Sparse Signal Recovery via Generalized Entropy Functions Minimization

Shuai Huang, Student Member, IEEE, and Trac D. Tran, Fellow, IEEE

Abstract—Compressive sensing relies on the sparse prior imposed on the signal to solve the ill-posed recovery problem in an under-determined linear system. The objective function that enforces the sparse prior information should be both effective and easily optimizable. Motivated by the entropy concept from information theory, in this paper we propose the generalized Shannon entropy function and Rényi entropy function of the signal as the sparsity promoting objectives. Both entropy functions are nonconvex, and their local minimums only occur on the boundaries of the spheres or the orthogonal space. Compared to other popular objective functions such as the $\|x\|_1$, $\|x\|_0$, minimizing the proposed entropy functions not only promotes sparsity in the recovered signals, but also encourages the signal energy to be concentrated towards a few significant entries. The corresponding optimization problem can be converted into a series of reweighted $l_1$ minimization problems and solved efficiently. Sparse signal recovery experiments on both the simulated and real data show the proposed entropy function minimization approaches are better than other popular approaches and achieve state-of-the-art performances.

Index Terms—Compressive sensing, entropy function minimization, entropy minimization, image recovery

I. INTRODUCTION

Nowadays, there is an increasing amount of digital information constantly generated from every aspect of our life and data that we work with grow in both size and variety. Fortunately, most of the data have sparse structures. Compressive sensing [1]–[3] offers us an efficient framework to not only collect most of the data have sparse structures. Compressive sensing

Various sparsity regularizers have been proposed as the

where $x$ is the sparse signal with mostly zero

is an upper bound on the noise contribution. This

share this sparse property. Following the well known Occam’s razor, we can use the $l_0$ norm as the criterion and choose the sparsest (simplest) one:

$P_0(x): \min_{x} \|x\|_0 \text{ subject to } \|y - Ax\|_2^2 \leq \epsilon$. \hspace{1cm} (2)

This rather naïve attempt is actually backed up by sound theories [4]–[8]. Under noiseless conditions, it can be shown that the sparsest solution is indeed the true signal when $x$ is sufficiently sparse and $A$ satisfies the corresponding restricted isometry property [4], [7].

$P_0(x)$ is a nonconvex NP-hard problem whose solutions requires an intractable combinatorial search [9]. In practice, two alternative approaches are usually employed to solve $P_0(x)$:

1) Greedy search under the constrain $\|x\|_0 \leq S$.
2) Relaxation of the $l_0$ norm $\|x\|_0$.

$S > 0$ is an upper bound on the number of nonzero entries in $x$. The greedy search approach leads to various matching pursuit methods [10]–[13], while the relaxation approach leads to methods that minimize different objective functions to promote sparsity in the recovered solution [8], [14]–[23]. Here we focus on studying the “relaxation” approach that tries to solve the following unconstrained recovery problem:

$P_g(x): \min_{x} \|y - Ax\|_2^2 + \lambda g(x)$, \hspace{1cm} (3)

where $\lambda > 0$ is the parameter that balances the trade-off between the data fidelity term $\|y - Ax\|_2^2$ and the sparsity regularizer $g(x)$. The sparse prior information is enforced via the regularizer $g(x)$, and a proper $g(x)$ is crucial to the success of the sparse signal recovery task: it should favor sparse solutions and make sure the problem $P_g(x)$ can be solved efficiently in the mean time.

The proposed $g(x)$ can be any function that encourages sparsity, but it is important to note that $g(x)$ should be nonconvex to ensure the existence of a unique solution. In this paper we propose the generalized Shannon entropy function and Rényi entropy function of the signal $x$ as two nonconvex sparsity regularizers.

This rather naïve attempt is actually backed up by sound theories [4]–[8]. Under noiseless conditions, it can be shown that the sparsest solution is indeed the true signal when $x$ is sufficiently sparse and $A$ satisfies the corresponding restricted isometry property [4], [7].

$P_0(x)$ is a nonconvex NP-hard problem whose solutions requires an intractable combinatorial search [9]. In practice, two alternative approaches are usually employed to solve $P_0(x)$:

1) Greedy search under the constrain $\|x\|_0 \leq S$.
2) Relaxation of the $l_0$ norm $\|x\|_0$.

$S > 0$ is an upper bound on the number of nonzero entries in $x$. The greedy search approach leads to various matching pursuit methods [10]–[13], while the relaxation approach leads to methods that minimize different objective functions to promote sparsity in the recovered solution [8], [14]–[23]. Here we focus on studying the “relaxation” approach that tries to solve the following unconstrained recovery problem:

$P_g(x): \min_{x} \|y - Ax\|_2^2 + \lambda g(x)$, \hspace{1cm} (3)

where $\lambda > 0$ is the parameter that balances the trade-off between the data fidelity term $\|y - Ax\|_2^2$ and the sparsity regularizer $g(x)$. The sparse prior information is enforced via the regularizer $g(x)$, and a proper $g(x)$ is crucial to the success of the sparse signal recovery task: it should favor sparse solutions and make sure the problem $P_g(x)$ can be solved efficiently in the mean time.

In this paper we propose the generalized Shannon entropy function and Rényi entropy function of the signal $x$ as new sparsity-regularization objectives, and show their effectiveness and advantages over other popular regularizers in promoting sparse solutions with both theoretical analyses and experimental evaluations.

A. Prior Work

Various sparsity regularizers have been proposed as the relaxation of the $l_0$ norm. Most popular among them are the convex $l_1$ norm and the nonconvex $l_p$ norm to the $p$-th power [14], [18], [20], [21], [24], [25]:

- $l_1$ norm: $\|x\|_1 = \sum_i |x_i|$.
- $l_p$ norm to the $p$-th power: $\|x\|_p^p = \sum_i |x_i|^p$, $0 < p < 1$. 

This work is supported by the National Science Foundation under grants NSF-CCF-1117545, NSF-CCF-1422995 and NSF-EECS-1443936. The authors are with the Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD, 21218 USA (email: shuaihuang@jhu.edu; trac@jhu.edu).
Both \( \|x\|_1 \) and \( \|x\|_p^p \) are separable functions, and strict error bounds on the recovered solutions from \( l_1 \)-minimization and \( l_p^p \)-minimization problems in (3) can be established [20], [21]. Compared to \( l_1 \)-minimization, \( l_p^p \)-minimization has a tighter error bound and better sparse recovery performances.

[24] uses the following “logarithm of energy” function as a measure of sparsity:

\[
g_1(x) = \sum_i \log |x_i|^2 = 2 \sum_i \log |x_i|
\]

Minimizing it is equivalent to the \( l_p^p \)-minimization method when \( p \to 0 \) [25]. [8] later proposed the reweighted \( l_1 \)-minimization algorithm as a way to enhance the sparsity, which is essentially the iterative minimization of the first-order approximation of a modified \( g_1(x) \) function: \( \sum_i \log(|x_i| + \epsilon), \epsilon > 0 \).

Entropy-based functionals have also been widely used to promote sparsity [14], [15], [17], [18], [24]. The Shannon entropy functions considered all share the following form:

\[
g_2(x) = -\sum_i \hat{x}_i \log \hat{x}_i, \hat{x}_i > 0, \sum_i \hat{x}_i = 1.
\]

\( \hat{x}_i \) is constructed from the sparse signal \( x \). In [15], [17], [18], [24], the case where \( \hat{x}_i = \frac{|x_i|}{\sum_i |x_i|} \) is studied. [17] also considers the cases where \( \hat{x}_i = |x_i| \) and \( \hat{x}_i = \frac{|x_i|}{\sum_i |x_i|} \), and the corresponding Rényi entropy functions. Choosing \( \hat{x}_i = |x_i| \) fails to promote sparsity in the signal \( x \) whose \( l_1 \)-norm is not 1, i.e. \( \|x\|_1 \neq 1 \). There are also some imprecisions in [17]’s analysis on the local minimums of the \( l_2^2 \)-normalized entropy functions, [17] states that the local minimums “occur just shy of the boundaries defined by the coordinate axes”. In section II-C, we can prove that the local minimums actually occur exactly on the coordinate axes.

The previously mentioned entropy functions are all nonconvex. [23] later proposes the following convex entropy function as an approximation to the \( l_1 \)-norm \( \|x\|_1 \):

\[
g_3(x) = \sum_i \left( \left| |x_i| + \frac{1}{2} \right| \log \left( |x_i| + \frac{1}{2} \right) + \frac{1}{2} \right).
\]

The entropy function \( g_3(x) \) by [23] maintains the strictly convex property of \( |x|_1 \) and is continuously differentiable in \( \mathbb{R} \). However, \( g_3(x) \) only produces concentrated but not truly sparse solutions.

The convex \( l_1 \)-norm minimization problem can be efficiently solved by many available algorithms [26]–[32], such as interior-point method, FISTA, AMP, etc. However, it is often quite difficult to directly minimize the aforementioned nonconvex sparsity regularizers. In this case, we can iteratively minimize the approximation or upper bound of the regularizer using the reweighted \( l_1 \) or \( l_2 \) approach [8], [16], [25], [33].

### B. Main Contribution

Compared to previously adapted entropy functions [15], [17], [18], [24], our proposed Shannon entropy function \( h_p(x) \) and Rényi entropy function \( h_{p,\alpha}(x) \) are more generalized. They are defined with respect to the probability distribution in (6) where \( p \) can choose any positive number. This gives us more freedom in constructing the proper entropy function for the sparse signal recovery task. As is evident from the experiments on both simulated and real data, a good choice of the \( p \) value enables us to fully exploit the sparsity-promoting property of the entropy functions and to achieve better performances over the state-of-the-art \( \|x\|_1 \) and \( \|x\|_p^p \) minimization approaches.

Previous works [17], [18] focus on the study of the Schur-concavity of the entropy functions with respect to \( x \) where \( p = 1, 2 \), and believe that the Shannon entropy function produces truly sparse solutions only when \( p = 1 \). Here we can show that it’s the concavity or Schur-concavity with respect to the distribution \( P(\cdot) \) that really matters in the sparsity promotion analysis. In fact, \( \forall p > 0 \) and \( 1 \neq \alpha > 0 \), we can prove that the local minimums of the proposed entropy functions only occur at the boundaries of the orthants in \( \mathbb{R}^N \). Hence minimizing \( h_p(x) \) or \( h_{p,\alpha}(x) \) in said orthant \( \emptyset \) will lead us to the solutions on its boundaries, i.e. sparser solutions. The Shannon entropy function with \( p = 1 \) is not the only case where truly sparse solutions can be obtained.

Additionally, minimizing the proposed entropy functions promotes large-magnitude entries in the recovered signal and encourages the energy of \( \hat{x} \) to be concentrated towards a few significant entries. Using the proximal approximation [34], [35] of the data fidelity term and the first order approximations of the entropy functions, we can convert the nonconvex minimization problems into a series of classical reweighted \( l_1 \) problems, and solve them efficiently.

### II. Sparsity-Regularization Entropy Function

#### A. Introduction to Entropy

We first introduce the entropy concepts in information theory [36], [37]. Both the Shannon entropy and Rényi entropy are defined with respect to the probability distribution \( P(V) \) of some random variable \( V \). Here we give the following definitions in terms of discrete probability distribution:

- Shannon entropy:
  \[
  H(V) = -\sum_{i=1}^{\left| V \right|} P(V_i) \log P(V_i). \tag{4}
  \]

- Rényi entropy:
  \[
  H_\alpha(V) = \frac{1}{1-\alpha} \log \left( \sum_{i=1}^{\left| V \right|} P(V_i)^\alpha \right), \tag{5}
  \]

where \( \alpha \geq 0 \) and \( \alpha \neq 1 \). When \( \alpha \in (0, 1) \), \( H_\alpha(V) \) is strictly concave with respect to \( P(V) \) [38]; when \( \alpha \in (1, \infty) \), \( H_\alpha(V) \) is strictly Schur concave with respect to \( P(V) \) [39].

We should make it clear that Shannon entropy \( H(V) \) is not a special case of the Rényi entropy, but the limiting value of the Rényi entropy \( H_\alpha(V) \) as \( \alpha \to 1 \) [40]. Hence we need to discuss them respectively in this paper.

#### B. Entropy Function of the Sparse Signal

Entropy measures the uncertainty about the random variable \( V \) with \( |V| = N \). The lower the entropy is, the more predictable the variable \( V \) is, which corresponds to a skewed

1For continuous distributions, the sum in (4,5) should be replaced with integration \( f \).

2The “log” in this paper is by default natural logarithm, i.e. base \( e \).
distribution \( \mathcal{P}(\mathcal{V}) \). The idea of a skewed distribution could translate naturally to the idea of a sparse probability vector \( \mathcal{P}_x = [\mathcal{P}(v_1), \ldots, \mathcal{P}(v_N)]^T \) in the sense that only a few probability values of \( \mathcal{P}_x \in \mathbb{R}^N \) are significant. In other words, the entropy can be used as a measure of how sparse the probability vector \( \mathcal{P}_x \) is. This observation motivates us to adapt the concept of entropy as a sparsity-measure for the general signal \( x \) and to use it as a regularizer in the sparse signal recovery task.

As we have mentioned before, the entropy is defined with respect to a probability distribution \( \mathcal{P}(\cdot) \). Here we can construct the following discrete probability distribution out of the signal \( x \in \mathbb{R}^N \):

\[
x \rightarrow \left[ \frac{|x_1|^p}{\|x\|^p}, \frac{|x_2|^p}{\|x\|^p}, \ldots, \frac{|x_N|^p}{\|x\|^p} \right],
\]

where \( p > 0 \). The adaptation from the classical entropy to the entropy function is then pretty straightforward. Specifically, the following two types of entropy functions are proposed:

1) Shannon entropy function:

\[
h_p(x) = -\sum_{i=1}^{N} \frac{|x_i|^p}{\|x\|^p} \log \left( \frac{|x_i|^p}{\|x\|^p} \right),
\]

where \( p > 0 \), \( h_p(x) \) is the “Shannon entropy function” of \( x \), it should not be confused with the “Shannon entropy” of \( x \) in (4); \( \mathcal{H}(x) = -\int_x \mathcal{P}(x) \log \mathcal{P}(x) \, dx \).

2) Rényi entropy function:

\[
h_{p,\alpha}(x) = \frac{1}{1-\alpha} \log \left( \sum_{i=1}^{N} \left( \frac{|x_i|^p}{\|x\|^p} \right)^\alpha \right),
\]

where \( p > 0 \), \( \alpha > 0 \) and \( \alpha \neq 1 \). Again, this should not be confused with the Rényi entropy of \( x \) in (5).

Both \( h_p(x) \) and \( h_{p,\alpha}(x) \) are nonconvex functions, their local minimums occur at the boundary of each orthant in the Euclidean space \( \mathbb{R}^N \), i.e. the axes. Take the Shannon entropy function for example, the 2-dimensional level plots of \( h_p(x) \) with \( p = \{0.5, 1, 2\} \) are shown in Fig. 1. We can see that the local minimums occur at the two axes in all three cases.

In order to promote sparsity in the recovered solutions, we would like to minimize the entropy functions. The sparse signal recovery problems in (3) based on the Shannon entropy function (SEF) minimization and the Rényi entropy function (REF) minimization then become:

\[
P_{p}(x) : \min_{x} \|y - Ax\|^2 + \lambda h_p(x)
\]
\[
P_{p,\alpha}(x) : \min_{x} \|y - Ax\|^2 + \lambda h_{p,\alpha}(x).
\]

### C. Sparsity Promotion Analysis

We next show that \( h_p(x) \) and \( h_{p,\alpha}(x) \) can be used as sparsity regularizers in the following sense: minimizing them in an orthant \( \mathbb{O} \) of the Euclidean space \( \mathbb{R}^N \) leads us to solutions on the boundary of said orthant, i.e. sparser solutions.

#### Noiseless recovery: In this case we are minimizing \( h_p(x) \) or \( h_{p,\alpha}(x) \) subject to the constrain \( y = Ax \). We first show that there is a one to one mapping in each orthant between \( x = [x_1, \ldots, x_N]^T \) and \( \tilde{x} = [\tilde{x}_1, \ldots, \tilde{x}_N]^T \), where \( \tilde{x}_i = \text{sign}(x_i) \cdot \|x_i\|_p \). This will be done in two steps: Lemma 1 and Lemma 2.

**Lemma 1.** If \( x \) is the solution to \( y = Ax \), \( y \neq 0 \), then there is a one to one mapping in each orthant between \( x \) and \( \tilde{x} = \frac{x}{\|x\|_p} \).

**Proof.** We just need to prove \( x \leftrightarrow \tilde{x} \):

- It is easy to verify that \( x \rightarrow \tilde{x} \).
- Suppose there are two solutions of \( y = Ax: x_1(1), x_2(2) \) in the same orthant, and they are both mapped to \( \tilde{x} \). We then have:

\[
\frac{x_1(1)}{\|x_1(1)\|_p} = \frac{x_2(2)}{\|x_2(2)\|_p} = \frac{\tilde{x}_1}{\|\tilde{x}_1\|_p} = \frac{\tilde{x}_2}{\|\tilde{x}_2\|_p}.
\]


\[
\frac{y}{\|x_1(1)\|_p} = \frac{Ax_1(1)}{\|x_1(1)\|_p} = \frac{Ax_2(2)}{\|x_2(2)\|_p} = \frac{\tilde{x}_1}{\|\tilde{x}_1\|_p} = \frac{y}{\|x_2(2)\|_p},
\]

\( \square \)

![Fig. 1: Shannon entropy function \( h_p(x) \) in the 2-dimensional space: (a) \( p = 0.5 \); (b) \( p = 1 \); (c) \( p = 2 \).](image-url)
which tells us \( \frac{y}{\|x(1)\|_p} = \frac{y}{\|x(2)\|_p} \). Since \( y \neq 0 \), we have \( \|x(1)\|_p = \|x(2)\|_p \). Using (11), we get \( x(1) = x(2) \). Hence \( x \leftarrow \hat{x} \).

**Lemma 2.** There is a one to one mapping in each orthant between \( \hat{x} \) and \( \bar{x} \)

**Proof.** We just need to prove \( \hat{x} \leftarrow \bar{x} \):

- We can rewrite \( \bar{x} \) in terms of \( \hat{x} \): \( \bar{x} = \text{sign}(\hat{x}) \cdot |\hat{x}|^p \). Hence \( \hat{x} \rightarrow \bar{x} \).
- Suppose there are two points \( \hat{x}(1), \hat{x}(2) \) in the same orthant mapped to the same \( \bar{x} \). We then have:
  \[
  \text{sign}(\hat{x}(1)) \cdot |\hat{x}(1)|^p = \bar{x} = \text{sign}(\hat{x}(2)) \cdot |\hat{x}(2)|^p, \tag{13}
  \]
  which tells us \( |\hat{x}(1)| = |\hat{x}(2)| \). Since \( \text{sign}(\hat{x}(1)) = \text{sign}(\hat{x}(2)) \), we get \( \hat{x}(1) = \hat{x}(2) \). Hence \( \hat{x} \leftarrow \bar{x} \).

Combining Lemma 1 and Lemma 2, we have \( x \leftarrow \hat{x} \), as is shown in Fig. 2. Let \( \mathcal{X} = \{x_1, x_2, \cdots\} \) be the solutions of \( y = Ax \) in one of the orthants \( \mathcal{O} \). Specifically, \( \mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2 \) and \( \mathcal{X}_1 \cap \mathcal{X}_2 = \emptyset \), where \( \mathcal{X}_1 \) contains solutions on the boundary of the orthant \( \mathcal{O} \) and \( \mathcal{X}_2 \) contains the rest solutions that are not on the boundary. The solution \( x \) is then mapped to \( \hat{x} \) one by one, producing the corresponding mapped sets \( \mathcal{X}_1, \mathcal{X}_2 \). We can verify that the solutions in \( \mathcal{X}_1 \) are sparser than those in \( \mathcal{X}_2 \), and we have the following Lemma 3:

**Lemma 3.** For every solution \( x \in \mathcal{X}_2 \), there is a solution \( x^* \in \mathcal{X}_1 \) on the boundary of the orthant \( \mathcal{O} \) such that \( h_p(x^*) < h_p(x) \) and \( h_{p,\alpha}(x^*) < h_{p,\alpha}(x) \).

**Proof.** By definition we have:

\[
\begin{align*}
  h_p(x) &= g(\hat{x}) = -\sum_{i=1}^N |\hat{x}_i| \log |\hat{x}_i|, \tag{14} \\
  h_{p,\alpha}(x) &= g_\alpha(\bar{x}) = \frac{1}{1-\alpha} \log \left( \sum_{i=1}^N |\bar{x}_i|^\alpha \right). \tag{15}
\end{align*}
\]

For the SEF \( h_p(x) \), we first study the local minimums on the plane \( \|\hat{x}\|_1 = 1 \). \( g(\hat{x}) \) is strictly concave with respect to \( \hat{x} \), and the local minimums of \( g(\hat{x}) \) are on the boundary of the orthant \( \mathcal{O} \). Hence for every \( \hat{x} \in \mathcal{X}_2 \), there is a \( \hat{x}^* \in \mathcal{X}_1 \) such that \( g(\hat{x}^*) < g(\hat{x}) \).

For the REF \( h_{p,\alpha}(x) \), when \( \alpha \in (0, 1) \), \( g_\alpha(\bar{x}) \) is strictly concave with respect to \( \bar{x} \), the local minimums of \( g_\alpha(\bar{x}) \) are on the boundary of the orthant \( \mathcal{O} \). When \( \alpha \in (1, \infty) \), \( g_\alpha(\bar{x}) \) is strictly Schur concave [39], since the boundary of the orthant \( \mathcal{O} \) majorizes the \( \bar{x} \) inside \( \mathcal{O} \), the local minimums of \( g_\alpha(\bar{x}) \) are also on the boundary of \( \mathcal{O} \). Hence for every \( \bar{x} \in \mathcal{X}_2 \), there also exists a \( \bar{x}^* \in \mathcal{X}_1 \) such that \( g_\alpha(\bar{x}^*) < g_\alpha(\bar{x}) \) for \( \alpha \in (0, 1) \cup (1, \infty) \).

There is a one to one mapping in \( \mathcal{O} \) between \( x \) and \( \hat{x} \): \( x \leftarrow \hat{x} \). Since \( h_p(x) = g(\hat{x}) \) and \( h_{p,\alpha}(x) = g_\alpha(\bar{x}) \), for every \( \bar{x} \in \mathcal{X}_2 \), there is a \( x^* \in \mathcal{X}_1 \) such that \( h_p(x^*) < h_p(x) \) and \( h_{p,\alpha}(x^*) < h_{p,\alpha}(x) \).

From Lemma 3 we can see that minimizing \( h_p(x) \) or \( h_{p,\alpha}(x) \) in the orthant \( \mathcal{O} \) will lead us to the sparser solutions in \( \mathcal{X}_1 \).

**Noisy recovery:** We can show similarly that minimizing \( h_p(x) \) or \( h_{p,\alpha}(x) \) subject to the constrain \( \|y - Ax\|_2^2 \leq \epsilon \) in an orthant \( \mathcal{O} \) of the Euclidean space \( \mathbb{R}^N \) also produces sparse solutions. First, we have the following Lemma 4:

**Lemma 4.** Let \( \mathcal{X}^c = \{x_1, x_2, \cdots\} \) are the nonzero solutions satisfying the constrain \( \|y - Ax\|_2^2 \leq \epsilon, y \neq 0 \) such that: \( \forall x_i \neq x_j, x_i = \tau x_j \) for some \( \tau > 0 \). Pick any \( x_i \in \mathcal{X}^c \), there is a one to one mapping in each orthant between the set \( \mathcal{X}^c \) and \( \mathcal{X}_i = \frac{x_i}{\|x_i\|_p} \).

**Proof.** We need to prove \( \mathcal{X}_i \leftarrow \hat{x}_i \):

- \( \forall x_j \in \mathcal{X}^c \setminus x_i \), we have \( \hat{x}_i = \frac{x_i}{\|x_i\|_p} = \frac{x_j}{\|x_j\|_p} = \frac{x_j}{\|x_j\|_p} = \frac{x_j}{\|x_j\|_p} = \hat{x}_j \). It is easy to verify that \( x_j \rightarrow \hat{x}_j \). Hence \( \mathcal{X}^c \rightarrow \hat{x}_i \).
the noisy case. We can see that minimizing
solutions that are not on the boundary. Lemma 3 also applies in
\(O\)
Lemma 5.

\[
\text{D. Energy Concentration Analysis}
\]

Apart from promoting sparsity in the recovered signal, minimizing the entropy functions \(h_p(x)\) and \(h_{p,\alpha}(x)\) also encourages the energy of \(x\) to be concentrated towards a few significant entries.

This energy concentration behavior can be best illustrated from an optimization point of view. Take the SEF \(h_p(x)\) for example, we use gradient descent to minimize it and the solution \(x\) is updated as follows:

\[
x = x - \eta \cdot \text{sign}(x) \cdot \nabla_{|x|} h_p(x),
\]

\(\tilde{x}_o\) and \(\rho(\tilde{x}_o)\) are further projected to \(\tilde{x}_o\) and \(\rho(\tilde{x}_o)\) respectively on the sphere \(|\tilde{x}_o| = 1\)
inside the same orthant \(\bigcirc\).

\[
\tilde{x}_o \leftrightarrow \tilde{x}_o \quad \text{and} \quad \rho(\tilde{x}_o) \leftrightarrow \rho(\tilde{x}_o).
\]

Consequently, \(\forall \tilde{x} \in \rho(\tilde{x}_o), h_p(\tilde{x}) \geq h_p(\tilde{x}_o)\) or \(h_{p,\alpha}(\tilde{x}) \geq h_{p,\alpha}(\tilde{x}_o). \tilde{x}_o\) is also a corresponding local minimum \(\tilde{x}_o\) on the sphere \(|\tilde{x}_o| = 1\)
inside the orthant \(\bigcirc\).

However, in the following we can show that such a \(\tilde{x}_o\) does not exist on the sphere \(|\tilde{x}_o| = 1\)
inside the orthant \(\bigcirc\):

1) SEF \(h_p(x)\) with \(p > 0\): It is strictly concave with respect to \(x\), there are no local minimums on the sphere \(|\tilde{x}_o| = 1\)
inside the orthant \(\bigcirc\), i.e. \(\tilde{x}_o\) does not exist.

2) REF \(h_{p,\alpha}(x)\) with \(p > 0, \alpha > 0\) and \(\alpha \neq 1\): When \(\alpha \in (0,1), h_{p,\alpha}(\tilde{x})\) is strictly concave with respect to \(\tilde{x}\). When \(\alpha \in (1,\infty), h_{p,\alpha}(\tilde{x})\) is strictly Schur concave with respect to \(\tilde{x}\). \(\tilde{x}_o\) also does not exist in this case.

We can thus see that there are no local minimums inside each orthant. Furthermore, the local minimums of \(h_p(x)\) and \(h_{p,\alpha}(x)\) occur at the boundaries of the sphere \(|\tilde{x}_o| = 1\).
Hence the local minimums of \(h_p(x)\) and \(h_{p,\alpha}(x)\) only occur at the boundaries of each orthant.

\[
\phi \text{ Discussion: In this section we have showed that minimizing the entropy functions (7,8) leads to sparser solutions. So what happens if we minimize the “true” entropy of } x? \text{ Take the Shannon entropy in (4) for example, it is defined with respect to some probability distribution } P(\cdot). \text{ Here we can assume the signal } x \text{ follows the pre-specified distribution } P(x|\theta) \text{ parameterized by } \theta. \text{ For the sparse signal recovery task, we rely on the posterior distribution } P(x|y, \theta) \text{ to perform the MAP estimation or MMSE estimation of the signal } \tilde{x}. \]

In practice the entries \(x\) are further assumed to be independently distributed given \(y\), and the following Shannon entropy will be minimized:

\[
\tilde{\theta} = \arg \min_{\theta} H(x|y, \theta)
\]

\[
= \arg \min_{\theta} - \int P(x|y, \theta) \log P(x|y, \theta) \, dx
\]

\[
= \arg \min_{\theta} - \sum_i \int P(x_i|y, \theta) \log P(x_i|y, \theta) \, dx_i.
\]

\(P(x_i|y, \theta)\) can be computed using the approximate message passing algorithms (AMP) [31], [41]–[43]. In this case (19) essentially serves as a parameter estimation step in the PE-GAMP algorithm [43].
Remark. Let \( \nabla |x|h_p(x) \) be the derivative with respect to the magnitude \( |x| \):

\[
\frac{\partial h_p(x)}{\partial |x|} = -p|x|^{|p-1|} \log |x|^p + p|x|^{(p-1)} \sum_l |x|^p \log |x|^p.
\]

Following similar derivation process, we can see that \( h_p(x) \) and \( h_{p,\alpha}(x) \) are non-convex, a good initialization is needed to ensure good performance. Here we will use the solution from \( l_1 \) norm-minimization as the initialization to our proposed algorithm. The sparsity-promotion analysis in section II-C shows that we are able to obtain sparse solutions by minimizing the entropy functions. In order to solve the problems in (9,10), the following two steps are repeated in alternation until convergence.

1) In the first step, the data fidelity term \( f(x) = \|y - Ax\|_2^2 \) is approximated: For the \( (t+1) \)-th iteration to solve the problems \( P_{h_p}(x) \) and \( P_{h_{p,\alpha}}(x) \), we use its quadratic approximation, a.k.a. proximal regularization [34], at the previous \( t \)-th iteration’s solution \( \hat{x}^{(t)} \):

\[
f(x) = \|y - Ax\|_2^2 \\
\leq f(\hat{x}^{(t)}) + \left< x - \hat{x}^{(t)}, \nabla f(\hat{x}^{(t)}) \right> + \frac{\kappa}{2} \| x - \hat{x}^{(t)} \|_2^2
\]

\[
= o(\hat{x}^{(t)}) + \frac{\kappa}{2} \| x - \left( \hat{x}^{(t)} - \frac{1}{\kappa} \nabla f(\hat{x}^{(t)}) \right) \|_2^2,
\]

where \( o(\hat{x}^{(t)}) \) is a relative constant depending on the previous solution \( \hat{x}^{(t)} \), \( \nabla f(\hat{x}^{(t)}) = 2(A^T A \hat{x}^{(t)} - A^T y) \), \( \kappa \) is the Lipschitz constant of the gradient \( \nabla f \) [44]. The smallest value \( \kappa \) can take is twice the largest eigenvalue of \( A^T A \) to ensure that \( f(x) \) is bounded by the proximal regularization. It can be viewed as a suitable step size.

III. Entropy Function Minimization

In this section we propose the algorithms to perform the sparse signal recovery tasks in (9,10). Specifically, the proximal regularization [34], [35] of the data fidelity term \( f(x) = \|y - Ax\|_2^2 \) and the first order approximations of the entropy functions \( h_p(x), h_{p,\alpha}(x) \) are minimized in alternation iteratively until convergence.

For the entries with relatively small magnitudes \( |x_i| > \nu \), the update makes their magnitudes even smaller. In this way minimizing \( h_p(x) \) promotes large-magnitude entries in the recovered signal, hence energy concentration.

For the REF \( h_{p,\alpha}(x) \), the derivative with respect to the magnitude \( |x| \) is:

\[
\frac{\partial h_{p,\alpha}(x)}{\partial |x|} = \frac{1}{1 - \alpha} \times \sum_{i=1}^{N} \left( \frac{|x_i|}{|x|} \right)^{\alpha} \times p \alpha \left[ |x|^p - |x_i - \hat{x}^{(t)}|^{p+\alpha} \right].
\]

Similarly we have the following remark:

Remark. Let \( \nu = \exp \left( \frac{1}{p\alpha - p} \log \frac{\|x\|_{\alpha,p}}{\|x\|_p} \right), \nu > 0 \), we then have:

\[
\frac{\partial h_{p,\alpha}(x)}{\partial |x|} \begin{cases} < 0 & \text{if } |x_i| > \nu \\ = 0 & \text{if } |x_i| = \nu \\ > 0 & \text{if } |x_i| < \nu. 
\end{cases}
\]

In Fig. 5, the 2-dimensional level plots of \( \|x\|_1 \) and \( \|x\|_{0.5} \) are shown. Following similar derivation process, we can see that the popular \( \|x\|_1 \) and \( \|x\|_p \) with \( 0 < p < 1 \) don’t have the energy concentration properties.
to ensure the upper bound on $f(x)$ in (25). When $\kappa$ is unknown or difficult to compute, we can use a backtracking strategy to find it.

The problems in (9, 10) then becomes:

$$
P_{h_p}^{(1)}(x) : \min_{x} \frac{\kappa}{2} \left\| x - \left( \hat{x}^{(t)} - \frac{1}{\kappa} \nabla f(\hat{x}^{(t)}) \right) \right\|^2_2 + \lambda h_p(x) \tag{26}
$$

$$
P_{h_{p,\alpha}}^{(1)}(x) : \min_{x} \frac{\kappa}{2} \left\| x - \left( \hat{x}^{(t)} - \frac{1}{\kappa} \nabla f(\hat{x}^{(t)}) \right) \right\|^2_2 + \lambda h_{p,\alpha}(x) \tag{27}
$$

2) In the second step, the problems $P_{h_p}^{(1)}(x)$ and $P_{h_{p,\alpha}}^{(1)}(x)$ are iteratively solved: In the “inner” $(r+1)$-th iteration to solve $P_{h_p}^{(1)}(x)$ and $P_{h_{p,\alpha}}^{(1)}(x)$ are approximated with their first order approximations with respect to $|\hat{x}^{(t+1,r)}|$ from the previous $r$-th iteration:

$$
h_p(x) \approx \left\{ |x| - |\hat{x}^{(t+1,r)}|, \nabla h_p(\hat{x}^{(t+1,r)}) \right\} + h_p(\hat{x}^{(t+1,r)}) \tag{28}
$$

$$
h_{p,\alpha}(x) \approx \left\{ |x| - |\hat{x}^{(t+1,r)}|, \nabla h_{p,\alpha}(\hat{x}^{(t+1,r)}) \right\} + h_{p,\alpha}(\hat{x}^{(t+1,r)}) \tag{29}
$$

where $\nabla h_p(x), \nabla h_{p,\alpha}(x)$ are the first order derivatives with respect to $|x|$ given in (21,23). Since $\log 0 = -\infty$, when computing $\nabla h_p(\hat{x}^{(t+1,r)})$, we add a small positive value $\epsilon = 1 - 12$ to $|\hat{x}^{(t+1,r)}|$ in case $|\hat{x}^{(t+1,r)}| = 0$. Ignoring the relative constant terms in (28,29) that depend on $\hat{x}(r)$, the problems $P_{h_p}^{(1)}(x)$ and $P_{h_{p,\alpha}}^{(1)}(x)$ then become:

$$
P_{h_p}^{(2)}(x) : \min_{x} \frac{\kappa}{2} \left\| x - \left( \hat{x}^{(t)} - \frac{1}{\kappa} \nabla f(\hat{x}^{(t)}) \right) \right\|^2_2 + \lambda \left\{ |x|, \nabla h_p(\hat{x}^{(t+1,r)}) \right\} \tag{30}
$$

$$
P_{h_{p,\alpha}}^{(2)}(x) : \min_{x} \frac{\kappa}{2} \left\| x - \left( \hat{x}^{(t)} - \frac{1}{\kappa} \nabla f(\hat{x}^{(t)}) \right) \right\|^2_2 + \lambda \left\{ |x|, \nabla h_{p,\alpha}(\hat{x}^{(t+1,r)}) \right\} \tag{31}
$$

$P_{h_p}^{(2)}(x)$ and $P_{h_{p,\alpha}}^{(2)}(x)$ are simple reweighted $l_1$ minimization problems that can be converted to a series of independent one-dimensional problems. The solutions $\hat{x}_i^{(t+1,r+1)}$ to the above problems can be obtained using the iterative shrinkage thresholding algorithm (ISTA):

$$
\hat{x}_i^{(t+1,r+1)} = \begin{cases} 
\hat{x}_i^{(t+1,r)} & \text{if } |\hat{x}_i^{(t+1,r)}| < \tau \\
\Gamma_{\kappa} \nabla h_p(|\hat{x}_i^{(t+1,r)}|) \left( \hat{x}_i^{(t)} - \frac{1}{\kappa} \nabla f(\hat{x}_i^{(t)}) \right) & \text{if } |\hat{x}_i^{(t+1,r)}| \geq \tau \end{cases} \tag{32}
$$

$$
\hat{x}_i^{(t+1,r+1)} = \begin{cases} 
\hat{x}_i^{(t+1,r)} & \text{if } |\hat{x}_i^{(t+1,r)}| < \tau \\
\Gamma_{\kappa} \nabla h_{p,\alpha}(|\hat{x}_i^{(t+1,r)}|) \left( \hat{x}_i^{(t)} - \frac{1}{\kappa} \nabla f(\hat{x}_i^{(t)}) \right) & \text{if } |\hat{x}_i^{(t+1,r)}| \geq \tau \end{cases} \tag{33}
$$

where $\Gamma_{\kappa} (\cdot)$ is the soft thresholding function, a.k.a. shrinkage operator, defined as follows:

$$
\Gamma_{\kappa} (x) = \begin{cases} 
0 & \text{if } |x| \leq \tau \\
(x - \tau) \cdot \text{sign}(x) & \text{if } |x| > \tau \end{cases} \tag{34}
$$

Conventional ISTA solves a convex problem and requires the threshold $\tau$ to be positive. However, the derivatives $\nabla h_p(\hat{x}^{(t+1,r)})$ and $\nabla h_{p,\alpha}(\hat{x}^{(t+1,r)})$ in (30,31) could be negative. In Appendix A we can show that the optimal solution can still be obtained using the soft thresholding operator given in (34), yet with a different derivation process. The proposed entropy function minimization approach can be summarized in Algorithm 1.

Algorithm 1 Sparse signal recovery via entropy function minimization

Require: $\{y, A\}, \lambda, \kappa, \{p, \alpha\}$

1: Initialize $\hat{x}^{(0)}$ with the solution from $l_1$ norm minimization;
2: for $t = \{0, 1, \ldots \}$ do
3: Compute $\hat{x}^{(t)} = \frac{1}{\kappa} \nabla f(\hat{x}^{(t)})$ in (25);
4: Initialize $\{\hat{x}^{(t+1,r)}, r = 0\}$ with $\hat{x}^{(t)}$;
5: for $r = \{0, 1, \ldots \}$ do
6: Compute $\nabla h_p(\hat{x}) \text{ or } \nabla h_{p,\alpha}(\hat{x})$ in (21,23);
7: Obtain $\hat{x}^{(t+1,r+1)}$ by solving $P_{h_p}^{(2)}(x)$ or $P_{h_{p,\alpha}}^{(2)}(x)$ in (32,33);
8: if $\hat{x}^{(t+1,r+1)}$ reaches convergence or the objective functions in (26,27) increase then
9: $\hat{x}^{(t+1)} = \hat{x}^{(t+1,r+1)}$;
10: break;
11: end if
12: end for
13: if $\hat{x}^{(t+1)}$ reaches convergence then
14: $\hat{x} = \hat{x}^{(t+1)}$;
15: break;
16: end if
17: end for
18: Return Output $\hat{x}$;

As is shown in Appendix B, Algorithm 1 produces a sequence $\{\hat{x}^{(t)}, t = 0, 1, \ldots\}$ that decreases objective functions in (9,10) monotonically. Since the data fidelity term $f(x) = \|y - Ax\|^2_2 \geq 0$ and the entropy functions $h_p(x) \geq 0, h_{p,\alpha}(x) \geq 0$ are all bounded from below, eventually Algorithm 1 is going to converge to some local minimums of the nonconvex objective functions. A proper initialization is thus need for best performance, here we use the solution from $l_1$ norm minimization to initialize $\hat{x}^{(0)}$.

ISTA usually converges slowly in practice. [29] proposes a fast iterative shrinkage thresholding algorithm (FISTA) to address this issue. Although FISTA is proposed for convex regularizers, we find that it could also speed up the convergence of the nonconvex regularizers such as $\|x\|_p^p$, the entropy functions, etc. by a great deal.

Naturally, choosing a proper $\lambda$ is the key to the success of sparse signal recovery. For noiseless signals, the solution can be obtained as $\lambda \to 0$ [32]. Here we use solve a series of minimization problems characterized by a decreasing
\( \lambda \) sequence. Take the Shannon entropy function \( h_p(x) \) for example, we have the following:

1: Start with a relatively large \( \lambda_0 \).

2: for \( k = \{0, 1, \cdots\} \) do

3: Solve the minimization problem characterized by \( \lambda_k \):

\[
P_{h_p}(x) : \min_{x} \| y - Ax \|_2^2 + \lambda_k h_p(x). \tag{35}
\]

4: Update \( \lambda_{k+1} = \rho \cdot \lambda_k \).

5: if convergence is reached then break.

6: end for

To ensure the best performance, \( \rho \) is chosen to be \( 0.9 \leq \rho < 1 \).

For noisy signals, the optimal \( \lambda \) depends on the noise level and is usually unknown. A fixed \( \lambda \) is tuned on some development set and used in practice.

### IV. Experimental Results

We compare the proposed Shannon entropy function (SEF) minimization and Rényi entropy function (REF) minimization approaches with the state-of-the-art \( l_1 \) norm (L1) minimization and \( l_p \) norm to the \( p \)-th power (Lp) minimization approaches on simulated and real datasets.

#### A. Simulated sparse signal recovery

For the noiseless sparse signal recovery experiments, we fix \( N = 1000 \) and vary the sampling ratio \( \sigma = \frac{M}{N} \in \{0.05, 0.1, 0.15, \cdots, 0.95\} \) and the sparsity ratio \( \rho = \frac{S}{N} \in \{0.05, 0.1, 0.15, \cdots, 0.95\} \), where \( S \) is the sparsity of the signal, i.e. the number of nonzero coefficients. For each combination of \( \sigma \) and \( \rho \), we randomly generate 100 pairs of \( \{x, A\} \); \( A \) is a \( M \times N \) random Gaussian matrix with normalized and centralized rows; the nonzero entries of the sparse signal \( x \in \mathbb{R}^N \) are i.i.d. generated according to the Gaussian distribution \( N(0, 1) \).

Given the measurement vector \( y = Ax \) and the sensing matrix \( A \), we try to recover the signal \( x \). If \( \epsilon = \|x - x^\ast\|_2/\|x^\ast\|_2 < 10^{-3} \), the recovery is considered to be a success. The parameters are selected to obtain best performance for each method: for the SEF minimization approach, \( p = 1.1 \); for the REF minimization approach, \( p = 1.1, \alpha = 1.1 \); for the Lp minimization approach, \( p = 0.5 \). FPC method [45] is used to approach the optimal \( \lambda = 0 \). Based on the 100 trials, we compute the success recovery rate for each combination of \( \sigma \) and \( \rho \) and plot the PTCs in Fig. 7.

The PTC is the contour that corresponds to the 0.5 success rate in the domain \( \{\sigma, \rho\} \in (0, 1)^2 \), it divides the domain into a “success” phase (lower right) and a “failure” phase (upper left). We can see that the proposed SEF minimization and REF minimization approaches generally perform equally well, and they both perform better than the L1 and Lp minimization approaches.

We next try to recover the sparse signal \( x \) from a noisy measurement vector \( y \). Specifically, we fix \( S = 100, N = 100 \) and...
1000 and increase the number of measurement $M$. $y \in \mathbb{R}^M$ is generated as follows:

Noisy measurements: $y = Ax + \nu w$,  

(36)

where $\nu > 0$ controls the amount of noise added to $y$, the entries of $w$ are i.i.d Gaussian $\mathcal{N}(0, 1)$. We choose $\nu = 0.05$, this creates a measurement $y$ with signal to noises ratio (SNR) around $25$ dB. We randomly generate 100 triples of $\{x, A, w\}$. The average SNRs of the recovered signals $\hat{x}$ are shown in Fig. 8. We can see that the proposed SEF/REF minimization approaches and the Lp minimization approach perform better than the L1 minimization approach. When $\sigma < 0.5$, the SEF and REF minimization approaches outperform the Lp minimization approach.

![Fig. 8: The signal-to-noise-ratio (SNR) of the recovered signal $\hat{x}$ using different sparsity regularization approaches in the noisy case.](image)

The entries of the noise $w$ are generated using i.i.d Gaussian distribution $\mathcal{N}(0, 1)$, $\nu$ is chosen to be $0.02$ so that the SNR of the measurement vector $y$ is around $30$ dB$^3$.

Take the SEF minimization for example, we have the following recovery problem:

$$
\min_{s} \|y - Us\|^2_{2} + \lambda h_p(V^Ts) .
$$

(41)

Since the recovery problem is with respect to $x$, we need to modify Algorithm 1: we also use the proximal regularization of the data fidelity term $\|y - Us\|^2_{2}$, the optimization problem in the $(t + 1)$-th iteration then becomes:

$$
\min_{s} \frac{\kappa}{2} \left\| s - \left( \frac{s^{(t)}}{\kappa} + 2U^T(Us^{(t)} - y) \right) \right\|^2_{2} + \lambda h_p(V^Ts) ,
$$

(42)

where $\kappa = 2$ for the chosen $U$ in (38). In the $(r + 1)$-th iteration to minimize (42), let $Q^{r+1,r}$ be a diagonal matrix whose diagonal entries are the partial derivative of $h_p(V^Ts)$.

3When $\nu$ is set to other values, the relative performances of the four methods are similar.

B. Real image recovery

Real images are considered to be approximately sparse under some proper basis, such as the DCT basis, wavelet basis, etc. Here we compare the recovery performances of the aforementioned sparsity regularization approaches based on varying noiseless and noisy measurements of the 4 real images in Fig. 9: Barbara, Boat, Lena, Peppers. Specifically, in order to reveal the sparse coefficients $x$ of the real images $s$, we use the sparsity averaging method by [46] to construct an over-complete wavelet basis by concatenating Db1-Db4 [47] as follows:

$$
V = \frac{1}{2} \times [V_{Db1} \ V_{Db2} \ V_{Db3} \ V_{Db4}] .
$$

(37)

It is easy to verify that $s = Vx$, and $x = V^Ts$. The sampling matrix $U$ is constructed using the structurally random matrix approach by [48]:

$$
U = DFR ,
$$

(38)

where $R$ is a uniform random permutation matrix that scrambles the signal’s sample locations globally while a diagonal matrix of Bernoulli random variables flips the signal’s sample locations locally. $F$ is an orthonormal DCT matrix that computes fast transforms, $D$ is a sub-sampling matrix that randomly selects a subset of the rows of the matrix $FR$.

The noiseless measurements $y$ of the image $s$ are obtained as follows:

Noisy measurements: $y = DFRV x = UVx = Us$.  

(39)

The entries of the noise $w$ are obtained using different sparsity regularization approaches in the noisy case.

Fig. 9: The real images used in the recovery experiments: (a) Barbara; (b) Boat; (c) Lena; (d) Peppers.
with respect to $|V^Ts|$ at the solution $\hat{s}^{(t+1,r)}$, the optimization problem is as follows:

$$\min_s \left( s - (\hat{s}^{(t)} - U^T(U\hat{s}^{(t)} - y)) \right)^2 + \lambda Q^{(t+1,r)} |V^Ts|.$$

(43) can be efficiently solved using the alternating split bregman shrinkage algorithm by [49].

Since the real images are only approximately sparse, both the noiseless and noisy recovery experiments are done using a fixed $\lambda$. The parameters are tuned to obtain best performance for each approach. For the L1 minimization approach, $\lambda = 0.1$; for the SEF minimization approach, $p = 1$, $\lambda = 5000$; for the REF minimization approach, $p = 0.9$, $\alpha = 1.1$, $\lambda = 10000$; for the Lp minimization approach, $p = 0.8$, $\lambda = 0.01$. The peak signal to noise ratios (PSNR) of the noiseless and noisy recovery experiments are shown in Fig. 10 and 11 respectively. We can see that the proposed SEF and REF entropy function minimization approaches perform equally well, and they give the best performances in terms of PSNR (dB).

V. CONCLUSION AND FUTURE WORK

In this paper we proposed the generalized Shannon entropy function $h_p(x)$ and Rényi entropy function $h_{p,\alpha}(x)$ as the sparsity regularizers for the sparse signal recovery task. Regardless of the values of $p, \alpha$, the local minimums of the entropy functions occur on the boundaries of the orthants in $\mathbb{R}^N$ and minimizing them promotes sparsity in the recovered signals. Both $h_p(x)$ and $h_{p,\alpha}(x)$ are nonconvex function, the corresponding minimization problems (9,10) can be solved by iteratively minimizing the their first order approximations until convergence. Compared to the popular $l_1$ norm $\|x\|_1$ and the $l_p$ norm to the $p$-th power $\|x\|_p^p$, minimizing the entropy functions encourages the energy of the recovered signal to be concentrated towards a few significant entries.

Sparse signal recovery experiments on both the simulated and real data show the proposed entropy function regularizations perform better than the popular $\|x\|_1$ and $\|x\|_p^p$ regularizations. This motivates us to explore theoretical guarantees of the advantage over other approaches in the future by establishing error bounds on the recovered signal $\hat{x}$ and providing sufficient conditions under which the successful recovery is warranted. Additionally, we would also like to apply the entropy function minimizations to other Compressive Sensing applications such as the SRC [50], RPCA [52–55], dictionary learning [56–58], etc.

APPENDIX A

GENERALIZED SOFT SHRINKAGE THRESHOLDING

Take $P_{h_p}^{(2)}(x)$ in (30) for example, let $r^{(t+1,r)} = \frac{1}{\lambda} \nabla h_p(|\hat{x}_i^{(t+1,r)}|)$, $\hat{x}_i^{(t)} = x_i^{(t)} - \frac{1}{\lambda} \nabla f(x_i^{(t)})$, we have the
following problem for \( x_i \):
\[
\min_{x_i} \frac{1}{2} \left\| x_i - \tilde{x}_i^{(t)} \right\|^2 + \tau_i^{(t+1,r)} |x_i| . \tag{44}
\]

When \( \tau_i^{(t+1,r)} \geq 0 \), (44) is a convex problem. Its solution is given by applying the shrinkage operator given in (34) on \( \tilde{x}_i^{(t)} \) with the threshold \( \tau_i^{(t+1,r)} \).

When \( \tau_i^{(t+1,r)} < 0 \), (44) is a not necessarily a convex problem. Luckily this is a simple one dimensional problem, its global optimal solution can be still found as follows:

1) When \( \tilde{x}_i^{(t)} < \tau_i^{(t+1,r)} \), as is shown in Fig. 12(a):
   For \( x_i \geq 0 \), we have:
   \[
   \min_{x_i} \left( x_i + \tau_i^{(t+1,r)} - \tilde{x}_i^{(t)} \right)^2 - \left( \tau_i^{(t+1,r)} - \tilde{x}_i^{(t)} \right)^2 + (\tilde{x}_i^{(t)})^2 . \tag{45}
   \]
   Since \( \tau_i^{(t+1,r)} - \tilde{x}_i^{(t)} > 0 \), the \( x_i \) that minimizes (45) is 0.

   For \( x_i < 0 \), we have:
   \[
   \min_{x_i} \left( x_i - \tau_i^{(t+1,r)} + \tilde{x}_i^{(t)} \right)^2 - \left( \tau_i^{(t+1,r)} + \tilde{x}_i^{(t)} \right)^2 + (\tilde{x}_i^{(t)})^2 . \tag{46}
   \]
   Since \( -\tau_i^{(t+1,r)} - \tilde{x}_i^{(t)} > 0 \), the \( x_i \) that minimizes (46) is \( -\tau_i^{(t+1,r)} + \tilde{x}_i^{(t)} \).

   (44) is continuous at the point \( x_i = 0 \). Hence the global minimum of (44) is obtained by \( x_i = \tau_i^{(t+1,r)} + \tilde{x}_i^{(t)} \).

2) When \( \tau_i^{(t+1,r)} \leq \tilde{x}_i^{(t)} < 0 \), as is shown in Fig. 12(b):
   For \( x_i \geq 0 \), we have (45). Since \( \tau_i^{(t+1,r)} - \tilde{x}_i^{(t)} < 0 \), the \( x_i \) that minimizes (45) is \( -\tau_i^{(t+1,r)} + \tilde{x}_i^{(t)} \).

   For \( x_i < 0 \), we have (46). Since \( -\tau_i^{(t+1,r)} - \tilde{x}_i^{(t)} > 0 \), the \( x_i \) that minimizes (46) is \( \tau_i^{(t+1,r)} + \tilde{x}_i^{(t)} \).

   It’s easy to verify that the minimum of (46) is smaller than the minimum of (45). Hence the global minimum of (44) is obtained by \( x_i = -\tau_i^{(t+1,r)} + \tilde{x}_i^{(t)} \).

3) When \( \tilde{x}_i^{(t)} \leq \tau_i^{(t+1,r)} \), as is shown in Fig. 12(c):
   For \( x_i \geq 0 \), we have (45). Since \( \tau_i^{(t+1,r)} - \tilde{x}_i^{(t)} < 0 \), the \( x_i \) that minimizes (45) is \( -\tau_i^{(t+1,r)} + \tilde{x}_i^{(t)} \).

   For \( x_i < 0 \), we have (46). Since \( -\tau_i^{(t+1,r)} - \tilde{x}_i^{(t)} > 0 \), the \( x_i \) that minimizes (46) is \( \tau_i^{(t+1,r)} + \tilde{x}_i^{(t)} \).

   Combining the above 4 different scenarios, we have the following results:
   1) When \( \tilde{x}_i^{(t)} \geq 0 \), the solution to (44) is \( \tilde{x}_i^{(t)} = \tau_i^{(t+1,r)} \).

   2) When \( \tilde{x}_i^{(t)} < 0 \), the solution to (44) is \( \tilde{x}_i^{(t)} = \tau_i^{(t+1,r)} \).

   This is exactly the shrinkage operator given in (34) on \( \tilde{x}_i^{(t)} \) with the threshold \( \tau_i^{(t+1,r)} \).

\section*{Appendix B

\textbf{MONOTONIC OBJECTIVE FUNCTION MINIMIZATION}}

Take Shannon entropy minimization (SEF) problem \( P_{h_p}(\mathbf{x}) \) in (9) for example, we can have the following:
\[
\begin{align*}
& f(\hat{\mathbf{x}}^{(t+1)}) + h_p(\hat{\mathbf{x}}^{(t+1)}) \\
\overset{(a)}{\leq} & f(\hat{\mathbf{x}}^{(t)}) + \left\langle \hat{\mathbf{x}}^{(t+1)} - \hat{\mathbf{x}}^{(t)}, \nabla f(\hat{\mathbf{x}}^{(t)}) \right\rangle \\
& + \frac{\kappa}{2} \left\| \hat{\mathbf{x}}^{(t+1)} - \hat{\mathbf{x}}^{(t)} \right\|_2^2 + h_p(\hat{\mathbf{x}}^{(t+1)}) \\
\overset{(b)}{\leq} & f(\hat{\mathbf{x}}^{(t)}) + \left\langle \hat{\mathbf{x}}^{(t)} - \hat{\mathbf{x}}^{(t)}, \nabla f(\hat{\mathbf{x}}^{(t)}) \right\rangle \\
& + \frac{\kappa}{2} \left\| \hat{\mathbf{x}}^{(t)} - \hat{\mathbf{x}}^{(t)} \right\|_2^2 + h_p(\hat{\mathbf{x}}^{(t)}) \\
= & f(\hat{\mathbf{x}}^{(t)}) + h_p(\hat{\mathbf{x}}^{(t)}).
\end{align*}
\]

The first inequality “(a)” in (47) is obtained using (25). The second inequality “(b)” in (47) holds since \( \hat{\mathbf{x}}^{(t+1)} \) is a solution of \( P_{h_p}^{(t)}(\mathbf{x}) \) in (26). In this case, the objective function \( f(\mathbf{x}) + h_p(\mathbf{x}) \) monotonically decreases on the sequence \( \{\hat{\mathbf{x}}^{(t)}, t = 0, 1, \cdots\} \). Similar remark can also be made for
the Rényi entropy function minimization problem $P_{h,p,\alpha}(x)$ in (10).

REFERENCES

[1] D. L. Donoho, “Compressed sensing.” *IEEE Trans. on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[2] E. J. Candès, “Compressive sampling.” *IEEE Proceedings of the International Congress of Mathematicians*, 2006.

[3] E. J. Candès and M. B. Wakin, “An introduction to compressive sampling.” *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 21–30, March 2008.

[4] E. Candès and T. Tao, “Decoding by linear programming.” *IEEE Trans. on Information Theory*, vol. 51(12), pp. 4203–4215, 2005.

[5] E. J. Candès, J. Romberg, and T. Tao, “Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information.” *IEEE Trans. on Information Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[6] E. J. Candès and T. Tao, “Near-optimal signal recovery from random projections: Universal encoding strategies?” *IEEE Trans. on Information Theory*, vol. 52, no. 12, pp. 5406–5425, 2006.

[7] E. J. Candès, “The restricted isometry property and its implications for compressed sensing.” *Comptes Rendus Mathématique*, vol. 346, no. 910, pp. 589–592, 2008.

[8] E. J. Candès, M. B. Wakin, and S. Boyd, “Enhancing sparsity by reweighted $\ell_1$-minimization.” *Journal of Fourier Analysis and Applications*, vol. 14, no. 5, pp. 877–905, 2008.

[9] B. K. Natarajan, “Sparse approximate solutions to linear systems.” *SIAM Journal on Computing*, vol. 24, no. 2, pp. 227–234, 1995.

[10] S. G. Mallat and Z. Zhang, “Matching pursuits with time-frequency dictionaries.” *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, 1993.

[11] J. A. Tropp and A. C. Gilbert, “Signal recovery from random measurements via orthogonal matching pursuit.” *IEEE Transactions on Information Theory*, vol. 53, no. 12, pp. 4655–4666, 2007.

[12] W. Dai and O. Milenkovic, “Subspace pursuit for compressive sensing signal reconstruction.” *IEEE Transactions on Information Theory*, vol. 55, no. 5, pp. 2233–2249, 2009.

[13] D. Needell and J. A. Tropp, “Cosamp: iterative signal recovery from incomplete and inaccurate samples.” *Applied and Computational Harmonic Analysis*, vol. 26, no. 3, pp. 301 – 321, 2009.

[14] D. L. Donoho, “On minimum entropy segmentation,” in *Wavelets: theory, algorithms, and applications*. 1994, pp. 233–269, Academic Press.

[15] R. R. Coifman and M. V. Wickerhauser, “Entropy-based algorithms for finding sparse solutions,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 317–329, April 2010.

[16] M. A. T. Figueiredo, R. D. Nowak, and S. J. Wright, “Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems.” *IEEE Journal of Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 586–597, Dec 2007.

[17] Y. Nesterov, “A method of solving a convex programming problem with convergence rate $O(1/k^2)$.” *Soviet Mathematics Doklady*, vol. 27, no. 2, pp. 372–376, 1983.

[18] S. Becker, J. Bobin, and E. J. Candès, “Nesta: A fast and accurate first-order method for sparse recovery.” *SIAM Journal on Imaging Sciences*, vol. 4, no. 1, pp. 1–39, 2011.

[19] A. Beck and M. Teboulle, “A fast iterative shrinkage-thresholding algorithm for linear inverse problems.” *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[20] J. Yang and Y. Zhang, “Alternating direction algorithms for $\ell_1$-problems in compressive sensing.” *SIAM Journal on Scientific Computing*, vol. 33, no. 1, pp. 250–278, 2011.

[21] D. L. Donoho, A. Maleki, and A. Montanari, “Message-passing algorithms for compressed sensing,” *Proceedings of the National Academy of Sciences*, vol. 106, no. 45, pp. 18914–18919, 2009.

[22] A. Y. Yang, Z. Zhou, A. G. Balasubramanian, S. S. Sastry, and Y. Ma, “Fast $\ell_1$-minimization algorithms for robust face recognition.” *IEEE Transactions on Image Processing*, vol. 22, no. 8, pp. 3234–3246, May 2013.

[23] R. Chartrand and W. Yin, “Iteratively reweighted algorithms for compressive sensing,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, March 2008, pp. 3869–3872.

[24] B. Martinet, “Brève communication. régularisation d’inéquations variationnelles par approximations successives,” *ESAIM: Mathematical Modelling and Numerical Analysis - Modélisation Mathématique et Analyse Numérique*, vol. 4, no. R3, pp. 154–158, 1970.

[25] N. Parikh and S. Boyd, “Proximal algorithms,” in *Foundations and Trends in Optimization*, vol. 1, no. 3, pp. 123–231, 2013.

[26] T. M. Cover and J. A. Thomas, *Elements of information theory*. Wiley-Interscience, 2006.

[27] A. Rényi, “On measures of information and entropy,” in *Proceedings of the 4th Berkeley symposium on mathematics, statistics and probability*, 1960, vol. 1, pp. 547–561.

[28] M. Ben-Bassat and J. Raviv, “Rényi’s entropy and the probability of error.” *IEEE Transactions on Information Theory*, vol. 24, no. 3, pp. 524–531, May 1978.

[29] G. M. Bosyk, S. Zozor, F. Holik, M. Portesi, and P. W. Lamberti, “A family of generalized quantum entropies: definition and properties,” *Quantum Information Processing*, vol. 15, no. 8, pp. 3393–3420, 2016.

[30] P. A. Bromiley, N. A. Thacker, and E. Bouhova-Thacker, “Shannon entropy, rényi entropy, and information.” Tech. Rep., Tina Mem No. 2004-004, 2004.

[31] M. J. Wainwright and M. I. Jordan, “Graphical models, exponential families, and variational inference.” *Foundations and Trends in Machine Learning*, vol. 1, no. 1-2, pp. 1–305, 2008.

[32] S. Rangan, “Generalized approximate message passing for estimation with random linear mixing,” in *IEEE International Symposium on Information Theory (ISIT)*, July 2011, pp. 2168–2172.

[33] S. Huang and T. D. Tran, “Sparse signal recovery using generalized approximate message passing with built-in parameter estimation,” *CoRR*, vol. abs/1606.00901, 2016, http://arxiv.org/abs/1606.00901/.

[34] K. Eriksson, D. Estep, and C. Johnson, “Applied mathematics body and soul: Vol i: derivatives and geometry in ir3,” Springer-Verlag Publishing, pp. 149–164, 2004.

[35] D. Wipf and S. Nagarajan, “Iterative reweighted $\ell_1$ and $\ell_2$ methods for finding sparse solutions.” *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 547–561, May 1978.

[36] R. Chartrand, “Exact reconstructions of sparse signals via nonconvex minimization.” *IEEE Signal Processing Letters*, vol. 14, no. 7, pp. 707–710, 2007.

[37] S. Foucart and M.-J. Lai, “Sparsest solutions of underdetermined linear systems via $\ell_2$-minimization for $0 < p ≤ 1$.” *Applied and Computational Harmonic Analysis*, vol. 26, no. 3, pp. 395 – 407, 2009.

[38] N. Hurley and S. Rickard, “Comparing measures of sparsity.” *IEEE Transactions on Information Theory*, vol. 55, no. 10, pp. 4723–4741, Oct 2009.

[39] K. Kose, O. Gunay, and A. E. Cetin, “Compressive sensing using the modified entropy functional.” *Digital Signal Processing*, vol. 24, pp. 63–70, 2014.

[40] M. V. Wickerhauser, *Adapted Wavelet Analysis from Theory to Software*, A. K. Peters, Ltd., Natick, MA, USA, 1994.
[50] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, “Robust face recognition via sparse representation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, Feb 2009.

[51] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, and S. Yan, “Sparse representation for computer vision and pattern recognition,” *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1031–1044, June 2010.

[52] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma, “Robust principal component analysis: exact recovery of corrupted low-rank matrices via convex optimization,” in *Advances in Neural Information Processing Systems*, pp. 2080–2088. Curran Associates, Inc., 2009.

[53] Z. Lin, M. Chen, and Y. Ma, “The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices,” *UIUC Technical Report*, , no. UILU-ENG-09-2215, 2009.

[54] Z. Lin, A. Ganesh, J. Wright, L. Wu, M. Chen, and Y. Ma, “Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix,” *UIUC Technical Report*, , no. UILU-ENG-09-2214, 2009.

[55] E. J. Candès, X. Li, Y. Ma, and J. Wright, “Robust principal component analysis?,” *Journal of the ACM*, vol. 58, no. 3, pp. 11:1–11:37, May 2011.

[56] M. Aharon, M. Elad, and A. Bruckstein, “k -svd: An algorithm for designing overcomplete dictionaries for sparse representation,” *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, Nov 2006.

[57] M. Elad and M. Aharon, “Image denoising via sparse and redundant representations over learned dictionaries,” *IEEE Transactions on Image Processing*, vol. 15, no. 12, pp. 3736–3745, Dec 2006.

[58] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, “Online dictionary learning for sparse coding,” in *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009, pp. 689–696.