

AN EXACT PENALTY METHOD FOR SPARSITY-CONSTRAINED OPTIMIZATION

Jianing Zhao[†] Junbin Liu^{†,‡} Anthony Man-Cho So[†] Wing-Kin Ma[‡]

[†] Dept. of Syst. Eng. & Eng. Manag., The Chinese Univ. of Hong Kong, Hong Kong SAR, China

[‡] Dept. of Electron. Eng., The Chinese Univ. of Hong Kong, Hong Kong SAR, China

ABSTRACT

Sparsity-constrained optimization appears in various applications where the solution exhibits specific sparsity patterns. These problems are challenging in general, mainly due to the nonconvex and combinatorial nature of the constraints. In this work, we develop a nonconvex penalty method based on the error bound principle that accommodates a wide range of sparsity structures, including vector sparsity, low-rank, and group sparsity, as well as their combinations with additional constraints such as nonnegativity. We show that the proposed penalty method can preserve both the global and local solution sets of the original constrained problem. Algorithmically, we design a nonconvex proximal gradient method and show that its proximal operator admits a simple closed-form solution, despite the nonconvexity. Taking sparse linear regression as an example, we compare our approach with other penalty methods, validating its potential empirically.

Index Terms— sparsity-constrained optimization, error bound, exact penalization, nonconvex proximal gradient method

1. INTRODUCTION

Sparsity is a widely studied and utilized property that manifests in various structured forms across numerous applications. For instance, spectral sparsity lies at the core of compressive sensing [1]; group sparsity is effectively employed in direction-of-arrival estimation [2]; singular value sparsity is fundamental to low-rank matrix problems such as matrix completion [3]; and more recently, sparsity structures have been explored in Transformer architectures to enhance scalability and computational efficiency [4].

Sparsity often appears as a constraint in optimization problems. However, sparsity constraints are generally challenging to handle. In this work, we explore penalty-based approaches for solving sparsity-constrained optimization problems. Our framework is built upon the error bound principle [5], a powerful tool that has been instrumental in the development of penalty methods; see, e.g., [6, 7] for recent applications to bilevel optimization. Penalized formulations derived using this principle are equivalent reformulations of the original problem, i.e., they share the same set of global optimal solutions. This theoretical guarantee makes the error bound principle particularly appealing. In addition, error bounds play a fundamental role in obtaining strong convergence rate results for iterative methods; see, e.g., [8, 9, 10]. Recently, the error bound principle has also served as a foundational basis for addressing a wide range of constraints, including binary and discrete phase constraints, binary selection constraints, sign-constrained Stiefel manifold constraints, and permutation matrix constraints [11, 12, 13]. However, the use of error bounds in handling sparsity constraints remains underexplored.

In this regard, we introduce a tailored error bound-based method that leads to a general penalty framework capable of handling a variety of sparsity structures, including vector sparsity, low-rank constraints, and group sparsity, as well as their possible coexistence with other additional constraints such as the nonnegativity constraint. Moreover, we establish strong equivalence guarantees between the penalized formulation and the original sparsity-constrained problem in terms of local optimal solution sets.

First-order optimization methods can readily tackle the resulting penalty formulations. In particular, we develop a nonconvex proximal gradient (PG) method, where the nonconvex proximal operator can be evaluated efficiently. To validate the proposed approach, we apply it to sparse linear regression and demonstrate its efficacy.

2. BACKGROUND

Sparsity-constrained optimization involves problems where the solution exhibits specific sparsity patterns, such as sparsity in vector elements, matrix rows or columns, or singular values. A classic formulation is

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) \quad \text{s.t.} \quad \|\mathbf{x}\|_0 \leq k, \quad (1)$$

where f is the objective function, $\|\mathbf{x}\|_0$ denotes the ℓ_0 pseudo-norm that counts the number of nonzero entries in \mathbf{x} , and k is a positive integer controlling the sparsity level. Problems of the form (1) are generally challenging due to the nonconvex and discrete nature of the sparsity constraint set. In fact, many instances of (1) have been shown to be NP-hard [14, 15, 16].

A substantial body of research has attempted to develop algorithmic approaches for solving sparsity-constrained problems, such as greedy algorithms [17] and approaches based on mixed-integer optimization [18]. In this work, we focus on penalty methods that reformulate (1) as

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) + \lambda h(\mathbf{x}), \quad (2)$$

where h is a penalty function and $\lambda > 0$ is given. Among penalty-based approaches, one of the most well-known is LASSO [19], proposed in the context of linear regression. LASSO replaces the nonconvex ℓ_0 pseudo-norm with the convex ℓ_1 -norm to promote sparsity. However, ℓ_1 -norm is convex and it may not always well approximate the pseudo-norm. This limitation has motivated the development of nonconvex penalties that more closely approximate the sparsity constraint, such as the smoothly clipped absolute deviation (SCAD) [20], the minimax concave penalty (MCP) [21], the pseudo-norm as $p < 1$ [22], the logarithmic penalty [23], and the Ky Fan k -norm-based nonconvex penalty [24].

In penalty methods, a central aspect is whether the penalized formulation shares the same set of optimal solutions as the original sparsity-constrained problem. Such an equivalence is termed *exact penalization*. Early studies have shown that, in the context of sparse

This work was supported by a General Research Fund (GRF) of Hong Kong Research Grant Council (RGC) under Project ID CUHK 14203721.

linear regression, the ℓ_1 -norm penalty yields exact penalization under certain conditions [25, 26]. For nonconvex penalties, such as the logarithmic penalty and the ℓ_p -norm penalty with $p \in (0, 1)$, weaker forms of exact penalization have been established [27, 28]. Specifically, under certain conditions, the solution sets of the penalized and original problems intersect. Exact penalization results have also been obtained for nonconvex penalties based on the Ky Fan k -norm [24], assuming certain regularity conditions on the objective function and the structure of the optimal solution set. All of the aforementioned results pertain to global optimality. By contrast, results concerning the equivalence of local optimal solution sets are relatively scarce. In practice, due to the nonconvex nature of the optimization landscape, algorithms may converge to local minima. However, these local solutions may not correspond to feasible solutions of the original problem without equivalence between local optimal solution sets. Therefore, establishing such an equivalence for nonconvex penalization formulations is of both theoretical and practical value.

3. PROPOSED METHOD

We start with a brief review of the error bound principle.

3.1. Error Bound Principle and Exact Penalization

Consider an optimization problem over a non-empty closed set \mathcal{A} :

$$\min_{\mathbf{x} \in \mathcal{A}} f(\mathbf{x}). \quad (3)$$

The definition of error bound function regarding the set \mathcal{A} is given below.

Definition 1. Let $\mathcal{C} \subseteq \mathbb{R}^n$ and $\mathcal{A} \subseteq \mathcal{C}$. A function $\psi : \mathbb{R}^n \rightarrow \mathbb{R}$ is called an **error bound function** of \mathcal{A} relative to \mathcal{C} if it satisfies

$$\begin{aligned} \text{dist}(\mathbf{x}, \mathcal{A}) &\leq \psi(\mathbf{x}), & \forall \mathbf{x} \in \mathcal{C}, \\ \psi(\mathbf{x}) &= 0, & \forall \mathbf{x} \in \mathcal{A}, \end{aligned}$$

where $\text{dist}(\mathbf{x}, \mathcal{A}) = \inf\{\|\mathbf{x} - \mathbf{z}\|_2 \mid \mathbf{z} \in \mathcal{A}\}$.

Building on this definition, the error bound function, when used as a penalty, can lead to exact penalization, as established in the following proposition.

Proposition 1. (Proposition 9.1.1 in [5]) Let $\mathcal{A} \subseteq \mathbb{R}^n$ be non-empty closed, let $\mathcal{C} \subseteq \mathbb{R}^n$ be some set such that $\mathcal{A} \subseteq \mathcal{C}$, let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be K -Lipschitz continuous on \mathcal{C} , and let $\psi : \mathbb{R}^n \rightarrow \mathbb{R}$ be an error bound function of \mathcal{A} relative to \mathcal{C} . Suppose that $\arg\min_{\mathbf{x} \in \mathcal{A}} f(\mathbf{x}) \neq \emptyset$.

For any $\lambda > K$, the following holds:

$$\arg\min_{\mathbf{x} \in \mathcal{A}} f(\mathbf{x}) = \arg\min_{\mathbf{x} \in \mathcal{C}} f(\mathbf{x}) + \lambda\psi(\mathbf{x}).$$

The key idea behind the error bound principle is to design a valid error bound function that is amenable to algorithmic design. We note that Proposition 1 characterizes the equivalence between global optimal solution sets.

3.2. Error Bound Principle for Sparsity Constraints

Using the error bound principle, we can derive penalties for sparsity constraints and their combinations with other constraints. We define some notation first. Let k be a given positive integer. Let $\|\mathbf{x}\|_{[k]}$ denote the Ky Fan k -norm that sums up the first k absolutely largest

elements of \mathbf{x} ; i.e., $\|\mathbf{x}\|_{[k]} = \sum_{i=1}^k |x_{[i]}|$, where $|x_{[i]}|$ denotes the i -th largest value of $|x_1|, \dots, |x_n|$. The mixed- (ℓ_p, ℓ_q) norm $\|\mathbf{X}\|_{p,q}$ is defined as $\|\mathbf{X}\|_{p,q} = \|\tilde{\mathbf{x}}\|_q$, where each entry of the vector $\tilde{\mathbf{x}} \in \mathbb{R}^m$ is given by $\tilde{x}_i = \|\tilde{\mathbf{x}}_i\|_p$ and $\tilde{\mathbf{x}}_i$ denotes the i -th row of \mathbf{X} . Similarly, we define $\|\mathbf{X}\|_{p,0} = \|\tilde{\mathbf{x}}\|_0$ and $\|\mathbf{X}\|_{p,[k]} = \|\tilde{\mathbf{x}}\|_{[k]}$. The vector $\boldsymbol{\sigma}(\mathbf{X})$ contains the singular values of the matrix \mathbf{X} .

Lemma 1. We have the following error bounds:

Vector Sparsity: Let $\mathcal{A} = \{\mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x}\|_0 \leq k\}$. The function $\mathbf{x} \mapsto \psi(\mathbf{x}) = \|\mathbf{x}\|_1 - \|\mathbf{x}\|_{[k]}$ is an error bound function of \mathcal{A} relative to \mathbb{R}^n .

Vector Sparsity and Nonnegativity: Let $\mathcal{A}_1 = \{\mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x}\|_0 \leq k\}$ and $\mathcal{A}_2 = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{x} \geq \mathbf{0}\}$. The function $\mathbf{x} \mapsto \psi(\mathbf{x}) = \|\mathbf{x}\|_1 - \|\mathbf{x}\|_{[k]}$ is an error bound function of $\mathcal{A}_1 \cap \mathcal{A}_2$ relative to \mathbb{R}_+^n .

Group Sparsity: Let $\mathcal{A} = \{\mathbf{X} \in \mathbb{R}^{m \times n} \mid \|\mathbf{X}\|_{2,0} \leq k\}$ where $k \leq \min\{m, n\}$. The function $\mathbf{X} \mapsto \psi(\mathbf{X}) = \|\mathbf{X}\|_{2,1} - \|\mathbf{X}\|_{2,[k]}$ is an error bound function of \mathcal{A} relative to $\mathbb{R}^{m \times n}$.

Group Sparsity and Nonnegativity: Let $\mathcal{A}_1 = \{\mathbf{X} \in \mathbb{R}^{m \times n} \mid \|\mathbf{X}\|_{2,0} \leq k\}$ where $k \leq \min\{m, n\}$, and $\mathcal{A}_2 = \{\mathbf{X} \in \mathbb{R}^{m \times n} \mid \mathbf{X} \geq \mathbf{0}\}$. The function $\mathbf{X} \mapsto \psi(\mathbf{X}) = \|\mathbf{X}\|_{2,1} - \|\mathbf{X}\|_{2,[k]}$ is an error bound function of $\mathcal{A}_1 \cap \mathcal{A}_2$ relative to $\mathbb{R}_+^{m \times n}$.

Low-Rank Constraint: Let $\mathcal{A} = \{\mathbf{X} \in \mathbb{R}^{m \times n} \mid \text{rank}(\mathbf{X}) \leq k\}$ where $k \leq \min\{m, n\}$. The function $\mathbf{X} \mapsto \psi(\mathbf{X}) = \|\boldsymbol{\sigma}(\mathbf{X})\|_1 - \|\boldsymbol{\sigma}(\mathbf{X})\|_{[k]}$ is an error bound function of \mathcal{A} relative to $\mathbb{R}^{m \times n}$.

Lemma 1 shows that the error bound principle enables straightforward derivation of penalty functions for various sparsity patterns, as well as their combinations with additional constraints, that achieve exact penalization. While we illustrate the inclusion of nonnegativity constraints, the framework can easily accommodate other types of constraints, for instance, sparsity within each row in the group sparsity case. Notably, this flexibility offers an advantage over existing penalty methods tailored solely to sparsity constraints. A proof for the vector sparsity case in Lemma 1 is provided in Appendix 6.1. We omit the proofs for the other cases in Lemma 1 due to space limitations.

More intriguingly, in the case of sparsity constraints, we are able to extend the general error bound exact penalization results and establish the equivalence between local optimal solution sets, as stated in the following proposition.

Proposition 2. Consider the sparsity-constrained problem (3) and its penalized formulation (2). Suppose that the objective function f is K -Lipschitz continuous. Then, for the sparsity constraints listed in Lemma 1, when using the associated error bound functions, both the **global** and **local** optimal solution sets of (3) and (2) are the same for $\lambda > K$.

The proof is relegated to Appendix 6.2.

3.3. Related Work

The penalty functions stated in Lemma 1, except for the nonnegative ones, were studied in [24]. In that prior work, the aforementioned penalty functions were shown to provide exact penalization in terms of the globally optimal solution set. However, the proof therein is not based on error bounds. It assumes that the original problem possesses a smooth objective function with a Lipschitz continuous gradient, and the penalized problem (2) has a bounded solution set. By contrast, our method is based on the error bound principle. It has no requirement on the solution set and assumes only Lipschitz continuity, which accommodates certain nonsmooth objectives. Moreover, we establish local optimality set equivalence—a result not seen

in [24]. Our framework is also more flexible, capable of straightforwardly incorporating additional constraints, such as nonnegativity, which was not considered in [24].

3.4. A Nonconvex Proximal Gradient Algorithm

Apart from the theoretical analysis of exact penalization, we take the typical sparse linear regression problem as an example and develop a nonconvex PG algorithm. The regression model can be written as

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{v}, \quad \text{s.t.} \quad \|\mathbf{x}\|_0 \leq k, \quad (4)$$

where $\mathbf{y} \in \mathbb{R}^m$ is the observation vector; $\mathbf{A} \in \mathbb{R}^{m \times n}$ is the known sensing matrix with $m \leq n$; \mathbf{v} denotes the noise vector. The task is to estimate the sparse vector \mathbf{x} given \mathbf{A} and \mathbf{y} .

This estimation is commonly formulated as an optimization problem with the objective function $\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2$. However, this function is not Lipschitz continuous over \mathbb{R}^n , which is a required condition for exact penalization as stated in Proposition 2. To address this issue, we develop a trick described in the following lemma. It essentially shows that for a class of differentiable functions that are not Lipschitz continuous over \mathbb{R}^n , applying a square root transformation can yield a Lipschitz continuous function.

Lemma 2. *Let $f : \mathbb{R}^n \rightarrow [0, \infty)$ be a differentiable function such that $\|\nabla f(\mathbf{x})\|_2^2 \leq C f(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^n$, where $C > 0$ is a constant. Define $g : \mathbb{R}^n \rightarrow (0, \infty)$ by $g(\mathbf{x}) = \sqrt{f(\mathbf{x}) + \epsilon}$ for some $\epsilon > 0$. Then, the function g is Lipschitz continuous on \mathbb{R}^n .*

It is obvious that minimizing $\|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2$ is the same as minimizing $\sqrt{\|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 + \epsilon}$. We can therefore formulate the sparse linear regression problem as

$$\min_{\mathbf{x} \in \mathbb{R}^n} \underbrace{\sqrt{\|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 + \epsilon}}_{=f(\mathbf{x})} + \lambda \underbrace{(\|\mathbf{x}\|_1 - \|\mathbf{x}\|_{[k]})}_{=h(\mathbf{x})}, \quad (P)$$

where $\epsilon > 0$. By Lemma 2, f is Lipschitz continuous on \mathbb{R}^n . Additionally, it can be shown that the Lipschitz constant of f is $\|\mathbf{A}\|_2$.

To handle problem (P), we consider the PG algorithm [29], whose iterations are given by

$$\mathbf{x}^{\ell+1} = \text{prox}_{\eta_\ell \lambda h}(\mathbf{x}^\ell - \eta_\ell \nabla f(\mathbf{x}^\ell)), \quad \ell = 0, 1, \dots, \quad (5)$$

where $\eta_\ell > 0$ denotes the step size at iteration ℓ . The proximal operator is defined as

$$\text{prox}_{\eta_\ell \lambda h}(\mathbf{z}) = \arg \min_{\mathbf{x} \in \mathbb{R}^n} \left\{ \frac{1}{2} \|\mathbf{z} - \mathbf{x}\|_2^2 + \eta_\ell \lambda h(\mathbf{x}) \right\}. \quad (6)$$

A key aspect of PG methods is whether the proximal operator can be evaluated efficiently. In our case, the penalty function is given by $\mathbf{x} \mapsto h(\mathbf{x}) = \|\mathbf{x}\|_1 - \|\mathbf{x}\|_{[k]}$, which is nonconvex. Nonetheless, we show in the following proposition that a global optimal solution to the nonconvex proximal problem (6) can be obtained efficiently.

Proposition 3. *Consider the following nonconvex optimization problem*

$$\min_{\mathbf{x} \in \mathbb{R}^n} \frac{1}{2} \|\mathbf{z} - \mathbf{x}\|_2^2 + \mu (\|\mathbf{x}\|_1 - \|\mathbf{x}\|_{[k]}), \quad (7)$$

where $\mu > 0$ is given. Let π be a permutation of $\{1, \dots, n\}$ such that $|z_{\pi(1)}| \geq |z_{\pi(2)}| \geq \dots \geq |z_{\pi(n)}|$. Then, an optimal solution $\tilde{\mathbf{x}}$ to (7) is given by

$$\tilde{x}_i = \begin{cases} z_{\pi(i)}, & \text{if } i \leq k, \\ \text{sign}(z_{\pi(i)}) \cdot \max(|z_{\pi(i)}| - \mu, 0), & \text{otherwise.} \end{cases} \quad (8)$$

Hence, each iteration of the proposed PG method admits a closed-form expression. Moreover, since f and h are proper closed semi-algebraic functions, $f + h$ satisfies the Kurdyka-Łojasiewicz property, ensuring convergence to a critical point [30, Theorem 5.1].

In practice, it has been observed that algorithms may get stuck at poor local minima when the penalty parameter λ is directly set to a large value for nonconvex penalties [31, 32]. To mitigate this issue, a common strategy is to anneal λ from a small initial value to a large one, gradually approaching the regime of exact penalization. We adopt this annealing strategy in our implementation. Additionally, we anneal the parameter ϵ from a relatively large value toward zero. Proposition 3 can be extended to the other sparsity cases in Lemma 1, and we omit the details due to space limitations.

4. NUMERICAL RESULTS

In this section, we make a numerical comparison with other penalty methods for sparse linear regression. The benchmarks are: 1) LASSO [19]; 2) the iterative hard thresholding (IHT) algorithm [33]; 3) smoothly clipped absolute deviation (SCAD) [21]; 4) the difference of convex algorithm (DCA) by [24]; and (5) the logarithmic penalty algorithm [23]. Synthetic data are generated following standard procedures. The sensing matrix $\mathbf{A} \in \mathbb{R}^{128 \times 256}$ has i.i.d. entries drawn from $\mathcal{N}(0, 1/k)$, where k is the sparsity level. The ground truth vector \mathbf{x}^* is k -sparse, with nonzero entries sampled uniformly from $[-1, 1]$ and positions selected uniformly at random. Observations are generated via the linear model $\mathbf{y} = \mathbf{A}\mathbf{x}^* + \mathbf{v}$, where \mathbf{v} is Gaussian noise with zero mean and variance 0.01. All algorithms are initialized with \mathbf{x} drawn uniformly from $[-1, 1]^n$. For our method, the parameters λ and ϵ are annealed by increasing λ exponentially by a factor of 1.5 per iteration and applying exponential decay to ϵ with a rate of 0.5. We conduct 50 independent trials and report average results.

k	Method	PG	IHT	DCA	SCAD	LOG	LASSO
16	Error	0.076	0.077	0.112	0.430	0.579	0.534
	Prob	0.962	0.950	0.890	0.710	0.765	0.875
	Supp	16.0	16.0	15.6	11.4	15.4	15.9
	Time	0.001	0.003	0.039	0.026	0.086	0.006
24	Error	0.166	0.171	0.334	0.732	1.081	1.285
	Prob	0.923	0.923	0.819	0.702	0.669	0.712
	Supp	24.0	24.0	23.8	18.9	22.9	21.7
	Time	0.001	0.002	0.017	0.025	0.114	0.015
32	Error	0.292	0.598	0.832	1.563	1.902	2.291
	Prob	0.893	0.828	0.705	0.605	0.553	0.520
	Supp	32.0	32.0	31.8	22.2	31.3	20.7
	Time	0.001	0.004	0.101	0.067	0.180	0.058

Table 1. Performance of algorithms with different sparsity k .

The performance of each benchmark is evaluated according to the following criteria: 1) error, measured as the ℓ_2 -norm distance between the numerical solution and the ground truth \mathbf{x}^* ; 2) recovery probability (prob), defined as the proportion of nonzero positions that are common to both the solution and to \mathbf{x}^* ; 3) support size (supp), the number of nonzero elements in the solution, indicating the feasibility of the result; and 4) time, the CPU time required to complete each trial. The numerical results for the comparison of different algorithms are shown in Table 1. As we can see, the proposed method demonstrates competitive performance relative to existing

approaches, providing empirical support for our theoretical developments.

5. CONCLUSION

In this work, we have proposed a nonconvex penalty framework for sparsity-constrained optimization by the error bound principle. This framework accommodates various structures, including vector, group, and low-rank sparsity. We have shown that it preserves the global and local solution sets of the original problem. The penalized problem is efficiently solvable by a nonconvex proximal gradient method with a closed-form proximal operator. We have demonstrated the effectiveness of our approach via experiments on sparse linear regression.

6. APPENDIX

6.1. Error Bound Function for Vector Sparsity Case

We consider the vector sparsity case $\mathcal{A} = \{\mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x}\|_0 \leq k\}$ as an example. Let $\mathbf{x} \in \mathbb{R}^n$ be given. Without loss of generality, assume that $|x_1| \geq |x_2| \geq \dots \geq |x_n|$. It is easy to verify that $\mathbf{y}^* = (x_1, \dots, x_k, 0, \dots, 0)$ is a solution to $\min_{\mathbf{y} \in \mathcal{A}} \|\mathbf{x} - \mathbf{y}\|_2$. It follows that

$$\begin{aligned} \text{dist}(\mathbf{x}, \mathcal{A}) &= \|\mathbf{x} - \mathbf{y}^*\|_2 \leq \|\mathbf{x} - \mathbf{y}^*\|_1 \\ &= \sum_{i=k+1}^n |x_i| = \|\mathbf{x}\|_1 - \|\mathbf{x}\|_{[k]}. \end{aligned} \quad (9)$$

6.2. Proof of Proposition 2

The equivalence of global optimal solution sets is supported by Proposition 1. To show the equivalence of local optimal solution sets, we first introduce the below lemma:

Lemma 2. (Proposition 9.1.2 in [5]) *Consider the same setting in Proposition 1:*

(a) *Given any scalar $\lambda > K$, we have*

$$\tilde{\mathbf{x}} \in \underset{\mathbf{x} \in \mathcal{A}}{\text{arglocmin}} f(\mathbf{x}) \implies \tilde{\mathbf{x}} \in \underset{\mathbf{x} \in \mathcal{C}}{\text{arglocmin}} f(\mathbf{x}) + \lambda\psi(\mathbf{x}).$$

(b) *If $\tilde{\mathbf{x}}$ is a point in \mathcal{A} , then we have the following implication:*

$$\tilde{\mathbf{x}} \in \underset{\mathbf{x} \in \mathcal{C}}{\text{arglocmin}} f(\mathbf{x}) + \lambda\psi(\mathbf{x}) \implies \tilde{\mathbf{x}} \in \underset{\mathbf{x} \in \mathcal{A}}{\text{arglocmin}} f(\mathbf{x}).$$

The above local optimality result is limited, as it assumes that if a solution to the penalized problem is feasible for the original constraint, then it is locally optimal for the original problem. This does not guarantee that a local solution returned by an algorithm is also locally optimal for the original problem. For sparsity constraints, we show that this assumption can be removed, yielding a stronger equivalence of local optima between the two formulations. Due to space constraints, we present the result for vector sparsity; the proofs for the other sparsity forms are essentially the same.

Based on Lemma 2, it suffices to prove that the local optimal points of (2) lie exactly in the feasible region of (3). Define $F_\lambda(\mathbf{x}) = f(\mathbf{x}) + \lambda h(\mathbf{x})$, where $h(\mathbf{x}) = \|\mathbf{x}\|_1 - \|\mathbf{x}\|_{[k]}$. Let $\tilde{\mathbf{x}}$ be a local minimum of F_λ . By definition, there exists a constant $\varepsilon > 0$ such that for any $\mathbf{x} \in \mathcal{N} := \{\mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x} - \tilde{\mathbf{x}}\|_2 \leq \varepsilon\}$, we have $F_\lambda(\tilde{\mathbf{x}}) \leq F_\lambda(\mathbf{x})$.

Suppose that $\tilde{\mathbf{x}} \notin \mathcal{A}$, which means that $\|\tilde{\mathbf{x}}\|_0 > k$. Without loss of generality, assume that $|\tilde{x}_1| \geq \dots \geq |\tilde{x}_n|$. We note that $\tilde{x}_{k+1} \neq 0$ due to $\|\tilde{\mathbf{x}}\|_0 > k$. Define

$$x_i = \begin{cases} \tilde{x}_i, & \text{if } i \neq k+1, \\ \text{sgn}(\tilde{x}_{k+1}) \cdot \max(0, |\tilde{x}_{k+1}| - \varepsilon), & \text{if } i = k+1. \end{cases}$$

It can be verified that $\|\tilde{\mathbf{x}} - \mathbf{x}\|_2 \leq \varepsilon$. Also,

$$\begin{aligned} h(\tilde{\mathbf{x}}) - h(\mathbf{x}) &= \sum_{i=k+1}^n |\tilde{x}_i| - \sum_{i=k+1}^n |x_i| = |\tilde{x}_{k+1}| - |x_{k+1}| \\ &= |\tilde{x}_{k+1} - x_{k+1}| = \|\tilde{\mathbf{x}} - \mathbf{x}\|_2. \end{aligned}$$

It follows that for $\lambda > K$,

$$\begin{aligned} F_\lambda(\tilde{\mathbf{x}}) - F_\lambda(\mathbf{x}) &= f(\tilde{\mathbf{x}}) - f(\mathbf{x}) + \lambda(h(\tilde{\mathbf{x}}) - h(\mathbf{x})) \\ &\geq -K\|\mathbf{x} - \tilde{\mathbf{x}}\|_2 + \lambda\|\tilde{\mathbf{x}} - \mathbf{x}\|_2 > 0, \end{aligned}$$

which contradicts the assumption that $\tilde{\mathbf{x}}$ is a locally optimal solution to (2). This implies that a locally optimal solution $\tilde{\mathbf{x}}$ to (2) must lie in \mathcal{A} , thus satisfying $\|\tilde{\mathbf{x}}\|_0 \leq k$.

6.3. Proof of Lemma 2

Since $\mathbf{x} \mapsto g(\mathbf{x}) = \sqrt{f(\mathbf{x}) + \varepsilon}$ is differentiable, by the Mean Value Theorem, in order to establish the Lipschitz continuity of g on \mathbb{R}^n , it suffices to show that $\|\nabla g(\mathbf{x})\|_2$ is bounded for all $\mathbf{x} \in \mathbb{R}^n$. For any $\mathbf{x} \in \mathbb{R}^n$, we have

$$\|g(\mathbf{x})\|_2 = \frac{\|f(\mathbf{x})\|_2}{\sqrt{f(\mathbf{x}) + \varepsilon}} \leq \frac{\sqrt{Cf(\mathbf{x})}}{\sqrt{f(\mathbf{x}) + \varepsilon}} < \sqrt{C}$$

for all $\mathbf{x} \in \mathbb{R}^n$. Hence, the function g is Lipschitz-continuous on \mathbb{R}^n .

6.4. Proof of Proposition 3

Problem (7) can be rewritten as

$$\min_{\mathbf{x}} \left\{ \frac{1}{2} \|\mathbf{x} - \mathbf{z}\|_2^2 + \mu \sum_{i \notin \mathcal{S}_x} |x_i| \right\}, \quad (10)$$

where \mathcal{S}_x is the index set for the k absolutely largest elements of \mathbf{x} . Here, Problem (10) is equivalent to minimizing over the vector \mathbf{x} and index set \mathcal{S} with fixed cardinality k at the same time:

$$\min_{\mathbf{x}, \mathcal{S}: |\mathcal{S}|=k} \left\{ \frac{1}{2} \|\mathbf{x} - \mathbf{z}\|_2^2 + \mu \sum_{i \notin \mathcal{S}} |x_i| \right\}. \quad (11)$$

Problem (11) is equivalent to minimizing over $\mathbf{x} \in \mathbb{R}^n$ for a fixed set \mathcal{S} , where the inner minimization problem over \mathbf{x} can be solved in an elementwise manner. For $i \in \mathcal{S}$, we need to minimize $\frac{1}{2}(x_i - z_i)^2$, which yields the solution $x_i = z_i$. For $i \notin \mathcal{S}$, it suffices to minimize $\frac{1}{2}(x_i - z_i)^2 + \lambda|x_i|$, which leads to $x_i = \text{sign}(z_i) \cdot \max(|z_i| - \mu, 0)$. The optimal value of the inner minimization problem is $\sum_{i \notin \mathcal{S}} \phi(z_i)$, where

$$\phi(t) = \begin{cases} \frac{1}{2}t^2, & \text{if } |t| \leq \mu, \\ \mu|t| - \frac{1}{2}\mu^2, & \text{if } |t| > \mu. \end{cases}$$

Notice that $t \mapsto \phi(|t|)$ is strictly increasing in $|t|$. Therefore, to solve $\min_{\mathcal{S}: |\mathcal{S}|=k} \sum_{i \notin \mathcal{S}} \phi(z_i)$, it can be verified that \mathcal{S} is exactly the indices of the k largest $|z_i|$. This leads to the optimal solution expressed in (8).

7. REFERENCES

- [1] Emmanuel J Candès, Justin Romberg, and Terence Tao, “Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information,” *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, 2006.
- [2] Wei Liu, Martin Haardt, Maria S Greco, Christoph F Mecklenbräuker, and Peter Willett, “Twenty-five years of sensor array and multichannel signal processing: A review of progress to date and potential research directions,” *IEEE Signal Process. Mag.*, vol. 40, no. 4, pp. 80–91, 2023.
- [3] Emmanuel Candes and Benjamin Recht, “Exact matrix completion via convex optimization,” *Commun. ACM*, vol. 55, no. 6, pp. 111–119, 2012.
- [4] Beidi Chen, Tri Dao, Eric Winsor, Zhao Song, Atri Rudra, and Christopher Ré, “Scatterbrain: Unifying sparse and low-rank attention,” *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 17413–17426, 2021.
- [5] Ying Cui and Jong-Shi Pang, *Modern Nonconvex Nondifferentiable Optimization*, SIAM, Philadelphia, PA, USA, 2022.
- [6] Pengyu Chen, Xu Shi, Rujun Jiang, and Jiulin Wang, “Penalty-based methods for simple bilevel optimization under Hölderian error bounds,” *Adv. Neural Inf. Process. Syst.*, vol. 37, pp. 140731–140765, 2024.
- [7] Han Shen, Quan Xiao, and Tianyi Chen, “On penalty-based bilevel gradient descent method,” *Math. Program.*, pp. 1–51, 2025.
- [8] Zhi-Quan Luo and Paul Tseng, “Error bounds and convergence analysis of feasible descent methods: A general approach,” *Ann. Oper. Res.*, vol. 46, no. 1, pp. 157–178, 1993.
- [9] Zirui Zhou and Anthony Man-Cho So, “A unified approach to error bounds for structured convex optimization problems,” *Math. Program.*, vol. 165, no. 2, pp. 689–728, 2017.
- [10] Man-Chung Yue, Zirui Zhou, and Anthony Man-Cho So, “On the quadratic convergence of the cubic regularization method under a local error bound condition,” *SIAM J. Optim.*, vol. 29, no. 1, pp. 904–932, 2019.
- [11] Junbin Liu, Ya Liu, Wing-Kin Ma, Mingjie Shao, and Anthony Man-Cho So, “Extreme point pursuit—Part I: A framework for constant modulus optimization,” *IEEE Trans. Signal Process.*, 2024.
- [12] Junbin Liu, Ya Liu, Wing-Kin Ma, Mingjie Shao, and Anthony Man-Cho So, “Extreme point pursuit—Part II: Further error bound analysis and applications,” *IEEE Trans. Signal Process.*, 2024.
- [13] Xiaojun Chen, Yifan He, and Zaikun Zhang, “Tight error bounds for the sign-constrained Stiefel manifold,” *SIAM J. Optim.*, vol. 35, no. 1, pp. 302–329, 2025.
- [14] Balas Kausik Natarajan, “Sparse approximate solutions to linear systems,” *SIAM J. Comput.*, vol. 24, no. 2, pp. 227–234, 1995.
- [15] Geoff Davis, Stephane Mallat, and Marco Avellaneda, “Adaptive greedy approximations,” *Constr. Approx.*, vol. 13, no. 1, pp. 57–98, 1997.
- [16] Xiaojun Chen, Dongdong Ge, Zizhuo Wang, and Yinyu Ye, “Complexity of unconstrained minimization,” *Math. Program.*, vol. 143, no. 1, pp. 371–383, 2014.
- [17] Stéphane G Mallat and Zhifeng Zhang, “Matching pursuits with time-frequency dictionaries,” *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, 1993.
- [18] Dimitris Bertsimas, Angela King, and Rahul Mazumder, “Best subset selection via a modern optimization lens,” *Ann. Stat.*, vol. 44, no. 2, pp. 813–852, 2016.
- [19] Robert Tibshirani, “Regression shrinkage and selection via the LASSO,” *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 58, no. 1, pp. 267–288, 1996.
- [20] Jianqing Fan and Runze Li, “Variable selection via nonconcave penalized likelihood and its oracle properties,” *J. Am. Stat. Assoc.*, vol. 96, no. 456, pp. 1348–1360, 2001.
- [21] Cun-Hui Zhang, “Nearly unbiased variable selection under minimax concave penalty,” *Ann. Stat.*, vol. 38, no. 2, pp. 894–942, 2010.
- [22] Wenjiang J Fu, “Penalized regressions: The bridge versus the LASSO,” *J. Comput. Graph. Stat.*, vol. 7, no. 3, pp. 397–416, 1998.
- [23] Jason Weston, André Elisseeff, Bernhard Schölkopf, and Mike Tipping, “Use of the zero-norm with linear models and kernel methods,” *J. Mach. Learn. Res.*, vol. 3, no. Mar, pp. 1439–1461, 2003.
- [24] Jun-ya Gotoh, Akiko Takeda, and Katsuya Tono, “DC formulations and algorithms for sparse optimization problems,” *Math. Program.*, vol. 169, no. 1, pp. 141–176, 2018.
- [25] Rémi Gribonval and Morten Nielsen, “Sparse representations in unions of bases,” *IEEE Trans. Inf. Theory*, vol. 49, no. 12, pp. 3320–3325, 2004.
- [26] Emmanuel J Candès, Justin K Romberg, and Terence Tao, “Stable signal recovery from incomplete and inaccurate measurements,” *Commun. Pure Appl. Math.*, vol. 59, no. 8, pp. 1207–1223, 2006.
- [27] Francesco Rinaldi, Fabio Schoen, and Marco Sciandrone, “Concave programming for minimizing the zero-norm over polyhedral sets,” *Comput. Optim. Appl.*, vol. 46, no. 3, pp. 467–486, 2010.
- [28] Paul S Bradley, Olvi L Mangasarian, and J Ben Rosen, “Parimonious least norm approximation,” *Comput. Optim. Appl.*, vol. 11, no. 1, pp. 5–21, 1998.
- [29] Amir Beck, *First-Order Methods in Optimization*, SIAM, 2017.
- [30] Hedy Attouch, Jérôme Bolte, and Benar Fux Svaiter, “Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods,” *Math. Program.*, vol. 137, no. 1, pp. 91–129, 2013.
- [31] Mingjie Shao and Wing-Kin Ma, “Binary MIMO detection via homotopy optimization and its deep adaptation,” *IEEE Trans. Signal Process.*, vol. 69, pp. 781–796, 2020.
- [32] Mikhail Zaslavskiy, Francis Bach, and Jean-Philippe Vert, “A path following algorithm for the graph matching problem,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 12, pp. 2227–2242, 2008.
- [33] Thomas Blumensath, Mehrdad Yaghoobi, and Mike E Davies, “Iterative hard thresholding and l0 regularization,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2007, vol. 3, pp. 877–880.