Exact Penalization for Cardinality and Rank Constrained Optimization Problems via Partial Regularization

Zhaosong Lu* Xiaorui Li† Shuhuang Xiang‡

September 26, 2021 (Revised: June 5, 2022)

Abstract

In this paper we consider a class of constrained optimization problems whose constraints involve a cardinality or rank constraint. The penalty formulation based on a partial regularization has recently been promoted in the literature to approximate these problems, which usually outperforms the penalty formulation based on a full regularization in terms of solution quality. Nevertheless, the relation between the penalty formulation with a partial regularizer and the original problem was not much studied yet. Under some suitable assumptions, we show that the penalty formulation based on a partial regularization is an exact reformulation of the original problem in the sense that they both share the same global minimizers. We also show that a local minimizer of the original problem is that of the penalty reformulation. These results provide some theoretical justification for the often-observed superior performance of the penalty model based on a partial regularizer over a corresponding full regularizer.

Keywords: sparse optimization, low-rank optimization, cardinality constraint, rank constraint, partial regularization, exact penalty.

AMS subject classifications: 65C60, 65K05, 90C26, 90C30

1 Introduction

Nowadays, there are numerous applications in which sparse or low-rank solutions are concerned. For example, in compressed sensing, a large sparse signal is decoded by using a relatively small number of linear measurements, which can be formulated as finding a sparse

*Department of Industrial and Systems Engineering, University of Minnesota, USA (email: zhaosong@umn.edu).

†Huawei Technologies Canada, Burnaby, Canada. (email: lixiaorui1986@gmail.com).

‡School of Mathematics and Statistics, Central South University, Changsha, Hunan, China 410083. (email: xiangsh@csu.edu.cn).

solution to a system of linear equalities and/or inequalities (e.g., see [32, 24, 38]). The similar ideas have also been widely used in sparse inverse covariance selection, sparse logistic regression, sparse multivariate regression, and image deblurring (e.g., see [33, 43, 48, 12]). Mathematically speaking, they can be formulated into the following cardinality constrained problem

$$\min_x \{f(x) : \|x\|_0 \le s, \ x \in S\} \tag{1}$$

for some integer $s \ge 0$ controlling the sparsity of the solutions, where S is a closed set in $\Re^n$, $f : \Re^n \to \Re \cup \{\infty\}$, and $\|x\|_0$ denotes the cardinality of the vector x. In addition, finding a low-rank solution to a system or an optimization problem has attracted a great deal of attention in science and engineering. Generally, it can be formulated into the following rank constrained optimization problem

$$\min_X \{g(X) : \mathrm{rank}(X) \le r, \ X \in \mathcal{X}\} \tag{2}$$

for some integer $r \ge 0$ controlling the rank of the solutions, where $\mathcal{X}$ is a closed set in $\Re^{m\times n}$, $g : \Re^{m\times n} \to \Re \cup \{\infty\}$, and $\mathrm{rank}(X)$ denotes the rank of the matrix X. Numerous application problems can be modeled by (2), including the low-rank matrix completion problem, the wireless sensor network localization problem, and the nearest low-rank correlation matrix problem (e.g., see [15, 4, 36, 20, 28]).

Given that $\|\cdot\|_0$ is an integer-valued, discontinuous and nonconvex function, it is generally hard to solve problem (1). One common approach in the literature (e.g., see [18, 37, 9, 19, 16, 9, 8, 47, 40, 22, 46]) is to approximate (1) by the problem

$$\min_{x\in S} f(x) + \lambda \sum_{i=1}^n \phi(|x_i|), \tag{3}$$

where $\lambda > 0$ is a parameter controlling the sparsity of the solution, and ϕ is a non-decreasing regularizer defined on $[0, \infty)$ satisfying that $\phi(0) = 0$ and $\phi(t) > 0$ for every $t > 0$. Some popular ϕ's are listed as follows:

(i) (ℓ_1 [37, 10, 7]): $\phi(t) = t \ \ \forall t \ge 0$;

(ii) (ℓ_p [18, 19]): $\phi(t) = t^p \ \ \forall t \ge 0$;

(iii) (Log [39, 8]): $\phi(t) = \log(t + \varepsilon) - \log(\varepsilon) \ \ \forall t \ge 0$;

(iv) (Capped-ℓ_1 [47]): $\phi(t) = \begin{cases} t & \text{if } 0 \le t < \nu, \\ \nu & \text{if } t \ge \nu; \end{cases}$

(v) (MCP [46]): $\phi(t) = \begin{cases} \lambda t - \frac{t^2}{2\alpha} & \text{if } 0 \le t < \lambda\alpha, \\ \frac{\lambda^2\alpha}{2} & \text{if } \ t \ge \lambda\alpha; \end{cases}$

(vi) (SCAD [16]): $\phi(t) = \begin{cases} \lambda t & \text{if } 0 \leq t \leq \lambda, \\ \frac{-t^2+2\beta\lambda t-\lambda^2}{2(\beta-1)} & \text{if } \lambda < t < \lambda\beta, \\ \frac{(\beta+1)\lambda^2}{2} & \text{if } \ t \geq \lambda\beta, \end{cases}$

where $0 < p < 1$, $\varepsilon > 0$, $\nu > 0$, $\lambda > 0$, $\alpha > 1$ and $\beta > 1$ are some parameters.

It is known that $\sum_{i=1}^n \phi(|x_i|)$ is generally a biased approximation to $\|x\|_0$. To neutralize the bias in $\sum_{i=1}^n \phi(|x_i|)$ incurred by some leading entries (in magnitude) of x, the partial regularizer $\sum_{i=s+1}^n \phi(|x|_{[i]})$ has recently been advocated in the literature (e.g., see [27, 23, 1, 34, 30, 2]), where $|x|_{[i]}$ is the ith largest element in $\{|x_1|, |x_2|, \ldots, |x_n|\}$. Lu and Li [30] showed that the partial ℓ_p-regularizer for $0 < p \leq 1$ is more capable than the full ℓ_p-regularizer in recovering the sparse solution of a linear system. In addition, the extensive numerical study in [27, 23, 30, 2] demonstrate that the partial regularizer $\sum_{i=s+1}^n \phi(|x|_{[i]})$ is usually more effective than the full regularizer $\sum_{i=1}^n \phi(|x_i|)$ in finding a sparse approximate solution. Inspired by these, in this paper we consider the following less biased approximation to (1):

$$\min_{x \in S} f(x) + \lambda \sum_{i=s+1}^{n} \phi(|x|_{[i]}). \tag{4}$$

Under some suitable assumptions, we show that problem (4) is an exact penalty reformulation of (1) for various ϕ including the popular ones listed above, that is, problems (1) and (4) share the same global minimizers. Moreover, we show that a local minimizer of (1) is also that of (4). We believe that such properties usually do not hold for problem (3) since $\sum_{i=1}^n \phi(|x_i|)$ is generally a biased approximation to $\|x\|_0$. These results provide some theoretical justification for the often-observed superior performance of a partial regularizer over a corresponding full regularizer in finding a sparse approximate solution.

Similar to (1), problem (2) is also generally difficult to solve due to the discontinuity and non-convexity of the rank function. One popular approach in the literature is to approximate problem (2) by the problem

$$\min_{X \in \mathcal{X}} g(X) + \lambda \sum_{i=1}^{q} \phi(\sigma_i(X)), \tag{5}$$

where $q = \min(m, n)$, $\sigma_i(X)$ denotes the ith largest singular value of X, and ϕ is some regularizer mentioned above (e.g., see [17, 44, 6, 28, 31, 41]). It is not hard to see that $\sum_{i=1}^q \phi(\sigma_i(X))$ is generally a biased approximation to $\text{rank}(X)$. To neutralize the bias in $\sum_{i=1}^q \phi(\sigma_i(X))$ incurred by some leading singular values of X, the partial regularizer $\sum_{i=r+1}^q \phi(\sigma_i(X))$ has recently been used in the literature (e.g., see [20, 21, 29]) to approximate the rank function. Numerical results in [21, 29] show that such a partial regularizer is usually more effective than the full regularizer $\sum_{i=1}^q \phi(\sigma_i(X))$ in finding a low-rank approximate solution. Inspired by these, in this paper we consider the following less biased approximation to (2):

$$\min_{X \in \mathcal{X}} g(X) + \lambda \sum_{i=r+1}^{q} \phi(\sigma_i(X)). \tag{6}$$

Under some suitable assumptions, we show that problem (6) is an exact penalty reformulation of (2) for various ϕ including the popular ones mentioned above, namely, problems (2) and (6) share the same global minimizers. Moreover, we show that a local minimizer of (2) is also that of (6). We believe that such properties usually do not hold for problem (5) since $\sum_{i=1}^q \phi(\sigma_i(X))$ is generally a biased approximation to $\mathrm{rank}(X)$. These results provide some theoretical justification on the often-observed superior performance of a partial regularizer over a corresponding full regularizer in finding a low-rank approximate solution.

The rest of this paper is organized as follows. In Section 2 we establish some exact penalty results for a general optimization problem. In Section 3 we show that under some suitable assumptions, problem (4) is an exact penalty reformulation of problem (1). In Section 4 we show that under some suitable assumptions, problem (6) is an exact penalization for problem (2). Finally we make some concluding remarks in Section 5.

1.1 Notations

In this paper the set of all nonnegative real numbers is denoted by $\Re_+$. The nonnegative orthant of $\Re^n$ is denoted by $\Re^n_+$. For any $x \in \Re^n$, $\|x\|_0$ and $\|x\|$ denote the cardinality (i.e., the number of nonzero entries) and the Euclidean norm of x, respectively. For any $p > 0$, let $\|x\|_p := (\sum_{i=1}^n |x_i|^p)^{1/p}$. In addition, $x_{[i]}$ denotes the ith largest entry of x for $i = 1, \ldots, n$, and $|x|$ stands for the n-dimensional vector whose ith entry is $|x_i|$ for all i. For any $J \subseteq \{1, \ldots, n\}$, x_J denotes the subvector of x indexed by J, J^{c} stands for the complement of J in $\{1, \ldots, n\}$, and $|J|$ denotes the cardinality of J. Given any $X \in \Re^{m \times n}$, $\mathrm{rank}(X)$ and $\|X\|$ stand for the rank and the spectral norm of X, respectively; $\sigma_i(X)$ denotes the ith largest singular value of X. The nuclear norm of X is denoted by $\|X\|_*$, that is, $\|X\|_* = \sum_i \sigma_i(X)$. In addition, $\mathscr{D}(X)$ is the $m \times n$ matrix whose ith diagonal entry is $\sigma_i(X)$ for all i and off-diagonal entries are 0, that is, $[\mathscr{D}(X)]_{ii} = \sigma_i(X)$ for $1 \le i \le q$ and $[\mathscr{D}(X)]_{ij} = 0$ for all $i \ne j$, where $q = \min(m, n)$. Let $\mathcal{U}$ be a normed vector space equipped with a norm $\|\cdot\|$. Given any $u \in \mathcal{U}$, $\mathcal{B}(u; \epsilon)$ stands for a closed ball in $\mathcal{U}$ centered at u with radius ϵ, that is, $\mathcal{B}(u; \epsilon) = \{v \in \mathcal{U} : \|v - u\| \le \epsilon\}$.

2 Some general penalization results

Exact penalization is an important technique for converting some sophisticated constrained optimization problems into simpler ones (e.g., see [13, 5, 42]). In this section we develop some results regarding exact penalization for a general optimization problem in the form of

$$\min_{u \in \Omega_1 \cap \Omega_2} h(u) \tag{7}$$

for some function $h : \mathcal{U} \to \Re \cup \{\infty\}$, where Ω_1 and Ω_2 are two nonempty closed sets in a normed vector space $\mathcal{U}$. These results are a generalization of the ones presented in [11], and will be used in the subsequent sections to study exact penalization for problems (1) and (2). To this end, assume throughout this section that problem (7) has at least one optimal solution and that $\Phi : \mathcal{U} \times \mathcal{U} \to \Re_+$ is a function satisfying the following assumption.

Assumption 1. *(a)* $\Omega_1 \cap \operatorname*{Argmin}_{v\in\Omega_2} \Phi(u,v) \neq \emptyset$ *for all* $u \in \Omega_1$.

(b) $\min_{v\in\Omega_2} \Phi(u,v) = 0$ *if and only if* $u \in \Omega_2$.

We next show that under some suitable assumptions, the following problem

$$\min_{u\in\Omega_1} \left\{ h(u) + \lambda \min_{v\in\Omega_2} \Phi(u,v) \right\} \tag{8}$$

for some $\lambda > 0$ is an exact penalty reformulation of problem (7). For convenience, we denote the objective function of (8) by $H_\lambda(u)$, that is, $H_\lambda(u) = h(u) + \lambda \min_{v\in\Omega_2} \Phi(u,v)$.

Theorem 2.1. *Suppose that the function* Φ *satisfies Assumption 1, and problem* (7) *has at least one optimal solution. Additionally, assume that there exists some constant* $L > 0$ *such that*

$$|h(u) - h(\tilde{u})| \leq L\Phi(u,\tilde{u}) \quad \forall u, \tilde{u} \in \Omega_1. \tag{9}$$

Then the following statements hold.

(i) Any global minimizer of problem (7) *is a global minimizer of problem* (8) *whenever* $\lambda \geq L$.

(ii) Any global minimizer of problem (8) *is a global minimizer of problem* (7) *whenever* $\lambda > L$.

Proof. (i) Suppose that $\lambda \geq L$. Let u^* be a global minimizer of problem (7) and $u \in \Omega_1$ be arbitrarily chosen. It then follows from Assumption 1(a) that $\Omega_1 \cap \operatorname{Argmin}_{v\in\Omega_2} \Phi(u,v) \neq \emptyset$. Let $\tilde{u} \in \Omega_1 \cap \operatorname{Argmin}_{v\in\Omega_2} \Phi(u,v)$ be arbitrarily chosen. By this and the definition of H_λ, one has

$$H_\lambda(u) = h(u) + \lambda \min_{v\in\Omega_2} \Phi(u,v) = h(u) + \lambda\Phi(u,\tilde{u}). \tag{10}$$

In addition, since $\tilde{u} \in \Omega_1$ and u^* is a global minimizer of (7), we have $h(\tilde{u}) \geq h(u^*)$. By $u^* \in \Omega_2$ and Assumption 1(b), one has $\min_{v\in\Omega_2} \Phi(u^*,v) = 0$ and hence $H_\lambda(u^*) = h(u^*)$. It then follows that $h(\tilde{u}) \geq H_\lambda(u^*)$. Using this and (9), we obtain that

$$H_\lambda(u^*) \leq h(\tilde{u}) \leq h(u) + L\Phi(u,\tilde{u}),$$

which together with (10), $\lambda \geq L$ and the nonnegativity of Φ yields $H_\lambda(u) \geq H_\lambda(u^*)$. By this, $u^* \in \Omega_1$ and the arbitrariness of $u \in \Omega_1$, one can conclude that u^* is a global minimizer of problem (8).

(ii) Suppose that $\lambda > L$. Let u^* be a global minimizer of problem (8). Then $u^* \in \Omega_1$. In view of this and Assumption 1(a), we know that $\Omega_1 \cap \operatorname{Argmin}_{v\in\Omega_2} \Phi(u^*,v) \neq \emptyset$. Let $\tilde{u}^* \in \Omega_1 \cap \operatorname{Argmin}_{v\in\Omega_2} \Phi(u^*,v)$. Clearly, $H_\lambda(u^*) \leq H_\lambda(\tilde{u}^*)$ and $\tilde{u}^* \in \Omega_1 \cap \Omega_2$. The latter relation together with Assumption 1(b) yields $\min_{v\in\Omega_2} \Phi(\tilde{u}^*,v) = 0$. It follows from this and the definition of H_λ that $H_\lambda(\tilde{u}^*) = h(\tilde{u}^*)$. In addition, one can observe that $H_\lambda(u^*) = h(u^*) + \lambda\Phi(u^*, \tilde{u}^*)$. By these, $u^*, \tilde{u}^* \in \Omega_1$ and (9), one has

$$H_\lambda(\tilde{u}^*) = h(\tilde{u}^*) \le h(u^*) + L\Phi(u^*, \tilde{u}^*) = H_\lambda(u^*) + (L - \lambda)\Phi(u^*, \tilde{u}^*),$$

which, together with $\lambda > L$, $H_\lambda(u^*) \le H_\lambda(\tilde{u}^*)$ and the nonnegativity of Φ, implies that

$$0 \le \Phi(u^*, \tilde{u}^*) \le (H_\lambda(u^*) - H_\lambda(\tilde{u}^*))/(\lambda - L) \le 0.$$

Hence, $\Phi(u^*, \tilde{u}^*) = 0$, which leads to $\min_{v\in\Omega_2} \Phi(u^*, v) = 0$. By this and Assumption 1(b), one can conclude $u^* \in \Omega_2$ and hence $u^* \in \Omega_1 \cap \Omega_2$. Notice from the definition of H_λ and Assumption 1(b) that $H_\lambda(u) = h(u)$ for all $u \in \Omega_1 \cap \Omega_2$. By these and the fact that u^* is a global minimizer of (8), we have that

$$h(u) = H_\lambda(u) \ge H_\lambda(u^*) = h(u^*), \quad \forall u \in \Omega_1 \cap \Omega_2.$$

It follows that u^* is a global minimizer of problem (7). ∎

In what follows, we show that under some suitable assumptions, a local minimizer of problem (7) is also that of problem (8). Before proceeding, we make some further assumptions on the function Φ below.

Assumption 2. *(a) For any $\epsilon > 0$, there exists some $\delta > 0$ such that $\|u - v\| \le \epsilon$ for any $u, v \in \Omega_1$ satisfying $\Phi(u, v) \le \delta$.*

(b) For any $\epsilon > 0$, there exists some $\delta > 0$ such that $\Phi(u, v) \le \epsilon$ for any $u, v \in \Omega_1$ satisfying $\|u - v\| \le \delta$.

Theorem 2.2. *Suppose that there is a function $\Phi : \mathcal{U} \times \mathcal{U} \to \Re_+$ satisfying Assumptions 1 and 2. Let u^* be a local minimizer of problem (7). Assume that there exist some $\epsilon > 0$ and $L > 0$ such that*

$$|h(u) - h(\hat{u})| \le L\Phi(u, \hat{u}) \quad \forall u, \hat{u} \in \mathcal{B}(u^*; \epsilon) \cap \Omega_1. \tag{11}$$

Then u^* is a local minimizer of problem (8) whenever $\lambda \ge L$.

Proof. Since u^* is a local minimizer of problem (7), there exists some $\tilde{\epsilon} > 0$ such that

$$h(u) \ge h(u^*) \quad \forall u \in \mathcal{B}(u^*; \tilde{\epsilon}) \cap \Omega_1 \cap \Omega_2. \tag{12}$$

Clearly, $u^* \in \Omega_1 \cap \Omega_2$. By Assumption 2(a), there exists some $\delta > 0$ such that $\|u - v\| \le \min(\epsilon, \tilde{\epsilon})/2$ for any $u, v \in \Omega_1$ satisfying $\Phi(u, v) \le \delta$. Let $u \in \Omega_1$ be such that $\Phi(u, u^*) \le \delta$. It then follows that $\|u - u^*\| \le \min(\epsilon, \tilde{\epsilon})/2$. In addition, by Assumption 1(a), we know that $\Omega_1 \cap \operatorname{Argmin}_{v\in\Omega_2} \Phi(u, v) \ne \emptyset$. Let $\tilde{u} \in \Omega_1 \cap \operatorname{Argmin}_{v\in\Omega_2} \Phi(u, v)$ be arbitrarily chosen. By this, $u^* \in \Omega_2$ and $\Phi(u, u^*) \le \delta$, one has $\Phi(u, \tilde{u}) \le \Phi(u, u^*) \le \delta$, which together with $u, \tilde{u} \in \Omega_1$ implies that $\|u - \tilde{u}\| \le \min(\epsilon, \tilde{\epsilon})/2$. Using this and $\|u - u^*\| \le \min(\epsilon, \tilde{\epsilon})/2$, we have

$$\|\tilde{u} - u^*\| \le \|u - u^*\| + \|u - \tilde{u}\| \le \min(\epsilon, \tilde{\epsilon}).$$

By this, (11), (12), $\|u - u^*\| \leq \min(\epsilon, \tilde{\epsilon})/2$, and $\tilde{u} \in \Omega_1 \cap \Omega_2$, one can see that

$$h(u^*) \leq h(\tilde{u}) \leq h(u) + L\Phi(u, \tilde{u}).$$

Using this, $\lambda \geq L$, $\tilde{u} \in \mathrm{Argmin}_{v \in \Omega_2} \Phi(u, v)$, and the nonnegativity of Φ, we obtain that

$$h(u^*) \leq h(u) + \lambda\Phi(u, \tilde{u}) = H_\lambda(u). \tag{13}$$

Recall that $u^* \in \Omega_2$. It then follows from Assumption 1(b) and the definition of H_λ that $H_\lambda(u^*) = h(u^*)$. Combining this with (13), we conclude that $H_\lambda(u) \geq H_\lambda(u^*)$ for any $u \in \Omega_1$ satisfying $\Phi(u, u^*) \leq \delta$. In addition, by Assumption 2(b), there exists some $\hat{\epsilon} > 0$ such that $\Phi(u, u^*) \leq \delta$ for any $u \in \mathcal{B}(u^*; \hat{\epsilon}) \cap \Omega_1$. It follows that $H_\lambda(u) \geq H_\lambda(u^*)$ for any $u \in \mathcal{B}(u^*; \hat{\epsilon}) \cap \Omega_1$. This along with $u^* \in \Omega_1$ implies that u^* is a local minimizer of problem (8). ■

3 Exact penalization for problem (1)

In this section we study the relation between the cardinality constrained problem (1) and the penalty model (4). Before proceeding, we make a mild assumption on the function ϕ associated with problem (4), which will be used frequently in this section. It holds for various regularizers such as ℓ_1, ℓ_p, Log, Capped-ℓ_1, MCP and SCAD that are presented in Section 1.

Assumption 3. *The function $\phi : \Re_+ \to \Re_+$ is non-decreasing. Moreover, $\phi(0) = 0$ and $\phi(t) > 0$ for every $t > 0$.*

The following result shows that under some suitable assumptions, problem (4) is an exact penalty reformulation of problem (1), that is, they share the same global minimizers.

Theorem 3.1. *Suppose that Assumption 3 holds for ϕ, and moreover,*

$$S \cap \operatorname*{Argmin}_{\|y\|_0 \leq s} \sum_{i=1}^{n} \phi(|x - y|_i) \neq \emptyset \qquad \forall x \in S. \tag{14}$$

Assume that there exists some constant $L > 0$ such that

$$|f(x) - f(y)| \leq L \sum_{i=1}^{n} \phi(|x - y|_i) \qquad \forall x, y \in S. \tag{15}$$

Then the following statements hold.

(i) If x^* is a global minimizer of problem (1), then x^* is a global minimizer of problem (4) whenever $\lambda \geq L$.

(ii) If x^* is a global minimizer of problem (4), then x^* is a global minimizer of problem (1) whenever $\lambda > L$.

Proof. We know from Assumption 3 that ϕ is non-decreasing on $\Re_+$. It thus follows that

$$\begin{aligned}\min_{\|y\|_0\le s}\sum_{i=1}^n \phi(|x-y|_i) &= \min_{\substack{J\subseteq\{1,\ldots,n\}\\ |J|\le s}}\left\{\min_{y_{J^c}=0}\sum_{i=1}^n \phi(|x-y|_i)\right\},\\ &= \min_{\substack{J\subseteq\{1,\ldots,n\}\\ |J|\le s}}\sum_{i\in J^c}\phi(|x|_i) = \sum_{i=s+1}^n \phi(|x|_{[i]}).\end{aligned} \tag{16}$$

Also, we see from Assumption 3 that $\phi(t)=0$ if and only if $t=0$. By this and (16), one can observe that

$$\min_{\|y\|_0\le s}\sum_{i=1}^n \phi(|x-y|_i)=0 \Longleftrightarrow \sum_{i=s+1}^n \phi(|x|_{[i]})=0 \Longleftrightarrow \|x\|_0\le s. \tag{17}$$

Let $\Omega_1=S$, $\Omega_2=\{x\in\Re^n:\|x\|_0\le s\}$, $\Phi(x,y)=\sum_{i=1}^n\phi(|x-y|_i)$ for all $x,y\in\Re^n$. In view of (14) and (17), one can see that Assumption 1 holds for such Φ, Ω_1 and Ω_2. The conclusion of this theorem then follows from (15), (16), and Theorem 2.1 with $h=f$. ■

Roughly speaking, the condition (14) requires that S satisfy a certain sparsity property, while the condition (15) requires that f be Lipschitz continuous relative to S. In general, it may be difficult to verify (14) directly. We next provide a sufficient yet simpler condition for (14) to hold.

Proposition 3.2. *Suppose that the set S satisfies*

$$S\cap \operatorname*{Argmin}_{\|y\|_0\le s}\|x-y\|\neq\emptyset \qquad \forall x\in S. \tag{18}$$

Then the condition (14) *holds.*

Proof. It is not hard to observe from (16) that

$$\operatorname*{Argmin}_{\|y\|_0\le s}\|x-y\| \;\subseteq\; \operatorname*{Argmin}_{\|y\|_0\le s}\sum_{i=1}^n\phi(|x-y|_i), \tag{19}$$

which together with (18) implies that (14) holds. ■

Remark 1 If ϕ is strictly increasing on $\Re_+$ such as ℓ_1, the two sets in (19) are equal and thus the conditons (14) and (18) are equivalent. Otherwise, (18) is stronger than (14), for example, when ϕ is Capped-ℓ_1, MCP or SCAD. In addition, the condition (18) holds for numerous sets S. For example, one can verify that it holds for the following sets.

(i) $S=\prod_{i=1}^n[a_i,b_i]$, where $-\infty\le a_i\le 0\le b_i\le\infty$ for $i=1,\ldots,n$.[1]

[1]When $b_i=\infty$, $[a_i,b_i]$ shall be changed to $[a_i,b_i)$. Similar change shall be made when $a_i=-\infty$.

(ii) $S = \{x \in \Re^n : \sum_i d_i |x_i|^p \le \gamma\}$ or $\{x \in \Re^n_+ : \sum_i d_i x_i^p \le \gamma\}$ for $p > 0$, $d_i > 0$ and $\gamma > 0$.

(iii) $S = \{x \in \Re^n : \psi(|x_1|, \dots, |x_n|) \le t\}$, where $t \in \Re$ and $\psi : \Re^n_+ \to \Re \cup \{\infty\}$ is non-decreasing, i.e., $\psi(x) \ge \psi(y)$ for all $x, y \in \Re^n_+$ satisfying $x \ge y$ componentwisely. [2]

The following result is an immediate consequence of Theorem 3.1 and Proposition 3.2.

Corollary 3.3. *Suppose that* (15), (18), *and Assumption 3 hold. Then the conclusion of Theorem 3.1 holds.*

As another consequence of Theorem 3.1 and Proposition 3.2 with $\phi(t) = t$ for all $t \in \Re_+$, we obtain a result regarding the exact penalization of problem (1) based on the partial ℓ_1 regularization.

Corollary 3.4. *Suppose that the set S satisfies* (18). *Assume that f is Lipschitz continuous on S, that is, there exists some $L_f > 0$ such that*

$$|f(x) - f(y)| \le L_f \|x - y\|_1 \qquad \forall x, y \in S. \tag{20}$$

Then the following statements hold.

(i) If x^* is a global minimizer of problem (1), *then x^* is a global minimizer of problem*

$$\min_{x \in S} f(x) + \lambda \sum_{i=s+1}^{n} |x|_{[i]}. \tag{21}$$

whenever $\lambda \ge L_f$.

(ii) If x^* is a global minimizer of problem (21), *then x^* is a global minimizer of problem* (1) *whenever $\lambda > L_f$.*

The following result is a consequence of Corollary 3.3. It holds for the partial regularization induced by various ϕ such as ℓ_1, ℓ_p, Log, Capped-ℓ_1, MCP and SCAD.

Corollary 3.5. *Suppose that Assumption 3 holds, f satisfies* (20), *and S is a non-singleton compact set satisfying* (18). *Assume additionally that $\liminf_{t \to 0^+} \phi(t)/t > 0$.*[3] *Then the conclusion of Theorem 3.1 holds with $L = L_f / M_\phi$, where*

$$M_\phi = \inf_{t \in [0,R]} \frac{\phi(t)}{t}, \qquad R = \max_{x,y \in S} \|x - y\|_\infty.$$

[2] The items (i) and (ii) can be viewed as a special case of the item (iii) with $t = 0$ and ψ being the indicator function of their set S.

[3] It is not hard to verify that the regularizers ℓ_1, ℓ_p, Log, Capped-ℓ_1, MCP and SCAD satisfy these assumptions.

Proof. We first show that (15) holds with $L = L_f/M_\phi$. Indeed, since S is a non-singleton compact set, it follows that $R \in (0, \infty)$. Using Assumption 3, we know that ϕ is non-decreasing and $\phi(t) > 0$ for all $t > 0$. By these and $\liminf_{t \to 0^+} \phi(t)/t > 0$, one can easily verify $M_\phi \in (0, \infty)$. In addition, notice from the definition of R that $|x - y|_i \le R$ for every i and $x, y \in S$. By this and the definition of M_ϕ, one has that $\phi(|x - y|_i) \ge M_\phi |x - y|_i$ for every i and $x, y \in S$. It follows from this and (20) that

$$|f(x) - f(y)| \le L_f \|x - y\|_1 \le \frac{L_f}{M_\phi} \sum_{i=1}^n \phi(|x - y|_i), \qquad \forall x, y \in S,$$

and hence (15) holds with $L = L_f/M_\phi$ as desired. The conclusion of this corollary then follows from Corollary 3.3. ■

We next study some relation between the local minimizers of problem (1) and those of problem (4). Before proceeding, we make a stronger assumption on ϕ than Assumption 3, which holds as well for various regularizers such as ℓ_1, ℓ_p, Log, Capped-ℓ_1, MCP and SCAD that are presented in Section 1.

Assumption 4. *The function ϕ is non-decreasing on $[0, \infty)$ and strictly increasing in a right neighborhood of 0. Moreover, $\phi(0) = 0$ and ϕ is right continuous at 0.*

The following result shows that under some suitable assumptions, a local minimizer of problem (1) is also that of problem (4).

Theorem 3.6. *Suppose that (14) and Assumption 4 hold for ϕ. Let x^* be a local minimizer of problem (1). Assume that there exist some $\epsilon > 0$ and $L > 0$ such that*

$$|f(x) - f(y)| \le L \sum_{i=1}^n \phi(|x - y|_i) \qquad \forall x, y \in \mathcal{B}(x^*; \epsilon) \cap S. \tag{22}$$

Then x^* is also a local minimizer of problem (4) for any $\lambda \ge L$.

Proof. Let $\Omega_1 = S$, $\Omega_2 = \{x \in \Re^n : \|x\|_0 \le s\}$ and $\Phi(x, y) = \sum_{i=1}^n \phi(|x - y|_i)$ for all $x, y \in \Re^n$. By (14), Assumption 4, and the same argument as in the proof of Theorem 3.1, one can see that Assumption 1 holds for such Φ, Ω_1 and Ω_2. Also, by Assumption 4, we know that $\phi(0) = 0$ and ϕ is right continuous at 0. It immediately follows that Assumption 2(b) holds for such Φ. In addition, since ϕ is strictly increasing in a right neighborhood of 0, there exists some $\bar{\epsilon} > 0$ such that ϕ is strictly increasing in $[0, \bar{\epsilon}]$. Let $\tilde{\epsilon} > 0$ be arbitrarily chosen, $\hat{\epsilon} = \min(\tilde{\epsilon}/\sqrt{n}, \bar{\epsilon})/2$ and $\delta = \sup_{t \in [0, \hat{\epsilon}]} \phi(t)$. Using these and the monotonicity of ϕ, we can see that $t \in [0, \tilde{\epsilon}/\sqrt{n}]$ if $\phi(t) \le \delta$. By this, the expression of Φ, and the nonnegativity of ϕ, one can observe that $|x - y|_i \le \tilde{\epsilon}/\sqrt{n}$ if $\Phi(x, y) \le \delta$. It then follows that $\|x - y\| \le \tilde{\epsilon}$ if $\Phi(x, y) \le \delta$. Hence, Assumption 2(a) holds for the above Φ. The conclusion of this theorem then follows from (22) and Theorem 2.2 with $h = f$. ■

The following result is a consequence of Theorem 3.6 in which the condition (14) is replaced by a simpler one.

Corollary 3.7. *Suppose that* (18) *and Assumption 4 hold. Let* x^* *be a local minimizer of problem* (1)*. Assume that there exist some* $\epsilon > 0$ *and* $L > 0$ *such that* (22) *holds. Then* x^* *is also a local minimizer of problem* (4) *for any* $\lambda \geq L$*.*

Proof. We know from the proof of Corollary 3.3 that (18) implies (14). The conclusion then follows from Theorem 3.6. ■

As an immediate consequence of Corollary 3.7, we obtain the following result.

Corollary 3.8. *Suppose that* (18)*, Assumption 4, and* $\liminf_{t\to 0^+} \phi(t)/t > 0$ *hold.*[4] *Let* x^* *be a local minimizer of problem* (1)*. Assume that there exist some* $\epsilon > 0$ *and* $L_f > 0$ *such that*

$$|f(x) - f(y)| \leq L_f \|x - y\|_1 \qquad \forall x, y \in \mathcal{B}(x^*; \epsilon) \cap S. \tag{23}$$

Then x^* *is a local minimizer of problem* (4) *for any* $\lambda \geq L_f/M_\phi$*, where* $M_\phi = \inf_{t\in[0,2\epsilon]} \phi(t)/t$*.*

Proof. Notice that $|x - y|_i \leq 2\epsilon$ for every i and $x, y \in \mathcal{B}(x^*; \epsilon) \cap S$. By this, (23) and a similar argument as in the proof of Corollary 3.5, one can verify that

$$|f(x) - f(y)| \leq \frac{L_f}{M_\phi} \sum_{i=1}^{n} \phi(|x - y|_i), \qquad \forall x, y \in \mathcal{B}(x^*; \epsilon) \cap S,$$

and hence (22) holds with $L = L_f/M_\phi$. The conclusion then follows from Corollary 3.7. ■

Remark 2 In many applications, f is often Lipschitz continuous on the set S satisfying (18). It then follows from Corollaries 3.4 and 3.8 that the partial ℓ_1 regularized model (21) is an exact penalty reformulation of problem (1). Further, if S is compact, it follows from Corollaries 3.5 and 3.8 that the partially regularized model (4) for various ϕ such as ℓ_p, Log, Capped-ℓ_1, MCP and SCAD is an exact penalty reformulation of (1). Therefore, Corollaries 3.4, 3.5 and 3.8 provide some theoretical justification for the often-observed superior performance of a partial regularizer over a corresponding full regularizer in finding a sparse approximate solution.

4 Exact penalization for problem (2)

Exact penalization technique has recently been studied for rank constrained problems (e.g., see [3, 35]). In this section, we study the relation between the rank constrained problem (2) and the penalty model (6). In particular, under some suitable assumptions, we show that problem (6) is an exact penalty reformulation of problem (2), that is, they share the same global minimizers. We also show that a local minimizer of problem (2) is that of problem (6).

Theorem 4.1. *Suppose that Assumption 3 holds for* ϕ*, and moreover,*

$$\mathcal{X} \cap \operatorname*{Argmin}_{\operatorname{rank}(Y)\leq r} \sum_{i=1}^{q} \phi(\sigma_i(X - Y)) \neq \emptyset \qquad \forall X \in \mathcal{X}, \tag{24}$$

[4] The regularizers ℓ_1, ℓ_p, Log, Capped-ℓ_1, MCP and SCAD satisfy these assumptions.

where $q = \min(m,n)$. Assume that there exists some $L > 0$ such that

$$|g(X) - g(Y)| \leq L \sum_{i=1}^{q} \phi(\sigma_i(X-Y)) \qquad \forall X, Y \in \mathcal{X}. \tag{25}$$

Then the following statements hold.

(i) If X^* is a global minimizer of problem (2), then X^* is a global minimizer of problem (6) whenever $\lambda \geq L$.

(ii) If X^* is a global minimizer of problem (6), then X^* is a global minimizer of problem (2) whenever $\lambda > L$.

Proof. Let $\Omega_1 = \mathcal{X}$, $\Omega_2 = \{X \in \Re^{m\times n} : \mathrm{rank}(X) \leq r\}$ and $\Phi(X,Y) = \sum_{i=1}^q \phi(\sigma_i(X-Y))$ for all $X, Y \in \Re^{m\times n}$. In view of (24), one can see that Assumption 1(a) holds for such Φ, Ω_1 and Ω_2. We next show that Assumption 1(b) also holds for them. To this end, we first claim that

$$\min_{\mathrm{rank}(Y)\leq r} \sum_{i=1}^{q} \phi(\sigma_i(X-Y)) \leq \sum_{i=r+1}^{q} \phi(\sigma_i(X)) \qquad \forall X \in \Re^{m\times n}.[5] \tag{26}$$

Indeed, let $X \in \Re^{m\times n}$ be arbitrarily chosen and $X = \sum_{i=1}^q \sigma_i(X) u_i v_i^T$ a singular value decomposition (SVD) of X. Also, let $Y = \sum_{i=1}^r \sigma_i(X) u_i v_i^T$. One can immediately see that

$$\mathrm{rank}(Y) \leq r, \qquad \sum_{i=1}^{q} \phi(\sigma_i(X-Y)) = \sum_{i=r+1}^{q} \phi(\sigma_i(X)),$$

which imply that (26) holds as claimed. Now suppose that $X \in \Omega_2$ is arbitrarily chosen. Then $\mathrm{rank}(X) \leq r$ and hence $\sigma_i(X) = 0$ for $r+1 \leq i \leq q$, which together with Assumption 3 yields $\sum_{i=r+1}^q \phi(\sigma_i(X)) = 0$. It follows from this and (26) that $\min\limits_{\mathrm{rank}(Y)\leq r} \sum_{i=1}^q \phi(\sigma_i(X-Y)) = 0$ for every $X \in \Omega_2$. On the other hand, suppose that $\min\limits_{\mathrm{rank}(Y)\leq r} \sum_{i=1}^q \phi(\sigma_i(X-Y)) = 0$ for some X. Then $\sum_{i=1}^q \phi(\sigma_i(X-Y)) = 0$ for some Y with $\mathrm{rank}(Y) \leq r$. By this and Assumption 3, one has that $\sigma_i(X-Y) = 0$ for $1 \leq i \leq q$. It follows that $X = Y$ and hence $\mathrm{rank}(X) \leq r$, that is, $X \in \Omega_2$. Therefore, Assumption 1(b) also holds for the above Φ and Ω_2. We are now ready to prove statements (i) and (ii).

(i) Suppose that X^* is a global minimizer of problem (2) and $\lambda \geq L$. It follows from (25) and Theorem 2.1 with $h = g$ that X^* is a global minimizer of the following problem

$$\min_{X\in\mathcal{X}} \left\{ g(X) + \lambda \min_{\mathrm{rank}(Y)\leq r} \sum_{i=1}^{q} \phi(\sigma_i(X-Y)) \right\}. \tag{27}$$

[5] Notice that the relation $\min\limits_{\mathrm{rank}(Y)\leq r} \sum_{i=1}^q \phi(\sigma_i(X-Y)) = \sum_{i=r+1}^q \phi(\sigma_i(X))$ generally does not hold. Therefore, the conclusion of this theorem cannot follow directly from Theorem 2.1.

In addition, notice that $\mathrm{rank}(X^*) \leq r$, which along with (26) implies that

$$\min_{\mathrm{rank}(Y)\leq r} \sum_{i=1}^{q} \phi(\sigma_i(X^* - Y)) = \sum_{i=r+1}^{q} \phi(\sigma_i(X^*)) = 0. \tag{28}$$

By (26) and (28), one can observe that the objective of (27) is majorized by that of (6), and moreover, they achieve the same value at X^*. Hence, X^* is also a global minimizer of problem (6).

(ii) Suppose that X^* is a global minimizer of problem (6) and $\lambda > L$. Claim that problems (6) and (27) have the same optimal value. To this end, let $\hat{X}^*$ be a global minimizer of problem (27). It follows from (25) and Theorem 2.1 with $h = g$ that $\hat{X}^*$ is also a global minimizer of problem (2). By this and $\lambda > L$, it follows from statement (i) that $\hat{X}^*$ is also a global minimizer of problem (6) and hence $\mathrm{rank}(\hat{X}^*) \leq r$. Using the latter relation and the same argument as in the proof of statement (i), we see that (28) holds with X^* replaced by $\hat{X}^*$. It then implies that problems (6) and (27) have equal objective value at $\hat{X}^*$. By this and the fact that $\hat{X}^*$ is a global minimizer of them, we conclude that they share the same optimal value as claimed. Using this and the supposition that X^* is a global minimizer of (6), we see that the optimal value of problems (6) and (27) is $g(X^*) + \lambda \sum_{i=r+1}^{q} \phi(\sigma_i(X^*))$, and hence

$$g(X^*) + \lambda \min_{\mathrm{rank}(Y)\leq r} \sum_{i=1}^{q} \phi(\sigma_i(X^* - Y)) \geq g(X^*) + \lambda \sum_{i=r+1}^{q} \phi(\sigma_i(X^*)).$$

It together with (26) leads to

$$\min_{\mathrm{rank}(Y)\leq r} \sum_{i=1}^{q} \phi(\sigma_i(X^* - Y)) = \sum_{i=r+1}^{q} \phi(\sigma_i(X^*)).$$

Hence, the optimal value of (27) is $g(X^*) + \lambda \min\limits_{\mathrm{rank}(Y)\leq r} \sum_{i=1}^{q} \phi(\sigma_i(X^* - Y))$, which implies that X^* is also a global minimizer of (27). By this, $\lambda > L$, (25) and Theorem 2.1 with $h = g$, we conclude that X^* is also a global minimizer of problem (2). ∎

Generally it may not be easy to check the relation (24). We next provide some sufficient conditions under which (24) holds. Before proceeding, we establish a technical lemma as follows. It appears that only part of this lemma is known and proved in the literature. We therefore provide a proof for the rest part, which is quite nontrivial.

Lemma 4.2. *Let $X \in \Re^{m\times n}$ be given, and $q = \min(m, n)$. Then*

$$\operatorname*{Argmin}_{\mathrm{rank}(Y)\leq r} \|X - Y\|_F = \left\{ \sum_{i=1}^{r} \sigma_i(X) u_i v_i^T : \sum_{i=1}^{q} \sigma_i(X) u_i v_i^T \text{ is an SVD of } X. \right\}, \tag{29}$$

where $\|\cdot\|_F$ is the Frobenius norm.

Proof. For convenience, let $\mathcal{M}_r$ denote the set on the right-hand side of (29). It follows from Eckart and Young's Theorem [14] that

$$\mathcal{M}_r \subseteq \operatorname*{Argmin}_{\mathrm{rank}(Y)\le r} \|X-Y\|_F, \qquad \min_{\mathrm{rank}(Y)\le r} \|X-Y\|_F^2 = \sum_{i=r+1}^{q} \sigma_i^2(X). \tag{30}$$

We next show that $\operatorname*{Argmin}_{\mathrm{rank}(Y)\le r} \|X-Y\|_F \subseteq \mathcal{M}_r$. To this end, let $\bar{Y} \in \operatorname*{Argmin}_{\mathrm{rank}(Y)\le r} \|X-Y\|_F$ be arbitrarily chosen. Then $\mathrm{rank}(\bar{Y}) \le r$, which implies $\sigma_i(\bar{Y}) = 0$ for $i \ge r+1$. By Weyl's Theorem (e.g., see [26, Theorem 3.3.16]), one knows that $\sigma_{i+j-1}(A+B) \le \sigma_i(A) + \sigma_j(B)$ for any $A, B \in \Re^{m\times n}$ and $1 \le i,j \le q$ with $i+j \le q+1$. Letting $A = X - \bar{Y}$, $B = \bar{Y}$ and $j = r+1$, we obtain that

$$\sigma_{i+r}(X) \le \sigma_i(X-\bar{Y}), \quad i = 1,\ldots,q-r. \tag{31}$$

In addition, it follows from (30) that

$$\sum_{i=1}^{q} \sigma_i^2(X-\bar{Y}) = \|X-\bar{Y}\|_F^2 = \sum_{i=r+1}^{q} \sigma_i^2(X),$$

which together with (31) implies that

$$\sigma_i(X-\bar{Y}) = \begin{cases} \sigma_{i+r}(X) & \text{if } 1 \le i \le q-r; \\ 0 & \text{otherwise.} \end{cases} \tag{32}$$

Also, we know from [25, Theorem 7.4.51] that

$$\|A-B\|' \ge \|\mathscr{D}(A) - \mathscr{D}(B)\|' \qquad \forall A, B \in \Re^{m\times n} \tag{33}$$

holds for any unitarily invariant norm $\|\cdot\|'$ on $\Re^{m\times n}$. Using this, (30), and $\sigma_i(\bar{Y}) = 0$ for $i \ge r+1$, we have

$$\sum_{i=r+1}^{q} \sigma_i^2(X) = \|X-\bar{Y}\|_F^2 \ge \sum_{i=1}^{q} (\sigma_i(X) - \sigma_i(\bar{Y}))^2 = \sum_{i=1}^{r} (\sigma_i(X) - \sigma_i(\bar{Y}))^2 + \sum_{i=r+1}^{q} \sigma_i^2(X),$$

which implies $\sigma_i(\bar{Y}) = \sigma_i(X)$ for $1 \le i \le r$. Recall that $\sigma_i(\bar{Y}) = 0$ for $i \ge r+1$. Due to these facts and (32), any SVD of $\bar{Y}$ and $X - \bar{Y}$ shall be in the form of

$$\bar{Y} = \sum_{i=1}^{r} \sigma_i(X) \bar{u}_i \bar{v}_i^T, \qquad X - \bar{Y} = \sum_{i=r+1}^{q} \sigma_i(X) \bar{u}_i \bar{v}_i^T \tag{34}$$

for some unit vectors $\bar{u}_i$ and $\bar{v}_i$ satisfying $\bar{u}_i^T \bar{u}_j = 0$ and $\bar{v}_i^T \bar{v}_j = 0$ for all $1 \le i \ne j \le r$ and $r+1 \le i \ne j \le q$. Clearly, $X = \sum_{i=1}^{q} \sigma_i(X) \bar{u}_i \bar{v}_i^T$. Claim that $\sum_{i=1}^{q} \sigma_i(X) \bar{u}_i \bar{v}_i^T$ is an SVD of

X. To this end, we first show that $\bar{u}_i^T\bar{u}_j = 0$ and $\bar{v}_i^T\bar{v}_j = 0$ for all $1 \le i \le r$ and $r+1 \le j \le q$. Indeed, since $\bar{Y} \in \operatorname*{Argmin}_{\operatorname{rank}(Y)\le r} \|X-Y\|_F$ and $\bar{Y} = \sum_{i=1}^r \sigma_i(X)\bar{u}_i\bar{v}_i^T$, it follows that

$$(\bar{u}_1,\dots,\bar{u}_r,\bar{v}_1,\dots,\bar{v}_r) \in \operatorname*{Argmin}_{u,v} \left\| X - \sum_{i=1}^r \sigma_i(X) u_i v_i^T \right\|_F^2.$$

Its first-order optimality conditions yield $(X-\bar{Y})\bar{v}_i = 0$ and $(X-\bar{Y})^T\bar{u}_i = 0$ for each $1 \le i \le r$. It follows that $\bar{v}_i \perp \mathscr{R}((X-\bar{Y})^T)$ and $\bar{u}_i \perp \mathscr{R}(X-\bar{Y})$ for every $1 \le i \le r$, where $\mathscr{R}(\cdot)$ denotes the range space of the associated matrix. Notice from (34) that $\bar{v}_j \in \mathscr{R}((X-\bar{Y})^T)$ and $\bar{u}_j \in \mathscr{R}(X-\bar{Y})$ for every $r+1 \le j \le q$. We immediately conclude that $\bar{u}_i^T\bar{u}_j = 0$ and $\bar{v}_i^T\bar{v}_j = 0$ for every $1 \le i \le r$ and $r+1 \le j \le q$ as claimed. Hence, $\{\bar{u}_1,\dots,\bar{u}_q\}$ and $\{\bar{v}_1,\dots,\bar{v}_q\}$ are two sets of orthonormal vectors, which together with $X = \sum_{i=1}^q \sigma_i(X)\bar{u}_i\bar{v}_i^T$ implies that $\sum_{i=1}^q \sigma_i(X)\bar{u}_i\bar{v}_i^T$ is an SVD of X. In view of this, $\bar{Y} = \sum_{i=1}^r \sigma_i(X)\bar{u}_i\bar{v}_i^T$ and the definition of $\mathcal{M}_r$, one can see that $\bar{Y} \in \mathcal{M}_r$. By the arbitrariness of $\bar{Y}$, we have $\operatorname*{Argmin}_{\operatorname{rank}(Y)\le r} \|X-Y\|_F \subseteq \mathcal{M}_r$. It together with (30) implies that (29) holds. ∎

We are now ready to provide a sufficient condition under which the relation (24) holds.

Proposition 4.3. *Let $q = \min(m,n)$. Suppose that Assumption 3 holds and there exists a strictly increasing function $\psi : \Re_+ \to \Re \cup \{\infty\}$ such that $\|x\|^\diamond := \psi(\sum_{i=1}^q \phi(|x_i|))$ is a norm on $\Re^q$.*[6] *Assume that*

$$\mathcal{X} \cap \operatorname*{Argmin}_{\operatorname{rank}(Y)\le r} \|X-Y\|_F \neq \emptyset \qquad \forall X \in \mathcal{X}. \tag{35}$$

Then the relation (24) holds.

Proof. Suppose that $\psi : \Re_+ \to \Re \cup \{\infty\}$ is a strictly increasing function such that $\|x\|^\diamond = \psi(\sum_{i=1}^q \phi(|x_i|))$ is a norm on $\Re^q$. Clearly, $\|x\|^\diamond = \||x|\|^\diamond$ and $\|x\|^\diamond = \|Px\|^\diamond$ for every $x \in \Re^q$ and $q \times q$ permutation matrix P. It follows from [26, Definition 3.5.17] that $\|\cdot\|^\diamond$ is a symmetric gauge function. This and [26, Theorem 3.5.18] imply that $\|X\|^\diamond$ is a unitarily invariant norm on $\Re^{m\times n}$, where

$$\|X\|^\diamond = \|\sigma(X)\|^\diamond, \quad \sigma(X) = (\sigma_1(X),\cdots,\sigma_q(x))^T \qquad \forall X \in \Re^{m\times n}.$$

Using this and (33), we obtain that for all $X, Y \in \Re^{m\times n}$,

$$\psi\left(\sum_{i=1}^q \phi(\sigma_i(X-Y))\right) = \|X-Y\|^\diamond \ge \|\mathscr{D}(X)-\mathscr{D}(Y)\|^\diamond = \psi\left(\sum_{i=1}^q \phi(|\sigma_i(X)-\sigma_i(Y)|)\right).$$

This together with the strict monotonicity of ψ implies that

$$\sum_{i=1}^q \phi(\sigma_i(X-Y)) \ge \sum_{i=1}^q \phi(|\sigma_i(X)-\sigma_i(Y)|) \qquad \forall X, Y \in \Re^{m\times n}. \tag{36}$$

[6]It is not hard to see that this assumption holds for $\phi(t) = t^p$ and $\psi(t) = t^{1/p}$ for all $t \ge 0$ and $p \ge 1$.

By Assumption 3 and a similar argument as for (16), it is not hard to observe that

$$\min_{\mathrm{rank}(Y)\le r} \sum_{i=1}^{q} \phi(|\sigma_i(X)-\sigma_i(Y)|) = \sum_{i=r+1}^{q} \phi(\sigma_i(X)) \qquad \forall X \in \Re^{m\times n}.$$

It follows from this and (36) that

$$\min_{\mathrm{rank}(Y)\le r} \sum_{i=1}^{q} \phi(\sigma_i(X-Y)) \ge \sum_{i=r+1}^{q} \phi(\sigma_i(X)) \qquad \forall X \in \Re^{m\times n}. \tag{37}$$

Let $X \in \Re^{m\times n}$ be arbitrarily chosen, and let $X = \sum_{i=1}^{q} \sigma_i(X) u_i v_i^T$ be an arbitrary SVD of X. Also, let $\bar{Y} = \sum_{i=1}^{r} \sigma_i(X) u_i v_i^T$. One can immediately see that

$$\mathrm{rank}(\bar{Y}) \le r, \qquad \sum_{i=1}^{q} \phi(\sigma_i(X-\bar{Y})) = \sum_{i=r+1}^{q} \phi(\sigma_i(X)),$$

which together with (37) implies $\bar{Y} \in \operatorname*{Argmin}_{\mathrm{rank}(Y)\le r} \sum_{i=1}^{q} \phi(\sigma_i(X-Y))$. By this, Lemma 4.2, $\bar{Y} = \sum_{i=1}^{r} \sigma_i(X) u_i v_i^T$, and the fact that $\sum_{i=1}^{q} \sigma_i(X) u_i v_i^T$ is an arbitrary SVD of X, one can see that

$$\operatorname*{Argmin}_{\mathrm{rank}(Y)\le r} \sum_{i=1}^{q} \|X-Y\|_F \subseteq \operatorname*{Argmin}_{\mathrm{rank}(Y)\le r} \sum_{i=1}^{q} \phi(\sigma_i(X-Y)) \qquad \forall X \in \Re^{m\times n}.$$

It then follows from this and (35) that the relation (24) holds. ∎

We next provide another sufficient condition under which the relation (24) holds.

Proposition 4.4. *Suppose that $\phi : \Re_+ \to \Re_+$ is concave and $\phi(0) = 0$. Assume that the set $\mathcal{X}$ satisfies* (35)*. Then the relation* (24) *holds.*

Proof. Since $\phi : \Re_+ \to \Re_+$ is concave and $\phi(0) = 0$, it follows from [45, Theorem 1] that

$$\sum_{i=1}^{q} \phi(\sigma_i(X-Y)) \ge \sum_{i=1}^{q} |\phi(\sigma_i(X)) - \phi(\sigma_i(Y))| \qquad \forall X, Y \in \Re^{m\times n}.$$

It leads to

$$\min_{\mathrm{rank}(Y)\le r} \sum_{i=1}^{q} \phi(\sigma_i(X-Y)) \ge \min_{\mathrm{rank}(Y)\le r} \sum_{i=1}^{q} |\phi(\sigma_i(X)) - \phi(\sigma_i(Y))| = \sum_{i=r+1}^{q} \phi(\sigma_i(X)),$$

where the equality is due to the nonnegativity of ϕ and the fact that $\mathrm{rank}(Y) \le r$ if and only if $\sigma_i(Y) = 0$ for every $i > r$. The rest of the proof is the same as that of Proposition 4.3. ∎

Remark 3 The relation (35) holds for numerous sets $\mathcal{X}$. For example, one can verify that it holds for the following sets with $q = \min(m, n)$.

(i) $\mathcal{X} = \{X \in \Re^{m\times n} : \sigma_i(X) \le b_i, i = 1,\dots,q\}$ for $b_i \in [0,\infty], i = 1,\dots,q$.[7]

(ii) $\mathcal{X} = \{X \in \Re^{m\times n} : \sum_{i=1}^q d_i[\sigma_i(X)]^p \le t\}$ for $p > 0$, $t \ge 0$ and $d_i > 0, i = 1,\dots,q$.

(iii) $\mathcal{X} = \{X \in \Re^{m\times n} : \psi(\sigma_1(X),\dots,\sigma_q(X)) \le t\}$, where $t \in \Re$ and $\psi : \Re^q_+ \to \Re \cup \{\infty\}$ is non-decreasing, that is, $\psi(x) \ge \psi(y)$ for all $x, y \in \Re^q_+$ with $x \ge y$.

As a consequence of Theorem 4.1 and Proposition 4.3, we have the following result.

Corollary 4.5. *Suppose that* (25), (35) *and Assumption 3 hold. Assume additionally that there exists a strictly increasing function* $\psi : \Re_+ \to \Re \cup \{\infty\}$ *such that* $\psi(\sum_{i=1}^q \phi(|x_i|))$ *is a norm on* $\Re^q$*, where* $q = \min(m,n)$*. Then the conclusion of Theorem 4.1 holds.*

The following result is an immediate consequence of Theorem 4.1 and Proposition 4.4.

Corollary 4.6. *Suppose that* (25), (35) *and Assumption 3 hold. Assume additionally that* ϕ *is concave in* $[0,\infty)$*. Then the conclusion of Theorem 4.1 holds.*

The following result is a consequence of Corollaries 4.5 and 4.6.

Corollary 4.7. *Suppose that the set* $\mathcal{X}$ *satisfies* (35)*. Assume that there exist some* $L > 0$ *and* $p > 0$ *such that*

$$|g(X) - g(Y)| \le L \sum_{i=1}^q [\sigma_i(X-Y)]^p \qquad \forall X, Y \in \mathcal{X}, \tag{38}$$

where $q = \min(m,n)$*. Then the following statements hold.*

(i) If X^* *is a global minimizer of problem* (2)*, then* X^* *is a global minimizer of problem* (6)

$$\min_{X\in\mathcal{X}} g(X) + \lambda \sum_{i=r+1}^q [\sigma_i(X)]^p \tag{39}$$

whenever $\lambda \ge L$*.*

(ii) If X^* *is a global minimizer of problem* (39)*, then* X^* *is a global minimizer of problem* (2) *whenever* $\lambda > L$*.*

Proof. Let $\phi(t) = t^p$ and $\psi(t) = t^{1/p}$ for all $t \ge 0$. We divide the proof into two cases.

Case 1) $p \ge 1$. Clearly, $\psi(\sum_{i=1}^q \phi(|x_i|)) = \|x\|_p$ is a norm on $\Re^q$. By this and the assumptions of this corollary, we see that the assumptions of Corollary 4.5 hold. Therefore, the conclusion of this corollary follows from Corollary 4.5.

Case 2) $0 < p < 1$. Notice that in this case ϕ is concave in $[0,\infty)$. By this and the assumptions of this corollary, we see that the assumptions of Corollary 4.6 hold. The conclusion of this corollary thus follows from Corollary 4.6. ∎

The following result is a consequence of Corollary 4.6, which holds for various regularizers ϕ such as ℓ_1, ℓ_p, Log, Capped-ℓ_1, MCP and SCAD. It can be viewed as a generalization of Corollary 3.5 from vector to matrix.

[7] When $b_i = \infty$, $\sigma_i(X) \le b_i$ shall be changed to $\sigma_i(X) < b_i$.

Corollary 4.8. *Suppose that $\mathcal{X}$ is a non-singleton compact set satisfying* (35)*, and ϕ is concave on $[0,\infty)$ satisfying Assumption 3 and $\liminf_{t\to 0^+} \phi(t)/t > 0$.*[8] *Assume that g is Lipschitz continuous on $\mathcal{X}$, that is, there exists some $L_g > 0$ such that*

$$|g(X) - g(Y)| \le L_g \|X - Y\|_* \qquad \forall X, Y \in \mathcal{X}. \tag{40}$$

Then the conclusion of Theorem 4.1 holds with $L = L_g/M_\phi$, where

$$M_\phi = \inf_{t\in[0,R]} \frac{\phi(t)}{t}, \qquad R = \max_{X,Y\in\mathcal{X}} \|X - Y\|.$$

Proof. By the definition of R, one can see that for every i,

$$\sigma_i(X - Y) \le \|X - Y\| \le R \qquad \forall X, Y \in \mathcal{X}.$$

By this and a similar argument as in the proof of Corollary 3.5, one has that for every i,

$$\phi(\sigma_i(X - Y)) \ge M_\phi \sigma_i(X - Y) \qquad \forall X, Y \in \mathcal{X}.$$

It follows from this and (40) that

$$|g(X) - g(Y)| \le L_g \|X - Y\|_* \le \frac{L_g}{M_\phi} \sum_{i=1}^{q} \phi(\sigma_i(X - Y)), \qquad \forall X, Y \in \mathcal{X},$$

and hence (25) holds with $L = L_f/M_\phi$, where $q = \min(m, n)$. The conclusion then follows from Corollary 4.6. ∎

Remark 4 It was established in [3] that problem (6) with $\phi(t) = t$ is an exact penalization of problem (2) with some simple closed set $\mathcal{X}$.

We next establish some relation between the local minimizers of problem (2) and those of problem (6). In particular, we show that under some suitable assumptions, a local minimizer of (2) is also that of (6).

Theorem 4.9. *Suppose that* (24) *and Assumption 4 hold. Let X^* be a local minimizer of problem* (2)*. Assume that there exist some $\epsilon > 0$ and $L > 0$ such that*

$$|g(X) - g(Y)| \le L \sum_{i=1}^{q} \phi(\sigma_i(X - Y)) \qquad \forall X, Y \in \mathcal{B}(X^*;\epsilon) \cap \mathcal{X}, \tag{41}$$

where $q = \min(m, n)$. Then X^* is a local minimizer of problem (6) *whenever $\lambda \ge L$.*

Proof. Let $\Omega_1 = \mathcal{X}$, $\Omega_2 = \{X \in \Re^{m\times n} : \text{rank}(X) \le r\}$ and $\Phi(X,Y) = \sum_{i=1}^q \phi(\sigma_i(X - Y))$ for all $X, Y \in \Re^{m\times n}$. In view of (24), Assumption 4 and the same argument as in the proof of Theorem 4.1, one can see that Assumption 1 holds for such Φ, Ω_1 and Ω_2. Notice from

[8]The regularizers ℓ_1, ℓ_p, Log, Capped-ℓ_1, MCP and SCAD satisfy these assumptions.

Assumption 4 that $\phi(0) = 0$ and ϕ is right continuous at 0. Using these and the fact that $\sigma_i(\cdot)$ is a continuous function for each i, we see that Assumption 2(b) holds for such Φ. We next show that Assumption 2(a) also holds for such Φ. To this end, let $\tilde{\epsilon} > 0$ be arbitrarily chosen. By a similar argument as in the proof of Theorem 3.6, there exists some $\delta > 0$ such that $t \in [0, \tilde{\epsilon}]$ whenever $\phi(t) \leq \delta$. By this, the expression of Φ, and the nonnegativity of ϕ, one can observe that $\sigma_i(X - Y) \leq \tilde{\epsilon}$ for all i if $\Phi(X, Y) \leq \delta$, which implies that $\|X - Y\| \leq \tilde{\epsilon}$ whenever $\Phi(X, Y) \leq \delta$. Hence, Assumption 2(a) holds for the above Φ. It then follows from (41) and Theorem 2.2 with $h = g$ that X^* is also a local minimizer of problem (27). Notice that $\mathrm{rank}(X^*) \leq r$. By the same argument as in the proof of Theorem 4.1, one can see that (28) holds at X^*. From this and the proof of Theorem 4.1, we know that the objective of (27) is majorized by that of (6) and they achieve the same value at X^*. By these and the result that X^* is a local minimizer of (27), one can conclude that X^* is also a local minimizer of problem (6). ■

As a consequence of Theorem 4.9 and Proposition 4.3, we have the following result.

Corollary 4.10. *Suppose that* (35)*,* (41) *and Assumption 4 hold. Assume that there exists a strictly increasing function* $\psi : \Re_+ \to \Re \cup \{\infty\}$ *such that* $\psi(\sum_{i=1}^q \phi(|x_i|))$ *is a norm on* $\Re^q$*, where* $q = \min(m, n)$*. Then the conclusion of Theorem 4.9 holds.*

The following result is an immediate consequence of Theorem 4.9 and Proposition 4.4.

Corollary 4.11. *Suppose that* (35)*,* (41) *and Assumption 4 hold. Assume Additionally that* ϕ *is concave in* $[0, \infty)$*. Then the conclusion of Theorem 4.9 holds.*

The following result is a consequence of Corollaries 4.10 and 4.11, whose proof is similar to that of Corollary 4.7 and thus omitted.

Corollary 4.12. *Suppose that the set* $\mathcal{X}$ *satisfies* (35)*. Let* X^* *be a local minimizer of problem* (2)*. Assume that there exist some* $\epsilon > 0$*,* $p > 0$*, and* $L > 0$ *such that*

$$|g(X) - g(Y)| \leq L \sum_{i=1}^{q} [\sigma_i(X - Y)]^p \qquad \forall X, Y \in \mathcal{B}(X^*; \epsilon) \cap \mathcal{X}.$$

Then X^* *is a local minimizer of problem* (39) *whenever* $\lambda \geq L$*.*

The following result is a consequence of Corollary 4.11, which holds for various ϕ such as ℓ_1, ℓ_p, Log, Capped-ℓ_1, MCP and SCAD. It can be viewed as a generalization of Corollary 3.8 from vector to matrix.

Corollary 4.13. *Suppose that* $\mathcal{X}$ *satisfies* (35)*, and* ϕ *is concave on* $[0, \infty)$ *satisfying Assumption 4 and* $\liminf_{t \to 0^+} \phi(t)/t > 0$*.*[9] *Assume that there exist some* $\epsilon > 0$ *and* $L_g > 0$ *such that*

$$|g(X) - g(Y)| \leq L_g \|X - Y\|_* \qquad \forall X, Y \in \mathcal{B}(X^*; \epsilon) \cap \mathcal{X}.$$

Then X^* *is a local minimizer of problem* (6) *for any* $\lambda \geq L_f / M_\phi$*, where* $M_\phi = \inf_{t \in [0, 2\epsilon]} \phi(t)/t$*.*

[9] The regularizers ϕ such as ℓ_1, ℓ_p, Log, Capped-ℓ_1, MCP and SCAD satisfy these assumptions.

Proof. One can observe that for every i,

$$\sigma_i(X-Y) \le \|X-Y\| \le \|X-X^*\| + \|Y-X^*\| \le 2\epsilon \qquad \forall X,Y \in \mathcal{B}(X^*;\epsilon)\cap\mathcal{X}.$$

By this and a similar argument as in the proof of Corollaries 3.8 and 4.8, one can have

$$|g(X)-g(Y)| \le \frac{L_g}{M_\phi}\sum_{i=1}^{q}\phi(\sigma_i(X-Y)) \qquad \forall X,Y \in \mathcal{B}(X^*;\epsilon)\cap\mathcal{X},$$

and hence (41) holds with $L = L_g/M_\phi$, where $q=\min(m,n)$. The conclusion then follows from Corollary 4.11. ∎

Remark 5 (i) In many applications, g is often Lipschitz continuous on the set $\mathcal{X}$ satisfying (35). It then follows from Corollaries 4.7 and 4.12 with $p=1$ that the partial nuclear-norm regularized model

$$\min_{X\in\mathcal{X}} g(X) + \lambda\sum_{i=r+1}^{q}\sigma_i(X)$$

is an exact penalty reformulation of problem (2) for some $\lambda>0$, where $q=\min(m,n)$. Further, if $\mathcal{X}$ is compact, it follows from Corollaries 4.8 and 4.13 that the partially regularized model (6) for various ϕ such as ℓ_p, Log, Capped-ℓ_1, MCP and SCAD is an exact penalty reformulation of (2). Therefore, Corollaries 4.7, 4.8, 4.12 and 4.13 provide some theoretical justification on the often-observed superior performance of a partial regularizer over a corresponding full regularizer in finding a low-rank approximate solution.

(ii) Corollaries 4.8 and 4.13 generalize the results of [29] that were developed for the case where $\phi(t)=t^p$ with $p\in(0,1]$, $\mathcal{X}$ is an semidefinite box, and g is Lipschitz continuous on $\mathcal{X}$.

5 Concluding remarks

In this paper we studied exact penalization for a class of cardinality and rank constrained optimization problems. In particular, under some suitable assumptions, we showed that the penalty model based on a partial regularization is an exact reformulation of the original problem. We also showed that a local minimizer of the latter problem is that of the former one. These properties, however, generally do not hold for the penalty model based on a full regularization. Our results provide some theoretical justification for the often-observed superior performance of a partial regularizer over a corresponding full regularizer. In addition, our results indicate some dependence of the partial regularizer on the objective of the considered problem, which can provide some guidance on how to choose a partial regularizer for a given problem.

References

[1] M. Ahn, J.S. Pang, and J. Xin. Difference-of-convex learning I: directional stationarity, optimality, and sparsity. *SIAM J. Optim.*, 27(3):1637-1665, 2017.

[2] G. Banjac and P. J. Goulart. A novel approach for solving convex problems with cardinality constraints. IFAC (International Federation of Automatic Control) PapersOnline 50(1):13182–13187, 2017. DOI: 10.1016/j.ifacol.2017.08.2174.

[3] S. Bi and S. Pan. Error bounds for rank constrained optimization problems and applications. *Oper. Res. Lett.*, 44(3):336-341, 2016.

[4] R. Borsdorf, N. J. Higham, and M. Raydan. Computing a nearest correlation matrix with factor structure. *SIAM J. Matrix Anal. & Appl.*, 31(5):2603–2622, 2010.

[5] J. V. Burke. Calmness and exact penalization. *SIAM J Control Optim.*, 29(2):493-497, 1991.

[6] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9(6):717–772, 2009.

[7] E. J. Candès and T. Tao. Decoding by linear programming. *IEEE T. Inform. Theory*, 51(12):4203–4215, 2005.

[8] E. J. Candès, M. B. Wakin, and S. P. Boyd. Enhancing sparsity by reweighted ℓ_1 minimization. *J. Fourier Anal. Appl.*, 14(5):877–905, 2008.

[9] R. Chartrand. Exact reconstruction of sparse signals via nonconvex minimization. *IEEE Signal Proc. Let.*, 14(10):707–710, 2007.

[10] S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, 20(1):33–61, 1998.

[11] X. Chen, Z. Lu, and T. K. Pong. Penalty methods for a class of non-Lipschitz optimization problems. *SIAM J. Optimiz.*, 26(3):1465–1492, 2016.

[12] X. Chen, L. Pan and N. Xiu. Solution sets of three sparse optimization problems for multivariate regression. *J. Glob. Optim.*, 2022. https://doi.org/10.1007/s10898-021-01124-w.

[13] F. H. Clarke. Optimization and nonsmooth analysis, SIAM, 1990.

[14] C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1:211–218, 1936.

[15] L. Eldén. Matrix Methods in Data Mining and Pattern Recognition. SIAM, Philadelphia, PA, USA, 2007.

[16] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96(456):1348–1360, 2001.

[17] M. Fazel, H. Hindi, and S. P. Boyd. A rank minimization heuristic with application to minimum order system approximation. In *Proceedings of the 2001 American Control Conference*, 6:4734–4739, 2001.

[18] I. Frank and J. Friedman. A statistical view of some chemometrics regression tools (with discussion). *Technometrics*, 35(2):109–148, 1993.

[19] W. J. Fu. Penalized regression: the bridge versus the lasso. *J. Comput. Graph. Statist.*, 7(3):397–416, 1998.

[20] Y. Gao. Structured low rank matrix optimization problems: a penalized approach. PhD Thesis, Department of Mathematics, National University of Singapore, 2010.

[21] Y. Gao and D. Sun. A majorized penalty approach for calibrating rank constrained correlation matrix problems. Technical report, Department of Mathematics, National University of Singapore, 2010.

[22] G. Gasso, A. Rakotomamonjy, and S. Canu. Recovering sparse signals with a certain family of nonconvex penalties and DC programming. *IEEE T. Signal Proces.*, 57(12):4686–4698, 2009.

[23] J. Gotoh, A. Takeda, and K. Tono. DC formulations and algorithms for sparse optimization problems. *Math. Program.*, 169(1):141–176, 2018.

[24] M. Herman and T. Strohmer. High-resolution radar via compressed sensing. *IEEE T. Signal Proces.*, 57(6):2275–2284, 2009.

[25] R. A. Horn and C. R. Johnson. Matrix Analysis. Cambridge University Press, New York, NY, USA, 1990.

[26] R. Horn and C. Johnson. Topics in Matrix Analysis. Cambridge University Press, , New York, NY, USA, 1991.

[27] X. L. Huang, L. Shi, and M. Yan. Nonconvex sorted ℓ_1 minimization for sparse approximation. *Journal of the Operations Research Society of China*, 3(2):207–229, 2015.

[28] S. Ji, K. F. Sze, Z. Zhou, A. M. C. So, and Y. Ye. Beyond convex relaxation: A polynomial-time non-convex optimization approach to network localization. In *2013 Proceedings IEEE INFOCOM*, 72–85, 2013.

[29] T. Liu, Z. Lu, X. Chen, and Y.-H. Dai. An exact penalty method for semidefinite-box constrained low-rank matrix optimization problems. *IMA J. Numer. Anal.*, 40(1):563–586, 2020.

[30] Z. Lu and X. Li. Sparse recovery via partial regularization: models, theory and algorithms. *Math. Oper. Res.*, 43(4): 1290–1316, 2018.

[31] Z. Lu, Y. Zhang, and J. Lu. ℓ_p regularized low-rank approximation via iterative reweighted singular value minimization. *Comput. Optim. Appl.*, 68(3):619–642, 2017.

[32] M. Lustig, D. L. Donoho, and J. M. Pauly. Sparse MRI: The application of compressed sensing for rapid MR imaging. *Magn. Reson. Med.*, 58(6):1182–1195, 2008.

[33] A. Y. Ng. Feature selection, L1 vs. L2 regularization, and rotational invariance. In *Proceedings of the Twenty-First International Conference on Machine Learning (ICML)*, 72–85, 2004.

[34] J. S. Pang. Difference-of-convex statistical learning: directional stationarity, optimality, and sparsity. Plenary talk at the Fifth International Conference on Continuous Optimization, Tokyo, Japan, August 11, 2016.

[35] Y. Qian and S. Pan. Calmness of partial perturbation to composite rank constraint systems and its applications. arXiv preprint arXiv:2102.10373, 2021.

[36] B. Recht, M. Fazel, and P. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.*, 52(3):471–501, 2010.

[37] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B*, 58(1):267–288, 1996.

[38] J. A. Tropp, J. N. Laska, M. F. Duarte, J. K. Romberg, and R. G. Baraniuk. Beyond nyquist: efficient sampling of sparse bandlimited signals. *IEEE T. Inform. Theory*, 56(1):520–544, 2010.

[39] J. Weston, A. Elisseeff, B. Scholkopf, and M. Tipping. The use of zero-norm with linear models and kernel methods. *J. Mach. Learn. Res.*, 3:1439–1461, 2003.

[40] S. J. Wright, R. D. Nowak and M. Figueiredo. Sparse reconstruction by separable approximation. *IEEE T. Signal Proces.*, 57(7):2479–2493, 2009.

[41] Q. Yao, J. T. Kwok, T. Wang, and T. Liu. Large-scale low-rank matrix learning with nonconvex regularizers. *IEEE T. Pattern Anal. Mach. Intell.*, early access, 2018.

[42] J. J. Ye, D. L. Zhu and Q. Zhu. Exact penalization and necessary optimality conditions for generalized bilevel programming problems. *SIAM J. Optim.*, 7(2):481-507, 1991.

[43] M. Yuan and Y. Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94:19–35, 2007.

[44] M. Yuan, A. Ekici, Z. Lu, and R. Monteiro. Dimension reduction and coefficient estimation in multivariate linear regression. *J. R. Statist. Soc. B*, 69(3):329–346, 2007.

[45] M.-C. Yue and A. M.-C. So. A perturbation inequality for concave functions of singular values and its applications in low-rank matrix recovery. *Appl. Comput. Harmon. Anal.*, 40(2):396–416, 2016.

[46] C. H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, 38(2):894–942, 2010.

[47] T. Zhang. Analysis of multi-stage convex relaxation for sparse regularization. *J. Mach. Learn. Res.*, 11:1081–1107, 2010.

[48] Y. Zhang, B. Dong, and Z. Lu. ℓ_0 minimization for wavelet frame based image restoration. *Math. Compu.*, 82(282):995–1015, 2013.