Mathematics of Sparsity and Entropy: Axioms, Core Functions and Sparse Recovery

Giancarlo Pastor, Inmaculada Mora-Jiménez, Riku Jäntti, and Antonio J. Caamaño

Abstract

Sparsity and entropy are pillar notions of modern theories in signal processing and information theory. However, there is no clear consensus among scientists on the characterization of these notions. Previous efforts have contributed to understand individually sparsity or entropy from specific research interests. This paper proposes a mathematical formalism, a joint axiomatic characterization, which contributes to comprehend (the beauty of) sparsity and entropy. The paper gathers and introduces inherent and first principles criteria as axioms and attributes that jointly characterize sparsity and entropy. The proposed set of axioms is constructive and allows to derive simple or core functions and further generalizations. Core sparsity generalizes the Hoyer measure, Gini index and $pq$-means. Core entropy generalizes the Rényi entropy and Tsallis entropy, both of which generalize Shannon entropy. Finally, core functions are successfully applied to compressed sensing and to minimum entropy given sample moments. More importantly, the (simplest) core sparsity adds theoretical support to the $\ell_1$-minimization approach in compressed sensing.

Index Terms

Sparsity, entropy, axioms, compressive sampling, compressed sensing, Rényi entropy, Tsallis entropy.

This paper was presented in part in the Proceedings of the Tenth International Symposium in Wireless Communication Systems (ISWCS 2013).

G.P. (corresponding author, giancarlo.pastor@aalto.fi) and R.J. are with the Department of Communications and Networking, Aalto University, Finland.

G.P., I.M.J. and A.J.C. are with the Department of Information and Communication Technologies, King Juan Carlos University, Spain.

This work was supported by the Spanish Ministry of Economy and Competitiveness (TEC2013-4839-C4-1-R) and the Finnish Cultural Foundation.
I. INTRODUCTION

For a given signal $x$, the uncertainty (randomness) of its elements defines the compressibility (compactness) of its coefficients $w$ in a given domain. This dependence, between the (elements’) uncertainty and (coefficients’) compressibility of a signal, suggests a connection between the two families of functions that measure these properties. But, what axioms do these families have in common? This paper adopts the following definitions with the aim to respond to this question. Let $p$ be the probability mass function (pmf) of signal $x$.

**Definition 1.1 (Compressibility or sparsity):** The property of to concentrate most of the energy in few coefficients of $w$.

**Definition 1.2 (Uncertainty or entropy):** The property of not to concentrate most of the probability mass in few atoms of $p$.

Sparsity and entropy functions quantify properties [1.1] and [1.2] respectively. One approach to evaluate the goodness of these functions is through validation of criteria (see [30] [33] for sparsity, [20] [28] for entropy, and [41] for both). Following this approach, the paper gathers and introduces inherent and first principles criteria as axioms and attributes that jointly characterize sparsity and entropy functions. As expected, it turns out that criteria of both families of functions are strictly complementary.

In [16] a set of conditions are identified under which variance and entropy order distributions in a similar way. This suggests that entropy could be extracted from the “shape” of distributions. Intuitively, for a given variance, random processes with similar sorted pmf’s should present similar levels of entropy. Hence, for a given signal with elements following a univariate distribution, the uncertainty of the undergoing process equals the non-compressibility of the pmf. The following toy examples illustrate this relation. Let r.v. $x \in \{x_1, x_2\}$ with pmf $p = (p_1, p_2)$.

1) Assume $p_1 > p_2$: if $p_1$ increases (thus $p_2$ decreases since $p_1, p_2 \geq 0$ and $\|p\|_1 = 1$), then $x_1$ is even more certain to appear, i.e. if the compressibility of $p$ increases then the uncertainty of $X$ decreases, 

\[
\text{sparsity}(p) \uparrow \equiv \text{entropy}(x) \downarrow .
\]

2) Assume $p = (p_1, p_2) = (1, 0)$ (a constant r.v.): if $p_2$ increases (thus $p_1$ decreases), then $x_1$ is not the unique possible outcome, i.e. if the compressibility of $p$ decreases then the uncertainty of $X$ increases, 

\[
\text{sparsity}(p) \downarrow \equiv \text{entropy}(x) \uparrow .
\]

**Relation to Previous Work**

Entropy functions (related to uncertainty, information, fuzziness or complexity measures) and sparsity functions (related to compressibility, fairness or dispersion measures) are not regarded as easily characterizable. Hence these functions have been thoroughly studied in numerous articles from different perspectives and research areas. The following present relevant references which contribute explicitly to the set of axioms and attributes of sparsity and entropy functions of Section [11].

The main method for axiomatic characterization of entropy functions consist on to treat inherent or satisfactory properties as axioms for (reference) Shannon entropy function. This started in 1948 with [5] which established...
continuity and monotonicity of entropy functions; [13] added concavity for fuzziness functions and [18] relaxed it to quasi-concavity for dispersion functions; [15] added concentration for information functions; [20] added maximality for entropy functions; and [28] added symmetry for information functions.

On the other hand, although sparse data and sparsity has attracted a lot of attention in recent years, concentration, scaling, homogeneous growth and replication were originally applied in 1920 by [2] in a financial setting to measure the inequity of wealth distribution; [18] added bounds for dispersion functions; [30] [33] added symmetry and continuity for sparsity and fairness functions, respectively; and [38] added quasi-convexity for reward-risk ratio functions.

Contribution of the Paper

For each property, previous works have separately established some axioms and relationships, and derived generalizations. The contribution of the paper is threefold.

1) The main contribution of this paper is a refined and constructive set of axioms (and attributes) of sparsity and entropy functions (Section II), which allows to derive function generalizations.

2) The derived sparsity functions explain the benefits of proposed sparsity functions in Compressed Sensing and, in particular, the effectiveness of the $\ell_1$-minimization approach (Section III).

3) The derived entropy functions generalize Rényi and Tsallis entropy, and offer simple formulations to the sparse recovery of probability measures (Section IV).

The rest of the paper is organized as follows: For sparsity and entropy functions, Section II presents the axioms; Sections III-IV describe existing functions and derive generalizations; Section V concludes; and the Appendix contains the proofs.

II. AXIOMS

This section gathers and introduces inherent and first principles criteria as axioms and attributes that jointly characterize sparsity and entropy functions.

Let $X$ be an independent and identically distributed (i.i.d.) discrete random variable (r.v.), defined on a countable sample space $\Omega = \{x^1, \ldots, x^b\} \subseteq \mathbb{R}$ ($b$ could be infinity), and let $x = (x_i) \in \Omega^n$ be a vector containing $n$ realizations of r.v. $X$. The definition of vector $x$ allows two possible interpretations. It is an arbitrary signal as a whole and a random process formed by its elements. Throughout this paper both interpretations are adopted, and different operations are applied on vector $x$ according to the analysis performed. The analysis of sparsity will be from the signal’s compressibility perspective. Analogously, the analysis of entropy will be from the process’s uncertainty perspective. The analogies between these perspectives are summarized in Table I. The following operations will provide adequate representations for the analyzes.

For the compressibility assessment of signal $x$, a domain transformation is applied using basis $\Phi \in \mathbb{R}^{n \times n}$,

$$\Phi : \mathbb{R}^n \rightarrow \mathbb{R}^n_{+} : x \mapsto w = \Phi x,$$

(3)
where \( w = (w_i) \) is the (vector of) coefficients of \( x \). For ease of notation, since compressibility is measured from magnitudes, the energy values of coefficients \( w \) are assumed non-negative, i.e. \( w \in \mathbb{R}_+^n \). Essentially, redefine coefficients (in terms of true coefficients) as \( w_i = |w_i| \). Under this model, a sparsity function, which measures the compactness of signal \( x \) under basis \( \Phi \), is
\[
s : \mathbb{R}_+^n \to \mathbb{R} : w = \Phi x \mapsto s(w). \tag{4}
\]

For the uncertainty calculation of process \( x \), a \( b \)-bins-based \( \ell_1 \)-normalized histogram method is applied using operator \( \Psi \),
\[
\Psi : \mathbb{R}^n \to [0, 1]^b : x \mapsto p = \Psi x, \tag{5}
\]
where \( p = (p_j) \) is the sample probability mass function (pmf) of \( x \), with \( \|p\|_1 = 1 \) and
\[
p_j = \frac{\text{cardinality}\{ i : x_i \in x^j \}}{n} \approx \mathbb{P}(X = x^j). \tag{6}
\]

Similarly, define \( \mathcal{S}_+^b = \{ p \in [0, 1]^b : \|p\|_1 = 1 \} \) (probability simplex). Then, for a given variance \( \sigma^2 \), an entropy function, which measures the randomness of process \( x \) from its outcomes’ occurrences using method \( \Psi \), is
\[
h^b : \mathcal{S}_+^b \to \mathbb{R} : p = \Psi x \mapsto h^b(p). \tag{7}
\]

Under this notation (see Table II for a summary), the following enumerates a collection of axioms and attributes of sparsity and entropy functions. Axioms and attributes describe the effect of different actions on the argument of functions, e.g. changes on its pattern, ratios or relative differences (see Table III for mathematical description).

In the case of entropy analysis, and in order to stay inside the probability simplex, the actions related to regularity II-B5 and homogeneous growth II-B6 require a subsequent normalization \( \|\tilde{p}\|_1 = 1 \), where \( \tilde{p} \) denotes \( p \) after the action. The histogram model described above makes this normalization step transparent.

A. Axioms

1) Continuity (II-A1):

S. (Sparsity) Slight energy perturbation retains the ratio of dominant and negligible coefficients. [33]

H. (Entropy) Slight probability mass perturbation retains the ratio of unlikely and likely events. [5]
Table II

| Symbol | Description |
|--------|-------------|
| $s(\cdot)$ | sparsity function |
| $h^b(\cdot)$ | entropy function given $b$ events (states) |
| $\Omega$ | countable sample space, $\Omega = \{x_1, \ldots, x^b\} \subseteq \mathbb{R}$ |
| $P$ | probability measure defined on $\Omega$ |
| $X$ | random variable, $X : \omega \mapsto X(\omega) = \omega$ |
| $x = (x_i)$ | vector with elements $x_i \in \Omega$ (realizations of $X$) |
| $n$ | length of vector $x$ |
| $\Phi$ | domain transformation, $\Phi \in \mathbb{R}^{n \times n}$ |
| $w = (w_i)$ | vector with coefficients $w_i$ of $x$, $w = \Phi x$ |
| $\Psi$ | $b$-bins histogram operator, $\Psi$ |
| $p = (p_j)$ | sample probability mass function of $x$, $p = \Psi x$ |
| $b$ | number of events (states) or bins, length of vector $p$ |
| $x^j$ | $j$-th event in the histogram method |
| $v = (v_k)$ | arbitrary non-null vector ($v \neq 0$) with elements $v_k$ |
| $\|\cdot\|_0$ | $\ell_0$-“norm”, $\|v\|_0 = \text{cardinality}(\text{supp}(v))$ |
| $\|\cdot\|_p$ | $\ell_p$-norm, $\|v\|_p = (\sum_k |v_k|^p)^{\frac{1}{p}}$, $p \in (0, \infty)$ |
| $\|\cdot\|_\infty$ | $\ell_\infty$-norm, $\|v\|_\infty = \max_k |v_k|$ |
| $\|\cdot\|_{1^*}$ | $\ell_{1^*}$ weighted norm, $\|v\|_{1^*} = \sum_k k|v_k|$, $v_i \geq v_j$, $i > j$ |
| $\cdot\cdot$ | concatenation operator of vectors |
| $\pi(\cdot)$ | permutation in the order of elements |
| $S^n_+$ | $n$-simplex, $S^n_+ = \{v \in [0, 1]^b : \|v\|_1 = n\}$ |
| $0_k$, $1_k$ | vector of all zeros and ones of length $k$, resp. |

2) **Symmetry (II-A2):**

S. Energy permutation retains the ratio of dominant and negligible coefficients. [18]

H. Probability mass permutation retains the ratio of unlikely and likely events. [28]

3) **Concentration (II-A3) (Dalton’s 1st Law):**

S. Moving energy from negligible to dominant coefficients shorten the set of dominant coefficients. [2]

H. Moving probability mass from unlikely to likely events shorten the set of likely events. [15]

4) **Scaling (II-A4) (Dalton’s modified 2nd Law):**

S. Scaled coefficients contain proportional amounts of energy in proportional number of coefficients. [2]

H. Scaled occurrences contain proportional amounts of probability mass in proportional number of events (or by the $\|p\|_1 = 1$ normalization).

5) **Replication (II-A5) (Dalton’s 4th Law):**

S. Concatenating replicas of (all) coefficients retains the ratio of dominant and negligible coefficients. [2]

H. Adding realizations following the same law retains the ratio of unlikely and likely events. This axiom
affirms that entropy is intrinsic to a given process (see Subsection IV-B for details).

B. Attributes

1) **Bounds (II-B1):**

S. The least (most) compressible coefficients allocate (all) its energy on all coefficients (on one single coefficient), e.g. a white noise (a single “tone”).

H. The most (least) uncertain distribution allocates (all) its probability mass on all events (on one single event), e.g. a uniform r.v. (a constant r.v.).

**Lemma 2.1:** II-A3 and II-A4 imply II-B1

2) **Quasi-convexity (II-B2):**

S. Quasi-convexity prefers extremes than averages.

H. Quasi-concavity encourages diversification.

**Lemma 2.2:** II-A1, II-A2, II-A3 and II-A4 imply II-B2

3) **Monotonicity (II-B3):** A simple version of II-A3

S. For two signals with non-zero pair of coefficients, the signal with the largest ratio highest-energy by lowest-energy is more compressible.

H. For two processes with non-impossible pair of events, the process with the largest ratio highest-mass by lowest-mass is less uncertain.

**Lemma 2.3:** II-A3 implies II-B3

4) **Completeness (II-B4):**

S. Adding zero-energy coefficients concentrates the energy in relatively less coefficients. (Babies)

H. (Conditioning on) having impossible (zero-mass) events concentrates the mass in relatively less events.

**Lemma 2.4:** II-A3, II-A4 and II-A5 imply II-B4

5) **Regularity (II-B5):**

S. Significant concentration of energy in one single coefficient makes the rest negligible. (Bill Gates)

H. Significant concentration of probability mass in one single event makes the rest unlikely.

**Lemma 2.5:** II-A3 and II-A4 imply II-B5

6) **Homogeneous growth (II-B6) (Dalton’s 3rd Law):**

S. Increasing relatively more the energy of negligible coefficients makes them relatively more dominant.

H. Increasing relatively more the probability mass of unlikely events makes them relatively more likely.

**Lemma 2.6:** II-A4, II-B1 and II-B2 imply II-B6

7) **Schur-convexity (II-B7):**

S. Schur-convexity holds.

H. Schur-concavity holds.

**Lemma 2.7:** II-A2 and II-B2 imply II-B7

January 23, 2015
Table III

| Name | Sparsity or Compressibility | Entropy or Uncertainty |
|------|-----------------------------|------------------------|
| II-A1 | Continuity | \( \tilde{w} \to w, s(\tilde{w}) \to s(w) \) | \( \tilde{p} \to p, h^h(\tilde{p}) \to h^h(p) \) |
| II-A2 | Symmetry | \( \tilde{w} = \pi(w), s(\tilde{w}) = s(w) \) | \( \tilde{p} = \pi(p), h^h(\tilde{p}) = h^h(p) \) |
| II-A3 | Concentration | \[ \begin{array}{c}
\tilde{w} = (\tilde{u}_k), \tilde{u}_i = u_i + \alpha, \tilde{u}_j = u_j - \alpha, w_i \geq w_j \\
\tilde{w}_k = w_k, k \neq \{i, j\}, s(\tilde{w}) > s(w)
\end{array} \] | \( \tilde{p} = (\tilde{p}_k), \tilde{p}_i = p_i + \alpha, \tilde{p}_j = p_j - \alpha, p_i \geq p_j \\
\tilde{p}_k = p_k, k \neq \{i, j\}, h^h(\tilde{p}) < h^h(p) \) |
| II-A4 | Scaling | \( \tilde{w} = \alpha w, s(\tilde{w}) = s(w) \) | \( \tilde{p} = \alpha p, h^h(\tilde{p}) = h^h(p) \) |
| II-A5 | Replication | \( \tilde{w} = w \| \ldots \| w, s(\tilde{w}) = s(w) \) | \( \tilde{p} = p \| \ldots \| p, h^h(\tilde{p}) = h^h(p) \) |
| II-B1 | Bounds | \( \tilde{w} = \alpha \mathbf{1}_n, \tilde{w} = 0_{n-1} \| \alpha, s(\tilde{w}) \leq s(w) \) | \( \tilde{p} = \alpha \mathbf{1}_n, \tilde{p} = 0_{n-1} \| \alpha, h^h(\tilde{p}) \geq h^h(p) \) |
| II-B2 | Quasi-convexity | \( \alpha < 1, s(\alpha w + (1 - \alpha)\tilde{w}) < \max\{s(w), s(\tilde{w})\} \) | \( \alpha < 1, h^h(\alpha p + (1 - \alpha)\tilde{p}) > \min\{h^h(p), h^h(\tilde{p})\} \) |
| II-B3 | Monotonicity | \( \frac{\alpha}{\tilde{w}} \leq w < \frac{\alpha}{\tilde{w}}, s(\tilde{w} - \alpha w) > s(w, \alpha - w) \) | \( \frac{\alpha}{\tilde{p}} \leq p < \frac{\alpha}{\tilde{p}}, h^h(\tilde{p}, \alpha - \tilde{p}) > h^h(p, \alpha - p) \) |
| II-B4 | Completeness | \( \tilde{w} = w \| 0, s(\tilde{w}) > s(w) \) | \( \tilde{p} = p \| 0, h^h(\tilde{p}) < h^h(p) \) |
| II-B5 | Regularity | \( \exists \beta, \tilde{w} = (\tilde{u}_k), \tilde{w}_i = \tilde{w}_i = \tilde{w}_k = w_k, k \neq i \) | \( \exists \beta, \tilde{p} = (\tilde{p}_k), \tilde{p}_i = \tilde{p}_k = p_k, k \neq i \) |
| II-B6 | Hom. growth | \( \tilde{w} = w + \alpha \mathbf{1}_n, s(\tilde{w}) < s(w) \) | \( \tilde{p} = p + \alpha \mathbf{1}_n, h^h(\tilde{p}) > h^h(p) \) |
| II-B7 | Schur-convexity | \( \tilde{w} = w, s(\tilde{w}) \geq s(w) \) | \( \tilde{p} \geq p, h^h(\tilde{p}) \leq h^h(p) \) |
| II-B8 | Triangle Ineq. | \( s(. \geq 0, s(w + \tilde{w}) \leq s(w) + s(\tilde{w}) \) | \( h^h(.) \geq 0, h^h(p + \tilde{p}) \geq h^h(p) + h^h(\tilde{p}) \) |

8) Triangle inequality (II-B8):

S. If \( s > 0 \), the triangle inequality holds.

H. If \( h < 0 \), the reverse triangle inequality holds.

Lemma 2.8: If \( s(.) > 0 \), II-A4 and II-B2 imply II-B8

III. Sparsity Functions

In signal processing, sparsity functions describe the efficiency of basis representation.

The \( \ell_0 \)-pseudo-norm is known as the canonical sparsity count. It is also known as strict or hard sparsity since it counts the cardinality of the support of a function, e.g. the number of non-zero elements of a vector. In the following, this sparsity count will be denoted \( \ell_0^{\text{num}} \). Its lack of useful derivative leads to combinatorial optimization when used as the objective functional in maximizing-sparsity problems. For instance, in Compressed Sensing [25], a signal \( x \in \mathbb{R}^n \), of assumed sparse coefficients \( w \in \mathbb{R}^n \), is recovered from samples \( y \in \mathbb{R}^m \), with \( m \ll n \). The recovery is based on searching the vector \( x^\# \) of sparsest representation or minimum complexity [25].

\[
\Phi x^\# = \arg \max_w \|w\|_s \quad \text{s.t.} \quad \Pi w = y,
\]

The quotation marks (" ") help to preserve and tries to alert about the misleading notation which treats \( \ell_0 \)-pseudo- (\( \ell_1 \)-) norm as a sparsity function. In fact, \( \ell_0 \)-pseudo- (\( \ell_1 \)-) norm measures non-sparsity. This misleading notation led [10] to treat \( \ell_0 \) (no quotation marks) as a sparsity function.
Table IV
SPARSITY FUNCTIONS.

| Function | Description | $s_{pq}$ family |
|----------|-------------|-----------------|
| "$\ell_0$" or sparsity count | $-\frac{1}{n} \|w\|_0$ | not a member |
| "$\ell_1$" (proxy) | $-\frac{1}{n} \|w\|_1$ | not a sparsity func. |
| Kurtosis | $n \left( \frac{\|w\|_4}{\|w\|_2} \right)^4$ | not a member |
| Gini index | $\frac{2}{n} \left( \frac{\|w\|_1}{\|w\|_2} \right) - 1$ | $p = 1^*, q = 1$ |
| Hoyer measure | $\frac{1}{\sqrt{n}} \left( \frac{\|w\|_2}{\|w\|_1} \right)$ | $p = 1, q = 2$ |
| $pq$-means | $-n^{\frac{1}{q}} \left( \frac{\|w\|_p}{\|w\|_q} \right)$ | $p \leq 1, q > 1$ |
| $s_{1,\infty}$ or max-sparsity | $-\frac{1}{n} \|w\|_1$ | $p = 1, q = \infty$ |

Table V
TEST RESULTS OF SPARSITY FUNCTIONS. (LEMMAS 2.#, PROOFS A.##.)

| II-A1 | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ |
| II-A2 | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ |
| II-A3 | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ |
| II-A4 | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ |
| II-A5 | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ |
| II-B1 | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ |
| II-B2 | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ |
| II-B3 | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ |
| II-B4 | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ |
| II-B5 | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ |
| II-B6 | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ |

where $\|\cdot\|_s$ measures sparsity, and $\Phi$ and $\Pi$ are the sparsifying and measurement matrices, respectively.

For $[8]$ and related problems, the adoption of the $\ell_1$-norm as a proxy function allows to derive efficient algorithms and performance guarantees $[21]$. "$\ell_1$ is a convex surrogate for $\ell_0$ count. It is the best surrogate in the sense that the $\ell_1$ ball is the smallest convex body containing all 1-sparse objects of the form $\pm e_i$" $[42]$. This function will be denoted "$\ell_1$". "$\ell_1$" is not though a sparsity function $[30]$ and it should not be treated as such. However, Subsection III-B will validate its adoption in $[8]$. Other sparsity functions proposed for $[8]$ will be discussed later. Among these functions, the kurtosis $[1]$, Gini index $[4]$, Hoyer measure $[19]$ and $pq$-means $\{p \leq 1, q > 1\}$ $[30]$ (denoted $pq$-means in the following) satisfy most of the six-rules criteria of $[30]$. [41] extends these criteria and derives max-sparsity. Table [IV] defines all these functions, and Table [V] shows the results of the test using axioms and attributes of Section II.
A. The $s_{pq}$ Core Sparsity

Further inspection of sparsity functions (see Table IV) reveals a parallel in their construction [41]. Metrics of inequality of wealth or fairness, i.e. non-sparsity, exhibit the same parallel, e.g. the Theil [8] and Atkinson [9] indexes. This simple observation allows to derive a generalization.

**Theorem 3.1:** [Core sparsity] The function

$$s_{pq}(w) = -n^{-\frac{1}{q}} \frac{\|w\|_p}{\|w\|_q}, \quad \{p \leq 1, q \geq 1, p \neq q\},$$

satisfies axioms II-A1-II-A5.

Core sparsity is continuous (except at the origin), symmetric, scale-invariant and semi-strictly quasi-convex.

The kurtosis ($p = 4, q = 2$) is out of the range for $p$ and $q$ in (9). This function failed the test using criteria of [30] hence it will not be further analyzed. Nevertheless, the Gini index (weighted mean $p = 1^*$, $q = 1$) and Hoyer measure ($p = 1, q = 2$) are positive affine transformations of core sparsity. These transformations are increasing hence they preserve the quasi-convexity attribute II-B2 by Theorem A.2 of [31].

More importantly, core sparsity extends $pq$-means. [22] presents $pq$-means with $\{p \leq 1, q = 2\}$. “This, ($\ell_p$) normalization by the $\ell_2$-norm may turn out to be the best sparseness criterion. This however, has yet to be further investigated. [22]”. And although [30] grants the origin (citation principle) of $pq$-means to [22], (anonymous) reviewers suggested these functions to [30] as stated in its acknowledgments. Thus, $pq$-means have no clear origin neither a formal derivation in the literature. Thus, the Appendix derives core sparsity (which extends $pq$-means) from axioms II-A1II-A5.

A key factor of core sparsity is its resemblance to “$\ell_0$”. Fig. 1 shows this resemblance. Figs. 1(d)-1(f) show the smoothness, symmetry and quasi-convexity of sparsity functions; and Figs. 1(g)-1(l) show that the unit balls of their respective “core” functions and of “$\ell_0$” are the same. A difference between functions in Fig. 1 is their behavior at the origin. “$\ell_0$” attains its maximum at the origin, which makes sense from a practical perspective since no information is required to reconstruct null vectors. However, core sparsity is not defined at the origin due to continuity, e.g. in Fig. 1

$$0 = \lim_{\epsilon \to 0} s_{pq}((\epsilon, \epsilon)) \neq \lim_{\epsilon \to 0} s_{pq}((\epsilon, 0)) = 1.$$ (10)

Notably, $p$ and $q$ parameters localize the domain of core sparsity. Fig. 2(b) presents this localization process for $w$ following a power law decay, i.e. $w = (w_i)$, with $w_i = ri^{-\frac{1}{2}}$. Specifically, Fig. 2(a) shows the versatility of max-sparsity (and of “$\ell_1$”) to assess the sparsity of arbitrarily-sparse coefficients, and the localized domain of the kurtosis, Gini index and Hoyer measure, which are better distinguishing among highly-sparse coefficients [41]. This would suggest the (slightly) superior performance, in terms of number of measurements $m$ (the dimension of $y$), of recovery strategies based on these last functions compared with the common $\ell_1$-minimization, as reported in [26] for $\ell_p$, with $p \in (0, 1)$; in [36] for Gini index, which stochastic algorithm tends to be unstable; in [27] for kurtosis, although the function selects inappropriate basis; and in [35] for Hoyer measure in the framework of image regularization. Thus, appropriate choice of parameters $p$ and $q$ would offer a localized-sparsity formulation of [8].
since core sparsity strongly encourages coefficients of a given (or estimated) level of sparsity. Further, core sparsity is differentiable (by construction) which is appropriate for gradient-based optimization, and by the quasi-convexity, its solution to (8) is attained at an extreme point of the polytope defined by linear (equality) constraints \[32\], e.g. \(\Pi w = y\). Then, a simple exact algorithm can solve (8) by walking along the extreme points of the feasible region. Still, the potential of core sparsity in sparse recovery has not been studied.

B. Essentials of \(\ell_1\)-minimization in Compressed Sensing

\(\ell_1\) fails the test using axioms and attributes as shown in Table \(\Box\) hence it is not a sparsity function. But \(\ell_1\) is able to solve (8) in which it commonly plays the role of objective functional. Further, it offers a computational efficient convex, actually linear, formulation. Certainly core sparsity can not improve this formulation due to its quasi-convex nature, which leads to quasi-convex maximization in (8). Nevertheless, maximization of sparsity via
s_{1\infty} core sparsity or max-sparsity\(^2\) validates the adoption of “\(\ell_1\)” in (8).

**Theorem 3.2:** In linear constrained (8), \(s_{1\infty}\)-maximization simplifies to “\(\ell_1\)”-maximization, i.e. \(\ell_1\)-minimization,
\[
\begin{align*}
\max_w s_{1\infty}(w) &\equiv \min_w \|w\|_1 \\
\text{s.t. } \Pi w &\equiv y
\end{align*}
\]
(11)

**Proof 3.3:** Let \(t = \frac{1}{|w_{\text{max}}|}, u_i = tw_i,\) and \(v_i = \frac{1}{t}u_i.\) Then,
\[
\begin{align*}
\max_w \frac{1}{n}\|w\|_1 &\equiv \min_w \frac{|w_1| + \cdots + |w_n|}{|w_{\text{max}}|} \\
\text{s.t. } \Pi w &\equiv y
\end{align*}
\]
(12)
\[
\begin{align*}
\min_{|u| \leq 1} |u_1| + \cdots + |u_n| &\equiv \min \Pi u = yt \\
\text{s.t. } t &\geq 0
\end{align*}
\]
(13)
\[
\begin{align*}
\min_{|v| \leq \frac{1}{t^p}} |v_1| + \cdots + |v_n| &\equiv \min \Pi v = y \\
\text{s.t. } t &\geq 0
\end{align*}
\]
(14)

**Theorem 3.2** does not (only) state that \(s_{1\infty}\)-maximization is equivalent to \(\ell_1\)-minimization, but that \(\ell_1\)-minimization can be formulated in (8) since it is a (simple) reformulation of \(s_{1\infty}\)-maximization: essentially, the linear constraint in (8) ensures the appropriate normalization (by the maximum) that allows “\(\ell_1\)” to perform as an actual sparsity function, e.g. Proof A.30. This result clearly extends to \(\ell_p\)-minimization, with \(p \in (0, 1)\) [26], which is a reformulation of \(s_{p\infty}\)-maximization, with \(p \in (0, 1).\)

More importantly, the formalism of the present paper and the generalized convexity framework developed by [31] offer an opportunity to face the non-linear-constrained version of (8), i.e. the non-linear Compressed Sensing problem [39].

**IV. Entropy Functions**

Entropy is a concept of physics and information theory that characterizes the unpredictability of systems’ states and processes’ events.

In physics, the Tsallis entropy [11] generalizes Boltzmann-Gibbs entropy. In information theory, Rényi entropy [6] generalizes entropy functions via its parameter \(p \geq 0\): Hartley entropy or log max entropy [3] \((p = 0)\) counts the cardinality of non-zero probability events; Shannon entropy \((p = 1)\) follows a similar expression than Boltzmann-Gibbs entropy to measure the average unpredictability of events; collision entropy \((p = 2)\); log min entropy [23] \((p = \infty)\) is the most conservative entropy function since it measures the unpredictability of the most likely event. Hence log min entropy is never greater than Shannon entropy (see Figs. 3-4). Table VI defines all these functions, and Table VII shows the results of the test using axioms and attributes of Section II.

\(^2\)max-sparsity measures the non-similarity between \(w\) and \(1_n\), the least compressible vector. It was originally denoted \(s_{1\infty}\) in [41].
Table VI

### ENTROPY FUNCTIONS

| Function            | Description                                           | $h_{pq}$ family |
|---------------------|-------------------------------------------------------|-----------------|
| log max entropy     | $\log_2 ||p||_0$                                      | $p = 0$, $q = 1$|
| Shannon entropy     | $- (p, \log_2 p)$                                    | $p = 1$, $q = 1$|
| Collision entropy   | $- 2 \log_2 ||p||_2$                                  | $p = 1$, $q = 2$|
| log min entropy     | $- \log_2 ||p||_\infty$                              | $p = 1$, $q = \infty$|
| Rényi entropy       | $\frac{p}{1-p} \log_2 ||p||_p$                     | $p \neq 1$, $q = 1$|
| Tsallis entropy     | $\frac{1}{p} \left(1 - ||p||_p^p\right)$           | $p \neq 1$, $q = 1$|
| $h_{1\infty}$ or min-entropy | $\frac{1}{||p||_\infty}$                       | $p = 1$, $q = \infty$|

Table VII

### TEST RESULTS OF ENTROPY FUNCTIONS. (LEMMAS 2.##, PROOFS A.##*)

| Function | Shannon | Rényi | Tsallis | min-h | $h_{pq}$ |
|----------|---------|-------|---------|-------|----------|
| II-A1    | ✓       | ✓     | ✓       | ✓     | ✓        |
| II-A2    | ✓       | ✓     | ✓       | ✓     | ✓        |
| II-A3    | ✓       | ✓     | ✓       | ✓     | ✓        |
| II-A4    | ✓       | ✓     | ✓       | ✓     | ✓        |
| II-A5    | ✓       | ✓     | ✓       | ✓     | ✓        |
| II-B1    | ✓       | ✓     | ✓       | ✓     | ✓        |
| II-B2    | ✓       | ✓     | ✓       | ✓     | ✓        |
| II-B3    | ✓       | ✓     | ✓       | ✓     | ✓        |
| II-B4    | ✓       | ✓     | ✓       | ✓     | ✓        |
| II-B5    | ✓       | ✓     | ✓       | ✓     | ✓        |
| II-B6    | ✓       | ✓     | ✓       | ✓     | ✓        |
| II-B7    | ✓       | ✓     | ✓       | ✓     | ✓        |
| II-B8    | ✓       | ✓     | ✓       | ✓     | ✓        |

In computer science, entropy characterizes the complexity of sequences. In this context, entropy functions are called complexity algorithms. Among these algorithms, Lempel-Ziv complexity [10] characterizes the randomness of a sequence of symbols by measuring its rate of (new) pattern generation. Following a distinct approach, Approximate entropy [12] examines finite-length time series for similar epochs, and Sample entropy [17] performs similarly but without counting self-matches. Both algorithms emerged from the formulation of Kolmogorov complexity for finite-sample approximations. Axioms and attributes of Section II are not applicable to these algorithms but they will be used for comparison (see Figs. 3-4).

**A. The $h_{pq}$ Core Entropy**

Although Rényi entropy and Tsallis entropy are concepts from different fields and are applied to different objects, both functions possess the same parallel that was identified in sparsity functions (see Tables IV and VI). The key
difference between Rényi entropy and Tsallis entropy (and sparsity functions of Section III) is the logarithm transformation which provides the information units to Rényi entropy and which, at the same time, preserves the quasi-concavity attribute by Theorem A.2 of [31]. The same applies to the power function in Tsallis entropy. Thus, following Section II (which states the complementary behavior of sparsity functions and entropy functions) and Section III (which introduces core sparsity), and removing all unit transformations, allows to derive a generalization.

**Theorem 4.1:** [Core entropy] The function
\[
\left\| p \right\|_{b_{pq}} = \frac{\left\| p \right\|_p}{\left\| p \right\|_q}, \quad \{ q \ge 1, p \ne q, p \ge 1 \},
\]
(15)
satisfies axioms II-A1-II-A5.

Core entropy is continuous (except at the origin which corresponds to the empty set), symmetric and semi-strictly quasi-concave. It follows the quotient-of-weighted-functions form of general entropy functions as in [14], e.g. Aczel-Daróczy entropy [7].

Interestingly, the application of function \( a(.) = -1 \) to core entropy increases the domain of its parameters \( p \) and \( q \) to \( \{ p \ge 1, q \le 1, p \ne q \} \) and makes core entropy negative. Then, the Tsallis entropy \( (p \ne 1, q = 1) \), Rényi entropy \( (p \ne 1, q = 1) \) and their special cases (Shannon, etc.) are transformations (by the logarithm and power functions) of core entropy. All these previous transformations (including \( a(.) \)) are increasing hence they preserve the (now) quasi-concavity attribute II-B2 by Theorem A.2 of [31]. Note that the domain extension of parameters \( p \) and \( q \) to \( \{ p \ge 1, q \le 1, p \ne q \} \) also applies to core sparsity, which will be now positive.

Some features of core entropy are shown in Fig. 3 which presents the entropy of a r.v. following the Bernoulli distribution with parameter \( \pi \), the canonical example of information theory textbooks [24]. Fig. 3(a) shows the smoothness, monotonicity, symmetry and quasi-concavity of entropy functions and complexity algorithms. Fig. 3(b) shows the same features in core entropy. Setting parameter \( p = 1 \) allows to appreciate the most interesting case: core entropy tends to Shannon entropy as \( q \downarrow 1 \). Unsurprisingly, the same tendency is present for \( q = 1 \) fixed and \( p \uparrow 1 \).

For general univariate distributions, [16] identifies conditions under which entropy defines an order (on distributions). Fig. 4 shows this ordering phenomenon for several distributions. For a better visual comparison, these
test distributions do not include the uniform distribution which reaches the maximum in entropy functions. Fig. 4(a) shows the ordering obtained using entropy functions and complexity algorithms. Although its simplicity, $h_{1\infty}$ core entropy or min-entropy respects the same trend ordering reported by other functions. As in Fig. 3(b) Fig. 4(b) shows that the ordering trend drawn by core entropy, with $p = 1$ fixed and $q \downarrow 1$, tends to resulting ordering obtained using Shannon entropy.

**B. On Relative and Absolute Functions**

Core sparsity is a relative measure and core entropy is an absolute measure. The following discusses this subtle difference and its consequences on the way both functions follow the axioms (and attributes) of Section II.

For entropy functions, replication II-A5 does not refer to a larger number of events or states. As the number of “replicas” of $p$ goes to infinity, replication II-A5 affirms that entropy is intrinsic to a process (system). Hence entropy functions measure an absolute property as is the uncertainty of a process (system) of fixed possible events (states). Analogously, sparsity functions measure a relative property as is the efficiency of a basis representation or compressibility. For example, coefficients $w = (1, 0)$ and $\tilde{w} = (1, 1, 0, 0)$ both require 50% of their content for an exact representation.

This subtle difference is characterized by the dimension normalization $n^{\frac{1}{q}-\frac{1}{p}}$, which is present only in core sparsity (cf. Proof A.10). Furthermore, (discrete) entropy is defined for discrete r.v.s on countable sample spaces which cardinality could be infinity, i.e. $b \rightarrow \infty$. Thus, normalization of core entropy by $b^{\frac{1}{q}-\frac{1}{p}}$ would not be possible in general.

More importantly, replication II-A5 and completeness II-B4 should not be confused with the following two common axioms of entropy functions [28].

**Lemma 4.2:** Core entropy satisfies

$$h_{pq}^b(p) = h_{pq}^{b+1}(p\|0).$$

**Proof 4.3:** Trivial by the definition of core entropy.
**Lemma 4.4:** Core entropy satisfies, for \( m < n \),

\[
h_{pq}^m(1_m) < h_{pq}^n(1_n).
\]

**Proof 4.5:**

\[
h_{pq}^m(1_m) \leq h_{pq}^n(1_m) \|0_{n-m}) < h_{pq}^n(1_n),
\]

which is also obvious given \( 1_m \|0_{n-m} \) is more “compressible”, hence less uncertain, than \( 1_n \). Or by simple calculation: 
\[
h_{pq}^m(1_m) = m^{\frac{1}{p} - \frac{1}{q}}, \quad h_{pq}^n(1_n) = n^{\frac{1}{p} - \frac{1}{q}}.
\]

**C. Sparse Recovery of Probability Measures**

Consider the minimum (Shannon) entropy problem on the probability simplex with sample moment constraints \[34\],

\[
\begin{aligned}
\min_{p \geq 0, \|p\|_1 = 1} & \quad \exp(h_{Shannon}^b(p)) \\
\text{s.t.} & \quad \mathbb{E}_p[X^k] = \mu_k, \quad 1 \leq k \leq m - 1,
\end{aligned}
\]

where \( \mathbb{E}_p \) denotes expectation with respect to desired \( p \).

Write \( A = (a_{ij}) \), with \( a_{ij} = x_j^{i-1}, \quad 1 \leq j \leq b \) and \( 1 \leq i \leq m \); and sample moments \( \mu = (1, \mu_1, \ldots, \mu_{m-1}) \). Since \( \log \max \) entropy upper bounds Shannon entropy, the solution of (20) is upper-bounded by the solution of

\[
\begin{aligned}
\min_{p \geq 0} & \quad h_{max}^b(p) \\
\text{s.t.} & \quad Ap = \mu
\end{aligned}
\]

where \( h_{max}^b \) denotes \( h_{01}^b \) core entropy or max-entropy. Deterministic measurement matrix \( A \) is Vandermonde \[29\], and linear constraint \( Ap = \mu \) includes the \( \|p\|_1 = 1 \) normalization. Then, under the common Compressed Sensing hypotheses, (21) is equivalent to

\[
\begin{aligned}
\min_{p \geq 0} & \quad \|p\|_1 \\
\text{s.t.} & \quad Ap = \mu
\end{aligned}
\]

However, formulation (22) is useless in this case since it produces a feasibility problem which solution is not unique if the problem is under-determined. Nevertheless, by the linearity of the constraints and Theorem 3.2, (22) is equivalent to \( s_{1\infty} \) maximization, which is equivalent to minimization of \( h_{1\infty}^b \) or min-entropy\( ^3 \) denoted in the following \( h_{min}^b \).

**Theorem 4.6:** Under an appropriate Vandermonde matrix \( A \), min-entropy minimization solves (20) uniquely.

\(^3\)min-entropy measures the similarity between \( p \) and \( 1_b \), the most uncertain distribution, i.e. uniform distribution \( \) \[41\].
Proof 4.7: Given the linear constraints of (20),

\[
\min h_{\min}^b(p) \leq \min h_{\text{Shannon}}^b(p) \leq \min h_{\max}^b(p).
\]

Hence, (21) not only upper bounds (20) (cf. [37]).

V. CONCLUSIONS

A complete characterization of sparsity and entropy offered new functions and tools to solve efficiently problems of sparse recovery. This paper proposed a formalism, a joint axiomatic characterization, for sparsity and entropy. The proposed set of axioms is constructive and allows to derive the core functions, core sparsity and core entropy. Both core functions are simple functional forms found in well-known sparsity and entropy functions. Finally, core functions were applied to the sparse recovery problem where they offered efficient formulations. More importantly, the (simplest) core sparsity validates the use of the \( \ell_1 \)-norm as a (non-)sparsity measure in compressed sensing problem which further provides insights to this problem.

Future work will follow two directions. Concerning entropy, to study the compatibility of proposed axioms (and attributes) and specific axioms of entropy, especially those related to relative and conditional entropy. Some of these specific axioms were already verified here. Concerning sparsity, to study the potential of core sparsity in a general optimization framework, where core sparsity can be “tuned” to match the optimal recovery strategy according to a sparsity level.

APPENDIX

In this appendix Lemmas 2.1-2.8 are proved, which derive attributes II-B1-II-B8 from axioms II-A1-II-A5. The appendix also contains the proofs listed in Tables V and VII where functions, applied to non-zero vectors, are evaluated using axioms and attributes of Section II. These proofs give a refinement of the formalism compared to proofs of [30] of which some proofs are corrected.

Tables VIII IX describe the notation adopted in the proofs related to sparsity and entropy functions, respectively. The proofs are based on the following key results of [31].

Theorem A.1: [31] Let \( f, g \) be defined on \( S \) convex, and

\[
s(w) = \frac{f(w)}{g(w)}.
\]

If any of the following properties hold

1) \( f \) non-negative convex, \( g \) positive concave; or
2) \( f \) non-positive convex, \( g \) positive convex; or
3) \( f \) convex, \( g \) positive affine.

Then, \( s \) is semi-strictly quasi-convex on \( S \).
Theorem A.2: Let $s$ be semi-strictly quasi-convex (-concave) defined on $S \subseteq \mathbb{R}^n$ convex. Let $a : A \rightarrow \mathbb{R}$ be increasing, with $s(S) \subseteq A$. Then, $a \circ s$ is semi-strictly quasi-convex (-concave) on $S$.

Examples of function $a(.)$ are the $\log(.)$ and $-\frac{1}{n}$ (which changes the sign and allows to prove Lemma 2.8).

A. Proofs of Lemmas

Proof A.3: [Lem. 2.1, Bounds] Let $w$ arbitrary with $\|w\| = 1$. Construct $\hat{w} = 0_{n-1} \parallel 1$ and $\tilde{w} = \frac{1}{n} 1_n$ from $w$ via concentration actions II-A3. Then,

$$s(\hat{w}) \leq s(w) \overset{II-A3}{\leq} s(\tilde{w})$$

and by scaling II-A4

$$s(\alpha 1_n) \overset{II-A3}{\leq} s(\hat{w})$$

$$s(\tilde{w}) \overset{II-A3}{\leq} s(0_{n-1} \parallel \alpha).$$

Proof A.4: [Lem. 2.2, Quasi-convexity] Let $s$ continuous II-A1, symmetric II-A2, homogeneous of degree 0 II-A4 and non-negative. Then, $s$ can be written as $f \circ g$, with $f$ and $g$ continuous, symmetric, homogeneous of degree 1 and non-negative. Construct $\tilde{w}$ from $w$ via a concentration action II-A3. Then, it holds that

$$f(\tilde{w}) > f(w)$$

and

$$g(\tilde{w}) < g(w).$$

For function $f$ (differentiable)

$$f(w) - f(\tilde{w}) \approx -\alpha \nabla_{e_i - e_j} f(w)$$

$$= \alpha \left( \frac{\partial f}{\partial w_j}(w) - \frac{\partial f}{\partial w_i}(w) \right).$$

Then, by the symmetry of $f$ and (28), each component function $f_k$ of $f$ satisfies

$$\lim_{\alpha \to 0} \frac{f_k(w_j) - f_k(w_j + \alpha)}{\alpha} < \lim_{\alpha \to 0} \frac{f_k(w_i) - f_k(w_i + \alpha)}{\alpha}$$

with $w_i > w_j$, which implies that $f$ is convex. Similarly, $g$ is concave. Now, by Theorem A.1, $s = \frac{f}{g}$ is semi-strictly quasi-convex. A similar proof holds if $s$ non-positive.

Proof A.5: [Lem. 2.3, Monotonicity] Construct $\tilde{w} = (\hat{w}, \alpha - \tilde{w})$ from $w = (w, \alpha - w)$ via concentration actions II-A3 with $\frac{\alpha}{2} \leq w \leq \tilde{w} \leq \alpha$.

Proof A.6: [Lem. 2.4, Completeness] Let $w$ arbitrary. Then,

$$s(w) \overset{II-A3}{\leq} s(w \parallel \ldots \parallel w) \overset{n+1 \text{ times}}{\leq} \overset{II-A3}{\leq} s \left( \left( 1 + \frac{1}{n} \right) w \parallel \ldots \parallel w \| 0_n \right)$$

$$\overset{II-A3}{\leq} s(\|w\| \ldots \|w\| 0_n) \overset{II-A3}{\leq} s(\|w\| 0).$$
Proof A.7: [Lem. 2.5 Regularity] Let \( \tilde{w} = [w_1, \ldots, w_i + \beta, \ldots, w_n] \) and \( 0 < \epsilon = 1 + \delta \). Construct \( \hat{w} \) from \( w = \epsilon \tilde{w} \) via concentration actions [II-A3]

\[
\hat{w}_j = \epsilon w_j - \delta w_j = w_j, \quad j \neq i, \quad (36)
\]

\[
\hat{w}_i = \epsilon (w_i + \beta) + \sum_{j \neq i} \delta w_j = w_i + \beta + \alpha, \quad (37)
\]

where \( \alpha = \delta (\beta + \sum_j w_j) \). Then,

\[
s(w^0)_{II-A4} \leq s(w^1)_{II-A3} < s(w^2). \quad (39)
\]

Proof A.8: [Lem. 2.6 Hom. growth] Let \( w \) arbitrary, \( \tilde{w} = 1_n \) and \( \alpha < 1 \). Then,

\[
s\left( w + \frac{1 - \alpha}{\alpha} \tilde{w} \right)_{II-A4} \leq s(\alpha w + (1 - \alpha) \tilde{w})_{II-B2} \leq \max\{s(w), s(\tilde{w})\} \quad (41)
\]

\[
= s(w). \quad (42)
\]

Proof A.9: [Lem. 2.8 Triangle ineq.] Let \( w \) and \( \tilde{w} \) arbitrary and \( \alpha < 1 \),

\[
s(\alpha w + (1 - \alpha) \tilde{w})_{II-B3} \leq \max\{s(w), s(\tilde{w})\} \quad (44)
\]

\[
s > 0 \leq s(w) + s(\tilde{w}) \quad (45)
\]

\[
= s(\alpha w) + s((1 - \alpha) \tilde{w}). \quad (46)
\]

B. Derivation of \( s_{pq} \) Core Functions

Proof A.10: [Thm. 3.1 Thm. 4.1] Since concentration [II-A3] implies quasi-convexity [II-B2] Theorems [A.1] and [A.2] with \( a(\cdot) = -\frac{1}{\alpha} \) increasing, state that sparsity \( s \) can be written as

\[
s(w) = -\xi(n) \frac{g(w)}{f(w)} \quad (47)
\]

with \( \xi(n) > 0, \forall n \), responsible of the dimension-invariance of \( s \) or replication [II-A5] and \( g > 0 \) concave and \( f > 0 \) convex, and both functions continuous [II-A1] symmetric [II-A2] and homogeneous of degree 1, such that \( s \) is homogeneous of degree 0 or scale-invariant [II-A4]. By the homogeneity, \( f(0_n) = 0 \) and \( g(0_n) = 0 \) (hence \( s \) cannot be defined at the origin). The previous description of \( f \) and \( q \) resembles vector norms and leads to the pair of candidate functions

\[
f(\cdot) = \|\cdot\|_q, \quad q \geq 1, \quad (48)
\]

\[
g(\cdot) = \|\cdot\|_p, \quad p \leq 1, \quad (49)
\]

with \( p \neq q \). Finally, normalization by \( \xi(n) = n^{\frac{1}{q} - \frac{1}{p}} \) allows \( s \) to satisfy replication [II-A5].

A similar proof holds for core entropy which does not need this last normalization.
Table VIII

| Table VIII |
|-----------------------------------------|
| **Summary of notation used in proofs related to sparsity functions.** |
| Assumptions | Desired result |
| Continuity | \( \tilde{w} \rightarrow w \) |
| Symmetry | \( \pi(w) \rightarrow s(\tilde{w}) = s(w) \) |
| Concentration | \( \tilde{w}_i = w_i + \alpha, \tilde{w}_j = w_j - \alpha \rightarrow s(\tilde{w}) > s(w) \) |
| Scaling | \( \alpha w \rightarrow s(\tilde{w}) = s(w) \) |
| Replication | \( w \rightarrow \infty \rightarrow n \rightarrow \infty \rightarrow s(\tilde{w}) \rightarrow s(w) \) |
| Bounds | \( \alpha < 1, s(w) \leq s(\tilde{w}) \rightarrow s(\tilde{w}) > s(w) \) |
| Quasiconvex | \( \tilde{w}_i = w_i + \beta, \tilde{w}_j = w_j + \beta + \alpha \rightarrow s(\tilde{w}) < s(w) \) |
| Completeness | \( w \rightarrow \infty \rightarrow s(\tilde{w}) > s(w) \) |

**C. Proofs for Sparsity Functions**

Proof A.11: [Continuity, symmetry] By the continuity and symmetry of \( \ell_p \)-norms.

1) **Proofs for \( \ell_0 \):** Denoted \( s_0 \) in the following.

Proof A.12: [Concentration](cf. correction of [30].)

\[
\| \tilde{w} \|_0 = \begin{cases} 
\| w \|_0 - 1, & \text{new null coefficient,} \\
\| w \|_0, & \text{otherwise}
\end{cases}
\]

i.e. \( s_0(\tilde{w}) \geq s_0(w) \).

Proof A.13: [Scaling](cf. correction of [30].)

\[
\| \tilde{w} \|_0 = \| w \|_0,
\]

i.e. \( s_0(\tilde{w}) = s_0(w) \).

Proof A.14: [Replication](cf. correction of [30].)

\[
\| \tilde{w} \|_0 = m \| w \|_0,
\]

i.e. \( s_0(\tilde{w}) = s_0(w) \).

Proof A.15: [Bounds]

\[
s_0(\tilde{w}) = -1,
\]

\[
s_0(\tilde{w}) = -\frac{1}{n},
\]

i.e. \( s_0(\tilde{w}) < s_0(w) \).

Proof A.16: [Quasi-convexity]

\[
\| \tilde{w} \|_0 \geq \min\{ \| w \|_0, \| \tilde{w} \|_0 \},
\]
i.e. \( s_0(\hat{w}) \leq \max\{s_0(w), s_0(\tilde{w})\} \).

Proof A.17: [Monotonicity]

\[
\|\tilde{w}\|_0 = \begin{cases} 2, & \tilde{w} < \alpha, \\ 1, & \text{otherwise}, \end{cases}
\]

i.e. \( s_0(\hat{w}) \leq s_0(\tilde{w}) \).

Proof A.18: [Completeness] (cf. correction of [30].)

\[
\|\tilde{w}\|_0 = \|w\|_0,
\]

i.e. \( s_0(\hat{w}) > s_0(w) \).

Proof A.19: [Regularity] (cf. correction of [30].)

\[
\|\tilde{w}\|_0 = \|\hat{w}\|_0,
\]

i.e. \( s_0(\hat{w}) = s_0(\tilde{w}) \).

Proof A.20: [Hom. growth] (cf. correction of [30].)

\[
\|\tilde{w}\|_0 = \begin{cases} \|w\|_0 + m, & (m) \text{ new non-null coefficients,} \\ \|w\|_0, & \text{otherwise}, \end{cases}
\]

i.e. \( s_0(\hat{w}) \leq s_0(w) \).

2) Proofs for "\( \ell_1 \)": Denoted \( s_1 \) in the following.

Proof A.21: [Concentration] (cf. counterexample in [30].)

\[
\|\tilde{w}\|_1 = \|w\|_1,
\]

i.e. \( s_1(\hat{w}) = s_1(w) \).

Proof A.22: [Scaling] (cf. counterexample in [30].)

\[
\|\tilde{w}\|_1 = \alpha\|w\|_1,
\]

i.e. \( s_1(\hat{w}) = \alpha s_1(w) \).

Proof A.23: [Replication] (Correction of [30].)

\[
\|\tilde{w}\|_1 = m\|w\|_1,
\]

i.e. \( s_1(\hat{w}) = s_1(w) \).

Proof A.24: [Bounds]

\[
s_1(\tilde{w}) = s_1(\hat{w}) = \frac{\alpha}{n},
\]

i.e. \( s_1(\hat{w}) = s_1(\tilde{w}) \).
Proof A.25: [Quasi-convexity] Assume \( \|\tilde{w}\|_1 \leq \|w\|_1 \), then, \( s_1(\tilde{w}) \geq s_1(w) \). Since \( w, \tilde{w} \) non-negative,
\[
\|\tilde{w}\|_1 \leq \alpha \|\tilde{w}\|_1 + (1 - \alpha) \|w\|_1
\]
\[
= \|\tilde{w}\|_1,
\]
i.e. \( s_1(\tilde{w}) \leq s_1(\tilde{w}) = \max\{s_1(w), s_1(\tilde{w})\} \).

Proof A.26: [Monotonicity]
\[
\|\tilde{w}\|_1 = \|\hat{w}\|_1,
\]
i.e. \( s_1(\tilde{w}) = s_1(\hat{w}) \).

Proof A.27: [Completeness](Correction of [30].)
\[
\|\tilde{w}\|_1 = \|w\|_1,
\]
i.e. \( s_1(\tilde{w}) > s_1(w) \).

Proof A.28: [Regularity](cf. counterexample in [30].)
\[
\|\tilde{w}\|_1 = \|\tilde{w}\|_1 + \alpha,
\]
i.e. \( s_1(\tilde{w}) > s_1(\hat{w}) \).

Proof A.29: [Hom. growth](cf. obvious in [30].)
\[
\|\tilde{w}\|_1 = \|w\|_1 + n\alpha,
\]
i.e. \( s_1(\tilde{w}) < s_1(w) \).

Proof A.30: [Triangle ineq.] Notice that \( \alpha, \beta \) such that \( \tilde{s}_1(v) = \alpha s_1(v) + \beta \geq 0 \), \( \forall v \) without knowledge of the maximum of \( w \) (\( \sup |w| = \|w\|_\infty \)). Then,
\[
\|\tilde{w}\| \leq \|w\|_1 + \|\tilde{w}\|_1,
\]
i.e. \( s_1(\tilde{w}) \geq s_1(w) + s_1(\tilde{w}) \).

3) Proofs for Hoyer:

Proof A.31: [Concentration](Correction of [30].)
\[
g(\alpha) = s_{\text{Hoyer}}(\tilde{w}(\alpha)) - s_{\text{Hoyer}}(w),
\]
where \( g(0) = 0 \). Then,
\[
g'(\alpha) = \frac{\|u\|_1}{\sqrt{n} - 1} \left( \frac{-1}{\|\tilde{w}(\alpha)\|_2} \right)'
\]
\[
= \frac{1}{\sqrt{n} - 1} \frac{\|w\|_1}{\|w\|_2^2} (w_i - w_j + 2\alpha) > 0,
\]
since \( w_i + \alpha > w_j - \alpha \), i.e. \( s_{\text{Hoyer}}(\tilde{w}) > s_{\text{Hoyer}}(w) \).

Proof A.32: [Scaling](cf. obvious in [30].)
\[
\frac{\|\tilde{w}\|_1}{\|w\|_2} = \frac{\|\tilde{w}\|_1}{\|w\|_2},
\]
i.e. $s_{\text{Hoyer}}(\hat{w}) = s_{\text{Hoyer}}(w)$.

**Proof A.33** [Replication](cf. counterexample in [30].)

$$s_{\text{Hoyer}}(\hat{w}) = \frac{\sqrt{n} - 1}{\sqrt{n} - \sqrt{n-1}} s_{\text{Hoyer}}(w),$$

(76)

where $\frac{\sqrt{n} - 1}{\sqrt{n} - \sqrt{n-1}} \to 1$ as $n \to \infty$, i.e. $s_{\text{Hoyer}}(\hat{w}) \to s_{\text{Hoyer}}(w)$.

**Proof A.34** [Bounds]

$$s_{\text{Hoyer}}(\hat{w}) = 0,$$

(77)

$$s_{\text{Hoyer}}(\hat{w}) = 1,$$

(78)

i.e. $s_{\text{Hoyer}}(\hat{w}) < s_{\text{Hoyer}}(\bar{w})$.

**Proof A.35** [Quasi-convexity] Let $p = 1$, $q > 1$, e.g. Hoyer with $q = 2$, $s_{1,\infty}$ with $q = \infty$, then

- $f(.) = \|\cdot\|_q$ convex,
- $g(.) = \|\cdot\|_p$ affine (or concave),
- by Theorem A.13 (or 1), $z(.) = \frac{f(.)}{g(.)}$ is semi-strictly quasi-convex,
- and, by Theorem A.2 with $a(.) = -\frac{1}{q}$, $a \circ z(.) = -\frac{g(.)}{f(.)}$ is semi-strictly quasi-convex.

**Proof A.36** [Monotonicity]

$$g(\alpha) = s_{\text{Hoyer}}(\hat{w}(\alpha)) - s_{\text{Hoyer}}(\bar{w}(\alpha))$$

(79)

$$= \frac{\alpha}{\sqrt{n} - 1} \left( \frac{1}{\|\cdot\|_2} - \frac{1}{\|\cdot\|_2} \right) > 0,$$

(80)

since $\|\cdot\|_2$ convex and $\frac{\alpha}{2} \leq w_1 < w_2 \leq \alpha$, i.e. $s_{\text{Hoyer}}(\hat{w}) > s_{\text{Hoyer}}(\bar{w})$.

**Proof A.37** [Completeness](cf. obvious in [30].)

$$g(w) = s_{\text{Hoyer}}(\hat{w}(w)) - s_{\text{Hoyer}}(w)$$

(81)

$$= \frac{\sqrt{n} + 1 - \sqrt{n}}{(\sqrt{n} + 1 - 1)(\sqrt{n} - 1)} \left( \|w\|_1 - \|w\|_2 \right) > 0,$$

(82)

since $\|v\|_2 < \|v\|_1$, i.e. $s_{\text{Hoyer}}(\hat{w}) > s_{\text{Hoyer}}(w)$.

**Proof A.38** [Regularity](Correction of [30].)

$$g(\alpha) = s_{\text{Hoyer}}(\hat{w}(\alpha)) - s_{\text{Hoyer}}(\bar{w}),$$

(83)

where $g(0) = 0$. Then,

$$g'(\alpha) = \frac{-1}{\sqrt{n} - 1} \left( \frac{\|w\|_1 + \alpha}{\|w(\alpha)\|_2} \right)'$$

(84)

$$= \frac{\|w\|_1 (\|w\|_\infty + \alpha) - \|w\|_2 - \|w\|_\infty \cdot \alpha}{(\sqrt{n} - 1)\|w\|_2^2} > 0,$$

(85)

since $\|v\|_2^2 < \|v\|_\infty \|v\|_1$ and $\|v\|_1 > \|v\|_\infty$, i.e. $s_{\text{Hoyer}}(\hat{w}) > s_{\text{Hoyer}}(\bar{w})$.

**Proof A.39** [Hom. growth](cf. [30].)

$$g(\alpha) = s_{\text{Hoyer}}(w) - s_{\text{Hoyer}}(\hat{w}(\alpha)),$$

(86)
where \( g(0) = 0 \). Then,
\[
\begin{align*}
g'(\alpha) &= \frac{1}{\sqrt{n} - 1} \left( \frac{\| w \|_1 + n\alpha}{\| \tilde{w}(\alpha) \|_2} \right)' \\
&= \frac{1}{\sqrt{n} - 1} \frac{n\| w \|_2^2 - \| w \|_1^2}{\| \tilde{w} \|_2^3} > 0,
\end{align*}
\]
(87)
since \( \| v \|_1 < \sqrt{n} \| v \|_2 \), i.e. \( s_{\text{Hoyer}}(w) > s_{\text{Hoyer}}(\tilde{w}) \).

4) **Proofs for Gini:** Recall that the order of indexes changes for this function, i.e. \( w_i > w_j \) as \( i > j \).

**Proof A.40:** [Concentration](cf. [30].)
\[
g(\alpha) = s_{\text{Gini}}(\tilde{w}(\alpha)) - s_{\text{Gini}}(w),
\]
(89)
where \( g(0) = 0 \). Then,
\[
\begin{align*}
g'(\alpha) &= \frac{2}{(1 + n)\| w \|_1} (\| \tilde{w}(\alpha) \|_1')' \\
&= \frac{2}{(1 + n)\| w \|_1} (i^*(w_i + \alpha) + j^*(w_j - \alpha))' \\
&= \frac{2}{(1 + n)\| w \|_1} (i^* - j^*) > 0,
\end{align*}
\]
(90)
(91)
(92)
where \( i^* \geq i \) and \( j^* \leq j \) are the new indexes (and weights) of coefficients \( w_i + \alpha \) and \( w_j - \alpha \), respectively; i.e. \( s_{\text{Gini}}(\tilde{w}) > s_{\text{Gini}}(w) \).

**Proof A.41:** [Scaling](cf. [30].)
\[
\frac{\| \tilde{w} \|_1}{\| w \|_1} = \frac{\| w \|_1}{\| w \|_1},
\]
(93)
i.e. \( s_{\text{Gini}}(\tilde{w}) = s_{\text{Gini}}(w) \).

**Proof A.42:** [Replication](Correction of [30].)
\[
s_{12}(\tilde{w}) = \frac{2}{1 + mn} \sum_{i=1}^{n} \sum_{k=(i-1)m+1}^{im} w_i - 1
\]
(94)
\[
= \frac{(1 + n)m}{1 + mn} s_{12}(w),
\]
(95)
where \( \frac{(1 + n)m}{1 + mn} \rightarrow 1 \) as \( n \rightarrow \infty \), i.e. \( s_{\text{Gini}}(\tilde{w}) \rightarrow s_{\text{Gini}}(w) \).

**Proof A.43:** [Bounds]
\[
s_{\text{Gini}}(\tilde{w}) = 0,
\]
(96)
\[
s_{\text{Gini}}(\tilde{w}) = \frac{n - 1}{n + 1},
\]
(97)
i.e. \( s_{\text{Gini}}(\tilde{w}) < s_{\text{Gini}}(w) \).

**Proof A.44:** [Quasi-convexity] Let \( p = 1^* \), \( q = 1 \), e.g. Gini, then
- \( f(.) = \| . \|_p \) affine (and convex)
- \( g(.) = \| . \|_q \) affine
- by Theorem [A.1(3)], \( z(.) = \frac{f(.)}{g(.)} \) is semi-strictly quasi-convex.
Proof A.45: [Monotonicity]

\[ g(\alpha) = s_{\text{Gini}}(\hat{w}(\alpha)) - s_{\text{Gini}}(\tilde{w}(\alpha)) \]
\[ = \frac{2(\hat{w} - \tilde{w})}{\alpha(1 + n)} > 0, \]

since \( \hat{w} > \tilde{w} \), i.e. \( s_{\text{Gini}}(\hat{w}) > s_{\text{Gini}}(\tilde{w}) \).

Proof A.46: [Completeness](cf. [30].)

\[ g(w) = s_{\text{Gini}}(\tilde{w}) - s_{\text{Gini}}(w), \]
\[ = \frac{2}{1 + n} \left( \frac{\sum_{i=1}^{n}(i + 1)w_i}{\|w\|_1} - \frac{\sum_{i=1}^{n}iw_i}{\|w\|_1} \right), \]
\[ = \frac{2}{1 + n} > 0, \]

i.e. \( s_{\text{Gini}}(\tilde{w}) > s_{\text{Gini}}(w) \).

Proof A.47: [Regularity](cf. [30].)

\[ g(\alpha) = s_{\text{Gini}}(\hat{w}(\alpha)) - s_{\text{Gini}}(\tilde{w}), \]

where \( g(0) = 0 \). Then,

\[ g'(\alpha) = \frac{2}{1 + n} \left( \frac{\|\tilde{w}\|_1 + n\alpha}{\|\tilde{w}\|_1 + \alpha} \right)' \]
\[ = \frac{2}{(1 + n)\|\tilde{w}\|_1^2} (n\|\tilde{w}\|_1 - \|\tilde{w}\|_1^*) > 0, \]

since \( n\|v\|_1 - \|v\|_1^* \), i.e. \( s_{\text{Gini}}(\tilde{w}) > s_{\text{Gini}}(\tilde{w}) \).

Proof A.48: [Hom. growth](cf. [30].)

\[ g(\alpha) = s_{\text{Gini}}(w) - s_{\text{Gini}}(\tilde{w}(\alpha)), \]

where \( g(0) = 0 \). Then,

\[ g'(\alpha) = \frac{-2n}{n + 1} \left( \frac{\|w\|_1 + \sum_{i=1}^{n}i\alpha}{\|w\|_1 + n\alpha} \right)' \]
\[ = \frac{2n}{(n + 1)\|\tilde{w}\|_1} \sum_{i=1}^{n}w_i(2i - 1 - n) \]
\[ = \frac{2n\|\tilde{w}\|_1^2}{(n + 1)} \sum_{i=1}^{\lfloor n\rfloor}(w_{n-i+1} - w_i)(n + 1 - 2i) > 0, \]

since \( w_{n-i+1} > w_i \) and \( n + 1 > 2i \) with \( i \leq \lfloor n/2 \rfloor \), i.e. \( s_{\text{Gini}}(w) > s_{\text{Gini}}(\tilde{w}) \).

5) Proofs for \( s_{1,\infty} \):

Proof A.49: [Concentration]

\[ g(\alpha) = s_{1,\infty}(\hat{w}(\alpha)) - s_{1,\infty}(w) \]
\[ = \frac{\|w\|_1}{n} \left( \frac{1}{\|w\|_\infty} - \frac{1}{\|\tilde{w}\|_\infty} \right) \geq 0, \]
since $\|\tilde{w}\|_\infty \geq \|w\|_\infty$, i.e. $s_{1\infty}(\tilde{w}) \geq s_{1\infty}(w)$.

Proof A.50: [Scaling]

$$\frac{\|\tilde{w}\|_1}{\|\tilde{w}\|_\infty} = \frac{\|w\|_1}{\|w\|_\infty},$$

(112)
i.e. $s_{1\infty}(\tilde{w}) = s_{1\infty}(w)$.

Proof A.51: [Replication]

$$s_{1\infty}(\tilde{w}) = s_{1\infty}(w).$$

(113)

Proof A.52: [Bounds]

$$s_{1\infty}(\tilde{w}) = 0$$

(114)

$$s_{1\infty}(\tilde{w}) = 1 - \frac{1}{n},$$

(115)
i.e. $s_{1\infty}(\tilde{w}) < s_{1\infty}(w)$. Further, let arbitrary $\tilde{w}$ with $\|\tilde{w}\|_1 = \beta$ and $\|\tilde{w}\|_\infty = \lambda\beta$, $\frac{1}{n} \leq \lambda \leq 1$. Then,

$$s_{1\infty}(\tilde{w}) = 1 - \frac{1}{\lambda n},$$

(116)
i.e. $s_{1\infty}(\tilde{w}) \leq s_{1\infty}(\tilde{w}) \leq s_{1\infty}(w)$.

Proof A.53: [Monotonicity]

$$s_{1\infty}(\tilde{w}) = 1 - \frac{\alpha}{2\tilde{w}},$$

(117)

$$s_{1\infty}(\tilde{w}) = 1 - \frac{\alpha}{2\tilde{w}},$$

(118)
i.e. $s_{1\infty}(\tilde{w}) < s_{1\infty}(\tilde{w})$.

Proof A.54: [Completeness]

$$g(w) = s_{1\infty}(\tilde{w}(w)) - s_{1\infty}(w)$$

$$= \frac{1}{n(n+1)} \|w\|_1 > 0,$$

(120)
i.e. $s_{1\infty}(\tilde{w}) > s_{1\infty}(w)$.

Proof A.55: [Regularity]

$$g(\alpha) = s_{1\infty}(\tilde{w}(\alpha)) - s_{1\infty}(\tilde{w}),$$

$$= \frac{\alpha(\|\tilde{w}\|_1 - \|\tilde{w}\|_\infty)}{n\|\tilde{w}\|_\infty} > 0,$$

(122)
since $\|v\|_1 > \|v\|_\infty$, i.e. $s_{1\infty}(\tilde{w}) > s_{1\infty}(\tilde{w})$.

Proof A.56: [Hom. growth]

$$g(\alpha) = s_{1\infty}(w) - s_{1\infty}(\tilde{w}(\alpha))$$

$$= \frac{\alpha(n\|w\|_\infty - \|w\|_1)}{n\|w\|_\infty\|\tilde{w}\|_\infty} > 0,$$

(124)
since $n\|v\|_\infty > \|v\|_1$, i.e. $s_{1\infty}(w) > s_{1\infty}(\tilde{w})$. 

January 23, 2015
6) Proofs for $s_{pq}$ compressibility: Let $p \leq 1 < q$.

Proof A.57: [Concentration](Correction of [30].)

\begin{equation}
\begin{aligned}
g(\alpha) &= s_{pq}(\tilde{w}(\alpha)) - s_{pq}(w) \\
&\approx \alpha \nabla_{e_1-e_j}s_{pq}(w),
\end{aligned}
\end{equation}

where $g(0) = 0$, $\nabla_{e_1-e_j}f(v) = \frac{\partial f}{\partial w_1}(v) - \frac{\partial f}{\partial w_j}(v)$, and

\begin{equation}
\begin{aligned}
\frac{\partial s_{pq}}{\partial w_k}(w) &= n^{\frac{1}{q}} - \frac{1}{q} \left( \frac{w^{p-1}_k}{\|w\|_q^q} - \frac{w^{p-1}_k}{\|w\|_p^p} \right)
\end{aligned}
\end{equation}

positive, since $v^q_k - \|v\|_p^q > v^{p-1}_k - \|v\|_q^q$ (expand to see); i.e. $s_{pq}(\tilde{w}) > s_{pq}(w)$.

Proof A.58: [Scaling](cf. [30].)

\begin{equation}
\begin{aligned}
\frac{\|\tilde{w}\|_p}{\|\tilde{w}\|_q} = \frac{\|w\|_p}{\|w\|_q}
\end{aligned}
\end{equation}

i.e. $s_{pq}(\tilde{w}) = s_{pq}(w)$.

Proof A.59: [Replication](cf. [30].)

\begin{equation}
\begin{aligned}
s_{pq}(\tilde{w}) = s_{pq}(w).
\end{aligned}
\end{equation}

Proof A.60: [Bounds]

\begin{equation}
\begin{aligned}
s_{pq}(\tilde{w}) &= -1, \\
s_{pq}(\tilde{w}) &= -n^{\frac{1}{q}} - \frac{1}{q},
\end{aligned}
\end{equation}

i.e. $s_{pq}(\tilde{w}) < s_{pq}(w)$.

Proof A.61: [Quasi-convexity] Let $p \leq 1$, $q > 1$, e.g. Hoyer with $q = 2$, $s_{1,\infty}$ with $q = \infty$, then

- $f(.) = -\|\cdot\|_p$ non-positive convex, \\
- $g(.) = \|\cdot\|_q$ positive convex, \\
- by Theorem A.1(2), $z(.) = \frac{f(.)}{g(.)}$ is semi-strictly quasi-convex;

Proof A.62: [Monotonicity] Let $\beta = \tilde{w} - \hat{w}$,

\begin{equation}
\begin{aligned}
g(\beta) &= s_{pq}(\tilde{w}) - s_{pq}(\hat{w}) \\
&\approx \beta \nabla_{e_1-e_2}s_{pq}(\tilde{w}),
\end{aligned}
\end{equation}

where $g(0) = 0$, $\nabla_{e_1-e_2}f(v) = \frac{\partial f}{\partial w_1}(v) - \frac{\partial f}{\partial w_2}(v)$, and

\begin{equation}
\begin{aligned}
\frac{\partial s_{pq}}{\partial \hat{w}_k}(\tilde{w}) &= n^{\frac{1}{q}} - \frac{1}{q} \left( \frac{\tilde{w}^{p-1}_k}{\|\tilde{w}\|_q^q} - \frac{\hat{w}^{p-1}_k}{\|\tilde{w}\|_p^p} \right)
\end{aligned}
\end{equation}

positive, since $v^q_k - \|v\|_p^q > v^{p-1}_k - \|v\|_q^q$ (expand to see); i.e. $s_{pq}(\tilde{w}) > s_{pq}(\hat{w})$.

Proof A.63: [Completeness](cf. [30].)

\begin{equation}
\begin{aligned}
g(w) &= s_{pq}(\tilde{w}(w)) - s_{pq}(w) \\
&= \frac{\|w\|_p}{\|w\|_q} (n^{\frac{1}{q}} - \frac{1}{q} - (n + 1)^{\frac{1}{q}} - \frac{1}{q}) > 0,
\end{aligned}
\end{equation}
i.e. \( s_{pq}(\bar{w}) > s_{pq}(w) \).

**Proof A.64:** [Regularity] cf. approx. in [30] under minor corrections.

\[
g(\alpha) = s_{pq}(\bar{w}) - s_{pq}(w) \\
\approx \alpha \nabla_{\alpha} s_{pq}(\bar{w}),
\]

where \( g(0) = 0 \), \( \nabla_{\alpha} f(v) = \frac{\partial f}{\partial \alpha}(v) \), and

\[
\frac{\partial s_{pq}(\bar{w})}{\partial \bar{w}_i} = n^{\frac{1}{n} - \frac{1}{p}} \frac{1}{\bar{w}_i} \| \bar{w} \|_p \left( \bar{w}_i^{q-1} - \bar{w}_i^{p-1} \| \bar{w} \|_q \right) \| \bar{w} \|_q^{-1}
\]

positive, since \( v_i^{q-1} \| v \|_p > v_i^{p-1} \| v \|_q \) (expand to see); i.e. \( s_{pq}(\bar{w}) > s_{pq}(w) \).

**Proof A.65:** [Hom. growth] cf. approx. in [30] which divides by non-necessarily non-null coefficients.

\[
g(\alpha) = s_{pq}(w) - s_{pq}(\bar{w}(\alpha)),
\]

where \( g(0) = 0 \). Then,

\[
g'(\alpha) = n^{\frac{1}{n} - \frac{1}{p}} \left( \frac{\| w + \alpha 1_n \|_p}{\| w + \alpha 1_n \|_q} \right) \ahun{\| w \|_p}^{p-1} \| w \|_p^{-1} \| \bar{w} \|_q^{-1} \| \bar{w} \|_q > 0 \]

since \( \| v \|_p^{p-1} \| v \|_q > \| v \|_p^{q-1} \| v \|_q^{q-1} \) (expand to see), i.e. \( s_{pq}(w) > s_{pq}(\bar{w}) \).

### D. Proofs for Entropy Functions

Regularity [II-B5] and homogeneous growth [II-B6] use the histogram model: in each bin \( k \), new \( q_k \) count is added to old count \( p_k \), then sample pmf \( \tilde{p} \) is obtained after normalization \( \| \cdot \|_1 = 1 \) (using the number of realizations \( \tilde{n} \)).
Completeness II-B4 requires normalization by the number of events. The simple trial \( \frac{1}{\xi} \) (instead of \( b^{\frac{1}{\xi} - \frac{1}{\xi}} \)) will do it.

Proof A.66: [Continuity, symmetry] By the continuity and symmetry of \( \ell_p \)-norms, and the continuity of the logarithm and power functions.

1) Proofs for Shannon entropy:

Proof A.67: [Concentration](cf. [15].)

\[
g(\alpha) = h_{\text{Shannon}}^b(p) - h_{\text{Shannon}}^b(\tilde{p}(\alpha)) \tag{143}
\]

where \( g(0) = 0 \). Then,

\[
g'(\alpha) = (\log_2(p_j - \alpha)^{p_j - \alpha} + \log_2(p_i + \alpha)^{p_i + \alpha})'
\]

\[
= \log_2 \frac{p_i + \alpha}{p_j - \alpha} > 0, \tag{144}
\]

since \( p_i + \alpha > p_j - \alpha \), i.e. \( h_{\text{Shannon}}^b(w) > h_{\text{Shannon}}^b(\tilde{w}) \).

Proof A.68: [Scaling] Scaling holds by the \( \ell_1 \) normalization of \( p \), i.e. \( \|p\|_1 = 1 \).

Proof A.69: [Replication] For entropy functions, replication equals scaling since each replica of the histogram allocates proportional occurrences to events.

Proof A.70: [Bounds] Shannon entropy attains its maximum with the uniform distribution and its minimum with a constant process [24].

\[
h_{\text{Shannon}}^b(\hat{p}) = \log_2 b, \tag{146}
\]

\[
h_{\text{Shannon}}^b(\check{p}) = 0, \tag{147}
\]

i.e. \( h_{\text{Shannon}}^b(\hat{p}) > h_{\text{Shannon}}^b(\check{p}) \).

Proof A.71: [Quasi-convexity] Shannon entropy is concave [24].

Proof A.72: [Monotonicity] By the Shannon entropy of the Bernoulli distribution [24].

Proof A.73: [Completeness]

\[
g(w) = h_{\text{Shannon}}^b(p) - h_{\text{Shannon}}^b(\tilde{p}(p)) \tag{148}
\]

\[
= \frac{1}{b + 1} h_{\text{Shannon}}^b(p) \geq 0, \tag{149}
\]

since \( h_{\text{Shannon}}^b(p) \geq 0 \), i.e. \( h_{\text{Shannon}}^b(p) \geq h_{\text{Shannon}}^b(\check{p}) \).

Proof A.74: [Regularity]

\[
g(\alpha) = h_{\text{Shannon}}^b(\tilde{p}(\alpha)) - h_{\text{Shannon}}^b(\hat{p}), \tag{150}
\]
where \( g(0) = 0 \). Then,

\[
\begin{aligned}
g'(\alpha) &= \sum_{k=1, k \neq i}^b \left( \frac{q_k - n}{n + \alpha} \log_2 \frac{q_k}{n + \alpha} \right) \\
&\quad + \left( \frac{\hat{q}_i + \alpha}{n + \alpha} \log_2 \frac{\hat{q}_i + \alpha}{n + \alpha} \right) \\
&= \frac{1}{n} (h_{\text{Shannon}}^b(\hat{p}) - h_{\text{Shannon}}^b(\tilde{p})) > 0,
\end{aligned}
\]

since \( h_{\text{Shannon}}^b(\hat{p}) > h_{\text{Shannon}}^b(\tilde{p}) \) (cf. Figs. 3 and 4), i.e. \( h_{\text{Shannon}}^b(\hat{p}) > h_{\text{Shannon}}^b(\tilde{p}) \).

**Proof A.75:** [Hom. growth]

\[
g(\alpha) = h_{\text{Shannon}}^b(\tilde{p}(\alpha)) - h_{\text{Shannon}}^b(p),
\]

where \( g(0) = 0 \). Then,

\[
\begin{aligned}
g'(\alpha) &= -\sum_{i=1}^b \left( \frac{p_i + \alpha}{n + b\alpha} \log_2 \frac{p_i + \alpha}{n + b\alpha} \right) \\
&= \sum_{k=1}^b \frac{q_k - n}{n} \log_2 \frac{q_k}{n} > 0,
\end{aligned}
\]

i.e. \( h_{\text{Shannon}}^b(\tilde{w}) \geq h_{\text{Shannon}}^b(w) \).

2) **Proofs for Rényi entropy**:

**Proof A.76:** [Concentration]

\[
g(\alpha) = h_{\text{Rényi}}^b(p) - h_{\text{Rényi}}^b(\tilde{p}(\alpha)),
\]

where \( g(0) = 0 \). Let \( \xi = \frac{1}{(1-q) \log_2} \). Then,

\[
\begin{aligned}
g'(\alpha) &= \xi \left( \log \left( \sum_{k=1, k \neq i,j}^b p_k^q + (p_i + \alpha)^q + (p_j - \alpha)^q \right) \right) \\
&= \frac{\xi}{\|p\|_q} ((p_i + \alpha)^q - 1 - (p_j - \alpha)^q - 1) > 0,
\end{aligned}
\]

since \( p_i + \alpha > p_j - \alpha \), i.e. \( h_{\text{Rényi}}^b(p) > h_{\text{Rényi}}^b(\tilde{p}) \).

**Proof A.77:** [Bounds]

\[
\begin{aligned}
h_{\text{Rényi}}^b(\tilde{p}) &= \log_2 b, \\
h_{\text{Rényi}}^b(\tilde{p}) &= 0,
\end{aligned}
\]

i.e. \( h_{\text{Rényi}}^b(\tilde{p}) > h_{\text{Rényi}}^b(\hat{p}) \).

**Proof A.78:** [Quasi-concavity] By Theorem A.2 with \( \log_2 \) increasing,

\[
\begin{aligned}
h_{\text{Rényi}}^b(\tilde{p}) &= \frac{q}{1-q} \log_2 \circ \|\cdot\|_q)(p),
\end{aligned}
\]

January 23, 2015
Proof A.79: [Monotonicity]

\[ g(1) = h^b_{\text{Renyi}}(\tilde{\mathbf{p}}) - h^b_{\text{Renyi}}(\hat{\mathbf{p}}) \]

\[ = \frac{1}{(1 - q)} \log_2 \frac{\|\tilde{\mathbf{p}}\|_q}{\|\hat{\mathbf{p}}\|_q}, \]

since \( \|\tilde{\mathbf{p}}\|_q > (\|\hat{\mathbf{p}}\|_q) \) by the concavity (convexity) when \( q < (>) 1 \), i.e. \( h^b_{\text{Renyi}}(\tilde{\mathbf{p}}) < h^b_{\text{Renyi}}(\hat{\mathbf{p}}) \).

Proof A.80: [Completeness]

\[ g(p) = h^b_{\text{Renyi}}(\mathbf{p}) - h^b_{\text{Renyi}}(\tilde{\mathbf{p}}(p)) \]

\[ = \frac{1}{b} h^b_{\text{Renyi}}(\mathbf{p}) \geq 0, \]

since \( h^b_{\text{Renyi}}(\mathbf{p}) \geq 0 \), i.e. \( h^b_{\text{Renyi}}(\mathbf{p}) > h^b_{\text{Renyi}}(\tilde{\mathbf{p}}) \).

Proof A.81: [Regularity]

\[ g(\alpha) = h^b_{\text{Renyi}}(\tilde{\mathbf{p}}) - h^b_{\text{Renyi}}(\hat{\mathbf{p}}(\alpha)), \]

where \( g(0) = 0 \). Then,

\[ g'(\alpha) = \frac{1}{(1 - q)} \log_2 \frac{\sum_{k=1}^{b} \left( \frac{q_k}{n + \alpha} \right)^q + \left( \frac{q_k + \alpha}{n + \alpha} \right)^q}{\sum_{k=1}^{b} \left( \frac{p_k}{n + b\alpha} \right)^q} \]

\[ = \frac{q}{(1 - q)} \log_2 \frac{\sum_{k=1}^{b} (\tilde{p}_k^q - 1)}{\sum_{k=1}^{b} (\hat{p}_k^q - 1)} > 0, \]

since \( \|\mathbf{v}\|_q^q > (\|\mathbf{v}\|_\infty^q - 1) \) when \( q < (>) 1 \), i.e. \( h^b_{\text{Renyi}}(\tilde{\mathbf{p}}) > h^b_{\text{Renyi}}(\hat{\mathbf{p}}) \).

Proof A.82: [Hom. growth]

\[ g(\alpha) = h^b_{\text{Renyi}}(\tilde{\mathbf{p}}) - h^b_{\text{Renyi}}(\hat{\mathbf{p}}), \]

where \( g(0) = 0 \). Then,

\[ g'(\alpha) = \frac{1}{(1 - q)} \log_2 \frac{\sum_{k=1}^{b} \left( \frac{q_k + \alpha}{n + b\alpha} \right)^q}{\sum_{k=1}^{b} \left( \frac{p_k}{n + b\alpha} \right)^q} \]

\[ = \frac{q}{(1 - q)} \log_2 \frac{\sum_{k=1}^{b} (\hat{p}_k^q - 1)}{\sum_{k=1}^{b} (\tilde{p}_k^q - 1)} > 0, \]

since \( \hat{p}_k^q - 1 < (>) 1 \) with \( q > (>) 1 \), i.e. \( h^b_{\text{Renyi}}(\tilde{\mathbf{p}}) > h^b_{\text{Renyi}}(\hat{\mathbf{p}}) \).

3) Proofs for Tsallis entropy:

Proof A.83: [Concentration]

\[ g(\alpha) = h^b_{\text{Tsallis}}(\mathbf{p}) - h^b_{\text{Tsallis}}(\tilde{\mathbf{p}}(\alpha)), \]

where \( g(0) = 0 \). Then,

\[ g'(\alpha) = \frac{1}{(q - 1)} ((p_i + \alpha)^q + (p_j - \alpha)^q)' \]

\[ = \frac{q}{(q - 1)} ((p_i + \alpha)^q - (p_j - \alpha)^q) > 0, \]
since $p_i + \alpha > p_j - \alpha$, i.e. $h^b_{\text{Tsalis}}(\mathbf{p}) > h^b_{\text{Tsalis}}(\hat{\mathbf{p}})$.

**Proof A.84:** [Bounds]

\[
h^b_{\text{Tsalis}}(\mathbf{p}) = \frac{1 - b^{1-q}}{(q-1)},
\]

(176)

\[
h^b_{\text{Tsalis}}(\hat{\mathbf{p}}) = 0,
\]

(177)

i.e. $h^b_{\text{Tsalis}}(\mathbf{p}) > h^b_{\text{Tsalis}}(\hat{\mathbf{p}})$.

**Proof A.85:** [Quasi-concavity] By Theorem A.2 with $(\cdot)^q$ increasing,

\[
h^b_{\text{Tsalis}}(\hat{\mathbf{p}}) = \frac{q}{q-1} (1 - (\cdot)^q \circ \|\|_q)(\mathbf{p}).
\]

(178)

\[
g(1) = h^b_{\text{Tsalis}}(\mathbf{p}) - h^b_{\text{Tsalis}}(\hat{\mathbf{p}})
\]

(179)

\[
= \frac{\|\mathbf{p}\|_q}{(q-1)} \left( \frac{\|\mathbf{p}\|_q}{\|\hat{\mathbf{p}}\|_q} - 1 \right),
\]

(180)

since $\|\mathbf{p}\|_q > (\cdot)\|\hat{\mathbf{p}}\|_q$ by the concavity (convexity) when $q < (>) 1$, i.e. $h^b_{\text{Tsalis}}(\mathbf{p}) < h^b_{\text{Tsalis}}(\hat{\mathbf{p}})$.

**Proof A.86:** [Monotonicity]

\[
g(1) = h^b_{\text{Tsalis}}(\mathbf{p}) - h^b_{\text{Tsalis}}(\hat{\mathbf{p}})
\]

(181)

\[
g(\mathbf{p}) = h^b_{\text{Tsalis}}(\mathbf{p}) - h^b_{\text{Tsalis}}(\hat{\mathbf{p}}(\mathbf{p}))
\]

(182)

\[
= \frac{b+1}{b+1} h^b_{\text{Tsalis}}(\mathbf{p}) \geq 0,
\]

since $h^b_{\text{Tsalis}}(\mathbf{p}) \geq 0$, i.e. $h^b_{\text{Tsalis}}(\mathbf{p}) > h^b_{\text{Tsalis}}(\hat{\mathbf{p}})$.

**Proof A.87:** [Completeness]

\[
g(\alpha) = h^b_{\text{Tsalis}}(\hat{\mathbf{p}}) - h^b_{\text{Tsalis}}(\hat{\mathbf{p}}(\alpha)),
\]

(183)

where $g(0) = 0$. Then,

\[
g'(\alpha) = \frac{1}{(1-q)} \sum_{k \neq i}^b \left( \frac{q_k}{n + \alpha} \right)^q + \left( \frac{\tilde{q}_i + \alpha}{\tilde{n} + \alpha} \right)^q,
\]

(184)

\[
= \frac{q}{(1-q)n} (\|\hat{\mathbf{p}}\|_{\infty} - \|\hat{\mathbf{p}}\|_q^2) > 0,
\]

(185)

since $\|\mathbf{v}\|_{\infty}^2 > (\cdot)\|\mathbf{v}\|_q^2$ with $q < (>) 1$, i.e. $h^b_{\text{Tsalis}}(\mathbf{p}) > h^b_{\text{Tsalis}}(\hat{\mathbf{p}})$.

**Proof A.88:** [Regularity]

\[
g(\alpha) = h^b_{\text{Tsalis}}(\hat{\mathbf{p}}) - h^b_{\text{Tsalis}}(\hat{\mathbf{p}}(\alpha)),
\]

(186)

where $g(0) = 0$. Then,

\[
g'(\alpha) = \frac{-1}{(1-q)} \sum_{k=1}^b \left( \frac{q_k + \alpha}{n + b\alpha} \right)^q,
\]

(187)

\[
= \frac{-q}{(q-1)n} (\|\hat{\mathbf{p}}\|_{q-1} - b\|\hat{\mathbf{p}}\|_q^q) > 0,
\]

(188)

since $\|\mathbf{v}\|_{q-1}^2 > (\cdot)\|\mathbf{v}\|_0\|\mathbf{v}\|_q^q$ with $q \leq (>) 1$, i.e. $h^b_{\text{Tsalis}}(\mathbf{p}) > h^b_{\text{Tsalis}}(\hat{\mathbf{p}})$. 

January 23, 2015
REFERENCES

[1] K. Pearson, “Das Fehlergesetz und Seine Verallgemeinerungen Durch Fechner und Pearson. A Rejoinder,” Biometrika, vol. 4, no. 1-2, pp. 169–212, June 1905.

[2] H. Dalton, “The Measurement of the Inequity of Incomes,” The Economic Journal, vol. 30, no. 119, pp. 348–361, September 1920.

[3] R. V. L. Hartley, “Transmission of Information,” Bell System Technical Journal, vol. 7, pp. 535–563, 1928.

[4] C. Gini, “On the Measure of Concentration with Special Reference to Income and Statistics,” Colorado College Publication, General Series, no. 208, pp. 73–79, 1936.

[5] C. E. Shannon, “A Mathematical Theory of Communication,” The Bell System Technical Journal, vol. 27, no. 3, pp. 379–423, 623–656, July, October 1948.

[6] A. Rényi, “On measures of entropy and information,” in Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics. University of California Press, 1961, pp. 547–561.

[7] J. Aczél and Z. Daróczy, “Charakterisierung der Entropien Positiver Ordnung und der Shannonschen Entropie,” Acta Mathematica Academiae Scientiarum Hungarica, vol. 14, no. 1-2, pp. 95–121, 1963.

[8] H. Theil, Economics and Information Theory. North-Holland Publishing Company, 1967.

[9] A. B. Atkinson, “On the Measurement of Inequality,” Journal of Economic Theory, vol. 2, no. 3, p. 244263, September 1970.

[10] A. Lempel and J. Ziv, “On the Complexity of Finite Sequences,” IEEE Trans. Inf. Theory, vol. 22, no. 1, pp. 75–81, January 1976.

[11] C. Tsallis, “Possible Generalization of Boltzmann-Gibbs Statistics,” Journal of Statistical Physics, vol. 52, 1988.

[12] S. M. Pincus, “Approximate Entropy as a Measure of System Complexity,” Proceedings of the National Academy of Sciences of the United States of America (Mathematics), vol. 88, no. 6, pp. 2297–2301, March 1991.

[13] N. R. Pal and J. C. Bezdek, “Measuring Fuzzy Uncertainty,” IEEE Trans. Fuzzy Syst., vol. 2, no. 2, pp. 107–118, May 1994.

[14] D. L. Donoho, “For Most Large Underdetermined Systems of Linear Equations the Minimal $\ell_1$-norm Solution is also the Sparsest Solution,” Communications on Pure and Applied Mathematics, vol. 59, pp. 797–829, 2004.

[15] R. Chartrand, “Robust Uncertainty Principles: Exact Signal Reconstruction From Highly Incomplete Frequency Information,” IEEE Trans. Inf. Theory, vol. 52, no. 2, pp. 489–509, February 2006.

[16] C. Akujuobi, O. Odejide, A. Annamalai, and G. Fudge, “Sparseness Measures of Signals for Compressive Sampling,” in Signal Processing and Information Technology, 2007 IEEE International Symposium on, December 2007, pp. 1042–1047.

[17] I. Csiszár, “Axiomatic Characterizations of Information Measures,” Entropy, vol. 10, no. 3, pp. 261–273, 2008.
[29] A. Hormati and M. Vetterli, “Distributed Compressed Sensing: Sparsity Models and Reconstruction Algorithms using Annihilating Filter,” in Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on, March 2008, pp. 5141–5144.

[30] N. Hurley and S. Rickard, “Comparing Measures of Sparsity,” IEEE Trans. Inf. Theory, vol. 55, no. 10, pp. 4723–4741, October 2009.

[31] A. Cambini and L. Martein, Generalized Convexity and Optimization. Springer, 2009, vol. 616.

[32] J. Haberl, “On Global Minima of Semistrictly Quasiconcave Functions,” Optimization Letters, vol. 3, no. 3, pp. 387–396, 2009.

[33] Tian Lan, D. Kao, Mung Chiang, and A. Sabharwal, “An Axiomatic Theory of Fairness in Network Resource Allocation,” in INFOCOM, 2010 Proceedings IEEE, March 2010, pp. 1–9.

[34] A. Cohen and A. Yeredor, “On the use of sparsity for recovering discrete probability distributions from their moments,” in Statistical Signal Processing Workshop (SSP), 2011 IEEE, June 2011, pp. 753–756.

[35] D. Krishnan, T. Tay, and R. Fergus, “Blind Deconvolution using a Normalized Sparsity Measure,” in Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, June 2011, pp. 233–240.

[36] D. Zonoobi, A. A. Kassim, and Y. V. Venkatesh, “Gini Index as Sparsity Measure for Signal Reconstruction from Compressive Samples,” Selected Topics in Signal Processing, IEEE Journal of, vol. 5, no. 5, pp. 927–932, September 2011.

[37] M. Pilanci, L. El Ghaoui, and V. Chandrasekaran, “Recovery of Sparse Probability Measures via Convex Programming,” in Proceedings of Advances in Neural Information Processing Systems (NIPS), 2012, pp. 2429–2437.

[38] P. Cheridito and E. Kromer, “Reward-Risk Ratios,” Journal of Investment Strategies, Forthcoming, October 2013.

[39] T. Blumensath, “Compressed Sensing with Nonlinear Observations and Related Nonlinear Optimization Problems,” IEEE. Trans. Inf. Theory, vol. 59, no. 6, pp. 3466–3474, June 2013.

[40] M. Lopes, “Estimating Unknown Sparsity in Compressed Sensing,” in Proceedings of the 30th International Conference on Machine Learning (ICML) (3), vol. 28, no. 3, May 2013, pp. 217–225.

[41] G. Pastor, I. Mora-Jiménez, R. Jánitti, and A. J. Caamaño, “Sparsity-Based Criteria for Entropy Measures,” in Wireless Communication Systems (ISWCS 2013), Proceedings of the Tenth International Symposium on, August 2013, pp. 1–5.

[42] E. J. Candès, “Mathematics of Sparsity (and a Few Other Things),” in Proceedings of the International Congress of Mathematicians, 2014.