Journal of Mathematical Imaging and Vision (2022) 64:379–394
https://doi.org/10.1007/s10851-022-01071-5

Bias Versus Non-Convexity in Compressed Sensing

Daniele Gerosa[1] · **Marcus Carlsson**[1] · **Carl Olsson**[1,2]

Received: 7 April 2021 / Accepted: 13 December 2021 / Published online: 8 March 2022

Abstract
Cardinality and rank functions are ideal ways of regularizing under-determined linear systems, but 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 ℓ^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. This well-known drawback has given rise to a fauna of non-convex alternatives, which usually features better global minima at the price of maybe getting stuck in undesired local minima. We focus in particular penalties based on the quadratic envelope, which have been shown to have global minima which even coincide with the "oracle solution," i.e., there is no bias at all. So, which one do we choose, convex with a definite bias, or non-convex with no bias but less predictability? In this article, we develop a framework which allows us to interpolate between these alternatives; that is, we construct sparsity inducing penalties where the degree of non-convexity/bias can be chosen according to the specifics of the particular problem.

Keywords Compressed sensing · Quadratic envelopes · Non-convex optimization

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

$$\min_{\mathbf{x}} \mu\|\mathbf{x}\|_0 + \|A\mathbf{x} - \mathbf{b}\|^2, \tag{1}$$

where $\|\mathbf{x}\|_0$ is the number of nonzero elements of $\mathbf{x}$. Optimization of (1) is difficult since the term $\|\mathbf{x}\|_0$ is non-convex and discontinuous at any point containing entries that are zero, which in particular applies to the sought *sparse* solution.

✉ Daniele Gerosa
daniele.gerosa@math.lu.se

Marcus Carlsson
marcus.carlsson@math.lu.se

Carl Olsson
carl.olsson@math.lth.se

[1] Centre for Mathematical Sciences and LTH, Lund University, Lund, Sweden

[2] Department of Electrical Engineering, Chalmers University of Technology, Göteborg, Sweden

A common practice is to replace $\|\mathbf{x}\|_0$ with the ℓ^1-norm, resulting in the convex relaxation (LASSO)

$$\min_{\mathbf{x}} \lambda\|\mathbf{x}\|_1 + \|A\mathbf{x} - \mathbf{b}\|^2. \tag{2}$$

However, it has been observed that (2) suffers from a shrinking bias [18,23], since the ℓ^1 term not only has the (desired) effect of forcing many entries in $\mathbf{x}$ to 0, but also the (undesired) effect of diminishing the size of the nonzero entries. This has led to a large amount of non-convex alternatives to replace the ℓ^1-penalty, see, e.g., [3,4,6,10,19,21,22,24,25,30–32]. Typically these come without global convergence guarantees. In [13], however, a non-convex alternative that provides optimality guarantees is studied. These papers propose to replace the term $\mu\|\mathbf{x}\|_0$ with $\mathcal{Q}_2(\mu\|\cdot\|_0)(\mathbf{x})$, where $\mathcal{Q}_2(f)$ is the so-called *quadratic envelope of* f, a functional transform studied in [12]. We recall here the definition of quadratic envelope [12]:

Definition 1.1 (*Quadratic envelope*) Let be $\mathcal{V}$ a real Hilbert space and $f : \mathcal{V} \to \mathbb{R}$ a functional. The *quadratic envelope of* f is defined as

$$\mathcal{Q}_2(f)(\mathbf{x}) = \sup_{\alpha\in\mathbb{R},\, \mathbf{v}\in\mathcal{V}} \left\{\alpha - \|\mathbf{x}-\mathbf{v}\|^2 \,:\, \alpha - \|\mathbf{x}-\mathbf{v}\|^2 \le f(\mathbf{x})\right\}.$$

It can be shown (Theorem 3.1 in [12]) that $\mathcal{Q}_2(f) + \|\cdot\|^2$ is the convex envelope of $f + \|\cdot\|^2$; this is useful for concrete calculations. For $f(\mathbf{x}) = \mu\|\mathbf{x}\|_0$, we obtain the objective

$$\mathcal{Q}_2(\mu\|\cdot\|_0)(\mathbf{x}) + \|A\mathbf{x} - \mathbf{b}\|^2 = \sum_i \mu - \max(\sqrt{\mu} - |x_i|, 0)^2 + \|A\mathbf{x} - \mathbf{b}\|^2 \tag{3}$$

where $\mathcal{Q}_2(\mu\|\cdot\|_0)(\mathbf{x})$ coincides, in fact, with the so-called minimax concave penalty (MCP) [34]; calculation details can be found in [11], Example 2.4. In [8], it was argued that the so-called oracle solution is the best one could possibly wish for, which is what we get if we minimize $\|A\mathbf{x} - \mathbf{b}\|^2$ over the "true" support of $\mathbf{x}_0$. It was shown in [13] that the oracle solution often is a global minimizer of (3), and moreover, that it is unique as a sparse minimizer, i.e., any local minimizer will necessarily have higher cardinality. This is true under the LRIP assumption (lower restricted isometry property, see [2]) on A which states that there should be a positive constant δ_K^- sufficiently close to 0 such that

$$(1 - \delta_K^-)\|\mathbf{x}\|^2 \leq \|A\mathbf{x}\|^2, \tag{4}$$

for all vectors $\mathbf{x}$ with $\|\mathbf{x}\|_0 \leq K$, and hence, this is a weaker assumption than the standard RIP estimates (see, e.g., [8]). It is noteworthy that the LRIP condition is not only less stringent than RIP, and the estimates for the corresponding constant are also easier to satisfy for the results in [13] to be valid. The same holds true for the present paper, where we will show similar results for a class of penalties intermediate between $\lambda\|\cdot\|_1$ and $\mathcal{Q}_2(\mu\|\cdot\|_0)$.

Before outlining the details, let us mention that there is a parallel theory for low-rank matrices. In this setting, we are seeking to minimize

$$\mu\ \mathrm{rank}(X) + \|\mathcal{A}X - \mathbf{b}\|_F^2 \tag{5}$$

$\mathcal{A} : \mathbb{R}^{m\times n} \to \mathbb{R}^p$ is a linear operator. Here the standard approach relies on replacing $\mu\ \mathrm{rank}(X)$ with the nuclear norm $\|X\|_*$, which is the ℓ^1-norm applied to the singular values $\sigma(X)$ of a given matrix X [7,26], whereas [24] proposed solving instead

$$\sum_i \mu - \max(\sqrt{\mu} - \sigma_i(X), 0)^2 + \|\mathcal{A}X - \mathbf{b}\|_F^2, \tag{6}$$

and showed a number of desirable features, which was further strengthened in [14]. As for the vector case, this paper provides penalties in between the two extremes.

While the non-convex relaxations (3) and (6) provide unbiased alternatives to the ℓ^1/nuclear norms which can be shown to only have one sparse/low-rank stationary point ([13,14,24]), it is clear that there always will be poor local minimizers. To see this, let $\mathbf{x}_h$ be a dense vector from the nullspace of A and $\mathbf{x}_p$ a minimizer of $\|A\mathbf{x} - \mathbf{b}\|$. Then by rescaling $\mathbf{x}_h$ so that all the elements of $\mathbf{x}_p + \mathbf{x}_h$ have magnitude strictly larger than $\sqrt{\mu}$ we obtain a vector that minimizes the data fit while the regularization $\mathcal{Q}_2(\mu\|\cdot\|_0)$ is (locally) constant around it.

We recall that (3) and (6) are usually solved with iterative solvers such as forward–backward splitting (FBS) or alternating direction method of multipliers (ADMM), which often are initialized by $\mathbf{0}$ or some rough approximation of the desired solution. We introduce the somewhat non-stringent concept *convergence basin*, by which we mean the set of initial points which lead to the global minimizer, without further specifying which algorithm or parameter choice is used. For example, the point $\mathbf{x}_p + \mathbf{x}_h$ above (and any point near it) lies outside the "convergence basin." In contrast, (2) (and its matrix counterpart) has the whole space as convergence basin. To summarize, the non-convex penalties enjoy better properties of the global minimizer but could have a small convergence basin, leading to suboptimal performance in practice.

To find a good trade-off between the benefits of both methods, we introduce here a sort of crossover. We will study relaxations of

$$\mu\|\mathbf{x}\|_0 + \lambda\|\mathbf{x}\|_1 + \|A\mathbf{x} - \mathbf{b}\|^2, \tag{7}$$

and

$$\mu\ \mathrm{rank}(X) + \lambda\|X\|_* + \|\mathcal{A}X - \mathbf{b}\|_F^2 \tag{8}$$

for sparsity and rank regularization, respectively. We propose to minimize these by replacing the penalties with their quadratic envelopes $\mathcal{Q}_2(\mu\|\cdot\|_0 + \lambda\|\cdot\|_1)$ and $\mathcal{Q}_2(\mu\ \mathrm{rank} + \lambda\|\cdot\|_*)$, respectively. A reason for this choice is that this regularization does not move the global minimizer, and hence, in many cases we actually find the minimizer of (7) and (8). Our formulation can be seen as a trade-off between small bias and improved optimization properties. While the terms $\lambda\|\mathbf{x}\|_1$ and $\lambda\|X\|_*$ introduce a small bias to solutions, they also increase the convergence basin.

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 from motion (NRSfM). Here each nonzero singular value corresponds to a mode of deformation. When choosing a smaller μ (larger rank) in order to capture all fine deformations the resulting problem is often ill-posed due to unobserved depths. As noted in [24], 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 Sect. 7.5.

The main contributions of this paper are

- We present a class of new regularizers that leverage the benefits of previous convex as well as unbiased non-convex formulations.
- We show that local minimizers of the relaxed functionals is a subset of local minimizers of those to (7).
- We introduce a concrete point called the λ-regularized oracle solution, which is a local minimizer of the relaxation of (7) (and coincides with the classical oracle solution for $\lambda = 0$, i.e., MCP). Moreover we show that all other stationary points necessarily have higher cardinality, so the λ-regularized oracle solution is in this sense unique.
- We show how to compute proximal operators of our regularization enabling fast optimization via splitting methods such as ADMM and FBS.
- We show by examples 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 (7) and (8). We will present our results in the context of the vector case (7). The corresponding matrix versions follow by applying the regularization term to the singular values, with similar proofs. Our first theorem shows that adding the term $\lambda\|\cdot\|_1$ before or after taking the quadratic envelope makes no difference. We say that a function is sign-invariant if the sign on any coordinate can be changed without affecting the function value.

Theorem 2.1 *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a lower semicontinuous sign-invariant function such that $f(\mathbf{x}+\mathbf{y}) \geq f(\mathbf{x})$ for every $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d_+$. Then*

$$\mathcal{Q}_2(f + \lambda\|\cdot\|_1)(\mathbf{x}) = \mathcal{Q}_2(f)(\mathbf{x}) + \lambda\|\mathbf{x}\|_1 \tag{9}$$

for every $\mathbf{x} \in \mathbb{R}^d$.

Proof In [12] (Proposition 3.1 and Theorem 3.1), it is shown that for a lower semicontinuous functional g we have $\mathcal{Q}_2(g)(\mathbf{x}) + \|\mathbf{x}\|^2 = (g(\cdot) + \|\cdot\|^2)^{**}(\mathbf{x})$ (Theorem 3.1), where * denotes the Fenchel conjugate, i.e., $g^*(\mathbf{x}) = \sup_{\mathbf{y}} \langle \mathbf{x}, \mathbf{y}\rangle - g(\mathbf{y})$. Setting $h(\mathbf{x}) := f(\mathbf{x}) + \|\mathbf{x}\|^2$, the result follows if we show that

$$(h(\mathbf{y}) + \lambda\|\mathbf{y}\|_1)^{**} = h^{**}(\mathbf{y}) + \lambda\|\mathbf{y}\|_1. \tag{10}$$

By the symmetry property of h, it suffices to consider $\mathbf{y} \in \mathbb{R}^d_+$. First notice that in

$$(h(\cdot) + \lambda\|\cdot\|_1)^*(\mathbf{y}) = \sup_{\mathbf{x}} \langle \mathbf{x}, \mathbf{y}\rangle - (h(\mathbf{x}) + \lambda\|\mathbf{x}\|_1), \tag{11}$$

only the term $\langle \mathbf{x}, \mathbf{y}\rangle$ depends on the signs of the elements of $\mathbf{x}$; thus, it is clear that any maximizing $\mathbf{x}$ will have $\text{sign}(x_i) = \text{sign}(y_i)$. Therefore, we may assume without loss of generality that $\mathbf{x} \in \mathbb{R}^d_+$ as well. We now have $\|\mathbf{x}\|_1 = \langle \mathbf{x}, \mathbf{1}\rangle$ which reduces (11) to

$$\sup_{\mathbf{x}\in\mathbb{R}^d_+} \langle \mathbf{x}, \mathbf{y} - \lambda\mathbf{1}\rangle - h(\mathbf{x}).$$

Note that if $y_j - \lambda < 0$ for some j, then for every $\mathbf{x} \in \mathbb{R}^d_+$ we have

$$\langle \mathbf{x} - \mathbf{e}_j x_j, \mathbf{y} - \lambda\mathbf{1}\rangle - h(\mathbf{x} - x_j\mathbf{e}_j) \geq \langle \mathbf{x}, \mathbf{y} - \lambda\mathbf{1}\rangle - h(\mathbf{x}),$$

where $\mathbf{e}_j$ is the jth vector of the canonical basis, which implies that the above supremum is the same if we only restrict attention to $\mathbf{x}$ with $\text{supp}(\mathbf{x}) \subset S$, where $S = \{i : y_i > \lambda\}$. Define $\chi_S\mathbf{x} = \sum_{k\in S} \mathbf{e}_k x_k$ and note that

$$\begin{aligned}
&\sup_{\mathbf{x}\in\mathbb{R}^d_+} \langle \mathbf{x}, \mathbf{y} - \lambda\mathbf{1}\rangle - h(\mathbf{x}) \\
&\quad = \sup_{\mathbf{x}\in\mathbb{R}^d_+} \langle \chi_S\mathbf{x}, \mathbf{y} - \lambda\mathbf{1}\rangle - h(\chi_S\mathbf{x}) \\
&\quad = \sup_{\mathbf{x}\in\mathbb{R}^d_+} \langle \chi_S\mathbf{x}, \mathbf{y} - \lambda\mathbf{1}\rangle - h(\mathbf{x}) \\
&\quad = \sup_{\mathbf{x}\in\mathbb{R}^d} \langle \mathbf{x}, \chi_S(\mathbf{y} - \lambda\mathbf{1})\rangle - h(\mathbf{x}) = h^*((\mathbf{y} - \lambda\mathbf{1})_+),
\end{aligned}$$

where $(\mathbf{x})_+$ denotes thresholding at 0, that is, $(\mathbf{x})_+ = (\max(x_1, 0), ..., \max(x_d, 0))$, which gives a more concrete expression for (11).

To compute the second Fenchel conjugate, first note that $h^*(\mathbf{x} + \mathbf{v}) \geq h^*(\mathbf{x})$ for $\mathbf{x}, \mathbf{v} \in \mathbb{R}^d_+$ since

$$\langle \mathbf{y}, \mathbf{x}\rangle - h(\mathbf{y}) \leq \langle \mathbf{y}, \mathbf{x} + \mathbf{v}\rangle - h(\mathbf{y})$$

for all $\mathbf{y} \in \mathbb{R}^d_+$. Moreover, in the supremum $\sup_{\mathbf{x}\in\mathbb{R}^d} \langle \mathbf{x}, \mathbf{y}\rangle - h^*((\mathbf{x} - \lambda\mathbf{1})_+)$ it clearly suffices to consider $\mathbf{x}$ with $x_j \geq \lambda$ for all $1 \leq j \leq d$. By this observation, we get that $(h + \lambda\|\cdot\|_1)^{**}(\mathbf{y})$ equals to

$$\begin{aligned}
&\sup_{\mathbf{x}\in\mathbb{R}^d} \langle \mathbf{x}, \mathbf{y}\rangle - h^*((\mathbf{x} - \lambda\mathbf{1})_+) \\
&\quad = \sup_{x_j\geq\lambda,\ 1\leq j\leq d} \langle \mathbf{x}, \mathbf{y}\rangle - h^*(\mathbf{x} - \lambda\mathbf{1}) \\
&\quad = \sup_{\mathbf{z}\in\mathbb{R}^d_+} \langle \mathbf{z} + \lambda\mathbf{1}, \mathbf{y}\rangle - h^*(\mathbf{z}) \\
&\quad = \lambda\|\mathbf{y}\|_1 + \sup_{\mathbf{z}\in\mathbb{R}^d} \langle \mathbf{z}, \mathbf{y}\rangle - h^*(\mathbf{z}),
\end{aligned}$$

which shows that $(h + \lambda\|\cdot\|_1)^{**}(\mathbf{y}) = \lambda\|\mathbf{y}\|_1 + h^{**}(\mathbf{y})$. □

The function f need not be $\mu\|\cdot\|_0$, if the sought cardinality is known one could for example incorporate this information

by letting f be the indicator functional of $\{\mathbf{x} : \|\mathbf{x}\|_0 \leq K\}$ (c.f. [13]) which leads to highly non-trivial non-separable new penalties. However, for the remainder of this paper we focus exclusively on $f(\mathbf{x}) = \mu\|\mathbf{x}\|_0$.

In view of the above and (3), it follows that $\mathcal{Q}_2(\mu\|\cdot\|_0 + \lambda\|\cdot\|_1) = r_{\mu,\lambda}$, where

$$r_{\mu,\lambda}(\mathbf{x}) = \sum_i \left(\mu - \max(\sqrt{\mu} - |x_i|, 0)^2\right) + \lambda\|\mathbf{x}\|_1. \quad (12)$$

We therefore propose to minimize the objective

$$r_{\mu,\lambda}(\mathbf{x}) + \|A\mathbf{x} - \mathbf{b}\|^2. \quad (13)$$

This is motivated by the following simple observation.

Lemma 2.2 *If A has columns of Euclidean norm (strictly) less than one, the local minimizers of (13) form a subset of those of (7); moreover, the global minimizers coincide.*

Proof Let $\mathbf{x}$ be a local minimizer of (13). If $0 < |x_i| < \sqrt{\mu}$ holds for some index i, then it follows by (12) that $\partial_i^2 r_{\mu,\lambda} = -2$. If $\mathbf{a}_i$ denotes the i:th column of A, we get on the other hand that $\partial_i^2\|A\mathbf{x} - \mathbf{b}\|^2 = 2\|\mathbf{a}_i\| < 2$, and hence, this point cannot be a local minimizer of (13), a contradiction. Hence, we either have $x_i = 0$ or $|x_i| \geq \sqrt{\mu}$ for all indices i. By (12), we get that $r_{\mu,\lambda}(\mathbf{x}) = \mu\|\mathbf{x}\|_0 + \lambda\|\mathbf{x}\|_1$, and hence, $\mathbf{x}$ must also be a local minimizer for (7) (since (13) is less than or equal to (7), but equal at the point $\mathbf{x}$; this follows from a general feature of the quadratic envelope $\mathcal{Q}_2$, cfr. Theorem 3.1 in [12] for additional details). □

We remark that the assumption on A always can be achieved by a rescaling of the problem[1]. The property of not moving minimizers is inherent to quadratic envelope regularizations, see [12]. Note that $r_{\mu,\lambda}(\mathbf{x}) + \|A\mathbf{x} - \mathbf{b}\|^2$ reduces to (2) if $\mu = 0$ and (3) if $\lambda = 0$. Figure 1 shows an illustration of $r_{\mu,\lambda}$ for a couple of different values of λ. When $\lambda = 0$ the function is constant for values larger than $\sqrt{\mu}$. Therefore, large elements give zero gradients which can result in algorithms getting stuck in poor local minimizers. Increasing λ makes the regularizer closer to being convex, which as we show numerically in Sect. 7, increases its convergence basin.

We conclude this section with a simple 2D illustration of the general principle. Consider $r_{\mu,\lambda}(\mathbf{x}) + \|A\mathbf{x} - \mathbf{b}\|^2$ for a two-dimensional problem with

$$A = \begin{pmatrix} 0.4 & 0 \\ 0 & 0.6 \end{pmatrix} \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} 0.8 \\ 1.8 \end{pmatrix}. \quad (14)$$

[1] Alternatively we can work with the original A and the more general transform $\mathcal{Q}_\bigcirc$ where $\gamma > 0$ is a parameter chosen with respect to the size of A (see [12]). We have chosen the above assumption on A and set $\gamma = 2$ for simplicity of the presentation.

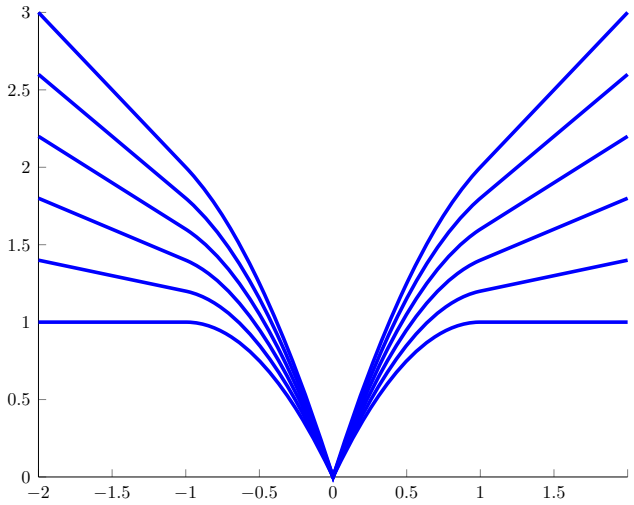

Fig. 1 Function (12) for $\mu = 1$ and $\lambda = 0, 0.2, 0.4, \ldots, 1$

Since the matrix is diagonal, the function is the sum of two functions of 1 variable, which are depicted in Fig. 2. The blue curves show the case $\mu = 1$ and $\lambda = 0$. It is easy to verify that the problem has two local minimizers $\mathbf{x} = (2, 3)$ and $\mathbf{x} = (0, 3)$ (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 ℓ^1 formulation (2), with $\lambda = 0.7$. Here we have used the smallest possible λ so that the optimum of the left residual is 0 while the right one is nonzero. The resulting solution $\mathbf{x} = (0, 2)$ has the correct support; however, the magnitude of the nonzero 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 $\mathbf{x} \approx (0, 2.5)$.

3 Oracle Solutions

For sparsity problems, the so-called oracle solution [8] is what we would obtain if we somehow knew the "true" support of the sought solution and we were to solve the least squares problem over the nonzero entries of $\mathbf{x}$. Candés et al. [8] showed that under certain RIP conditions the solution (2) (i.e., LASSO) approximates the oracle solution. In [13], it was then shown that (3) often gives exactly the oracle solution. In this section, we will show that our relaxation solves a similar ℓ^1-regularized least squares problem. In particular, for $\mu = 0$ this gives a concrete characterization of the LASSO minimizer.

Let x_0 be the so-called ground truth, i.e., a sparse vector that we wish to recover using the measurement $\mathbf{b} = A\mathbf{x}_0 + \epsilon$

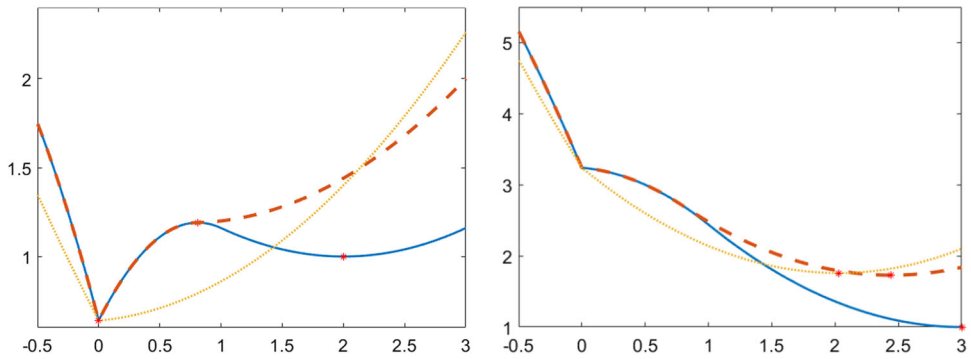

Fig. 2 Illustration; yellow "only" ℓ^1, blue "only" MCP, red dashed shows an intermediate penalty based on $r_{\mu,\lambda}$ (Color figure online)

where ϵ denotes noise. Furthermore, let S be the set of nonzero indices of x_0, let K be the cardinality of S and suppose that $\delta_K^- \in [0, 1)$, which simply means that any K columns of A are linearly independent. 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. We refer to the λ-regularized oracle solution as the minimizer of

$$\mathbf{x}_\lambda = \arg\min_{\mathbf{x}} \lambda\|\mathbf{x}\|_1 + \|A_S\mathbf{x} - \mathbf{b}\|^2. \tag{15}$$

Note that $\mathbf{x}_\lambda$ indeed gets support in S (since we use A_S instead of A above), and hence, that the minimization problem (15) has a unique solution (due to the assumption that $\delta_K^- \in [0, 1)$ which implies Condition 1 in [33]). It is easy to see that in the limit $\lambda \to 0^+$, this becomes the classical oracle solution, i.e., the least squares solution over the correct support. For a nonzero λ, the ℓ^1 norm modifies the solution by adding a shrinking bias.

We now show that the solutions to (15) also are stationary points of (13). For non-convex functions, we will say that a point is stationary if its Fréchet subdifferential $\hat{\partial}$ includes $\mathbf{0}$, we refer to Section 3 of [13] for more details.

Theorem 3.1 *Suppose that* $\mathbf{x}_\lambda$ *of* (15) *fulfills* $|x_{\lambda,i}| \geq \sqrt{\mu}$ *for all* $i \in S$, *that* $\delta_K^- \in [0, 1)$, *that* A *has columns of Euclidean norm less than one, and that the residual errors* $\epsilon = A\mathbf{x}_\lambda - \mathbf{b}$ *satisfy* $\|\epsilon\|_2 < \sqrt{\mu} + \lambda/2$. *Then* $\mathbf{x}_\lambda$ *is a* stationary point *of* (13).

Proof It is easy to see that $r_{\mu,\lambda}(\mathbf{x}) + \|\mathbf{x}\|^2$ can be written as the convex function $\sum_j \max(2\sqrt{\mu}|x_j|, \mu + x_j^2 + \lambda|x_j|)$. Hence, (13) is the difference of this convex function and the smooth function $\|\mathbf{x}\|^2 - \|A\mathbf{x} - \mathbf{b}\|^2$, and for such functions, it is easy to see that a point $\mathbf{x}$ is stationary if and only if the gradient of the smooth part is a member of the subdifferential of the convex part. Since the latter can be written as a cartesian product $\times_i A_i$ with $A_i \subseteq \mathbb{R}$, this condition can be verified coordinate-wise. For $\mathbf{x} = \mathbf{x}_\lambda$ and j such that $x_j = 0$, we have

$$\nabla_j(\|\mathbf{x}\|^2 - \|A\mathbf{x} - \mathbf{b}\|^2) = 2\langle A\mathbf{x} - \mathbf{b}, a_j\rangle,$$

where a_j denotes the j :th column of A, whereas the corresponding interval for the subgradient of the convex part is $[-2\sqrt{\mu} - \lambda, 2\sqrt{\mu} + \lambda]$. Since $|\langle A\mathbf{x} - \mathbf{b}, a_j\rangle| \leq \|\epsilon\|\|a_j\| < \|\epsilon\|$ the former point is a member of this subset. It remains to check the nonzero x_j's, i.e., for $j \in S$. (This follows by the definition of $\mathbf{x}_\lambda$ and the assumption $|x_{\lambda,i}| \geq \sqrt{\mu}$ for all $j \in S$.) Let $\#S$ denote the cardinality of S and note that by assumption on $\mathbf{x}_\lambda$ we have

$$r_{\mu,\lambda}(\mathbf{x}) = \mu\#S + \lambda\|\mathbf{x}\|_1$$

for $\mathbf{x}$ in a vicinity of $\mathbf{x}_\lambda$ with supp $\mathbf{x} \subset S$. This, in combination with the fact that $\mathbf{x}_\lambda$ solves (15), shows that

$$\begin{aligned}\mathbf{0} &\in \partial_j\left(\lambda\|\mathbf{x}_\lambda\|_1 + \|A_S\mathbf{x}_\lambda - \mathbf{b}\|^2\right)\\ &= \partial_j\left(r_{\mu,\lambda}(\mathbf{x}_\lambda) + \|A_S\mathbf{x}_\lambda - \mathbf{b}\|^2\right),\end{aligned}$$

as desired. □

Whether $\mathbf{x}_\lambda$ is the global optimum or not depends on the problem instance. However, for the particular case of $\mu = 0$, the problem is convex and hence a stationary point is a global minimizer. In other words, the theorem says that the $\lambda-$regularized solution is often the solution of the classical ℓ^1-problem (2) (LASSO). For $\mu > 0$, we will in the following sections show that under a sufficiently strong RLIP it is the sparsest possible stationary point.

In Fig. 3, we illustrate with an experiment, the setup is very similar to the one described in Sect. 7.3: a random matrix A, together with a ground truth $\mathbf{x}_0$ and a set of noisy measurements $\mathbf{b}$ are fixed; the parameter μ is also kept fixed and chosen such that $x_{S,i} = x_{0,i} \geq \sqrt{\mu}$ for all $i \in S$. We study the the impact of an increasingly bigger value of λ on the reconstruction performances, and we draw quantitative conclusions. Blue graphs relate to solving LASSO (2) for various values of λ (solution denoted $\mathbf{x}_{\ell^1}$), and the red graphs show the same but for (13) (denoted $\mathbf{x}_{r_{\mu,\lambda}}$). The noise is fixed at noise level 30%. The yellow line shows distance from $\mathbf{x}_\lambda$ to ground truth $\mathbf{x}_0$. Clearly, this deviates from $\mathbf{x}_0$ at a linear rate, as expected, demonstrating the need to keep λ small.

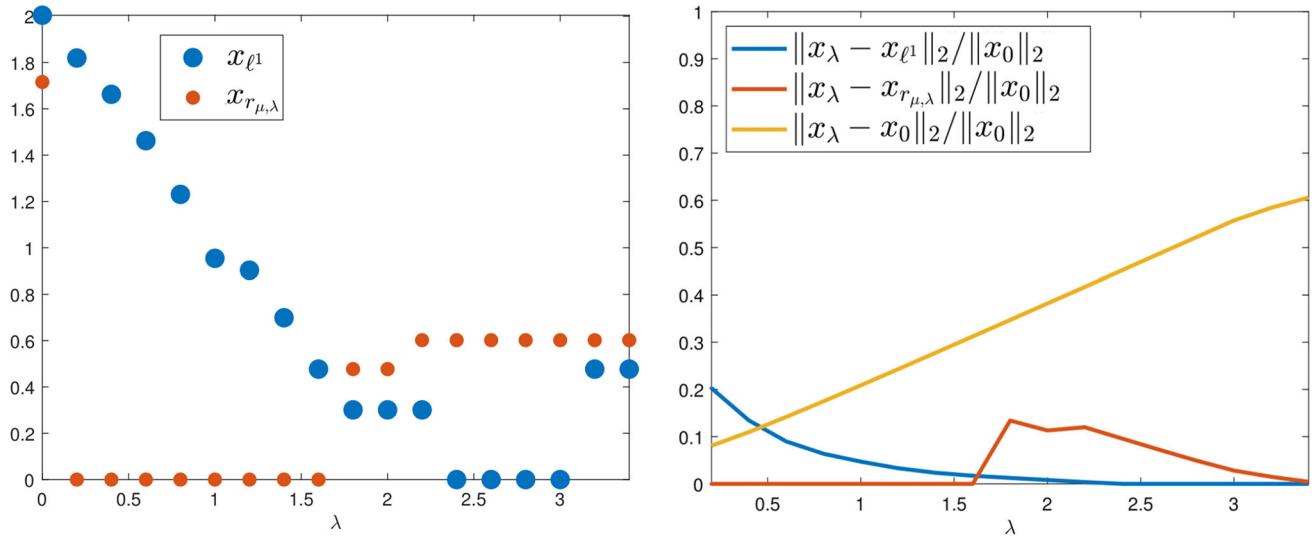

Fig. 3 Left: support misfit (in form of $\log_{10}(1 + \#\{\text{misfit}\})$) for (2) (blue) and (13) (red). Right: normalized distance to $\mathbf{x}_\lambda$ as well as $\|\mathbf{x}_\lambda - \mathbf{x}_0\|_2$ (yellow) (Color figure online)

The left graph shows $\log_{10}(1 + \#\{\text{misfit}\})$ where $\#\{\text{misfit}\}$ is counting the amount of wrong positions in the support of the estimated sparse solution (so value 0 indicates perfect support recovery). As it is plain to see, ℓ^1 finds the correct support only for very high values of λ, and in this regime, we also have $\mathbf{x}_{\ell^1} = \mathbf{x}_\lambda$ as predicted by Theorem 3.1 (when $\mu = 0$), but here $\mathbf{x}_\lambda$ is very far from the ground truth. This regime is also small and therefore hard to find in practice. On the contrary, $\mathbf{x}_{r_{\mu,\lambda}}$ fails only for $\lambda = 0$ (the algorithm is initialized at the least squares solution, which is a local minimum) and very high values of λ. Another interesting observation is that $\mathbf{x}_{r_{\mu,\lambda}}$ stops having the right support as soon as the condition $|x_{\lambda,i}| \geq \sqrt{\mu}$ from Theorem 3.1 is violated for some $i \in S$.

4 Separation of Stationary Points

A feature of the left red graph in Fig. 3 which is not explained by Theorem 3.1 is the fact that when $\mathbf{x}_{r_{\mu,\lambda}}$ fails to find $\mathbf{x}_\lambda$ for $\lambda = 0$, it has a very large support. In this section, we aim at providing theoretical support also for this fact. More precisely, we study the stationary points of the objective function (13) under the assumption that A fulfills the RLIP condition (4) with decent values of δ_K^-. We will extend the results of [13,24] to our class of functionals. Specifically, we show that under some technical conditions two stationary points $\mathbf{x}'$ and $\mathbf{x}''$ have to be separated by $\|\mathbf{x}'' - \mathbf{x}'\|_0 > 2K$. From a practical point of view, this means that if we find a stationary point with $\|\mathbf{x}'\|_0 \leq K$ 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 $\mathbf{z}$ which depends on the data. As in [13] we introduce the auxiliary function $\mathcal{G}_{\mu,\lambda}(\mathbf{x}) = \frac{1}{2}(r_{\mu,\lambda}(\mathbf{x}) + \|\mathbf{x}\|^2)$, i.e., $2\mathcal{G}_{\mu,\lambda}(\mathbf{x})$ equals the l.s.c. convex envelope of $\mu\|\cdot\|_0 + \lambda\|\cdot\|_1 + \|\cdot\|_2$, by Theorem 2.1 and the design of $\mathcal{Q}_2$. Notice that the function $\mathcal{G}_{\mu,\lambda}$is convex and proper, and thus, the object $\partial\mathcal{G}_{\mu,\lambda}$ is the classical subdifferential from convex analysis.

Given a point $\mathbf{x}$ (stationary or not), we introduce the auxiliary point

$$\mathbf{z}(\mathbf{x}) = (I - A^t A)\mathbf{x} + A^t\mathbf{b}; \tag{16}$$

in the following proofs, we will use the shorter notations $\mathbf{z} = \mathbf{z}(\mathbf{x})$, $\mathbf{z}' = \mathbf{z}(\mathbf{x}')$ and $\mathbf{z}'' = \mathbf{z}(\mathbf{x}'')$.

Proposition 4.1 *The point $\mathbf{x}'$ is stationary for (13) if and only if $\mathbf{z}' \in \partial\mathcal{G}_{\mu,\lambda}(\mathbf{x}')$ and if and only if*

$$\mathbf{x}' \in \arg\min_{\mathbf{x}} r_{\mu,\lambda}(\mathbf{x}) + \|\mathbf{x} - \mathbf{z}'\|^2. \tag{17}$$

Proof First note the identity

$$r_{\mu,\lambda}(\mathbf{x}) + \|A\mathbf{x} - \mathbf{b}\|^2 = 2\mathcal{G}_{\mu,\lambda}(\mathbf{x}) + \|A\mathbf{x} - \mathbf{b}\|^2 - \|\mathbf{x}\|^2.$$

By differentiating, we see that $\mathbf{x}'$ is stationary in (13) if and only if $0 \in 2\partial\mathcal{G}_{\mu,\lambda}(\mathbf{x}') + 2(A^tA\mathbf{x}' - A^t\mathbf{b} - \mathbf{x}')$ which reordered becomes $\mathbf{z}' \in \partial\mathcal{G}_{\mu,\lambda}(\mathbf{x}')$. Similarly, differentiating (17) we see that $\mathbf{x}'$ is stationary in (17) if and only if $\mathbf{z}' \in \partial\mathcal{G}_{\mu,\lambda}(\mathbf{x}')$. Now recall that by construction the functional in (17) is

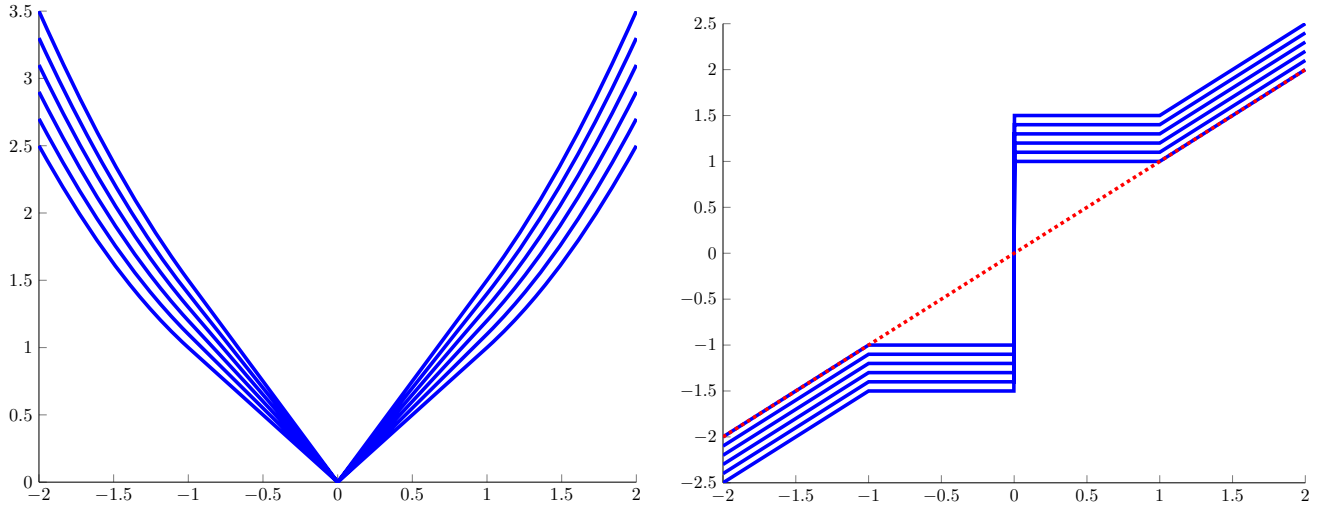

Fig. 4 Function $g_{\mu,\lambda}(x)$ (left) and the subdifferential $\partial g_{\mu,\lambda}(x)$ (right) for $\mu = 1$ and $\lambda = 0, 0.2, 0.4, \ldots, 1$. (For comparison we also plot the red dotted curve $y = x$)

convex and therefore $\mathbf{x}'$ being stationary is equivalent to solving (17). □

We will use properties of the vector $\mathbf{z}'$ to establish conditions that ensure that $\mathbf{x}'$ is the sparsest possible stationary point of (13). The overall idea which follows [11,24] is to show that the subdifferential $\partial\mathcal{G}_{\mu,\lambda}$ grows faster than $\mathbf{z}$, as a function of $\mathbf{x}$, and therefore, we can only have $\mathbf{z} \in \partial\mathcal{G}_{\mu,\lambda}(\mathbf{x})$ in one (sparse) point. The result requires that the elements of the vector $\mathbf{z}'$ are not too close to the threshold $\sqrt{\mu} + \frac{\lambda}{2}$.

Theorem 4.2 *Let δ^-_{2K} be the RLIP constant (4) for the matrix A for cardinality $2K$. Assume that $\mathbf{x}'$ is a stationary point of (13) and that each element of $\mathbf{z}'$ (defined as in (16)) fulfills $|z'_i| \notin [(1-\delta^-_{2K})\sqrt{\mu} + \frac{\lambda}{2}, \frac{\sqrt{\mu}}{(1-\delta^-_{2K})} + \frac{\lambda}{2}]$. If $\mathbf{x}''$ is another stationary point then $\|\mathbf{x}'' - \mathbf{x}'\|_0 > 2K$.*

The proof of Theorem 4.2 requires an estimate of the growth of the subgradients of $\mathcal{G}_{\mu,\lambda}$ which we now present. The function $\mathcal{G}_{\mu,\lambda}$ is separable and can be evaluated separately for each element of $\mathbf{x}$. To separate the notation, we write $g_{\mu,\lambda}$ for $\mathcal{G}_{\mu,\lambda}$ restricted to one real parameter x. The subdifferential of $g_{\mu,\lambda}$ then becomes

$$\partial g_{\mu,\lambda}(x) = \begin{cases} \{x + \lambda/2\mathrm{sign}(x)\} & |x| \geq \sqrt{\mu} \\ \{(\sqrt{\mu} + \lambda/2)\mathrm{sign}(x)\} & 0 < |x| \leq \sqrt{\mu} \\ [-\sqrt{\mu} - \lambda/2, \sqrt{\mu} + \lambda/2] & x = 0. \end{cases} . \tag{18}$$

Figure 4 shows the function $g_{\mu,\lambda}$ and $\partial g_{\mu,\lambda}$. The parameter λ adds a constant offset to the positive values of $\partial g_{\mu,\lambda}(x)$ and subtracts the same value for all negative values.

It is clear from Fig. 4 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

$$\langle \partial g_{\mu,\lambda}(x'') - \partial g_{\mu,\lambda}(x'), x'' - x' \rangle = |x'' - x'|^2. \tag{19}$$

For the other parts, the subdifferential grows less. To ensure a certain growth, we need to add some assumptions on the subdifferential which is done in the following result.

Lemma 4.3 *Assume that $\mathbf{x}'$ is such that $\mathbf{z}' \in \partial\mathcal{G}_{\mu,\lambda}(\mathbf{x}')$ and $\beta > 0$, where again $\mathbf{z}'$ is defined by (16). If the elements z'_i fulfill $|z'_i| \notin [\beta^2\sqrt{\mu} + \frac{\lambda}{2}, \frac{\sqrt{\mu}}{\beta^2} + \frac{\lambda}{2}]$ for every i, then for any $\mathbf{x}''$ with $\mathbf{z}'' \in \partial\mathcal{G}_{\mu,\lambda}(\mathbf{x}'')$ we have*

$$\langle \mathbf{z}'' - \mathbf{z}', \mathbf{x}'' - \mathbf{x}' \rangle > (1 - \beta^2)\|\mathbf{x}'' - \mathbf{x}'\|^2, \tag{20}$$

as long as $\mathbf{x}' \neq \mathbf{x}''$.

Proof We first consider the scalar case: $z' \in \partial g_{\mu,\lambda}(x')$; by the symmetry of (18), we may assume that $z' \geq 0$.

First assume that $z' > \frac{\sqrt{\mu}}{\beta^2} + \frac{\lambda}{2}$. In view of (18), we then have $x' = z' - \frac{\lambda}{2} > \frac{\sqrt{\mu}}{\beta^2}$. Now consider the linear function

$$l(x) = (1 - \beta^2)(x - x') + z' = (1 - \beta^2)x + \beta^2 x' + \frac{\lambda}{2}.$$

Since $l(x') = z'$ and $l(0) = \beta^2 x' + \frac{\lambda}{2} > \sqrt{\mu} + \frac{\lambda}{2}$, Fig. 4 shows that $l(x'') > z''$ for all $x'' < x'$. Therefore,

$$z' - z'' > z' - l(x'') = (1 - \beta^2)(x' - x''),$$

for all $x'' < x'$. Additionally, for $x'' > x'$ we clearly have that

$$z'' - z' = x'' - x' > (1 - \beta^2)(x'' - x');$$

in both scenarios ($x'' > x'$ and $x' > x''$), we obtain

$$(z'' - z')(x'' - x') > (1 - \beta^2)(x'' - x')^2. \tag{21}$$

Now assume that $0 \leq z' \leq \beta^2\sqrt{\mu} + \frac{\lambda}{2}$; this implies $x' = 0$, which follows from the structure of the subgradient of $g_{\mu,\lambda}$ (18). If we define another linear function $p(x) = (1-\beta^2)x + z'$, we have that

$$p(\sqrt{\mu}) = (1 - \beta^2)\sqrt{\mu} + z' < \sqrt{\mu} + \frac{\lambda}{2};$$

and it is clear that, if $x'' > 0$, then $p(x'') < z''$ (there are no hypothesis on z'' here). Therefore,

$$z'' - z' > p(x'') - z' = (1 - \beta^2)x'' = (1 - \beta^2)(x'' - x').$$

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

$$\begin{aligned} z' - z'' > z' - p(x'') &= -(1 - \beta^2)x'' \\ &= (1 - \beta^2)(x' - x''), \end{aligned}$$

which again yields (21). To obtain (20), we now sum over the nonzero entries of $\mathbf{x}'' - \mathbf{x}'$. □

Proof of Theorem 4.2 By Proposition 4.1, we have $\mathbf{z}' \in \partial\mathcal{G}_{\mu,\lambda}(\mathbf{x}')$ so Lemma 4.3 applies to $\mathbf{x}'$, $\mathbf{z}'$. Let $\mathbf{z}''$ be related to $\mathbf{x}''$ via (16). Then

$$\mathbf{z}'' - \mathbf{z}' = (I - A^t A)(\mathbf{x}'' - \mathbf{x}'),$$

which gives

$$\langle \mathbf{z}'' - \mathbf{z}', \mathbf{x}'' - \mathbf{x}' \rangle = \|\mathbf{x}'' - \mathbf{x}'\|^2 - \|A(\mathbf{x}'' - \mathbf{x}')\|^2.$$

Since A satisfies the RLIP condition this is less than or equal to $\delta_{2K}^{-}\|\mathbf{x}'' - \mathbf{x}'\|^2$ whenever $\|\mathbf{x}'' - \mathbf{x}'\|_0 \leq 2K$, which is impossible by Lemma 4.3. □

Let us summarize our conclusions so far: We have introduced a relaxed functional (13) which is intermediate between the classical LASSO and MCP penalties. We have shown that the local minimizers of (13) are a subset of those of (7), and we have concretely characterized one such minimizer $\mathbf{x}_\lambda$. This is the sought solution and, although it may not be unique, it is unique as a sparse solution. In other words, if Theorem 4.2 applies with K sufficiently big and the algorithm gets stuck in an undesired local minimum, this will be visible by its high cardinality. It is clear that the bias of $\mathbf{x}_\lambda$ scales linearly with λ, but a small λ in LASSO gives a too big support, and this is where the $\mu-$parameter comes handy. Ideally, one should pick $\lambda = 0$ for then the oracle solution is among the local minimizers (in fact, it is often the unique global minimizer, see [13]), but in practice a trade-off may be more reliable due to the risk of getting stuck in undesired local minima of MCP.

Let us also underline that although we have studied one concrete and relatively simple separable penalty $r_{\mu,\lambda}$, the idea to extend the convergence basin of non-convex penalties applies to a whole array of sparsity inducing penalties such as those studied in [20].

5 Optimization

The optimization of functions of the type (13) is straightforward and can be done either via ADMM or FBS, once the proximal operator is known, which we now compute. Both have been proven to converge in the present setting, in the former case see [29] and in the latter one needs to combine the results of [12] with [1]. We have also run both algorithms in parallel and found that they almost always converge to the same point, despite the non-convex landscape. To generalize these algorithms to the matrix case is also straightforward, one basically needs to apply the vector proximal operator to the singular values, see [14].

5.1 The Proximal Operator

The proximal operator of $r_{\mu,\lambda}/\rho$, where ρ is a step length parameter, is defined by

$$\mathsf{prox}_{\frac{r_{\mu,\lambda}}{\rho}}(\mathbf{y}) = \arg\min_{\mathbf{x}} r_{\mu,\lambda}(\mathbf{x}) + \frac{\rho}{2}\|\mathbf{x} - \mathbf{y}\|^2 \tag{22}$$

where $r_{\mu,\lambda}(\mathbf{x}) = \mathcal{Q}_2(\mu\|\cdot\|_0 + \lambda\|\cdot\|_1)(\mathbf{x})$. The following result shows that in general the proximal operator of $\mathcal{Q}_2(f + \lambda\|\cdot\|_1)$ is easy to compute if the proximal operator of $\mathcal{Q}_2(f)$ is known. Note that $\mathcal{Q}_2(f)$ is a non-convex functional with maximum negative curvature -2 (see [12]), and hence we must require that $\rho > 2$ in order for the proximal operator to be single valued (Fig. 5).

We recall that a function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be *sign-invariant* if

$$f(\mathbf{x}) = f(S\mathbf{x})$$

for all $\mathbf{x} \in \mathbb{R}^d$ and all diagonal $d \times d$ matrices S with only 1 and -1 on the main diagonal.

Proposition 5.1 *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a lower semicontinuous sign-invariant function such that $f(\mathbf{x}+\mathbf{y}) \geq f(\mathbf{x})$ for every $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d_+$. Then*

$$\mathsf{prox}_{\mathcal{Q}_2(f+\lambda\|\cdot\|_1)/\rho}(\mathbf{y}) = \mathsf{prox}_{\mathcal{Q}_2(f)/\rho}(\mathsf{prox}_{\lambda\|\cdot\|_1/\rho}(\mathbf{y}))$$

for every $\mathbf{y} \in \mathbb{R}^d$ and $\rho \geq 2$.

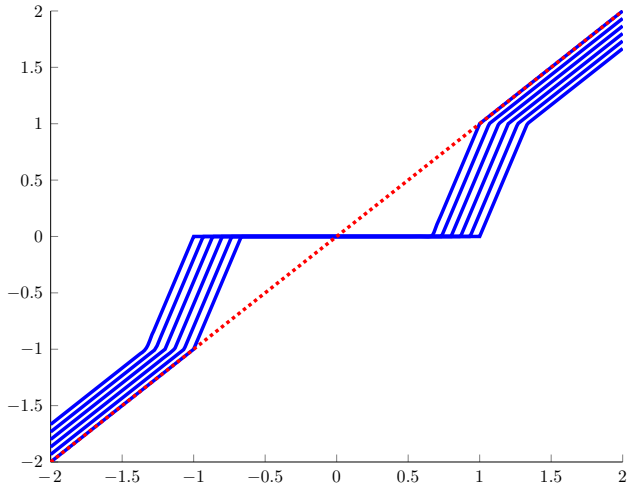

Fig. 5 Proximal operator given by (23) for $\rho = 3$, $\mu = 1$ and $\lambda = 0, 0.2, 0.4, \ldots, 1$

Proof It is enough to compute the proximal operator of the function $\mathcal{Q}_2(f)(\cdot)+\lambda\|\cdot\|_1$ as per Theorem 2.1. Without loss of generality we assume that $\mathbf{y} \in \mathbb{R}^d_+$. With the same notation as in the proof of Theorem 2.1 we have

$$\begin{aligned}
&\mathsf{prox}_{(\mathcal{Q}_2(f)+\lambda\|\cdot\|_1)/\rho}(\mathbf{y}) \\
&= \arg\min_{\mathbf{x}\in\mathbb{R}^d} \frac{\mathcal{Q}_2(f)(\mathbf{x})}{\rho} + \frac{\lambda}{\rho}\|\mathbf{x}\|_1 + \frac{1}{2}\|\mathbf{x}-\mathbf{y}\|^2 \\
&= \arg\min_{\mathbf{x}\in\mathbb{R}^d_+} \frac{\mathcal{Q}_2(f)(\mathbf{x})}{\rho} + \frac{\|\mathbf{x}\|^2}{2} \\
&\quad -\langle \mathbf{x}, (\mathbf{y}-\frac{\lambda}{\rho}\mathbf{1})_+\rangle + \frac{\|\mathbf{y}\|^2}{2}
\end{aligned}$$

because the quantity $\lambda\|\mathbf{x}\|_1/\rho - \langle \mathbf{x}, \mathbf{y}\rangle$ will be minimized with an $\mathbf{x}$ with the same signs of $\mathbf{y}$. i.e., $\mathbf{x} \in \mathbb{R}^d_+$. Moreover

$$\lambda\|\mathbf{x}\|_1/\rho - \langle \mathbf{x}, \mathbf{y}\rangle = -\langle \mathbf{x}, \mathbf{y} - \lambda\mathbf{1}/\rho\rangle$$

and again we want the latter to be as small as possible and thus we pick $\mathbf{x}$ such that $x_j = 0$ if $(\mathbf{y} - \lambda\mathbf{1}/\rho)_j < 0$. Since $\|\mathbf{x}\|^2 - 2\langle \mathbf{x}, (\mathbf{y}-\frac{\lambda}{\rho}\mathbf{1})_+\rangle = \|\mathbf{x}-(\mathbf{y}-\frac{\lambda}{\rho}\mathbf{1})_+\|^2 - \|(\mathbf{y}-\frac{\lambda}{\rho}\mathbf{1})_+\|^2$ and the terms in $\mathbf{y}$ are constant (since the minimization is over $\mathbf{x}$), we see that $\mathbf{x}$ also solves

$$\arg\min_{\mathbf{x}\in\mathbb{R}^d_+} \frac{\mathcal{Q}_2(f)(\mathbf{x})}{\rho} + \frac{1}{2}\|\mathbf{x}-(\mathbf{y}-\frac{\lambda}{\rho}\mathbf{1})_+\|^2.$$

Note that $(\mathbf{y}-\frac{\lambda}{\rho}\mathbf{1})_+ = \mathsf{prox}_{\lambda\|\cdot\|_1/\rho}(\mathbf{y})$ since $\mathbf{y} \in \mathbb{R}^d_+$. Also, since the elements of $(\mathbf{y}-\frac{\lambda}{\rho}\mathbf{1})_+$ are nonnegative it is clear that minimizing over $\mathbf{x} \in \mathbb{R}^d$ instead of $\mathbb{R}^d_+$ does not change the optimizer and therefore

$$\mathsf{prox}_{(\mathcal{Q}_2(f)+\lambda\|\cdot\|_1)/\rho}(\mathbf{y}) = \mathsf{prox}_{\mathcal{Q}_2(f)/\rho}((\mathbf{y}-\frac{\lambda}{\rho}\mathbf{1})_+).$$

□

For our particular case, $f(\mathbf{x}) = \mu\|\mathbf{x}\|_0$, the proximal operator is separable and each element of the vector $\mathbf{x}$ can be treated independently. As usual the soft thresholding operator is given by $\mathrm{sign}(y_i)\max(|y_i|-\lambda/\rho, 0)$. The computations of $\mathbf{x} = \mathsf{prox}_{\frac{\mathcal{Q}_2(\mu\|\cdot\|_0)}{\rho}}(\mathbf{y})$ are fairly straightforward and can be found, e.g., in [11,20]. For $\rho > 2$ we get

$$\begin{aligned}
&(\mathsf{prox}_{(\mathcal{Q}_2(f)+\lambda\|\cdot\|_1)/\rho}(\mathbf{y}))_i \\
&= \begin{cases} y_i - \lambda/\rho & |y_i| \geq \lambda/\rho + \sqrt{\mu} \\ \frac{\rho y_i - \lambda - 2\sqrt{\mu}\,\mathrm{sign}(y_i)}{\rho-2} & \frac{\lambda+2\sqrt{\mu}}{\rho} \leq |y_i| \leq \frac{\lambda}{\rho} + \sqrt{\mu} \\ 0 & |y_i| \leq \frac{\lambda+2\sqrt{\mu}}{\rho}. \end{cases}
\end{aligned} \tag{23}$$

6 Matrix Framework

In this section, we briefly show how the theory can be lifted to the matrix framework. We let $\sigma(X)$ denote the singular values of a given matrix X. Note that $\|\sigma(X)\|_0 = \mathrm{rank}(X)$ and that $\|\cdot\|_1$ applied to the singular values gives the nuclear norm $\|X\|_*$, which is a rank-reducing penalty, see the discussion around (5) and (6). Analogously we can consider $r_{\mu,\lambda}(\sigma(X))$ which is a rank-reducing penalty with less of a bias than $\|X\|_*$. For the case $\lambda = 0$, it has been shown in [14] how to lift basically any statement about vectors to a corresponding statement for matrices, and along these lines, we could develop a theory for matrices parallel to the results in Sects. 2–5. We refrain from this and focus here on providing the necessary details to apply this framework in practice. We recall that a function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be *absolutely symmetric* if $f(|\mathbf{x}|) = f(\mathbf{x})$ and $f(\Pi\mathbf{x}) = f(\mathbf{x})$ for all permutations Π and all $\mathbf{x} \in \mathbb{R}^d$.

Proposition 6.1 *Suppose that f is an absolutely symmetric functional on $\mathbb{R}^d$, $d = \min(n_1, n_2)$, and that $F(Y) = f(\sigma(Y))$, $Y \in \mathbb{R}^{n_1\times n_2}$. Then*

$$\mathcal{Q}_2(F)(Y) = \mathcal{Q}_2(f)(\sigma(Y)).$$

Proof See Proposition 4.1 of [14]. □

In a similar fashion, "lifted" proximal operators can be computed:

Proposition 6.2 *Let f be an absolutely symmetric function on $\mathbb{R}^d$ and set as in the previous proposition $F(Y) = f(\sigma(Y))$. Then for $\rho > 2$*

$$\mathsf{prox}_{\mathcal{Q}_2(F)/\rho}(X) = U\,\mathrm{diag}(\mathsf{prox}_{\mathcal{Q}_2(f)/\rho}(\sigma(X)))V^*$$

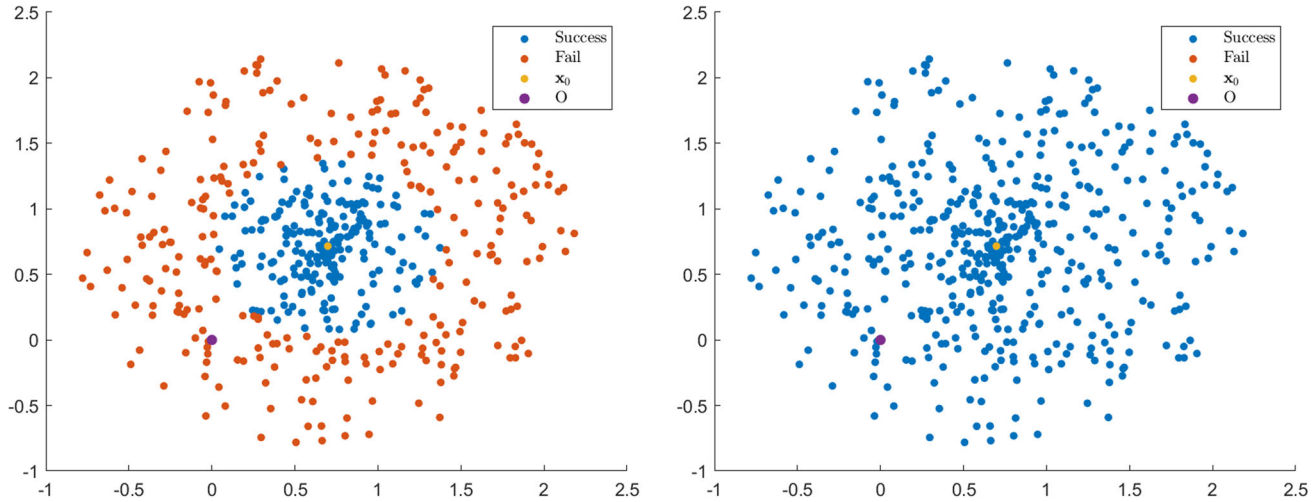

Fig. 6 Starting points clouds. The image on the left shows the outcome for the functional $\mathcal{Q}_2(\mu\|\cdot\|_0)(\mathbf{x}) + \|A\mathbf{x}-\mathbf{b}\|^2$ while the image on the right shows the outcome for (13)

where $U\,\mathrm{diag}(\sigma(X))V^*$ *is the singular value decomposition of* X.

Proof See Proposition 2.1 of [15]. □

7 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.

7.1 Convergence Basins

One of the drawbacks of using non-convex penalties is that overall performances might be poor when the problem to be solved is particularly ill-posed. The ideas that we presented in this note and that we want to highlight in the present section is that some issues related to non-convexity can be mitigated by means of adding a small convex "perturbation." In this subsection, we empirically demonstrate how the convergence basin can be greatly enlarged when $r_{\mu,\lambda}$ is employed instead of $r_{\mu,0}$, with λ small; i.e., we show that the reconstruction algorithm seems less prone to get stuck in undesired stationary points.

For this purpose, we constructed a "ground truth" $\mathbf{x}_0$ that is not a sparse signal, but most of its magnitude is concentrated in the largest coefficients (more precisely, roughly 90% of the signal is distributed on 5% of the entries). The sensing matrix A was here a 500×4096 random (with Gaussian distribution) matrix with normalized columns. The measurements $\mathbf{b}$ were considered perturbed by additive Gaussian white noise ϵ such that $\|\epsilon\| = 0.05\|A\mathbf{x}_0\|$.

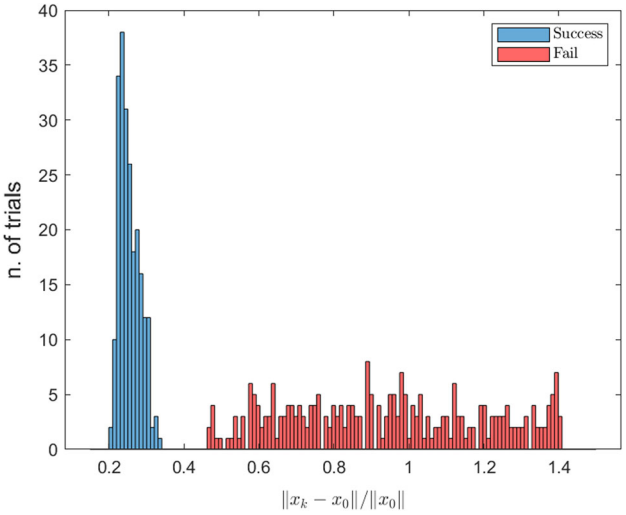

Fig. 7 Regrouping of the quantities $\|\mathbf{x}_k(\mathbf{x}_{SP}) - \mathbf{x}_0\|/\|\mathbf{x}_0\|$ for the different starting points $\mathbf{x}_{SP}$ generated. The cut between the blue and the red groups determined our definition of "success" (Color figure online)

We generated 500 different random (with uniform distribution) points belonging to the ball $B_{1.5}(\mathbf{x}_0)$ (with center $\mathbf{x}_0$ and radius 1.5, where $\|x_0\| = 1$) and used each of them as starting point for the FISTA algorithm, first to minimize the functional $\mathcal{Q}_2(\mu\|\cdot\|_0)(\mathbf{x}) + \|A\mathbf{x}-\mathbf{b}\|^2$ and then the relaxation (13), with $\lambda = 0.01$ and $\mu = 0.1$. The algorithm terminates when $\|\mathbf{x}_k - \mathbf{x}_{k+1}\| < 10^{-14}$ and the convergence point is simply called $\mathbf{x}_k$ in the below figures, or $\mathbf{x}_k(\mathbf{x}_{SP})$ if we want to specify the particular Starting Point $\mathbf{x}_{SP}$. The outcome is illustrated in Figs. 6 and 7. We say that a starting point $\mathbf{x}_{SP}$ "is successful" if $\mathbf{x}_k(\mathbf{x}_{SP})$ is such that $\|\mathbf{x}_k(\mathbf{x}_{SP}) - \mathbf{x}_0\|/\|\mathbf{x}_0\| \approx \|\mathbf{x}_k(\mathbf{x}_0) - \mathbf{x}_0\|/\|\mathbf{x}_0\| = 0.23$, since $\mathbf{x}_k(\mathbf{x}_0)$ likely is the best one could expect. The successful starting points are depicted in blue, the others in red. There

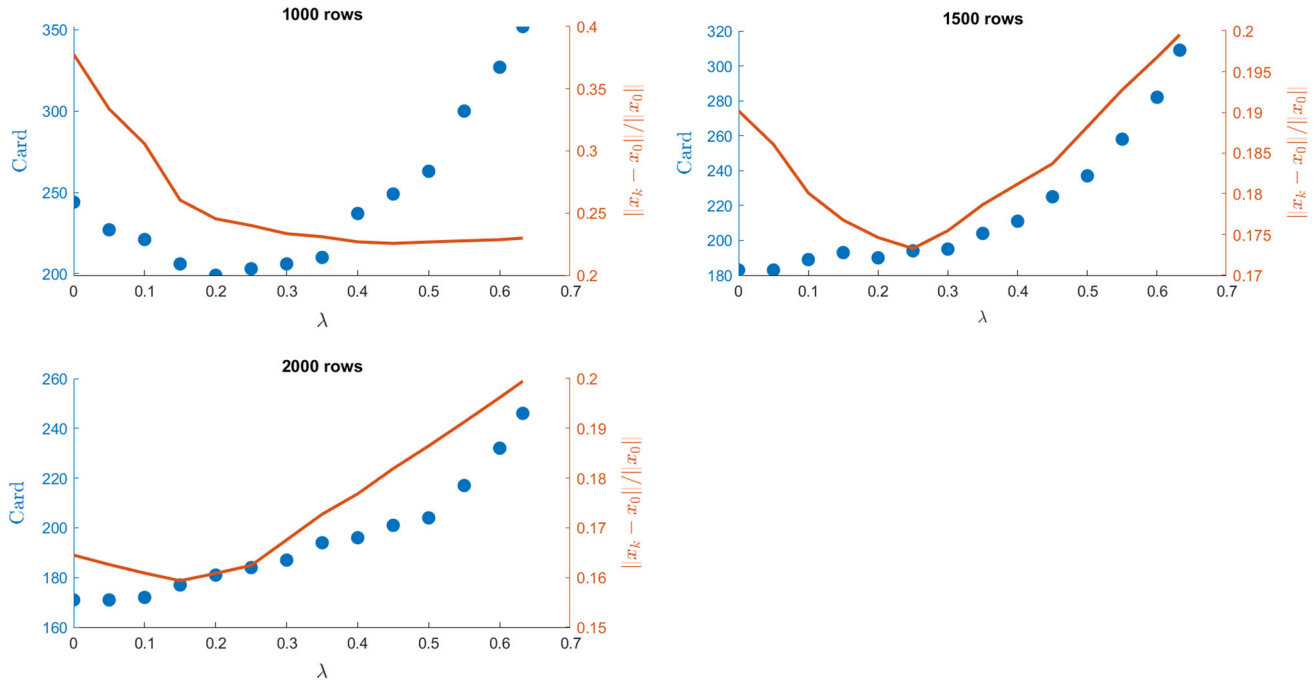

Fig. 8 Cardinality versus precision of the reconstructions in three different scenarios, as functions of λ

is a clear cut between what can be considered as "success" and what to be consider as "fail," as the next histogram shows;

Figure 6 illustrates the cloud of the starting points - angles are random for graphical representation purposes while distances to $\mathbf{x}_0$ are exact. Notice that the 0, often used as starting point, would lead to a fail when $\lambda = 0$ in the above example.

7.2 Well-Posedness vs Ill-Posedness

As already mentioned in the previous sections, the relaxation (13) shows its effectiveness with highly ill-posed problems. In this subsection we experimentally investigate more on this aspect. We consider $\mathbf{x}_0$ as in Sect. 7.1 and real random matrices with 1000, 1500 and 2000 rows, respectively (and 4096 normalized columns), since fewer rows leads to more ill-posedness.

The following pictures show the reconstruction precision in these three different scenarios as well as the cardinality of the retrieved approximate solutions along the segment $\sqrt{\mu} + \lambda/2 = \sqrt{0.1}$ for $\mu \in [0, 0.1]$. The rationale behind this parameter choice stems from the observation that the cardinality of the solution to

$$\arg\min_{\mathbf{x}} \mathcal{Q}_2(\mu\|\cdot\|_0)(\mathbf{x}) + \lambda\|\mathbf{x}\|_{\ell^1} + \|I\mathbf{x} - \mathbf{y}\|^2$$

is essentially determined by the number $\sqrt{\mu} + \lambda/2$, as seen by setting $\rho = 2$ in (23). When the identity I is replaced with a matrix A this might not be true any longer, but we still expect the cardinality to be roughly determined by the quantity $\sqrt{\mu} + \lambda/2$ (when A has normalized columns). The blue axis shows cardinality of the reconstruction and the red axis shows reconstruction misfit to ground truth for values of λ in the range 0 to $2\sqrt{0.1} \approx 0.63$ (where $\lambda = 0.63$ represents traditional LASSO (2)) (Fig. 8).

When the problem is ill-posed (1000 rows) we see the proposed crossover method at work: for λ in the range 0.2 to 0.4 the reconstruction precision is good while keeping the cardinality roughly constant. For bigger values of λ the reconstruction precision is still good, but at the cost of a higher cardinality. For 1500 both reconstruction quality and cardinality are optimal at $\lambda = 0.25$, while LASSO gives a significantly worse output with respect to both parameters. With 2000 rows the problem is well posed enough that optimal performance is found for very small λ, i.e., one may just as well skip the ℓ^1-penalty and only work with (3), as reported previously in [13]. Summing up, the $r_{\mu,\lambda}$-penalty does a better job than ℓ^1 in the entire range.

7.3 Random Matrices

In this section, we compare the robustness to local minima of the relaxations (2), (3) and (13). Note that (2) and (3) are special cases of (13), obtained by letting λ or μ equal to 0 (by Theorem 2.1).

We generated A-matrices of size 100×200 by drawing the columns from a uniform distribution over the unit sphere in $\mathbb{R}^{100}$, and the vector $\mathbf{x}_0$ was selected to have 10 random nonzero elements with random magnitudes between 2 and

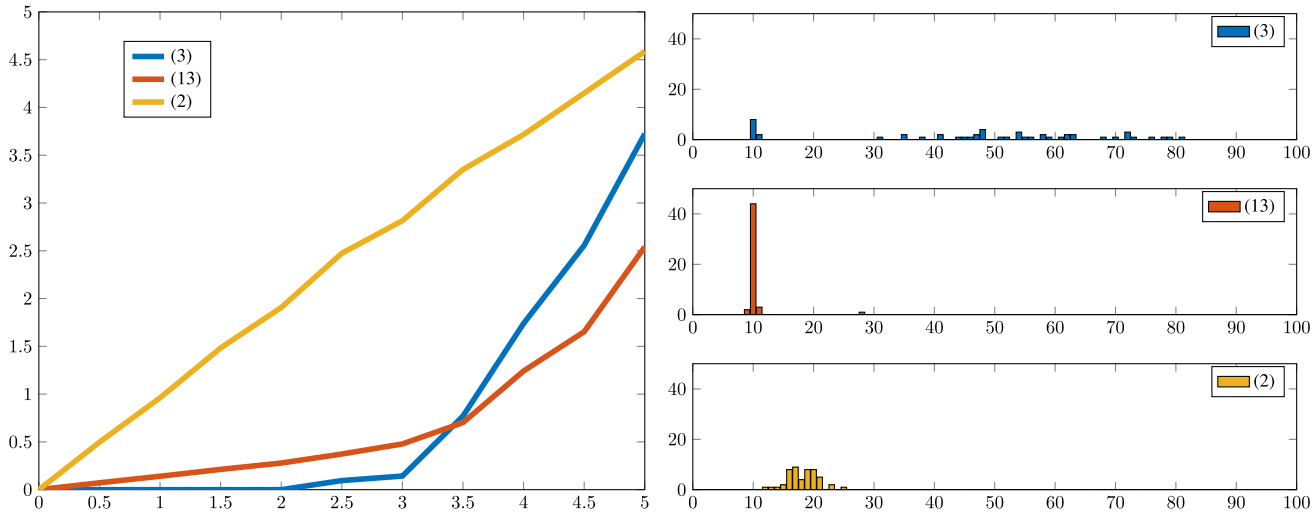

Fig. 9 Left: noise level $\|\epsilon\|$ (x-axis) versus distance $\|\mathbf{x} - \mathbf{x}_S\|$ (y-axis) between the obtained solution $\mathbf{x}$ and the oracle solution $\mathbf{x}_S$ for the three methods (3), (13) and (2). Right: cardinality of the retrieved vectors (x-axis) vs number of retrieved vectors with that cardinality (y-axis)

4, resulting in $\|\mathbf{x}_0\| \approx 10$. We then computed $\mathbf{b} = A\mathbf{x}_0 + \epsilon$ for different values of random noise with $\|\epsilon\|$ ranging from 0 to 5. For (3) we used $\mu = 1$ and for (2) we used $\lambda_{\ell^1} = 2\frac{\sqrt{2\log(200)}}{\sqrt{200}}\|\epsilon\| \approx 0.5\|\epsilon\|$; see [13] for the rationale behind these choices. For (13) we again chose $\mu = 1$ but used $\lambda = \lambda_{\ell^1}/6$. Figure 9 plots $\|\mathbf{x} - \mathbf{x}_S\|$ for the estimated $\mathbf{x}$ with the three methods, as a function of $\|\epsilon\|$; $\mathbf{x}_S$ is here the oracle solution to the linear system of equations $A\mathbf{x} = \mathbf{b}$ [13]. Both (3) and (13) do better than traditional ℓ^1 in the entire range, (3) finds $\mathbf{x}_S$ with 100% accuracy until around $\|\epsilon\| \approx 3$, where (13) starts to perform better. This is likely due to the fact that the small ℓ^1 term helps the (non-convex) minimization of (13) 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, which is known to be close to many local minima (we usually use 0 as initial point). The histograms to the right in Fig. 9 show the cardinality of the found solution. Adding the $\lambda\|\mathbf{x}\|_1$ enabled the algorithm to avoid almost all of these high cardinality solutions, in perfect harmony with Theorem 4.2 and Fig. 3.

7.4 Point-Set Registration with Outliers

Next we consider registration of 2D point clouds. We assume that we have a set of model points $\{\mathbf{p}_i\}_{i=1}^N$ that should be registered to $\{\mathbf{q}_i\}_{i=1}^N$ by minimizing $\sum_{i=1}^N \|sR\mathbf{p}_i + \mathbf{t} - \mathbf{q}_i\|^2$. 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 $a, b, \mathbf{t}$, we can by column-stacking them write the problem as $\|M\mathbf{y} - \mathbf{v}\|^2$, where the vector $\mathbf{y}$ contains the unknowns $a, b, \mathbf{t}$. We further assume that the point matches contain outliers that needs to be removed. Therefore we add a sparse vector $\mathbf{x}$ whose nonzero entries allows the solution to have large errors. We thus want to solve

$$\min_{\mathbf{x},\mathbf{y}} \mu\|\mathbf{x}\|_0 + \|M\mathbf{y} - \mathbf{v} + \mathbf{x}\|^2. \tag{24}$$

The minimization over $\mathbf{y}$ can be carried out in closed form by noting that $\mathbf{y} = (M^tM)^{-1}M^t(\mathbf{v} - \mathbf{x})$. Inserting into (24) which gives the objective function (1), where $A = I - M(M^TM)^{-1}M^t$ and $\mathbf{b} = A\mathbf{v}$. The matrix A is a projection onto the complement of the column space of M, and therefore has a four-dimensional null space.

Figure 10 shows the results of a synthetic experiment with 500 problem instances. The data were 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 $\mathbf{t}$ to be 2D Gaussian with mean $(0, 0)$ and covariance $5I$. We compare the three relaxations (2) with $\lambda = 2$, (3) with $\mu = 1$ and (13) with $\mu = 1$ and $\lambda = 0.5$. (The reason for using $\lambda = 2$ in (2) and $\mu = 1$ in (3) is that this gives the same threshold in the corresponding proximal operators.)

All methods where initialized with the least squares solution. In the left histogram of Fig. 10, we plot the data fit with respect to the inlier residuals (corresponding to the first 60 points, that supports the larger hypothesis). In other words we reorder the data points so that $(\|sR\mathbf{p}_i + \mathbf{t} - \mathbf{q}_i\|)_{i=1}^{100}$ is

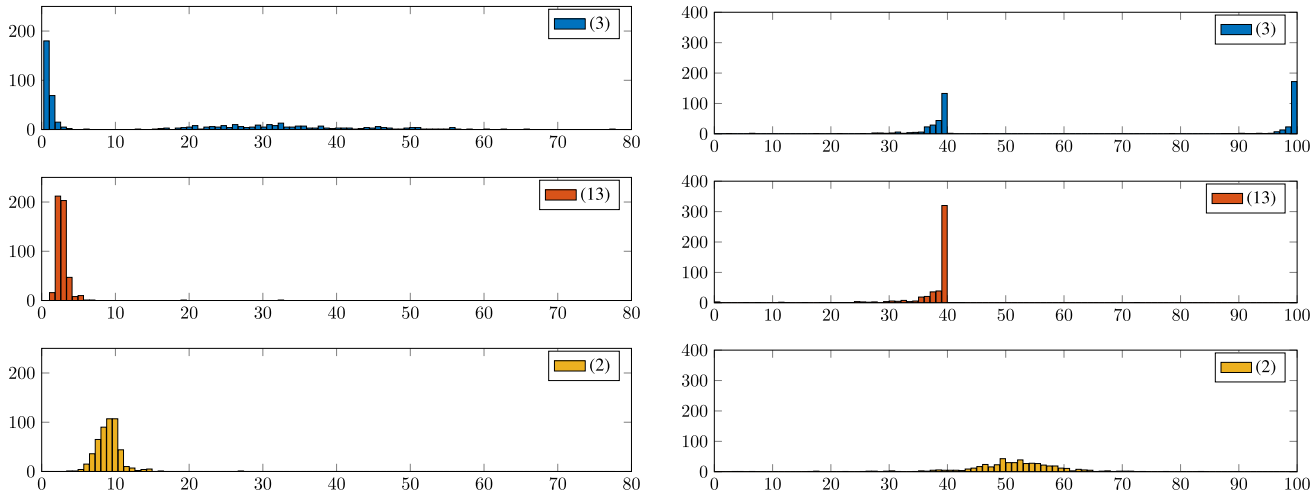

Fig. 10 Results from the synthetic registration experiment. Left: Data fit of the resulting estimation to the true inliers. Right: Number of estimated outliers

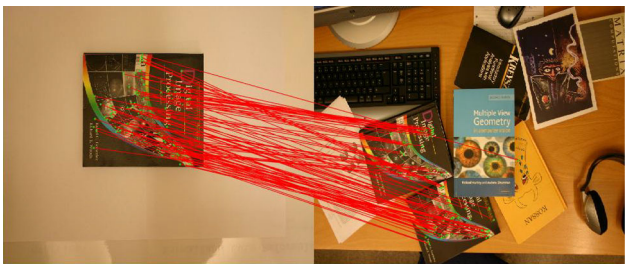

Fig. 11 Matches between two of the images used in Fig. 12

increasing, and then compute $\sum_{i=1}^{60} \|sR\mathbf{p}_i + \mathbf{t} - \mathbf{q}_i\|^2$. The histograms were produced with 500 trials, and a low value on the x-axis thus indicates a good fit. In the right histogram, the x-axis displays the number of residuals determined to be outliers (via a thresholding rule), and thus, a value near 40 indicates success. When starting from the least squares initialization the formulation (3) 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 ℓ^1 norm formulation (2) due to its lack of bias. The added ℓ^1 term helps the sequence generated by the minimization of (13) to converge to the correct solution with a good data fit.

Note that the number of outliers are in many cases is 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 Fig. 11, to generate the two point sets $\{\mathbf{p}_i\}_{i=1}^N$ and $\{\mathbf{q}_i\}_{i=1}^N$. We then registered the points sets using the formulations (3) with $\mu = 20^2$ and (2) with $\lambda = 10$ (which in both cases corresponds to a 20 pixel outlier threshold in a 3072×2048 image). For (13) we used $\mu = 20^2$ and $\lambda = 5$.

The results are shown in Fig. 12. In the first problem instance (first row) we used an image which generates one strong hypothesis. Here both (13) and (2) produce good results. In contrast, (3) 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 (13). As a result, the registration obtained by (13) is better than that of (3) and the number of determined inliers is larger.

7.5 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. [16] and let

Fig. 12 Results from the two real registration experiment. From left to right: (3), (13), (2). Red means that the point was classified as outlier, green inlier. White frame shows registration of the model book under the estimated transformation (Color figure online)

$$X = \begin{bmatrix} X_1 \\ Y_1 \\ Z_1 \\ \vdots \\ X_F \\ Y_F \\ Z_F \end{bmatrix} \text{ and } X^{\#} = \begin{bmatrix} X_1 & Y_1 & Z_1 \\ \vdots & \vdots & \vdots \\ X_F & Y_F & Z_F \end{bmatrix}, \tag{25}$$

where X_i,Y_i,Z_i are $1 \times m$ matrices containing the x-,y- and 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 2×3 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 [5] with r deformation modes, the matrix $X^{\#}$ can be written $X^{\#} = CB$, where B is $r \times 3m$, and therefore $\text{rank}(X^{\#}) = r$. We search for the matrix $X^{\#}$ of rank r that minimizes the residual error $\|PX - M\|_F^2$.

In Fig. 14 we compare the three relaxations

$$r_{0,\mu}(\boldsymbol{\sigma}(X^{\#})) + \|RX - M\|_F^2, \tag{26}$$

$$r_{\mu,\lambda}(\boldsymbol{\sigma}(X^{\#})) + \|RX - M\|_F^2. \tag{27}$$

$$2\sqrt{\mu}\|X^{\#}\|_* + \|RX - M\|_F^2, \tag{28}$$

on the four MOCAP sequences displayed in Fig. 13, obtained from [16]. These consist of real motion capture data and therefore the ground truth solution is only approximatively of low rank. Figure 14 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 (27) we kept $\lambda = 5$ fixed.) All three formulations were given the same (random) starting solution.

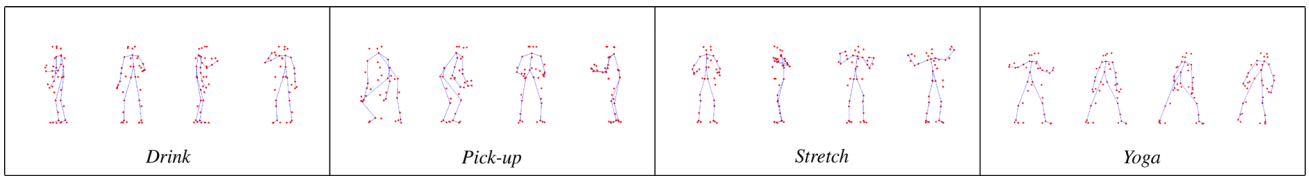

Fig. 13 Four images from each of the MOCAP datasets

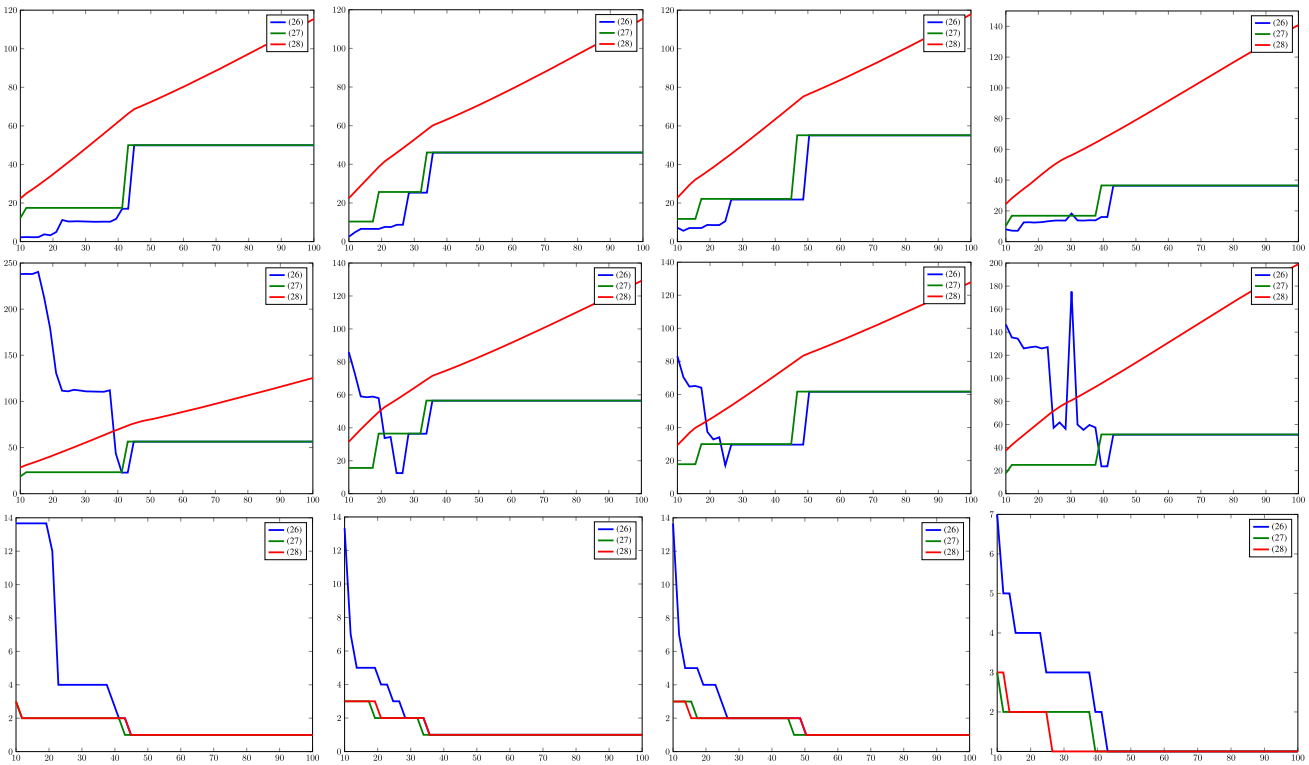

Fig. 14 Result of the four MOCAP experiments (columns 1–4). Top: Regularization strength μ versus data fit $\|RX - M\|_F$. Middle: Regularization strength μ versus ground truth distance $\|X - X_{gt}\|_F$. Bottom: Regularization strength μ versus $\text{rank}(X^{\#})$

The same tendencies are visible for all four sequences. While (26) generally gives a better data fit than (28), due to the nuclear norms shrinking bias, the distance to the ground truth is larger for low values of μ or equivalently large ranks where the problem gets ill-posed. The relaxed functional (27) consistently outperforms (28) in terms of both data fit and distance to ground truth. In addition, its performance is similar to (26) for high values of μ while it does not exhibit the same unstable behavior for high ranks.

Funding Open access funding provided by Lund University.

References

1. Attouch, H., Bolte, J., Svaiter, B.F.: Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized gauss-seidel methods. Math. Program. **137**(1–2), 91–129 (2013)
2. Blanchard, J.D., Cartis, C., Tanner, J.: Compressed sensing: how sharp is the restricted isometry property? SIAM Rev. **53**(1), 105–125 (2011)
3. Blumensath, T., Davies, M.E.: Iterative thresholding for sparse approximations. J. Fourier Anal. Appl. **14**(5–6), 629–654 (2008)
4. Bredies, K., Lorenz, D.A., Reiterer, S.: Minimization of non-smooth, non-convex functionals by iterative thresholding. J. Optim. Theory Appl. **165**(1), 78–112 (2015)
5. Bregler, C., Hertzmann, A., Biermann, H.: Recovering non-rigid 3d shape from image streams. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2000)
6. Breheny, P., Huang, J.: Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. Ann Appl Stat **5**(1), 232–253 (2011)
7. Candès, E.J., Li, X., Ma, Y., Wright, J.: Robust principal component analysis? J. ACM **58**(3), 11:1-11:37 (2011)
8. Candes, E.J., Romberg, J.K., Tao, T.: Stable signal recovery from incomplete and inaccurate measurements. Commun. Pure Appl. Math. **59**(8), 1207–1223 (2006)
9. Candes, E.J., Tao, T.: Near-optimal signal recovery from random projections: Universal encoding strategies? IEEE Trans. Inf. Theory **52**(12), 5406–5425 (2006)
10. Candes, E.J., Wakin, M.B., Boyd, S.P.: Enhancing sparsity by reweighted l^1 minimization. J. Fourier Anal. Appl. **14**(5–6), 877–905 (2008)
11. Carlsson, M.: On convexification/optimization of functionals including an l2-misfit term. arXiv preprint arXiv:1609.09378 (2016)
12. Carlsson, M.: On convex envelopes and regularization of non-convex functionals without moving global minima. J. Optim. Theory Appl. **183**, 66–84 (2019)
13. Carlsson, M., Gerosa, D., Olsson, C.: An unbiased approach to compressed sensing. Inverse Prob. **36**(11), 115014 (2020)
14. Carlsson, M., Gerosa, D., Olsson, C.: An un-biased approach to low rank recovery. arXiv preprint arXiv:1909.13363 (2019)
15. Carlsson, M., Gerosa, D.: On phase retrieval via matrix completion and the estimation of low rank PSD matrices. Inverse Prob. **36**(1), 015006 (2020)
16. Dai, Y., Li, H., He, M.: A simple prior-free method for non-rigid structure-from-motion factorization. Int. J. Comput. Vis. **107**(2), 101–122 (2014)
17. Donoho, D.L., Elad, M.: Optimally sparse representation in general (non-orthogonal) dictionaries via ℓ^1 minimization. Proc. Natl Acad. Sci. USA **100**, 2197–202 (2002)
18. Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. **96**(456), 1348–1360 (2001)
19. Fan, J., Xue, L., Zou, H.: Strong oracle optimality of folded concave penalized estimation. Ann. Stat. **42**(3), 819–849 (2014)
20. Larsson, V., Olsson, C.: Convex low rank approximation. Int. J. Comput. Vis. **120**(2), 194–214 (2016)
21. Loh, P.-L., Wainwright, M.J.: Regularized m-estimators with nonconvexity: statistical and algorithmic theory for local optima. J. Mach. Learn. Res. **16**(19), 559–616 (2015)
22. Loh, P.-L., Wainwright, M.J., et al.: Support recovery without incoherence: a case for nonconvex regularization. Ann. Stat. **45**(6), 2455–2482 (2017)
23. Mazumder, R., Friedman, J.H., Hastie, T.: Sparsenet: coordinate descent with nonconvex penalties. J. Am. Stat. Assoc. **106**(495), 1125–1138 (2011)
24. Olsson, C., Carlsson, M., Andersson, F., Larsson, V.: Non-convex rank/sparsity regularization and local minima. Proc. Int. Conf. Comput. Vis. (2017)
25. Pan, Z., Zhang, C.: Relaxed sparse eigenvalue conditions for sparse estimation via non-convex regularized regression. Pattern Recogn. **48**(1), 231–243 (2015)
26. Recht, B., Fazel, M., Parrilo, P.A.: Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. SIAM Rev. **52**(3), 471–501 (2010)
27. Tropp, J.A.: Just relax: Convex programming methods for identifying sparse signals in noise. IEEE Trans. Inf. Theory **52**(3), 1030–1051 (2006)
28. Tropp, J.A.: Convex recovery of a structured signal from independent random linear measurements. In: Sampling Theory: A Renaissance, pp. 67–101 (2015)
29. Wang, Y., Yin, W., Zeng, J.: Global convergence of admm in nonconvex nonsmooth optimization. J. Sci. Comput. **78**(1), 29–63 (2019)
30. Wang, Z., Liu, H., Zhang, T.: Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. Ann. Stat. **42**(6), 2164–2201 (2014)
31. Zhang, C.-H., Zhang, T.: A general theory of concave regularization for high-dimensional sparse estimation problems. Stat. Sci. 576–593 (2012)
32. Zou, H., Li, R.: One-step sparse estimates in nonconcave penalized likelihood models. Ann. Stat. **36**(4), 1509–1533 (2008)
33. Zhang, H., Yin, W., Cheng, L.: Necessary and sufficient conditions of solution uniqueness in 1-norm minimization. J. Optim. Theory Appl. **164**, 109–122 (2014)
34. Zhang, C.-H.: Nearly unbiased variable selection under minimax concave penalty. Ann. Stat. **38**(2), 894–942 (2010)

Daniele Gerosa is a senior PhD student in Mathematics at Lund University (Sweden). He holds a MSc and a BSc in Mathematics from Padova University (Italy).

Marcus Carlsson received his MSc degree in Mathematics in 2002, and his PhD degree in Mathematics in 2007, both from Lund University, Lund, Sweden. From 2007 to 2010, he was a Research Associate Professor with the Department of Mathematics and with the Geo-Mathematical Imaging Group at Purdue University, Indiana, USA. In 2011, he worked as Research Professor at the Department of Mathematics at Universidad de Santiago de Chile. Since 2012, he has held a tenure track research position at the Centre for Mathematical Sciences at Lund University.

Carl Olsson received his MSc degree in Electrical Engineering in 2004, and his PhD degree in Mathematics in 2009, both from Lund University, Lund, Sweden. He was Assistant Professor at Lund University (2009–2014) and since 2014 he is there Associate Professor. Since 2017 he is also Senior Researcher at Chalmers University of Technology, Göteborg, Sweden.