Supplementary Material:
Non-Convex Rank/Sparsity Regularization and Local Minima

Carl Olsson[1,2] Marcus Carlsson[2] Fredrik Andersson[2] Viktor Larsson[2]

[1]Department of Electrical Engineering
Chalmers University of Technology

[2]Centre for Mathematical Sciences
Lund University

{calle,mc,fa,viktorl}@maths.lth.se

1. Bounding the differential $\partial g(x)$

In this section we prove the growth estimate given in Lemma 3.2 (of the main paper). Recall that the sub-differential of g is given by

$$\partial g(x) = \begin{cases} \{2x\} & |x| \geq \sqrt{\mu} \\ \{2\sqrt{\mu}\text{sign}(x)\} & 0 < |x| \leq \sqrt{\mu} \\ [-2\sqrt{\mu}, 2\sqrt{\mu}] & x = 0 \end{cases}. \tag{1}$$

Lemma S.1. *Assume that* $2\mathbf{z} \in \partial g(\mathbf{x})$. *If*

$$|z_i| > \frac{\sqrt{\mu}}{1-\delta_c} \tag{2}$$

then for any $\mathbf{z}'$ *with* $2\mathbf{z}' \in \partial g(\mathbf{x}+\mathbf{v})$ *we have*

$$z_i' - z_i > \delta_c v_i \quad \textit{if } v_i > 0 \tag{3}$$

and

$$z_i' - z_i < \delta_c v_i \quad \textit{if } v_i < 0. \tag{4}$$

Proof. We first assume that $x_i > 0$. Because of (2) and (1) we have $x_i = z_i > \frac{\sqrt{\mu}}{1-\delta_c}$. There are now two possibilities:

- If $v_i > 0$ then $x_i + v_i > \sqrt{\mu}$ and by (1) we therefore must have that $z_i' = x_i + v_i = z_i + v_i > z_i + \delta_c v_i$.
- If $v_i < 0$ we consider the line

  $$l(x) = 2z_i + 2\delta_c(x - x_i). \tag{5}$$

  See the left graph of Figure 1. We will show that this line is an upper bound on the sub-gradients for all $v_i < 0$.

  We note that for $x < x_i$ we have

  $$l(x) = \underbrace{2z_i - 2x_i}_{=0} + \underbrace{2(1-\delta_c)x_i}_{>2(1-\delta_c)x} + 2\delta_c x > 2x. \tag{6}$$

  Furthermore

  $$l(x) = 2(1-\delta_c)x_i + 2\delta_c x > 2\sqrt{\mu} + 2\delta_c x. \tag{7}$$

  The right hand side is clearly larger than both $2\sqrt{\mu}$ for $x \geq 0$. For $-\sqrt{\mu} \leq x \leq 0$ we have $2\sqrt{\mu} + 2\delta_c x > 2\sqrt{\mu} + 2x \geq 0 \geq -2\sqrt{\mu}$. This shows that the line $l(x)$ is an upper bound on the subgradients of g for every $x < x_i$, that is $l(x_i + v_i) > 2z_i'$ for all $v_i < 0$ and since $l(x_i + v_i) = 2z_i + 2\delta_c v_i$ we get $2z_i' < 2z_i + 2\delta_c v_i$.

The proof for the case $x_i < 0$ is similar. □

Lemma S.2. *Assume that* $2\mathbf{z} \in \partial g(\mathbf{x})$. *If*

$$|z_i| < (1-\delta_c)\sqrt{\mu} \tag{8}$$

then for any $\mathbf{z}'$ *with* $2\mathbf{z}' \in \partial g(\mathbf{x}+\mathbf{v})$ *we have*

$$z_i' - z_i > \delta_c v_i \quad \textit{if } v_i > 0 \tag{9}$$

and

$$z_i' - z_i < \delta_c v_i \quad \textit{if } v_i < 0. \tag{10}$$

Proof. By (8) we see that $x_i = 0$. We first assume that $v_i > 0$ and consider the line $l(x) = 2z_i + 2\delta_c x$, see the right graph of Figure 1. We have that

$$l(x) < 2(1-\delta_c)\sqrt{\mu} + 2\delta_c x. \tag{11}$$

The right hand side is less than $2(1-\delta_c)\sqrt{\mu} + 2\delta_c\sqrt{\mu} = 2\sqrt{\mu}$ when $0 < x \leq \sqrt{\mu}$ and less than $2(1-\delta_c)x + 2\delta_c x = 2x$ when $x > \sqrt{\mu}$. Therefore $l(x)$ is a lower bound on the subgradients of g for all $x > 0$ which gives $l(v_i) < 2z_i'$ for $v_i > 0$ and since $l(v_i) = 2z_i + 2\delta_c v_i$ we get $2z_i' > 2z_i + 2\delta_c v_i$. The case $v_i < 0$ is similar. □

Proof of Lemma 3.2. The proof of the growth estimate is now an immediate consequence of the previous two results.

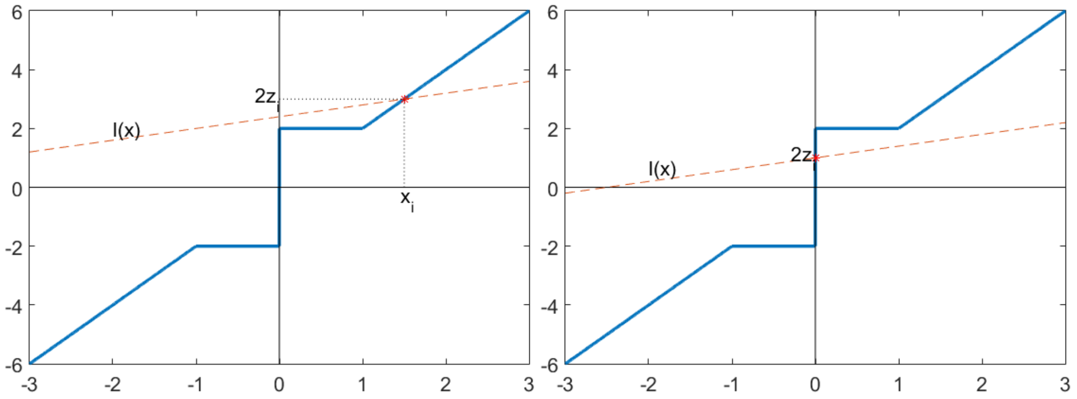


Figure 1: Illustration of the subdifferential $\partial g(x)$ and the line $l(x)$, when (2) holds (left) and when (8) holds (right).

Under the assumptions of Lemma 3.2 we have according to Lemmas 1 and 2 that

$$(z_i' - z_i)v_i > \delta_c v_i^2, \tag{12}$$

for all i with $v_i \neq 0$. Since $v_i = 0$ gives $(z_i' - z_i)v_i = 0$ summing over i gives

$$\langle \mathbf{z}' - \mathbf{z}, \mathbf{v} \rangle > \delta_c \|\mathbf{v}\|^2, \tag{13}$$

as long as $\|\mathbf{v}\| \neq 0$.

A one dimensional example. We conclude this section with a simple one dimensional example which shows that the bounds (2) and (8) cannot be made sharper. Figure 2 shows the function $r_1(x) + (\frac{1}{\sqrt{2}}x - b)^2$ for different values of $b \geq 0$. It is not difficult to verify that this function can have three stationary points (when $b \geq 0$). The point $x = 0$ is stationary if $b \leq \sqrt{2}$, $x = 2 - \sqrt{2}b$ if $\frac{1}{\sqrt{2}} < b < \sqrt{2}$ and $x = \sqrt{2}b$ if $b \geq \frac{1}{\sqrt{2}}$, see Figure 2. For this example $A = \frac{1}{\sqrt{2}}$ and therefore $(1-\delta)|x|^2 \leq |Ax|^2 \leq (1+\delta)|x|^2$ holds with $1 - \delta = \frac{1}{2}$.

Now suppose that $b \leq \sqrt{2}$ and that we, using some algorithm, find the stationary point $x = 0$. We then have

$$z = (1 - A^T A)x + A^T b = \frac{1}{\sqrt{2}}b. \tag{14}$$

Theorem 3.3 now tells us that $x = 0$ is the unique stationary point if

$$\frac{1}{\sqrt{2}}b \notin \left[1 - \delta, \frac{1}{1-\delta}\right] \Leftrightarrow b \notin \left[\frac{1}{\sqrt{2}}, 2\sqrt{2}\right]. \tag{15}$$

Note that the lower interval bound $b < \frac{1}{\sqrt{2}}$ is precisely when $x = 0$ is unique, see the leftmost graph in Figure 2.

Similarly suppose that $b \geq \frac{1}{\sqrt{2}}$. For the point $x = \sqrt{2}b$ we get

$$\frac{1}{2}z = (1 - A^T A)x + A^T b = \frac{1}{2}\sqrt{2}b + \frac{1}{\sqrt{2}}b = \sqrt{2}b. \tag{16}$$

Theorem 3.3 now shows that $x = \sqrt{2}b$ is unique if

$$\sqrt{2}b \notin \left[1 - \delta, \frac{1}{1-\delta}\right] \Leftrightarrow b \notin \left[\frac{1}{2\sqrt{2}}, \sqrt{2}\right]. \tag{17}$$

Here the upper interval bound $b > \sqrt{2}$ is precisely when $x = \sqrt{2}b$ is unique, see rightmost graph in Figure 2. Hence for this example Theorem 3.3 is tight in the sense that it would be able to verify uniqueness of the stationary point for every b where this holds.

2. Sparsity Experiments

In this section we evaluate the proposed sparsity formulation on synthetic data. We compare the two formulations

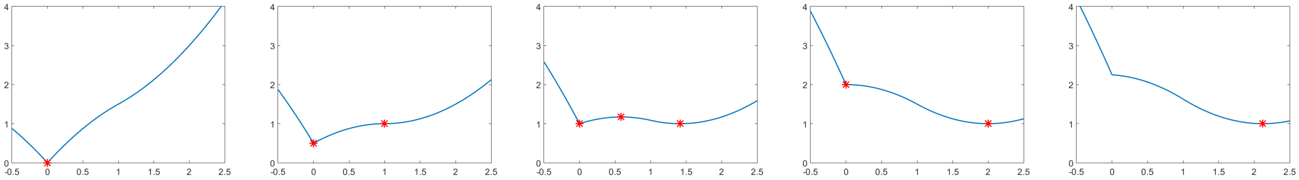
$$\mu' \|\mathbf{x}\|_1 + \|A\mathbf{x} - \mathbf{b}\|^2. \tag{18}$$

$$r_\mu(\mathbf{x}) + \|A\mathbf{x} - \mathbf{b}\|^2 \tag{19}$$

for low rank recovery for varying regularization strengths μ and μ'. Similarly to the rank case, the proximal operator of the ℓ_1-norm, $\arg\min_{\mathbf{x}} \mu' \|\mathbf{x}\|_1 + \|\mathbf{x} - \mathbf{z}\|^2$, performs soft thresholding at $\frac{\mu'}{2}$ while that of r_μ, $\arg\min_{\mathbf{x}} \mu r_\mu(\mathbf{x}) + \|\mathbf{x} - \mathbf{z}\|^2$, thresholds at $\sqrt{\mu}$ [3]. We therefore use $\mu' = 2\sqrt{\mu}$ in (18).

2.1. Optimization

For minimizing (19) we use a GIST [2] approach similar to the one described for low rank recovery. In each step we



$b = 0$ $b = \frac{1}{\sqrt{2}}$ $b = 1$ $b = \sqrt{2}$ $b = 1.5$

Figure 2: The function $r_1(x) + (\frac{1}{\sqrt{2}}x - b)^2$ and its stationary points (red) for different values of b. When b is close to the threshold 1 the function has multiple stationary points.

find $\mathbf{x}_{k+1}$ by minimizing

$$r_\mu(\mathbf{x}) + \tau_k \left\| \mathbf{x} - \underbrace{\left(\mathbf{x}_k - \frac{1}{\tau_k}(A^T A\mathbf{x}_k - A^T\mathbf{b}) \right)}_{:=m} \right\|^2 . \quad (20)$$

The optimization is separable and for each element x_i we minimize $-\max(\sqrt{\mu} - |x_i|, 0)^2 + \tau_k(x_i - m_i)^2$. It is easy to show that there are four possible choices $x_i = m_i$, $x_i = \frac{\tau_k m_i \pm \sqrt{\mu}}{\tau_k - 1}$ and $x_i = 0$ that can be optimal. In our implementation we simply test which one of these yields the smallest objective value. (If $\tau_k = 1$ it is enough to test $x_i = m_i$ and $x_i = 0$.) For initialization we use $\mathbf{x}_0 = 0$.

2.2. Sparse Recovery

For Figure 3 (a)-(c) we randomly generated problem instances for sparse recovery. Each instance uses a matrix A of size 200×200 with $\delta = 0.2$ which was generated by first randomly sampling the elements of a matrix $\tilde{A}$ a Gaussian $\mathcal{N}(0, 1)$ distribution. The matrix A was then constructed from $\tilde{A}$ by modifying the singular values to be evenly distributed between $\sqrt{1-\delta}$ and $\sqrt{1+\delta}$. To generate a ground truth solution and a $\mathbf{b}$ vector we then randomly select values for 10 nonzero elements of $\mathbf{x}$ and computed $\mathbf{b} = A\mathbf{x} + \epsilon$, where all elements of ϵ are $\mathcal{N}(0, \sigma^2)$.

The averaged results (over 50 random instances for each (σ, μ) setting) are shown in Figure 3 (a)-(c). Similar to the matrix case it is quite clear that the ℓ_1 norm (a) suffers from shrinking bias. It consistently gives the best agreement with the ground truth data for values of μ that are not big enough to generate low cardinality. In contrast, (19) gives the best fit at the correct cardinality for all noise levels. This fit was consistently better than that of (18) for all noise levels. In Figure 3 (c) we show the fraction of problem instances that could be verified to be optimal.

In Figure 3 (d) and (e) we tested the case where the elements of an $m \times n$ matrix A are sampled from $\mathcal{N}(0, \frac{1}{m})$ [1]. Here we let A be random 150×200 matrices and generated the ground truth solution and $\mathbf{b}$ vector as described previously. Here (19) consistently outperformed (18) which exhibits the same tendency to achieve a better fit for non-sparse solutions.

3. Proof of Corollary 4.2

Here we present the technical details of the continuity argument in the proof of Corollary 4.2. Recall that if $\boldsymbol{\sigma}(X') \neq \boldsymbol{\sigma}(X)$, $2Z \in \partial G(X)$, $2Z' \in \partial G(X')$ and the singular values of the matrix Z fulfill $z_i \notin [(1 - \delta_r - \epsilon)\sqrt{\mu}, \frac{\sqrt{\mu}}{1-\delta_r-\epsilon}]$, then for any $2Z' \in \partial G(X')$ we have

$$\langle Z' - Z, X' - X \rangle > (\delta_r + \epsilon)\|X' - X\|_F^2, \quad (21)$$

for some $\epsilon > 0$.

We now assume that $\|X' - X\|_F \neq 0$ and $\boldsymbol{\sigma}(X) = \boldsymbol{\sigma}(X')$. We must have $\sigma_1(X) > 0$ since otherwise $X = X' = 0$ and therefore $\|X' - X\|_F = 0$. By the definition of the sub differential we therefore know that $z_1 \geq \sqrt{\mu}$ and by the assumptions of the lemma we have that $z_1 > \frac{\sqrt{\mu}}{1-\delta_r-\epsilon}$.

If $X = UD_{\boldsymbol{\sigma}(X)}V^T$ we now define $\bar{X}(t) = UD_{\boldsymbol{\sigma}(\bar{X}(t))}V^T$, where

$$\sigma_i(\bar{X}(t)) = \begin{cases} \sigma_1(X) + t & \text{if } i = 1 \\ \sigma_i(X) & \text{otherwise} \end{cases}. \quad (22)$$

Similarly we define $\bar{Z}(t) = UD_{\bar{\mathbf{z}}(t)}V^T$, where

$$\bar{z}_i(t) = \begin{cases} z_1 + t & \text{if } i = 1 \\ z_i & \text{otherwise} \end{cases}. \quad (23)$$

It is now clear that $2\bar{Z}(t) \in \partial G(\bar{X}(t))$ and $\bar{z}_i(t) \notin [(1 - \delta_r - \epsilon)\sqrt{\mu}, \frac{\sqrt{\mu}}{1-\delta_r-\epsilon}]$, for all $t \geq 0$. Further more $\bar{Z}(t) \rightarrow \bar{Z}(0) = Z$ and $\bar{X}(t) \rightarrow \bar{X}(0) = X$ when $t \rightarrow 0^+$. Since $\boldsymbol{\sigma}(\bar{X}(t)) \neq \boldsymbol{\sigma}(X')$ for $t > 0$ we have by (25) that

$$\langle Z' - \bar{Z}(t), X' - \bar{X}(t) \rangle > (\delta_r + \epsilon)\|X' - \bar{X}(t)\|_F^2, \quad (24)$$

for all $t > 0$. By continuity of the Frobenius norm and the scalar product we can now conclude that

$$\langle Z' - Z, X' - X \rangle \geq (\delta_r + \epsilon)\|X' - X\|_F^2. \quad (25)$$

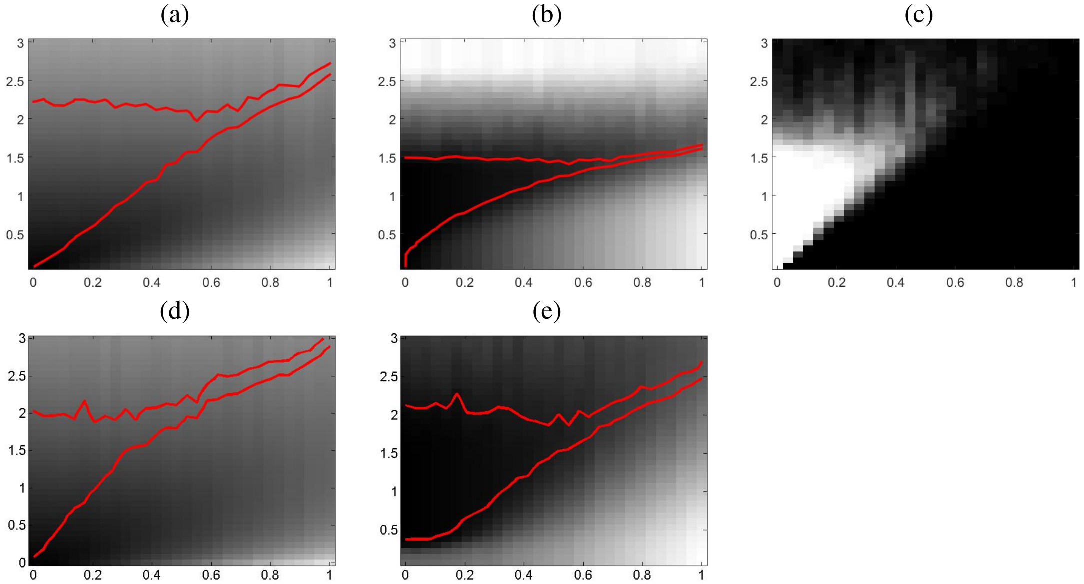


Figure 3: Sparse recovery results for varying noise level (x-axis) and regularization strength (y-axis). Top row: Random 200×200 A with $\delta = 0.2$. Bottom row: Random 150×200 A (and unknown δ). Plots (a) and (d) show the average distance between the ℓ_1 regularized and the ground truth solutions for values of μ between 0 and 3. (red curves marks the area where the obtained solution has $\operatorname{card}(\mathbf{x}) = 10$.) Plots (b) and (e) show the average distance between (19) and the ground truth solutions. Plot (c) shows the number of instances where our method could be verified to provide the global optima for $\delta = 0.2$ (white = all, black = none).

References

[1] E. J. Candes, J. K. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, 59(8):1207–1223, 2006. 3

[2] P. Gong, C. Zhang, Z. Lu, J. Huang, and J. Ye. A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. In *International Conference on Machine Learning (ICML)*, pages 37–45, 2013. 2

[3] V. Larsson and C. Olsson. Convex low rank approximation. *International Journal of Computer Vision*, 120(2):194–214, 2016. 2