Bias Reduction in Compressed Sensing

Carl Olsson\textsuperscript{1,2} Marcus Carlsson\textsuperscript{2} Daniele Gerosa\textsuperscript{2}
\textsuperscript{1}Department of Electrical Engineering Chalmers University of Technology
\textsuperscript{2}Centre for Mathematical Sciences Lund University
caois@chalmers.se mc@maths.lth.se daniele.gerosa@math.lu.se

Abstract

Sparsity and rank functions are important ways of regularizing under-determined linear systems. Optimization of the resulting formulations is made difficult since both these penalties are non-convex and discontinuous. The most common remedy is to instead use the $\ell^1$ and nuclear-norms. While these are convex and can therefore be reliably optimized they suffer from a shrinking bias that degrades the solution quality in the presence of noise.

In this paper we combine recently developed bias free non-convex alternatives with the nuclear- and $\ell^1$-penalties. This reduces bias and still enables reliable optimization properties. We develop an efficient minimization scheme using derived proximal operators and evaluate the method on several real and synthetic computer vision applications with promising results.

1. Introduction and Background

Sparsity and rank penalties are common tools for regularizing ill posed linear problems. The sparsity regularized problem is often formulated as

$$\min_x \mu \|x\|_0 + \|Ax - b\|^2,$$

where $\|x\|_0$ is the number of non-zero elements of $x$. Optimization of (1) is difficult since the term $\|x\|_0$ is non-convex and discontinuous. Moreover, the sought solutions should be sparse and therefore the minimizer is typically in the vicinity of discontinuities. A common practice is to replace $\|x\|_0$ with the $\ell^1$-norm, resulting in the convex relaxation

$$\min_x \lambda \|x\|_1 + \|Ax - b\|^2.$$  \hspace{1cm} (2)

The seminal works \cite{CandesRombergTao2006, Romberg2005, CandesRombergTao2006b, CandesRomberg2007} gave performance guarantees for this approach. The notion of Restricted Isometry Property (RIP), which was introduced in \cite{CandesRombergTao2006, CandesRomberg2007}, states that $A$ obeys a RIP if

$$(1 - \delta_K)\|x\|^2 \leq \|Ax\|^2 \leq (1 + \delta_K)\|x\|^2,$$ \hspace{1cm} (3)

for all vectors $x$ with $\|x\|_0 \leq K$. In \cite{CandesRombergTao2006}, Candès, Romberg and Tao proved the surprising result that, if $A$ fulfills $\delta_{3K} + 3\delta_{4K} < 2$, given a sparse vector $x_0$ and a measurement $b = Ax_0 + \epsilon$ where $\epsilon$ is Gaussian noise, solving (2) yields a vector $\hat{x}$ that satisfies

$$\|\hat{x} - x_0\| < C_K \|\epsilon\|,$$ \hspace{1cm} (4)

where $C_K$ is a constant. While \cite{CandesRombergTao2006} argued that it is impossible to beat a linear noise dependence, it has been observed that (2) suffers from a shrinking bias \cite{CandesRombergTao2012, Donoho2006}. Since the $\ell^1$ term not only has the (desired) effect of forcing many entries in $x$ to 0, but also the (undesired) effect of diminishing the size of the non-zero entries. This has led to a large amount of non-convex alternatives to replace the $\ell^1$-penalty, see e.g. \cite{CandesRombergTao2006a, CandesRomberg2007a, CandesRombergTao2007, CandesRomberg2008, CandesRombergTao2008, CandesRomberg2009, CandesRomberg2010, CandesRomberg2011, CandesRomberg2012, CandesRomberg2013}. Typically these come without global convergence guarantees.

In \cite{CandesRombergTao2006b, CandesRombergTao2007a, CandesRombergTao2008} a non-convex alternative that provides optimality guarantees is studied. These papers propose to replace the term $\mu \|x\|_1$ with $Q_2(\mu \|x\|_0)(x)$, where $Q_2(f)$ is the so called quadratic envelope of $f$, which is defined so that $Q_2(f)(x) + \frac{\gamma}{2}\|x\|^2$ is the convex envelope of $f(x) + \frac{\gamma}{2}\|x\|^2$ \cite{CandesRomberg2007b}. For $f(x) = \mu \|x\|_0$ this results in the objective

$$\sum_i \mu - \max(\sqrt{\mu} - \|x_i\|, 0)^2 + \|Ax - b\|^2.$$ \hspace{1cm} (5)

The resulting relaxation is non-convex. However, it was shown in \cite{CandesRombergTao2007a, CandesRombergTao2008} that if a RIP constraint holds then sparse minimizers of this formulation are generally unique (and globally optimal) even though there may exist additional non-sparse local minima. Furthermore, under the slightly more general RLIP condition, which is basically the lower bound in (3), \cite{CandesRombergTao2007a} showed that an estimate similar to (4) holds but with a much smaller constant exists. Additionally, in contrast to (2) the global optimizer of (5) is the so called ”oracle solution” \cite{CandesRombergTao2006}, which is what we get if we minimize $\|Ax - b\|^2$ over the ”true” support of $x_0$.

Rank regularization methods are largely analogous to the sparsity approaches. If $A : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}^p$ is a linear operator
we are seeking to minimize
\[ \mu \text{rank}(X) + \|AX - b\|^2_F. \] (6)

The matrix version of the RIP counterpart
\[ (1 - \delta_r)\|X\|^2_F \leq \|AX\|^2 \leq (1 + \delta_r)\|X\|^2_F, \] (7)
for all matrices \(X\) with \(\text{rank}(X) \leq r\), was introduced in [29]. Similar to the sparsity setting [29, 6] propose to replace \(\text{rank}(X)\) with the convex nuclear norm \(\|X\|_* = \sum_i \sigma_i(X)\), where \(\sigma_i(X), i = 1, ..., N\) are the singular values of \(X\). Since then a number of generalizations that give performance guarantees for the nuclear norm relaxation have appeared, e.g. [29, 27, 6, 7]. The nuclear norm penalizes both small and large singular values and therefore exhibits the same kind of shrinking bias as the \(\ell^1\)-norm. To address this non-convex alternatives that are locally optimized have been shown to improve performance [26, 23, 18, 24]. In [19] it was shown how to compute the convex envelope and proximal operator of
\[ g(\text{rank}(X)) + \|X - X_0\|^2_F, \] (8)
where \(g\) is a non-decreasing convex function. Setting \(g(\text{rank}(X)) = \mu \text{rank}(X)\) shows that \(Q_2(\mu \text{rank}(X)) = \sum_i \mu - \max(\sqrt{\mu} - \sigma_i(X)), 0)^2\). Therefore [25] proposed solving
\[ \sum_i \mu - \max(\sqrt{\mu} - \sigma_i(X)), 0)^2 + \|AX - b\|^2, \] (9)
and gave optimality conditions under RIP (7), similar to the sparsity case.

While the non-convex relaxations (5) and (9) provide unbiased alternatives to the \(\ell^1\)-nuclear-norms and are guaranteed [25, 13] to only have one sparse/low-rank stationary point when a sufficiently strong RIP holds, it is clear that there can still be poor local minimizers. For example, let \(x_h\) be a dense vector from the nullspace of \(A\) and \(x_b\) a minimizer of \(\|Ax - b\|\). Then by rescaling \(x_b\) so that all the elements of \(x_p + x_b\) have magnitude strictly larger than \(\sqrt{\mu}\) we obtain a vector that minimizes the data fit while the regularization \(Q_2(\mu\|\cdot\|_0)\) is (locally) constant around it. This construction also provides local minimizers for the original formulations (6), (1) and any other unbiased formulation that leaves elements/singular values larger than some threshold unpenalized. In this paper we remedy this by introducing a small penalty for large singular values. We study relaxations of
\[ \mu \|x\|_0 + \lambda \|x\|_1 + \|Ax - b\|^2_F, \] (10)
and
\[ \mu \text{rank}(X) + \lambda \|X\|_* + \|AX - b\|^2_F, \] (11)
for sparsity and rank regularization respectively. We propose to solve these by replacing the relaxation terms with their quadratic envelopes \(Q_2(\mu\|\cdot\|_0 + \lambda\|\cdot\|_1)\) and \(Q_2(\mu \text{rank} + \lambda\|\cdot\|_*).\) Our formulation can be seen as a trade-off between small bias and improved optimization properties. While the terms \(\gamma\|x\|_1\) and \(\gamma\|X\|_*\) introduce a bias to small solutions they also increase the convergence basin for cases where the RLIP is not strong enough to ensure global convergence of (5) and (9).

Simple optimization is often related to good modeling. Adding a weak shrinking factor may also make sense from a modeling perspective for certain applications. In this paper we exemplify with non-rigid structure form motion (NRSfM). Here each non-zero singular value corresponds to a mode of deformation. When choosing a weaker \(\mu\) (larger rank) in order to capture all fine deformations the resulting problem is often ill posed due to unobserved depths. As noted in [25] this may result in a large difference to the true reconstruction despite good data fit. The addition of the \(\lambda\|X\|_*\) allows us to separately incorporate a variable bias restricting the size of the deformations, which regularizes the problem further, see Section 6.3.

The main contributions of this paper are

- We present a class of new regularizers that leverage the benefits of previous convex and unbiased formulations.
- We characterize the stationary points of the proposed formulation, and give theoretical results that guarantee the uniqueness of a sparse stationary point.
- We show how to compute proximal operators of our regularization enabling fast optimization via splitting methods such as ADMM [2].
- We show that our new formulations generate better solutions in cases where a weak or no RIP holds.

2. Relaxations and Shrinking Bias

In this section we will study properties of our proposed relaxations of (10) and (11). We will present our results in the context of the vector case (10). The corresponding matrix versions follow by applying the regularization term to the singular values, with similar almost identical proofs. Our first theorem (which is proven in Appendix A) shows that adding the term \(\|\cdot\|_1\) before or after taking the quadratic envelop makes no difference.

**Theorem 1.** \(f : \mathbb{R}^d \to \mathbb{R}\) be a lower semicontinuous sign-invariant function such that \(f(0) = 0\) and \(f(x + y) \geq f(x)\) for every \(x, y \in \mathbb{R}^d\). Then
\[ Q_2(f + \lambda\|\cdot\|_1)(x) = Q_2(f)(x) + \lambda\|x\|_1 \] (12)
for every \(x \in \mathbb{R}^d\).
In view of the above it is clear that \( Q_2(\gamma \cdot \|0 + \lambda \cdot \|_1) = r_{\lambda, \mu} \), where
\[
 r_{\mu, \lambda}(x) = \sum_i (\mu - \max(\sqrt{\mu} - |x_i|, 0)^2) + \lambda\|x\|_1. \quad (13)
\]
We therefore propose to minimize the objective
\[
 r_{\mu, \lambda}(x) + \|Ax - b\|^2. \quad (14)
\]
Note that \( r_{\mu, \lambda}(x) + \|Ax - b\|^2 \) reduces to (2) if \( \mu = 0 \) and (5) if \( \lambda = 0 \).

Figure 1 shows an illustration of \( r_{\mu, \lambda} \) for a couple of different values of \( \lambda \). When \( \lambda = 0 \) the function is constant for values larger than \( \sqrt{\mu} \). Therefore large elements give zero gradients which can result in local algorithms getting stuck in poor local minimizers. Note that if RIP holds the results of [13, 25] show that such non-global minimizers can not be sparse under moderate noise conditions, and therefore this situation could in practice be detected. However, for a general problem instance increasing \( \lambda \) makes the regularizer closer to being convex, which as we show in Section 6, increases its convergence basin.

Figure 1. Left: The function (13) for \( \mu = 1 \) and \( \lambda = 0, 0.2, 0.4, ..., 1 \). Right: Illustration of (17) for the same \( \mu \) and \( \lambda \). The graphs show \( z_i \) (x-axis) versus \( x_i^* \) (y-axis). If \( z_i = \sqrt{\mu} + \frac{c}{2} \) all \( x_i^* \in [0, \sqrt{\mu}] \) are optimal.

To characterize the kind of shrinking bias that we can expect from this family of relaxations we now consider the problem of minimizing
\[
 \min_{x} r_{\mu, \lambda}(x) + \|x - z\|^2. \quad (15)
\]
The minimization is separable in the elements of \( x \) and therefore we are able to solve it by considering
\[
 \min_{x_i} \mu - (\max(\sqrt{\mu} - |x_i|, 0))^2 + \lambda|x_i| + (x_i - z_i^*)^2. \quad (16)
\]
This is a one dimensional problem that can easily be solved by computing stationary points, see Appendix B for details. This resulting minimizer is given by
\[
 x_i^* = \begin{cases} 
 z_i - \text{sign}(z_i) \frac{c}{2} & |z_i| > \sqrt{\mu} + \frac{c}{2} \\
 0, \sqrt{\mu} \text{sign}(z_i) & |z_i| = \sqrt{\mu} + \frac{c}{2} \\
 \{0\} & |z_i| < \sqrt{\mu} + \frac{c}{2} 
\end{cases} \quad (17)
\]
Figure 1 illustrates the solution set (17) for \( \mu = 1 \) and \( \lambda = 0, 0.2, 0.4, ..., 1 \). The shrinking bias comes from the subtraction of \( \frac{c}{2} \) from the magnitude of the large elements. Since we would ideally like these to remain unchanged it is essential to keep \( \lambda \) small. On the other hand a larger \( \lambda \) makes the regularization function \( r_{\mu, \lambda} \) "more convex" which simplifies optimization.

We conclude this section with a simple 2D-illustration of the different regularizes.

\[
 r_{\mu, \lambda}(x) + \|Ax - b\|^2 \quad \text{for a two dimensional problem with}
\]
\[
 A = \begin{pmatrix} 0.4 & 0 \\ 0 & 0.6 \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} 0.8 \\ 1.8 \end{pmatrix}. \quad (18)
\]
The blue curves show the two error residuals for \( \mu = 1 \) and \( \lambda = 0 \). It is easy to verify that the problem has two local minimizers \( x = (2, 3) \) and \( x = (0, 3) \) (which is also global). These points (and in addition \((0, 0)\) and \( (2, 0) \) are also local minima to (1) with \( \mu = 1 \). The yellow curve shows the effect of using the convex \( \ell_1 \) formulation (2), with \( \lambda = 0.7 \). Here we have used the smallest possible \( \lambda \) so that the optimum of the left residual is 0 while the right one is non-zero. The resulting solution \( x = (0, 2) \) has the correct support however, the magnitude of the non-zero element is reduced from 3 to 2 due to the shrinking bias. With our approach it is possible to chose an objective which has less bias but still a single local minimizer. Setting \( \mu = 0.7 \) and \( \lambda = 0.4 \) gives the red dashed curves with optimal point \( x \approx (0, 2.5) \).

3. Oracle Solutions

For sparsity problems the so called oracle solution [9] is what we would obtain if we somehow knew the "true" support of the solution and we were to solve the least squares problem over the non-zero entries of \( x \). We will use the notation \( A_S \) to denote the matrix which has the same entries as \( A \) in the columns indexed by \( S \) and zeros otherwise. Similarly \( A_{S^c} \) have zeros in the columns in \( S \) and \( x_S \) is a vector with zero elements in \( S^c \). The oracle solution is then
\[
 \min_{x_S} \|A_S x_S - b\|^2. \quad (19)
\]
Since \(x_S\) has support in \(S\) we have \(Ax_S = A_Sx_S\). Furthermore, since \(A = A_S + A_S^T\) equation (24) can be written
\[
u_1 + \lambda u_2 = 2x_S - 2(A_S^T A_S + A_S^T A_S)x_S + 2(A_S^T + A_S^T \epsilon),
\]
where \(u_2 = v_S\) reduces to
\[
u_1 = 2x_S + A_S^T (A_Sx_S - b) = 2x_S + 2A_S^T \epsilon.
\]
On \(S\) we have \(u_1 = 2x_S\). Since the non-zero elements of \(x_S\) are larger than \(\sqrt{\mu}\) the first case of (25) holds on \(S\). Similarly, on \(\overline{S}\) we have \(u_1 = 2A_S^T \epsilon\) with elements smaller than \(2\sqrt{\mu}\). Since \(x_S\) is zero on \(\overline{S}\), case 3 of (25) holds here, which shows that \(u_1 \in \partial g_0(x_S)\).

Whether \(x_S\) is the global optimum or not depends on the problem instance. In the following sections we will show that under a sufficiently strong RIP it is the sparsest possible stationary point.

### 4. Separation of Stationary Points

In this section we study the stationary points of the objective function (14) under the assumption that \(A\) fulfills the RIP condition (3) for all \(x\) with \(\|x\|_0 \leq K\). We will extend the results of [13, 25] to our class of functionals. Specifically, we show that under some technical conditions two stationary points \(x'\) and \(x''\) have to be separated by \(\|x' - x''\|_0 > K\). From a practical point of view this means that if we find a stationary point with \(\|x'\|_0 \leq K\) we can be certain that this is the sparsest one possible.

#### 4.1. Stationary Points and Local Approximation

We will first characterize a stationary point as being a thresholded version of a noisy vector \(z\) which depends on the data. Let \(g_\lambda\) and \(b\) be defined as in the previous section.

**Lemma 1.** If \(z' = (I - A^TA)x' + A^Tb\) the point \(x'\) is stationary in (14) if and only if \(2z' \in \partial g_\lambda(x')\) and if and only if
\[
x' \in \arg \min_x r_{\mu,\lambda}(x) + \|x - z'\|^2.
\]

**Proof.** By differentiating \(g_\lambda(x) + h(x)\) we see that \(x'\) is stationary in (14) if and only if \(2z' \in \partial g_\lambda(x')\). Similarly, differentiating (28) we see that \(x'\) is stationary in (28) if and only if \(2z' \in \partial g_\lambda(x')\). Now recall that by definition (28) is convex and therefore \(x'\) being stationary is equivalent to solving (28).

We will use properties of the vector \(z'\) to establish conditions that ensure that \(x'\) is the sparsest possible stationary point to (14). The overall idea which follows [25, 12] is to show that subdifferential \(\partial g_\lambda\) grows faster than \(-\nabla h\) and therefore we can only have \(-\nabla h(x) \in \partial g_\lambda(x)\) in one
(sparse) point. This requires an estimate of the growth of the subgradients of \(g_\lambda\) which we now present.

The function \(g_\lambda\) is separable and can be evaluated separately for each element of \(x\). The subdifferential \(\partial g_\lambda(x)\) can be written

\[
\partial g_\lambda(x) = \begin{cases} 
2x + \lambda \text{sign}(x) & |x| \geq \sqrt{\mu} \\
(2\sqrt{\mu} + \lambda) \text{sign}(x) & 0 < |x| \leq \sqrt{\mu} \\
[-2\sqrt{\mu} - \lambda, 2\sqrt{\mu} + \lambda] & x = 0
\end{cases}
\]

Figure 3 shows the function \(g_\lambda\) and \(\partial g_\lambda\). The parameter \(\lambda\) adds a constant offset to the positive values of \(\partial g_\lambda(x)\) and subtracts the same value for all negative values. It is clear from Figure 3 that in \((-\infty, -\sqrt{\mu}]\) and \([\sqrt{\mu}, \infty)\) the subdifferential contains a single element. In addition for any two elements \(x'', x'\) in one of these intervals we have

\[
\langle \partial g_\lambda(x''), \partial g_\lambda(x') \rangle = \langle x'', x' \rangle^2.
\]

We are now ready to prove that stationary points of (14) can be separated in terms of the cardinality of their difference. The result requires that the elements of the vector \(z'\) are not too close to the threshold \(\sqrt{\mu} + 2\sqrt{\mu}\). This condition is a natural restriction since the vector \(z\) is related to the noise of the problem [13], and for very large noise levels we can expect that there will be multiple solutions.

**Theorem 3.** Assume that \(x'\) is a stationary point with \(2z' \in \partial g_\lambda(x')\) and that each element of \(z'\) fulfills \(|z'_i| \notin [(1 - \delta_K)\sqrt{\mu} + \frac{\lambda}{2}, \sqrt{\mu} + \frac{\lambda}{2}]\). If \(x''\) is another stationary point then \(|x'' - x'||_0 > K\). If, in addition, \(|x'||_0 < \frac{b}{2}\) then

\[
x' \in \arg\min_{\|x\| \leq \frac{b}{2}} r_{\mu, \lambda}(x) + \|Ax - b\|^2.
\]

**Proof.** Assume that \(|x'' - x'||_0 \leq K\). Since \(x'\) is a stationary point we have \(2z' + \nabla h(x') = 0\), where \(2z' \in \partial g_\lambda(x')\).

We first show that \(2z'' + \nabla h(x'') \neq 0\) if \(2z'' \in \partial g_\lambda(x'')\). If \(v = x'' - x'\) has \(|v|_0 \leq K\) then according to Lemma 2 and (33) we have

\[
\langle 2z'' + \nabla h(x''), v \rangle = \langle 2(z'' - z'), v \rangle + \langle \nabla h(x + v) - \nabla h(x), v \rangle > 0.
\]

Thus \(x''\) cannot be stationary.

Since \(g_\lambda\) is convex its directional derivative exists and is given by

\[
g'_{\mu, \lambda}(x) = \max_{2z' \in \partial g_\lambda(x)} \langle 2z', v \rangle.
\]

Since \(h'(x) = \langle \nabla h(x), v \rangle\), (35) shows that \(g'_{\mu, \lambda}(x' + tv) + h'(x' + tv) > 0\) for any \(t > 0\) if \(|v|_0 \leq K\). Now, if \(|x'|_0 < \frac{b}{2}\) then \(g_\lambda + h\) is increasing on every line segment between \(x'\) and any other point \(x''\) with \(|x''|_0 \leq \frac{b}{2}\).

\[
5. Optimization
\]

In this section we present a simple algorithm for optimizing objective functions of the type (14). We restrict ourselves to sparsity problems since the same approach works for rank regularization with minimal changes.

**5.1. Algorithm**

For optimization we employ the popular ADMM [2] approach. This is a splitting scheme that uses two copies of the \(x\) and enforces them to be equal using dual variables. The augmented Lagrangian for the problem is

\[
L(x, y, \eta) = r_{\mu, \lambda}(x) + \rho\|x - y + \eta\|^2 + \|Ay - b\|^2 - \rho\|\eta\|^2.
\]

In each iteration \(t\) of ADMM the variable updates are given by

\[
x_{t+1} = \arg\min_x r_{\mu, \lambda}(x) + \rho\|x - y_t + \eta_t\|^2,
\]

\[
y_{t+1} = \arg\min_y \rho\|x_{t+1} - y + \eta_t\|^2 + \|Ay - b\|^2,
\]

\[
\eta_{t+1} = \eta_t + x_{t+1} - y_{t+1}.
\]
The update in equation (39) is a simple least squares minimization with closed form solution. Problem (38) is solved using the proximal operator of $r_{\mu,\lambda}$ which we present in the next section. We refer to [2] for details on ADMM.

5.2. The proximal operator

To solve (43) we can use the proximal operator
\[
\text{prox}_{\frac{r_{\mu,\lambda}}{\rho}}(y) = \arg\min_x \frac{1}{\rho} r_{\mu,\lambda}(x) + \|x - y\|^2. \quad (41)
\]

The following result (which is proven in Appendix D) shows that in general the proximal operator of $Q_\gamma(f + \lambda \| \cdot \|_1)$ is easy to compute if the proximal operator of $Q_\gamma(f)$ is known.

**Proposition 1.** Let $f : \mathbb{R}^d \to \mathbb{R}$ be a lower semicontinuous sign-invariant function such that $f(0) = 0$ and $f(x + y) \geq f(x)$ for every $x,y \in \mathbb{R}^d$. Then
\[
\text{prox}_{Q_\gamma(f + \lambda \| \cdot \|_1)/\rho}(y) = \text{prox}_{Q_\gamma(f)/\rho}(\text{prox}_{\lambda \| \cdot \|_1/\rho}(y))
\]
for every $y \in \mathbb{R}^d$.

For our case $f(x) = \mu \|x\|_0$ the proximal operator is separable and each element of the vector $x$ can be treated independently. Adding soft thresholding to the results of [19] we get that for $\rho > 1$ the proximal solution is given by
\[
z_i = \begin{cases} 
\max(y_i - \lambda/2\rho, 0) & y_i \geq 0 \\
\min(y_i + \lambda/2\rho, 0) & y_i \leq 0 
\end{cases} \quad (43)
\]
and
\[
x_i = \begin{cases} 
\frac{z_i}{\rho} & \frac{\sqrt{\mu}}{\rho + 1} \leq |z_i| \\
\frac{\sqrt{\mu}}{\rho + 1} & |z_i| \leq \frac{\sqrt{\mu}}{\rho + 1} \\
0 & |z_i| \leq \frac{\sqrt{\mu}}{\rho + 1} 
\end{cases} 
\]

6. Experiments

In this section we test the proposed formulation on a number of real and synthetic experiments. Our focus is to evaluate the proposed method’s robustness to local minima and the effects of its shrinking bias.

6.1. Random Matrices

In this section we compare the robustness to local minima of the relaxations (2), (5) and (14). Note that (2) and (5) are special cases of (14), obtained by letting $\lambda$ or $\mu$ equal 0 (by Theorem 1).

We generated $A$-matrices of size $100 \times 200$ by drawing the columns from a uniform distribution over the unit sphere $S^{99}$ in $\mathbb{R}^{100}$, and the vector $x_0$ was selected to have 10 random nonzero elements with random magnitudes between 2 and 4, resulting in $\|x_0\| \approx 10$. We then computed $b = Ax_0 + \epsilon$ for different values of random noise with $\|\epsilon\|$ ranging from 0 to 5. For (5) we used $\mu = 1$ and for (2) we used $\lambda_{\ell_1} = 2 \sqrt{2 \log(200)} \|\epsilon\| \approx 0.5 \|\epsilon\|$; see [13] for the rationale behind these choices. For (14) we again chose $\mu = 1$ but used $\lambda = \lambda_{\ell_1}/6$. Figure (4) plots $\|x - x_S\|$ for the estimated $x$ with the three methods, as a function of $\|\epsilon\|$. Both (5) and (14) do better than traditional $\ell_1$ in the entire range, (5) finds $x_S$ with 100% accuracy until around $\|\epsilon\| \approx 3$, where (14) starts to perform better. This is likely due to the fact that the small $\ell_1$ term helps the (non-convex) method (14) to not get stuck in local minima. To test this conjecture, we ran the same experiment for 50 iterations for the fixed noise level $\|\epsilon\| = 3.5$ and chose as initial point the least squares solution $\arg\min_x \|Ax - b\|^2$, which is known to be close to many local minima. The histograms to the right in Figure 4 show the cardinality of the found solution.

![Figure 4](image-url)

**Figure 4.** Left: Noise level $\|\epsilon\|$ versus distance $\|x - x_S\|$ between the obtained solution $x$ and the oracle solution $x_S$ for the three methods (5), (14) and (2). Right, histogram of the number of non-zero elements in the obtained solutions.

6.2. Point-set Registration with Outliers

Next we consider registration of 2D point clouds. We assume that we have a set of model points $\{p_i\}_{i=1}^N$ that should be registered to $\{q_i\}_{i=1}^N$ by minimizing
\[
\sum_{i=1}^N \|sR(p_i + t - q_i)\|^2. \quad (45)
\]

Here $sR$ is a scaled rotation of the form $\begin{pmatrix} a & -b \\ b & a \end{pmatrix}$ and $t \in \mathbb{R}^2$ is a translation vector. Since the residuals are linear in the parameters $a, b, t$, we can by column-stacking them write the problem as $\|My - v\|^2$, where the vector $y$ contains the unknowns $a, b, t$. We further assume that the point matches contain outliers that needs to be removed. Therefore we add a sparse vector $x$ whose non-zero entries allows the solution to have large errors. We thus want to solve
\[
\min_{x,y} \mu \|x\|_0 + \|My - v + x\|^2. \quad (45)
\]
The minimization over $y$ can be carried out in closed form by noting that $y = (M^TM)^{-1}MT(v - x)$. Inserting into (45) which gives the objective function (4), where $A = I - M(M^TM)^{-1}MT$ and $b = Av$. The matrix $A$ is a
projection onto the complement of the column space of $M$, and therefore has a 4 dimensional null space.

Figure 5 shows the results of a synthetic experiment with 500 problem instances. The data was generated by first selecting 100 random Gaussian 2D points. We then divided these into two groups of 60 and 40 respectively and transformed these using two different random similarity transformations. This way the data supports two strong hypotheses which yields a problem which is much more difficult than what adding random uniformly distributed outliers does. The transformations were generated by taking $a$ and $b$ to be Gaussian with mean 0 and variance 1, and selecting $t$ to be 2D-Gaussian with mean $(0, 0)$ and covariance $5I$. We compare the three relaxations (2) with $\lambda = 2$, (5) with $\mu = 1$ and (14) with $\mu = 1$ and $\lambda = 0.5$. (The reason for using $\lambda = 2$ in (2) and $\mu = 1$ in (5) is that this gives the same threshold in the corresponding proximal operators.)

All methods where initialized with the least squares solution $\min_{x} \|Ax - b\|^2$. In the left histogram of Figure 5 we plot the data fit with respect to the inlier residuals (corresponding to the first 60 points, that supports the larger hypothesis). In the right one we plot the number of residuals determined to be outliers. When starting from the least squares initialization the formulation (5) frequently gets stuck in solutions with poor data fit that are dense and close to the least squares solution. However when it converges to the correct solution it gives a much better data fit then the $\ell^1$ norm formulation (2) due to its lack of bias. The added $\ell^1$ term to helps (14) converge to the correct solution with a good data fit. Note that the number of outliers are in many cases smaller than 40 due to the randomness of the data.

We also include a few problem instances with real data. Here we matched SIFT descriptors between two images, as shown in Figure 6, to generate the two point sets $\{p_i\}_{i=1}^N$ and $\{q_i\}_{i=1}^N$. We then registered the point sets using the formulations (5) with $\mu = 20^2$ and (2) with $\lambda = 10$ (which in both cases corresponds to a 20 pixel outlier threshold in a $3072 \times 2048$ image). For (14) we used $\mu = 20^2$ and $\lambda = 5$.

The results are shown in Figure 6. In the first problem instance (first row) we used an image which generates one strong hypothesis. Here both (14) and (2) produce good results. In contrast (5) immediately gets stuck in the least squares solution for which all residuals are above the threshold. In the second instance there are two strong hypotheses. The incorrect one introduces a systematic bias that effects (2) more than (14). As a result the registration obtained by (14) is better than that of (5) and the number of determined inliers is larger.

Figure 7. Results from the two real registration experiment. From left to right: (5), (14), (2). Red means that the point was classified as outlier, green inlier. White frame shows registration of the model book under the estimated transformation.

6.3. Non-rigid Structure from Motion

In our final experiment we consider non rigid structure form motion with a rank prior. We follow the aproach of Dai et al. [14] and let

$$X = \begin{bmatrix} X_1 \\ Y_1 \\ Z_1 \\ \vdots \\ X_F \\ Y_F \\ Z_F \end{bmatrix} \quad \text{and} \quad X^\# = \begin{bmatrix} X_1 & Y_1 & Z_1 \\ \vdots & \vdots & \vdots \\ X_F & Y_F & Z_F \end{bmatrix} , \quad (46)$$

where $X_i, Y_i, Z_i$ are $1 \times m$ matrices containing the $x$-$y$-$z$-coordinates of tracked image points in frame $i$. With an orthographic camera the projection of the 3D points can be written $M = RX$, where $R$ is a $2F \times 3F$ block diagonal
matrix with $2 \times 3$ blocks $R_i$, consisting of two orthogonal rows that encode the camera orientation in image $i$. The resulting $2F \times m$ measurement matrix $M$ consists of the $x$- and $y$-image coordinates of the tracked points. Under the assumption of a linear shape basis model [4] with $r$ deformation modes, the matrix $X^\#$ can be written $X^\# = CB$, where $B$ is $r \times 3n$, and therefore $\text{rank}(X^\#) = r$. We search for the matrix $X^\#$ of rank $r$ that minimizes the residual error $\|PX - M\|_F^2$.

The linear operator defined by $A(X^\#) = RX$ does by itself not obey (7) since there are matrices of rank 1 in its nullspace, see [25].

In Figure 9 we compare the three relaxations

$$r_{0,\mu}(\sigma(X^\#)) + \|RX - M\|_F^2,$$  \hspace{1cm} (47)
$$r_{\mu,\lambda}(\sigma(X^\#)) + \|RX - M\|_F^2,$$  \hspace{1cm} (48)
$$2\sqrt{\mu}\|X^\#\|_* + \|RX - M\|_F^2,$$  \hspace{1cm} (49)

on the four MOCAP sequences displayed in Figure 8, obtained from [14]. These consist of real motion capture data and therefore the ground truth solution is only approximatively of low rank. Figure 9 shows results for the three methods. We solved the problem for 50 values of $\sqrt{\mu}$ between 10 and 100 (orange curve) and computed the resulting rank and datafit. (For (48) we kept $\lambda = 5$ fixed.) All three formulations were given the same (random) starting solution.

The same tendencies are visible for all four sequences. While (47) generally gives a better data fit than (49), due to the nuclear norms shrinking bias, the distance to the ground truth is larger for low values of $\mu$ or equivalently large ranks where the problem gets ill posed. The relaxation (48) consistently outperforms (49) both in terms of data fit and distance to ground truth. In addition its performance is similar to (47) for high values of $\mu$ while it does not exhibit the same unstable behavior for high ranks.

References

[1] T. Blumensath and M. E. Davies, Iterative thresholding for sparse approximations. Journal of Fourier analysis and Ap-
[2] M. Carlsson, D. Gerosa, and C. Olsson. An unbi-
E. J. Candes and T. Tao. Near-optimal signal recovery
J. Fan and R. Li. Variable selection via nonconcave penalized
J. Fan, L. Xue, and H. Zou. Strong oracle optimality of
Y. Hu, D. Zhang, J. Ye, X. Li, and X. He. Fast and accurate
IEEE Transactions on Pattern Analysis and Machine Intelli-
gence, 35(9):2117–2130, 2013. 2

[8] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Dis-
K. Bredies, D. A. Lorenz, and S. Reiterer. Minimization
Y. Dai, H. Li, and M. He. A simple prior-free method for
E. J. Cand `es, J. Romberg, and T. Tao. Robust uncertainty
E. J. Cand `es and B. Recht. Exact matrix completion via con-
E. J. Cand `es, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? J. ACM, 58(3):11:1–11:37, 2011. 2
E. J. Cand `es and B. Recht. Exact matrix completion via con-
E. J. Cand `es, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incom-
E. J. Candes, J. K. Romberg, and T. Tao. Stable signal re-
E. J. Candes and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? IEEE transactions on information theory, 52(12):5406–5425, 2006. 1
E. J. Candes, M. B. Wakin, and S. P. Boyd. Enhancing spar-
M. Carlsson. On convexification/optimization of func-
M. Carlsson, D. Gerosa, and C. Olsson. An unbi-
Y. Dai, H. Li, and M. He. A simple prior-free method for
D. L. Donoho and M. Elad. Optimally sparse representation in general (non-orthogonal) dictionaries via \ell^1 minimization. In PROC. NATLACAD. SCI. USA 100 2197–2002, 2002. 1
J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American statistical Association, 96(456):1348–1360, 2001. 1
J. Fan, L. Xue, and H. Zou. Strong oracle optimality of folded concave penalized estimation. Annals of statistics, 42(3):819, 2014. 1
[18] Y. Hu, D. Zhang, J. Ye, X. Li, and X. He. Fast and accurate matrix completion via truncated nuclear norm regularization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(9):2117–2130, 2013. 2

[19] V. Larsson and C. Olsson. Convex low rank approximation. International Journal of Computer Vision, 120(2):194–214, 2016. 2, 6
[20] P.-L. Loh and M. J. Wainwright. Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for lo-
cal optima. In Advances in Neural Information Processing Systems, pages 476–484, 2013. 1
[21] P.-L. Loh, M. J. Wainwright, et al. Support recovery without
[22] R. Mazumder, J. H. Friedman, and T. Hastie. Sparsenet: Coordinate descent with nonconvex penalties. Journal of the American Statistical Association, 106(495):1125–1138, 2011. 1
[23] K. Mohan and M. Fazel. Iterative reweighted least squares for matrix rank minimization. In Annual Allerton Conference on Communication, Control, and Computing, pages 653–661, 2010. 2
[24] T. H. Oh, Y. W. Tai, J. C. Bazin, H. Kim, and I. S. Kweon. Partial sum minimization of singular values in robust pca: Algorithm and applications. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(4):744–758, 2016. 2
[25] C. Olsson, M. Carlsson, F. Andersson, and V. Larsson. Non-
[26] S. Oymak, A. Jalali, M. Fazel, Y. C. Eldar, and B. Hassibi. Simultaneously structured models with application to sparse and low-rank matrices. IEEE Transactions on Information Theory, 61(5):2886–2908, 2015. 2
[27] S. Oymak, K. Mohan, M. Fazel, and B. Hassibi. A simplified approach to recovery conditions for low rank matrices. In IEEE International Symposium on Information Theory Proceedings (ISIT), pages 2318–2322, 2011. 2
[28] Z. Pan and C. Zhang. Relaxed sparse eigenvalue conditions for sparse estimation via non-convex regularized regression. Pattern Recognition, 48(1):231–243, 2015. 1
[29] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-
rank solutions of linear matrix equations via nuclear norm minimization. SIAM Rev., 52(3):471–501, Aug. 2010. 2
[30] J. A. Tropp. Just relax: Convex programming methods for identifying sparse signals in noise. IEEE transactions on information theory, 52(3):1030–1051, 2006. 1
[31] J. A. Tropp. Convex recovery of a structured signal from independent random linear measurements. In Sampling Theory, a Renaissance, pages 67–101. 2015. 1
[32] Z. Wang, H. Liu, and T. Zhang. Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. Annals of statistics, 42(6):2164, 2014. 1
[33] C.-H. Zhang and T. Zhang. A general theory of concave reg-
[34] H. Zou and R. Li. One-step sparse estimates in non-

[1] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American statistical Association, 96(456):1348–1360, 2001. 1
[17] J. Fan, L. Xue, and H. Zou. Strong oracle optimality of folded concave penalized estimation. Annals of statistics, 42(3):819, 2014. 1
A. Proof of Theorem 1

Proof. We need to prove that
\[
(f(y) + \lambda \|y\|_1)'' = f''(y) + \lambda \|y\|_1, \tag{50}
\]
where $\ast$ denotes the Fenchel conjugate, i.e. $g^*(x) = \sup_y \langle x, y \rangle - g(y)$. For a general function $g$ the convex envelope can be computed using the bi-conjugate $g^{**}$.

By symmetry it suffices to consider $y \in \mathbb{R}^d_+$. First notice that in
\[
(f(\cdot) + \lambda \cdot \|\cdot\|_1)''(y) = \sup_x \langle x, y \rangle - (f(x) + \lambda \|x\|_1), \tag{51}
\]
only the term $\langle x, y \rangle$ depends on the signs of the elements of $x$. It is clear that any maximizing $x$ will have $\text{sign}(x_i) = \text{sign}(y_i)$ Therefore we may assume without loss of generality that $x \in \mathbb{R}^d_+$ as well. We now have $\|x\|_1 = \langle x, 1 \rangle$ which reduces (51) to
\[
\sup_{x \in \mathbb{R}^d_+} \langle x, y - \lambda 1 \rangle - f(x). \tag{52}
\]
Note that if $y_j - \lambda < 0$ for some $j$, then for every $x \in \mathbb{R}^d_+$ we have
\[
\langle x - e_j x_j, y - \lambda 1 \rangle - f(x - x_j e_j) \geq \langle x, y - \lambda 1 \rangle - f(x), \tag{53}
\]
where $e_j$ is the $j$th vector of the canonical basis. Therefore we introduce the set $S = \{i : y_i < \lambda \}$ and the notation $\chi_S x = \sum_{k \in S} e_k x_k$. We then have
\[
\sup_{x \in \mathbb{R}^d_+} \langle x, y - \lambda 1 \rangle - f(x) = \sup_{x \in \mathbb{R}^d_+} \langle \chi_S x, y - \lambda 1 \rangle - f(x) \tag{54}
\]
\[
= \sup_{x \in \mathbb{R}^d} \langle x, \chi_S (y - \lambda 1) \rangle - f(x) = f''((y - \lambda 1)_+),
\]
where $(x)_+$ denotes thresholding at 0, that is, $(x)_+ = (\max(x_1, 0), \ldots, \max(x_d, 0))$.

To compute the second Fenchel-conjugate, first note that $f''(x + v) \geq f''(x)$ for $x, v \in \mathbb{R}^d_+$ since
\[
\langle y, x \rangle - f(y) - \|y\|_1 \leq \langle y, x + v \rangle - f(y) - \|y\|_1 \tag{55}
\]
for all $y \in \mathbb{R}^d_+$. Therefore we assume wlog $y \in \mathbb{R}^d_+$. Note that the supremum in $\sup_{x \in \mathbb{R}^d}(\langle x, y \rangle - f''((x - \lambda 1)_+))$ is clearly attained for an $x$ with $x_j \geq \lambda$ for all $1 \leq j \leq d$. By this observation we get $\langle f + \lambda \|\cdot\|_1 \rangle''(y) = \sup_{x \geq \lambda} \langle x, y \rangle - f''((x - \lambda 1)_+) \tag{56}$
\[
= \sup_{x \geq \lambda} \langle z + \lambda 1, y \rangle - f''(z) = \lambda\|y\|_1 + \sup_{z \in \mathbb{R}^d_+} \langle z, y \rangle - f''(z), \tag{56}
\]
which shows that $(f + \lambda \|\cdot\|_1)''(y) = \lambda\|y\|_1 + f''(y)$. \(\square\)

Since the proof for the matrix case looks somewhat different we present it here. It uses properties of the $\mathcal{S}_\gamma$-transform [12] which is defined as
\[
\mathcal{S}_\gamma(f)(x) = (f + \frac{\gamma}{2}\|\cdot\|^2)''(\gamma x) - \frac{\gamma}{2}\|x\|^2 \tag{57}
\]
The connection to $\mathcal{Q}_\gamma$ is that $\mathcal{Q}_\gamma(f)(x) = \mathcal{S}_\gamma(S_f(f)(x))$. Hence it is enough to show that two functions have the same $\mathcal{S}_\gamma$ transform to show that they have the same $\mathcal{Q}_\gamma$ transform.

Proposition 2. Suppose that $f$ is a permutation and sign invariant $[0, \infty]$-valued functional on $\mathbb{R}^d$, $d = \min(n_1, n_2)$, and that $F(X) = f(\sigma(X))$, $X \in \mathbb{R}^{n_1 \times n_2}$. Then
\[
\mathcal{S}_\gamma(F)(Y) = \mathcal{S}_\gamma(f)(\sigma(Y)). \tag{58}
\]
Proof. From (57) it can be seen (by completing squares) that
\[
\mathcal{S}_\gamma(F)(Y) = \left( \sup_{\nu \geq \nu_2 \geq \ldots} - f(\nu) - \frac{\gamma}{2}\|\nu - \sigma(Y)\|^2 \right). \tag{59}
\]
Due to the permutation and sign invariance of $f$, we can drop the restrictions on $\nu$ which turns (59) into
\[
\mathcal{S}_\gamma(f)(\sigma(Y)). \tag{59}
\]
\[\square\]

B. Minimizers of (28)

In this section we compute the minimizers of the one dimensional problem
\[
\min_{x_i} \mu - (\max(\sqrt{\mu} - |x_i|, 0))^2 + \lambda|x_i| + (x_i - z_i)^2. \tag{60}
\]
Because of symmetry of $(\max(\sqrt{\mu} - |x_i|, 0))^2 + \lambda|x_i|$ it is clear that the optimal $x_i \geq 0$ if $z_i \geq 0$. We therefore first consider
\[
\mu - (\max(\sqrt{\mu} - x_i, 0))^2 + \lambda|x_i| + (x_i - z_i)^2, \tag{61}
\]
on $x_i \geq 0$. It is clear that this function is differentiable on $x_i > 0$ and goes to infinity when $x_i \to \infty$. Therefore the optimizer is either in a stationary point in $x_i > 0$ or in $x_i = 0$. Dividing into the two cases $0 < x_i < \sqrt{\mu}$ and $x_i \geq \sqrt{\mu}$ and differentiating shows that:
• There are no stationary points if \( z_i < \sqrt{\mu + \frac{\lambda}{2}} \). Thus the minimum is in \( x_i = 0 \).

• All points in \((0, \sqrt{\mu})\) are stationary and have objective value \( z_i^2 \) if \( z_i = \sqrt{\mu + \frac{\lambda}{2}} \). Since \( x_i = 0 \) gives the same objective value any \( x_i \in [0, \sqrt{\mu}] \) will be optimal.

• If \( z_i > \sqrt{\mu + \frac{\lambda}{2}} \) then the point \( x_i = z_i - \frac{\lambda}{2} \) is stationary with objective value

\[
\mu + \lambda z_i - \frac{\lambda^2}{4} < \left( \sqrt{\mu + \frac{\lambda}{2}} \right)^2 < z_i^2.
\]

Since \( x_i = 0 \) gives the objective \( z_i^2 \) the optimizer is \( x_i = z_i - \frac{\lambda}{2} \).

Because of symmetry we now get that the minimizers of (28) is given by This resulting minimizer is given by

\[
\{ z_i - \text{sign}(z_i) \frac{\lambda}{2} \mid |z_i| > \sqrt{\mu + \frac{\lambda}{2}} \times (0, \sqrt{\mu}) \text{sign}(z_i) \mid |z_i| = \sqrt{\mu + \frac{\lambda}{2}} \times \{0\} \mid |z_i| < \sqrt{\mu + \frac{\lambda}{2}} \}
\]

C. Proof of Lemma 2

Proof. We first consider the scalar case \( 2z' \in \partial g_\lambda(x') \) where \( z' \geq 0 \). If \( z' > \frac{\sqrt{\mu}}{1 - \delta_K} + \frac{\lambda}{2} \) then in view of (29) we have \( x' = z' - \frac{\lambda}{2} > \frac{\sqrt{\mu}}{1 - \delta_K} \). Now consider the linear function

\[
l(x) = \delta_K (x - x') + z' = \delta_K x + (1 - \delta_K)x' + \frac{\lambda}{2}.
\]

Since \( l(x') = z' \) and \( l(0) = (1 - \delta_K)x' + \frac{\lambda}{2} > \sqrt{\mu + \frac{\lambda}{2}} \) it is clear from Figure 3 that \( l(x') > z'' \) for all \( x'' < x' \). Therefore

\[
z'' - z' > l(x'') - z' = \delta_K (x'' - x'),
\]

for all \( x'' < x' \). Additionally, for \( x'' > x' \) we clearly have

\[
z'' - z' = x'' - x' > \delta_K (x'' - x').
\]

Now assume that \( 2z' \in \partial g_\lambda(x') \) where \( 0 \leq z' \leq (1 - \delta_K)\sqrt{\mu + \frac{\lambda}{2}} \). Then because of (29) \( x' = 0 \). We let \( l(x) = \delta_K x + z' \). Since \( l(0) = z' < \sqrt{\mu + \frac{\lambda}{2}} \) and

\[
l(\sqrt{\mu}) = \delta_K \sqrt{\mu} + z' < \delta_K \sqrt{\mu} + (1 - \delta_K)\sqrt{\mu} + \frac{\lambda}{2}
\]

it is clear that \( l(x'') < z'' \) for all \( x'' > 0 \). Therefore

\[
z'' - z' > l(x'') - z' = \delta_K x'' - \delta_K (x'' - x').
\]

Similarly, it is easy to see that \( l(x'') > z'' \) if \( x'' < 0 \) and therefore

\[
z' - z'' > z' - l(x'') = -\delta_K x'' = \delta_K (x' - x'').
\]

which shows that

\[
(z'' - z')(x'' - x') > \delta_K (x'' - x')^2.
\]

Because of symmetry of \( \partial g_\lambda \) we conclude that the same holds if \( z < 0 \). To obtain (31) we now sum over the non-zero entries of \( x \).

We conclude this section by noting that the proof of the corresponding lemma for the matrix case is somewhat more complicated since the U and V matrices of the singular value decomposition of X have to be accounted for. This can be done by combining the above result with Lemma 4.1 of [25].

D. Proof of Proposition 1

Proof. It is enough to compute the proximal operator of the function \( S^2_\gamma(f)(\cdot) + \lambda \| \cdot \|_1 \). Wlog we assume that \( y \in \mathbb{R}^d_+ \). With the same notation as in the proof of Theorem 1 we have

\[
\frac{\text{prox}_{\gamma}(f)(y)}{\rho} = \frac{\text{arg min}_{x \in \mathbb{R}^d} \| \gamma x \|_1 + \| x - y \|_2^2}{\rho} \quad \text{and} \quad \frac{\text{arg min}_{x \in \mathbb{R}^d} \| \gamma x \|_1 + \| x - y \|_2^2}{\rho}.
\]

Since \( \| \gamma x \|_1 + \| x - y \|_2^2 = \| x - (\frac{\lambda}{2\rho} 1)_+ \|_2^2 \) and \( y \) is constant in the minimization of \( x \) we see that \( x \) also solves

\[
\frac{\text{arg min}_{x \in \mathbb{R}^d} \| \gamma x \|_1 + \| x - y \|_2^2}{\rho}.
\]

Note that \( y - \frac{\lambda}{2\rho} 1 \) is the \( \text{prox}_{\gamma}(f)(y) \) since \( y \in \mathbb{R}^d_+ \). Also, since the elements of \( y - \frac{\lambda}{2\rho} 1 \) are non-negative it is clear that minimizing over \( x \in \mathbb{R}^d \) instead of \( \mathbb{R}^d_+ \) does not change the optimizer and therefore

\[
\frac{\text{prox}_{\gamma}(f)(y)}{\rho} = \frac{\text{prox}_{\gamma}(f)(y)}{\rho} = \frac{\text{prox}_{\gamma}(f)(y)}{\rho}.
\]