, M\}$, a set $E \subseteq \mathbb{R}^M$ is said to be *B-countably m-rectifiable* if E equals the union of a family of Borel sets $\{E_i\}_{i\in\mathcal{J}}$ with $\mathcal{J} \subseteq \mathbb{N}$, where for each $i \in \mathcal{J}$ there exist a bounded set $A_i \subseteq \mathbb{R}^m$ and a Lipschitz function $f_i \colon A_i \to \mathbb{R}^M$ with $f_i(A_i) = E_i$. A set E is said to be *B-countably 0-rectifiable* (case $m = 0$) if it is countable.

B-countable m-rectifiability is a stronger notion than countable m-rectifiability as defined in [7, §3.2.14(2)], because it additionally requires the sets E_i, $i \in \mathcal{J}$ to be Borel. This restriction turns out to be useful for our considerations.

Example: An m-dimensional C^1 submanifold $E \subseteq \mathbb{R}^M$ is B-countably m-rectifiable. Indeed, by the Lindelöf property of Euclidean space [8, Th. 30.3], countably many C^1 embeddings $\phi_i \colon A_i \to E$, $i \in \mathcal{J}$ with $A_i \subseteq \mathbb{R}^m$ suffice to cover E. We may also assume that the sets A_i are compact so that the sets $\phi_i(A_i)$ are also compact and, hence, Borel. Furthermore, a C^1 embedding ϕ_i is locally Lipschitz by [9, Prop. 2.2 in Ch. 6] and, thus, $E = \bigcup_{i\in\mathcal{J}} \phi_i(A_i)$ is B-countably m-rectifiable.

Definition 2: Let $\mathcal{H}^m$ denote the m-dimensional Hausdorff measure on $\mathbb{R}^M$ [7, §2.10.2(1)]. A Borel probability measure μ on $\mathbb{R}^M$ is said to be *B-countably m-rectifiable* if there exists a B-countably m-rectifiable set E such that μ is absolutely continuous with respect to (w.r.t.) $\mathcal{H}^m|_E$, in symbols $\mu \ll \mathcal{H}^m|_E$. The set E is called a *support of* μ.

Note that $\mu(E^c) = 0$. The support E of a B-countably m-rectifiable measure μ is not unique. For example, adding an arbitrary countable set results in another support.

Since $\mathcal{H}^0$ equals the counting measure [7, §2.10.2(1)] and $\mathcal{H}^M = \lambda^M$, where λ^M denotes the M-dimensional Lebesgue measure [7, Th. 2.10.35], it can be easily shown that Definition 2 encompasses discrete ($m = 0$) and continuous ($m = M$) probability measures.

III. Entropy

We will use the definition of relative entropy from [10, Sec. 2.4] in a generalized form where the reference measure is not necessarily a probability measure. Let μ be a probability measure and ν a σ-finite measure on the same measurable space, with $\mu \ll \nu$. The *relative entropy of μ w.r.t. ν* is defined as

$$h_\nu(\mu) \triangleq -\int \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right) \mathrm{d}\mu = -\int \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right) \mathrm{d}\nu \quad (1)$$

where $\frac{\mathrm{d}\mu}{\mathrm{d}\nu}$ denotes the Radon-Nikodym (RN) derivative. Furthermore, if μ_X is the distribution of a random variable X, the *relative entropy of X w.r.t. ν* is defined as $h_\nu(\mu_X)$.

A. Definition of Entropy

We now apply (1) with $\nu = \mathcal{H}^m|_E$ to a B-countably m-rectifiable probability measure μ.

Definition 3: The entropy of a B-countably m-rectifiable probability measure μ is defined as

$$h(\mu) \triangleq h_{\mathcal{H}^m|_E}(\mu) = -\int \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\mathcal{H}^m|_E}\right) \mathrm{d}\mu \quad (2)$$

$$= -\int \frac{\mathrm{d}\mu}{\mathrm{d}\mathcal{H}^m|_E} \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\mathcal{H}^m|_E}\right) \mathrm{d}\mathcal{H}^m|_E \quad (3)$$

where E is an arbitrary support of μ (cf. Definition 2).

To show that this entropy is well-defined, we have to prove that $\mathcal{H}^m|_E$ is σ-finite (to ensure that the RN derivative exists) and that (2) is independent of the particular choice of the support E. By [11, Prop. 2.4.7], the image $f_i(A_i)$ of a bounded set $A_i \subseteq \mathbb{R}^m$ under a Lipschitz mapping f_i has finite $\mathcal{H}^m$ measure. Thus, the set E can be covered by a countable number of sets of finite $\mathcal{H}^m$ measure (cf. Definition 1). This implies that $\mathcal{H}^m|_E$ is σ-finite. To see that (2) is independent of E, assume that E' is another support of μ. Then $\widetilde{E} \triangleq E \cap E'$ is also a support of μ (cf. Definition 2). By the uniqueness of the RN derivative, $\frac{\mathrm{d}\mu}{\mathrm{d}\mathcal{H}^m|_E}(x) = \frac{\mathrm{d}\mu}{\mathrm{d}\mathcal{H}^m|_{\widetilde{E}}}(x)$ holds for all $x \in A$ for some Borel set $A \subseteq \mathbb{R}^M$ with $\mu(A^c) = 0$ (briefly written as μ-a.e.), because $E \setminus \widetilde{E}$ is a μ-null set. Using (2), this implies $h_{\mathcal{H}^m|_E}(\mu) = h_{\mathcal{H}^m|_{\widetilde{E}}}(\mu)$. Analogously, we obtain $h_{\mathcal{H}^m|_{E'}}(\mu) = h_{\mathcal{H}^m|_{\widetilde{E}}}(\mu)$, which proves that $h_{\mathcal{H}^m|_E}(\mu) = h_{\mathcal{H}^m|_{E'}}(\mu)$.

We mention two important special cases of Definition 3. If μ is a probability measure that is absolutely continuous w.r.t. λ^M (i.e., $m = M$), then (3) yields $h(\mu) = -\int f \log(f) \mathrm{d}\lambda^M$ (with $f = \frac{\mathrm{d}\mu}{\mathrm{d}\lambda^M}$ the density of μ), i.e., the differential entropy of the continuous random variable with distribution μ. If μ is a discrete probability measure (i.e., $m = 0$), then (3) becomes a sum over the elements of the countable set E, $h(\mu) = -\sum_{x\in E} \mu(\{x\}) \log(\mu(\{x\}))$, i.e., the entropy of the discrete random variable with distribution μ. Thus, Definition 3 generalizes both the entropy of a discrete random variable and the differential entropy of a continuous random variable.

Remark: An AEP analogous to [1, Sec. 8.2] holds for the entropy in (2). The proof is essentially the same, but (using the notation of [1, Sec. 8.2]) $\mathrm{Vol}(A)$ is replaced by the Hausdorff measure $\mathcal{H}^{mn}(A)$.

B. Expression in Terms of Hausdorff Density

The definition of relative entropy in (1) is based on the RN theorem [12, Th. 4.1.1], which is nonconstructive in that, in general, it does not provide an explicit expression of the RN derivative $\frac{\mathrm{d}\mu}{\mathrm{d}\mathcal{H}^m|_E}$. However, if μ is a B-countably m-rectifiable measure, then the Besicovitch differentiation theorem [13, Th. 2.22] can be used to show that $\frac{\mathrm{d}\mu}{\mathrm{d}\mathcal{H}^m|_E}$ equals the Hausdorff density of μ. The m-dimensional Hausdorff density of a Borel measure μ on $\mathbb{R}^M$ is defined as [11, Def. 2.2.1]

$$\Theta^m(\mu, x) \triangleq \lim_{r\downarrow 0} \frac{\mu(\overline{B_r(x)})}{\alpha(m)\, r^m} \quad (4)$$

if this limit exists. Here, $\overline{B_r(x)}$ denotes the closed ball of radius $r \geq 0$ centered at $x \in \mathbb{R}^M$ and $\alpha(m)$ denotes the volume of the unit sphere in $\mathbb{R}^m$, i.e., $\alpha(m) \triangleq \left(\Gamma(\frac{1}{2})\right)^m / \Gamma(\frac{m}{2}+1)$. The following theorem states that in our setting, the Hausdorff density is well defined $\mathcal{H}^m|_E$-a.e. and equal to the desired RN derivative $\frac{\mathrm{d}\mu}{\mathrm{d}\mathcal{H}^m|_E}$.

Theorem 4: For a B-countably m-rectifiable probability measure μ with support E,

$$\frac{\mathrm{d}\mu}{\mathrm{d}\mathcal{H}^m|_E} = \Theta^m(\mu, \cdot\,), \quad \mathcal{H}^m|_E\text{-a.e.} \quad (5)$$

and, thus,

$$h(\mu) = -\int \Theta^m(\mu, \cdot\,) \log(\Theta^m(\mu, \cdot\,))\, \mathrm{d}\mathcal{H}^m|_E\,.$$

Proof: We will prove this theorem for the case where $E = \bigcup_{i\in\mathcal{J}} E_i$ in Definition 1 degenerates to $E = E_1$, i.e., there is just one Lipschitz function $f\colon A \to \mathbb{R}^M$ with a bounded Borel set $A \subseteq \mathbb{R}^m$ such that $f(A) = E$. Our proof can be easily generalized [14, Cor. 3.2.6].

Both μ and $\mathcal{H}^m|_E$ are Radon measures [13, Def. 1.40]. We can therefore apply the Besicovitch differentiation theorem [13, Th. 2.22]. This theorem guarantees the existence of

$$\widetilde{f}(x) \triangleq \lim_{r\downarrow 0} \frac{\mu(\overline{B_r(x)})}{\mathcal{H}^m|_E(\overline{B_r(x)})} = \lim_{r\downarrow 0} \frac{\mu(\overline{B_r(x)})}{\mathcal{H}^m(E \cap \overline{B_r(x)})} \quad (6)$$

$\mathcal{H}^m|_E$-a.e., and it ensures that $\widetilde{f}$ is the RN derivative of μ w.r.t. $\mathcal{H}^m|_E$, i.e.,

$$\frac{\mathrm{d}\mu}{\mathrm{d}\mathcal{H}^m|_E} = \widetilde{f}\,, \quad \mathcal{H}^m|_E\text{-a.e.} \quad (7)$$

By [7, Th. 3.2.19], we have that

$$\Theta^m(\mathcal{H}^m|_E, \cdot\,) = 1\,, \quad \mathcal{H}^m|_E\text{-a.e.} \quad (8)$$

We obtain for $\widetilde{f}(x)$ in (6)

$$\begin{aligned}\widetilde{f}(x) &\overset{(8)}{=} \lim_{r\downarrow 0} \frac{\mu(\overline{B_r(x)})}{\mathcal{H}^m(E \cap \overline{B_r(x)})}\, \Theta^m(\mathcal{H}^m|_E, x) \\ &\overset{(4)}{=} \lim_{r\downarrow 0} \frac{\mu(\overline{B_r(x)})}{\mathcal{H}^m(E \cap \overline{B_r(x)})} \lim_{r\downarrow 0} \frac{\mathcal{H}^m(E \cap \overline{B_r(x)})}{\alpha(m)\, r^m} \\ &\overset{(4)}{=} \Theta^m(\mu, x)\end{aligned}$$

$\mathcal{H}^m|_E$-a.e., which together with (7) implies (5). ■

C. Transformation by a Locally Lipschitz Function

We now consider the entropy (2) for a random variable that is the transformation of a continuous random variable on $\mathbb{R}^m$ by some locally Lipschitz function. A proof of the next result is provided in the appendix.

Theorem 5: Let μ be a Borel probability measure on $\mathbb{R}^m$ that is absolutely continuous w.r.t. the Lebesgue measure λ^m, i.e., $\mu \ll \lambda^m$. Furthermore, let $f\colon \mathbb{R}^m \to \mathbb{R}^M$ with $m \leq M$ and let $W \subseteq \mathbb{R}^m$ be a Borel set with $\lambda^m(W^c) = 0$ such that for all $x \in W$: (i) there exists some open neighborhood U_x of x such that $f|_{U_x}$ is Lipschitz, and (ii) $J_f^m(x) > 0$, where J_f^m denotes the m-dimensional Jacobian of f [11, Def. 5.1.3]. Then the induced probability measure μ_f on $\mathbb{R}^M$, defined as $\mu_f(A) \triangleq \mu(f^{-1}(A))$ for a Borel set $A \subseteq \mathbb{R}^M$, is B-countably m-rectifiable. Furthermore, there exists a support E of μ_f such that

$$\frac{\mathrm{d}\mu_f}{\mathrm{d}\mathcal{H}^m|_E}(y) = \sum_{x \in f^{-1}(\{y\})} \frac{\mathrm{d}\mu}{\mathrm{d}\lambda^m}(x) \frac{1}{J_f^m(x)}\,, \quad \mathcal{H}^m|_E\text{-a.e.} \quad (9)$$

If additionally f is one-to-one, the entropy of μ_f is given by

$$h(\mu_f) = h(\mu) + \int \frac{\mathrm{d}\mu}{\mathrm{d}\lambda^m}(x) \log(J_f^m(x))\, \mathrm{d}\lambda^m(x) \quad (10)$$

provided that the right-hand side of (10) exists.

Note that $h(\mu)$ in (10) is the differential entropy of μ on $\mathbb{R}^m$.

IV. Joint Entropy, Conditional Entropy, and Mutual Information

We will now extend our treatment to rectifiable measures on the product space $\mathbb{R}^{M_1} \times \mathbb{R}^{M_2} \cong \mathbb{R}^{M_1+M_2}$. This will allow us to define joint entropy and conditional entropy and, in turn, to establish relations with mutual information.

A. Joint Entropy

Definition 6: For $m_1 \in \{0, 1, \dots, M_1\}$ and $m_2 \in \{0, 1, \dots, M_2\}$, a Borel probability measure μ on $\mathbb{R}^{M_1+M_2}$ is said to be *combined B-countably* (m_1, m_2)*-rectifiable* if there exist a B-countably m_1-rectifiable set $C \subseteq \mathbb{R}^{M_1}$ and a B-countably m_2-rectifiable set $D \subseteq \mathbb{R}^{M_2}$ such that $\mu \ll \mathcal{H}^{m_1}|_C \times \mathcal{H}^{m_2}|_D$. The sets C and D are said to be a *first and second support of* μ, respectively.

The following corollary of [7, Th. 3.2.23] states that a combined B-countably (m_1, m_2)-rectifiable measure is also B-countably $(m_1 + m_2)$-rectifiable.

Corollary 7: Let the sets $C \subseteq \mathbb{R}^{M_1}$ and $D \subseteq \mathbb{R}^{M_2}$ be B-countably m_1-rectifiable and B-countably m_2-rectifiable, respectively. Then the set $C \times D$ is B-countably $(m_1 + m_2)$-rectifiable and we have the following equality of measures:

$$\mathcal{H}^{m_1+m_2}|_{C\times D} = \mathcal{H}^{m_1}|_C \times \mathcal{H}^{m_2}|_D\,. \quad (11)$$

Furthermore, a combined B-countably (m_1, m_2)-rectifiable measure μ on $\mathbb{R}^{M_1+M_2}$ with first support C and second support D is also B-countably $(m_1 + m_2)$-rectifiable on $\mathbb{R}^{M_1+M_2}$ with support $C \times D$.

Proof: The case where $m_1 = 0$ or $m_2 = 0$ follows directly from the definition of the Hausdorff measure and is not treated here; a detailed account can be found in [14, Lem. 4.1.4]. Hereafter, we assume $m_1, m_2 \neq 0$. Without loss of generality, the covers (cf. Definition 1) $C = \bigcup_i f_i(A_i)$ and $D = \bigcup_j g_j(B_j)$ are assumed disjoint, i.e., they are partitions. Let $E \triangleq C \times D$, $C_i \triangleq f_i(A_i)$, and $D_j \triangleq g_j(B_j)$. We first show that $\mathcal{H}^{m_1+m_2}|_E(U \times V) = \mathcal{H}^{m_1}|_C(U)\,\mathcal{H}^{m_2}|_D(V)$ for any two Borel sets $U \subseteq \mathbb{R}^{M_1}$ and $V \subseteq \mathbb{R}^{M_2}$. By[1] [7, Th. 3.2.23] and the σ-additivity of measures,

$$\mathcal{H}^{m_1+m_2}|_E = \sum_{i,j} \mathcal{H}^{m_1+m_2}|_{C_i\times D_j} = \sum_{i,j} \mathcal{H}^{m_1}|_{C_i} \times \mathcal{H}^{m_2}|_{D_j}\,.$$

Thus,

$$\begin{aligned}\mathcal{H}^{m_1+m_2}|_E(U\times V) &= \sum_{i,j} \mathcal{H}^{m_1}|_{C_i}(U)\,\mathcal{H}^{m_2}|_{D_j}(V)\\ &= \Big(\sum_i \mathcal{H}^{m_1}|_{C_i}(U)\Big)\Big(\sum_j \mathcal{H}^{m_2}|_{D_j}(V)\Big)\\ &= \mathcal{H}^{m_1}|_C(U)\,\mathcal{H}^{m_2}|_D(V)\end{aligned}$$

which proves (11). Next, let $\phi_{i,j}(x, y) \triangleq (f_i(x), g_j(y))$. As (cf. Definition 1) the sets $A_i \times B_j$ are bounded and $\phi_{i,j}$ is Lipschitz, with Lipschitz constant $\mathrm{Lip}(\phi_{i,j}) = \max(\mathrm{Lip}(f_i), \mathrm{Lip}(g_j))$, the union E of the Borel sets $\phi_{i,j}(A_i \times B_j) = C_i \times D_j$ is B-countably $(m_1 + m_2)$-rectifiable. Using (11), we obtain $\mu \ll \mathcal{H}^{m_1+m_2}|_{C\times D}$ (cf. Definition 6), and thus μ is B-countably $(m_1 + m_2)$-rectifiable (cf. Definition 2). ■

A combined B-countably (m_1, m_2)-rectifiable probability measure μ can be interpreted as the joint distribution $\mu_{X,Y}$ induced by two random variables X and Y, one being "m_1-dimensional" and the other "m_2-dimensional." Indeed, interpreting the Euclidean space $\mathbb{R}^{M_1+M_2}$ with the Borel σ-algebra as a measurable space, we define the two random variables $X\colon \mathbb{R}^{M_1+M_2} \to \mathbb{R}^{M_1}$ and $Y\colon \mathbb{R}^{M_1+M_2} \to \mathbb{R}^{M_2}$ as the canonical projections onto the first M_1 and last M_2 coordinates, respectively. The marginal distributions μ_X and μ_Y are thus given by $\mu_X(A) = \mu(A \times \mathbb{R}^{M_2})$ and $\mu_Y(B) = \mu(\mathbb{R}^{M_1} \times B)$ for Borel sets $A \subseteq \mathbb{R}^{M_1}$ and $B \subseteq \mathbb{R}^{M_2}$.

Hereafter, we denote by $\mu_{X,Y} = \mu$ a combined B-countably (m_1, m_2)-rectifiable probability measure on $\mathbb{R}^{M_1} \times \mathbb{R}^{M_2}$ with first support C and second support D. By Corollary 7, we can directly apply Definition 3 to $\mu_{X,Y}$ and define the *joint entropy of* X *and* Y as

$$h(\mu_{X,Y}) = h_{\mathcal{H}^{m_1+m_2}|_{C\times D}}(\mu)\,.$$

Similarly, the following lemma allows us to apply Definition 3 to the marginal distributions μ_X and μ_Y.

Lemma 8: The marginal distributions μ_X and μ_Y are B-countably m_1-rectifiable with support C and B-countably m_2-rectifiable with support D, respectively.

Proof: The set C is B-countably m_1-rectifiable by Definition 6. Furthermore, $\mu_X \ll \mathcal{H}^{m_1}|_C$. (Indeed, for any Borel set $A \subseteq \mathbb{R}^{M_1}$, $\mathcal{H}^{m_1}|_C(A) = 0$ implies $(\mathcal{H}^{m_1}|_C \times \mathcal{H}^{m_2}|_D)(A \times \mathbb{R}^{M_2}) = 0$, which in turn implies $\mu_{X,Y}(A \times \mathbb{R}^{M_2}) = 0$, which is equivalent to $\mu_X(A) = 0$.) Hence, μ_X is B-countably m_1-rectifiable with support C (cf. Definition 2). The proof for μ_Y is analogous. ■

[1] When applying [7, Th. 3.2.23], the fact that the sets are B-countably m_1-/m_2-rectifiable and not merely countably m_1-/m_2-rectifiable (as defined in [7, §3.2.14(2)]) is used.

B. Conditional Entropy

Next, we use the concept of a regular conditional probability measure to define conditional entropy. Since we consider probability measures in Euclidean space, a regular conditional probability measure $\mu_{X|Y}(\,\cdot\,|\,\cdot\,)$ always exists [15, Cor. 5.8.1].

Lemma 9: We have μ_Y-a.e. that the measure $\mu_{X|Y}(\,\cdot\,|y)$ is B-countably m_1-rectifiable with support C and with RN derivative

$$\frac{\mathrm{d}\mu_{X|Y}(\,\cdot\,|y)}{\mathrm{d}\mathcal{H}^{m_1}|_C}(x) = \frac{\frac{\mathrm{d}\mu_{X,Y}}{\mathrm{d}\mathcal{H}^{m_1+m_2}|_{C\times D}}(x,y)}{\frac{\mathrm{d}\mu_Y}{\mathrm{d}\mathcal{H}^{m_2}|_D}(y)}, \quad \mathcal{H}^{m_1}|_C\text{-a.e.} \tag{12}$$

Proof: Because C is a B-countably m_1-rectifiable set (cf. Definition 6), it suffices to show that $\mu_{X|Y}(\,\cdot\,|y) \ll \mathcal{H}^{m_1}|_C$. The RN derivatives $\frac{\mathrm{d}\mu_{X,Y}}{\mathrm{d}\mathcal{H}^{m_1+m_2}|_{C\times D}}$ and $\frac{\mathrm{d}\mu_Y}{\mathrm{d}\mathcal{H}^{m_2}|_D}$ exist by Corollary 7 and Lemma 8, respectively. Thus, for any Borel sets $A \subseteq \mathbb{R}^{M_1}$ and $B \subseteq \mathbb{R}^{M_2}$,

$$\begin{aligned}
&\int_B \int_A \frac{\mathrm{d}\mu_{X,Y}}{\mathrm{d}\mathcal{H}^{m_1+m_2}|_{C\times D}}(x,y)\,\mathrm{d}\mathcal{H}^{m_1}|_C(x)\,\mathrm{d}\mathcal{H}^{m_2}|_D(y)\\
&\quad\overset{(11)}{=} \mu_{X,Y}(A\times B)\\
&\quad\overset{(*)}{=} \int_B \mu_{X|Y}(A|y)\,\mathrm{d}\mu_Y(y)\\
&\quad= \int_B \frac{\mathrm{d}\mu_Y}{\mathrm{d}\mathcal{H}^{m_2}|_D}(y)\,\mu_{X|Y}(A|y)\,\mathrm{d}\mathcal{H}^{m_2}|_D(y)
\end{aligned} \tag{13}$$

where $(*)$ follows from the definition of regular conditional probability [15, Sec. 5.8]. Because (13) holds for arbitrary B, we obtain $\mathcal{H}^{m_2}|_D$-a.e. that $\frac{\mathrm{d}\mu_Y}{\mathrm{d}\mathcal{H}^{m_2}|_D}(y)\,\mu_{X|Y}(\,\cdot\,|y) \ll \mathcal{H}^{m_1}|_C$. Clearly, we have μ_Y-a.e. that $\frac{\mathrm{d}\mu_Y}{\mathrm{d}\mathcal{H}^{m_2}|_D}(y) \neq 0$. We thus conclude that $\mu_{X|Y}(\,\cdot\,|y) \ll \mathcal{H}^{m_1}|_C$ and, hence, that $\mu_{X|Y}(\,\cdot\,|y)$ is B-countably m_1-rectifiable with support C, μ_Y-a.e. By (13), the RN derivative of $\frac{\mathrm{d}\mu_Y}{\mathrm{d}\mathcal{H}^{m_2}|_D}(y)\,\mu_{X|Y}(\,\cdot\,|y)$ w.r.t. $\mathcal{H}^{m_1}|_C$ is $\frac{\mathrm{d}\mu_{X,Y}}{\mathrm{d}\mathcal{H}^{m_1+m_2}|_{C\times D}}(\,\cdot\,,y)$, which immediately gives (12). ■

Based on Lemma 9, we can define $h(\mu_{X|Y}(\,\cdot\,|y))$ μ_Y-a.e. by substituting $\mu_{X|Y}(\,\cdot\,|y)$ for μ and (12) for $\frac{\mathrm{d}\mu}{\mathrm{d}\mathcal{H}^m|_E}$ in Definition 3. Furthermore, we define the *conditional entropy of X given Y* as

$$h(\mu_{X|Y}) \triangleq \int h(\mu_{X|Y}(\,\cdot\,|y))\,\mathrm{d}\mu_Y(y)\,. \tag{14}$$

Theorem 10: The conditional entropy satisfies

$$h(\mu_{X|Y}) = h(\mu_{X,Y}) - h(\mu_Y) \tag{15}$$

provided that the right-hand side of (15) exists.

Proof: By (12), the right-hand side of (14) can be written as

$$\begin{aligned}
&-\iint \frac{\mathrm{d}\mu_{X,Y}}{\mathrm{d}\mathcal{H}^{m_1+m_2}|_{C\times D}}(x,y)\\
&\qquad\cdot \log\left(\frac{\frac{\mathrm{d}\mu_{X,Y}}{\mathrm{d}\mathcal{H}^{m_1+m_2}|_{C\times D}}(x,y)}{\frac{\mathrm{d}\mu_Y}{\mathrm{d}\mathcal{H}^{m_2}|_D}(y)}\right)\mathrm{d}\mathcal{H}^{m_1}|_C(x)\,\mathrm{d}\mathcal{H}^{m_2}|_D(y)\,.
\end{aligned}$$

By splitting the logarithm and applying (11), this becomes $h(\mu_{X,Y}) - h(\mu_Y)$. ■

C. Mutual Information

The *mutual information of X and Y* is defined as [16, Sec. 5.5]

$$I(X;Y) \triangleq \int \log\left(\frac{\mathrm{d}\mu_{X,Y}}{\mathrm{d}(\mu_X\times\mu_Y)}\right)\mathrm{d}\mu_{X,Y} \tag{16}$$

if $\mu_{X,Y} \ll \mu_X\times\mu_Y$ and $I(X;Y) \triangleq \infty$ otherwise. Note that by the Perez-Yaglom-Gelfand theorem [16, Lem. 5.2.3], this definition is consistent with the definition of information in [10, Def. 2.1] .

Theorem 11: If $\mu_{X,Y} \ll \mu_X\times\mu_Y$, then

$$\begin{aligned}
I(X;Y) &= h(\mu_X) + h(\mu_Y) - h(\mu_{X,Y}) &(17)\\
&= h(\mu_X) - h(\mu_{X|Y}) &(18)
\end{aligned}$$

provided that the right-hand sides of (17) and (18) exist.

Proof: With $E \triangleq C\times D$, one can write

$$\frac{\mathrm{d}\mu_{X,Y}}{\mathrm{d}(\mu_X\times\mu_Y)} = \frac{\mathrm{d}\mu_{X,Y}}{\mathrm{d}\mathcal{H}^{m_1+m_2}|_E}\left(\frac{\mathrm{d}(\mu_X\times\mu_Y)}{\mathrm{d}\mathcal{H}^{m_1+m_2}|_E}\right)^{-1}.$$

(Note that the involved RN derivatives exist due to $\mu_{X,Y} \ll \mu_X\times\mu_Y$.) Inserting in (16) yields

$$\begin{aligned}
&I(X;Y)\\
&= \int \log\left(\frac{\mathrm{d}\mu_{X,Y}}{\mathrm{d}\mathcal{H}^{m_1+m_2}|_E}\left(\frac{\mathrm{d}(\mu_X\times\mu_Y)}{\mathrm{d}\mathcal{H}^{m_1+m_2}|_E}\right)^{-1}\right)\mathrm{d}\mu_{X,Y}\\
&= -h(\mu_{X,Y}) - \int \log\left(\frac{\mathrm{d}(\mu_X\times\mu_Y)}{\mathrm{d}\mathcal{H}^{m_1+m_2}|_E}\right)\mathrm{d}\mu_{X,Y}\\
&\overset{(*)}{=} -h(\mu_{X,Y})\\
&\qquad - \int \log\left(\frac{\mathrm{d}\mu_X}{\mathrm{d}\mathcal{H}^{m_1}|_C}(x)\,\frac{\mathrm{d}\mu_Y}{\mathrm{d}\mathcal{H}^{m_2}|_D}(y)\right)\mathrm{d}\mu_{X,Y}(x,y)
\end{aligned}$$

where $(*)$ follows upon using (11) and Fubini's theorem. By splitting the logarithm, we obtain (17). Expression (18) then follows upon using (15). ■

V. Conclusion

We have established a definition of entropy for a broad and practically relevant class of singular probability measures. This definition encompasses the entropy of discrete random variables and the differential entropy of continuous random variables as special cases. We have shown that our entropy transforms in a natural manner under Lipschitz functions and that it conveys useful expressions of the mutual information for a broad class of random variables. Further results and more detailed proofs are provided in [14], including an extension of our entropy definition to mixtures of rectifiable measures of possibly different dimensions.

An extension of our approach to noninteger dimension m seems difficult. Indeed, rectifiability is only defined for integer dimension m and the convenient property of the Hausdorff measure in (11) has only been shown for rectifiable sets (see [7, §3.2.24] for a counterexample). Moreover, the convergence of (4) on a Borel set of positive measure requires m to be integer [17, Th. 3.1].

An interesting topic for further research is the connection of our entropy to the entropy proposed in [2]. The latter is defined via a limit of the expectation of the logarithm of discretizations of the random variable, whereas our entropy definition can be interpreted as the expectation of the logarithm of a limit of discretizations of the random variable (cf. (4)).

There are cases where the entropy of [2] is defined and our entropy is not (e.g., for Cantor-like distributions). However, as the example in [2, pp. 197–198] shows, there are also cases where differential entropy (which coincides with our entropy definition for continuous random variables) exists but the entropy of [2] does not.

Appendix: Proof of Theorem 5

We first show that μ_f is B-countably m-rectifiable. According to Definition 2, we have to show that there exists a B-countably m-rectifiable set E such that $\mu_f \ll \mathcal{H}^m|_E$.

For each $x \in W$, we can find an open ball $V_x \triangleq B_{r_x}(x) \subseteq \overline{B_{r_x}(x)} \subseteq U_x$. The family $\{V_x\}_{x\in W}$ constitutes an open cover of W. By the Lindelöf property, a countable subfamily $\{V_{x_i}\}_{i\in\mathcal{J}}$ suffices to cover W. The sets $\overline{V_{x_i}}$ are compact, thus the sets $E_i \triangleq f(\overline{V_{x_i}})$ are compact as well. Consequently, E_i is Borel and, hence, the set $E \triangleq \bigcup_{i\in\mathcal{J}} E_i$ is B-countably m-rectifiable.

To show that $\mu_f \ll \mathcal{H}^m|_E$, let $S \subseteq \mathbb{R}^M$ be a Borel set such that $\mathcal{H}^m|_E(S) = 0$. We need to show that $\mu_f(S) = \mu(f^{-1}(S)) = 0$. To this end, we will first show that $\lambda^m(f^{-1}(S)) = 0$. Because the sets V_{x_i} cover W, we have $\lambda^m(f^{-1}(S)) \leq \sum_{i\in\mathcal{J}} \lambda^m(f^{-1}(S)\cap V_{x_i})$, and thus it suffices to show that $\lambda^m(f^{-1}(S)\cap V_{x_i}) = 0$ for all $i \in \mathcal{J}$. Using Kirszbraun's theorem [7, Th. 2.10.43], we can extend $f|_{V_{x_i}}$ to a Lipschitz function $\widetilde{f}$ with domain $\mathbb{R}^m$ and apply the area formula [7, Th. 3.2.3], which yields

$$\int_{f^{-1}(S)\cap V_{x_i}} J^m_{\widetilde{f}} \,\mathrm{d}\lambda^m = 0\,. \tag{19}$$

As (by our assumption) $J^m_{\widetilde{f}}(x) > 0$ for $x \in V_{x_i} \cap W$, (19) implies that $\lambda^m(f^{-1}(S) \cap V_{x_i}) = 0$. Thus, $\lambda^m(f^{-1}(S)) = 0$ and, using the assumption $\mu \ll \lambda^m$, we conclude that $\mu(f^{-1}(S)) = 0$. Hence, $\mu_f \ll \mathcal{H}^m|_E$.

Next, we show (9). Consider the disjoint partition $\{\widetilde{V}_i\}_{i\in\mathcal{J}}$ of W that is constructed as $\widetilde{V}_i \triangleq \overline{V_{x_i}} \setminus \bigcup_{i'<i} \overline{V_{x_{i'}}}$ (i.e., all preceding sets are removed). For every $i \in \mathcal{J}$, $f|_{\widetilde{V}_i}$ is Lipschitz. We apply Kirszbraun's theorem and the area formula to $f|_{U_{x_i}}$, yielding

$$\int_{\widetilde{V}_i} u(x)\, J^m_f(x) \,\mathrm{d}\lambda^m(x) = \int_{f(\widetilde{V}_i)} \Bigg(\sum_{x\in f^{-1}(\{y\})} u(x) \Bigg) \mathrm{d}\mathcal{H}^m(y) \tag{20}$$

for any Borel-measurable function $u\colon \mathbb{R}^m \to \mathbb{R}^+$. Observing that $J^m_f(x) > 0$ for $x \in W$, we choose $u(x) = \mathbb{1}_{f^{-1}(Z)}(x) \frac{\mathrm{d}\mu}{\mathrm{d}\lambda^m}(x) \frac{1}{J^m_f(x)}$, where $Z \subseteq \mathbb{R}^M$ is an arbitrary Borel set and $\mathbb{1}_A$ denotes the indicator function of the set A. The left-hand side of (20) then becomes

$$\int_{\widetilde{V}_i} \mathbb{1}_{f^{-1}(Z)}(x) \frac{\mathrm{d}\mu}{\mathrm{d}\lambda^m}(x) \,\mathrm{d}\lambda^m(x) \;=\; \mu(f^{-1}(Z) \cap \widetilde{V}_i) \tag{21}$$

and the right-hand side of (20) becomes

$$\int_{f(\widetilde{V}_i)\cap Z} \Bigg(\sum_{x\in f^{-1}(\{y\})} \frac{\mathrm{d}\mu}{\mathrm{d}\lambda^m}(x) \frac{1}{J^m_f(x)} \Bigg) \mathrm{d}\mathcal{H}^m(y)\,. \tag{22}$$

We then obtain

$$\begin{aligned}
\mu_f(Z) &= \mu(f^{-1}(Z)) \\
&\overset{(a)}{=} \sum_{i\in\mathcal{J}} \mu(f^{-1}(Z) \cap \widetilde{V}_i) \\
&\overset{(b)}{=} \sum_{i\in\mathcal{J}} \int_{f(\widetilde{V}_i)\cap Z} \Bigg(\sum_{x\in f^{-1}(\{y\})} \frac{\mathrm{d}\mu}{\mathrm{d}\lambda^m}(x) \frac{1}{J^m_f(x)} \Bigg) \mathrm{d}\mathcal{H}^m(y) \\
&\overset{(c)}{=} \int_{E\cap Z} \Bigg(\sum_{x\in f^{-1}(\{y\})} \frac{\mathrm{d}\mu}{\mathrm{d}\lambda^m}(x) \frac{1}{J^m_f(x)} \Bigg) \mathrm{d}\mathcal{H}^m(y)
\end{aligned} \tag{23}$$

where (a) is due to $\mu\big((\bigcup_{i\in\mathcal{J}} \widetilde{V}_i)^c\big) = 0$, (b) follows by the equality of (21) and (22), and (c) follows because $\bigcup_{i\in\mathcal{J}} f(\widetilde{V}_i) = E$. Since Z is an arbitrary Borel set, (23) implies (9).

Finally, we consider the case where f is one-to-one. Starting from (3), we have

$$\begin{aligned}
h(\mu_f) &= -\int \frac{\mathrm{d}\mu_f}{\mathrm{d}\mathcal{H}^m|_E}(y) \log\bigg(\frac{\mathrm{d}\mu_f}{\mathrm{d}\mathcal{H}^m|_E}(y) \bigg) \mathrm{d}\mathcal{H}^m|_E(y) \\
&= -\int \frac{\mathrm{d}\mu}{\mathrm{d}\lambda^m}(x) \log\bigg(\frac{\mathrm{d}\mu}{\mathrm{d}\lambda^m}(x) \frac{1}{J^m_f(x)} \bigg) \mathrm{d}\lambda^m(x)
\end{aligned} \tag{24}$$

where the last step follows from [7, Th. 3.2.3(2)] considering (9). By the additivity of the Lebesgue integral, (24) is equal to the right-hand side of (10).

References

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New York, NY: Wiley, 2006.

[2] A. Rényi, "On the dimension and entropy of probability distributions," *Acta Math. Hung.*, vol. 10, no. 1, pp. 193–215, 1959.

[3] D. Stotz and H. Bölcskei, "Degrees of freedom in vector interference channels," in *Proc. Allerton Conf. Comm. Control Comput.*, Monticello, IL, Oct. 2012, pp. 1755–1760.

[4] Y. Wu and S. Verdú, "Rényi information dimension: Fundamental limits of almost lossless analog compression," *IEEE Trans. Inf. Theory*, vol. 56, no. 8, pp. 3721–3748, Aug. 2010.

[5] R. Palanki, "Iterative decoding for wireless networks," Ph.D. dissertation, California Inst. of Technology, Pasadena, CA, 2004.

[6] G. Koliander, E. Riegler, G. Durisi, and F. Hlawatsch, "Degrees of freedom of generic block-fading MIMO channels without a priori channel state information," *IEEE Trans. Inf. Theory*, submitted, 2013, available online: arXiv:1310.2490.

[7] H. Federer, *Geometric Measure Theory*. New York, NY: Springer, 1969.

[8] J. R. Munkres, *Topology*. Upper Saddle River, NJ: Prentice Hall, 2000.

[9] L. J. Corwin and R. H. Szczarba, *Calculus in Vector Spaces*. New York, NY: Marcel Dekker, 1979.

[10] M. S. Pinsker, *Information and Information Stability of Random Variables and Processes*. San Francisco, CA: Holden-Day, 1964.

[11] S. G. Krantz and H. R. Parks, *Geometric Integration Theory*. Boston, MA: Birkhäuser, 2008.

[12] K. B. Athreya and S. N. Lahiri, *Measure Theory and Probability Theory*. New York, NY: Springer, 2006.

[13] L. Ambrosio, N. Fusco, and D. Pallara, *Functions of Bounded Variation and Free Discontinuity Problems*. New York, NY: Oxford Univ. Press, 2000.

[14] G. Pichler, "Information theory on rectifiable sets," Master's thesis, Vienna Univ. of Technology, Vienna, Austria, 2013.

[15] R. M. Gray, *Probability, Random Processes, and Ergodic Properties*, 1st ed. New York, NY: Springer, 1988, corrected version available online: http://ee.stanford.edu/~gray/arp.html.

[16] R. M. Gray, *Entropy and Information Theory*, 1st ed. New York, NY: Springer, 1990, corrected version available online: http://ee.stanford.edu/~gray/it.html.

[17] C. De Lellis, *Rectifiable Sets, Densities and Tangent Measures*. Zurich, Switzerland: EMS Publishing House, 2008.