Non-Convex Rank/Sparsity Regularization and Local Minima

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

[1]Department of Electrical Engineering
Chalmers University of Technology

[2]Centre for Mathematical Sciences
Lund University

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

Abstract

This paper considers the problem of recovering either a low rank matrix or a sparse vector from observations of linear combinations of the vector or matrix elements. Recent methods replace the non-convex regularization with ℓ_1 or nuclear norm relaxations. It is well known that this approach recovers near optimal solutions if a so called restricted isometry property (RIP) holds. On the other hand it also has a shrinking bias which can degrade the solution.

In this paper we study an alternative non-convex regularization term that does not suffer from this bias. Our main theoretical results show that if a RIP holds then the stationary points are often well separated, in the sense that their differences must be of high cardinality/rank. Thus, with a suitable initial solution the approach is unlikely to fall into a bad local minimum. Our numerical tests show that the approach is likely to converge to a better solution than standard ℓ_1/nuclear-norm relaxation even when starting from trivial initializations. In many cases our results can also be used to verify global optimality of our method. [1]

1. Introduction

Sparsity penalties are important priors for regularizing linear systems. Typically one tries to solve a formulation that minimizes a trade-off between sparsity and residual error such as

$$\mu \operatorname{card}(\mathbf{x}) + \|A\mathbf{x} - \mathbf{b}\|^2, \tag{1}$$

where $\operatorname{card}(\mathbf{x})$ is the number of non-zero elements in $\mathbf{x}$, and the matrix A is of size $m \times n$. Direct minimization of (1) is generally considered difficult because of the properties of the card function, which is non-convex and discontinuous. The method that has by now become the standard approach is to replace $\operatorname{card}(\mathbf{x})$ with the convex ℓ_1 norm $\|\mathbf{x}\|_1$ [32, 31, 8, 9, 14]. This choice can be justified with the ℓ_1 norm being the convex envelope of the card function on the set $\{\mathbf{x}; \|\mathbf{x}\|_\infty \leq 1\}$. Furthermore, strong performance guarantees can be derived [8, 9] if A obeys a RIP

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

for all vectors $\mathbf{x}$ with $\operatorname{card}(\mathbf{x}) \leq c$, where c is a bound on the number of non-zero terms in the sought solution. The ℓ_1 approach suffers from a shrinking bias since it penalizes both small elements of $\mathbf{x}$, assumed to stem from measurement noise, and large elements, assumed to make up the true signal, equally. In some sense the suppression of noise also requires an equal suppression of signal. Therefore non-convex alternatives able to penalize small components proportionally harder have been considered [13, 10]. Convergence to the global optimum is however not guaranteed.

This paper considers the non-convex relaxation

$$f(x) = r_\mu(\mathbf{x}) + \|A\mathbf{x} - \mathbf{b}\|^2, \tag{3}$$

where $r_\mu(\mathbf{x}) = \sum_i \left(\mu - \max(\sqrt{\mu} - |x_i|, 0)^2\right)$. Figure 1 shows one dimensional illustrations of the card-function, ℓ_1-norm and r_μ term. The regularizer r_μ is a particular case of the minmax concave penalty (MCP) introduced in [34]. Optimization of MCP-regularized systems have been addressed in a number of works e.g. [29, 5, 34], typically using methods only guaranteeing convergence to a stationary point. It can be shown [18, 30, 19] that the convex envelope of

$$\mu \operatorname{card}(\mathbf{x}) + \|\mathbf{x} - \mathbf{z}\|^2, \tag{4}$$

where $\mathbf{z}$ is some given vector, is

$$r_\mu(\mathbf{x}) + \|\mathbf{x} - \mathbf{z}\|^2. \tag{5}$$

Note that similarly to the card function r_μ does not penalize elements that are larger than $\sqrt{\mu}$. In fact it is easy to show that the minimizer $\mathbf{x}^*$ of both (4) and (5) is given by thresholding of $\mathbf{z}$, that is, $x_i^* = z_i$ if $|z_i| > \sqrt{\mu}$ and $x_i^* = 0$ if $|z_i| < \sqrt{\mu}$. If there is an i such that $|z_i| = \sqrt{\mu}$ then the

[1] This work has been funded by the Swedish Research Council (grants no. 2012-4213, 2015-03780 and 2016-04445) and the Swedish Foundation for Strategic Research (Semantic Mapping and Visual Navigation for Smart Robots).

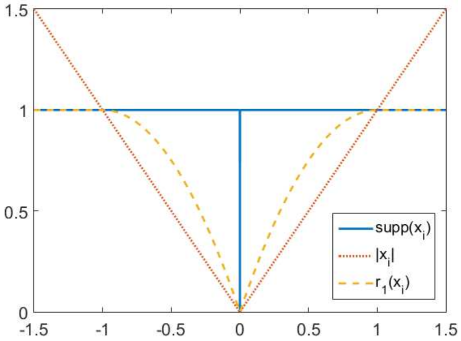

Figure 1: One dimensional illustrations of the three regularization terms (when $\mu = 1$).

minimizer is not unique. In (4) x_i^* can take either the value 0 or $\sqrt{\mu}$ and in (5) any convex combination of these.

Assuming that A fulfills (2) it is natural to wonder about convexity properties of (3). Intuitively $\|A\mathbf{x}\|^2$ seems to behave like $\|\mathbf{x}\|^2$ which combined with $r_\mu(\mathbf{x})$ only has one local minimum. In this paper we make this reasoning formal and study the stationary points of (3). We show that if $\mathbf{x}_s$ is a stationary point of (3) and the elements of the vector $\mathbf{z} = (I - A^T A)\mathbf{x}_s + A^T\mathbf{b}$ fulfill $|z_i| \notin [\sqrt{\mu}(1-\delta_c), \frac{\sqrt{\mu}}{1-\delta_c}]$ then for any other stationary point $\mathbf{x}_s'$ we have $\operatorname{card}(\mathbf{x}_s - \mathbf{x}_s') > c$. A simple consequence is that if we for example find such a local minimizer with $\operatorname{card}(\mathbf{x}_s) < c/2$ then this is the sparsest possible one.

The meaning of the vector $\mathbf{z}$ can in some sense be understood by seeing that the stationary point $\mathbf{x}_s$ fulfills $\mathbf{x}_s \in \arg\min_{\mathbf{x}} r_\mu(\mathbf{x}) + \|\mathbf{x} - \mathbf{z}\|^2$ (see Section 3). Hence $\mathbf{x}_s$ can be obtained through thresholding of the vector $\mathbf{z}$. Our results essentially state that if the elements $|z_i|$ are not too close to the threshold $\sqrt{\mu}$ then $\operatorname{card}(\mathbf{x}_s - \mathbf{x}_s') > c$ holds for all other stationary points $\mathbf{x}_s'$.

In a recent related work [21] the stationary points of MCP-regularized linear least squares are studied. Using the the RIP-like notion of *restricted strong convexity* they show that if A is *sub-Gaussian* (and some additional technical assumptions hold) then with high probability there will be a unique stationary point. In two very recent papers [30, 11] the relationship between (both local and global) minimizers of (3) and (1) is studied. Among other things [11] shows that if $\|A\| \leq 1$ then any local minimizer of (3) is also a local minimizer of (1), and that their global minimizers coincide. Hence results about the stationary points of (3) are also relevant to the original discontinuous objective (1).

The theory of rank minimization largely parallels that of sparsity with the elements x_i of the vector $\mathbf{x}$ replaced by the singular values $\sigma_i(X)$ of the matrix X. Typically we want to solve a problem of the type

$$\mu \operatorname{rank}(X) + \|\mathcal{A}X - \mathbf{b}\|^2, \tag{6}$$

where $\mathcal{A} : \mathbb{R}^{m\times n} \mapsto \mathbb{R}^p$ is some linear operator on the set of $m \times n$ matrices. In this context the standard approach is to replace the rank function with the convex nuclear norm $\|X\|_* = \sum_i \sigma_i(X)$ [28, 6]. It was first observed that this is the convex envelope of the rank function over the set $\{X; \sigma_1(X) \leq 1\}$ in [15]. In [28] the notion of RIP was generalized to the matrix setting requiring that $\mathcal{A}$ is a linear operator $\mathbb{R}^{m\times n} \rightarrow \mathbb{R}^k$ fulfilling

$$(1-\delta_r)\|X\|_F^2 \leq \|\mathcal{A}X\|^2 \leq (1+\delta_r)\|X\|_F^2, \tag{7}$$

for all X with $\operatorname{rank}(X) \leq r$. Since then a number of generalizations that give performance guarantees for the nuclear norm relaxation have appeared [26, 6, 7]. Non-convex alternatives have also been shown to improve performance [25, 24].

Analogous to the vector setting it was recently shown [19] that the convex envelope of $\mu \operatorname{rank}(X) + \|X - M\|_F^2$, is given by

$$r_\mu(\boldsymbol{\sigma}(X)) + \|X - M\|_F^2, \tag{8}$$

where $\boldsymbol{\sigma}(X)$ is the vector of singular values of X. In [1] an efficient fixed-point algorithm is developed for objective functions of the type $r_\mu(\boldsymbol{\sigma}(X)) + q\|X - M\|_F^2$ with linear constraints. The approach is illustrated to work well even when $q < 1$ which gives a non-convex objective.

In this paper we consider

$$F(X) = r_\mu(\boldsymbol{\sigma}(X)) + \|\mathcal{A}X - \mathbf{b}\|^2, \tag{9}$$

where $\mathcal{A}$ obeys (7). The objective (9) is a special case of the MCP framework considered in e.g [33, 22, 23]. Our main result states that if X_s is a stationary point of (9) and $Z = (I - \mathcal{A}^*\mathcal{A})X_s + \mathcal{A}^*\mathbf{b}$ has no singular values in the interval $[\sqrt{\mu}(1-\delta_r), \frac{\sqrt{\mu}}{1-\delta_r}]$ then for any other stationary point we have $\operatorname{rank}(X_s - X_s') > r$.

A number of recent papers [3, 27, 16] parametrize X using UV^T and studies the local minimizers of $\|\mathcal{A}(UV^T) - b\|^2$ under the RIP constraint. They essentially bound the distance between the global and any local solution in terms of the residual error. As a consequence they are able to show that (with hight probability) all local minima are close to the global solution (and in the noise free case they are all optimal). In contrast to our results this cannot be used to rule out the existence of multiple minima (in the noisy case) and additionally requires that the rank of the sought solution is known beforehand.

2. Notation and Preliminaries

In this section we introduce some preliminary material and notation. In general we will use boldface to denote a vector $\mathbf{x}$ and its ith element x_i. By $\|\mathbf{x}\|$ we denote the standard euclidean norm $\|\mathbf{x}\| = \sqrt{\mathbf{x}^T\mathbf{x}}$. We use $\sigma_i(X)$ to denote the ith singular value of a matrix X. The vector of all singular values is denoted $\boldsymbol{\sigma}(X)$. A diagonal matrix with diagonal elmenents $\mathbf{x}$ will be denoted $D_{\mathbf{x}}$. The scalar product is defined as $\langle X, Y \rangle = \operatorname{tr}(X^T Y)$, where tr is the trace function, and the Frobenius norm $\|X\|_F = \sqrt{\langle X, X \rangle} = \sqrt{\sum_{i=1}^{n} \sigma_i^2(X)}$. The adjoint of a linear matrix operator $\mathcal{A}$ is denoted $\mathcal{A}^*$. For functions taking values in $\mathbb{R}$ such as r_μ we will frequently use the convention that $r_\mu(\mathbf{x}) = \sum_i r_\mu(x_i)$.

The function $g(x) = r_\mu(x) + x^2$ will be useful when considering stationary points, since it is convex with a well defined sub-differential. We can write g as

$$g(x) = \begin{cases} \mu + x^2 & |x| \geq \sqrt{\mu} \\ 2\sqrt{\mu}|x| & 0 \leq |x| \leq \sqrt{\mu} \end{cases}. \tag{10}$$

Its sub-differential is given by

$$\partial g(x) = \begin{cases} \{2x\} & |x| \geq \sqrt{\mu} \\ \{2\sqrt{\mu}\operatorname{sign}(x)\} & 0 < |x| \leq \sqrt{\mu} \\ [-2\sqrt{\mu}, 2\sqrt{\mu}] & x = 0 \end{cases}. \tag{11}$$

Note that the sub-differential consists of a single point for each $x \neq 0$. By $\partial g(\mathbf{x})$ we mean the set of vectors $\{\mathbf{z}; z_i \in \partial g(x_i), \ \forall i\}$. Figure 2 illustrates g and its sub-differential. For the matrix case we similarly define $G(X) = r_\mu(\boldsymbol{\sigma}(X)) + \|X\|_F^2$. It can be shown [20] that a matrix Z is in the sub-differential of G at X if and only if

$$Z = U D_{\mathbf{z}} V^T, \text{ where } \mathbf{z} \in \partial g(\boldsymbol{\sigma}(X)) \tag{12}$$

and $U D_{\boldsymbol{\sigma}(X)} V^T$ is the SVD of X.

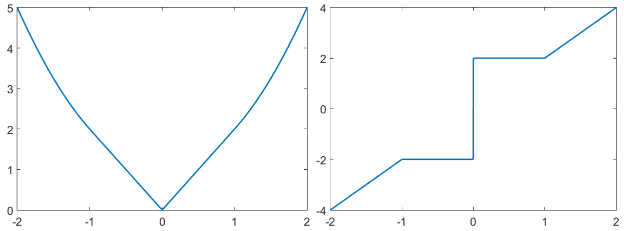

Figure 2: The function $g(x)$ (left) and its sub-differential $\partial g(x)$ (right). Note that the sub-differential contains a unique element everywhere except at $x = 0$.

In Section 4 we utilize the notion of doubly substochastic (DSS) matrices [2]. A matrix M is DSS if its rows and columns fulfill $\sum_i |m_{ij}| \leq 1$ and $\sum_j |m_{ij}| \leq 1$. The DSS matrices are closely related to permutations. Let π denote a permutation and $M_{\pi,\mathbf{v}}$ the matrix with elements $m_{i,\pi(i)} = v_i$ and zeros otherwise. It is shown in [2] (Lemma 3.1) that an $m \times m$ matrix is DSS if and only if it lies in the convex hull of the set $\{M_{\pi,\mathbf{v}}; \pi \text{ is a permutation}, \ |v_i| = 1 \ \forall i\}$. The result is actually proven for matrices with complex entries, but the proof is identical for real matrices.

3. Sparsity Regularization

In this section we consider stationary points of the proposed sparsity formulation (3). The function f can equivalently be written

$$f(\mathbf{x}) = g(\mathbf{x}) + \mathbf{x}^T(A^T A - I)\mathbf{x} - 2\mathbf{x}^T A^T \mathbf{b} + \mathbf{b}^T \mathbf{b}. \tag{13}$$

Taking derivatives we see that the stationary points solve

$$2(I - A^T A)\mathbf{x}_s + 2A^T \mathbf{b} \in \partial g(\mathbf{x}_s). \tag{14}$$

The following lemma clarifies the connection between a stationary point $\mathbf{x}_s$ and the vector $\mathbf{z} = (I - A^T A)\mathbf{x}_s + A^T \mathbf{b}$.

Lemma 3.1. *The point $\mathbf{x}_s$ is stationary in f if and only if $2\mathbf{z} \in \partial g(\mathbf{x}_s)$ and if and only if*

$$\mathbf{x}_s \in \arg\min_{\mathbf{x}} r_\mu(\mathbf{x}) + \|\mathbf{x} - \mathbf{z}\|^2. \tag{15}$$

Proof. By (14) we know that $\mathbf{x}_s$ is stationary in f if and only if $2\mathbf{z} \in \partial g(\mathbf{x}_s)$. Similarly, inserting $A = I$ and $\mathbf{b} = \mathbf{z}$ in (14) shows that $\mathbf{x}_s$ is stationary in $r_\mu(\mathbf{x}) + \|\mathbf{x} - \mathbf{z}\|^2$ if and only if $2\mathbf{z} \in \partial g(\mathbf{x}_s)$. Since $r_\mu(\mathbf{x}) + \|\mathbf{x} - \mathbf{z}\|^2$ is convex in $\mathbf{x}$, this is equivalent to solving (15). □

The above result shows that stationary points of f are sparse approximations of $\mathbf{z}$ in the sense that small elements are suppressed. The elements of $\mathbf{x}_s$ are either zero or have magnitude larger than $\sqrt{\mu}$ assuming that the vector $\mathbf{z}$ has no elements that are precisely $\pm\sqrt{\mu}$.

3.1. Stationary points under the RIP constraint

We now assume that A is a matrix fulfilling the RIP (2) and write

$$f(\mathbf{x}) = g(\mathbf{x}) - \delta_c \|\mathbf{x}\|^2 + h(\mathbf{x}) + \|\mathbf{b}\|^2, \tag{16}$$

where

$$h(\mathbf{x}) = \delta_c \|\mathbf{x}\|^2 + \left(\|A\mathbf{x}\|^2 - \|\mathbf{x}\|^2\right) - 2\mathbf{x}^T A^T \mathbf{b}. \tag{17}$$

The term $\|\mathbf{b}\|^2$ is constant with respect to $\mathbf{x}$ and we can therefore drop it without affecting the optimizers. The point $\mathbf{x}$ is a stationary point of f if $2\delta_c \mathbf{x} - \nabla h(\mathbf{x}) \in \partial g(\mathbf{x})$, that is there is a vector $2\mathbf{z} \in \partial g(\mathbf{x})$ such that $2\delta_c \mathbf{x} - \nabla h(\mathbf{x}) = 2\mathbf{z}$.

Our goal is now to find constraints that assure that this system of equations have only one sparse solution. Before getting into the details we outline the overall idea. For simplicity consider two differentiable strictly convex functions $\tilde{h}$ and $\tilde{g}$. Their sum is minimized by the stationary point $\mathbf{x}_s$ fulfilling $-\nabla \tilde{h}(\mathbf{x}_s) = \nabla \tilde{g}(\mathbf{x}_s)$. Since $\tilde{g}$ is strictly convex its directional derivative $\langle \nabla \tilde{g}(\mathbf{x}_s + t\mathbf{v}), \mathbf{v} \rangle$ is increasing for all directions $\mathbf{v} \neq 0$. Similarly $\langle -\nabla \tilde{h}(\mathbf{x}_s + t\mathbf{v}), \mathbf{v} \rangle$ is decreasing for all $\mathbf{v} \neq 0$ since $-\tilde{h}$ is strictly concave. Therefore $\langle -\nabla \tilde{h}(\mathbf{x}_s + t\mathbf{v}), \mathbf{v} \rangle < \langle \nabla \tilde{g}(\mathbf{x}_s + t\mathbf{v}), \mathbf{v} \rangle$ which means that $\mathbf{x}_s$ is the only stationary point. In what follows we will estimate the growth of the directional derivatives of the functions involved in (16) in order to show a similar contradiction. For the function h we do not have convexity, however due to (2) we shall see that it behaves essentially like a convex function for sparse vectors $\mathbf{x}$. Additionally, because of the non-convex perturbation $-\delta_c\|\mathbf{x}\|^2$ we need somewhat sharper estimates than just growth of the directional derivatives of g.

We first consider the estimate for the derivatives of h. Note that $\nabla h(\mathbf{x}) = 2\delta_c\mathbf{x} + 2(A^TA - I)\mathbf{x} - 2A^Tb$, and therefore

$$\langle \nabla h(\mathbf{x}+\mathbf{v}) - \nabla h(\mathbf{x}), \mathbf{v}\rangle = 2\delta_c\|\mathbf{v}\|^2 + 2\left(\|A\mathbf{v}\|^2 - \|\mathbf{v}\|^2\right).$$

Applying (2) now shows that

$$\langle \nabla h(\mathbf{x}+\mathbf{v}) - \nabla h(\mathbf{x}), \mathbf{v}\rangle \geq 2\delta_c\|\mathbf{v}\|^2 - 2\delta_c\|\mathbf{v}\|^2 = 0, \quad (18)$$

when $\operatorname{card}(\mathbf{v}) \leq c$.

Next we need a similar bound on the sub-gradients of g. In order to guarantee uniqueness of a sparse stationary point we need to show that they grow faster than $2\delta_c\|\mathbf{v}\|^2$.

Lemma 3.2. *Assume that $2\mathbf{z} \in \partial g(\mathbf{x})$. If the elements z_i fulfill $|z_i| \notin [(1-\delta_c)\sqrt{\mu}, \frac{\sqrt{\mu}}{1-\delta_c}]$ for every i, then for any $\mathbf{z}'$ with $2\mathbf{z}' \in \partial g(\mathbf{x}+\mathbf{v})$ we have*

$$\langle \mathbf{z}' - \mathbf{z}, \mathbf{v}\rangle > \delta_c\|\mathbf{v}\|^2, \quad (19)$$

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

The proof, which is somewhat technical, is given in the supplementary material. We are now ready to consider the distribution of stationary points. Set $\mathbf{z} = (I - A^TA)\mathbf{x}_s + A^T\mathbf{b}$ and recall that $2\mathbf{z} \in \partial g(\mathbf{x}_s)$ for stationary points $\mathbf{x}_s$ (Lemma 3.1).

Theorem 3.3. *Assume that $\mathbf{x}_s$ is a stationary point of f and that each element z_i fulfills $|z_i| \notin [(1-\delta_c)\sqrt{\mu}, \frac{\sqrt{\mu}}{1-\delta_c}]$. If $\mathbf{x}_s'$ is another stationary point of f then $\operatorname{card}(\mathbf{x}_s' - \mathbf{x}_s) > c$.*

Proof. Assume that $\operatorname{card}(\mathbf{x}_s' - \mathbf{x}_s) \leq c$. We first note that

$$2\delta_c\mathbf{x} - \nabla h(\mathbf{x}) = 2(I - A^TA)\mathbf{x} + 2A^T\mathbf{b}. \quad (20)$$

Since $\mathbf{x}_s$ and $\mathbf{x}_s'$ are both stationary points we have $2\delta_c\mathbf{x}_s - \nabla h(\mathbf{x}_s) = 2\mathbf{z}$ and $2\delta_c\mathbf{x}_s' - \nabla h(\mathbf{x}_s') = 2\mathbf{z}'$, where $2\mathbf{z} \in \partial g(\mathbf{x}_s)$ and $2\mathbf{z}' \in \partial g(\mathbf{x}_s')$. Taking the difference between the two equations gives

$$2\delta(\mathbf{x}_s' - \mathbf{x}_s) - (\nabla h(\mathbf{x}_s') - \nabla h(\mathbf{x}_s)) = 2(\mathbf{z}' - \mathbf{z}), \quad (21)$$

which implies

$$2\delta_c\|\mathbf{v}\|^2 - \langle \nabla h(\mathbf{x}+\mathbf{v}) - \nabla h(\mathbf{x}), \mathbf{v}\rangle = 2\langle \mathbf{z}' - \mathbf{z}, \mathbf{v}\rangle, \quad (22)$$

where $\mathbf{v} = \mathbf{x}_s' - \mathbf{x}_s$. However, according to (18) the left hand side is less than $2\delta\|\mathbf{v}\|^2$ if $\operatorname{card}(\mathbf{v}) \leq c$ which contradicts Lemma 3.2. □

In the supplementary material we give a simple example that shows that the constraint $|z_i| \notin [(1-\delta_c)\sqrt{\mu}, \frac{\sqrt{\mu}}{1-\delta_c}]$ is necessary and cannot be made tighter.

4. Low Rank Regularization

Next we generalize the vector formulation from the previous section to a matrix setting. We let F be as in (9) and assume that $\mathcal{A}$ is a linear operator $\mathbb{R}^{m\times n} \rightarrow \mathbb{R}^k$ fulfilling (7) for all X with $\operatorname{rank}(X) \leq r$. As in the vector case we can (ignoring constants) equivalently write

$$F(X) = G(X) - \delta_r\|X\|_F^2 + H(X), \quad (23)$$

where $H(X) = \delta_r\|X\|_F^2 + \left(\|\mathcal{A}X\|^2 - \|X\|_F^2\right) - 2\langle X, \mathcal{A}^*\mathbf{b}\rangle$ and $G(X) = g(\boldsymbol{\sigma}(X))$, with g as in (10). The first estimate

$$\langle \nabla H(X+V) - \nabla H(V), V\rangle \geq 0 \quad (24)$$

follows directly from (7) if $\operatorname{rank}(V) \leq r$. Our goal is now to show a matrix version of Lemma 3.2.

Lemma 4.1. *Let $\mathbf{x}$,$\mathbf{x}'$,$\mathbf{z}$,$\mathbf{z}'$ be fixed vectors with non-increasing and non-negative elements such that $\mathbf{x} \neq \mathbf{x}'$, $2\mathbf{z} \in \partial g(\mathbf{x})$ and $2\mathbf{z}' \in \partial g(\mathbf{x}')$. Define $X' = U'D_{\mathbf{x}'}V'^T$, $X = UD_{\mathbf{x}}V^T$, $Z' = U'D_{\mathbf{z}'}V'^T$, and $Z = UD_{\mathbf{z}}V^T$ as functions of unknown orthogonal matrices U, V, U' and V'. If*

$$a^* = \min_{U,V,U',V'} \frac{\langle Z' - Z, X' - X\rangle}{\|X' - X\|_F^2} \leq 1 \quad (25)$$

then

$$a^* = \min_{\pi} \frac{\langle M_{\pi,\mathbb{1}}\mathbf{z}' - \mathbf{z}, M_{\pi,\mathbb{1}}\mathbf{x}' - \mathbf{x}\rangle}{\|M_{\pi,\mathbb{1}}\mathbf{x}' - \mathbf{x}\|^2}, \quad (26)$$

where $\mathbb{1}$ is a vector of all ones.

Proof. We may assume that $U = I_{m\times m}$ and $V = I_{n\times n}$. We first note that (U', V') is a minimizer of (25) if and only if

$$\langle Z' - Z, X' - X\rangle \leq a^*\|X' - X\|_F^2. \quad (27)$$

This constraint can equivalently be written

$$C - \langle Z' - a^*X', X\rangle - \langle Z - a^*X, X'\rangle \leq 0, \quad (28)$$

where $C = \langle Z', X'\rangle + \langle Z, X\rangle - a^*(\|X'\|_F^2 + \|X\|_F^2)$ is independent of U' and V'. Thus any minimizer of (25) must also maximize

$$\langle U'D_{\mathbf{z}'-a^*\mathbf{x}'}V'^T, D_{\mathbf{x}}\rangle + \langle D_{\mathbf{z}-a^*\mathbf{x}}, U'D_{\mathbf{x}'}V'^T\rangle. \quad (29)$$

For ease of notation we now assume that $m \leq n$, that is, the number of rows are less than the columns (the opposite case can be handled by transposing). Equation (29) can now be written

$$\mathbf{x}^TM(\mathbf{z}' - a^*\mathbf{x}') + (\mathbf{z} - a^*\mathbf{x})^TM\mathbf{x}', \quad (30)$$

where $M = U' \odot V'_{1,1}$, $V'_{1,1}$ is the upper left $m \times m$ block of V' and $\odot$ denotes element wise multiplication. Since both U' and V' are orthogonal it is easily shown (using the Cauchy-Schwartz inequality) that M is DSS.

Note that objective (30) is linear in M and therefore optimization over the set of DSS matrices is guaranteed to have an extreme point $M_{\pi,\mathbf{v}}$ that is optimal. Furthermore, since $a^* \leq 1$ the vectors $\mathbf{x}$,$\mathbf{x}'$,$\mathbf{z} - a^*\mathbf{x}$ and $\mathbf{z}' - a^*\mathbf{x}'$ all have positive entries, and therefore the maximizing matrix has to be $M_{\pi,\mathbb{1}}$ for some permutation π. Since $M_{\pi,\mathbb{1}}$ is orthogonal and $M_{\pi,\mathbb{1}} = M_{\pi,\mathbb{1}} \odot M_{\pi,\mathbb{1}}$, $U' = M_{\pi,\mathbb{1}}$ and $V'_{1,1} = M_{\pi,\mathbb{1}}$ will be optimal when maximizing (30) over U' and $V'_{1,1}$. An optimal V' in (29) can now be chosen to be $V' = \begin{bmatrix} M_{\pi,\mathbb{1}} & 0 \\ 0 & I \end{bmatrix}$. Note that this choice is somewhat arbitrary since only the upper left block of V' affects the value of (29). The matrices $U'Z'V'^T$ and $U'X'V'^T$ are now diagonal, with diagonal elements $M_{\pi,\mathbb{1}}\mathbf{z}'$ and $M_{\pi,\mathbb{1}}\mathbf{x}'$, which concludes the proof. □

Corollary 4.2. *Assume that $2Z \in \partial G(X)$. If the singular values of the matrix Z fulfill $z_i \notin [(1-\delta_r)\sqrt{\mu}, \frac{\sqrt{\mu}}{1-\delta_r}]$, then for any $2Z' \in \partial G(X')$ we have*

$$\langle Z' - Z, X' - X \rangle > \delta_r \|X' - X\|_F^2, \tag{31}$$

as long as $\|X' - X\|_F \neq 0$.

Proof. We will first prove the result under the assumption that $\boldsymbol{\sigma}(X) \neq \boldsymbol{\sigma}(X')$ and then generalize to the general case using a continuity argument. For this purpose we need to extend the infeasible interval somewhat. Since $\delta_r < 1$ and the complement of $[(1-\delta_r)\sqrt{\mu}, \frac{\sqrt{\mu}}{1-\delta_r}]$ is open there is an $\epsilon > 0$ such that $z_i \notin [(1-\delta_r-\epsilon)\sqrt{\mu}, \frac{\sqrt{\mu}}{1-\delta_r-\epsilon}]$ and $\delta_r + \epsilon < 1$. Now assume that $a^* > 1$ in (25), then clearly

$$\langle Z' - Z, X' - X \rangle > (\delta_r + \epsilon)\|X' - X\|_F^2, \tag{32}$$

since $\delta_r + \epsilon < 1$. Otherwise $a^* \leq 1$ and we have

$$\frac{\langle Z' - Z, X' - X \rangle}{\|X' - X\|_F^2} \geq \frac{\langle M_{\pi,\mathbb{1}}\mathbf{z}' - \mathbf{z}, M_{\pi,\mathbb{1}}\boldsymbol{\sigma}(X') - \boldsymbol{\sigma}(X)\rangle}{\|M_{\pi,\mathbb{1}}\boldsymbol{\sigma}(X') - \boldsymbol{\sigma}(X)\|^2}. \tag{33}$$

According to Lemma 3.2 the right hand side is strictly larger than $\delta_r + \epsilon$, which proves that (32) holds for all X' with $\boldsymbol{\sigma}(X') \neq \boldsymbol{\sigma}(X)$.

For the case $\boldsymbol{\sigma}(X') = \boldsymbol{\sigma}(X)$ and $\|X' - X\|_F \neq 0$ it can now be proven that

$$\langle Z' - Z, X' - X \rangle \geq (\delta_r + \epsilon)\|X' - X\|_F^2, \tag{34}$$

using continuity of the scalar product and the Frobenius norm. (The technical details of this argument are given in the supplementary material.) Since $\epsilon > 0$ this proves the result. □

Theorem 4.3. *Assume that X_s is a stationary point of F, that is, $(I - \mathcal{A}^*\mathcal{A})X_s + \mathcal{A}^*\mathbf{b} = Z$, where $2Z \in \partial G(X)$ and the singular values of Z fulfill $\sigma_i(Z) \notin [(1-\delta_r)\sqrt{\mu}, \frac{\sqrt{\mu}}{1-\delta_r}]$. If X'_s is another stationary point then rank$(X'_s - X_s) > r$.*

The proof is similar to that of Theorem 3.3 and therefore we omit it.

5. Experiments

In this section we evaluate the proposed formulation on a few synthetic experiments. We compare the two formulations

$$\mu'\|X\|_* + \|\mathcal{A}X - \mathbf{b}\|^2 \tag{35}$$
$$r_\mu(\boldsymbol{\sigma}(X)) + \|\mathcal{A}X - \mathbf{b}\|^2 \tag{36}$$

for low rank recovery for varying regularization strengths μ and μ'. Note that the proximal operator of the nuclear norm, $\arg\min_X \mu'\|X\|_* + \|X - W\|^2$, performs soft thresholding at $\frac{\mu'}{2}$ while that of r_μ, $\arg\min_X r_\mu(\boldsymbol{\sigma}(X)) + \|X - W\|^2$, thresholds at $\sqrt{\mu}$ [19]. In order for the methods to roughly suppress an equal amount of noise we therefore use $\mu' = 2\sqrt{\mu}$ in (35). For completeness we also include a similar sparse recovery experiment in the supplementary material.

5.1. Optimization Method

Because of its simplicity we use the GIST approach from [17]. Given a current iterate X_k this method uses a trust region formulation that approximates the data term $\|\mathcal{A}X - \mathbf{b}\|^2$ with the linear function $2\langle \mathcal{A}^*\mathcal{A}X_k - \mathcal{A}^*\mathbf{b}, X\rangle$. In each step the algorithm therefore finds X_{k+1} by solving

$$\min_X r_\mu(\boldsymbol{\sigma}(X)) + 2\langle \mathcal{A}^*\mathcal{A}X_k - \mathcal{A}^*\mathbf{b}, X\rangle + \tau_k\|X - X_k\|^2. \tag{37}$$

Here the third term $\tau_k\|X - X_k\|^2$ restricts the search to a neighborhood around X_k. Completing squares shows that the above problem is equivalent to

$$\min_X r_\mu(\boldsymbol{\sigma}(X)) + \tau_k \|X - M\|^2, \tag{38}$$

where $M = X_k - \frac{1}{\tau_k}(\mathcal{A}^*\mathcal{A}X_k - \mathcal{A}^*\mathbf{b})$. Note that if $\tau_k = 1$ then any fixed point of (38) is a stationary point by Lemma 3.1. The optimization of (38) will be separable in the singular values of X. For each i we minimize $-\max(\sqrt{\mu} - \sigma_i(X), 0)^2 + \tau_k(\sigma_i(X) - \sigma_i(M))^2$. Since singular values are always positive there are three possible minimizers: $\sigma_i(X) = \sigma_i(M)$, $\sigma_i(X) = 0$ and $\sigma_i(X) = \frac{\tau_k\sigma_i(M) - \sqrt{\mu}}{\tau_k - 1}$. In our implementation we simply test which one of these yields the lowest objective value. (If $\tau_k = 1$ it is enough to test $\sigma_i(X) = \sigma_i(M)$ and $\sigma_i(X) = 0$.) For initialization we use $X_0 = 0$.

In summary our algorithm consists of repeatedly solving (38) for a sequence of $\{\tau_k\}$. In the experiments we noted that using $\tau_k = 1$ for all k sometimes resulted in divergence of the method due to large step sizes. We therefore start from a larger value ($\tau_0 = 5$ in our implementation) and reduce towards 1 as long as this results in decreasing objective values. Specifically we set $\tau_{k+1} = \frac{\tau_k - 1}{1.1} + 1$ if the previous step was successful in reducing the objective value. Otherwise we increase τ according to $\tau_{k+1} = 1.5(\tau_k - 1) + 1$.

5.2. Low Rank Recovery

In this section we test the proposed method on synthetic data. We generate 20×20 ground truth matrices of rank 5 by randomly selecting 20×5 matrices U and V with $\mathcal{N}(0, 1)$ elements and multiplying $X = UV^T$. By column stacking matrices the linear mapping $\mathcal{A} : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}^p$ can be represented by a $p \times mn$ matrix $\hat{A}$. For a given rank $r < \min(m, n)$ it is a difficult problem to determine the exact δ_r for which (7) holds. However if we consider unrestricted solutions ($r = \min(m, n)$) (7) reduces to a singular value bound. It is easy to see that if $p \geq mn$ and $\sqrt{1 - \delta_r} \leq \sigma_i(\hat{A}) \leq \sqrt{1 - \delta_r}$ for all i then (7) clearly holds for all X. For under-determined systems finding the value of δ_r is much more difficult. However for a number of random matrix families it can be proven that (7) will hold with high probability when the matrix size tends to infinity. For example [9, 28] mentions random matrices with Gaussian entries, Fourier ensembles, random projections and matrices with Bernoulli distributed elements.

For Figure 3 we randomly generated problem instances for low rank recovery. Each instance uses a matrix $\hat{A}$ of size $20^2 \times 20^2$ with $\delta = 0.2$ which was generated by first randomly sampling the elements of a matrix $\tilde{A}$ a Gaussian $\mathcal{N}(0, 1)$ distribution. The matrix $\hat{A}$ was then constructed from $\tilde{A}$ by modifying the singular values to be evenly distributed between $\sqrt{1 - \delta}$ and $\sqrt{1 + \delta}$. To generate a ground truth solution and a $\mathbf{b}$ vector we then computed $\mathbf{b} = \mathcal{A}X + \epsilon$, where all elements of ϵ are $\mathcal{N}(0, \sigma^2)$. We then solved (35) and (36) for varying noise level σ and regularization strength μ and computed the distance between the obtained solution and the ground truth.

The averaged results (over 50 random instances for each (σ, μ) setting) are shown in Figure 3. (Here black means low and white means a high errors. Note that the colormaps of left and middle image are the same. The red curves show the area where the computed solution has the correct rank.) From Figure 3 it is quite clear that the nuclear norm suffers from a shrinking bias. It consistently gives the best agreement with the ground truth data for values of μ that are not big enough to generate low rank. The effect becomes more visible as the noise level increases since a larger μ is required to suppress the noise. In contrast, (9) gives the best fit at the correct rank for all noise levels. This fit was consistently better than that of (35) for all noise levels. To the right in Figure 3 we show the fraction of problem instances that could be verified to be optimal (by computing $Z = (I - \mathcal{A}^*\mathcal{A})X_s + \mathcal{A}^*\mathbf{b}$ and checking that $\sigma_i(Z) \notin [(1 - \delta)\sqrt{\mu}, \frac{\sqrt{\mu}}{1-\delta}]$). It is not unexpected that verification works best when the noise level is moderate and a solution with the correct rank has been recovered. In such cases the recovered Z is likely to be close to low rank. Note for example that in the noise free case, that is, $\mathbf{b} = \mathcal{A}X_0$ for some low rank X_0 then $Z = (I - \mathcal{A}^*\mathcal{A})X_0 + \mathcal{A}^*\mathcal{A}X_0 = X_0$.

In Figure 4 we randomly generated under-determined problems with $\hat{A}$ of size 300×20^2 with Gaussian $\mathcal{N}(0, \frac{1}{300})$ elements and $\mathbf{b}$ vector as described previously. Even though we could not verify the optimality of the obtained solution (since δ is unknown) our approach consistently outperformed nuclear norm regularization which exhibits the same tendency to achieve a better fit for non-sparse solutions. In this setting (35) performed quite poorly, failing to simultaneously achieve a good fit and a correct rank (even for low noise levels).

5.3. Non-rigid Reconstruction

Given projections of a number 3D points on a deforming object, tracked through several images, the goal of non-rigid SfM is to reconstruct the 3D positions of the points. The problem is typically regularized by assuming that all possible object shapes are spanned by a low dimensional linear basis [4]. Specifically, let X_f be a $3 \times n$-matrix containing the coordinates of the 3D points when image f was taken. Here column i of X_f contains the x-,y- and z-coordinates of point i. Under the linearity assumption there is a set of basis shapes B_k, $k = 1, ..., K$ such that

$$X_f = \sum_{k=1}^{K} c_{fk} B_k. \tag{39}$$

Here the basis shapes B_k are of size $3 \times n$ and the coefficients c_{fk} are scalars. The projection of the 3D shape X_f into the image is modeled by $x_f = R_f X_f$. The 2×3 matrix R_f contains two rows from an orthogonal matrix which encodes camera orientation.

Dai *et al.* [12] observed that (39) can be interpreted as a low rank constraint by reshaping the matrices. First, let $X_f^\#$ be the $1 \times 3n$ matrix obtained by concatenation of the 3 rows X_f. Second, let $X^\#$ be $F \times 3n$ with rows $X_f^\#$, $f = 1, ..., F$. Then (39) can be written $X^\# = CB^\#$, where C is the $F \times K$ matrix containing the coefficients c_{fk} and $B^\#$ is a $K \times 3n$ matrix constructed from the basis in the same way as $X^\#$. The matrix $X^\#$ is thus of at most rank K. Furthermore, the complexity of the deformation can be constrained by penalizing the rank of $X^\#$.

To define an objective function we let the $2F \times n$ matrix M be the concatenation of all the projections x_f, $f = 1, ..., F$. Similarly we let the $3F \times n$ matrix X be the con-

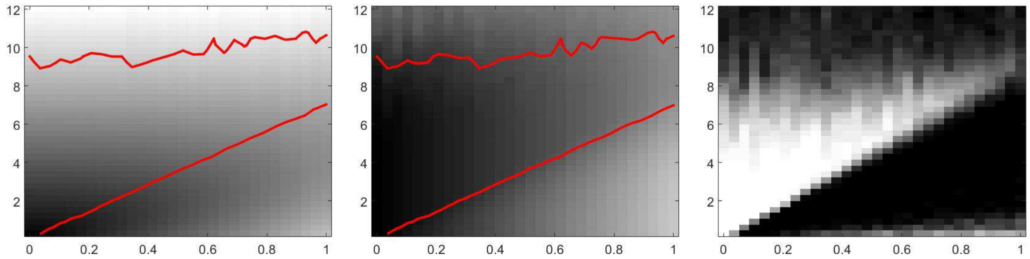

Figure 3: Low rank recovery results for varying noise level (x-axis) and regularization strength (y-axis) with random 400×400 $\mathcal{A}$ with $\delta = 0.2$. *Left*: Average distances between (35) and the ground truth for μ between 0 and 12. (red curves marks the area where the obtained solution has $\operatorname{rank}(X) = 5$). *Middle*: Average distances between (36) the ground truth. *Right*: Number of instances where (36) could be verified to be optimal for $\delta = 0.2$ (white = all, black = none).

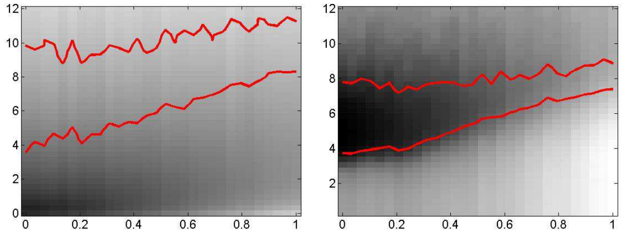

Figure 4: Low rank recovery results varying noise level (x-axis) and regularization strength (y-axis) with random 300×400 $\mathcal{A}$ (and unknown δ). *Left*: Average distances between (35) and the ground truth. (red curves marks the area where the obtained solution has $\operatorname{rank}(X) = 5$). *Middle*: Average distances between (36) the ground truth.

catenation of the X_f matrices. The objective function proposed by [12] is then

$$\mu \operatorname{rank}(X^{\#}) + \|RX - M\|_F^2, \tag{40}$$

where R is a $2F \times 3F$ block-diagonal matrix containing the R_f, $f = 1, ..., F$ matrices. Dai *et al.* proposed to solve (40) by replacing the rank penalty with $\|X^{\#}\|_*$. In this section we compare this to our approach that instead uses $r_\mu(\boldsymbol{\sigma}(X^{\#}))$. We test the approach on the 4 MOCAP sequences *Drink*, *Pick-up*, *Stretch* and *Yoga* used in [12], see Figure 5. Note that the MOCAP data is generated from motions recorded using real motion-capture-systems and the ground truth is therefore not of low rank. In Figure 6 we compare the two relaxations

$$r_\mu(\boldsymbol{\sigma}(X^{\#})) + \|RX - M\|_F^2 \tag{41}$$

and

$$2\sqrt{\mu}\|X^{\#}\|_* + \|RX - M\|_F^2, \tag{42}$$

for varying values of μ. We plot the obtained data fit versus the obtained rank for $\mu = 1, ..., 50$. The stair case shape of the blue curve is due to the nuclear norm's bias to small solutions. When μ is modified the strength of this bias changes and modifies the value of the data fit even if the modification is not big enough to change the rank. In contrast the data fit seems to take a (roughly) unique value for each rank when using (41).

The relaxation (41) consistently generates better data fit for all ranks and as an approximation of (40) it clearly performs better than (42). This is however not the whole truth. In Figure 7 we also plotted the distance to the ground truth solution. When the obtained solutions are not of very low rank (42) is generally better than (41) despite consistently generating a worse data fit. A feasible explanation is that when the rank is larger than roughly 3-4 there are multiple solutions with the same rank giving the same projections (witch also implies that the RIP (7) does not hold). Note that in Figure 6 the data fit seems to take a unique value for every rank. In short; when the space of feasible deformations becomes too large we cannot uniquely reconstruct the object from image data without additional priors. In contrast the ground truth distance can take several values for a given rank in Figure 7. The nuclear norm's bias to small solutions seems to have a regularizing effect.

Dai *et al.* [12] also suggested to further regularize the problem by penalizing derivatives of the 3D trajectories. For this they use a term $\|DX^{\#}\|_F^2$, where the matrix $D : \mathbb{R}^F \to \mathbb{R}^{F-1}$ is a first order difference operator. For completeness we add this term and compare

$$r_\mu(\boldsymbol{\sigma}(X^{\#})) + \|RX - M\|_F^2 + \|DX^{\#}\|_F^2 \tag{43}$$

and

$$2\sqrt{\mu}\|X^{\#}\|_* + \|RX - M\|_F^2 + \|DX^{\#}\|_F^2. \tag{44}$$

Figures 8 and 9 show the results. Our relaxation (43) generally finds better data fit at lower rank than what (44) does. Additionally, for low ranks (43) provides solutions that are closer to ground truth. When the rank increases most of the regularization becomes more dependent on the derivative prior leading to both methods providing similar results.

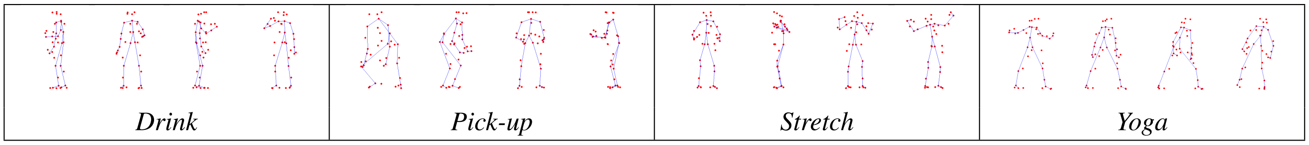

Figure 5: Four images from each of the MOCAP data sets.

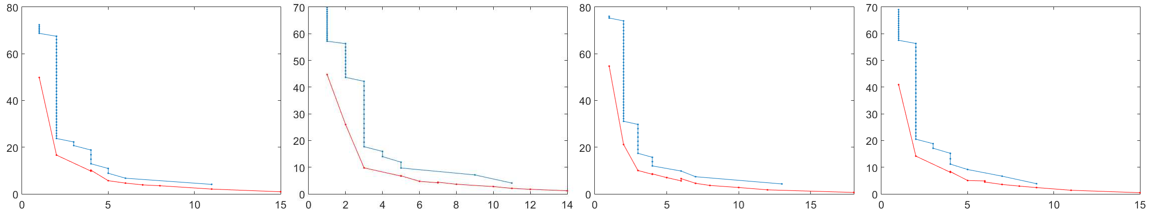

Figure 6: Results obtained with (41) and (42) for the four sequences. Data fit $\|RX - M\|_F$ (y-axis) versus $\operatorname{rank}(X^\#)$ (x-axis) is plotted for various regularization strengths. Blue curve uses $2\sqrt{\mu}\|X^\#\|_*$ and red curve $r_\mu(\boldsymbol{\sigma}(X^\#))$ with $\mu = 1, ..., 50$.

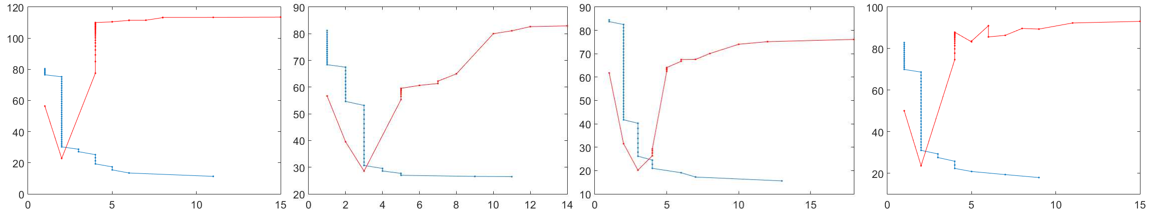

Figure 7: Results obtained with (41) and (42) for the four sequences. Distance to ground truth $\|X - X_{gt}\|_F$ (y-axis) versus $\operatorname{rank}(X^\#)$ (x-axis) is plotted for various regularization strengths. Blue curve uses $2\sqrt{\mu}\|X^\#\|_*$ and red curve $r_\mu(\boldsymbol{\sigma}(X^\#))$ with $\mu = 1, ..., 50$.

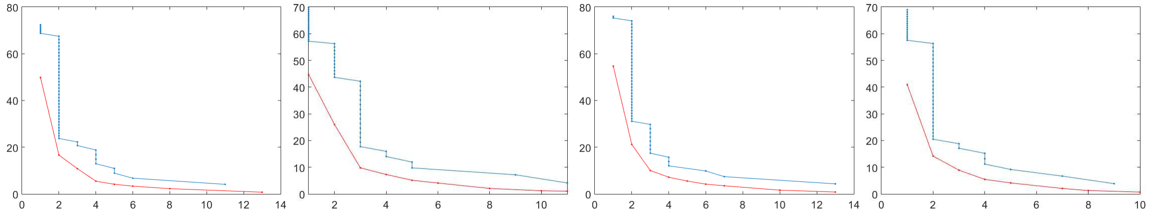

Figure 8: Results obtained with (43) and (44) for the four sequences. Data fit $\|RX - M\|_F$ versus $\operatorname{rank}(X^\#)$ is plotted for various regularization strengths. Blue curve uses $2\sqrt{\mu}\|X^\#\|_*$ and red curve $r_\mu(\boldsymbol{\sigma}(X^\#))$ with $\mu = 1, ..., 50$.

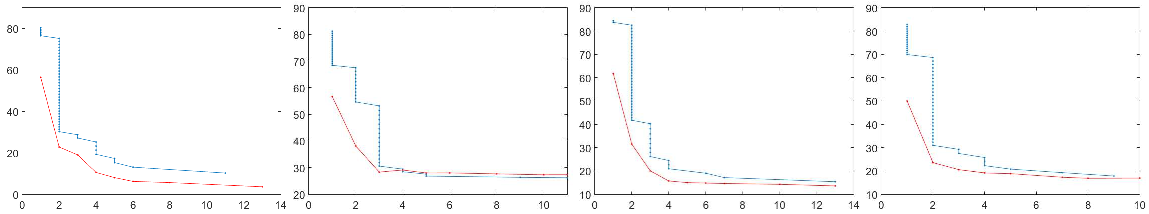

Figure 9: Results obtained with (43) and (44)for the four sequences. Distance to ground truth $\|X - X_{gt}\|_F$ (y-axis) versus $\operatorname{rank}(X^\#)$ (x-axis) is plotted for various regularization strengths. Blue curve uses $2\sqrt{\mu}\|X^\#\|_*$ and red curve $r_\mu(\boldsymbol{\sigma}(X^\#))$ with $\mu = 1, ..., 50$.

References

[1] F. Andersson and M. Carlsson. Fixed-point algorithms for frequency estimation and structured low rank approximation. *arXiv preprint arXiv:1601.01242*, 2016. 2

[2] F. Andersson, M. Carlsson, and C.-M. Perfekt. Operator-lipschitz estimates for the singular value functional calculus. *Proc. Amer. Math. Soc.*, 144:1867–1875, 2016. 3

[3] S. Bhojanapalli, B. Neyshabur, and N. Srebro. Global optimality of local search for low rank matrix recovery. In *Annual Conference in Neural Information Processing Systems (NIPS)*. 2016. 2

[4] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3d shape from image streams. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2000. 6

[5] P. Breheny and J. Huang. Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and Computing*, 25(2):173–187, Mar 2015. 1

[6] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *J. ACM*, 58(3):11:1–11:37, 2011. 2

[7] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009. 2

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

[9] 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, 6

[10] E. J. Candes, M. B. Wakin, and S. P. Boyd. Enhancing sparsity by reweighted 1 minimization. *Journal of Fourier analysis and applications*, 14(5-6):877–905, 2008. 1

[11] M. Carlsson. On convexification/optimization of functionals including an l2-misfit term. *arXiv preprint arXiv:1609.09378*, 2016. 2

[12] Y. Dai, H. Li, and M. He. A simple prior-free method for non-rigid structure-from-motion factorization. *International Journal of Computer Vision*, 107(2):101–122, 2014. 6, 7

[13] I. Daubechies, R. Devore, M. Fornasier, and C. Gntrk. Iteratively reweighted least squares minimization for sparse recovery. *Communications on Pure and Applied Mathematics*, 63(1):1–38, 1 2010. 1

[14] D. L. Donoho and M. Elad. Optimally sparse representation in general (non-orthogonal) dictionaries via minimization. In *PROC. NATL ACAD. SCI. USA 100 2197202*, 2002. 1

[15] M. Fazel, H. Hindi, and S. P. Boyd. A rank minimization heuristic with application to minimum order system approximation. In *American Control Conference*, 2001. 2

[16] R. Ge, C. Jin, and Y. Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. *arXiv preprint*, arxiv:1704.00708, 2017. 2

[17] P. Gong, C. Zhang, Z. Lu, J. Huang, and J. Ye. A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. In *International Conference on Machine Learning (ICML)*, pages 37–45, 2013. 5

[18] V. Jojic, S. Saria, and D. Koller. Convex envelopes of complexity controlling penalties: the case against premature envelopment. In *International Conference on Artificial Intelligence and Statistics*, 2011. 1

[19] V. Larsson and C. Olsson. Convex low rank approximation. *International Journal of Computer Vision*, 120(2):194–214, 2016. 1, 2, 5

[20] A. S. Lewis. The convex analysis of unitarily invariant matrix functions, 1995. 3

[21] P. Loh and M. J. Wainwright. Support recovery without incoherence: A case for nonconvex regularization. *arXiv preprint*, arXiv:1412.5632, 2014. 2

[22] C. Lu, J. Tang, S. Y. Yan, and Z. Lin. Generalized nonconvex nonsmooth low-rank minimization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 2

[23] C. Lu, C. Zhu, C. Xu, S. Yan, and Z. Lin. Generalized singular value thresholding. In *AAAI*, 2015. 2

[24] 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

[25] 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

[26] 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

[27] D. Park, A. Kyrillidis, C. Caramanis, and S. Sanghavi. Non-square matrix sensing without spurious local minima via the burer-monteiro approach. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS*, pages 65–74, 2017. 2

[28] 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, 6

[29] J. Shi, X. Ren, G. Dai, J. Wang, and Z. Zhang. A nonconvex relaxation approach to sparse dictionary learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011. 1

[30] E. Soubies, L. Blanc-Féraud, and G. Aubert. A continuous exact l0 penalty (cel0) for least squares regularized problem. *SIAM Journal on Imaging Sciences*, 8(3):1607–1639, 2015. 1, 2

[31] 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

[32] J. A. Tropp. Convex recovery of a structured signal from independent random linear measurements. In *Sampling Theory, a Renaissance*, pages 67–101. 2015. 1

[33] S. Wang, D. Liu, and Z. Zhang. Nonconvex relaxation approaches to robust matrix recovery. In *International Joint Conference on Artificial Intelligence*, 2013. 2

[34] C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 04 2010. 1