SUPPLEMENTAL MATERIAL: GLOBAL OPTIMALITY FOR NON-LINEAR CONSTRAINED RESTORATION PROBLEMS VIA INVEXITY

Samuel Pinilla, & Jeyan Thiyagalingam
[*]Scientific Computing Department, Science and Technology Facilities Council
samuel.pinilla@stfc.ac.uk

A PROOF OF THEOREM 3

In order to establish a versatile framework for invex functions, we provide proof of invexity for a specific class of summation functions. The key result is encapsulated in the following lemma, which reveals a fundamental connection: the invexity of the summation function, obtained by applying a unidimensional real function to distinct entries of a vector, hinges upon the invexity of this underlying function.

Lemma 1 (Invexity of Function Sum). Let $g : \mathbb{R}^n \rightarrow \mathbb{R}$ be a function defined as

$$g(\boldsymbol{x}) = \sum_{i=1}^{n} r(\boldsymbol{x}[i]), \tag{1}$$

where $r : \mathbb{R} \rightarrow \mathbb{R}$. If $r(w)$ is an invex function then $g(\boldsymbol{x})$ is also invex.

Proof. Let $g : \mathbb{R}^n \rightarrow \mathbb{R}$ be a function defined as $g(\boldsymbol{x}) = \sum_{i=1}^{n} r(\boldsymbol{x}[i])$. Assume $r : \mathbb{R} \rightarrow \mathbb{R}$ is invex. Then, from the invexity of $r(w)$ we have that there exists $\eta_r : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ such that

$$r(w_1) - r(w_2) \geq \zeta_{w_2} \cdot \eta_r(w_1, w_2), \tag{2}$$

for all $w_1, w_2 \in \mathbb{R}$, and any $\zeta_{w_2} \in \partial r(w_2)$. Take $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. Then, we have

$$r(\boldsymbol{x}[i]) - r(\boldsymbol{y}[i]) \geq \zeta_{\boldsymbol{y}[i]} \cdot \eta_r(\boldsymbol{x}[i], \boldsymbol{y}[i]), \tag{3}$$

for any $i = 1, \ldots, n$, and $\forall \zeta_{\boldsymbol{y}[i]} \in \partial r(\boldsymbol{y}[i])$. From the above inequality we conclude that for any $\boldsymbol{\zeta} \in \partial g(\boldsymbol{y})$

$$\begin{aligned}
\sum_{i=1}^{n} r(\boldsymbol{x}[i]) - r(\boldsymbol{y}[i]) &\geq \sum_{i=1}^{n} \zeta_{\boldsymbol{y}[i]} \cdot \eta_r(\boldsymbol{x}[i], \boldsymbol{y}[i]) \\
g(\boldsymbol{x}) - g(\boldsymbol{y}) &\geq \sum_{i=1}^{n} \zeta_{\boldsymbol{y}[i]} \cdot \eta_r(\boldsymbol{x}[i], \boldsymbol{y}[i]) \\
g(\boldsymbol{x}) - g(\boldsymbol{y}) &\geq \boldsymbol{\zeta}^T \eta(\boldsymbol{x}, \boldsymbol{y}),
\end{aligned} \tag{4}$$

such that $\eta(\boldsymbol{x}, \boldsymbol{y}) = [\eta_r(\boldsymbol{x}[1], \boldsymbol{y}[1]), \ldots, \eta_r(\boldsymbol{x}[n], \boldsymbol{y}[n])]^T$, and $\boldsymbol{\zeta} = [\zeta_{\boldsymbol{y}[1]}, \ldots, \zeta_{\boldsymbol{y}[n]}]^T$. Thus, from equation (4) the results holds. □

Now we prove Theorem 3 using the result in Lemma 1

Proof. We proceed by cases.

- Let f be an admissible function. Then according to Definition 4 we know that $f(\boldsymbol{x}) = \sum_{i=1}^{n} s(|\boldsymbol{x}[i]|)$, for some $s : [0, \infty) \rightarrow [0, \infty)$ with $s'(w) > 0$ where $t \in (0, \infty)$. Since the structure of f fits assumption in Lemma 1 then if $s(w)$ is invex then $f(\boldsymbol{x})$ is invex. Take $w_1, w_2 \in (0, \infty)$, and define $\eta : (0, \infty)^2 \rightarrow \mathbb{R}$ as

$$\eta(w_1, w_2) = \left\{ \begin{array}{ll} 0 & \text{if } s(w_1) > s(w_2) \\ \frac{s(w_1) - s(w_2)}{(\zeta_*)^2} \zeta_* & \text{otherwise ,} \end{array} \right. \tag{5}$$

[*] Rutherford Appleton Laboratory, Harwell, UK.

where ζ_* is an element in $\partial s(w_2)$ of minimal absolute value which satisfies $\frac{\zeta_* \cdot \zeta}{(\zeta_*)^2} \geq 1$ for all $\zeta \in \partial s(w_2)$. The existence of ζ_* is guaranteed because $s'(w) > 0$ and therefore $0 \notin \partial s(w_2)$ (Mishra & Giorgi, 2008, Page 64). From the above equation it is clear that for all w_1, w_2 we have

$$s(w_1) - s(w_2) \geq \eta(w_1, w_2) \cdot \zeta_{w_2}, \ \ \forall \zeta_{w_2} \in \partial s(w_2) \tag{6}$$

which means s is invex. Therefore $f(\boldsymbol{x})$ is invex.

- Let f, g be an admissible functions. Considering the result in previous statement, it is enough to show that $h = \beta f + \alpha g$ is an admissible function for any $\beta, \alpha \geq 0$. By definition we have that $h(0) = 0$, and $h'(w) > 0$. In addition, since both f, g are positive functions it implies that $h(w)/w^2$ is non-increasing. Thus the result holds.
- Let $f, g : \mathbb{R}^n \to \mathbb{R}$ be two admissible functions as in Definition 4, such that $f(\boldsymbol{x}) = \sum_{i=1}^n s_f(|\boldsymbol{x}[i]|)$, and $g(\boldsymbol{x}) = \sum_{i=1}^n s_g(|\boldsymbol{x}[i]|)$. Define $h_c(\boldsymbol{x}) = \sum_{i=1}^n (s_f \circ s_g)(|\boldsymbol{x}[i]|)$. Then, observe that $(s_f \circ s_g)(0) = s_f(s_g(0)) = s_f(0) = 0$. Additionally, since $s'_f(w), s'_g(w) > 0$, then $(s_f \circ s_g)'(w) > 0$ (by the chain rule) for all $t \in (0, \infty)$. Finally, we know that $s'_f(w), s'_g(w) > 0$ implies $s_f(w), s_g(w) > 0$ to be strictly increasing. Therefore, for any $w_1 < w_2$ we know $(s_f \circ s_g)(w_1) < (s_f \circ s_g)(w_2)$, which implies $\frac{(s_f \circ s_g)(w_1)}{(w_1)^2} > \frac{(s_f \circ s_g)(w_2)}{(w_2)^2}$. Thus the result holds.

Thus we proved the first part of this theorem. □

Now we proceed to prove the second part of this theorem.

Proof. Let $f : \mathbb{R}^n \to \mathbb{R}$ be two admissible functions as in Definition 4, such that $f(\boldsymbol{x}) = \sum_{i=1}^n s_f(|\boldsymbol{x}[i]|)$, and $s_f(w) \geq 0$. Define the set $\mathcal{D} = \{t \in (0, \infty) | s'(w) > 0\}$. In consequence, from first statement of this theorem proved above we know that $s_f(w)$ is invex in $\mathcal{D}$ for some function η_D. Then, define for any $w_1, w_2 \in (0, \infty)$ the following function

$$\eta(w_1, w_2) = \begin{cases} \eta_D(w_1, w_2) & \text{if } w_1, w_2 \in \mathcal{D} \\ 0 & \text{otherwise .} \end{cases} \tag{7}$$

Considering the above $\eta(w_1, w_2)$ function (which is non-zero), take $w_1, w_2 \in (0, \infty)$, and assume

$$s_f(w_1) - s_f(w_2) \leq 0. \tag{8}$$

If $w_1, w_2 \in \mathcal{D}$, then we know $s_f(w)$ is invex, which implies that

$$\begin{aligned} s_f(w_1) - s_f(w_2) \geq \eta(w_1, w_2) \cdot \zeta_{w_2} &= \eta_D(w_1, w_2) \cdot \zeta_{w_2} \\ 0 \geq \eta_D(w_1, w_2) \cdot \zeta_{w_2}&, \end{aligned} \tag{9}$$

for all $\zeta_{w_2} \in \partial s_f(w_2)$, and therefore quasi-invex. Otherwise, if either $w_1 \notin \mathcal{D}$ or $w_2 \notin \mathcal{D}$, it is clear to conclude that $\eta(w_1, w_2) \cdot \zeta_{w_2} = 0$, for all $\zeta_{w_2} \in \partial s_f(w_2)$. Thus, the result holds.

Finally, to establish the validity of the remaining two statements when f is quasi-invex, we follow a similar procedure as described above. In the interest of conciseness, we omit the detailed exposition. Thus the results holds. □

B Proof of Theorem 4

We prove Theorem 4 proceeding by cases, exploiting the result in Lemma 1, because equation (3) is the sum of unidimensional real functions applied to different entries of a vector.

Equation (3)

Proof. From the definition of function $g(\boldsymbol{x})$ in equation (3) it is clear that the only aspect we have to prove is that $\log(1 + w^p)/w^2$ is non-increasing for any $w > 0$, and fixed $p \in (0, 1)$. Take $r(w) = \log(1 + w^p)$. Observe that the first derivative of $h(w) = r(w)/w^2$ is given by $h'(w) = \frac{1}{w^3}\left(\frac{pw^p}{1+w^p} - 2\log(1 + w^p)\right)$. Since $\frac{pw^p}{1+w^p} - 2\log(1 + w^p) < 0$, then we have that $h'(w) < 0$, which leads to conclude that $r(w)/w^2$ is non-increasing on $(0, \infty)$. Then it is clear that $r(w)/w^2$ is non-increasing on $(0, \infty)$. □

Now we prove the second part of this theorem.

Proof. define the first-order difference matrix $\boldsymbol{D} \in \mathbb{R}^{(n-1)\times n}$ as

$$\boldsymbol{D} = \begin{bmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{bmatrix}. \tag{10}$$

Observe that it is clear that $\boldsymbol{D}$ is full row rank. Then, define $g_{TV}(\boldsymbol{x}) = g(\boldsymbol{D}\boldsymbol{x})$ where g is an admissible function. It is worth noticing that since g is point-wise non-increasing when divided by a quadratic mapping, then g is continuously differentiable. In addition, due to the fact that $\boldsymbol{D}$ is full row rank, we appeal to (Pinilla et al., 2022, Lemma 2) that ensures $g_{TV}(\boldsymbol{x})$ is invex. Thus the result holds. □

B.1 Additional Discussion on Theorem 4

In this section we discuss the 2D and 3D total-variation-like extensions of invex functions.

2D Total-variation-like regularizer

Let be g an admissible function. Then, the 2D total-variation-like regularizer based on is defined as

$$g_{TV}(\boldsymbol{x}) = g(\boldsymbol{D}_h\boldsymbol{x}) + g(\boldsymbol{D}_v\boldsymbol{x}), \tag{11}$$

where matrix $\boldsymbol{D}_h$, and $\boldsymbol{D}_v$ model the discrete derivative in the horizontal and vertical directions, respectively. Therefore, $\boldsymbol{D}_h$ is defined as in equation (10) and $\boldsymbol{D}_v$ is given as

$$\boldsymbol{D}_v = \begin{bmatrix} -1 & \cdots & 1 & & \\ & -1 & \cdots & 1 & \\ & & \ddots & \ddots & \\ & & -1 & \cdots & 1 \end{bmatrix}. \tag{12}$$

Now we prove that equation (11) is quasi-invex as follows.

Proof. Recall in the previous section we showed that each term $g(\boldsymbol{D}_h\boldsymbol{x})$ and $g(\boldsymbol{D}_v\boldsymbol{x})$ are invex (since $\boldsymbol{D}_h$, and $\boldsymbol{D}_v$ are full row-rank). Therefore, there exits $\eta_h(\boldsymbol{x},\boldsymbol{y})$ and $\eta_v(\boldsymbol{x},\boldsymbol{y})$ for any $\boldsymbol{x},\boldsymbol{y} \in \mathbb{R}^n$ such that

$$g(\boldsymbol{D}_h\boldsymbol{x}) - g(\boldsymbol{D}_h\boldsymbol{y}) \geq \boldsymbol{\zeta}_h^T\eta_h(\boldsymbol{x},\boldsymbol{y}) \tag{13}$$

for all $\boldsymbol{\zeta}_h \in \partial g(\boldsymbol{D}_h\boldsymbol{y})$ and

$$g(\boldsymbol{D}_v\boldsymbol{x}) - g(\boldsymbol{D}_v\boldsymbol{y}) \geq \boldsymbol{\zeta}_v^T\eta_v(\boldsymbol{x},\boldsymbol{y}) \tag{14}$$

for all $\boldsymbol{\zeta}_v \in \partial g(\boldsymbol{D}_v\boldsymbol{y})$. Then, assume

$$g_{TV}(\boldsymbol{x}) - g_{TV}(\boldsymbol{y}) \leq 0. \tag{15}$$

This leads to

$$\begin{aligned} g(\boldsymbol{D}_h\boldsymbol{x}) - g(\boldsymbol{D}_h\boldsymbol{y}) + g(\boldsymbol{D}_v\boldsymbol{x}) - g(\boldsymbol{D}_v\boldsymbol{y}) &\leq 0 \\ \boldsymbol{\zeta}_h^T\eta_h(\boldsymbol{x},\boldsymbol{y}) + \boldsymbol{\zeta}_v^T\eta_v(\boldsymbol{x},\boldsymbol{y}) &\leq 0 \\ \boldsymbol{\zeta}_g^T\eta_g(\boldsymbol{x},\boldsymbol{y}) &\leq 0 \end{aligned} \tag{16}$$

where $\boldsymbol{\zeta}_g = [\boldsymbol{\zeta}_h^T, \boldsymbol{\zeta}_v^T]^T \in \partial g_{TV}(\boldsymbol{y})$, and $\eta_{g_{TV}}(\boldsymbol{x},\boldsymbol{y}) = [\eta_h^T(\boldsymbol{x},\boldsymbol{y}), \eta_v^T(\boldsymbol{x},\boldsymbol{y})]^T$. Thus, from the above inequality we have that $g_{TV}(\boldsymbol{x})$ is quasi-invex. □

3D Total-variation-like regularizer

Let g be an admissible function. Then, the 3D total-variation-like regularizer is defined as

$$g_{TV}(\boldsymbol{x}) = g(\boldsymbol{D}_h\boldsymbol{x}) + g(\boldsymbol{D}_v\boldsymbol{x}) + g(\boldsymbol{D}_t\boldsymbol{x}), \tag{17}$$

where matrix $\boldsymbol{D}_h$, $\boldsymbol{D}_v$, $\boldsymbol{D}_t$ model the discrete derivative in the horizontal, vertical, and transversal (third dimension) directions, respectively. Therefore, $\boldsymbol{D}_h$ is defined as in equation (10) and $\boldsymbol{D}_v$ as in equation (12), and $\boldsymbol{D}_t$ is given as

$$\boldsymbol{D}_t = \begin{bmatrix} -1 & \cdots & \cdots & 1 & & & \\ & -1 & \cdots & \cdots & 1 & & \\ & & \ddots & \ddots & \ddots & & \\ & & & -1 & \cdots & \cdots & 1 \end{bmatrix}. \tag{18}$$

Now we prove that equation (11) is quasi-invex as follows.

Proof. Recall in the previous section we showed that each term $g(\boldsymbol{D}_h\boldsymbol{x})$, $g(\boldsymbol{D}_v\boldsymbol{x})$, and $g(\boldsymbol{D}_t\boldsymbol{x})$ are invex (since $\boldsymbol{D}_h$, $\boldsymbol{D}_v$, and $\boldsymbol{D}_t$ are full row-rank). Therefore, there exits $\eta_h(\boldsymbol{x},\boldsymbol{y})$, $\eta_v(\boldsymbol{x},\boldsymbol{y})$, and $\eta_t(\boldsymbol{x},\boldsymbol{y})$ for any $\boldsymbol{x},\boldsymbol{y} \in \mathbb{R}^n$ such that

$$g(\boldsymbol{D}_h\boldsymbol{x}) - g(\boldsymbol{D}_h\boldsymbol{y}) \geq \boldsymbol{\zeta}_h^T\eta_h(\boldsymbol{x},\boldsymbol{y}) \tag{19}$$

for all $\boldsymbol{\zeta}_h \in \partial g(\boldsymbol{D}_h\boldsymbol{y})$

$$g(\boldsymbol{D}_v\boldsymbol{x}) - g(\boldsymbol{D}_v\boldsymbol{y}) \geq \boldsymbol{\zeta}_v^T\eta_v(\boldsymbol{x},\boldsymbol{y}) \tag{20}$$

for all $\boldsymbol{\zeta}_v \in \partial g(\boldsymbol{D}_v\boldsymbol{y})$, and

$$g(\boldsymbol{D}_t\boldsymbol{x}) - g(\boldsymbol{D}_t\boldsymbol{y}) \geq \boldsymbol{\zeta}_t^T\eta_t(\boldsymbol{x},\boldsymbol{y}) \tag{21}$$

for all $\boldsymbol{\zeta}_t \in \partial g(\boldsymbol{D}_t\boldsymbol{y})$. Then, assume

$$g_{TV}(\boldsymbol{x}) - g_{TV}(\boldsymbol{y}) \leq 0. \tag{22}$$

This leads to

$$\begin{aligned} g(\boldsymbol{D}_h\boldsymbol{x}) - g(\boldsymbol{D}_h\boldsymbol{y}) + g(\boldsymbol{D}_v\boldsymbol{x}) - g(\boldsymbol{D}_v\boldsymbol{y}) + g(\boldsymbol{D}_t\boldsymbol{x}) - g(\boldsymbol{D}_t\boldsymbol{y}) &\leq 0 \\ \boldsymbol{\zeta}_h^T\eta_h(\boldsymbol{x},\boldsymbol{y}) + \boldsymbol{\zeta}_v^T\eta_v(\boldsymbol{x},\boldsymbol{y}) + \boldsymbol{\zeta}_t^T\eta_t(\boldsymbol{x},\boldsymbol{y}) &\leq 0 \\ \boldsymbol{\zeta}_g^T\eta_g(\boldsymbol{x},\boldsymbol{y}) &\leq 0 \end{aligned} \tag{23}$$

where $\boldsymbol{\zeta}_g = [\zeta_h^T, \zeta_v^T, \zeta_t^T]^T \in \partial g_{TV}(\boldsymbol{y})$, and $\eta_{g_{TV}}(\boldsymbol{x},\boldsymbol{y}) = [\eta_h^T(\boldsymbol{x},\boldsymbol{y}), \eta_v^T(\boldsymbol{x},\boldsymbol{y}), \eta_g^T(\boldsymbol{x},\boldsymbol{y})]^T$. Thus, from the above inequality we have that $g_{TV}(\boldsymbol{x})$ is quasi-invex. □

C Proof of Theorem 5

In this proof we seek to guarantee that the list of functions in Theorem 4 are admissible functions, and we proceed by cases.

Equation (4)

Proof. Take $r(w) = \log(1 + \frac{w^2}{\delta^2})$ for any $w \neq 0$, and fixed $\delta \in \mathbb{R}$. It is trivial to see that $r(0) = 0$, that $r(w)$ it is not identically zero, and non-decreasing on $(0, \infty)$. Then, we just need to show that $r(w)/w^2$ is non-increasing on $(0, \infty)$. Observe that the first derivative of $h(w) = r(w)/w^2$ is given by $h'(w) = \frac{2\left(\frac{w^2}{\delta^2+w^2} - \log(1+\frac{w^2}{\delta^2})\right)}{w^3}$. Since $\frac{w^2}{\delta^2+w^2} - \log(1 + \frac{w^2}{\delta^2}) < 0$, then we have that $h'(w) < 0$, which leads to conclude that $r(w)/w^2$ is non-increasing on $(0, \infty)$. Then it is clear that $r(w)/w^2$ is non-increasing on $(0, \infty)$. □

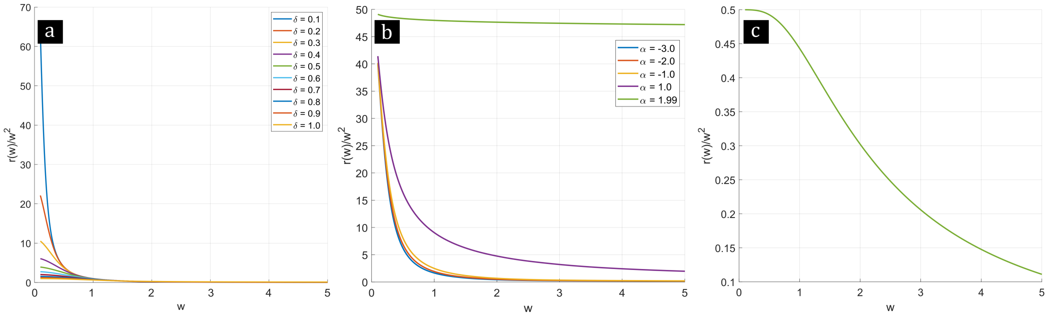


Figure 1: Plot of $r(w)/w$ for $r(w)$ being (a) equation (6), (b) equation (7) for $c = 0.1$, and (c) equation (8) and $w > 0$ to check that $r(w)/w^2$ is non-increasing on $(0, \infty)$

EQUATION (5)

Proof. Take $r(w) = \frac{2w^2}{w^2+4\delta^2}$ for any $w \neq 0$, and fixed $\delta \in \mathbb{R}$. It is trivial to see that $r(0) = 0$, that $r(w)$ it is not identically zero, and non-decreasing on $(0, \infty)$. Then, we just need to show that $r(w)/w^2$ is non-increasing on $(0, \infty)$. Observe that $h(w) = r(w)/w^2$ is given by $h(w) = \frac{2}{w^2+4\delta^2}$, which leads to conclude that $r(w)/w^2$ is non-increasing on $(0, \infty)$. Then it is clear that $r(w)/w^2$ is non-increasing on $(0, \infty)$. □

EQUATION (6)

Take $r(w) = 1 - \exp(-w^2/\delta^2)$ for any $w \neq 0$, and fixed $\delta \in \mathbb{R}$. It is trivial to see that $r(0) = 0$, that $r(w)$ it is not identically zero, and non-decreasing on $(0, \infty)$. Then, we just need to show that $r(w)/w^2$ is non-increasing on $(0, \infty)$. For easy of exposition we present in Figure 1(a) the plot of $r(w)/w^2$. Then it is clear that $r(w)/w^2$ is non-increasing on $(0, \infty)$.

EQUATION (7)

Take $r(w) = \frac{|\alpha-2|}{\alpha}\left(\left(\frac{(w/c)^2}{|\alpha-2|}+1\right)^{\alpha/2}-1\right)$ for any $w \neq 0$, and fixed $\alpha \in \mathbb{R}$, $c > 0$. It is trivial to see that $r(0) = 0$, that $r(w)$ it is not identically zero, and non-decreasing on $(0, \infty)$. Then, we just need to show that $r(w)/w^2$ is non-increasing on $(0, \infty)$. For easy of exposition we present in Figure 1(b) the plot of $r(w)/w^2$. Then it is clear that $r(w)/w^2$ is non-increasing on $(0, \infty)$.

EQUATION (8)

Take $r(w) = \log\left(1 + w^2\right) - \frac{w^2}{2w^2+2}$ for any $w \neq 0$. It is trivial to see that $r(0) = 0$, that $r(w)$ it is not identically zero, and non-decreasing on $(0, \infty)$. Then, we just need to show that $r(w)/w^2$ is non-increasing on $(0, \infty)$. For easy of exposition we present in Figure 1(c) the plot of $r(w)/w^2$. Then it is clear that $r(w)/w^2$ is non-increasing on $(0, \infty)$.

D Proof of Theorem 6

To prove that $g_\theta(\boldsymbol{x})$ is quasi-invex, for regularizers in Theorem 6, we show the existence of $\eta : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ satisfying Definition 3 for $g_\theta(\boldsymbol{x})$.

EQUATION (9)

Proof. Let $g_\theta(\boldsymbol{x}) = \lambda \sum_{i=1}^n r_\theta(\boldsymbol{x}[i])$, for $\theta \geq 1$, $\lambda \in (0,1]$, and

$$r_\theta(w) = \begin{cases} \lambda|w| - w^2/(2\theta) & \text{if } |w| < \theta\lambda \\ \theta\lambda^2/2 & \text{if } |w| \geq \theta\lambda \end{cases}. \tag{24}$$

First, we will show that if $|w| < \theta\lambda$ then $r_\theta(w)$ is invex. In consequence, $g_\theta(\boldsymbol{x})$ is invex for any $\boldsymbol{x} \in \mathcal{D}$ (see Lemma 1), with

$$\mathcal{D} = \{\boldsymbol{z} \in \mathbb{R}^n : |\boldsymbol{z}[i]| < \theta\lambda, \forall i \in \{1, \dots, n\}\}.$$

Once we prove this, we are able to build $\eta : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ such that $g_\theta(\boldsymbol{x})$ is quasi-invex for any $\boldsymbol{x} \in \mathbb{R}^n$.

Observe that $r'_\theta(w)$ for $0 < |w| < \theta\lambda$ is given by

$$r'_\theta(w) = \lambda \frac{w}{|w|} - \frac{w}{\theta}. \tag{25}$$

The above equation implies that $0 \in \partial r_\theta(w)$ only when $w = 0$. Further, since $r_\theta(0) < r_\theta(w)$ for any $0 < |w| < \theta\lambda$, then $w = 0$ is a global minimizer of $r_\theta(w)$, in the restricted domain, implying its invexity. Therefore, appealing to Lemma 1 there exists $\eta_r : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ such that $g_\theta(\boldsymbol{x})$ is invex in $\mathcal{D}$. It is worth mentioning that $\mathcal{D}$ is an open set, because it is the Cartesian product of n open sets. This observation is important because invexity requires to compute the subgradient which is only possible on open sets. Thus, the invexity of $g_\theta(\boldsymbol{x})$ is well defined. And therefore, $g_\theta(\boldsymbol{x})$ is quasi-invex in $\mathcal{D}$.

In consequence, define the following $\eta : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ as

$$\eta(\boldsymbol{x}, \boldsymbol{y}) = \begin{cases} \eta_r(\boldsymbol{x}, \boldsymbol{y}) & \text{if } \boldsymbol{x}, \boldsymbol{y} \in \mathcal{D} \\ \boldsymbol{0} & \text{otherwise} \end{cases}. \tag{26}$$

Considering the above $\eta(\boldsymbol{x}, \boldsymbol{y})$ function (which is non-zero), take $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, and assume

$$g_\theta(\boldsymbol{x}) - g_\theta(\boldsymbol{y}) \leq 0. \tag{27}$$

If $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{D}$, then we know $g_\theta(\boldsymbol{x})$ is invex, which implies that

$$\begin{aligned} g_\theta(\boldsymbol{x}) - g_\theta(\boldsymbol{y}) &\geq \boldsymbol{\zeta}^T \eta(\boldsymbol{x}, \boldsymbol{y}) \\ 0 &\geq \boldsymbol{\zeta}^T \eta(\boldsymbol{x}, \boldsymbol{y}), \end{aligned} \tag{28}$$

for all $\boldsymbol{\zeta} \in \partial g_\theta(\boldsymbol{y})$, and therefore quasi-invex. Otherwise, if either $\boldsymbol{x} \notin \mathcal{D}$ or $\boldsymbol{y} \notin \mathcal{D}$, it is clear to conclude that $\boldsymbol{\zeta}^T \eta(\boldsymbol{x}, \boldsymbol{y}) = 0$, for all $\boldsymbol{\zeta} \in \partial g_\theta(\boldsymbol{y})$. Thus, the result holds. □

EQUATION (10)

Proof. Let $g_\theta(\boldsymbol{x}) = \sum_{i=1}^n r_\theta(\boldsymbol{x}[i])$ be the regularizer in equation (10) for $\theta > 0$, and $r_\theta(\boldsymbol{x}[i]) = \min(r(\boldsymbol{x}[i]), \theta)$ where function $r(w)$ takes the form of

$$r(w) = \log(1 + (|w| + \epsilon)^p), p \in (0,1), \epsilon > 0. \tag{29}$$

To prove that equation (10) is quasi-invex we first show that the unidimensional function $r(w)$ is invex and then we proceed to show the construction of $\eta : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$. To prove that function $r(w)$ is invex observe that if $w > 0$ then we have that $\partial r(w) = \left\{\frac{p}{(|w|+\epsilon)^{1-p} + (|w|+\epsilon)}\right\}$, which means that $0 \notin \partial r(w)$. Conversely, if $w < 0$ then $\partial r(w) = \left\{\frac{-p}{(|w|+\epsilon)^{1-p} + (|w|+\epsilon)}\right\}$, leading to $0 \notin \partial r(w)$. Lets examine $w^* = 0$. Note that $\lim_{w \to 0^+} r'(w) = \frac{p}{\epsilon^{1-p} + \epsilon}$, and that $\lim_{w \to 0^-} r'(w) = \frac{-p}{\epsilon^{1-p} + \epsilon}$. Additionally, since $r(w)$ is a Lipschitz continuous function, then appealing to Theorem 1 we have that $\partial r(w^* = 0) = \text{conv}\left\{\frac{-p}{\epsilon^{1-p}+\epsilon}, \frac{p}{\epsilon^{1-p}+\epsilon}\right\} = \left[\frac{-p}{\epsilon^{1-p}+\epsilon}, \frac{p}{\epsilon^{1-p}+\epsilon}\right]$. This means that $0 \in \partial r(0)$. Further, given the fact that $r(0) \leq r(w)$ for all $w \in \mathbb{R}$, then $w^* = 0$ is a global minimizer of $r(w)$. Therefore, the function $r(w)$ is invex.

Now we show the existence of $\eta : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ satisfying Definition 3 for $g_\theta(\boldsymbol{x})$. First, observe that if function $r(w)$ does not reach the value of θ for any w, then it is clear the existence of $\eta_g : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ such that $g_\theta(\boldsymbol{x}) = \sum_{i=1}^n r(\boldsymbol{x}[i])$ is invex according to Lemma 1, and therefore quasi-invex. On the other hand, if we assume $r(w)$ reaches the value of θ, then there exists $w^* \geq 0$ such that $r(w) < \theta$ for any $|w| < w^*$. Let $\mathcal{D} = \{\boldsymbol{z} \in \mathbb{R}^n : |\boldsymbol{z}[i]| < w^*, \forall i \in \{1, \dots, n\}\}$. It is worth mentioning that $\mathcal{D}$ is an open set, because it is the Cartesian product of n open sets. This observation is important because invexity requires to compute the subgradient which is only possible on open sets. Thus, the invexity of $g_\theta(\boldsymbol{x})$ in $\mathcal{D}$ is well defined. Now, we define the following $\eta : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ as

$$\eta(\boldsymbol{x}, \boldsymbol{y}) = \begin{cases} \eta_g(\boldsymbol{x}, \boldsymbol{y}) & \text{if } \boldsymbol{x}, \boldsymbol{y} \in \mathcal{D} \\ \boldsymbol{0} & \text{otherwise} \end{cases} . \tag{30}$$

Considering the above $\eta(\boldsymbol{x}, \boldsymbol{y})$ (which is non-zero), take $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, and assume

$$g_\theta(\boldsymbol{x}) - g_\theta(\boldsymbol{y}) \leq 0. \tag{31}$$

If $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{D}$, then we know $g_\theta(\boldsymbol{x})$ is invex, which implies that

$$\begin{aligned} g_\theta(\boldsymbol{x}) - g_\theta(\boldsymbol{y}) &\geq \boldsymbol{\zeta}^T \eta(\boldsymbol{x}, \boldsymbol{y}) \\ 0 &\geq \boldsymbol{\zeta}^T \eta(\boldsymbol{x}, \boldsymbol{y}), \end{aligned} \tag{32}$$

for all $\boldsymbol{\zeta} \in \partial g_\theta(\boldsymbol{y})$, and therefore quasi-invex. Otherwise, if either $\boldsymbol{x} \notin \mathcal{D}$ or $\boldsymbol{y} \notin \mathcal{D}$, it is clear to conclude that $\boldsymbol{\zeta}^T \eta(\boldsymbol{x}, \boldsymbol{y}) = 0$, for all $\boldsymbol{\zeta} \in \partial g_\theta(\boldsymbol{y})$. Thus, the result holds. □

D.1 Third Part of Theorem 6

Proof. Let g be an admissible function such that $g(\boldsymbol{x}) = \sum_{i=1}^{n-1} s_g(|\boldsymbol{x}[i]|)$, and $s_g(w) \geq 0$. Define $g_{TV}(\boldsymbol{x}) = g(\boldsymbol{D}\boldsymbol{x})$ where $\boldsymbol{D}$ as defined in equation (10). Then, according to Theorem 3 we have that $g(\boldsymbol{x})$ is quasi-invex for some $\eta_g(\boldsymbol{x}, \boldsymbol{y})$, which implies that

$$g_{TV}(\boldsymbol{x}) - g_{TV}(\boldsymbol{y}) \leq 0 \implies \boldsymbol{\zeta}^T \eta_g(\boldsymbol{D}\boldsymbol{x}, \boldsymbol{D}\boldsymbol{y}) \leq 0, \tag{33}$$

for all $\boldsymbol{\zeta} \in \partial g(\boldsymbol{D}\boldsymbol{y})$. We know that matrix $\boldsymbol{D}$ is full row-rank which means that $(\boldsymbol{D}\boldsymbol{D}^T)^{-1}$ exits and $(\boldsymbol{D}\boldsymbol{D}^T)(\boldsymbol{D}\boldsymbol{D}^T)^{-1} = \boldsymbol{I}_{n-1}$ where $\boldsymbol{I}_{n-1}$ is the identity matrix. Then, from the above inequality we obtain

$$\begin{aligned} g_{TV}(\boldsymbol{x}) - g_{TV}(\boldsymbol{y}) \leq 0 &\implies \boldsymbol{\zeta}^T (\boldsymbol{D}\boldsymbol{D}^T)(\boldsymbol{D}\boldsymbol{D}^T)^{-1} \eta_g(\boldsymbol{D}\boldsymbol{x}, \boldsymbol{D}\boldsymbol{y}) \leq 0 \\ &\implies \boldsymbol{\zeta}^T \boldsymbol{D} \left(\boldsymbol{D}^T (\boldsymbol{D}\boldsymbol{D}^T)^{-1} \eta_g(\boldsymbol{D}\boldsymbol{x}, \boldsymbol{D}\boldsymbol{y})\right) \\ &\implies (\boldsymbol{D}^T \boldsymbol{\zeta})^T \left(\boldsymbol{D}^T (\boldsymbol{D}\boldsymbol{D}^T)^{-1} \eta_g(\boldsymbol{D}\boldsymbol{x}, \boldsymbol{D}\boldsymbol{y})\right) \\ &\implies (\boldsymbol{D}^T \boldsymbol{\zeta})^T \eta(\boldsymbol{x}, \boldsymbol{y}), \end{aligned} \tag{34}$$

where $\eta(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{D}^T (\boldsymbol{D}\boldsymbol{D}^T)^{-1} \eta_g(\boldsymbol{D}\boldsymbol{x}, \boldsymbol{D}\boldsymbol{y})$. Further, since g is continuously differentiable concluded from (Bagirov et al., 2014, Theorem 3.20) we know that $\boldsymbol{D}^T \boldsymbol{\zeta} \in \partial g_{TV}(\boldsymbol{y})$. Thus, from equation (34) we get

$$g_{TV}(\boldsymbol{x}) - g_{TV}(\boldsymbol{y}) \leq 0 \implies \boldsymbol{\zeta}^T \eta(\boldsymbol{x}, \boldsymbol{y}), \tag{35}$$

for all $\boldsymbol{\zeta} \in \partial g_{TV}(\boldsymbol{y})$, which implies $g_{TV}(\boldsymbol{x})$ is quasi-invex. Therefore, the result holds. □

D.2 Proof SCAD is Quasi-invex

Proof. Let $g_\theta(\boldsymbol{x}) = \lambda \sum_{i=1}^n r_\theta(\boldsymbol{x}[i])$, for $\theta > 2$, $\lambda \in (0, 1]$, and

$$r_\theta(w) = \begin{cases} \lambda|w| & \text{if } |w| \leq \lambda \\ \frac{-w^2 + 2\theta\lambda|w| - \lambda^2}{2(\theta - 1)} & \text{if } \lambda < |w| < \theta\lambda \\ (\theta + 1)\lambda^2/2 & \text{otherwise} \end{cases} . \tag{36}$$

First, we will show that if $|w| < \theta\lambda$ then $r_\theta(w)$ is invex. In consequence, $g_\theta(\boldsymbol{x})$ is invex for any $\boldsymbol{x} \in \mathcal{D}$ (see Lemma 1), with

$$\mathcal{D} = \{\boldsymbol{z} \in \mathbb{R}^n : |\boldsymbol{z}[i]| < \theta\lambda, \forall i \in \{1, \dots, n\}\}.$$

Once we prove this, we are able to build $\eta : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ such that $g_\theta(\boldsymbol{x})$ is quasi-invex for any $\boldsymbol{x} \in \mathbb{R}^n$.

Observe that $r'_\theta(w)$ for $0 < |w| < \theta\lambda$ is given by

$$r'_\theta(w) = \begin{cases} \lambda \frac{w}{|w|} & \text{if } |w| \leq \lambda \\ \frac{-w + \frac{\theta\lambda w}{|w|}}{\theta - 1} & \text{if } \lambda < |w| < \theta\lambda. \end{cases} \tag{37}$$

The above equation implies that $0 \in \partial r_\theta(w)$ only when $w = 0$. Further, since $r_\theta(0) < r_\theta(w)$ for any $0 < |w| < \theta\lambda$, then $w = 0$ is a global minimizer of $r_\theta(w)$, in the restricted domain, implying its invexity. Therefore, appealing to Lemma 1 there exists $\eta_r : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ such that $g_\theta(\boldsymbol{x})$ is invex in $\mathcal{D}$. It is worth mentioning that $\mathcal{D}$ is an open set, because it is the Cartesian product of n open sets. This observation is important because invexity requires to compute the subgradient which is only possible on open sets. Thus, the invexity of $g_\theta(\boldsymbol{x})$ is well defined. And therefore, $g_\theta(\boldsymbol{x})$ is quasi-invex in $\mathcal{D}$.

In consequence, define the following $\eta : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ as

$$\eta(\boldsymbol{x}, \boldsymbol{y}) = \begin{cases} \eta_r(\boldsymbol{x}, \boldsymbol{y}) & \text{if } \boldsymbol{x}, \boldsymbol{y} \in \mathcal{D} \\ \mathbf{0} & \text{otherwise} \end{cases} . \tag{38}$$

Considering the above $\eta(\boldsymbol{x}, \boldsymbol{y})$ function (which is non-zero), take $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, and assume

$$g_\theta(\boldsymbol{x}) - g_\theta(\boldsymbol{y}) \leq 0. \tag{39}$$

If $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{D}$, then we know $g_\theta(\boldsymbol{x})$ is invex, which implies that

$$\begin{aligned} g_\theta(\boldsymbol{x}) - g_\theta(\boldsymbol{y}) &\geq \boldsymbol{\zeta}^T \eta(\boldsymbol{x}, \boldsymbol{y}) \\ 0 &\geq \boldsymbol{\zeta}^T \eta(\boldsymbol{x}, \boldsymbol{y}), \end{aligned} \tag{40}$$

for all $\boldsymbol{\zeta} \in \partial g_\theta(\boldsymbol{y})$, and therefore quasi-invex. Otherwise, if either $\boldsymbol{x} \notin \mathcal{D}$ or $\boldsymbol{y} \notin \mathcal{D}$, it is clear to conclude that $\boldsymbol{\zeta}^T \eta(\boldsymbol{x}, \boldsymbol{y}) = 0$, for all $\boldsymbol{\zeta} \in \partial g_\theta(\boldsymbol{y})$. Thus, the result holds. □

D.3 Proof Capped-ℓ_1 is Quasi-invex

Proof. Let $g_\theta(\boldsymbol{x}) = \lambda \sum_{i=1}^n r_\theta(\boldsymbol{x}[i])$, for $\theta > 0$, $\lambda \in (0, 1]$, and

$$r_\theta(w) = \min(|w|, \theta).$$

Now we show the existence of $\eta : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ satisfying Definition 3 for $g_\theta(\boldsymbol{x})$. Let $\mathcal{D} = \{\boldsymbol{z} \in \mathbb{R}^n : |\boldsymbol{z}[i]| < \theta, \forall i \in \{1, \ldots, n\}\}$. Thus, we define $\eta : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ as

$$\eta(\boldsymbol{x}, \boldsymbol{y}) = \begin{cases} \boldsymbol{x} - \boldsymbol{y} & \text{if } \boldsymbol{x}, \boldsymbol{y} \in \mathcal{D} \\ \mathbf{0} & \text{otherwise} \end{cases} . \tag{41}$$

Considering the above $\eta(\boldsymbol{x}, \boldsymbol{y})$ (which is non-zero), take $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, and assume

$$g_\theta(\boldsymbol{x}) - g_\theta(\boldsymbol{y}) \leq 0. \tag{42}$$

If $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{D}$, then we know $g_\theta(\boldsymbol{x})$ is the ℓ_1-norm, which implies that

$$\begin{aligned} g_\theta(\boldsymbol{x}) - g_\theta(\boldsymbol{y}) &\geq \boldsymbol{\zeta}^T (\boldsymbol{x} - \boldsymbol{y}) \\ 0 &\geq \boldsymbol{\zeta}^T (\boldsymbol{x} - \boldsymbol{y}), \end{aligned} \tag{43}$$

for all $\boldsymbol{\zeta} \in \partial g_\theta(\boldsymbol{y})$, and therefore quasi-invex. Otherwise, if either $\boldsymbol{x} \notin \mathcal{D}$ or $\boldsymbol{y} \notin \mathcal{D}$, it is clear to conclude that $\boldsymbol{\zeta}^T \eta(\boldsymbol{x}, \boldsymbol{y}) = 0$, for all $\boldsymbol{\zeta} \in \partial g_\theta(\boldsymbol{y})$. Thus, the result holds. □

E Proof of Lemma 1

Proof. Let $f, h : \mathbb{R}^n \to \mathbb{R}$ be two invex functions (not necessarily with respect to the same η). Define the function $q(\boldsymbol{x}) = \max(f(\boldsymbol{x}), h(\boldsymbol{x}))$. Now we will show that $q(\boldsymbol{x})$ satisfies Definition 3. Then, take $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ and assume

$$q(\boldsymbol{x}) - q(\boldsymbol{y}) \leq 0. \tag{44}$$

The above inequality implies that

$$\begin{aligned} \max(f(\boldsymbol{x}), h(\boldsymbol{x})) - \max(f(\boldsymbol{y}), h(\boldsymbol{y})) \leq 0 \\ \max(f(\boldsymbol{x}), h(\boldsymbol{x})) \leq \max(f(\boldsymbol{y}), h(\boldsymbol{y})). \end{aligned} \tag{45}$$

Now we proceed by cases.

1. Assume $q(\boldsymbol{y}) = \max(f(\boldsymbol{y}), h(\boldsymbol{y})) = f(\boldsymbol{y})$, then from equation (45) we have

   $$\begin{aligned} f(\boldsymbol{x}) \leq \max(f(\boldsymbol{x}), h(\boldsymbol{x})) \leq \max(f(\boldsymbol{y}), h(\boldsymbol{y})) = f(\boldsymbol{y}) \\ f(\boldsymbol{x}) \leq f(\boldsymbol{y}). \end{aligned} \tag{46}$$

   Since f is an invex function, then we know there exits $\eta_f : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ such that equation (46) implies

   $$\begin{aligned} f(\boldsymbol{x}) - f(\boldsymbol{y}) \geq \boldsymbol{\zeta}^T \eta_f(\boldsymbol{x}, \boldsymbol{y}) \\ 0 \geq \boldsymbol{\zeta}^T \eta_f(\boldsymbol{x}, \boldsymbol{y}) \end{aligned} \tag{47}$$

   for all $\boldsymbol{\zeta} \in \partial f(\boldsymbol{y}) = \partial q(\boldsymbol{y})$.

2. Assume $q(\boldsymbol{y}) = \max(f(\boldsymbol{y}), h(\boldsymbol{y})) = h(\boldsymbol{y})$, then from equation (45) we have

   $$\begin{aligned} h(\boldsymbol{x}) \leq \max(f(\boldsymbol{x}), h(\boldsymbol{x})) \leq \max(f(\boldsymbol{y}), h(\boldsymbol{y})) = h(\boldsymbol{y}) \\ h(\boldsymbol{x}) \leq h(\boldsymbol{y}). \end{aligned} \tag{48}$$

   Since h is an invex function, then we know there exits $\eta_h : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ such that equation (48) implies

   $$\begin{aligned} h(\boldsymbol{x}) - h(\boldsymbol{y}) \geq \boldsymbol{\zeta}^T \eta_h(\boldsymbol{x}, \boldsymbol{y}) \\ 0 \geq \boldsymbol{\zeta}^T \eta_h(\boldsymbol{x}, \boldsymbol{y}) \end{aligned} \tag{49}$$

   for all $\boldsymbol{\zeta} \in \partial h(\boldsymbol{y}) = \partial q(\boldsymbol{y})$.

From the above discussed cases, we build $\eta : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ as

$$\eta(\boldsymbol{x}, \boldsymbol{y}) = \begin{cases} \eta_f(\boldsymbol{x}, \boldsymbol{y}) & \text{if } f(\boldsymbol{y}) > h(\boldsymbol{y}) \\ \mathbf{0} & \text{if } f(\boldsymbol{y}) = h(\boldsymbol{y}) \\ \eta_h(\boldsymbol{x}, \boldsymbol{y}) & \text{otherwise} \end{cases} . \tag{50}$$

Thus, from equation (46), equation (48), and the above definition of $\eta(\boldsymbol{x}, \boldsymbol{y})$, we conclude

$$q(\boldsymbol{x}) - q(\boldsymbol{y}) \leq 0 \implies \boldsymbol{\zeta}^T \eta(\boldsymbol{x}, \boldsymbol{y}) \leq 0, \tag{51}$$

$\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, and $\forall \boldsymbol{\zeta} \in \partial q(\boldsymbol{y})$. Therefore the desired result holds. □

F Global optima guarantees of program (1)

To prove the global optima guarantees of optimization problem in (1) we formulate and prove the following theorem.

Theorem 1. Let $f, g : \mathbb{R}^n \to \mathbb{R}$ be admissible functions as in Definition 4. Assume $\boldsymbol{x}^\star$ is a solution of program (1) for $\epsilon = 0$, then $\boldsymbol{x}^\star$ is a global minimizer.

Proof. Assume $\boldsymbol{x}^*$ is a minimizer of constrained optimization problem in equation (1), and define $\mathcal{D} = \{\boldsymbol{x} \in \mathbb{R}^n | f(\boldsymbol{x}) = 0\}$. This means that $\boldsymbol{x}^*$ is a minimizer of $g(\boldsymbol{x})$ in $\mathcal{D}$ i.e. $\boldsymbol{x}^* \in \mathcal{D}$. Since, $f(\boldsymbol{x})$ is an admissible function, then is invex. Therefore, $\boldsymbol{x}^*$ must be a global minimizer. Thus the result holds. □

G Proof of Theorem 7

In order to prove Theorem 7 we introduce an auxiliary definition and lemmata as follows.

Definition 1. (*Sparseness measure* (Gribonval & Nielsen, 2007)) Let $g : \mathbb{R}^n \to \mathbb{R}$ such that $g(\boldsymbol{x}) = \sum_{i=1}^{n} r(\boldsymbol{x}[i])$, where $r : [0, \infty) \to [0, \infty)$ and increasing. If r, not identically zero, with $r(0) = 0$ such that $r(w)/w$ is non-increasing on $(0, \infty)$, then $g(\boldsymbol{x})$ is said to be a *sparseness measure*.

Considering the above definitions we obtain the following conclusions.

Lemma 2. (Gribonval & Nielsen, 2007, Proposition 1) Let $g : \mathbb{R}^n \to \mathbb{R}$ be a function satisfying Definition 1. Then, g satisfies the triangle inequality.

Observe that the above lemma provides a practical methodology to verify when the function $g(\boldsymbol{x})$ satisfies the triangle inequality.

Lemma 3. Let $f : \mathbb{R}^n \to \mathbb{R}$ be an admissible function. Take $\boldsymbol{x} \in \mathbb{R}^n$ such that $\|\boldsymbol{x}\|_\infty < \infty$. If $f(\boldsymbol{x}) < \infty$, then $\|\boldsymbol{x}\|_2^2 < \frac{1}{c(\boldsymbol{x})} f(\boldsymbol{x})$ for a constant $c(\boldsymbol{x}) > 0$.

Proof. Since f is an admissible function, then $f(\boldsymbol{x}) = \sum_{i=1}^{n} s(\boldsymbol{x}[i])$ and $s(w)/w^2$ is nonincreasing. Therefore, we have for all $i = 1, \ldots, n$

$$\frac{s(\boldsymbol{x}[i])}{\boldsymbol{x}^2[i]} \geq \frac{s(\|\boldsymbol{x}\|_\infty)}{\|\boldsymbol{x}\|_\infty^2} = c(\boldsymbol{x}) > 0. \tag{52}$$

Thus, we obtain

$$\|\boldsymbol{x}\|_2^2 \leq \frac{1}{c(\boldsymbol{x})} f(\boldsymbol{x}). \tag{53}$$

Thus the result holds. □

With the above definitions and lemmata we proceed to prove Theorem 7. This proof assumes the regularizer $g(\boldsymbol{x})$ satisfies Definition 1. This assumption is proved at the end of this section.

Proof. From hypothesis we assume matrix $\boldsymbol{A}$ satisfies the RIP condition for any k-sparse vector with $\delta_{2k} < \frac{4}{\sqrt{41}}$. Then, taking $\mathcal{S} \subset \{1, \cdots, n\}$ the k-support of $\boldsymbol{x}$, from (Foucart & Rauhut, 2013, Theorem 6.13) and for constants $\rho \in (0, 1)$, $\tau > 0$ we have

$$\|\boldsymbol{x}_{\mathcal{S}}\|_1 \leq \rho \|\boldsymbol{x}_{\mathcal{S}^c}\|_1 + \tau \|\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{y}\|_2, \tag{54}$$

where the notation means that $\boldsymbol{x}_{\mathcal{S}}$ coincides with $\boldsymbol{x}$ on the indices in $\mathcal{S}$ and is extended to zero in its complement $\mathcal{S}^c$. Observe that by adding the term $\rho\|\boldsymbol{x}_{\mathcal{S}}\|_1$ to equation (54) we conclude

$$\|\boldsymbol{x}_{\mathcal{S}}\|_1 \leq \frac{\rho}{1+\rho} \|\boldsymbol{x}\|_1 + \frac{\tau}{1+\rho} \|\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{y}\|_2. \tag{55}$$

Since we assume $g(\boldsymbol{x})$ satisfies Definition 1, from (Gribonval & Nielsen, 2007, Theorem 5) we know that $\frac{g(\boldsymbol{x}_{\mathcal{S}})}{g(\boldsymbol{x})} \leq \frac{\|\boldsymbol{x}_{\mathcal{S}}\|_1}{\|\boldsymbol{x}\|_1}$. Thus, combining this inequality with equation (55) we get

$$\begin{aligned} g(\boldsymbol{x}_{\mathcal{S}}) &\leq \frac{\rho}{1+\rho} g(\boldsymbol{x}) + \frac{\tau'}{1+\rho} \|\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{y}\|_2 \\ g(\boldsymbol{x}_{\mathcal{S}}) &\leq \rho g(\boldsymbol{x}_{\mathcal{S}^c}) + \tau' \|\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{y}\|_2, \end{aligned} \tag{56}$$

where $\tau' = \tau \frac{g(\boldsymbol{x})}{\|\boldsymbol{x}\|_1} < \infty$. Observe that since $g(\boldsymbol{x})$ holds Lemma 2, then from (Woodworth & Chartrand, 2016, Proposition 4.4), and equation (56) we get

$$g(\boldsymbol{x}^* - \boldsymbol{x}) \leq 2\frac{1+\rho}{1-\rho} g(\boldsymbol{x}_{\mathcal{S}^c}) + 4\frac{\tau'}{1-\rho} \|\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{y}\|_2. \tag{57}$$

Now, since $f(\boldsymbol{x})$ is an admissible function, from Lemma 3, and equation (57) we obtain

$$g(\boldsymbol{x}^* - \boldsymbol{x}) \leq 2\frac{1+\rho}{1-\rho} g(\boldsymbol{x}_{\mathcal{S}^c}) + 4\frac{\tau'}{1-\rho} \sqrt{\epsilon} \sqrt{1/c(\boldsymbol{\eta})}, \tag{58}$$

where constant $c(\boldsymbol{\eta})$ as in equation (52). It is important to remark that the above inequality is only relevant for $g(\boldsymbol{x}) = \sum_{i=1}^{n} s_g(|\boldsymbol{x}[i]|)$ quasi-invex when $s_g(\|\boldsymbol{x}\|_\infty) \leq \gamma$ (which is assumed in this proof) for $\gamma = \min\{t \in (0,\infty) | s_g'(w) = 0\}$, otherwise $s_g(w)$ does not reflect the behaviour of $\boldsymbol{x}$. This assumption also applies to $f(\boldsymbol{\eta})$ quasi-invex which implies the constrain $s_f(\|\boldsymbol{\eta}\|_\infty) \leq \gamma$. In addition, for regularizers $g(\boldsymbol{x})$ satisfying Definition 1 we also know from (Gribonval & Nielsen, 2007, Lemma 1) that exists a constant $b(\boldsymbol{x})$ such that $\|\boldsymbol{x}\|_1 \leq \frac{1}{b(\boldsymbol{x})} g(\boldsymbol{x})$. Then, from equation (58) we conclude

$$\|\boldsymbol{x}^* - \boldsymbol{x}\|_1 \leq C\beta_{k,g}(\boldsymbol{x}) + D\sqrt{\epsilon}\upsilon_{f,\eta}(\boldsymbol{x}), \tag{59}$$

where $C = 2\frac{1+\rho}{1-\rho}$, $\beta_{k,g}(\boldsymbol{x}) = \frac{g(\boldsymbol{x}_{\mathcal{S}^c})}{b(\boldsymbol{x})}$, $D = 4\frac{\tau'}{1-\rho}$, and $\upsilon_{f,\eta}(\boldsymbol{x}) = \sqrt{1/c(\boldsymbol{\eta})}/b(\boldsymbol{x})$. Thus the result holds. □

SATISFYING DEFINITION 1

In this section we show that regularizers $g(\boldsymbol{x})$ in equation (3), equation (9), equation (10) satisfy Definition 1.

EQUATION (3)

Proof. Take $r(w) = \log(1 + |w|^p)$ for any $w \neq 0$, and fixed $p \in (0,1)$. It is trivial to see that $r(0) = 0$, that $r(w)$ it is not identically zero, and non-decreasing on $(0,\infty)$. Then, we just need to show that $r(w)/w$ is non-increasing on $(0,\infty)$. Observe that the first derivative of $h(w) = r(w)/w$ is given by $h'(w) = \frac{1}{w^2}\left(\frac{pw^p}{1+w^p} - \log(1 + w^p)\right)$. Since $\frac{pw^p}{1+w^p} - \log(1+w^p) < 0$, then we have that $h'(w) < 0$, which leads to conclude that $r(w)/w$ is non-increasing on $(0,\infty)$. Then it is clear that $r(w)/w$ is non-increasing on $(0,\infty)$. □

EQUATION (9)

Proof. Take

$$r_\theta(w) = \left\{ \begin{array}{ll} \lambda|w| - w^2/(2\theta) & \text{if } |w| < \theta\lambda \\ \theta\lambda^2/2 & \text{if } |w| \geq \theta\lambda \end{array} \right. ,$$

for any $w \in \mathbb{R}$ and fixed $\lambda > 0$, $\theta \geq 1$. It is trivial to see that $r_\theta(0) = 0$, and that $r_\theta(w)$ it is not identically zero. Then, we just need to show that $h(w) = r_\theta(w)/w$ is non-increasing on $(0,\infty)$. Observe that the first derivative of $h(w) = r(w)/w$ is given by

$$h'(w) = \left\{ \begin{array}{ll} -1/(2\theta) & \text{if } |w| < \theta\lambda \\ -\theta\lambda^2/(2w^2) & \text{if } |w| \geq \theta\lambda \end{array} \right. .$$

Since $h'(w) < 0$, then $r(w)/w$ is non-increasing on $(0,\infty)$. Thus it is clear that $r(w)/w$ is non-increasing on $(0,\infty)$. □

H PROOF OF THEOREM 8

Proof. Let $g : \mathbb{R}^n \to \mathbb{R}$ be an admissible function. Then from Lemma 3 we know that $g(\boldsymbol{x}) \geq c\|\boldsymbol{x}\|_2^2$ for some $c > 0$, and therefore $g_{TV}(\boldsymbol{x}) \geq c'\|\boldsymbol{x}\|_2^2$ for some $c' > 0$. Then from (Pallaschke & Rolewicz, 2013, Proposition 5.2.13) we know there exits constant $d' > 0$ such that program (11) has a unique minimizer for any $\lambda \leq \frac{1}{2d'}$. This fact implies that program (11) is invex. It is easy to verify that the TV-regularizer version of mappings in equation (4), equation (9), and equation (10) the constant d' is equal to one.

Now for the second part of this theorem, we have that $h(\boldsymbol{x}) = g_{TV}(\boldsymbol{x}) + \frac{1}{2\lambda}\|\boldsymbol{x} - \boldsymbol{u}\|_2^2$ for some $\boldsymbol{u} \in \mathbb{R}^n$ is an invex function (this is also true for $h(\boldsymbol{x}) = g(\boldsymbol{x}) + \frac{1}{2\lambda}\|\boldsymbol{x} - \boldsymbol{u}\|_2^2$ with $\lambda \leq \frac{1}{2d}$), then Theorem 2 states that any global minimizer $\boldsymbol{y}$ of h satisfies that $\boldsymbol{0} \in \partial h(\boldsymbol{y})$. This condition implies that $\boldsymbol{0} \in \partial g_{TV}(\boldsymbol{y}) + \frac{1}{\lambda}(\boldsymbol{y} - \boldsymbol{u})$, from which we obtain that $\boldsymbol{y} \in (\lambda\partial g_{TV} + \mathbf{I})^{-1}(\boldsymbol{u})$. Thus, we have that $\mathbf{prox}_{g_{TV}}(\boldsymbol{u}) = (\lambda\partial g_{TV} + \mathbf{I})^{-1}(\boldsymbol{u})$ (this is also true for $g(\boldsymbol{x})$) from which the result holds. □

H.1 Remarks on Prox-regular Functions

In this section we prove a prox-regular function is quasi-invex. To that end, we introduce the following definition first.

Definition 2. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a lower semi-continuous function, and $\boldsymbol{u} \in \mathbb{R}^n$. Then f is said to be *prox-regular* if $f(\boldsymbol{x}) + \frac{1}{2\lambda}\|\boldsymbol{x} - \boldsymbol{u}\|_2^2$ is convex for some $\lambda > 0$.

Now we proceed with the proof.

Proof. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a prox-regular function for some $\lambda > 0$. Then we know that for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$

$$f(\boldsymbol{x}) - f(\boldsymbol{y}) \geq \boldsymbol{\zeta}^T(\boldsymbol{x} - \boldsymbol{y}) - \frac{1}{2\lambda}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \tag{60}$$

for all $\boldsymbol{\zeta} \in \partial f(\boldsymbol{y})$. Define function $\eta(\boldsymbol{x}, \boldsymbol{y})$ as

$$\eta(\boldsymbol{x}, \boldsymbol{y}) = \begin{cases} \boldsymbol{0} & \text{if } \boldsymbol{0} \in \partial f(\boldsymbol{y}) \\ \boldsymbol{x} - \boldsymbol{y} - \frac{\|\boldsymbol{x}-\boldsymbol{y}\|_2^2}{2\lambda\|\boldsymbol{\zeta}^*\|_2^2}\boldsymbol{\zeta}^* & \text{otherwise} \end{cases}, \tag{61}$$

where $\boldsymbol{\zeta}^*$ is an element in $\partial f(\boldsymbol{y})$ of minimum norm. Take $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ and assume

$$f(\boldsymbol{x}) - f(\boldsymbol{y}) \leq 0. \tag{62}$$

Observe that if $\boldsymbol{0} \in \partial f(\boldsymbol{y})$ then we get $\boldsymbol{\zeta}^T\eta(\boldsymbol{x}, \boldsymbol{y}) = 0$, for all $\boldsymbol{\zeta} \in \partial f(\boldsymbol{y})$. Additionally, if $\boldsymbol{0} \notin \partial f(\boldsymbol{y})$, then from equation (60) we obtain

$$\begin{aligned} 0 &\geq \boldsymbol{\zeta}^T(\boldsymbol{x} - \boldsymbol{y}) - \frac{1}{2\lambda}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \\ &\geq \boldsymbol{\zeta}^T\left(\boldsymbol{x} - \boldsymbol{y} - \frac{\|\boldsymbol{x} - \boldsymbol{y}\|_2^2}{2\lambda\|\boldsymbol{\zeta}^*\|_2^2}\boldsymbol{\zeta}^*\right) = \boldsymbol{\zeta}^T\eta(\boldsymbol{x}, \boldsymbol{y}), \end{aligned} \tag{63}$$

where the second inequality comes from the fact that $\boldsymbol{\zeta}^*$ is an element in $\partial f(\boldsymbol{y})$ of minimum norm i.e. $\frac{\boldsymbol{\zeta}^T\boldsymbol{\zeta}^*}{\|\boldsymbol{\zeta}^*\|_2^2} \geq 1$ for all $\boldsymbol{\zeta} \in \partial f(\boldsymbol{y})$ (Bazaraa & Shetty, 2012, Theorem 2.4.4). From the above inequality the result holds. □

I Proof of Theorem 9

Here we prove Theorem 9 following a similar strategy as presented in (Parikh & Boyd, 2014).

Proof. Since $(\boldsymbol{x}^*, \boldsymbol{z}^*, \boldsymbol{v}^*)$ is a saddle point for $\mathcal{L}_0$, we have

$$\mathcal{L}_0(\boldsymbol{x}^*, \boldsymbol{z}^*, \boldsymbol{v}^*) \leq \mathcal{L}_0(\boldsymbol{x}^{(t+1)}, \boldsymbol{z}^{(t+1)}, \boldsymbol{v}^*). \tag{64}$$

Using $\boldsymbol{S}\boldsymbol{x}^* + \boldsymbol{P}\boldsymbol{z}^* = \boldsymbol{y}$ the left hand side is $h^* = \inf\{h_1(\boldsymbol{x}) + h_2(\boldsymbol{z}) \mid \boldsymbol{S}\boldsymbol{x} + \boldsymbol{P}\boldsymbol{z} = \boldsymbol{y}\}$. With $h^{(t+1)} = h_1(\boldsymbol{x}^{(t+1)}) + h_2(\boldsymbol{z}^{(t+1)})$, this can be written as

$$h^* \leq h^{(t+1)} + (\boldsymbol{v}^*)^T\boldsymbol{q}^{(t+1)}, \tag{65}$$

for $\boldsymbol{q}^{(t+1)} = \boldsymbol{S}\boldsymbol{x}^{(t+1)} + \boldsymbol{P}\boldsymbol{z}^{(t+1)} - \boldsymbol{y}$. Now, by definition, $\boldsymbol{x}^{(t+1)}$ minimizes $\mathcal{L}_\rho(\boldsymbol{x}, \boldsymbol{z}^{(t)}, \boldsymbol{v}^{(t)})$. From Appendix H and the fact that $\rho\sigma_n(\boldsymbol{S}) \geq 1$, $\rho\sigma_p(\boldsymbol{P}) \geq 1$, we know that the (necessary and sufficient) optimality condition for $\mathcal{L}_\rho(\boldsymbol{x}, \boldsymbol{z}^{(t)}, \boldsymbol{v}^{(t)})$ is given by

$$\boldsymbol{0} \in \partial\mathcal{L}_\rho(\boldsymbol{x}^{(t+1)}, \boldsymbol{z}^{(t)}, \boldsymbol{z}^{(t)}) = \partial h_1(\boldsymbol{x}^{(t+1)}) + \boldsymbol{S}^T\boldsymbol{v}^{(t)} + \rho\boldsymbol{S}^T(\boldsymbol{S}\boldsymbol{x}^{(t+1)} + \boldsymbol{P}\boldsymbol{z}^{(t)} - \boldsymbol{y}). \tag{66}$$

Since $\boldsymbol{v}^{(t+1)} = \boldsymbol{v}^{(t)} + \rho\boldsymbol{q}^{(t+1)}$, we can plug in $\boldsymbol{v}^{(t)} = \boldsymbol{v}^{(t+1)} - \rho\boldsymbol{q}^{(t+1)}$ and rearrange to obtain

$$\boldsymbol{0} \in \partial h_1(\boldsymbol{x}^{(t+1)}) + \boldsymbol{S}^T(\boldsymbol{v}^{(t+1)} - \rho\boldsymbol{P}(\boldsymbol{z}^{(t+1)} - \boldsymbol{z}^{(t)})). \tag{67}$$

From Appendix H, this implies that $\boldsymbol{x}^{(t+1)}$ uniquely minimizes

$$h_1(\boldsymbol{x}) + (\boldsymbol{v}^{(t+1)} - \rho\boldsymbol{P}(\boldsymbol{z}^{(t+1)} - \boldsymbol{z}^{(t)}))^T\boldsymbol{S}\boldsymbol{x}. \tag{68}$$

A similar argument shows that $\boldsymbol{z}^{(t+1)}$ uniquely minimizes $h_2(\boldsymbol{z}) + (\boldsymbol{v}^{(t+1)})^T \boldsymbol{P}\boldsymbol{z}$. It follows that

$$h_1(\boldsymbol{x}^{(t+1)}) + (\boldsymbol{v}^{(t+1)} - \rho\boldsymbol{P}(\boldsymbol{z}^{(t+1)} - \boldsymbol{z}^{(t)}))^T \boldsymbol{S}\boldsymbol{x}^{(t+1)} \\ \leq h_1(\boldsymbol{x}^*) + (\boldsymbol{v}^{(t+1)} - \rho\boldsymbol{P}(\boldsymbol{z}^{(t+1)} - \boldsymbol{z}^{(t)}))^T \boldsymbol{S}\boldsymbol{x}^*, \tag{69}$$

and that

$$h_2(\boldsymbol{z}^{(t+1)}) + (\boldsymbol{v}^{(t+1)})^T \boldsymbol{P}\boldsymbol{z}^{(t+1)} \leq h_2(\boldsymbol{z}^*) + (\boldsymbol{v}^{(t+1)})^T \boldsymbol{P}\boldsymbol{z}^*. \tag{70}$$

Adding the two inequalities above, using $\boldsymbol{S}\boldsymbol{x}^* + \boldsymbol{P}\boldsymbol{z}^* = \boldsymbol{y}$, and rearranging, we obtain

$$h^{(t+1)} - h^* \leq -(\boldsymbol{v}^{(t+1)})^T \boldsymbol{q}^{(t+1)} - \rho(\boldsymbol{P}(\boldsymbol{z}^{(t+1)} - \boldsymbol{z}^{(t)}))^T(-\boldsymbol{q}^{(t+1)} + \boldsymbol{P}(\boldsymbol{z}^{(t+1)} - \boldsymbol{z}^*)). \tag{71}$$

On the other hand, adding equation (65), and equation (71), regrouping terms, and multiplying through by 2 gives

$$2(\boldsymbol{v}^{(t+1)} - \boldsymbol{v}^*)^T \boldsymbol{q}^{(t+1)} - 2\rho(\boldsymbol{P}(\boldsymbol{z}^{(t+1)} - \boldsymbol{z}^{(t)}))^T \boldsymbol{q}^{(t+1)} \\ + 2\rho(\boldsymbol{P}(\boldsymbol{z}^{(t+1)} - \boldsymbol{z}^{(t)}))^T(\boldsymbol{P}(\boldsymbol{z}^{(t+1)} - \boldsymbol{z}^*)) \leq 0. \tag{72}$$

Now by rewriting the first term in equation (72), and substituting $\boldsymbol{v}^{(t+1)} = \boldsymbol{v}^{(t)} + \rho\boldsymbol{q}^{(t+1)}$ it gives

$$2(\boldsymbol{v}^{(t+1)} - \boldsymbol{v}^*)^T \boldsymbol{q}^{(t+1)} + \rho\|\boldsymbol{q}^{(t+1)}\|_2^2 + \rho\|\boldsymbol{q}^{(t+1)}\|_2^2, \tag{73}$$

and substituting $\boldsymbol{q}^{(t+1)} = (1/\rho)(\boldsymbol{v}^{(t+1)} - \boldsymbol{v}^{(t)})$ in the first two terms gives

$$(2/\rho)(\boldsymbol{v}^{(t)} - \boldsymbol{v}^*)^T(\boldsymbol{v}^{(t+1)} - \boldsymbol{v}^{(t)}) + (1/\rho)\|\boldsymbol{v}^{(t+1)} - \boldsymbol{v}^{(t)}\|_2^2 + \rho\|\boldsymbol{q}^{(t+1)}\|_2^2. \tag{74}$$

Since $\boldsymbol{q}^{(t+1)} - \boldsymbol{q}^{(t)} = (\boldsymbol{q}^{(t+1)} - \boldsymbol{q}^*) - (\boldsymbol{q}^{(t)} - \boldsymbol{q}^*)$, this can be written as

$$(1/\rho)(\|\boldsymbol{v}^{(t+1)} - \boldsymbol{v}^*\|_2^2 - \|\boldsymbol{v}^{(t)} - \boldsymbol{v}^*\|_2^2) + \rho\|\boldsymbol{q}^{(t+1)}\|_2^2. \tag{75}$$

We now rewrite the remaining terms

$$\rho\|\boldsymbol{q}^{(t+1)}\|_2^2 - 2\rho(\boldsymbol{P}(\boldsymbol{z}^{(t+1)} - \boldsymbol{z}^{(t)}))^T \boldsymbol{q}^{(t+1)} + 2\rho(\boldsymbol{P}(\boldsymbol{z}^{(t+1)} - \boldsymbol{z}^{(t)}))^T(\boldsymbol{P}(\boldsymbol{z}^{(t+1)} - \boldsymbol{z}^*)), \tag{76}$$

where $\rho\|\boldsymbol{q}^{(t+1)}\|_2^2$ is taken from equation (75). Substituting

$$\boldsymbol{z}^{(t+1)} - \boldsymbol{z}^* = (\boldsymbol{z}^{(t+1)} - \boldsymbol{z}^{(t)}) + (\boldsymbol{z}^{(t)} - \boldsymbol{z}^*), \tag{77}$$

in the last term gives

$$\rho\|\boldsymbol{q}^{(t+1)} - \boldsymbol{P}(\boldsymbol{z}^{(t+1)} - \boldsymbol{z}^{(t)})\|_2^2 + \rho\|\boldsymbol{P}(\boldsymbol{z}^{(t+1)} - \boldsymbol{z}^{(t)})\|_2^2 \\ + 2\rho(\boldsymbol{P}(\boldsymbol{z}^{(t+1)} - \boldsymbol{z}^{(t)}))^T(\boldsymbol{P}(\boldsymbol{z}^{(t+1)} - \boldsymbol{z}^*)), \tag{78}$$

and substituting

$$\boldsymbol{z}^{(t+1)} - \boldsymbol{z}^{(t)} = (\boldsymbol{z}^{(t+1)} - \boldsymbol{z}^*) - (\boldsymbol{z}^{(t)} - \boldsymbol{z}^*), \tag{79}$$

in the last two terms, we get

$$\rho\|\boldsymbol{q}^{(t+1)} - \boldsymbol{P}(\boldsymbol{z}^{(t+1)} - \boldsymbol{z}^{(t)})\|_2^2 + \rho\left(\|\boldsymbol{P}(\boldsymbol{z}^{(t+1)} - \boldsymbol{z}^*)\|_2^2 - \|\boldsymbol{P}(\boldsymbol{z}^{(t)} - \boldsymbol{z}^*)\|_2^2\right). \tag{80}$$

With the previous step, this implies that equation (72) can be written as

$$V^{(t)} - V^{(t+1)} \geq \rho\|\boldsymbol{q}^{(t+1)} - \boldsymbol{P}(\boldsymbol{z}^{(t+1)} - \boldsymbol{z}^{(t)})\|_2^2, \tag{81}$$

where $V^{(t)} = (1/\rho)\|\boldsymbol{v}^{(t)} - \boldsymbol{v}^*\|_2^2 + \rho\|\boldsymbol{P}(\boldsymbol{z}^{(t)} - \boldsymbol{z}^*)\|_2^2$.

Now, we show that the middle term $-2\rho(\boldsymbol{q}^{(t+1)})^T(\boldsymbol{P}(\boldsymbol{z}^{(t+1)} - \boldsymbol{z}^{(t)}))$ of the expanded right hand side of equation (81) is positive. To see this, recall that $\boldsymbol{z}^{(t+1)}$ minimizes $h_2(\boldsymbol{z}) + (\boldsymbol{v}^{(t+1)})^T \boldsymbol{P}\boldsymbol{z}$, and $\boldsymbol{z}^{(t)}$ minimizes $h_2(\boldsymbol{z}) + (\boldsymbol{v}^{(t)})^T \boldsymbol{P}\boldsymbol{z}$, so we can add

$$h_2(\boldsymbol{z}^{(t+1)}) + (\boldsymbol{v}^{(t+1)})^T \boldsymbol{P}\boldsymbol{z}^{(t+1)} \leq h_2(\boldsymbol{z}^{(t)}) + (\boldsymbol{v}^{(t+1)})^T \boldsymbol{P}\boldsymbol{z}^{(t)}, \tag{82}$$

and

$$h_2(\boldsymbol{z}^{(t)}) + (\boldsymbol{v}^{(t)})^T \boldsymbol{P}\boldsymbol{z}^{(t)} \leq h_2(\boldsymbol{z}^{(t+1)}) + (\boldsymbol{v}^{(t)})^T \boldsymbol{P}\boldsymbol{z}^{(t+1)}, \tag{83}$$

to get that

$$(\boldsymbol{v}^{(t+1)} - \boldsymbol{v}^{(t)})^T \boldsymbol{P}(\boldsymbol{z}^{(t+1)} - \boldsymbol{z}^{(t)}) \leq 0. \tag{84}$$

Substituting $\boldsymbol{v}^{(t+1)} - \boldsymbol{v}^{(t)} = \rho \boldsymbol{q}^{(t+1)}$ gives the result, since $\rho > 0$. Thus, from equation (81), and equation (84) we obtain

$$V^{(t+1)} \leq V^{(t)} - \rho\|\boldsymbol{q}^{(t+1)}\|_2^2 - \rho\|\boldsymbol{P}(\boldsymbol{z}^{(t+1)} - \boldsymbol{z}^{(t)})\|_2^2, \tag{85}$$

which states that $V^{(t)}$ decreases in each iteration by an amount that depends on the norm of the residual $\boldsymbol{q}^{(t)}$ and on the change in $\boldsymbol{z}^{(t)}$ over one iteration. Then, because $V^{(t)} \leq V^{(0)}$, it follows that $\boldsymbol{v}^{(t)}$ and $\boldsymbol{P}\boldsymbol{z}^{(t)}$ are bounded. Iterating the inequality above gives that

$$\rho \sum_{t=0}^{\infty} \Big(\|\boldsymbol{q}^{(t+1)}\|_2^2 + \|\boldsymbol{P}(\boldsymbol{z}^{(t+1)} - \boldsymbol{z}^{(t)})\|_2^2 \Big) \leq V^{(0)}, \tag{86}$$

which implies that $\boldsymbol{q}^{(t)} = \boldsymbol{S}\boldsymbol{x}^{(t)} + \boldsymbol{P}\boldsymbol{z}^{(t)} - \boldsymbol{y} \to 0$, and $\boldsymbol{P}(\boldsymbol{z}^{(t+1)} - \boldsymbol{z}^{(t)}) \to 0$ as $t \to \infty$. Additionally, applying (Deng et al., 2017, Lemma 1.2) on equation (86) we obtain a convergence rate for $\boldsymbol{q}^{(t)}$, $\boldsymbol{P}(\boldsymbol{z}^{(t+1)} - \boldsymbol{z}^{(t)})$ to zero of $\mathcal{O}(1/t)$. equation (86) also implies that the right hand side in equation (71) goes to zero as $t \to \infty$, because $\boldsymbol{P}(\boldsymbol{z}^{(t+1)} - \boldsymbol{z}^*)$ is bounded and both $\boldsymbol{q}^{(t+1)}$ and $\boldsymbol{P}(\boldsymbol{z}^{(t+1)} - \boldsymbol{z}^{(t)})$ go to zero. The right hand side in equation (65) goes to zero as $t \to \infty$, since $\boldsymbol{q}^{(t)}$ goes to zero. Thus we have $\lim_{t\to\infty} h^{(t)} = h^*$, i.e., objective convergence. Therefore the result of Theorem 9 holds. □

I.1 Remarks on Stability of ADMM

In this section we wish to emphasize that the stability and reliability (of the ADMM) can be established from the following: 1) ADMM decomposes the overall optimization problem into a number of simpler subproblems that have a unique solution (as rigorously demonstrated in previous section), aiding convergence and stability. 2) As shown in previous section, the sequences $\boldsymbol{x}^{(t+1)}$, $\boldsymbol{z}^{(t+1)}$, and $\boldsymbol{v}^{(t+1)}$, constructed by ADMM algorithm, always converge to global optima irrespective of the initial states $\boldsymbol{x}^{(0)}$, $\boldsymbol{z}^{(0)}$, and $\boldsymbol{v}^{(0)}$, conferring a steadfast assurance of reliable attainment of optimal solutions, and finally, 3) The effectiveness of the ADMM is well demonstrated in the literature using a range of real-world applications (Boyd et al., 2011; Wang et al., 2019b; Glowinski, 2014; Xie et al., 2019), which we believe can reaffirm the reliability (and potentially the stability) of ADMM.

Remarks on optimization problem in 1

In this section we show how program 1 is expressed as equation (12), where the non-linear constrain is in the form of $f(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y})$. From (Chen et al., 2001; Cai & Wang, 2011), we know that

$$\text{minimize } g(\boldsymbol{x}) \text{ subject to } f(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}) < \epsilon,$$

can be equivalently rewritten as

$$\text{minimize } g(\boldsymbol{x}) + \lambda f(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}),$$

for some $\lambda > 0$ that helps to satisfy the condition $f(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}) < \epsilon$. Thus, if introducing a variable $\boldsymbol{z} = \boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}$, program (1) can be finally rewritten as

$$\text{minimize } g(\boldsymbol{x}) + \lambda f(\boldsymbol{z}) \text{ subject to } \boldsymbol{z} = \boldsymbol{A}\boldsymbol{x} - \boldsymbol{y},$$

which is in the optimization form of ADMM as pointed out in equation (12). In this way, optimization problem in (1) is converted from non-linear constraints into linear ones.

I.2 Additional Discussion on Theorem 9

In this section, we discuss why the assumptions $\rho\sigma_n(\boldsymbol{S}) \geq 1$, and $\rho\sigma_p(\boldsymbol{P}) \geq 1$ for $\rho > 0$ in Theorem 9 are mild, by showing practical imaging examples that satisfy these conditions. We list those applications in the following.

1. **Image Fusion:** Image fusion has been receiving increasing attention in the research community to investigate general formal solutions to a broad spectrum of applications (Vargas et al., 2019). For example, in the remote sensing field, the increasing availability of spaceborne imaging sensors operating in various ground scales and spectral bands undoubtedly provides strong motivations (Camacho et al., 2022). Because of the trade-off imposed by the physical constraint between spatial and spectral resolutions, spatial enhancement of poor-resolution multispectral (MS) data is desirable (Camacho et al., 2022). Another example is medical imaging, in which the goal of image fusion is to create new images more suitable for human visual perception (Meyer-Bäse et al., 2004).

Mathematically the image fusion problem in spectral imaging can be modeled as in equation (12) defining matrices $\boldsymbol{S}$, and $\boldsymbol{P}$ as (Vargas et al., 2019)

$$\boldsymbol{S} = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{I} \\ \boldsymbol{0} \\ \boldsymbol{B}_s \\ \boldsymbol{G} \\ \boldsymbol{D} \\ \boldsymbol{I} \end{bmatrix}, \boldsymbol{P} = \begin{bmatrix} \boldsymbol{I} & -\boldsymbol{R}_\lambda\overline{\boldsymbol{M}} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I} & -\overline{\boldsymbol{M}} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I} \end{bmatrix} = \boldsymbol{I} + \boldsymbol{H}. \tag{87}$$

Observe that $\sigma_n(\boldsymbol{S}) = 2 + \delta > 1$ for $\delta > 0$ where is the smallest singular value of $\boldsymbol{B}_s^T\boldsymbol{B}_s + \boldsymbol{G}^T\boldsymbol{G} + \boldsymbol{D}^T\boldsymbol{D}$. In addition, $\sigma_n(\boldsymbol{P}) = 1 + \zeta > 0$ where ζ is the smallest singular value of $\boldsymbol{H} + \boldsymbol{H}^T + \boldsymbol{H}^T\boldsymbol{H}$. Thus, taking $\rho \geq \max\{0.5, 1/(1+\zeta)\}$ the needed conditions in Theorem 9 are satisfied.

2. **Spectral Imaging:** Spectral imaging combines two disciplines - spectroscopy and photography - to sample image data at many wavelength bands. In general, spectral imaging is separated into either multispectral (¡ 20 wavelength bands sampled) or hyperspectral (¿ 20 wavelength bands) (Vargas et al., 2018). Mathematically the spectral imaging problem can be modeled as in equation (12) defining matrices $\boldsymbol{S}$, and $\boldsymbol{P}$ are given by (Vargas et al., 2018)

$$\boldsymbol{S} = \left[\underbrace{\boldsymbol{I}, \ldots, \boldsymbol{I}}_{K}, \boldsymbol{\Psi}, \boldsymbol{L}^T\right]^T, \boldsymbol{P} = -\boldsymbol{I}, \tag{88}$$

where $\boldsymbol{\Psi}$ is an orthogonal matrix. Observe that $\sigma_n(\boldsymbol{S}) = K + 1 + \delta > 1$ for $\delta > 0$ where is the smallest singular value of $\boldsymbol{L}^T\boldsymbol{L}$. In addition, $\sigma_n(\boldsymbol{P}) = 1$. Thus, taking $\rho \geq 1$ the needed conditions in Theorem 9 are satisfied.

3. **Computer Tomography:** Computer tomography refers to a computerized X-ray imaging procedure in which a narrow beam of X-rays is aimed at a patient and quickly rotates around the body, producing signals processed by the machine's computer to generate cross-sectional images (Wang et al., 2019a). These slices are called tomographic images and can give a clinician more detailed information than conventional X-rays. Mathematically, the computer tomography problem can be modeled as in equation (12) defining matrices $\boldsymbol{S}$, and $\boldsymbol{P}$ are given by $\boldsymbol{S} = \boldsymbol{I}$, and $\boldsymbol{P} = -\boldsymbol{I}$ (Vargas et al., 2018). Thus, taking $\rho \geq 1$ the needed conditions in Theorem 9 are satisfied.

4. **Magnetic Resonance Imaging:** Magnetic Resonance Imaging (MRI) is a non-invasive imaging technique providing both functional and anatomical information for clinical diagnosis. Imaging speed is a fundamental challenge. Fast MRI techniques are essentially demanded to accelerate data acquisition while still reconstructing a high-quality image (Sun et al., 2016). Mathematically, the magnetic resonance imaging problem can be modeled as in equation (12) defining matrices $\boldsymbol{S}$, and $\boldsymbol{P}$ are given by $\boldsymbol{S} = \boldsymbol{I}$, and $\boldsymbol{P} = -\boldsymbol{\Psi}$ (Vargas et al., 2018), for $\boldsymbol{\Psi}$ an orthogonal matrix. Thus, taking $\rho \geq 1$ the needed conditions in Theorem 9 are satisfied.

5. **Compressive Sensing:** Compressive sensing is a recent highly applicative approach. It enables efficient data sampling at a much lower rate than the requirements indicated by the Nyquist theorem. Compressive sensing possesses several advantages, such as the much smaller need for sensory

devices, much less memory storage, higher data transmission rate, many times less power consumption (Ramirez et al., 2021). Due to all these advantages, compressive sensing has been used in a wide range of applications. Mathematically, the compressive sensing problem can be modeled as in equation (12) defining matrices $\boldsymbol{S}$, and $\boldsymbol{P}$ are given by $\boldsymbol{S} = \boldsymbol{\Psi}$, and $\boldsymbol{P} = -\boldsymbol{I}$ (Ramirez et al., 2021), for $\boldsymbol{\Psi}$ an orthogonal matrix. Thus, taking $\rho \geq 1$ the needed conditions in Theorem 9 are satisfied.

6. **Stepped-Frequency Radar:** Step-frequency is a radar waveform consisting of a series of sine waves with linearly increasing frequency. The radar measures the phase and amplitude on each frequency and used an inverse Fourier transform of these data to build a time domain profile (Johnston et al., 2021). Mathematically, the Stepped-Frequency Radar problem can be modeled as in equation (12) defining matrices $\boldsymbol{S}$, and $\boldsymbol{P}$ are given by $\boldsymbol{S} = \boldsymbol{I}$, and $\boldsymbol{P} = -\boldsymbol{I}$ (Johnston et al., 2021). Thus, taking $\rho \geq 1$ the needed conditions in Theorem 9 are satisfied.

I.3 Proximal Solutions

In this section we present the proximal operator for the studied invex/quasi-invex functions in this work, summarized in the following Table 4.

Table 4: Proximal operator for regularizers ℓ_p-quasinorm, and equation (3), equation (9) ($\lambda \in (0, 1]$ is a thresholding parameter).

Ref	Invex function	Proximal operator
(Marjanovic & Solo, 2012)	$g_\lambda(w) = \lambda\lvert w\rvert^p$, $p \in (0, 1)$, $w \neq 0$.	$\text{Prox}_{g_\lambda}(w) = \begin{cases} 0 & \lvert t\rvert < \tau \\ \{0, \text{sign}(w)\beta\} & \lvert t\rvert = \tau \\ \text{sign}(w)y & \lvert t\rvert > \tau \end{cases}$ where $\beta = [2\lambda(1-p)]^{1/(2-p)}$, $\tau = \beta + \lambda p \beta^{p-1}$, $h(y) = \lambda p y^{p-1} + y - \lvert t\rvert = 0$, $y \in [\beta, \lvert t\rvert]$
-	$g_\lambda(w) = \lambda\log(1 + \lvert w\rvert^p), p \in (0, 1), x \neq 0$	$\text{Prox}_{g_\lambda}(w) = \begin{cases} 0 & \lvert w\rvert = 0 \\ \text{sign}(w)\beta & \text{otherwise} \end{cases}$ where $\beta(\beta - \lvert w\rvert)(\beta^{-p} + 1) + \lambda p = 0$, $\beta > 0$. We solve this equation using standard bisection method.
(Zhang, 2010)	MCP	$\text{Prox}_{\lambda g_\theta}(w) = \begin{cases} \frac{\theta w - \text{sign}(w)\theta\lambda}{\theta - 1} & \lvert w\rvert < \theta\lambda \\ w & \lvert w\rvert \geq \theta\lambda \end{cases}$

Proximal of function in equation (3)

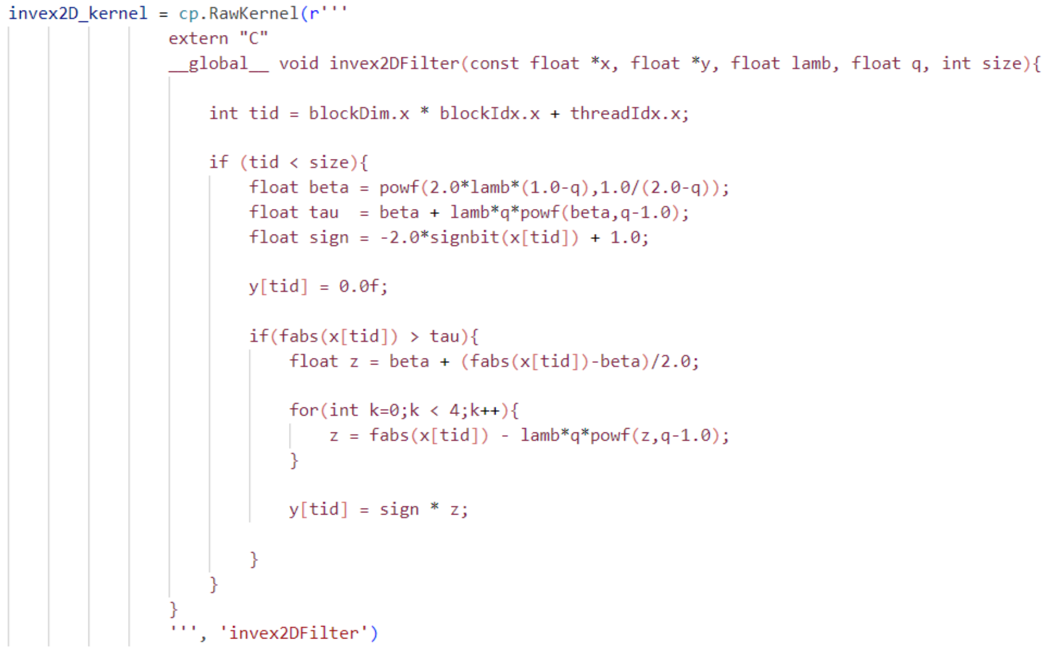
Consider $h(w) = \lambda \log(1 + |w|^p) + \frac{1}{2}(w - u)^2$ for $\lambda \in (0, 1]$, $w \neq 0$, and fixed $u \in \mathbb{R}$, $p \in (0, 1)$. We note first that we only consider $w's$ for which $\text{sign}(w) = \text{sign}(u)$, otherwise $h(w) = \lambda \log(1 + |w|^p) + \frac{1}{2}w^2 + |u||w| + \frac{1}{2}u^2$ which is clearly minimized at $w = 0$. Then, since with $\text{sign}(w) = \text{sign}(u)$ we have $(w-u)^2 = (|w|-|u|)^2$, we replace u with $|u|$ and take $w \geq 0$. As $h(w)$ is differentiable for $w > 0$, re-arranging $h'(w) = 0$ gives $\psi_\lambda(w) \triangleq \frac{\lambda p w^{p-1}}{1+w^p} + w = |u|$. Observe that $\psi'_\lambda(w)$ is always positive then it means that $\psi_\lambda(w)$ is monotonically increasing. Thus, the equation $\psi_\lambda(w) = |u|$ has unique solution i.e. at some point the quality holds. Thus, solving $\psi_\lambda(w) = |u|$ is equivalent to

$$w(w - |u|)(w^{-p} + 1) + \lambda p = 0. \tag{89}$$

We solve equation (89) using standard bisection method (Burden & Faires, 1985).

Implementation details: In order to efficiently compute the proximal solutions, which is an entry-wise operation, in Table 4, we use the Python library CuPy (Lib). The reason is that this library allows the creation of GPU kernels in the language C++ from Python (quick guide on how to create and use these kernels (ker)). Therefore, inside this GPU kernel, we implemented the entry-wise operation from the Proximal Operator column of Table 4. We report as an example how to implement the proximal solution for ℓ_p-quasinorm. We remark that this implementation is efficient because it runs in parallel and at the GPU speed. Additional features of the CuPy library are the flexibility to connect both TensorFlow and PyTorch libraries for training neural networks (pyt).

Lastly, the running time to compute the proximal of invex/quasi-invex functions is also essential to compare with its convex competitor, i.e., ℓ_1-norm. The reason is the desire to improve imaging quality, keeping the same computational complexity to obtain it. Therefore, the following Table 5 reports the running time to compute the proximal (in GPU) of equation (3), ℓ_p-quasinorm, and equation (9) for an image of 2048×2048 pixels. Table 5 suggests that computing the proximal of the ℓ_1-norm



```
invex2D_kernel = cp.RawKernel(r'''
                extern "C"
                __global__ void invex2DFilter(const float *x, float *y, float lamb, float q, int size){

                    int tid = blockDim.x * blockIdx.x + threadIdx.x;

                    if (tid < size){
                        float beta = powf(2.0*lamb*(1.0-q),1.0/(2.0-q));
                        float tau  = beta + lamb*q*powf(beta,q-1.0);
                        float sign = -2.0*signbit(x[tid]) + 1.0;

                        y[tid] = 0.0f;

                        if(fabs(x[tid]) > tau){
                            float z = beta + (fabs(x[tid])-beta)/2.0;

                            for(int k=0;k < 4;k++){
                                z = fabs(x[tid]) - lamb*q*powf(z,q-1.0);
                            }

                            y[tid] = sign * z;

                        }
                    }
                }
                ''', 'invex2DFilter')
```

is faster than the proximal of invex regularizers. However, this difference is given in milliseconds, making it negligible in practice.

Table 5: Time to compute the proximal for all invex and convex regularizers, of an image with 2048×2048 pixels. The reported time is the averaged over 256 trials. For ℓ_p-quasinorm we select $p = 0.5$, and $\epsilon = (p(1-p))^{\frac{1}{2-p}}$.

	equation (3)	ℓ_p-quasinorm	equation (9)	ℓ_1-norm
Time	$2.8ms$	$1.06ms$	$0.86ms$	$0.66ms$

J Extended Accelerated Proximal Gradient Method (APGM)

The accelerated proximal gradient method (APGM) (Li & Lin, 2015) has been shown to be effective in solving program (1) by minimizing $F(\boldsymbol{x}) = g(\boldsymbol{x}) + \hat{f}(\boldsymbol{x})$ where $\hat{f}(\boldsymbol{x}) = \frac{1}{2}\left(\max\{f(\boldsymbol{x}) - \epsilon, 0\}\right)^2$, achieving better quality in less iterations than its predecessors (Beck & Teboulle, 2009; Frankel et al., 2015; Boţ et al., 2016), and been frequently used by recent imaging works (Wang & Chen, 2022; Mai et al., 2022). Note that $\hat{f}$ is invex with the same η as f (which is also the same of g according to Theorem 3), since $h(w) = (\max(w - \epsilon, 0))^2$ is an increasing convex function (Mishra & Giorgi, 2008, Theorem 2.14). Therefore, $F(\boldsymbol{x})$ is invex. The iterative process performed by APGM is summarized as

$$\begin{aligned}
\boldsymbol{y}^{(t)} &= \boldsymbol{x}^{(t)} + \frac{r_{t-1}}{r_t}(\boldsymbol{z}^{(t)} - \boldsymbol{x}^{(t)}) + \frac{r_{t-1} - 1}{r_t}(\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t-1)}) \\
\boldsymbol{z}^{(t+1)} &= \text{prox}_{\alpha_2 g}(\boldsymbol{y}^{(t)} - \alpha_2 \nabla \hat{f}(\boldsymbol{y}^{(t)})) \\
\boldsymbol{v}^{(t+1)} &= \text{prox}_{\alpha_1 g}(\boldsymbol{x}^{(t)} - \alpha_1 \nabla \hat{f}(\boldsymbol{x}^{(t)})), r_{t+1} = \frac{\sqrt{4(r_t)^2 + 1} + 1}{2} \\
\boldsymbol{x}^{(t+1)} &= \begin{cases} \boldsymbol{z}^{(t+1)}, & \text{if } \hat{f}(\boldsymbol{z}^{(t+1)}) + g(\boldsymbol{z}^{(t+1)}) \leq \hat{f}(\boldsymbol{v}^{(t+1)}) + g(\boldsymbol{v}^{(t+1)}) \\ \boldsymbol{v}^{(t+1)}, & \text{otherwise} \end{cases},
\end{aligned} \tag{90}$$

for some positive constants α_1, α_2, and assuming $\hat{f}$ is L-smooth. The reported convergence guarantees of $\boldsymbol{x}^{(t+1)}$ to global optima of APGM has been only stated for convex losses (Li & Lin, 2015). For non-convex cases, convergence to a critical point has been stated (Li & Lin, 2015). Here, we extend the general convergence guarantees of APGM to invex/quasi-invex settings in Theorem 3 so that its benefits are available to the signal restoration problems.

Theorem 2. Let $f, g : \mathbb{R}^n \to \mathbb{R}$ be admissible functions as in Definition 4, where f is L-smooth. Assume $\boldsymbol{x}^\star$ a global minimizer of $F(\boldsymbol{x}) = g(\boldsymbol{x}) + \hat{f}(\boldsymbol{x})$ for $\epsilon = 0$, then

$$F(\boldsymbol{x}^{(T+1)}) - F(\boldsymbol{x}^*) \leq \frac{2}{\alpha'(T+1)^2}\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\|_2^2, \tag{91}$$

where T is number of iterations, $\alpha_1, \alpha_2 < \frac{\kappa}{L\kappa+4}$, $\frac{1}{\alpha'} = \left(\frac{1}{\alpha_1} - \frac{2}{\kappa}\right) > 0$, with κ a constant that depends on f, g.

Proof of Theorem 2 relies on the properties of admissible functions including Theorem 8. The proof is presented below. It is worth mentioning that the convergence rate for admissible functions as in Theorem 2 is the same as in the convex case (see (Li & Lin, 2015)).

Proof. Step 2 in the APGM is given by

$$\boldsymbol{z}^{(t+1)} = \operatorname*{arg\,min}_{\boldsymbol{x}\in\mathbb{R}^n} \left\langle \nabla f(\boldsymbol{y}^{(t)}), \boldsymbol{x} - \boldsymbol{y}^{(t)} \right\rangle + \frac{1}{2\alpha_1}\|\boldsymbol{x} - \boldsymbol{y}^{(t)}\|_2^2 + g(\boldsymbol{x}). \tag{92}$$

From Theorem 8 we know that equation (92) is unique for a $\alpha_1 < \frac{1}{\kappa_g}$ for some $\kappa_g > 0$. Then the optimality condition for equation (92) is given by

$$\mathbf{0} \in \nabla f(\boldsymbol{y}^{(t)}) + \frac{1}{\alpha_1}(\boldsymbol{z}^{(t+1)} - \boldsymbol{y}^{(t)}) + \partial g(\boldsymbol{z}^{(t+1)}). \tag{93}$$

Additionally, since $g(\boldsymbol{x})$ is an admissible function, then from (Pallaschke & Rolewicz, 2013, Proposition 5.2.14) we have that

$$g(\boldsymbol{x}) - g(\boldsymbol{z}^{(t+1)}) \geq \left\langle -\nabla f(\boldsymbol{y}^{(t)}) - \frac{1}{\alpha_1}(\boldsymbol{z}^{(t+1)} - \boldsymbol{y}^{(t)}), \boldsymbol{x} - \boldsymbol{z}^{(t+1)} \right\rangle - \frac{1}{\kappa_g}\|\boldsymbol{x} - \boldsymbol{z}^{(t+1)}\|_2^2. \tag{94}$$

for all $\boldsymbol{x}$. From the Lipschitz continuous of ∇f, we have for all $\boldsymbol{x}$ and $\kappa_f < \frac{1}{L}$ that

$$f(\boldsymbol{x}) - f(\boldsymbol{y}^{(t)}) \geq \left\langle \nabla f(\boldsymbol{y}^{(t)}), \boldsymbol{x} - \boldsymbol{y}^{(t)} \right\rangle - \frac{1}{\kappa_f}\|\boldsymbol{x} - \boldsymbol{y}^{(t)}\|_2^2. \tag{95}$$

In addition of the Lipschitz continuity of ∇f we have that

$$\begin{aligned} F(\boldsymbol{z}^{(t+1)}) &\leq g(\boldsymbol{z}^{(t+1)}) + f(\boldsymbol{y}^{(t)}) + \left\langle \nabla f(\boldsymbol{y}^{(t)}), \boldsymbol{z}^{(t+1)} - \boldsymbol{y}^{(t)} \right\rangle + \frac{L}{2}\|\boldsymbol{z}^{(t+1)} - \boldsymbol{y}^{(t)}\|_2^2 \\ &= g(\boldsymbol{z}^{(t+1)}) + f(\boldsymbol{y}^{(t)}) + \left\langle \nabla f(\boldsymbol{y}^{(t)}), \boldsymbol{x} - \boldsymbol{y}^{(t)} \right\rangle + \left\langle \nabla f(\boldsymbol{y}^{(t)}), \boldsymbol{z}^{(t+1)} - \boldsymbol{x} \right\rangle \\ &+ \frac{L}{2}\|\boldsymbol{z}^{(t+1)} - \boldsymbol{y}^{(t)}\|_2^2. \end{aligned} \tag{96}$$

Then, plugging equation (94), and equation (95) into equation (96) we obtain that

$$\begin{aligned} F(\boldsymbol{z}^{(t+1)}) &\leq g(\boldsymbol{z}^{(t+1)}) + f(\boldsymbol{x}) + \left\langle \nabla f(\boldsymbol{y}^{(t)}), \boldsymbol{z}^{(t+1)} - \boldsymbol{x} \right\rangle + \frac{1}{\kappa_f}\|\boldsymbol{x} - \boldsymbol{y}^{(t)}\|_2^2 + \frac{L}{2}\|\boldsymbol{z}^{(t+1)} - \boldsymbol{y}^{(t)}\|_2^2 \\ &\leq F(\boldsymbol{x}) + \left\langle \nabla f(\boldsymbol{y}^{(t)}) + \frac{1}{\alpha_1}(\boldsymbol{z}^{(t+1)} - \boldsymbol{y}^{(t)}), \boldsymbol{x} - \boldsymbol{z}^{(t+1)} \right\rangle + \left\langle \nabla f(\boldsymbol{y}^{(t)}), \boldsymbol{z}^{(t+1)} - \boldsymbol{x} \right\rangle \\ &+ \frac{1}{\kappa_f}\|\boldsymbol{x} - \boldsymbol{y}^{(t)}\|_2^2 + \frac{1}{\kappa_g}\|\boldsymbol{x} - \boldsymbol{z}^{(t+1)}\|_2^2 + \frac{L}{2}\|\boldsymbol{z}^{(t+1)} - \boldsymbol{y}^{(t)}\|_2^2 \\ &\leq F(\boldsymbol{x}) + \frac{1}{\alpha_1}\left\langle \boldsymbol{z}^{(t+1)} - \boldsymbol{y}^{(t)}, \boldsymbol{x} - \boldsymbol{z}^{(t+1)} \right\rangle + \frac{L}{2}\|\boldsymbol{z}^{(t+1)} - \boldsymbol{y}^{(t)}\|_2^2 \\ &+ \frac{1}{\kappa}\left(\|\boldsymbol{x} - \boldsymbol{y}^{(t)}\|_2^2 + \|\boldsymbol{x} - \boldsymbol{z}^{(t+1)}\|_2^2\right), \end{aligned} \tag{97}$$

where $\frac{1}{\kappa} = \max\{\frac{1}{\kappa_f}, \frac{1}{\kappa_g}\}$. Observe that the last term in the above equation (97) satisfies

$$\frac{1}{\kappa}\left(\|\boldsymbol{x} - \boldsymbol{y}^{(t)}\|_2^2 + \|\boldsymbol{x} - \boldsymbol{z}^{(t+1)}\|_2^2\right) \leq \frac{1}{\kappa}\|\boldsymbol{z}^{(t+1)} - \boldsymbol{y}^{(t)}\|_2^2 - \frac{2}{\kappa}\left\langle \boldsymbol{z}^{(t+1)} - \boldsymbol{y}^{(t)}, \boldsymbol{x} - \boldsymbol{z}^{(t+1)} \right\rangle. \tag{98}$$

Combining equation (97) and equation (98) we obtain

$$\begin{aligned} F(\boldsymbol{z}^{(t+1)}) &\leq F(\boldsymbol{x}) + \left(\frac{1}{\alpha_1} - \frac{2}{\kappa}\right)\left\langle \boldsymbol{z}^{(t+1)} - \boldsymbol{y}^{(t)}, \boldsymbol{x} - \boldsymbol{z}^{(t+1)}\right\rangle + \left(\frac{L}{2} + \frac{1}{\kappa}\right)\|\boldsymbol{z}^{(t+1)} - \boldsymbol{y}^{(t)}\|_2^2 \\ &= F(\boldsymbol{x}) + \left(\frac{1}{\alpha_1} - \frac{2}{\kappa}\right)\left\langle \boldsymbol{z}^{(t+1)} - \boldsymbol{y}^{(t)}, \boldsymbol{x} - \boldsymbol{y}^{(t)}\right\rangle - \left(\frac{1}{\alpha_1} - \frac{L}{2} - \frac{3}{\kappa}\right)\|\boldsymbol{z}^{(t+1)} - \boldsymbol{y}^{(t)}\|_2^2. \end{aligned} \tag{99}$$

Thus, if $\alpha_1 < \frac{\kappa}{L\kappa+4}$ then we have

$$F(\boldsymbol{z}^{(t+1)}) \leq F(\boldsymbol{x}) + \left(\frac{1}{\alpha_1} - \frac{2}{\kappa}\right)\left\langle \boldsymbol{z}^{(t+1)} - \boldsymbol{y}^{(t)}, \boldsymbol{x} - \boldsymbol{y}^{(t)}\right\rangle - \left(\frac{1}{2\alpha_1} - \frac{1}{\kappa}\right)\|\boldsymbol{z}^{(t+1)} - \boldsymbol{y}^{(t)}\|_2^2. \tag{100}$$

Let $\boldsymbol{x} = \boldsymbol{x}^{(t)}$ and $\boldsymbol{x}^*$, we have

$$F(\boldsymbol{z}^{(t+1)}) \leq F(\boldsymbol{x}^{(t)}) + \left(\frac{1}{\alpha_1} - \frac{2}{\kappa}\right)\left\langle \boldsymbol{z}^{(t+1)} - \boldsymbol{y}^{(t)}, \boldsymbol{x}^{(t)} - \boldsymbol{y}^{(t)}\right\rangle - \left(\frac{1}{2\alpha_1} - \frac{1}{\kappa}\right)\|\boldsymbol{z}^{(t+1)} - \boldsymbol{y}^{(t)}\|_2^2 \tag{101}$$

$$F(\boldsymbol{z}^{(t+1)}) \leq F(\boldsymbol{x}^*) + \left(\frac{1}{\alpha_1} - \frac{2}{\kappa}\right)\left\langle \boldsymbol{z}^{(t+1)} - \boldsymbol{y}^{(t)}, \boldsymbol{x}^* - \boldsymbol{y}^{(t)}\right\rangle - \left(\frac{1}{2\alpha_1} - \frac{1}{\kappa}\right)\|\boldsymbol{z}^{(t+1)} - \boldsymbol{y}^{(t)}\|_2^2, \tag{102}$$

Define $\frac{1}{\alpha'} = \left(\frac{1}{\alpha_1} - \frac{2}{\kappa}\right)$. Then, multiplying equation (101) by $r_t - 1$ and adding equation (102) we have

$$\begin{aligned} &r_t F(\boldsymbol{z}^{(t+1)}) - (r_t - 1)F(\boldsymbol{x}^{(t)}) - F(\boldsymbol{x}^*) \\ &\leq \frac{1}{\alpha'}\left\langle \boldsymbol{z}^{(t+1)} - \boldsymbol{y}^{(t)}, (r_t - 1)(\boldsymbol{x}^{(t)} - \boldsymbol{y}^{(t)}) + \boldsymbol{x}^* - \boldsymbol{y}^{(t)}\right\rangle - \frac{r_t}{2\alpha'}\|\boldsymbol{z}^{(t+1)} - \boldsymbol{y}^{(t)}\|_2^2. \end{aligned} \tag{103}$$

So we have

$$\begin{aligned} &r_t\left(F(\boldsymbol{z}^{(t+1)}) - F(\boldsymbol{x}^*)\right) - (r_t - 1)\left(F(\boldsymbol{x}^{(t)}) - F(\boldsymbol{x}^*)\right) \\ &\leq \frac{1}{\alpha'}\left\langle \boldsymbol{z}^{(t+1)} - \boldsymbol{y}^{(t)}, (r_t - 1)(\boldsymbol{x}^{(t)} - \boldsymbol{y}^{(t)}) + \boldsymbol{x}^* - \boldsymbol{y}^{(t)}\right\rangle - \frac{r_t}{2\alpha'}\|\boldsymbol{z}^{(t+1)} - \boldsymbol{y}^{(t)}\|_2^2. \end{aligned} \tag{104}$$

Multiplying both sides by r_t and using $(r_t)^2 - r_t = (r_{t-1})^2$ from APGM we have

$$\begin{aligned} &(r_t)^2\left(F(\boldsymbol{z}^{(t+1)}) - F(\boldsymbol{x}^*)\right) - (r_{t-1})^2\left(F(\boldsymbol{x}^{(t)}) - F(\boldsymbol{x}^*)\right) \\ &\leq \frac{1}{\alpha'}\left\langle r_t\left(\boldsymbol{z}^{(t+1)} - \boldsymbol{y}^{(t)}\right), (r_t - 1)\left(\boldsymbol{x}^{(t)} - \boldsymbol{y}^{(t)}\right) + \boldsymbol{x}^* - \boldsymbol{y}^{(t)}\right\rangle - \frac{1}{2\alpha'}\|r_t(\boldsymbol{z}^{(t+1)} - \boldsymbol{y}^{(t)})\|_2^2 \\ &= \frac{1}{\alpha'}\left\langle r_t\left(\boldsymbol{z}^{(t+1)} - \boldsymbol{y}^{(t)}\right), (r_t - 1)\boldsymbol{x}^{(t)} - r_t\boldsymbol{y}^{(t)} + \boldsymbol{x}^*\right\rangle - \frac{1}{2\alpha'}\|r_t(\boldsymbol{z}^{(t+1)} - \boldsymbol{y}^{(t)})\|_2^2 \\ &= \frac{1}{2\alpha'}\left(\|(r_t - 1)\boldsymbol{x}^{(t)} - r_t\boldsymbol{y}^{(t)} + \boldsymbol{x}^*\|_2^2 - \|(r_t - 1)\boldsymbol{x}^{(t)} - r_t\boldsymbol{z}^{(t+1)} + \boldsymbol{x}^*\|_2^2\right). \end{aligned} \tag{105}$$

Define

$$U^{(t+1)} = r_t\boldsymbol{z}^{(t+1)} - (r_t - 1)\boldsymbol{x}^{(t)} - \boldsymbol{x}^*. \tag{106}$$

Let

$$U^{(t)} = r_{t-1}\boldsymbol{z}^{(t)} - (r_{t-1} - 1)\boldsymbol{x}^{(t-1)} - \boldsymbol{x}^* = r_t\boldsymbol{y}^{(t)} - (r_t - 1)\boldsymbol{x}^{(t)} - \boldsymbol{x}^*. \tag{107}$$

We have

$$\begin{aligned} \boldsymbol{y}^{(t)} &= \frac{r_{t-1}\boldsymbol{z}^{(t)} - (r_{t-1} - 1)\boldsymbol{x}^{(t-1)} + (r_t - 1)\boldsymbol{x}^{(t)}}{r_t} \\ &= \boldsymbol{x}^{(t)} + \frac{r_{t-1}}{r_t}(\boldsymbol{z}^{(t)} - \boldsymbol{x}^{(t)}) + \frac{r_{t-1} - 1}{r_t}(\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t-1)}), \end{aligned} \tag{108}$$

which is the same in Step 1 of APGM. So we have

$$(r_t)^2 \left(F(\boldsymbol{z}^{(t+1)}) - F(\boldsymbol{x}^*)\right) - (r_{t-1})^2 \left(F(\boldsymbol{x}^{(t)}) - F(\boldsymbol{x}^*)\right) \leq \frac{1}{2\alpha'} \left(\|U^{(t)}\|_2^2 - \|U^{(t+1)}\|_2^2\right). \tag{109}$$

If $F(\boldsymbol{z}^{(t+1)}) \leq F(\boldsymbol{v}^{(t+1)})$, then $\boldsymbol{x}^{(t+1)} = \boldsymbol{z}^{(t+1)}$

$$\begin{aligned}&(r_t)^2 \left(F(\boldsymbol{z}^{(t+1)}) - F(\boldsymbol{x}^*)\right) - (r_{t-1})^2 \left(F(\boldsymbol{x}^{(t)}) - F(\boldsymbol{x}^*)\right)\\ =&(r_t)^2 \left(F(\boldsymbol{x}^{(t+1)}) - F(\boldsymbol{x}^*)\right) - (r_{t-1})^2 \left(F(\boldsymbol{x}^{(t)}) - F(\boldsymbol{x}^*)\right)\\ \leq&\frac{1}{2\alpha'} \left(\|U^{(t)}\|_2^2 - \|U^{(t+1)}\|_2^2\right).\end{aligned} \tag{110}$$

If $F(\boldsymbol{z}^{(t+1)}) > F(\boldsymbol{v}^{(t+1)})$, then $\boldsymbol{x}^{(t+1)} = \boldsymbol{v}^{(t+1)}$. So,

$$\begin{aligned}&(r_t)^2 \left(F(\boldsymbol{x}^{(t+1)}) - F(\boldsymbol{x}^*)\right) - (r_{t-1})^2 \left(F(\boldsymbol{x}^{(t)}) - F(\boldsymbol{x}^*)\right)\\ \leq&(r_t)^2 \left(F(\boldsymbol{z}^{(t+1)}) - F(\boldsymbol{x}^*)\right) - (r_{t-1})^2 \left(F(\boldsymbol{x}^{(t)}) - F(\boldsymbol{x}^*)\right)\\ \leq&\frac{1}{2\alpha'} \left(\|U^{(t)}\|_2^2 - \|U^{(t+1)}\|_2^2\right).\end{aligned} \tag{111}$$

Summing equation (111) over $t = 1, \dots, T$ we have

$$\begin{aligned}&(r_{T+1})^2(F(\boldsymbol{x}^{(T+1)}) - F(\boldsymbol{x}^*))\\ =&(r_{T+1})^2(F(\boldsymbol{x}^{(T+1)}) - F(\boldsymbol{x}^*)) - (r_0)^2(F(\boldsymbol{x}^{(1)}) - F(\boldsymbol{x}^*))\\ \leq&\frac{1}{2\alpha'} \left(\|U^{(1)}\|_2^2 - \|U^{(T+1)}\|_2^2\right) \leq \frac{1}{2\alpha'}\|U^{(1)}\|_2^2 = \frac{1}{2\alpha'}\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\|_2^2.\end{aligned} \tag{112}$$

From APGM we can easily have that $r_t \geq \frac{t+1}{2}$. So we have

$$F(\boldsymbol{x}^{(T+1)}) - F(\boldsymbol{x}^*) \leq \frac{2}{\alpha'(T+1)^2}\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\|_2^2, \tag{113}$$

where $\frac{1}{\alpha'} = \left(\frac{1}{\alpha_1} - \frac{2}{\kappa}\right)$, and $\alpha_1 < \frac{\kappa}{L\kappa+4}$, with $\frac{1}{\kappa} = \max\{\frac{1}{\kappa_f}, \frac{1}{\kappa_g}\}$ for $k_f < \frac{1}{L}$. □

Remarks on Stability of APGM

The APGM stability and reliability come from the pivotal factors: 1) APGM employs two proximal steps that have a unique solution (as rigorously demonstrated in the section above), aiding convergence and stability. 2) As shown in the section above, the sequence $\boldsymbol{x}^{(t+1)}$, constructed by APGM algorithm, always converge to global optima irrespective of the initial state $\boldsymbol{x}^{(0)}$, conferring a steadfast assurance of reliable attainment of optimal solutions. 3) The remarkable effectiveness demonstrated across an array of real-world applications (Beck & Teboulle, 2009; Ochs et al., 2014; Boţ et al., 2016) reaffirms APGM's reliability.

K Additional Implementation Details and Results

In this section we present additional numerical results and details to complement the three experiments in Section 5. For Experiments 1, and 2 the constants c, α of the adaptive invex loss in equation (7) are implemented as

$$c = \text{softplus}(c_{var}) + c_{min} \tag{114}$$

$$\alpha = (\alpha_{max} - \alpha_{min})\text{sigmoid}(\alpha_{var}) + \alpha_{min} \tag{115}$$

where c_{var}, and α_{var} are parameters to be learned simultaneously with the network weights using the same Adam optimizer instance. In addition, we fix $\alpha_{max} = 1.99$, $\alpha_{min} = 0.0$, and $c_{min} = 10^{-8}$.

K.1 Experiment 1

To train FISTA-Net[1] for the CT experiment using convex, invex and quasi-invex regularizers, we use a batch size of 64, optimized using Adam, and with seven hidden layers. Datasets used for training, validation and testing were normalized to the interval of $[0, 1]$. In addition to the experiments reported in Table 3 for Experiment 1, we show the performance of invex regularizers reported in (Pinilla et al., 2022) under FISTA-Net. These results are summarized in Table 6.

Table 6: Performance Comparison: Lower the better. We highlight the best, and the worst results for each metric. (Green: Best, and Red: Worst).

Experiment 1 (**Combination of functions used to train FISTA-Net**)							
Regularizer	Metrics	equation (4)	equation (5)	equation (6)	equation (7)	equation (8)	ℓ_2-norm
equation 7 in (Pinilla et al., 2022)	AbsError	0.2380	0.2106	0.2212	0.1867	0.1973	0.2556
	PSNR	38.36	39.09	38.65	39.13	39.08	38.08
	RMSE	0.0150	0.0124	0.0123	0.0105	0.0095	0.0174
	SSIM	0.9550	0.9574	0.9566	0.9590	0.9582	0.9558
equation 8 (Pinilla et al., 2022)	AbsError	0.2768	0.2501	0.2352	0.2240	0.2586	0.2941
	PSNR	38.05	38.33	38.72	38.87	38.78	37.79
	RMSE	0.0179	0.0104	0.0145	0.0137	0.0117	0.0217
	SSIM	0.9537	0.9554	0.9563	0.9580	0.9571	0.9545
equation 10 (Pinilla et al., 2022)	AbsError	0.2088	0.1819	0.1700	0.1601	0.1933	0.2259
	PSNR	38.68	39.40	39.32	39.54	38.99	38.38
	RMSE	0.0129	0.0087	0.0111	0.0108	0.0095	0.0146
	SSIM	0.9563	0.9585	0.9578	0.9600	0.9593	0.9571

K.2 Experiment 2

To train MST++[2] using the invex loss functions we use a batch size of 20, optimized using Adam with parameters $\beta_1 = 0.9$, and $\beta_2 = 0.999$, and the cosine Annealing scheme is adopted for 300 epochs. Datasets used for training, validation and testing were normalized to the interval of $[0, 1]$. To complement results in Table 3 we present Figure 2 which reports some restored spectral images using MST++ trained with losses equation (4)-equation (8). To evaluate the performance we employ the absolute error.

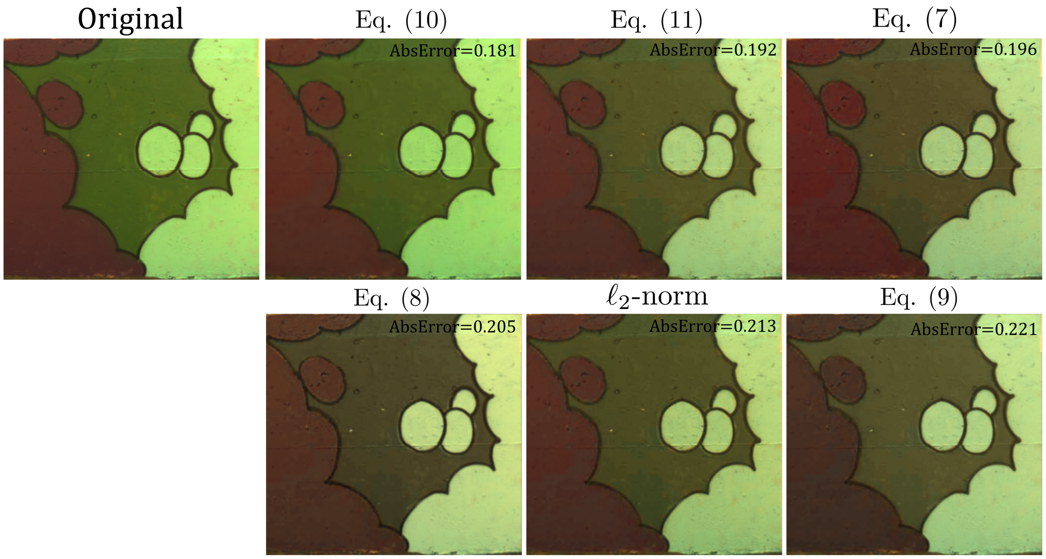


Figure 2: Reconstructed spectral images with MST++ trained with losses ℓ_2-norm, and the invex functions equation (4)-equation (8). To evaluate the performance we employ the absolute error.

[1] code of can be found at `https://github.com/jinxixiang/FISTA-Net`

[2] Implementation can be found in `https://github.com/caiyuanhao1998/MST-plus-plus`

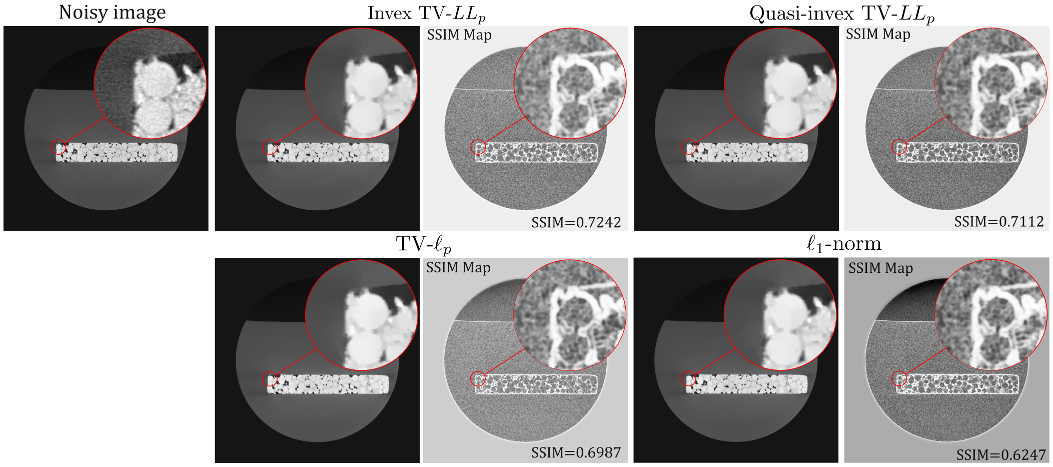


Figure 3: Restored images using the ADMM iterative process in equation (116) to solve program (11) using total-variation-like invex/quasi-invex regularizers g of ℓ_p-quasinorm, and equation (3) equation (10). To evaluate the performance we employ the structural similarity index measure (SSIM) by reporting the SSIM map for each image and its averaged value. Recall that SSIM is reported in the range [0,1] where 1 is the best achievable quality and 0 the worst. In the SSIM map small values of SSIM appear as dark pixels.

K.3 Experiment 3

In order to implement the ADMM iterative process in equation (13) to solve the total variation filtering optimization problem in (11), we use a linear approximation to update $\boldsymbol{x}^{(t+1)}$ based on (Ouyang et al., 2015). The reason of this approximation comes from the fact that updating $\boldsymbol{x}^{(t+1)}$ for the total variation filtering optimization problem in (11) requires the computation of $(\rho\boldsymbol{D}^T\boldsymbol{D}+\boldsymbol{I})^{-1}$ which is a computationally expensive step. Therefore, we implement the following ADMM iteration process to solve program (11) for the total-variation-like invex/quasi-invex regularizers g of ℓ_p-quasinorm, and equation (3) equation (10)

$$\begin{aligned}
\boldsymbol{x}_2^{(t+1)} &= (1-\alpha)\boldsymbol{x}^{(t)} + \alpha\boldsymbol{x}_1^{(t)} \\
\boldsymbol{x}_1^{(t+1)} &= \boldsymbol{x}_1^{(t)} - \frac{1}{2\beta}\left(\rho\boldsymbol{D}^T(\boldsymbol{D}\boldsymbol{x}_2^{(t+1)} - \boldsymbol{z}^{(t)} + \boldsymbol{v}^{(t)}) + \boldsymbol{x}_2^{(t+1)} - \boldsymbol{u}\right) \\
\boldsymbol{x}^{(t+1)} &= (1-\alpha)\boldsymbol{x}^{(t)} + \alpha\boldsymbol{x}_1^{(t+1)} \\
\boldsymbol{z}^{(t+1)} &= \text{Prox}_{\lambda/\rho g}(\boldsymbol{D}\boldsymbol{x}^{(t+1)} + \boldsymbol{v}^{(t)}/\rho) \\
\boldsymbol{v}^{(t+1)} &= \boldsymbol{v}^{(t)} + \rho(\boldsymbol{D}\boldsymbol{x}^{(t+1)} - \boldsymbol{z}^{(t+1)}),
\end{aligned} \tag{116}$$

where $\boldsymbol{D}$ is the discrete spatial derivative (see Section B for more details), the first three equality is the linear approximation to estimate $\boldsymbol{x}^{(t+1)}$ following (Ouyang et al., 2015), $\boldsymbol{u}$ is the noisy image, and α, ρ, β are positive constants chosen to be the best for each analyzed function determined by cross-validation. This iterative procedure is initialized as $\boldsymbol{x}^{(0)} = \boldsymbol{u}$, $\boldsymbol{z}^{(0)} = \boldsymbol{D}\boldsymbol{x}^{(0)}$, and $\boldsymbol{x}_1^{(0)} = \boldsymbol{x}_2^{(0)} = \boldsymbol{v}^{(0)} = \boldsymbol{0}$. Further, the proximal of regularizers g to compute $\boldsymbol{z}^{(t+1)}$ are given in Table 4. To complement results in Table 3 we present Figure 3 which reports some restored images obtained by total-variation-like invex/quasi-invex regularizers g of ℓ_p-quasinorm, and equation (3) equation (10). To numerically evaluate their performance we employ the structural similarity index measure (SSIM) by reporting the SSIM map for each denoised image and its averaged value. Recall that SSIM is reported in the range [0,1] where 1 is the best achievable quality and 0 the worst. In the SSIM map small values of SSIM appear as dark pixels.

K.4 Experiments with APGM and validity of Theorem 7

In this section we assess the performance of APGM (Li & Lin, 2015) solving compressive sensing problems. Specifically, this section provides two experiments summarized in Tables 7, and 8. The aim of the first experiment is to numerically validate Theorem 7 and the second it is to show the performance of APGM in a realistic setting when invex/quasi-invex are employed.

For the first experiment, we generate noisy measurements $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\eta}$ where $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ follows a Gaussian random matrix with ℓ_2-normalized columns, $\boldsymbol{x} \in \mathbb{R}^n$ is a k-sparse vector, and $\boldsymbol{\eta} \in \mathbb{R}^m$ models the noise. Specifically, the size n and the sparsity k of the unknown underlying signal $\boldsymbol{x}$ are fixed as $n = 1024$, and $k = 0.03n$. Therefore, appealing to the theoretical result in (Foucart & Rauhut, 2013, Theorem 9.13), we determined from the fixed values of n and k the number of measurements m as $m = 0.45n$ in order to satisfy equation (54) for $\tau = 0.3$ and $\rho = 0.4$. In consequence, the constant C and D of equation (59) are given by $C = 4.6$, and $D = 2\frac{g(\boldsymbol{x})}{\|\boldsymbol{x}\|_1}$. Under this setting, we are ready to validate inequality in equation (59) numerically (i.e. validity of Theorem 7) where the regularizer $g(\boldsymbol{x})$ in program (1) takes the form of equations (3), (9),(10), and the data fidelity term $f(\boldsymbol{x})$ takes the form of equations (5),(7),(8) (these were chosen due their performance in Table 3). Further, in order to solve program (1) and estimate the distance $\|\boldsymbol{x} - \boldsymbol{x}^*\|_1$ in equation (59), we use APGM (see Appendix J) where $\boldsymbol{x}^*$ is the solution returned by APGM. We consider additive white Gaussian noise in the measurements data vector with three different levels of SNR (Signal-to-Noise Ratio) $= 20, 30$. The number of iterations T of the APGM is fixed to $T = 1000$. The positive constants α_1, α_2 of APGM, and c, α, p of equations (7),(10) are chosen to be the best for each analyzed combination of functions by cross-validation.

We summarize the results in Table 7. From these numerical results, we observe that invex/quasi-invex mappings for both regularizer and fidelity term perform better than using the combination of ℓ_1-norm and ℓ_2-norm because the upper bound in equation (59) is smaller. Additionally, the results in Table 7 suggest this better lower upper bound is due to the regularizer. This is an expected behavior since this is a compressive sensing test, and the regularizer is the mapping that determines the reconstruction quality (Candès & Wakin, 2008).

Table 7: Performance Results: Best: green, and the worst: red. Column named ”dist” reports $\|\boldsymbol{x} - \boldsymbol{x}^*\|_1$ where $\boldsymbol{x}^*$ is the solution returned by APGM for each combination of $g(\boldsymbol{x})$, and $f(\boldsymbol{x})$ mappings. Additionally, column named ”bound” reports $C\beta_{k,g}(\boldsymbol{x}) + D\sqrt{\epsilon}\upsilon_{f,\eta}(\boldsymbol{x})$ in equation (59) for each combination of $g(\boldsymbol{x})$, and $f(\boldsymbol{x})$ mappings.

		Tested $f(\boldsymbol{x})$ functions for APGM							
		(5)		(7)		(8)		ℓ_2-norm	
$g(\boldsymbol{x})$	SNR	dist	bound	dist	bound	dist	bound	dist	bound
(3)	20dB	0.401	1.330	0.420	1.330	0.442	1.330	0.467	1.330
	30dB	0.218	0.453	0.188	0.453	0.217	0.453	0.260	0.453
(9)	20dB	0.588	5.255	0.671	5.255	0.783	5.255	0.941	5.255
	30dB	0.374	1.632	0.332	1.632	0.370	1.632	0.426	1.632
(10)	20dB	0.477	1.330	0.516	1.330	0.565	1.330	0.624	1.330
	30dB	0.275	0.453	0.240	0.453	0.273	0.453	0.323	0.453
ℓ_1-norm	20dB	0.768	5.431	0.959	5.431	1.278	5.431	1.916	5.431
	30dB	0.585	1.857	0.541	1.857	0.573	1.857	0.627	1.857

For the second experiment, we use 24 images of size 256 × 256 (i.e. $n = 65536$) from the Kodak dataset (kod), and we convert them to grayscale. In order to have noisy measurements $\boldsymbol{y}$ in a compressive sensing setup, we compute the Hadamard transform over these images $\boldsymbol{x}$ (which is sparse), after we add Cauchy distributed noise with a dispersion $\sigma = 1$, and we randomly select $m = 32000$ noisy transformed coefficients. Therefore, assuming $\boldsymbol{A}$ is a submatrix of size $m \times n$ of the Hadamard transform, then $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\eta}$ where $\boldsymbol{\eta}$ is the Cauchy distributed noise. Thus, these experiments are intended to recover $\boldsymbol{x}$ from the noisy measurements $\boldsymbol{y}$ using APGM, where the data fidelity functions $f(\boldsymbol{x})$ are the ℓ_2-norm, and equation (4)-equation (8), and the regularizers $g(\boldsymbol{x})$ are the ℓ_1-norm and equation (3), equation (9), equation (10). We point out that the combination of ℓ_2-norm as data fidelity term and ℓ_1-norm as regularizer is taken as the convex state-of-the-art. The number of iterations T of the APGM is fixed to $T = 1000$. The positive constants α_1, α_2 of APGM, and c, α, p of equations (7),(10) are chosen to be the best for each analyzed combination of functions by cross-validation. The performance for each combination of functions is measured by averaging the peak-signal-to-noise-ratio (PSNR) in dB over the image set. These results are summarized in

Table 8. These results suggest that the best result is obtained employing regularizer in equation (3) and fidelity term in equation (7). The intuition behind the superiority combining these two mapping comes from the possibility of adjusting the value of p for equation (3) and c, α for equation (7) in data-dependent manner (Wu & Chen, 2013; Barron, 2019).

Table 8: Performance Results: Best: green, and the worst: red.

Tested $f(\boldsymbol{x})$ functions for APGM							
Regularizer	Metric	equation (4)	equation (5)	equation (6)	equation (7)	equation (8)	ℓ_2-norm
equation (3)	PSNR	20.53	17.48	16.27	22.50	18.88	15.25
equation (9)	PSNR	18.03	15.64	14.67	19.52	16.76	13.83
equation (10)	PSNR	19.20	16.51	15.43	20.90	17.75	14.50
ℓ_1-norm	PSNR	17.00	14.87	13.99	18.32	15.87	13.22

REFERENCES

Cupy library. https://cupy.dev/. Accessed: 2023-09-11.

Cupy kernel. https://docs.cupy.dev/en/stable/user_guide/kernel.html. Accessed: 2023-09-11.

Kodak image dataset. http://r0k.us/graphics/kodak/. Accessed: 2023-08-08.

Pytorch and cupy. https://docs.cupy.dev/en/stable/user_guide/interoperability.html#using-custom-kernels-in-pytorch. Accessed: 2023-09-11.

Adil Bagirov, Napsu Karmitsa, and Marko M Mäkelä. *Introduction to Nonsmooth Optimization: theory, practice and software*. Springer, 2014.

Mokhtar S Bazaraa and Chitharanjan M Shetty. *Foundations of optimization*, volume 122. Springer Science & Business Media, 2012.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

Radu Ioan Boţ, Ernö Robert Csetnek, and Szilárd Csaba László. An inertial forward–backward algorithm for the minimization of the sum of two nonconvex functions. *EURO Journal on Computational Optimization*, 4(1):3–25, 2016.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

Richard L Burden and J Douglas Faires. 2.1 the bisection algorithm. *Numerical analysis*, 3, 1985.

T Tony Cai and Lie Wang. Orthogonal matching pursuit for sparse signal recovery with noise. *IEEE Transactions on Information theory*, 57(7):4680–4688, 2011.

Ariolfo Camacho, Edwin Vargas, and Henry Arguello. Hyperspectral and multispectral image fusion addressing spectral variability by an augmented linear mixing model. *International Journal of Remote Sensing*, 43(5):1577–1608, 2022.

Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM review*, 43(1):129–159, 2001.

Wei Deng, Ming-Jun Lai, Zhimin Peng, and Wotao Yin. Parallel multi-block admm with $\mathcal{O}(1/k)$ convergence. *Journal of Scientific Computing*, 71:712–736, 2017.

Simon Foucart and Holger Rauhut. *A Mathematical Introduction to Compressive Sensing*. 2013. ISBN 9780817649487 0817649484. doi: 10.1007/978-0-8176-4948-7.

Pierre Frankel, Guillaume Garrigos, and Juan Peypouquet. Splitting methods with variable metric for kurdyka–łojasiewicz functions and general convergence rates. *Journal of Optimization Theory and Applications*, 165(3):874–900, 2015.

Roland Glowinski. On alternating direction methods of multipliers: a historical perspective. *Modeling, simulation and optimization for science and technology*, pp. 59–82, 2014.

Rémi Gribonval and Morten Nielsen. Highly sparse representations from dictionaries are unique and independent of the sparseness measure. *Applied and Computational Harmonic Analysis*, 22 (3):335–355, 2007.

Jeremy Johnston, Yinchuan Li, Marco Lops, and Xiaodong Wang. ADMM-Net for communication interference removal in stepped-frequency radar. *IEEE Transactions on Signal Processing*, 69: 2818–2832, 2021.

Huan Li and Zhouchen Lin. Accelerated proximal gradient methods for nonconvex programming. *Advances in neural information processing systems*, 28, 2015.

Trang C Mai, Hien Quoc Ngo, and Le-Nam Tran. Energy-efficient power allocation in cell-free massive MIMO with zero-forcing: First order methods. *Physical Communication*, 51:101540, 2022.

Goran Marjanovic and Victor Solo. On ℓ_q optimization and matrix completion. *IEEE Transactions on signal processing*, 60(11):5714–5724, 2012.

Anke Meyer-Bäse, Anke Meyer-Baese, and Volker J Schmid. *Pattern Recognition and Signal Analysis in Medical Imaging*. Academic Press, 2004.

Shashi K Mishra and Giorgio Giorgi. *Invexity and optimization*, volume 88. Springer Science & Business Media, 2008.

Peter Ochs, Yunjin Chen, Thomas Brox, and Thomas Pock. iPiano: Inertial proximal algorithm for nonconvex optimization. *SIAM Journal on Imaging Sciences*, 7(2):1388–1419, 2014.

Yuyuan Ouyang, Yunmei Chen, Guanghui Lan, and Eduardo Pasiliao Jr. An accelerated linearized alternating direction method of multipliers. *SIAM Journal on Imaging Sciences*, 8(1):644–681, 2015.

Diethard Ernst Pallaschke and Stefan Rolewicz. *Foundations of mathematical optimization: convex analysis without linearity*, volume 388. Springer Science & Business Media, 2013.

Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 1(3): 127–239, 2014.

Samuel Pinilla, Tingting Mu, Neil Bourne, and Jeyan Thiyagalingam. Improved imaging by invex regularizers with global optima guarantees. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 10780–10794. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/45f7927942098d14e473fc5d000031e2-Paper-Conference.pdf.

Juan Marcos Ramirez, José Ignacio Martínez-Torre, and Henry Arguello. LADMM-Net: An unrolled deep network for spectral image fusion from compressive data. *Signal Processing*, 189: 108239, 2021.

Jian Sun, Huibin Li, Zongben Xu, et al. Deep ADMM-Net for compressive sensing MRI. *Advances in neural information processing systems*, 29, 2016.

Edwin Vargas, Oscar Espitia, Henry Arguello, and Jean-Yves Tourneret. Spectral image fusion from compressive measurements. *IEEE Transactions on Image Processing*, 28(5):2271–2282, 2018.

Edwin Vargas, Henry Arguello, and Jean-Yves Tourneret. Spectral image fusion from compressive measurements using spectral unmixing and a sparse representation of abundance maps. *IEEE Transactions on Geoscience and Remote Sensing*, 57(7):5043–5053, 2019.

Jiaxi Wang, Li Zeng, Chengxiang Wang, and Yumeng Guo. Admm-based deep reconstruction for limited-angle CT. *Physics in Medicine & Biology*, 64(11):115011, 2019a.

Weina Wang and Yunmei Chen. An accelerated smoothing gradient method for nonconvex nonsmooth minimization in image processing. *Journal of Scientific Computing*, 90(1):1–28, 2022.

Yu Wang, Wotao Yin, and Jinshan Zeng. Global convergence of ADMM in nonconvex nonsmooth optimization. *Journal of Scientific Computing*, 78:29–63, 2019b.

Joseph Woodworth and Rick Chartrand. Compressed sensing recovery via nonconvex shrinkage penalties. *Inverse Problems*, 32(7):075004, 2016.

Xingyu Xie, Jianlong Wu, Guangcan Liu, Zhisheng Zhong, and Zhouchen Lin. Differentiable linearized ADMM. In *International Conference on Machine Learning*, pp. 6902–6911. PMLR, 2019.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942, 2010.