# Least Sparsity of $p$-norm based Optimization Problems with $p > 1$

Jinglai Shen* and Seyedahmad Mousavi

Original version: July, 2017; Revision: February, 2018

## Abstract

Motivated by $\ell_p$-optimization arising from sparse optimization, high dimensional data analytics and statistics, this paper studies sparse properties of a wide range of $p$-norm based optimization problems with $p > 1$, including generalized basis pursuit, basis pursuit denoising, ridge regression, and elastic net. It is well known that when $p > 1$, these optimization problems lead to less sparse solutions. However, the quantitative characterization of the adverse sparse properties is not available. In this paper, by exploiting optimization and matrix analysis techniques, we give a systematic treatment of a broad class of $p$-norm based optimization problems for a general $p > 1$ and show that optimal solutions to these problems attain full support, and thus have the least sparsity, for almost all measurement matrices and measurement vectors. Comparison to $\ell_p$-optimization with $0 < p \leq 1$ and implications to robustness as well as extensions to the complex setting are also given. These results shed light on analysis and computation of general $p$-norm based optimization problems in various applications.

## 1 Introduction

Sparse optimization arises from various important applications of contemporary interest, e.g., compressed sensing, high dimensional data analytics and statistics, machine learning, and signal and image processing [7, 14, 17, 27]. The goal of sparse optimization is to recover the sparsest vector from observed data which are possibly subject to noise or errors, and it can be formulated as the $\ell_0$-optimization problem [2, 5]. Since the $\ell_0$-optimization problem is NP-hard, it is a folklore in sparse optimization to use the $p$-norm or $p$-quasi-norm $\|\cdot\|_p$ with $p \in (0,1]$ to approximate the $\ell_0$-norm to recover sparse signals [12, 13]. Representative optimization problems involving such the $p$-norm include basis pursuit, basis pursuit denoising, LASSO, and elastic net; see Section 2 for the details of these problems. In particular, when $p = 1$, it gives rise to a convex $\ell_1$-optimization problem which attains efficient numerical algorithms [14, 30]; when $0 < p < 1$, it yields a non-convex and non-Lipschitz optimization problem whose local optimal solutions can be effectively computed [9, 15, 16, 19]. It is worth mentioning that despite numerical challenges, $\ell_p$-minimization with $0 < p < 1$ often leads to improved and more stable recovery results, even under measurement noise and errors [24, 25, 29].

When $p > 1$, it is well known that the $p$-norm formulation will not lead to sparse solutions [4, 14]; see Figure 1 for illustration and comparison with $\ell_p$ minimization with $0 < p \leq 1$. However, to the best of our knowledge, a formal justification of this fact for a general setting with an arbitrary $p > 1$ is not available, except an intuitive and straightforward geometric interpretation for special cases, e.g., basis pursuit; see [31] for a certain adverse sparse property for $p$-norm based ridge regression with $p > 1$ from an algorithmic perspective. Besides, when different norms are used in objective functions of optimization problems, e.g., the ridge regression and elastic net, it is difficult to obtain a simple geometric interpretation. Moreover, for an arbitrary $p > 1$, there lacks a *quantitative* characterization of how less

---

*Department of Mathematics and Statistics, University of Maryland Baltimore County, Baltimore, MD 21250, U.S.A. Emails: shenj@umbc.edu and smousav1@umbc.edu.
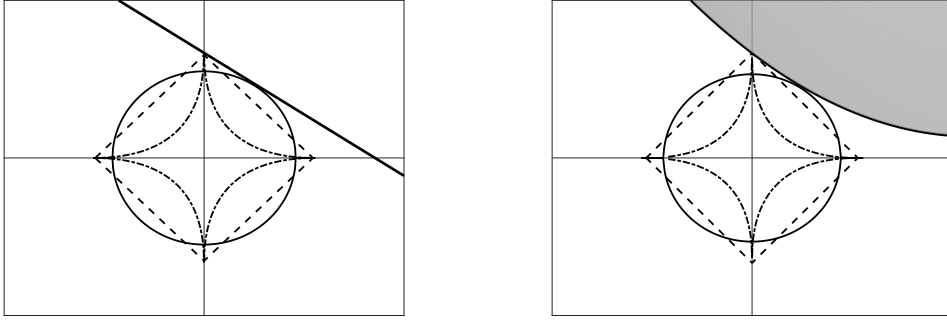
Figure 1: Left: basis pursuit, thick solid line: $\{x \mid Ax = y\}$; Right: basis pursuit denoising, shaded area: $\{x \mid \|Ax - y\|_2 \leq \varepsilon\}$. Solid line: $\ell_2$ ball; dashed line: $\ell_1$ ball; dash-dot line: $\ell_{1/2}$ ball.

sparse such solutions are and how these less sparse solutions depend on a measurement matrix and a measurement vector, in comparison with the related problems for $0 < p \leq 1$. In addition to theoretical interest, these questions are also of practical values, since the $p$-norm based optimization with $p > 1$ and its matrix norm extensions find applications in graph optimization [10], machine learning, and image processing [20]. It is also related to the $\ell_p$-programming with $p > 1$ coined by Terlaky [26]. Motivated by the aforementioned questions and their implications in applications, we give a formal argument for a broad class of $p$-norm based optimization problems with $p > 1$ generalized from sparse optimization and other fields. When $p > 1$, our results show that these problems not only fail to achieve sparse solutions but also yield the *least* sparse solutions generically. Specifically, when $p > 1$, for almost all measurement matrices $A \in \mathbb{R}^{m \times N}$ and measurement vectors $y \in \mathbb{R}^m$, solutions to these $p$-norm based optimization problems have full support, i.e., the support size is $N$; see Theorems 4.1, 4.2, 4.3, 4.4, and 4.5 for formal statements. The proofs for these results turn out to be nontrivial, since except $p = 2$, the optimality conditions of these optimization problems yield highly nonlinear equations and there are no closed form expressions of optimal solutions in terms of $A$ and $y$. To overcome these technical difficulties, we exploit techniques from optimization and matrix analysis and give a systematic treatment to a broad class of $p$-norm based optimization problems originally from sparse optimization and other related fields, including generalized basis pursuit, basis pursuit denoising, ridge regression, and elastic net. The results developed in this paper will also deepen the understanding of general $p$-norm based optimization problems emerging from many applications and shed light on their computation and numerical analysis.

The rest of the paper is organized as follows. In Section 2, we introduce generalized $p$-norm based optimization problems and show the solution existence and uniqueness. When $p > 1$, a lower sparsity bound and other preliminary results are established in Section 3. Section 4 develops the main results of the paper, namely, the least sparsity of $p$-norm optimization based generalized basis pursuit, generalized ridge regression and elastic net, and generalized basis pursuit denoising for $p > 1$. In Section 5, we extend the least sparsity results to measurement vectors restricted to a subspace of the range of $A$, possibly subject to noise, and compare this result with $\ell_p$-optimization for $0 < p \leq 1$ arising from compressed sensing; extensions to the complex setting are also given. Finally, conclusions are made in Section 6.

*Notation.* Let $A = [a_1, \ldots, a_N]$ be an $m \times N$ real matrix with $N > m$, where $a_i \in \mathbb{R}^m$ denotes the $i$th column of $A$. For a given vector $x \in \mathbb{R}^n$, $\text{supp}(x)$ denotes the support of $x$. For any index set $\mathcal{I} \subseteq \{1, \ldots, N\}$, let $|\mathcal{I}|$ denote the cardinality of $\mathcal{I}$, and $A_{\bullet \mathcal{I}} = [a_i]_{i \in \mathcal{I}}$ be the submatrix of $A$ formed by the columns of $A$ indexed by elements of $\mathcal{I}$. For a given matrix $M$, $R(M)$ and $N(M)$ denote the range and null space of $M$ respectively. Let $\text{sgn}(\cdot)$ denote the signum function with $\text{sgn}(0) := 0$. Let $\succeq$ denote the positive semi-definite order, i.e., for two real symmetric matrices $P$ and $Q$, $P \succeq Q$ means that $(P - Q)$ is positive semi-definite. The gradient of a real-valued differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is given by $\nabla f(x) = \left( \frac{\partial f(x)}{\partial x_1}, \ldots, \frac{\partial f(x)}{\partial x_n} \right)^T \in \mathbb{R}^n$. Let $F : \mathbb{R}^n \times \mathbb{R}^r \to \mathbb{R}^s$ be a differentiable function given

by $F(x, z) = (F_1(x, z), \ldots, F_s(x, z))^T$ with $F_i : \mathbb{R}^n \times \mathbb{R}^r \to \mathbb{R}$ for $i = 1, \ldots, s$. The Jacobian of $F$ with respect to $x = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$ is

$$\mathbf{J}_x F(x, z) = \begin{bmatrix} \frac{\partial F_1(x,z)}{\partial x_1} & \cdots & \cdots & \frac{\partial F_1(x,z)}{\partial x_n} \\ \frac{\partial F_2(x,z)}{\partial x_1} & \cdots & \cdots & \frac{\partial F_2(x,z)}{\partial x_n} \\ \vdots & & & \vdots \\ \frac{\partial F_s(x,z)}{\partial x_1} & \cdots & \cdots & \frac{\partial F_s(x,z)}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{s \times n}.$$

By convention, we also use $\nabla_x F(x, z)$ to denote $\mathbf{J}_x F(x, z)$. Furthermore, by saying that a statement (P) holds for almost all $x$ in a finite dimensional real vector space $E$, we mean that (P) holds on a set $W \subseteq E$ whose complement $W^c$ has zero Lebesgue measure. Lastly, for two vectors $u, v \in \mathbb{R}^q$, $u \perp v$ denotes the orthogonality of $u$ and $v$, i.e., $u^T v = 0$.

## 2 Generalized $p$-norm based Optimization Problems

In this section, we introduce a broad class of widely studied $p$-norm based optimization problems emerging from sparse optimization, statistics and other fields, and we discuss their generalizations. Throughout this section, we let the constant $p > 0$, the matrix $A \in \mathbb{R}^{m \times N}$ and the vector $y \in \mathbb{R}^m$. For any $p > 0$ and $x = (x_1, \ldots, x_N)^T \in \mathbb{R}^N$, define $\|x\|_p := \left( \sum_{i=1}^N |x_i|^p \right)^{1/p}$.

• **Generalized Basis Pursuit**. Consider the following linear equality constrained optimization problem whose objective function is given by the $p$-norm (or quasi-norm):

$$\text{BP}_p : \quad \min_{x \in \mathbb{R}^N} \|x\|_p \quad \text{subject to} \quad Ax = y, \tag{1}$$

where $y \in R(A)$. Geometrically, this problem seeks to minimize the $p$-norm distance from the origin to the affine set defined by $Ax = y$. When $p = 1$, it becomes the standard basis pursuit [6, 8, 14].

• **Generalized Basis Pursuit Denoising**. Consider the following constrained optimization problem which incorporates noisy signals:

$$\text{BPDN}_p : \quad \min_{x \in \mathbb{R}^N} \|x\|_p \quad \text{subject to} \quad \|Ax - y\|_2 \leq \varepsilon, \tag{2}$$

where $\varepsilon > 0$ characterizes the bound of noise or errors. When $p = 1$, it becomes the standard basis pursuit denoising (or quadratically constrained basis pursuit) [4, 14, 28]. Another version of the generalized basis pursuit denoising is given by the following optimization problem:

$$\min_{x \in \mathbb{R}^N} \|Ax - y\|_2 \quad \text{subject to} \quad \|x\|_p \leq \eta, \tag{3}$$

where the bound $\eta > 0$. Similarly, when $p = 1$, the optimization problem (3) pertains to a relevant formulation of basis pursuit denoising [14, 28].

• **Generalized Ridge Regression and Elastic Net**. Consider the following unconstrained optimization problem:

$$\text{RR}_p : \quad \min_{x \in \mathbb{R}^N} \frac{1}{2} \|Ax - y\|_2^2 + \lambda \|x\|_p^p, \tag{4}$$

where $\lambda > 0$ is the penalty parameter. When $p = 2$, it becomes the standard ridge regression extensively studied in statistics [17, 18]; when $p = 1$, it yields the least absolute shrinkage and selection operator (LASSO) with the $\ell_1$-norm penalty [27]. The $\text{RR}_p$ (4) is closely related to the maximum a posteriori (MAP) estimator when the prior takes the generalized normal distribution. A related optimization problem is the generalized elastic net arising from statistics:

$$\text{EN}_p : \quad \min_{x \in \mathbb{R}^N} \frac{1}{2} \|Ax - y\|_2^2 + \lambda_1 \|x\|_p^r + \lambda_2 \|x\|_2^2, \tag{5}$$

where $r > 0$ and $\lambda_1, \lambda_2$ are positive penalty parameters. When $p = r = 1$, the $\mathrm{EN}_p$ (5) becomes the standard elastic net formulation which combines the $\ell_1$ and $\ell_2$ penalties in regression [32]. Moreover, if we allow $\lambda_2$ to be non-negative, then the $\mathrm{RR}_p$ (4) can be treated as a special case of the $\mathrm{EN}_p$ (5) with $r = p$, $\lambda = \lambda_1 > 0$, and $\lambda_2 = 0$.

In the sequel, we show the existence and uniqueness of optimal solutions for the generalized optimization problems introduced above.

**Proposition 2.1.** *Fix an arbitrary $p > 0$. Each of the optimization problems (1), (2), (3), (4), and (5) attains an optimal solution for any given $A$, $y$, $\varepsilon > 0$, $\eta > 0$, $\lambda > 0$, $r > 0$, $\lambda_1 > 0$ and $\lambda_2 \geq 0$ as long as the associated constraint sets are nonempty. Further, when $p > 1$, each of (1), (2), and (4) has a unique optimal solution. Besides, when $p \geq 1$, $r \geq 1$, $\lambda_1 > 0$, and $\lambda_2 > 0$, (5) has a unique optimal solution.*

*Proof.* For any $p > 0$, the optimization problems (1), (2), (4), and (5) attain optimal solutions since their objective functions are continuous and coercive and the constraint sets (if nonempty) are closed. The optimization problem (3) also attains a solution because it has a continuous objective function and a compact constraint set.

When $p \geq 1$, (1) and (2) are convex optimization problems, and they are equivalent to $\min_{Ax=y} \|x\|_p^p$ and $\min_{\|Ax-y\|_2 \leq \varepsilon} \|x\|_p^p$ respectively. Note that when $p > 1$, the function $\| \cdot \|_p^p$ is strictly convex on $\mathbb{R}^N$; see the proof in Appendix (c.f. Section 7). Therefore, each of (1), (2) and (4) has a unique optimal solution. When $p \geq 1$, $r \geq 1$, $\lambda_1 > 0$, and $\lambda_2 > 0$, the generalized elastic net (5) is a convex optimization problem with a strictly convex objective function and thus has a unique optimal solution. $\qquad\square$

# 3 Preliminary Results on Sparsity of $p$-norm based Optimization with $p > 1$

This section develops key preliminary results for the global sparsity analysis of $p$-norm based optimization problems when $p > 1$.

## 3.1 Lower Bound on Sparsity of $p$-norm based Optimization with $p > 1$

We first establish a lower bound on the sparsity of optimal solutions arising from the $p$-norm based optimization with $p > 1$. Specifically, we show that when $p > 1$, for almost all $(A, y) \in \mathbb{R}^{m \times N} \times \mathbb{R}^m$, any (nonzero) optimal solution has at least $(N - m + 1)$ nonzero elements and thus is far from sparse when $N \gg m$. This result is critical to show in the subsequent section that for almost all $(A, y)$, an optimal solution achieves a full support; see the proofs of Propositions 4.1 and 4.3, and Theorem 4.2. Toward this end, we define the following set in $\mathbb{R}^{m \times N} \times \mathbb{R}^m$ with $N \geq m$:

$$S := \left\{ (A, y) \in \mathbb{R}^{m \times N} \times \mathbb{R}^m \mid \text{every } m \times m \text{ submatrix of } A \text{ is invertible, and } y \neq 0 \right\}. \quad (6)$$

Clearly, $S$ is open and its complement $S^c$ has zero measure in $\mathbb{R}^{m \times N} \times \mathbb{R}^m$. Note that a matrix $A$ satisfying the condition in (6) is called to be of completely full rank [19]. To emphasize the dependence of optimal solutions on the measurement matrix $A$ and the measurement vector $y$, we write an optimal solution as $x^*_{(A,y)}$ or $x^*(A, y)$ below; the latter notation is used when $x^*$ is unique for any given $(A, y)$ so that $x^*$ is a function of $(A, y)$.

**Proposition 3.1.** *Let $p > 1$. For any $(A, y) \in S$, the following statements hold:*

(i) *The optimal solution $x^*_{(A,y)}$ to the $\mathrm{BP}_p$ (1) satisfies $|supp(x^*_{(A,y)})| \geq N - m + 1$;*

(ii) *If $0 < \varepsilon < \|y\|_2$, then the optimal solution $x^*_{(A,y)}$ to the $\mathrm{BPDN}_p$ (2) satisfies $|supp(x^*_{(A,y)})| \geq N - m + 1$;*

*(iii) For any $\lambda > 0$, the optimal solution $x^*_{(A,y)}$ to the $RR_p$ (4) satisfies $|supp(x^*_{(A,y)})| \geq N - m + 1$;*

*(iv) For any $r > 0$, $\lambda_1 > 0$ and $\lambda_2 \geq 0$, each nonzero optimal solution $x^*_{(A,y)}$ to the $EN_p$ (5) satisfies $|supp(x^*_{(A,y)})| \geq N - m + 1$.*

We give two remarks on the conditions stated in the proposition before presenting its proof:

(a) Note that if $\varepsilon \geq \|y\|_2$ in statement (ii), then $x = 0$ is feasible such that the $BPDN_p$ (2) attains the trivial (unique) optimal solution $x^* = 0$. For this reason, we impose the assumption $0 < \varepsilon < \|y\|_2$.

(b) When $0 < r < 1$ in statement (iv) with $\lambda_1 > 0$ and $\lambda_2 \geq 0$, the $EN_p$ (5) has a non-convex objective function and it may have multiple optimal solutions. Statement (iv) says that any of such nonzero optimal solutions has the sparsity of at least $N - m + 1$.

*Proof.* Fix $(A, y) \in S$. We write an optimal solution $x^*_{(A,y)}$ as $x^*$ for notational simplicity in the proof. Furthermore, let $f(x) := \|x\|_p^p$. Clearly, when $p > 1$, $f$ is continuously differentiable on $\mathbb{R}^N$.

(i) Consider the $BP_p$ (1). Note that $0 \neq y \in R(A)$ for any $(A, y) \in S$. By Proposition 2.1, the $BP_p$ (1) (1) has a unique optimal solution $x^*$ for each $(A, y) \in S$. In view of $x^* = \operatorname{argmin}_{Ax=y} f(x)$, the necessary and sufficient optimality condition for $x^*$ is given by the following KKT condition:

$$\nabla f(x^*) - A^T \nu = 0, \qquad A\,x^* = y,$$

where $\nu \in \mathbb{R}^m$ is the Lagrange multiplier, and $(\nabla f(x))_i = p \cdot \operatorname{sgn}(x_i) \cdot |x_i|^{p-1}$ for each $i = 1, \dots, N$. Note that $\nabla f(x)$ is positively homogeneous in $x$ and each $(\nabla f(x))_i$ depends on $x_i$ only. Suppose that $x^*$ has at least $m$ zero elements. Hence, $\nabla f(x^*)$ has at least $m$ zero elements. By the first equation in the KKT condition, we deduce that there is an $m \times m$ submatrix $A_1$ of $A$ such that $A_1^T \nu = 0$. Since $A_1$ is invertible, we have $\nu = 0$ such that $\nabla f(x^*) = 0$. This further implies that $x^* = 0$. This contradicts $A\,x^* = y \neq 0$. Therefore, $|supp(x^*)| \geq N - m + 1$ for all $(A, y) \in S$.

(ii) Consider the $BPDN_p$ (2). Note that for any given $(A, y) \in S$ and $0 < \varepsilon < \|y\|_2$, the $BPDN_p$ (2) has a unique nonzero optimal solution $x^*$. Let $g(x) := \|Ax - y\|_2^2 - \varepsilon^2$. Since $A$ has full row rank, there exists $\overline{x} \in \mathbb{R}^N$ such that $g(\overline{x}) < 0$. As $g(\cdot)$ is a convex function, the Slater's constraint qualification holds for the equivalent convex optimization problem $\min_{g(x) \leq 0} f(x)$. Hence $x^*$ satisfies the KKT condition with the Lagrange multiplier $\mu \in \mathbb{R}$, where $\perp$ denotes the orthogonality,

$$\nabla f(x^*) + \mu \nabla g(x^*) = 0, \qquad 0 \leq \mu \perp g(x^*) \leq 0.$$

We claim that $\mu > 0$. Suppose not. Then it follows from the first equation in the KKT condition that $\nabla f(x^*) = 0$, which implies $x^* = 0$. This yields $g(x^*) = \|y\|_2^2 - \varepsilon^2 > 0$, contradiction. Therefore $\mu > 0$ such that $g(x^*) = 0$. Using $\nabla g(x^*) = 2A^T(Ax^* - y)$, we have $\nabla f(x^*) + 2\mu A^T(Ax^* - y) = 0$. Suppose, by contradiction, that $x^*$ has at least $m$ zero elements. Without loss of generality, we assume that the first $m$ elements of $x^*$ are zeros. Partition the matrix $A$ into $A = [A_1 \ A_2]$, where $A_1 \in \mathbb{R}^{m \times m}$ and $A_2 \in \mathbb{R}^{m \times (N-m)}$. Similarly, $x^* = [0; \widetilde{x}^*]$, where $\widetilde{x}^* \in \mathbb{R}^{N-m}$. Hence, the first $m$ elements of $\nabla f(x^*)$ are zero. By the first equation in the KKT condition, we derive $2\mu A_1^T(Ax^* - y) = 0$. Since $\mu > 0$ and $A_1$ is invertible, we obtain $Ax^* - y = 0$. This shows that $g(x^*) = -\varepsilon^2 < 0$, contradiction to $g(x^*) = 0$.

(iii) Consider the $RR_p$ (4). The unique optimal solution $x^*$ is characterized by the optimality condition: $A^T(Ax^* - y) + \lambda \nabla f(x^*) = 0$, where $\lambda > 0$. Suppose, by contradiction, that $x^*$ has at least $m$ zero elements. Using the similar argument for Case (ii), we derive that $Ax^* - y = 0$. In view of the optimality condition, we thus have $\nabla f(x^*) = 0$. This implies that $x^* = 0$. Substituting $x^* = 0$ into the optimality condition yields $A^T y = 0$. Since $A$ has full row rank, we obtain $y = 0$. This leads to a contradiction. Hence $|supp(x^*)| \geq N - m + 1$ for all $(A, y) \in S$.

(iv) Consider the $EN_p$ (5) with the exponent $r > 0$ and the penalty parameters $\lambda_1 > 0$ and $\lambda_2 \geq 0$. When $\lambda_2 = 0$, it is closely related to the $RR_p$ with the exponent on $\|x\|_p$ replaced by an arbitrary

$r > 0$. For any $(A, y) \in S$, let $x^*$ be a (possibly non-unique) nonzero optimal solution which satisfies the optimality condition: $A^T(Ax^* - y) + r\lambda_1 \cdot \|x^*\|_p^{r-1} \cdot \nabla\|x^*\|_p + 2\lambda_2 x^* = 0$, where for any nonzero $x \in \mathbb{R}^N$,

$$\nabla\|x\|_p = \frac{1}{\|x\|_p^{p-1}} \left( \text{sgn}(x_1)|x_1|^{p-1}, \ \dots, \ \text{sgn}(x_N)|x_N|^{p-1} \right)^T = \frac{\nabla\|x\|_p^p}{p \cdot \|x\|_p^{p-1}}.$$

The optimality condition can be equivalently written as

$$p \cdot A^T(Ax^* - y) + r\lambda_1 \cdot \|x^*\|_p^{r-p} \cdot \nabla f(x^*) + 2p\lambda_2 \, x^* = 0. \tag{7}$$

Consider two cases: (iv.1) $\lambda_2 = 0$. By the similar argument for Case (iii), it is easy to show that $|\text{supp}(x^*)| \geq N - m + 1$ for all $(A, y) \in S$; (iv.2) $\lambda_2 > 0$. In this case, suppose, by contradiction, that $x^*$ has at least $m$ zero elements. As before, let $A = [A_1 \ A_2]$ and $x^* = [0; \widetilde{x}^*]$ with $A_1 \in \mathbb{R}^{m \times m}$ and $\widetilde{x}^* \in \mathbb{R}^{N-m}$. Hence, the optimality condition leads to $p \cdot A_1^T(Ax^* - y) = 0$. This implies that $Ax^* - y = 0$ such that $r\lambda_1 \cdot \|x^*\|_p^{r-p} \cdot \nabla f(x^*) + 2p\lambda_2 \, x^* = 0$. Since $(\nabla f(x))_i = p \cdot \text{sgn}(x_i) \cdot |x_i|^{p-1}$ for each $i = 1, \dots, N$, we obtain $r\lambda_1 \cdot \|x^*\|_p^{r-p} \cdot |x_i^*|^{p-1} + 2\lambda_2|x_i^*| = 0$ for each $i$. Hence $x^* = 0$, a contradiction. We thus conclude that $|\text{supp}(x^*)| \geq N - m + 1$ for all $(A, y) \in S$. $\qquad\square$

We discuss an extension of the sparsity lower bound developed in Proposition 3.1 to another formulation of the basis pursuit denoising given in (3). It is noted that if $\eta \geq \min_{Ax=y} \|x\|_p$ (which implies $y \in R(A)$), then the optimal value of (3) is zero and can be achieved at some feasible $x^*$ satisfying $Ax^* = y$. Hence any optimal solution $x'$ must satisfy $Ax' = y$ so that the optimal solution set is given by $\{x \in \mathbb{R}^N \,|\, Ax = y, \ \|x\|_p \leq \eta\}$, which is closely related to the BP$_p$ (1). This means that if $\eta \geq \min_{Ax=y} \|x\|_p$, the optimization problem (3) can be converted to a reduced and simpler problem. For this reason, we assume that $0 < \eta < \min_{Ax=y} \|x\|_p$ for (3). The following proposition presents important results under this assumption; these results will be used for the proof of Theorem 4.5.

**Proposition 3.2.** *The following hold for the optimization problem (3) with $p > 1$:*

(i) *If $A$ has full row rank and $0 < \eta < \min_{Ax=y} \|x\|_p$, then (3) attains a unique optimal solution with a unique positive Lagrange multiplier;*

(ii) *For any $(A, y)$ in the set $S$ defined in (6) and $0 < \eta < \min_{Ax=y} \|x\|_p$, the unique optimal solution $x^*_{(A,y)}$ satisfies $|supp(x^*_{(A,y)})| \geq N - m + 1$.*

*Proof.* (i) Let $A$ be of full row rank. Hence, $y \in R(A)$ so that $\eta$ is well defined. Let $x^*$ be an arbitrary optimal solution to (3) with the specified $\eta > 0$. Hence $x^* = \text{argmin}_{f(x) \leq \eta^p} \frac{1}{2}\|Ax - y\|_2^2$, where we recall that $f(x) = \|x\|_p^p$. Clearly, the Slater's constraint qualification holds for the convex optimization problem (3). Therefore, $x^*$ satisfies the following KKT condition:

$$A^T(Ax^* - y) + \mu\nabla f(x^*) = 0, \qquad 0 \leq \mu \perp f(x^*) - \eta^p \leq 0, \tag{8}$$

where $\mu \in \mathbb{R}$ is the Lagrange multiplier. We claim that $\mu$ must be positive. Suppose not, i.e., $\mu = 0$. By the first equation in (8), we obtain $A^T(Ax^* - y) = 0$. Since $A$ has full row rank, we have $Ax^* = y$. Based on the assumption on $\eta$, we further have $\|x^*\|_p > \eta$, contradiction to $f(x^*) \leq \eta^p$. This proves the claim. Since $\mu > 0$, it follows from the second equation in (8) that any optimal solution $x^*$ satisfies $f(x^*) = \eta^p$ or equivalently $\|x^*\|_p = \eta$. To prove the uniqueness of optimal solution, suppose, by contradiction, that $x^*$ and $x'$ are two distinct optimal solutions for the given $(A, y)$. Thus $\|x^*\|_p = \|x'\|_p = \eta$. Since (3) is a convex optimization problem, the optimal solution set is convex so that $\lambda x^* + (1 - \lambda)x'$ is an optimal solution for any $\lambda \in [0, 1]$. Hence, $\|\lambda x^* + (1 - \lambda)x'\|_p = \eta, \forall \lambda \in [0, 1]$. Since $\|\cdot\|_p^p$ is strictly convex when $p > 1$, we have $\eta^p = \|\lambda x^* + (1 - \lambda)x'\|_p^p < \lambda\|x^*\|_p^p + (1 - \lambda)\|x'\|_p^p = \eta^p$ for each $\lambda \in (0, 1)$. This yields a contradiction. We thus conclude that (3) attains a unique optimal solution with $\mu > 0$.

6

(ii) Let $(A, y) \in S$. Clearly, $A$ has full row rank so that (3) has a unique optimal solution $x^*$ with a positive Lagrange multiplier $\mu$. Suppose $x^*$ has at least $m$ zero elements. It follows from the first equation in (8) and the similar argument for Case (iii) of Proposition 3.1 that $Ax^* = y$. In light of the assumption on $\eta$, we have $\|x^*\|_p > \eta$, a contradiction. Therefore $|\text{supp}(x^*)| \geq N - m + 1$ for any $(A, y) \in S$. $\qquad\square$

## 3.2 Technical Result on Measure of the Zero Set of $C^1$-functions

As shown in Proposition 2.1, when $p > 1$, each of the $\text{BP}_p$ (1), $\text{BPDN}_p$ (2), and $\text{RR}_p$ (4) has a unique optimal solution $x^*$ for any given $(A, y)$. Under additional conditions, each of the $\text{EN}_p$ (5) and the optimization problem (3) also attains a unique optimal solution. Hence, for each of these problems, the optimal solution $x^*$ is a function of $(A, y)$, and each component of $x^*$ becomes a real-valued function $x_i^*(A, y)$. Therefore, the global sparsity of $x^*$ can be characterized by the zero set of each $x_i^*(A, y)$. The following technical lemma gives a key result on the measure of the zero set of a real-valued $C^1$-function under a suitable assumption.

**Lemma 3.1.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable (i.e., $C^1$) on an open set $W \subseteq \mathbb{R}^n$ whose complement $W^c$ has zero measure in $\mathbb{R}^n$. Suppose $\nabla f(x) \neq 0$ for any $x \in W$ with $f(x) = 0$. Then the zero set $f^{-1}(\{0\}) := \{x \in \mathbb{R}^n \mid f(x) = 0\}$ has zero measure.*

*Proof.* Consider an arbitrary $x^* \in W$. If $f(x^*) = 0$, then $\nabla f(x^*) \neq 0$. Without loss of generality, we assume that $\frac{\partial f}{\partial x_n}(x^*) \neq 0$. Let $z := (x_1, \ldots, x_{n-1})^T \in \mathbb{R}^{n-1}$. By the implicit function theorem, there exist a neighborhood $U \subset \mathbb{R}^{n-1}$ of $z^* := (x_1^*, \ldots, x_{n-1}^*)^T$, a neighborhood $V \subset \mathbb{R}$ of $x_n^*$, and a unique $C^1$ function $g : U \to V$ such that $f(z, g(z)) = 0$ for all $z \in U$. The set $f^{-1}(\{0\}) \cap (U \times V) = \{(z, g(z)) \mid z \in U\}$ has zero measure in $\mathbb{R}^n$ since it is an $(n-1)$ dimensional manifold in the open set $U \times V \subset \mathbb{R}^n$. Moreover, in view of the continuity of $f$, we deduce that for any $x^* \in W$ with $f(x^*) \neq 0$, there exists an open set $\mathcal{B}(x^*)$ of $x^*$ such that $f(x) \neq 0, \forall x \in \mathcal{B}(x^*)$. Combining these results, it is seen that for any $x \in W$, there exists an open set $\mathcal{B}(x)$ of $x$ such that $f^{-1}(\{0\}) \cap \mathcal{B}(x)$ has zero measure. Clearly, the family of these open sets given by $\{\mathcal{B}(x)\}_{x \in W}$ forms an open cover of $W$. Since $\mathbb{R}^n$ is a topologically separable metric space, so is $W \subset \mathbb{R}^n$ and thus it is a Lindelöf space [22, 23]. Hence, this open cover attains a countable sub-cover $\{\mathcal{B}(x^i)\}_{i \in \mathbb{N}}$ of $W$, where each $x^i \in W$. Since $f^{-1}(\{0\}) \cap \mathcal{B}(x^i)$ has zero measure for each $i \in \mathbb{N}$, the set $W \cap f^{-1}(\{0\})$ has zero measure. Besides, since $f^{-1}(\{0\}) \subseteq W^c \cup (W \cap f^{-1}(\{0\}))$ and both $W^c$ and $W \cap f^{-1}(\{0\})$ have zero measure, we conclude that $f^{-1}(\{0\})$ has zero measure. $\qquad\square$

## 4 Least Sparsity of $p$-norm based Optimization Problems with $p > 1$

In this section, we establish the main results of the paper, namely, when $p > 1$, the $p$-norm based optimization problems yield *least* sparse solutions for almost all $(A, y)$. We introduce more notation to be used through this section. Let $f(x) := \|x\|_p^p$ for $x \in \mathbb{R}^N$, and when $p > 1$, we define for each $z \in \mathbb{R}$,

$$g(z) := p \cdot \text{sgn}(z) \cdot |z|^{p-1}, \qquad h(z) := \text{sgn}(z) \cdot \left|\frac{z}{p}\right|^{\frac{1}{p-1}}, \tag{9}$$

where $\text{sgn}(\cdot)$ denotes the signum function with $\text{sgn}(0) := 0$. Direct calculation shows that (i) when $p > 1$, $g(z) = (|z|^p)', \forall z \in \mathbb{R}$, and $h(z)$ is the inverse function of $g(z)$; (ii) when $p \geq 2$, $g$ is continuously differentiable and $g'(z) = p(p-1) \cdot |z|^{p-2}, \forall z \in \mathbb{R}$; and (iii) when $1 < p \leq 2$, $h$ is continuously differentiable and $h'(z) = |z|^{\frac{2-p}{p-1}}/[(p-1) \cdot p^{1/(p-1)}], \forall z \in \mathbb{R}$. Furthermore, when $p > 1$, $\nabla f(x) = (g(x_1), \ldots, g(x_N))^T$.

The proofs for the least sparsity developed in the rest of the section share the similar methodologies. To facilitate the reading, we give an overview of main ideas of these proofs and comment on certain key steps in the proofs. As indicated at the beginning of Section 3.2, the goal is to show that the zero set of each component of an optimal solution $x^*$, which is a real-valued function of $(A, y)$, has zero measure.

To achieve this goal, we first show using the KKT conditions and the implicit function theorem that $x^*$, possibly along with a Lagrange multiplier if applicable, is a $C^1$ function of $(A, y)$ on a suitable open set $S'$ in $\mathbb{R}^{m \times N} \times \mathbb{R}^m$ whose complement has zero measure. We then show that for each $i = 1, \ldots, N$, if $x_i^*$ is vanishing at $(A, y) \in S'$, then its gradient evaluated at $(A, y)$ is nonzero. In view of Lemma 3.1, this leads to the desired result. Moreover, for each of the generalized optimization problems with $p > 1$, i.e., the $\text{BP}_p$, $\text{BPDN}_p$, $\text{RR}_p$, and $\text{EN}_p$, we divide their proofs into two separate cases: (i) $p \geq 2$; and (ii) $1 < p \leq 2$. This is because each case invokes the derivative of $g(\cdot)$ or its inverse function $h(\cdot)$ defined in (9). When $p \geq 2$, the derivative $g'(\cdot)$ is globally well defined. On the contrary, when $1 < p \leq 2$, $g'(\cdot)$ is not defined at zero. Hence, we use $h(\cdot)$ instead, since $h'(\cdot)$ is globally well defined in this case. The different choice of $g$ or $h$ gives rise to different arguments in the following proofs, and the proofs for $1 < p \leq 2$ are typically more involved.

## 4.1 Least Sparsity of the Generalized Basis Pursuit with $p > 1$

We consider the case where $p \geq 2$ first.

**Proposition 4.1.** *Let $p \geq 2$ and $N \geq m$. For almost all $(A, y) \in \mathbb{R}^{m \times N} \times \mathbb{R}^m$, the unique optimal solution $x_{(A,y)}^*$ to the $\text{BP}_p$ (1) satisfies $|supp(x_{(A,y)}^*)| = N$.*

*Proof.* Recall that for any $(A, y) \in \mathbb{R}^{m \times N} \times \mathbb{R}^m$, the necessary and sufficient optimality condition for $x^*$ is given by the following KKT condition given in Proposition 3.1:

$$\nabla f(x^*) - A^T \nu = 0, \qquad A x^* = y,$$

where $\nu \in \mathbb{R}^m$ is the Lagrange multiplier, and $(\nabla f(x))_i = g(x_i)$ for each $i = 1, \ldots, N$. Here $g$ is defined in (9). When $p = 2$, $x^* = A^T(AA^T)^{-1}y$ for any $(A, y) \in S$. As $x_i^*(A, y) = 0$ yields a polynomial equation whose solution set has zero measure in $\mathbb{R}^{m \times N} \times \mathbb{R}^m$, the desired result is easily established. We consider $p > 2$ as follows, and show that for any $(A^\diamond, y^\diamond)$ in the open set $S$ defined in (6), $x^*(A, y)$ is continuously differentiable at $(A^\diamond, y^\diamond)$ and that each $x_i^*$ with $x_i^*(A^\diamond, y^\diamond) = 0$ has nonzero gradient at $(A^\diamond, y^\diamond)$.

Recall that $x^*$ is unique for any $(A, y)$. Besides, for each $(A, y) \in S$, $A^T$ has full column rank such that $\nu$ is also unique in view of the first equation of the KKT condition. Therefore, $(x^*, \nu)$ is a function of $(A, y) \in S$. For notational simplicity, let $x^\diamond := x^*(A^\diamond, y^\diamond)$ and $\nu^\diamond := \nu(A^\diamond, y^\diamond)$. Define the index set $\mathcal{J} := \{i \,|\, x_i^\diamond \neq 0\}$. By Proposition 3.1, we see that $\mathcal{J}$ is nonempty and $|\mathcal{J}^c| \leq m - 1$. Further, in light of the KKT condition, $(x^*, \nu) \in \mathbb{R}^N \times \mathbb{R}^m$ satisfies the following equation:

$$F(x, \nu, A, y) := \begin{bmatrix} \nabla f(x) - A^T \nu \\ Ax - y \end{bmatrix} = 0.$$

Clearly, $F : \mathbb{R}^N \times \mathbb{R}^m \times \mathbb{R}^{m \times N} \times \mathbb{R}^m \to \mathbb{R}^{N+m}$ is $C^1$, and its Jacobian with respect to $(x, \nu)$ is

$$\mathbf{J}_{(x,\nu)} F(x, \nu, A, y) = \begin{bmatrix} \Lambda(x) & -A^T \\ A & 0 \end{bmatrix},$$

where the diagonal matrix $\Lambda(x) := \text{diag}(g'(x_1), \ldots, g'(x_N))$. Partition $\Lambda^\diamond := \Lambda(x^\diamond)$ and $A$ as $\Lambda^\diamond = \text{diag}(\Lambda_1, \Lambda_2)$ and $A^\diamond = \begin{bmatrix} A_1 & A_2 \end{bmatrix}$ respectively, where $\Lambda_1 := \text{diag}(g'(x_i^\diamond))_{i \in \mathcal{J}^c} = 0$ as $p > 2$, $\Lambda_2 := \text{diag}(g'(x_i^\diamond))_{i \in \mathcal{J}}$ is positive definite, $A_1 := A_{\bullet \mathcal{J}^c}^\diamond$, and $A_2 := A_{\bullet \mathcal{J}}^\diamond$. We claim that the following matrix is invertible:

$$W := \mathbf{J}_{(x,\nu)} F(x^\diamond, \nu^\diamond, A^\diamond, y^\diamond) = \begin{bmatrix} \Lambda_1 & 0 & -A_1^T \\ 0 & \Lambda_2 & -A_2^T \\ A_1 & A_2 & 0 \end{bmatrix} \in \mathbb{R}^{(N+m) \times (N+m)}.$$

In fact, let $z := [u_1; u_2; v] \in \mathbb{R}^{N+m}$ be such that $Wz = 0$. Since $\Lambda_1 = 0$ and $\Lambda_2$ is positive definite, we have $A_1^T v = 0$, $u_2 = \Lambda_2^{-1} A_2^T v$, and $A_1 u_1 + A_2 u_2 = 0$. Therefore, $0 = v^T(A_1 u_1 + A_2 u_2) = v^T A_2 \Lambda_2^{-1} A_2^T v$,

8

which implies that $A_2^T v = 0$ such that $u_2 = 0$ and $A_1 u_1 = 0$. Since $|\mathcal{J}^c| \leq m - 1$ and any $m \times m$ submatrix of $A$ is invertible, the columns of $A_1$ are linearly independent such that $u_1 = 0$. This implies that $A^T v = 0$. Since $A$ has full row rank, we have $v = 0$ and thus $z = 0$. This proves that $W$ is invertible. By the implicit function theorem, there are local $C^1$ functions $G_1, G_2, H$ such that $x^* = (x_{\mathcal{J}^c}^*, x_{\mathcal{J}}^*) = (G_1(A, y), G_2(A, y)) := G(A, y)$, $\nu = H(A, y)$, and $F(G(A, y), H(A, y), A, y) = 0$ for all $(A, y)$ in a neighborhood of $(A^\diamond, y^\diamond)$.

By the chain rule, we have

$$\underbrace{\mathbf{J}_{(x,\nu)} F(x^\diamond, \nu^\diamond, A^\diamond, y^\diamond)}_{=W} \cdot \begin{bmatrix} \nabla_y G_1(A^\diamond, y^\diamond) \\ \nabla_y G_2(A^\diamond, y^\diamond) \\ \nabla_y H(A^\diamond, y^\diamond) \end{bmatrix} + \mathbf{J}_y F(x^\diamond, \nu^\diamond, A^\diamond, y^\diamond) = 0,$$

where

$$\mathbf{J}_y F(x^\diamond, \nu^\diamond, A^\diamond, y^\diamond) = \begin{bmatrix} 0 \\ 0 \\ -I \end{bmatrix}, \qquad W^{-1} := P = \begin{bmatrix} P_{11} & P_{12} & P_{13} \\ P_{21} & P_{22} & P_{23} \\ P_{31} & P_{32} & P_{33} \end{bmatrix}.$$

It is easy to verify that $\nabla_y G_1(A^\diamond, y^\diamond) = P_{13}$ and $P_{13} A_1 = I$ by virtue of $PW = I$. The latter equation shows that each row of $P_{13}$ is nonzero, so is each row of $\nabla_y G_1(A^\diamond, y^\diamond)$. Thus each row of $\nabla_{(A,y)} G_1(A^\diamond, y^\diamond)$ is nonzero. Hence, for each $i = 1, \ldots, N$, $x_i^*(A, y)$ is $C^1$ on the open set $S$, and when $x_i^*(A^\diamond, y^\diamond) = 0$ at $(A^\diamond, y^\diamond) \in S$, its gradient is nonzero. By Lemma 3.1, we see that for each $i = 1, \ldots, N$, the zero set of $x_i^*(A, y)$ has zero measure. This shows that $|\mathrm{supp}(x^*(A, y))| = N$ for almost all $(A, y) \in \mathbb{R}^{m \times N} \times \mathbb{R}^m$. $\square$

The next result addresses the case where $1 < p \leq 2$. In this case, it can be shown that if $x_i^*$ is vanishing at some $(A^\diamond, y^\diamond)$ in a certain open set, then the gradient of $x_i^*$ evaluated at $(A^\diamond, y^\diamond)$ also vanishes. This prevents us from applying Lemma 3.1 directly. To overcome this difficulty, we introduce a suitable function which has exactly the same sign of $x_i^*$ and to which Lemma 3.1 is applicable. This technique is also used in other proofs for $1 < p \leq 2$; see Theorems 4.2, 4.3, 4.5, and Proposition 4.4.

**Proposition 4.2.** *Let* $1 < p \leq 2$ *and* $N \geq 2m - 1$. *For almost all* $(A, y) \in \mathbb{R}^{m \times N} \times \mathbb{R}^m$, *the unique optimal solution* $x_{(A,y)}^*$ *to the BP$_p$ (1) satisfies* $|supp(x_{(A,y)}^*)| = N$.

*Proof.* Let $\widetilde{S}$ be the set of all $(A, y) \in \mathbb{R}^{m \times N} \times \mathbb{R}^m$ satisfying the following conditions: (i) $y \neq 0$; (ii) each column of $A$ is nonzero; and (iii) for any index set $\mathcal{I} \subseteq \{1, \ldots, N\}$ with $|\mathcal{I}^c| \geq m$ and $\mathrm{rank}(A_{\bullet\mathcal{I}}) < m$, $\mathrm{rank}(A_{\bullet\mathcal{I}^c}) = m$. Note that such an $A$ has full row rank, i.e., $\mathrm{rank}(A) = m$. It is easy to see that $\widetilde{S}$ is open and its complement $\widetilde{S}^c$ has zero measure. Note that the set $S$ given in (6) is a proper subset of $\widetilde{S}$.

Let $A = [a_1, \ldots, a_N]$, where $a_i \in \mathbb{R}^m$ is the $i$th column of $A$. It follows from the KKT condition $\nabla f(x^*) - A^T \nu = 0$ that $x_i^* = h(a_i^T \nu)$ for each $i = 1, \ldots, N$, where the function $h$ is defined in (9). Along with the equation $Ax^* = y$, we obtain the following equation for $(\nu, A, y)$:

$$F(\nu, A, y) := \sum_{i=1}^N a_i h(a_i^T \nu) - y = 0,$$

where $F : \mathbb{R}^m \times \mathbb{R}^{m \times N} \times \mathbb{R}^m \to \mathbb{R}^m$ is $C^1$ and its Jacobian with respect to $\nu$ is

$$\mathbf{J}_\nu F(\nu, A, y) = \begin{bmatrix} a_1 & \cdots & a_N \end{bmatrix} \begin{bmatrix} h'(a_1^T \nu) & & & \\ & h'(a_2^T \nu) & & \\ & & \ddots & \\ & & & h'(a_N^T \nu) \end{bmatrix} \begin{bmatrix} a_1^T \\ \vdots \\ a_N^T \end{bmatrix} \in \mathbb{R}^{m \times m}.$$

We show next that for any $(A^\diamond, y^\diamond) \in \widetilde{S}$ with (unique) $\nu$ satisfying $F(\nu, A^\diamond, y^\diamond) = 0$, the Jacobian $Q := \mathbf{J}_\nu F(\nu, A^\diamond, y^\diamond)$ is positive definite. This result is trivial for $p = 2$ since $h'(a_i^T \nu) = 1/2$ for each $i$. To

show it for $1 < p < 2$, we first note that $\nu \neq 0$ since otherwise $x^* = 0$ so that $Ax^* = 0 = y$, contradiction to $y \neq 0$. Using the formula for $h'(\cdot)$ given below (9), we have

$$w^T Q w = \sum_{i=1}^N (a_i^T w)^2 \cdot h'(a_i^T \nu) = \frac{1}{(p-1) \cdot p^{1/(p-1)}} \sum_{i=1}^N \left(a_i^T w\right)^2 \cdot \left|a_i^T \nu\right|^{\frac{2-p}{p-1}}, \qquad \forall\, w \in \mathbb{R}^m. \qquad (10)$$

Clearly, $Q$ is positive semi-definite. Suppose, by contradiction, that there exists $w \neq 0$ such that $w^T Q w = 0$. Define the index set $\mathcal{I} := \{i \mid a_i^T w = 0\}$. Note that $\mathcal{I}$ must be nonempty because otherwise, it follows from (10) that $A^T \nu = 0$, which contradicts $\mathrm{rank}(A) = m$ and $\nu \neq 0$. Similarly, $\mathcal{I}^c$ is nonempty in view of $w \neq 0$. Hence we have $(A_{\bullet\mathcal{I}})^T w = 0$ and $(A_{\bullet\mathcal{I}^c})^T \nu = 0$. Since $\mathcal{I} \cup \mathcal{I}^c = \{1, \ldots, N\}$, $\mathcal{I} \cap \mathcal{I}^c = \emptyset$ and $N \geq 2m-1$, we must have either $|\mathcal{I}| \geq m$ or $|\mathcal{I}^c| \geq m$. Consider the case of $|\mathcal{I}| \geq m$ first. As $(A_{\bullet\mathcal{I}^c})^T \nu = 0$, we see that $\nu$ is orthogonal to $R(A_{\bullet\mathcal{I}^c})$. Since $\nu \in \mathbb{R}^m$ is nonzero, we obtain $\mathrm{rank}(A_{\bullet\mathcal{I}^c}) < m$. Thus it follows from the properties of $A$ that $\mathrm{rank}(A_{\bullet\mathcal{I}}) = m$, but this contradicts $(A_{\bullet\mathcal{I}})^T w = 0$ for the nonzero $w$. Using the similar argument, it can be shown that the case of $|\mathcal{I}^c| \geq m$ also yields a contradiction. Consequently, $Q$ is positive definite. By the implicit function theorem, there exists a local $C^1$ function $H$ such that $\nu = H(A, y)$ and $F(H(A, y), A, y) = 0$ for all $(A, y)$ in a neighborhood of $(A^\diamond, y^\diamond)$. Let $\nu^\diamond := H(A^\diamond, y^\diamond)$. Using the chain rule, we have

$$\underbrace{\mathbf{J}_\nu F(\nu^\diamond, A^\diamond, y^\diamond)}_{:=Q} \cdot \nabla_y H(A^\diamond, y^\diamond) + \mathbf{J}_y F(\nu^\diamond, A^\diamond, y^\diamond) = 0.$$

Since $\mathbf{J}_y F(\nu^\diamond, A^\diamond, y^\diamond) = -I$, we have $\nabla_y H(A^\diamond, y^\diamond) = Q^{-1}$.

Observing that $x_i^* = h(a_i^T \nu)$ for each $i = 1, \ldots, N$, we deduce via the property of the function $h$ in (9) that $\mathrm{sgn}(x_i^*) = \mathrm{sgn}(a_i^T \nu)$ for each $i$. Therefore, in order to show that the zero set of $x_i^*(A, y)$ has zero measure for each $i = 1, \ldots, N$, it suffices to show that the zero set of $a_i^T \nu(A, y)$ has zero measure for each $i$. It follows from the previous development that for any $(A^\diamond, y^\diamond) \in \widetilde{S}$, $\nu = H(A, y)$ for a local $C^1$ function $H$ in a neighborhood of $(A^\diamond, y^\diamond)$. Hence, $\nabla_y(a_i^T \nu)(A^\diamond, y^\diamond) = (a_i^\diamond)^T \cdot \nabla_y H(A^\diamond, y^\diamond) = (a_i^\diamond)^T \cdot Q^{-1}$, where $Q := \mathbf{J}_\nu F(\nu^\diamond, A^\diamond, y^\diamond)$ is invertible. Since each $a_i^\diamond \neq 0$, we have $\nabla_y(a_i^T \nu)(A^\diamond, y^\diamond) \neq 0$ for each $i = 1, \ldots, N$. In light of Lemma 3.1, the zero set of $a_i^T \nu(A, y)$ has zero measure for each $i$. Hence $|\mathrm{supp}(x^*(A, y))| = N$ for almost all $(A, y) \in \mathbb{R}^{m \times N} \times \mathbb{R}^m$. $\qquad \square$

Combining Propositions 4.1 and 4.2, we obtain the following result for the generalized basis pursuit.

**Theorem 4.1.** *Let $p > 1$ and $N \geq 2m - 1$. For almost all $(A, y) \in \mathbb{R}^{m \times N} \times \mathbb{R}^m$, the unique optimal solution $x^*_{(A,y)}$ to the $BP_p$ (1) satisfies $|\mathrm{supp}(x^*_{(A,y)})| = N$.*

Motivated by Theorem 4.1, we present the following corollary for a certain fixed measurement matrix $A$ whereas the measurement vector $y$ varies. This result will be used for Theorem 5.1 in Section 5.

**Corollary 4.1.** *Let $p > 1$ and $N \geq 2m - 1$. Let $A$ be a fixed $m \times N$ matrix such that any of its $m \times m$ submatrix is invertible. For almost all $y \in \mathbb{R}^m$, the unique optimal solution $x^*_y$ to the $BP_p$ (1) satisfies $|\mathrm{supp}(x^*_y)| = N$.*

*Proof.* Consider $p \geq 2$ first. For any $y \in \mathbb{R}^m$, let $x^*(y)$ be the unique optimal solution to the $BP_p$ (1). It follows from the similar argument for Propositions 4.1 that for each $i = 1, \ldots, N$, $x_i^*$ is a $C^1$ function of $y$ on $\mathbb{R}^m \setminus \{0\}$, and that if $x_i^*(y) = 0$ for any $y \neq 0$, then the gradient $\nabla_y x_i^*(y) \neq 0$. By Lemma 3.1, $|\mathrm{supp}(x^*(y))| = N$ for almost all $y \in \mathbb{R}^m$. When $1 < p \leq 2$, we note that the given matrix $A$ satisfies the required conditions on $A$ in the set $\widetilde{S}$ introduced in the proof of Proposition 4.2, since $S$ defined in (6) is a proper subset of $\widetilde{S}$ as indicated at the end of the first paragraph of the proof of Proposition 4.2. Therefore, by the similar argument for Proposition 4.2, we have that for any $y \neq 0$, the gradient $\nabla_y x_i^*(y) \neq 0$. Therefore, the desired result follows. $\qquad \square$

## 4.2 Least Sparsity of the Generalized Ridge Regression and Generalized Elastic Net with $p > 1$

We first establish the least sparsity of the generalized ridge regression in (4) as follows.

**Theorem 4.2.** *Let $p > 1$, $N \geq m$, and $\lambda > 0$. For almost all $(A, y) \in \mathbb{R}^{m \times N} \times \mathbb{R}^m$, the unique optimal solution $x^*_{(A,y)}$ to the $RR_p$ (4) satisfies $|supp(x^*_{(A,y)})| = N$.*

*Proof.* Recall that for any given $(A, y) \in \mathbb{R}^{m \times N} \times \mathbb{R}^m$, the unique optimal solution $x^*$ to the $RR_p$ (4) is described by the optimality condition: $A^T(Ax^* - y) + \lambda \nabla f(x^*) = 0$, where $\lambda > 0$ is a penalty parameter.

(i) $p \geq 2$. The case of $p = 2$ is trivial by using $x^* = (2\lambda I + A^T A)^{-1} A^T y$, and we thus consider $p > 2$ as follows. Define the function $F(x, A, y) := \lambda \nabla f(x) + A^T(Ax - y)$, where $\nabla f(x) = (g(x_1), \ldots, g(x_N))^T$ with $g$ given in (9). Hence, the optimal solution $x^*$, as the function of $(A, y)$, satisfies the equation $F(x^*, A, y) = 0$. Obviously, $F$ is $C^1$ and its Jacobian with respect to $x$ is given by

$$\mathbf{J}_x F(x, A, y) = \lambda D(x) + A^T A,$$

where the diagonal matrix $D(x) := \text{diag}(g'(x_1), \ldots, g'(x_N))$. Since each $g'(x_i) \geq 0$, we see that $\lambda D(x) + A^T A$ is positive semi-definite for all $A$'s and $x$'s.

We show below that for any $(A^\diamond, y^\diamond)$ in the set $S$ defined in (6), the matrix $\mathbf{J}_x F(x^\diamond, A^\diamond, y^\diamond)$ is positive definite, where $x^\diamond := x^*(A^\diamond, y^\diamond)$. For this purpose, define the index set $\mathcal{J} := \{i \mid x_i^\diamond \neq 0\}$. Partition $D^\diamond := D(x^\diamond)$ and $A^\diamond$ as $D^\diamond = \text{diag}(D_1, D_2)$ and $A^\diamond = [A_1 \ A_2]$ respectively, where $D_1 := \text{diag}(g'(x_i^\diamond))_{i \in \mathcal{J}^c} = 0$, $D_2 := \text{diag}(g'(x_i^\diamond))_{i \in \mathcal{J}}$ is positive definite, $A_1 := A^\diamond_{\bullet \mathcal{J}^c}$, and $A_2 := A^\diamond_{\bullet \mathcal{J}}$. It follows from Proposition 3.1 that $|\mathcal{J}^c| \leq m - 1$ such that the columns of $A_1$ are linearly independent. Suppose there exists a vector $z \in \mathbb{R}^N$ such that $z^T[\lambda D^\diamond + (A^\diamond)^T A^\diamond]z = 0$. Let $u := z_{\mathcal{J}^c}$ and $v := z_{\mathcal{J}}$. Since $z^T[\lambda D^\diamond + (A^\diamond)^T A^\diamond]z \geq z^T \lambda D^\diamond z = \lambda v^T D_2 v \geq 0$ and $D_2$ is positive definite, we have $v = 0$. Hence, $z^T[\lambda D^\diamond + (A^\diamond)^T A^\diamond]z \geq z^T(A^\diamond)^T A^\diamond z = \|A^\diamond z\|_2^2 = \|A_1 u\|_2^2 \geq 0$. Since the columns of $A_1$ are linearly independent, we have $u = 0$ and thus $z = 0$. Therefore, $\mathbf{J}_x F(x^\diamond, A^\diamond, y^\diamond) = \lambda D^\diamond + (A^\diamond)^T A^\diamond$ is positive definite.

By the implicit function theorem, there are local $C^1$ functions $G_1$ and $G_2$ such that $x^* = (x^*_{\mathcal{J}}, x^*_{\mathcal{J}^c}) = (G_1(A, y), G_2(A, y)) := G(A, y)$ and $F(G(A, y), A, y) = 0$ for all $(A, y)$ in a neighborhood of $(A^\diamond, y^\diamond) \in S$. By the chain rule, we have

$$\mathbf{J}_x F(x^\diamond, A^\diamond, y^\diamond) \cdot \begin{bmatrix} \nabla_y G_1(A^\diamond, y^\diamond) \\ \nabla_y G_2(A^\diamond, y^\diamond) \end{bmatrix} + \mathbf{J}_y F(x^\diamond, A^\diamond, y^\diamond) = 0,$$

where $\mathbf{J}_y F(x^\diamond, A^\diamond, y^\diamond) = -(A^\diamond)^T = -[A_1 \ A_2]^T$. Let $P$ be the inverse of $\mathbf{J}_x F(x^\diamond, A^\diamond, y^\diamond)$, i.e.,

$$P = \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} = \begin{bmatrix} \lambda D_1 + A_1^T A_1 & A_1^T A_2 \\ A_2^T A_1 & \lambda D_2 + A_2^T A_2 \end{bmatrix}^{-1}.$$

Since $D_1 = 0$, we obtain $P_{11} A_1^T A_1 + P_{12} A_2^T A_1 = I$. Further, since $\nabla_y G_1(A^\diamond, y^\diamond) = P_{11} A_1^T + P_{12} A_2^T$, we have $\nabla_y G_1(A^\diamond, y^\diamond) \cdot A_1 = I$. This shows that each row of $\nabla_y G_1(A^\diamond, y^\diamond)$ is nonzero or equivalently the gradient of $x_i^*(A, y)$ is nonzero at $(A^\diamond, y^\diamond)$ for each $i \in \mathcal{J}^c$. By virtue of Lemma 3.1, $|supp(x^*(A, y))| = N$ for almost all $(A, y) \in \mathbb{R}^{m \times N} \times \mathbb{R}^m$.

(ii) $1 < p < 2$. Let $\widehat{S}$ be the set of all $(A, y) \in \mathbb{R}^{m \times N} \times \mathbb{R}^m$ such that each column of $A$ is nonzero and $A^T y \neq 0$. Obviously, $\widehat{S}$ is open and its complement has zero measure. Further, for any $(A, y) \in \widehat{S}$, it follows from the optimality condition and $A^T y \neq 0$ that the unique optimal solution $x^* \neq 0$.

Define the function

$$F(x, A, y) := \begin{bmatrix} x_1 + h(\lambda^{-1} a_1^T(Ax - y)) \\ \vdots \\ x_N + h(\lambda^{-1} a_N^T(Ax - y)) \end{bmatrix},$$

where $h$ is defined in (9). For any $(A, y) \in \mathbb{R}^{m \times N} \times \mathbb{R}^m$, the unique optimal solution $x^*$, as the function of $(A, y)$, satisfies $F(x^*, A, y) = 0$. Further, $F$ is $C^1$ and its Jacobian with respect to $x$ is

$$\mathbf{J}_x F(x, A, y) = I + \lambda^{-1} \cdot \Gamma(x, A, y) A^T A,$$

where the diagonal matrix $\Gamma(x, A, y) = \operatorname{diag}\big(h'(\lambda^{-1} a_1^T (Ax - y)), \ldots, h'(\lambda^{-1} a_N^T (Ax - y))\big)$.

We show next that for any $(A^\diamond, y^\diamond) \in \widehat{S}$, the matrix $\mathbf{J}_x F(x^\diamond, A^\diamond, y^\diamond)$ is invertible, where $x^\diamond := x^*(A^\diamond, y^\diamond)$. As before, define the index set $\mathcal{J} := \{i \mid x_i^\diamond \neq 0\}$ as before. Since $(A^\diamond, y^\diamond) \in \widehat{S}$ implies that $x^\diamond \neq 0$, the set $\mathcal{J}$ is nonempty. Partition $\Gamma^\diamond := \Gamma(x^\diamond, A^\diamond, y^\diamond)$ and $A^\diamond$ as $\Gamma^\diamond = \operatorname{diag}(\Gamma_1, \Gamma_2)$ and $A^\diamond = \begin{bmatrix} A_1 & A_2 \end{bmatrix}$ respectively, where $\Gamma_1 := \operatorname{diag}(h'(\lambda^{-1} (a_i^\diamond)^T (A^\diamond x^\diamond - y^\diamond)))_{i \in \mathcal{J}^c} = 0$ (due to $1 < p < 2$), $\Gamma_2 := \operatorname{diag}(h'(\lambda^{-1} (a_i^\diamond)^T (A^\diamond x^\diamond - y^\diamond)))_{i \in \mathcal{J}}$ is positive definite, $a_i^\diamond$ is the $i$th column of $A^\diamond$, $A_1 := A^\diamond_{\bullet \mathcal{J}^c}$, and $A_2 := A^\diamond_{\bullet \mathcal{J}}$. Therefore, we obtain

$$\mathbf{J}_x F(x^\diamond, A^\diamond, y^\diamond) = \begin{bmatrix} I & 0 \\ \lambda^{-1} \Gamma_2 A_2^T A_1 & I + \lambda^{-1} \Gamma_2 A_2^T A_2 \end{bmatrix}. \tag{11}$$

Since $\Gamma_2$ is positive definite, we deduce that $I + \lambda^{-1} \Gamma_2 A_2^T A_2 = \Gamma_2(\Gamma_2^{-1} + \lambda^{-1} A_2^T A_2)$ is invertible. Hence $\mathbf{J}_x F(x^\diamond, A^\diamond, y^\diamond)$ is invertible. By the implicit function theorem, there are local $C^1$ functions $G_1$ and $G_2$ such that $x^* = (x_{\mathcal{J}}^*, x_{\mathcal{J}^c}^*) = (G_1(A, y), G_2(A, y)) := G(A, y)$ and $F(G(A, y), A, y) = 0$ for all $(A, y)$ in a neighborhood of $(A^\diamond, y^\diamond) \in \mathbb{R}^{m \times N} \times \mathbb{R}^m$. By the chain rule, we have

$$\mathbf{J}_x F(x^\diamond, A^\diamond, y^\diamond) \cdot \begin{bmatrix} \nabla_y G_1(A^\diamond, y^\diamond) \\ \nabla_y G_2(A^\diamond, y^\diamond) \end{bmatrix} = -\mathbf{J}_y F(x^\diamond, A^\diamond, y^\diamond) = \begin{bmatrix} \Gamma_1 \lambda^{-1} A_1^T \\ \Gamma_2 \lambda^{-1} A_2^T \end{bmatrix} \in \mathbb{R}^{N \times m}.$$

In view of (11), $\Gamma_1 = 0$, and the invertibility of $\mathbf{J}_x F(x^\diamond, A^\diamond, y^\diamond)$, we obtain $\nabla_y G_1(A^\diamond, y^\diamond) = 0$ and $\nabla_y G_2(A^\diamond, y^\diamond) = (I + \lambda^{-1} \Gamma_2 A_2^T A_2)^{-1} \Gamma_2 \lambda^{-1} A_2^T$.

Noting that $x_i^* = -h(\lambda^{-1} a_i^T (Ax^* - y))$ for each $i = 1, \ldots, N$, we deduce via the property of the function $h$ in (9) that $\operatorname{sgn}(x_i^*) = \operatorname{sgn}(a_i^T (y - Ax^*))$ for each $i$. Therefore, it suffices to show that the zero set of $a_i^T (y - Ax^*)$ has zero measure for each $i = 1, \ldots, N$. It follows from the previous development that for any $(A^\diamond, y^\diamond) \in \widehat{S}$, $(x_{\mathcal{J}^c}^*, x_{\mathcal{J}}^*) = (G_1(A, y), G_2(A, y))$ in a neighborhood of $(A^\diamond, y^\diamond)$ for local $C^1$ functions $G_1$ and $G_2$. For each $i \in \mathcal{J}^c$, define

$$q_i(A, y) := a_i^T \big(y - A \cdot x^*(A, y)\big).$$

Then $\nabla_y q_i(A^\diamond, y^\diamond) = (a_i^\diamond)^T (I - A_2 \cdot \nabla_y G_2(A^\diamond, y^\diamond))$. Note that by the Sherman-Morrison-Woodbury formula [21, Section 3.8], we have

$$A_2 \cdot \nabla_y G_2(A^\diamond, y^\diamond) = A_2(I + \lambda^{-1} \Gamma_2 A_2^T A_2)^{-1} \Gamma_2 \lambda^{-1} A_2^T = I - (I + \lambda^{-1} A_2 \Gamma_2 A_2^T)^{-1},$$

where it is easy to see that $I + \lambda^{-1} A_2 \Gamma_2 A_2^T$ is invertible. Hence, for any $(A^\diamond, y^\diamond) \in \widehat{S}$, we deduce via $a_i^\diamond \neq 0$ that $\nabla_y q_i(A^\diamond, y^\diamond) = (a_i^\diamond)^T (I + (2\lambda)^{-1} A_2 \Gamma_2 A_2^T)^{-1} \neq 0$ for each $i \in \mathcal{J}^c$. In view of Lemma 3.1, $|\operatorname{supp}(x^*(A, y))| = N$ for almost all $(A, y) \in \mathbb{R}^{m \times N} \times \mathbb{R}^m$. $\qquad \square$

The next result pertains to the generalized elastic net (5).

**Theorem 4.3.** *Let $p > 1$, $N \geq m$, $r \geq 1$, and $\lambda_1, \lambda_2 > 0$. For almost all $(A, y) \in \mathbb{R}^{m \times N} \times \mathbb{R}^m$, the unique optimal solution $x^*_{(A, y)}$ to the $EN_p$ (5) satisfies $|\operatorname{supp}(x^*_{(A, y)})| = N$.*

*Proof.* Recall that for any given $(A, y) \in \mathbb{R}^{m \times N} \times \mathbb{R}^m$, the unique optimal solution $x^*$ to the $EN_p$ (5) is characterized by equation (7):

$$A^T(Ax^* - y) + p^{-1} r \lambda_1 \cdot \|x^*\|_p^{r-p} \cdot \nabla f(x^*) + 2\lambda_2 x^* = 0, \tag{12}$$

12

where $r \geq 1$, and $\lambda_1, \lambda_2 > 0$ are the penalty parameters.

(i) $p \geq 2$. Consider the open set $S$ defined in (6). For any $(A, y) \in S$, since $A$ has full row rank and $y \neq 0$, we have $A^T y \neq 0$. Hence, it follows from the optimality condition (12) and $A^T y \neq 0$ that the unique optimal solution $x^* \neq 0$ for any $(A, y) \in S$.

Define the function $F(x, A, y) := A^T(Ax - y) + p^{-1}r\lambda_1 \cdot \|x\|_p^{r-p} \cdot \nabla f(x) + 2\lambda_2 x$, where $\nabla f(x) = (g(x_1), \ldots, g(x_N))^T$. Hence, the optimal solution $x^*$, as the function of $(A, y)$, satisfies the equation $F(x^*, A, y) = 0$. Since $\|\cdot\|_p$ is $C^2$ on $\mathbb{R}^N \setminus \{0\}$, we see that $F$ is $C^1$ on the open set $(\mathbb{R}^N \setminus \{0\}) \times \mathbb{R}^{m \times N} \times \mathbb{R}^m$, and its Jacobian with respect to $x$ is given by

$$\mathbf{J}_x F(x, A, y) = A^T A + \lambda_1 \mathbf{H}(\|x\|_p^r) + 2\lambda_2 I,$$

where $\mathbf{H}(\|x\|_p^r)$ denotes the Hessian of $\|\cdot\|_p^r$ at any nonzero $x$. Since $r \geq 1$, $\|\cdot\|_p^r$ is a convex function and its Hessian at any nonzero $x$ must be positive semi-definite. This shows that for any $(A^\diamond, y^\diamond) \in S$, $\mathbf{J}_x F(x^\diamond, A^\diamond, y^\diamond)$ is positive definite, where $x^\diamond := x^*(A^\diamond, y^\diamond) \neq 0$. Hence, there exists a local $C^1$ function $G$ such that $x^* = G(A, y)$ with $F(G(A, y), A, y) = 0$ for all $(A, y)$ in a neighborhood of $(A^\diamond, y^\diamond)$.

Let $\Lambda(x) := \mathrm{diag}(g'(x_1), \ldots, g'(x_N))$. For a given $(A^\diamond, y^\diamond) \in S$, define the (nonempty) index set $\mathcal{J} := \{i \mid x_i^\diamond \neq 0\}$. Partition $\Lambda^\diamond := \Lambda(x^\diamond)$ and $A^\diamond$ as $\Lambda^\diamond = \mathrm{diag}(\Lambda_1, \Lambda_2)$ and $A^\diamond = \begin{bmatrix} A_1 & A_2 \end{bmatrix}$ respectively, where $\Lambda_1 := \mathrm{diag}(g'(x_i^\diamond))_{i \in \mathcal{J}^c}$, $\Lambda_2 := \mathrm{diag}(g'(x_i^\diamond))_{i \in \mathcal{J}}$ is positive definite, $A_1 := A_{\bullet \mathcal{J}^c}^\diamond$, and $A_2 := A_{\bullet \mathcal{J}}^\diamond$. Thus $\Lambda_1 = 0$ for $p > 2$, and $\Lambda_1 = 2I$ for $p = 2$. Using $\nabla(\|x\|_p) = (p\|x\|_p^{p-1})^{-1} \cdot \nabla(\|x\|_p^p)$ for any $x \neq 0$, we have

$$\mathbf{H}(\|x\|_p^r) = \frac{r}{p} \cdot \|x\|_p^{r-p} \cdot \left[ \Lambda(x) + \frac{r-p}{p\|x\|_p^p} \cdot \nabla f(x)(\nabla f(x))^T \right], \quad \forall\, x \neq 0.$$

Based on the partition given above, we have $\mathbf{H}(\|x^\diamond\|_p^{r-p}) = \mathrm{diag}(H_1, H_2)$, where the matrix $H_2$ is positive semi-definite, and $H_1 = 0$ for $p > 2$, and $H_1 = r\|x^\diamond\|_2^{r-2} \cdot I$ for $p = 2$. Therefore, we obtain

$$\mathbf{J}_x F(x^\diamond, A^\diamond, y^\diamond) = \begin{bmatrix} A_1^T A_1 + 2\lambda_2 I + \lambda_1 H_1 & A_1^T A_2 \\ A_2^T A_1 & A_2^T A_2 + 2\lambda_2 I + \lambda_1 H_2 \end{bmatrix}, \quad \mathbf{J}_y F(x^\diamond, A^\diamond, y^\diamond) = - \begin{bmatrix} A_1^T \\ A_2^T \end{bmatrix}.$$

Let $Q$ be the inverse of $\mathbf{J}_x F(x^\diamond, A^\diamond, y^\diamond)$, i.e., $Q = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix}$. Hence, we have

$$(Q_{11} A_1^T + Q_{12} A_2^T) A_2 + Q_{12}(2\lambda_2 I + \lambda_1 H_2) = 0. \tag{13}$$

We claim that each row of $Q_{11} A_1^T + Q_{12} A_2^T$ is nonzero. Suppose not, i.e., $(Q_{11} A_1^T + Q_{12} A_2^T)_{i\bullet} = 0$ for some $i$. Then it follows from (13) that $(Q_{12})_{i\bullet}(2\lambda_2 I + \lambda_1 H_2) = 0$. Since $2\lambda_2 I + \lambda_1 H_2$ is positive definite, we have $(Q_{12})_{i\bullet} = 0$. By $(Q_{11} A_1^T + Q_{12} A_2^T)_{i\bullet} = 0$, we obtain $(Q_{11})_{i\bullet} A_1^T = 0$. It follows from Proposition 3.1 that $|\mathcal{J}^c| \leq m - 1$ such that the columns of $A_1$ are linearly independent. Hence, we have $(Q_{11})_{i\bullet} = 0$ or equivalently $Q_{j\bullet} = 0$ for some $j$. This contradicts the invertibility of $Q$, and thus completes the proof of the claim. Furthermore, let $G_1, G_2$ be local $C^1$ functions such that $x^* = (x_{\mathcal{J}^c}^*, x_{\mathcal{J}}^*) = (G_1(A, y), G_2(A, y))$ for all $(A, y)$ in a neighborhood of $(A^\diamond, y^\diamond) \in S$. By the similar argument as before, we see that $\nabla_y G_1(A^\diamond, y^\diamond) = Q_{11} A_1^T + Q_{12} A_2^T$. Therefore, we deduce that the gradient of $x_i^*(A, y)$ at $(A^\diamond, y^\diamond)$ is nonzero for each $i \in \mathcal{J}^c$. By Lemma 3.1, $|\mathrm{supp}(x^*(A, y))| = N$ for almost all $(A, y)$.

(ii) $1 < p < 2$. Let $\widehat{S}$ be the set defined in Case (ii) of Theorem 4.2, i.e., $\widehat{S}$ is the set of all $(A, y) \in \mathbb{R}^{m \times N} \times \mathbb{R}^m$ such that each column of $A$ is nonzero and $A^T y \neq 0$. The set $\widehat{S}$ is open and its complement has zero measure. For any $(A, y) \in \widehat{S}$, the unique optimal solution $x^*$, treated as a function of $(A, y)$, is nonzero. By the optimality condition (12) and the definition of the function $h$ in (9), we see that $x^*$ satisfies the following equation for any $(A, y) \in \widehat{S}$:

$$F(x, A, y) := \begin{bmatrix} x_1 + h(w_1) \\ \vdots \\ x_N + h(w_N) \end{bmatrix} = 0, \tag{14}$$

13

where $w_i := p\|x\|_p^{p-r} \cdot \left[a_i^T(Ax-y) + 2\lambda_2 x_i\right]/(r\lambda_1)$ for each $i = 1, \ldots, N$. It is easy to show that $F$ is $C^1$ on $(\mathbb{R}^N \setminus \{0\}) \times \mathbb{R}^{m \times N} \times \mathbb{R}^m$ and its Jacobian with respect to $x$ is

$$\mathbf{J}_x F(x, A, y) = I + \frac{p}{r\lambda_1} \cdot \Gamma(x, A, y) \cdot \left\{ [A^T(Ax-y) + 2\lambda_2 x] \cdot [\nabla(\|x\|_r^{p-r})]^T + \|x\|_p^{p-r}(A^T A + 2\lambda_2 I) \right\},$$

where the diagonal matrix $\Gamma(x, A, y) := \mathrm{diag}\big(h'(w_1), \ldots, h'(w_N)\big)$ and $\nabla(\|x\|_r^{p-r}) = (p-r)\nabla f(x)/[p \cdot \|x\|_p^r]$.

We show next that $\mathbf{J}_x F(x^\diamond, A^\diamond, y^\diamond)$ is invertible for any $(A^\diamond, y^\diamond) \in \widehat{S}$, where $x^\diamond := x^*(A^\diamond, y^\diamond)$. Define the (nonempty) index set $\mathcal{J} := \{i \mid x_i^\diamond \neq 0\}$. Partition $\Gamma^\diamond := \Gamma(x^\diamond, A^\diamond, y^\diamond)$ and $A^\diamond$ as $\Gamma^\diamond = \mathrm{diag}(\Gamma_1, \Gamma_2)$ and $A^\diamond = \begin{bmatrix} A_1 & A_2 \end{bmatrix}$ respectively, where $\Gamma_1 := \mathrm{diag}(h'(w_i))_{i \in \mathcal{J}^c} = 0$, $\Gamma_2 := \mathrm{diag}(h'(w_i))_{i \in \mathcal{J}}$ is positive definite, $a_i^\diamond$ is the $i$th column of $A^\diamond$, $A_1 := A_{\bullet \mathcal{J}^c}^\diamond$, and $A_2 := A_{\bullet \mathcal{J}}^\diamond$. Therefore,

$$W := \mathbf{J}_x F(x^\diamond, A^\diamond, y^\diamond) = \begin{bmatrix} I & 0 \\ \star & W_{22} \end{bmatrix},$$

where by letting the vector $\widetilde{b} := (\nabla f(x^\diamond))_{\mathcal{J}}$,

$$W_{22} := I + \frac{p}{r\lambda_1} \cdot \Gamma_2 \cdot \left\{ [A_2^T(A_2 x_{\mathcal{J}}^\diamond - y) + 2\lambda_2 x_{\mathcal{J}}^\diamond] \cdot \frac{(p-r)\widetilde{b}^T}{p\|x^\diamond\|_p^r} + \|x^\diamond\|_p^{p-r}(A_2^T A_2 + 2\lambda_2 I) \right\}.$$

It follows from (12) that $\frac{p}{r\lambda_1} \cdot [A_2^T(A_2 x_{\mathcal{J}}^\diamond - y) + 2\lambda_2 x_{\mathcal{J}}^\diamond] = -\|x^\diamond\|_p^{r-p} \cdot \widetilde{b}$. Hence,

$$\Gamma_2^{-1} \cdot W_{22} = \underbrace{\Gamma_2^{-1} + \frac{r-p}{p\|x^\diamond\|_p^p} \cdot \widetilde{b} \cdot \widetilde{b}^T}_{:= U} + \frac{p\|x^\diamond\|_p^{p-r}}{r\lambda_1} \cdot (A_2^T A_2 + 2\lambda_2 I). \tag{15}$$

Clearly, when $r \geq p > 1$, the matrix $\Gamma_2^{-1} W_{22}$ is positive definite. In what follows, we consider the case where $2 > p > r \geq 1$. Let the vector $b := (\mathrm{sgn}(x_i^\diamond) \cdot |x_i^\diamond|^{p-1})_{i \in \mathcal{J}}$ so that $\widetilde{b} = p \cdot b$. In view of (14), we have $w_i = h^{-1}(-x_i^\diamond) = p \cdot \mathrm{sgn}(-x_i)|x_i^\diamond|^{p-1}$ for each $i$. Therefore, using the formula for $h'(\cdot)$ given below (9), we obtain that for each $i \in \mathcal{J}$,

$$h'(w_i) = \frac{|w_i|^{\frac{2-p}{p-1}}}{(p-1) \cdot p^{\frac{1}{p-1}}} = \frac{\big(p \cdot |x_i^\diamond|^{p-1}\big)^{\frac{2-p}{p-1}}}{(p-1) \cdot p^{\frac{1}{p-1}}} = \frac{|x_i^\diamond|^{2-p}}{(p-1) \cdot p}.$$

This implies that $\Gamma_2^{-1} = p(p-1)D$, where the diagonal matrix $D := \mathrm{diag}(|x_i^\diamond|^{p-2})_{i \in \mathcal{J}}$. Clearly, $D$ is positive definite. We thus have, via $p - 1 \geq p - r > 0$,

$$U = \Gamma_2^{-1} + \frac{r-p}{p\|x^\diamond\|_p^p} \cdot \widetilde{b} \cdot \widetilde{b}^T = p(p-1)\left(D - \frac{p-r}{p-1} \cdot \frac{b \cdot b^T}{\|x\|_p^p}\right) \succcurlyeq p(p-1)\left(D - \frac{b \cdot b^T}{\|x\|_p^p}\right),$$

where $\succcurlyeq$ denotes the positive semi-definite order. Since the diagonal matrix $D$ is positive definite, we further have

$$D - \frac{b \cdot b^T}{\|x\|_p^p} = D^{1/2}\left(I - \frac{D^{-1/2}b \cdot b^T D^{-1/2}}{\|x\|_p^p}\right)D^{1/2} = D^{1/2}\left(I - \frac{u \cdot u^T}{\|u\|_2^2}\right)D^{1/2},$$

where $u := D^{-1/2} \cdot b = \big(\mathrm{sgn}(x_i^\diamond)|x_i^\diamond|^{p/2}\big)_{i \in \mathcal{J}}$ such that $\|u\|_2^2 = \|x\|_p^p$. Since $I - \frac{u \cdot u^T}{\|u\|_2^2}$ is positive semi-definite, so is $D - \frac{b \cdot b^T}{\|x\|_p^p}$. This shows that $U$ in (15) is positive semi-definite. Since the last term on the right hand side of (15) is positive definite, $\Gamma_2^{-1} W_{22}$ is positive definite. Therefore, $W_{22}$ is invertible, and so is $W$ for all $1 < p < 2$ and $r \geq 1$. By the implicit function theorem, there are local $C^1$ functions $G_1$ and $G_2$

14

such that $x^* = (x^*_{\mathcal{J}}, x^*_{\mathcal{J}^c}) = (G_1(A, y), G_2(A, y)) := G(A, y)$ and $F(G(A, y), A, y) = 0$ for all $(A, y)$ in a neighborhood of $(A^\diamond, y^\diamond) \in \mathbb{R}^{m \times N} \times \mathbb{R}^m$. Moreover, we have

$$\mathbf{J}_x F(x^\diamond, A^\diamond, y^\diamond) \cdot \begin{bmatrix} \nabla_y G_1(A^\diamond, y^\diamond) \\ \nabla_y G_2(A^\diamond, y^\diamond) \end{bmatrix} = -\mathbf{J}_y F(x^\diamond, A^\diamond, y^\diamond) = \frac{p\|x^\diamond\|_p^{p-r}}{r\lambda_1} \begin{bmatrix} \Gamma_1 A_1^T \\ \Gamma_2 A_2^T \end{bmatrix} \in \mathbb{R}^{N \times m}.$$

In view of the invertibility of $\mathbf{J}_x F(x^\diamond, A^\diamond, y^\diamond)$ and $\Gamma_1 = 0$, we obtain

$$\nabla_y G_1(A^\diamond, y^\diamond) = 0, \qquad \nabla_y G_2(A^\diamond, y^\diamond) = \left(\frac{p\|x^\diamond\|_p^{p-r}}{r\lambda_1}\right) W_{22}^{-1} \Gamma_2 A_2^T.$$

Since $x_i^* = -h(w_i)$ for each $i = 1, \ldots, N$, where $w_i$ is defined below (14), we deduce via the positivity of $\|x\|_p$ and the property of the function $h$ in (9) that $\operatorname{sgn}(x_i^*) = \operatorname{sgn}(a_i^T(y - Ax^*) - 2\lambda_2 x_i^*)$ for each $i$. For each $i \in \mathcal{J}^c$, define

$$q_i(A, y) := a_i^T\big(y - A \cdot x^*(A, y)\big) - 2\lambda_2 x_i^*(A, y).$$

In what follows, we show that for each $i \in \mathcal{J}^c$, the gradient of $q_i(A, y)$ at $(A^\diamond, y^\diamond) \in \widehat{S}$ is nonzero. It follows from the previous development that for any $(A^\diamond, y^\diamond) \in \widehat{S}$, $(x^*_{\mathcal{J}^c}, x^*_{\mathcal{J}}) = (G_1(A, y), G_2(A, y))$ in a neighborhood of $(A^\diamond, y^\diamond)$ for local $C^1$ functions $G_1$ and $G_2$. Using $\nabla_y G_1(A^\diamond, y^\diamond) = 0$, we have $\nabla_y q_i(A^\diamond, y^\diamond) = (a_i^\diamond)^T(I - A_2 \cdot \nabla_y G_2(A^\diamond, y^\diamond))$. Letting $\alpha := \frac{p\|x^\diamond\|_p^{p-r}}{r\lambda_1} > 0$ and by (15), we have

$$A_2 \cdot \nabla_y G_2(A^\diamond, y^\diamond) = \alpha A_2 W_{22}^{-1} \Gamma_2 A_2^T = \alpha A_2 (\Gamma_2^{-1} W_{22})^{-1} A_2^T = \alpha A_2 \big[U + \alpha(A_2^T A_2 + 2\lambda_2 I)\big]^{-1} A_2^T.$$

Since $U$ is positive semi-definite and $A_2^T A_2 + 2\lambda_2 I$ is positive definite, it can be shown that

$$A_2 \big(A_2^T A_2 + 2\lambda_2 I\big)^{-1} A_2^T \succcurlyeq \alpha A_2 \big[U + \alpha(A_2^T A_2 + 2\lambda_2 I)\big]^{-1} A_2^T \succcurlyeq 0.$$

Since each eigenvalue of $A_2(A_2^T A_2 + 2\lambda_2 I)^{-1} A_2^T$ is strictly less than one, we conclude that $A_2 \cdot \nabla_y G_2(A^\diamond, y^\diamond)$ is positive semi-definite and each of its eigenvalues is strictly less than one. Therefore, $I - A_2 \cdot \nabla_y G_2(A^\diamond, y^\diamond)$ is invertible. Since each $a_i^\diamond \neq 0$, we have $\nabla_y q_i(A^\diamond, y^\diamond) \neq 0$ for each $i \in \mathcal{J}^c$. In view of Lemma 3.1, $|\operatorname{supp}(x^*(A, y))| = N$ for almost all $(A, y) \in \mathbb{R}^{m \times N} \times \mathbb{R}^m$. $\qquad\square$

## 4.3 Least Sparsity of the Generalized Basis Pursuit Denoising with $p > 1$

We consider the case where $p \geq 2$ first.

**Proposition 4.3.** *Let $p \geq 2$ and $N \geq m$. For almost all $(A, y) \in \mathbb{R}^{m \times N} \times \mathbb{R}^m$ with $y \neq 0$, if $0 < \varepsilon < \|y\|_2$, then the unique optimal solution $x^*_{(A,y)}$ to the $BPDN_p$ (2) satisfies $|\operatorname{supp}(x^*_{(A,y)})| = N$.*

*Proof.* Consider the set $S$ defined in (6). It follows from the proof for Case (ii) in Proposition 3.1 that for any given $(A, y) \in S$ and any $\varepsilon > 0$ with $\varepsilon < \|y\|_2$, the unique optimal solution $x^*$ satisfies the optimality conditions: $\nabla f(x^*) + 2\mu A^T(Ax^* - y) = 0$ for a unique positive $\mu$, and $\|Ax^* - y\|_2^2 = \varepsilon^2$. Hence, $(x^*, \mu) \in \mathbb{R}^{N+1}$ is a function of $(A, y)$ on $S$ and satisfies the following equation:

$$F(x, \mu, A, y) := \begin{bmatrix} g(x_1) + 2\mu\, a_1^T(Ax - y) \\ \vdots \\ g(x_N) + 2\mu\, a_N^T(Ax - y) \\ \|Ax - y\|_2^2 - \varepsilon^2 \end{bmatrix} = 0.$$

Clearly, $F : \mathbb{R}^N \times \mathbb{R} \times \mathbb{R}^{m \times N} \times \mathbb{R}^m \to \mathbb{R}^{N+1}$ is $C^1$ and its Jacobian with respect to $(x, \mu)$ is given by

$$\mathbf{J}_{(x,\mu)} F(x, \mu, A, y) = \begin{bmatrix} M(x, \mu, A) & 2A^T(Ax - y) \\ 2(Ax - y)^T A & 0 \end{bmatrix},$$

where $M(x, \mu, A) := \Lambda(x) + 2\mu A^T A$, and the diagonal matrix $\Lambda(x) := \text{diag}(g'(x_1), \ldots, g'(x_N))$ is positive semi-definite. Given $(A^\diamond, y^\diamond) \in S$, define $x^\diamond := x^*(A^\diamond, y^\diamond)$ and $\mu^\diamond := \mu(A^\diamond, y^\diamond) > 0$. We claim that $\mathbf{J}_{(x,\mu)}F(x^\diamond, \mu^\diamond, A^\diamond, y^\diamond)$ is invertible for any $(A^\diamond, y^\diamond) \in S$. To show it, define the index set $\mathcal{J} := \{i \,|\, x_i^\diamond \neq 0\}$. Note that $\mathcal{J}$ is nonempty by virtue of Proposition 3.1. Partition $\Lambda^\diamond := \Lambda(x^\diamond)$ and $A$ as $\Lambda^\diamond = \text{diag}(\Lambda_1, \Lambda_2)$ and $A^\diamond = [A_1 \; A_2]$ respectively, where $\Lambda_1 := \text{diag}(g'(x_i^\diamond))_{i \in \mathcal{J}^c}$, $\Lambda_2 := \text{diag}(g'(x_i^\diamond))_{i \in \mathcal{J}}$ is positive definite, $A_1 := A^\diamond_{\bullet \mathcal{J}^c}$, and $A_2 := A^\diamond_{\bullet \mathcal{J}}$. Here, $A_1^T(A^\diamond x^\diamond - y^\diamond) = 0$, $\Lambda_1 = 0$ for $p > 2$, and $\Lambda^\diamond = \text{diag}(\Lambda_1, \Lambda_2) = 2I$ for $p = 2$. It follows from the similar argument for Case (i) in Theorem 4.2 that $M^\diamond := M(x^\diamond, \mu^\diamond, A^\diamond)$ is positive definite. Moreover, it has been shown in the proof of Proposition 3.1 that $b := 2(A^\diamond)^T(A^\diamond x^\diamond - y^\diamond) \in \mathbb{R}^N$ is nonzero. Hence, for any $z = [z_1; z_2] \in \mathbb{R}^{N+1}$ with $z_1 \in \mathbb{R}^N$ and $z_2 \in \mathbb{R}$, we have

$$\mathbf{J}_{(x,\mu)}F(x^\diamond, \mu^\diamond, A^\diamond, y^\diamond)z = \begin{bmatrix} M^\diamond & b \\ b^T & 0 \end{bmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = 0 \; \Rightarrow \; M^\diamond z_1 + b z_2 = 0, \; b^T z_1 = 0 \; \Rightarrow \; b^T(M^\diamond)^{-1} b z_2 = 0.$$

This implies that $z_2 = 0$ and further $z_1 = 0$. Therefore, $\mathbf{J}_{(x,\mu)}F(x^\diamond, \mu^\diamond, A^\diamond, y^\diamond)$ is invertible. By the implicit function theorem, there are local $C^1$ functions $G_1, G_2, H$ such that $x^* = (x^*_{\mathcal{J}^c}, x^*_{\mathcal{J}}) = (G_1(A, y), G_2(A, y)) := G(A, y)$, $\mu = H(A, y)$, and $F(G(A, y), H(A, y), A, y) = 0$ for all $(A, y)$ in a neighborhood of $(A^\diamond, y^\diamond)$. By the chain rule, we have

$$\underbrace{\mathbf{J}_{(x,\mu)}F(x^\diamond, \mu^\diamond, A^\diamond, y^\diamond)}_{:=V} \cdot \begin{bmatrix} \nabla_y G_1(A^\diamond, y^\diamond) \\ \nabla_y G_2(A^\diamond, y^\diamond) \\ \nabla_y H(A^\diamond, y^\diamond) \end{bmatrix} = -\mathbf{J}_y F(x^\diamond, \mu^\diamond, A^\diamond, y^\diamond) = \begin{bmatrix} 2\mu^\diamond A_1^T \\ 2\mu^\diamond A_2^T \\ 2(A^\diamond x^\diamond - y^\diamond)^T \end{bmatrix},$$

where

$$V = \begin{bmatrix} \Lambda_1 + 2\mu^\diamond A_1^T A_1 & 2\mu^\diamond A_1^T A_2 & 2A_1^T(A^\diamond x^\diamond - y^\diamond) \\ 2\mu^\diamond A_2^T A_1 & \Lambda_2 + 2\mu^\diamond A_2^T A_2 & 2A_2^T(A^\diamond x^\diamond - y^\diamond) \\ 2(A^\diamond x^\diamond - y^\diamond)^T A_1 & 2(A^\diamond x^\diamond - y^\diamond)^T A_2 & 0 \end{bmatrix}.$$

Let $P$ be the inverse of $V$ given by the symmetric matrix $P = \begin{bmatrix} P_{11} & P_{12} & P_{13} \\ P_{12}^T & P_{22} & P_{23} \\ P_{13}^T & P_{23}^T & P_{33} \end{bmatrix}$. Consider $p > 2$ first. In this case, $\Lambda_1 = 0$ such that $P_{11} 2\mu^\diamond A_1^T A_1 + P_{12} 2\mu^\diamond A_2^T A_1 + P_{13} 2(A^\diamond x^\diamond - y^\diamond)^T A_1 = I_m$. Since

$$\nabla_y G_1(A^\diamond, y^\diamond) = -[P_{11} \; P_{12} \; P_{13}] \cdot \mathbf{J}_y F(x^\diamond, \mu^\diamond, A^\diamond, y^\diamond) = P_{11} 2\mu^\diamond A_1^T + P_{12} 2\mu^\diamond A_2^T + P_{13} 2(A^\diamond x^\diamond - y^\diamond)^T,$$

we have $\nabla_y G_1(A^\diamond, y^\diamond) \cdot A_1 = I_m$. This shows that each row of $\nabla_y G_1(A^\diamond, y^\diamond)$ is nonzero. We then consider $p = 2$. In this case, letting $B := \nabla_y G_1(A^\diamond, y^\diamond) = P_{11} 2\mu^\diamond A_1^T + P_{12} 2\mu^\diamond A_2^T + P_{13} 2(A^\diamond x^\diamond - y^\diamond)^T$ and using $\text{diag}(\Lambda_1, \Lambda_2) = 2I$, we have $2P_{11} + BA_1 = I$ and $2P_{12} + BA_2 = 0$. Suppose, by contradiction, that the $i$th row of $B$ is zero, i.e., $B_{i\bullet} = 0$. Then $(P_{11})_{i\bullet} = \mathbf{e}_i^T/2$ and $(P_{12})_{i\bullet} = 0$, where $\mathbf{e}_i$ denotes the $i$th column of $I$. Substituting these results into $B$ and using $A_1^T(A^\diamond x^\diamond - y^\diamond) = 0$, we have $0 = B_{i\bullet}(A^\diamond x^\diamond - y^\diamond) = 2\mu^\diamond(P_{11})_{i\bullet}A_1^T(A^\diamond x^\diamond - y^\diamond) + 2(P_{13})_i\|A^\diamond x^\diamond - y^\diamond\|_2^2 = 2(P_{13})_i\|A^\diamond x^\diamond - y^\diamond\|_2^2$, which implies that the real number $(P_{13})_i = 0$ as $A^\diamond x^\diamond - y^\diamond \neq 0$. In view of the symmetry of $P$, $V = P^{-1}$ and $\Lambda_1 = 2I$, we have $(2I + 2\mu^\diamond A_1^T A_1)_{ii} = 2$, which yields $(A_1)_{\bullet i} = 0$, a contradiction. Therefore, each row of $\nabla_y G_1(A^\diamond, y^\diamond)$ is nonzero. Consequently, by Lemma 3.1, $|\text{supp}(x^*(A, y))| = N$ for almost all $(A, y) \in \mathbb{R}^{m \times N} \times \mathbb{R}^m$. $\qquad \square$

In what follows, we consider the case where $1 < p < 2$.

**Proposition 4.4.** *Let $1 < p < 2$ and $N \geq m$. For almost all $(A, y) \in \mathbb{R}^{m \times N} \times \mathbb{R}^m$ with $y \neq 0$, if $0 < \varepsilon < \|y\|_2$, then the unique optimal solution $x^*_{(A,y)}$ to the $\text{BPDN}_p$ (2) satisfies $|\text{supp}(x^*_{(A,y)})| = N$.*

*Proof.* Let $\check{S}$ be the set of all $(A, y) \in \mathbb{R}^{m \times N} \times \mathbb{R}^m$ such that each column of $A$ is nonzero and $y \neq 0$. Obviously, $\check{S}$ is open in $\mathbb{R}^{m \times N} \times \mathbb{R}^m$ and its complement has zero measure. For any $(A, y) \in \check{S}$ and any

positive $\varepsilon$ with $\varepsilon < \|y\|_2$, it follows from the proof for Case (ii) in Proposition 3.1 that the unique optimal solution $x^* \neq 0$ with a unique positive $\mu$. Further, $(x^*, \mu) \in \mathbb{R}^{N+1}$ satisfies the following equation:

$$F(x, \mu, A, y) := \begin{bmatrix} x_1 + h(2\mu a_1^T(Ax - y)) \\ \vdots \\ x_N + h(2\mu a_N^T(Ax - y)) \\ \|Ax - y\|_2^2 - \varepsilon^2 \end{bmatrix} = 0,$$

where $h$ is defined in (9). Hence, $F : \mathbb{R}^N \times \mathbb{R} \times \mathbb{R}^{m \times N} \times \mathbb{R}^m \to \mathbb{R}^{N+1}$ is $C^1$ and its Jacobian with respect to $(x, \mu)$ is given by

$$\mathbf{J}_{(x,\mu)}F(x, \mu, A, y) = \begin{bmatrix} V(x, \mu, A, y) & 2\Gamma(x, \mu, A, y)A^T(Ax - y) \\ 2(Ax - y)^T A & 0 \end{bmatrix} \in \mathbb{R}^{(N+1) \times (N+1)},$$

where $\Gamma(x, \mu, A, y) := \text{diag}(h'(2\mu a_1^T(Ax - y)), \dots, h'(2\mu a_N^T(Ax - y))) \in \mathbb{R}^{N \times N}$, and $V(x, \mu, A, y) := I + \Gamma(x, \mu, A, y)\, 2\mu A^T A$.

We use the same notation $x^\diamond$ and $\mu^\diamond$ as before. For any $(A^\diamond, y^\diamond) \in \mathbb{R}^{m \times N} \times \mathbb{R}^m$, define the index set $\mathcal{J} := \{i \,|\, x_i^\diamond \neq 0\}$. Note that $\mathcal{J}$ is nonempty as $\|y\|_2 > \varepsilon$. Partition $\Gamma^\diamond := \Gamma(x^\diamond, \mu^\diamond, A^\diamond, y^\diamond)$ and $A^\diamond$ as $\Gamma^\diamond = \text{diag}(\Gamma_1, \Gamma_2)$ and $A^\diamond = \begin{bmatrix} A_1 & A_2 \end{bmatrix}$ respectively, where $\Gamma_1 := \text{diag}(h'(2\mu^\diamond(a_i^\diamond)^T(A^\diamond x^\diamond - y^\diamond)))_{i \in \mathcal{J}^c} = 0$ (due to $1 < p < 2$), $\Gamma_2 := \text{diag}(h'(2\mu^\diamond(a_i^\diamond)^T(A^\diamond x^\diamond - y^\diamond)))_{i \in \mathcal{J}}$ is positive definite, $A_1 := A^\diamond_{\bullet \mathcal{J}^c}$, and $A_2 := A^\diamond_{\bullet \mathcal{J}}$. Therefore, using the fact that $\Gamma_1 = 0$ and $\Gamma_2$ is positive definite, we obtain

$$\mathbf{J}_{(x,\mu)}F(x^\diamond, \mu^\diamond, A^\diamond, y^\diamond) = \begin{bmatrix} I & 0 & 0 \\ 2\mu^\diamond \Gamma_2 A_2^T A_1 & I + 2\mu^\diamond \Gamma_2 A_2^T A_2 & 2\Gamma_2 A_2^T(A^\diamond x^\diamond - y^\diamond) \\ 2(A^\diamond x^\diamond - y^\diamond)^T A_1 & 2(A^\diamond x^\diamond - y^\diamond)^T A_2 & 0 \end{bmatrix}. \tag{16}$$

Since $\Gamma_2$ is positive definite, the lower diagonal block in $\mathbf{J}_{(x,\mu)}F(x^\diamond, \mu^\diamond, A^\diamond, y^\diamond)$ becomes

$$\begin{bmatrix} I + 2\mu^\diamond \Gamma_2 A_2^T A_2 & 2\Gamma_2 A_2^T(A^\diamond x^\diamond - y^\diamond) \\ 2(A^\diamond x^\diamond - y^\diamond)^T A_2 & 0 \end{bmatrix} = \begin{bmatrix} \Gamma_2 & 0 \\ 0 & I \end{bmatrix} \cdot \underbrace{\begin{bmatrix} \Gamma_2^{-1} + 2\mu^\diamond A_2^T A_2 & 2A_2^T(A^\diamond x^\diamond - y^\diamond) \\ 2(A^\diamond x^\diamond - y^\diamond)^T A_2 & 0 \end{bmatrix}}_{:=Q}. \tag{17}$$

Clearly, $\Gamma_2^{-1} + 2\mu^\diamond A_2^T A_2$ is positive definite. Further, since $\mu^\diamond > 0$ and $x_i^\diamond \neq 0, \forall i \in \mathcal{J}$, we have $A_2^T(A^\diamond x^\diamond - y^\diamond) \neq 0$. Hence, by the similar argument for Proposition 4.3, we see that the matrix $Q$ is invertible such that $\mathbf{J}_{(x,\mu)}F(x^\diamond, \mu^\diamond, A^\diamond, y^\diamond)$ is invertible. By the implicit function theorem, there are local $C^1$ functions $G_1, G_2, H$ such that $x^* = (x^*_{\mathcal{J}^c}, x^*_{\mathcal{J}}) = (G_1(A, y), G_2(A, y)) := G(A, y)$, $\mu = H(A, y)$, and $F(G(A, y), H(A, y), A, y) = 0$ for all $(A, y)$ in a neighborhood of $(A^\diamond, y^\diamond)$. By the chain rule, we obtain

$$\mathbf{J}_{(x,\mu)}F(x^\diamond, \mu^\diamond, A^\diamond, y^\diamond) \cdot \begin{bmatrix} \nabla_y G_1(A^\diamond, y^\diamond) \\ \nabla_y G_2(A^\diamond, y^\diamond) \\ \nabla_y H(A^\diamond, y^\diamond) \end{bmatrix} = -\mathbf{J}_y F(x^\diamond, \mu^\diamond, A^\diamond, y^\diamond) = \begin{bmatrix} 0 \\ \Gamma_2 \, 2\mu^\diamond A_2^T \\ 2(A^\diamond x^\diamond - y^\diamond)^T \end{bmatrix},$$

where we use the fact that $\Gamma_1 = 0$. In view of (16) and the above results, we have $\nabla_y G_1(A^\diamond, y^\diamond) = 0$, and we deduce via (17) that

$$\begin{pmatrix} \nabla_y G_2(A^\diamond, y^\diamond) \\ \nabla_y H(A^\diamond, y^\diamond) \end{pmatrix} = \begin{bmatrix} \Gamma_2^{-1} + 2\mu^\diamond A_2^T A_2 & 2A_2^T(A^\diamond x^\diamond - y^\diamond) \\ 2(A^\diamond x^\diamond - y^\diamond)^T A_2 & 0 \end{bmatrix}^{-1} \cdot \begin{pmatrix} 2\mu^\diamond A_2^T \\ 2(A^\diamond x^\diamond - y^\diamond)^T \end{pmatrix}, \tag{18}$$

where $A^\diamond x^\diamond - y^\diamond \neq 0$ because otherwise $\|A^\diamond x^\diamond - y^\diamond\|_2^2 - \varepsilon^2 \neq 0$.

For each $i \in \mathcal{J}^c$, define $q_i(A, y) := a_i^T(y - A \cdot x^*(A, y))$. It follows from $\text{sgn}(x_i^*(A, y)) = \text{sgn}(q_i(A, y))$ and the previous argument that it suffices to show that $\nabla_y q_i(A^\diamond, y^\diamond) \neq 0$ for each $i \in \mathcal{J}^c$, where

$\nabla_y q_i(A^\diamond, y^\diamond) = (a_i^\diamond)^T (I - A_2 \cdot \nabla_y G_2(A^\diamond, y^\diamond))$. Toward this end, we see, by using $(a_i^\diamond)^T (A^\diamond x^\diamond - y^\diamond) = 0, \forall i \in \mathcal{J}^c$ and (18), that for each $i \in \mathcal{J}^c$,

$$
\begin{aligned}
(a_i^\diamond)^T A_2 \nabla_y G_2(A^\diamond, y^\diamond) &= \frac{(a_i^\diamond)^T}{2\mu^\diamond} \left[ 2\mu^\diamond A_2^\diamond \quad 2(A^\diamond x^\diamond - y^\diamond) \right] \cdot \begin{pmatrix} \nabla_y G_2(A^\diamond, y^\diamond) \\ \nabla_y H(A^\diamond, y^\diamond) \end{pmatrix} \\
&= \frac{(a_i^\diamond)^T}{2\mu^\diamond} \left[ 2\mu^\diamond A_2^\diamond \quad 2(A^\diamond x^\diamond - y^\diamond) \right] \cdot \begin{bmatrix} \Gamma_2^{-1} + 2\mu^\diamond A_2^T A_2 & 2A_2^T(A^\diamond x^\diamond - y^\diamond) \\ 2(A^\diamond x^\diamond - y^\diamond)^T A_2 & 0 \end{bmatrix}^{-1} \cdot \begin{bmatrix} 2\mu^\diamond A_2^T \\ 2(A^\diamond x^\diamond - y^\diamond)^T \end{bmatrix} \cdot
\end{aligned}
$$

Define

$$
d := A^\diamond x^\diamond - y^\diamond \neq 0, \quad C := \left[ \sqrt{2\mu^\diamond} \cdot A_2, \quad \sqrt{\frac{2}{\mu^\diamond}} \cdot d \right], \quad D := \begin{bmatrix} \Gamma_2^{-1} & 0 \\ 0 & -\frac{2}{\mu^\diamond} \|d\|_2^2 \end{bmatrix}.
$$

It is easy to verify that

$$
\begin{bmatrix} \Gamma_2^{-1} + 2\mu^\diamond A_2^T A_2 & 2A_2^T(A^\diamond x^\diamond - y^\diamond) \\ 2(A^\diamond x^\diamond - y^\diamond)^T A_2 & 0 \end{bmatrix} = D + C^T C.
$$

Therefore, we obtain

$$
(a_i^\diamond)^T (I - A_2 \cdot \nabla_y G_2(A^\diamond, y^\diamond)) = (a_i^\diamond)^T - (a_i^\diamond)^T C (D + C^T C)^{-1} C^T.
$$

Recall that $\mathcal{J}$ is nonempty such that $A_2$ exists and $A_2^T d \neq 0$. Since $A_2 \Gamma_2 A_2^T$ and $I - \frac{dd^T}{\|d\|_2^2}$ are both positive semi-definite and $N(I - \frac{dd^T}{\|d\|_2^2}) = \text{span}\{d\}$, it is easy to see that $N(A_2 \Gamma_2 A_2^T) \cap N(I - \frac{dd^T}{\|d\|_2^2}) = \{0\}$. Hence, the following matrix is positive definite:

$$
I + C D^{-1} C^T = 2\mu^\diamond A_2 \Gamma_2 A_2^T + I - \frac{dd^T}{\|d\|_2^2}.
$$

By the Sherman-Morrison-Woodbury formula [21, Section 3.8], we have

$$
C (D + C^T C)^{-1} C^T = I - (I + C D^{-1} C^T)^{-1}.
$$

Consequently, for any $(A^\diamond, y^\diamond) \in \check{S}$, it follows from that $a_i^\diamond \neq 0, \forall i$ that for each $i \in \mathcal{J}^c$,

$$
\begin{aligned}
\nabla_y q_i(A^\diamond, y^\diamond) &= (a_i^\diamond)^T (I - A_2 \cdot \nabla_y G_2(A^\diamond, y^\diamond)) = (a_i^\diamond)^T - (a_i^\diamond)^T C (D + C^T C)^{-1} C^T \\
&= (a_i^\diamond)^T (I + C D^{-1} C^T)^{-1} \neq 0.
\end{aligned}
$$

By Lemma 3.1, the zero set of $x_i^*(A, y)$ has zero measure for each $i = 1, \ldots, N$. Therefore, $|\text{supp}(x^*(A, y))| = N$ for almost all $(A, y) \in \mathbb{R}^{m \times N} \times \mathbb{R}^m$. $\qquad \square$

Putting Propositions 4.3 and 4.4 together, we obtain the following result.

**Theorem 4.4.** *Let $p > 1$ and $N \geq m$. For almost all $(A, y) \in \mathbb{R}^{m \times N} \times \mathbb{R}^m$ with $y \neq 0$, if $0 < \varepsilon < \|y\|_2$, then the unique optimal solution $x_{(A,y)}^*$ to the $BPDN_p$ (2) satisfies $|supp(x_{(A,y)}^*)| = N$.*

Next, we extend the above result to the optimization problem (3) pertaining to another version of the generalized basis pursuit denoising under a suitable assumption on $\eta$. Since its proof follows an argument similar to that for Theorem 4.4, we will be concise on the overlapping parts.

**Theorem 4.5.** *Let $p > 1$ and $N \geq m$. For almost all $(A, y) \in \mathbb{R}^{m \times N} \times \mathbb{R}^m$ with $y \in R(A)$, if $0 < \eta < \min_{Ax=y} \|x\|_p$, then the unique optimal solution $x_{(A,y)}^*$ to (3) satisfies $|supp(x_{(A,y)}^*)| = N$.*

*Proof.* We consider two cases as follows: (i) $p \geq 2$, and (ii) $1 < p < 2$.

(i) $p \geq 2$. Consider the set $S$ defined in (6). Clearly, $A$ has full row rank and $y \in R(A)$ for any $(A, y) \in S$. It follows from Proposition 3.2 that the optimal solution $x^*$ is unique and the associated unique Lagrange multiplier $\mu$ is positive. Define $\widetilde{\mu} := 1/\mu > 0$. Hence, $(x^*, \widetilde{\mu})$ is a function of $(A, y)$ on $S$ and satisfies the following equation obtained from (8):

$$F(x, \widetilde{\mu}, A, y) := \begin{bmatrix} g(x_1) + \widetilde{\mu}\, a_1^T(Ax - y) \\ \vdots \\ g(x_N) + \widetilde{\mu}\, a_N^T(Ax - y) \\ f(x) - \eta^p \end{bmatrix} = 0.$$

Clearly, $F : \mathbb{R}^N \times \mathbb{R} \times \mathbb{R}^{m \times N} \times \mathbb{R}^m \to \mathbb{R}^{N+1}$ is $C^1$ and its Jacobian with respect to $(x, \widetilde{\mu})$ is

$$\mathbf{J}_{(x,\widetilde{\mu})}F(x, \widetilde{\mu}, A, y) = \begin{bmatrix} M(x, \widetilde{\mu}, A) & A^T(Ax - y) \\ (\nabla f(x))^T & 0 \end{bmatrix},$$

where $M(x, \mu, A) := \Lambda(x) + \widetilde{\mu}A^TA$, and the diagonal matrix $\Lambda(x) := \mathrm{diag}(g'(x_1), \ldots, g'(x_N))$ is positive semi-definite. For any $(A^\diamond, y^\diamond) \in S$, we use the same notation $x^\diamond$, $\widetilde{\mu}^\diamond$, $\mathcal{J}$, $\Lambda^\diamond = \mathrm{diag}(\Lambda_1, \Lambda_2)$ and $A^\diamond = [A_1\ A_2]$ as before, where $\Lambda_1 = 0$ for $p > 2$, and $\mathrm{diag}(\Lambda_1, \Lambda_2) = 2I$ for $p = 2$. In light of $N \geq m$ and the second statement of Proposition 3.2, we have $|\mathrm{supp}(x^\diamond)| \geq N - m + 1 \geq 1$ such that the index set $\mathcal{J}$ is nonempty. It follows from (8) that $\nabla f(x^\diamond) + \widetilde{\mu}^\diamond \cdot (A^\diamond)^T(A^\diamond x^\diamond - y^\diamond) = 0$. Further, $\nabla f(x^\diamond) \neq 0$ and $(A^\diamond)^T(A^\diamond x^\diamond - y^\diamond) \neq 0$. Therefore, using the similar argument as in Proposition 4.3, we deduce that $\mathbf{J}_{(x,\widetilde{\mu})}F(x^\diamond, \widetilde{\mu}^\diamond, A^\diamond, y^\diamond)$ is invertible for any $(A^\diamond, y^\diamond) \in S$. By the implicit function theorem, there are local $C^1$ functions $G_1, G_2, H$ such that $x^* = (x^*_{\mathcal{J}^c}, x^*_{\mathcal{J}}) = (G_1(A, y), G_2(A, y)) := G(A, y)$, $\widetilde{\mu} = H(A, y)$, and $F(G(A, y), H(A, y), A, y) = 0$ for all $(A, y)$ in a neighborhood of $(A^\diamond, y^\diamond)$. By the chain rule, we have

$$\mathbf{J}_{(x,\widetilde{\mu})}F(x^\diamond, \widetilde{\mu}^\diamond, A^\diamond, y^\diamond) \cdot \begin{bmatrix} \nabla_y G_1(A^\diamond, y^\diamond) \\ \nabla_y G_2(A^\diamond, y^\diamond) \\ \nabla_y H(A^\diamond, y^\diamond) \end{bmatrix} = -\mathbf{J}_y F(x^\diamond, \widetilde{\mu}^\diamond, A^\diamond, y^\diamond) = \begin{bmatrix} \widetilde{\mu}^\diamond A_1^T \\ \widetilde{\mu}^\diamond A_2^T \\ 0 \end{bmatrix}.$$

Since $\nabla f(x^\diamond) = -\widetilde{\mu}^\diamond \cdot (A^\diamond)^T(A^\diamond x^\diamond - y^\diamond)$ and $g(x_i^\diamond) = 0, \forall\, i \in \mathcal{J}^c$, we have $A_1^T(A^\diamond x^\diamond - y^\diamond) = 0$, and

$$\mathbf{J}_{(x,\widetilde{\mu})}F(x^\diamond, \widetilde{\mu}^\diamond, A^\diamond, y^\diamond) = \begin{bmatrix} \Lambda_1 + \widetilde{\mu}^\diamond A_1^T A_1 & \widetilde{\mu}^\diamond A_1^T A_2 & 0 \\ \widetilde{\mu}^\diamond A_2^T A_1 & \Lambda_2 + \widetilde{\mu}^\diamond A_2^T A_2 & \star \\ 0 & -\widetilde{\mu}^\diamond (A^\diamond x^\diamond - y^\diamond)^T A_2 & 0 \end{bmatrix}.$$

Let $P = \begin{bmatrix} P_{11} & P_{12} & P_{13} \\ P_{21} & P_{22} & P_{23} \\ P_{31} & P_{32} & P_{33} \end{bmatrix}$ be the inverse of the above Jacobian. We consider $p > 2$ first. In this case, $\Lambda_1 = 0$ such that $P_{11}\widetilde{\mu}^\diamond A_1^T A_1 + P_{12}\widetilde{\mu}^\diamond A_2^T A_1 = I_m$. Since $B := \nabla_y G_1(A^\diamond, y^\diamond) = -[P_{11}\ P_{12}\ P_{13}] \cdot \mathbf{J}_y F(x^\diamond, \mu^\diamond, A^\diamond, y^\diamond) = P_{11}\widetilde{\mu}^\diamond A_1^T + P_{12}\widetilde{\mu}^\diamond A_2^T$, we have $\nabla_y G_1(A^\diamond, y^\diamond) \cdot A_1 = I_m$. This shows that each row of $\nabla_y G_1(A^\diamond, y^\diamond)$ is nonzero. We next consider $p = 2$, where $\mathrm{diag}(\Lambda_1, \Lambda_2) = 2I$. Hence, $2P_{11} + BA_1 = I$ and $2P_{12} + BA_2 - \widetilde{\mu}^\diamond P_{13}(A^\diamond x^\diamond - y^\diamond)^T A_2 = 0$. Suppose, by contradiction, that the $i$th row of $B$ is zero, i.e., $B_{i\bullet} = 0$. Then $(P_{11})_{i\bullet} = \mathbf{e}_i^T/2$ and $(P_{12})_{i\bullet} = (\widetilde{\mu}^\diamond(P_{13})_i)/2 \cdot (A^\diamond x^\diamond - y^\diamond)^T A_2$. Using $B_{i\bullet} = \widetilde{\mu}^\diamond[(P_{11})_{i\bullet}A_1^T + (P_{12})_{i\bullet}A_2^T]$, we have $0 = B_{i\bullet}(A^\diamond x^\diamond - y^\diamond) = \widetilde{\mu}^\diamond[(P_{11})_{i\bullet}A_1^T(A^\diamond x^\diamond - y^\diamond) + (\widetilde{\mu}^\diamond(P_{13})_i)/2 \cdot \|A_2^T(A^\diamond x^\diamond - y^\diamond)\|_2^2]$. Since $A_1^T(A^\diamond x^\diamond - y^\diamond) = 0$ and $A_2^T(A^\diamond x^\diamond - y^\diamond) \neq 0$, we have $(P_{13})_i = 0$ and $(P_{12})_{i\bullet} = 0$. Following the similar argument as in Proposition 4.3, we reach a contradiction. Thus each row of $\nabla_y G_1(A^\diamond, y^\diamond)$ is nonzero. Consequently, we deduce that $|\mathrm{supp}(x^*(A, y))| = N$ for almost all $(A, y) \in \mathbb{R}^{m \times N} \times \mathbb{R}^m$.

(ii) $1 < p < 2$. Let $\grave{S}$ be the set of $(A, y) \in \mathbb{R}^{m \times N} \times \mathbb{R}^m$ with $N \geq m$ such that $y \neq 0$, each column of $A$ is nonzero, and $A$ has full row rank. Hence, $y \in R(A)$ for any $(A, y) \in \grave{S}$. By Proposition 3.2, we see

that (3) attains a unique optimal solution $x^*$ and a unique Lagrange multiplier $\mu > 0$ for any $(A, y) \in \mathring{S}$. Define $\widetilde{\mu} := 1/\mu > 0$. Hence, $(x^*, \widetilde{\mu})$, as a function of $(A, y)$ on $\mathring{S}$, satisfies the following equation:

$$F(x, \mu, A, y) := \begin{bmatrix} x_1 + h(\widetilde{\mu} a_1^T (Ax - y)) \\ \vdots \\ x_N + h(\widetilde{\mu} a_N^T (Ax - y)) \\ f(x) - \eta^p \end{bmatrix} = 0.$$

Here $F : \mathbb{R}^N \times \mathbb{R} \times \mathbb{R}^{m \times N} \times \mathbb{R}^m \to \mathbb{R}^{N+1}$ is $C^1$ and its Jacobian with respect to $(x, \widetilde{\mu})$ is given by

$$\mathbf{J}_{(x,\widetilde{\mu})} F(x, \widetilde{\mu}, A, y) = \begin{bmatrix} V(x, \widetilde{\mu}, A, y) & \Gamma(x, \widetilde{\mu}, A, y) A^T (Ax - y) \\ (\nabla f(x))^T & 0 \end{bmatrix} \in \mathbb{R}^{(N+1) \times (N+1)},$$

where $\Gamma(x, \widetilde{\mu}, A, y) := \mathrm{diag}\big(h'(\widetilde{\mu} a_1^T (Ax - y)), \dots, h'(\widetilde{\mu} a_N^T (Ax - y))\big) \in \mathbb{R}^{N \times N}$, and $V(x, \widetilde{\mu}, A, y) := I + \Gamma(x, \widetilde{\mu}, A, y) \widetilde{\mu} A^T A$. Using the same notation introduced in Proposition 4.4, we deduce that for any $(A^\diamond, y^\diamond) \in \mathring{S}$,

$$\mathbf{J}_{(x,\widetilde{\mu})} F(x^\diamond, \widetilde{\mu}^\diamond, A^\diamond, y^\diamond) = \begin{bmatrix} I & 0 & 0 \\ \star & I + \widetilde{\mu}^\diamond \Gamma_2 A_2^T A_2 & \Gamma_2 A_2^T (A^\diamond x^\diamond - y^\diamond) \\ \star & v^T & 0 \end{bmatrix},$$

where the column vector $v := (g(x_i^\diamond))_{i \in \mathcal{J}}$. Note that the index set $\mathcal{J}$ is nonempty since $y \neq 0$ and $A$ has full row rank such that $A^T y \neq 0$. In view of (8), we have $v = -\widetilde{\mu}^\diamond A_2^T (A^\diamond x^\diamond - y^\diamond)$, where $\widetilde{\mu}^\diamond > 0$. This result, along with the similar argument for (17), shows that $\mathbf{J}_{(x,\widetilde{\mu})} F(x^\diamond, \widetilde{\mu}^\diamond, A^\diamond, y^\diamond)$ is invertible. Therefore, there are local $C^1$ functions $G_1, G_2, H$ such that $x^* = (x_{\mathcal{J}^c}^*, x_{\mathcal{J}}^*) = (G_1(A, y), G_2(A, y)) := G(A, y)$, $\widetilde{\mu} = H(A, y)$, and $F(G(A, y), H(A, y), A, y) = 0$ for all $(A, y)$ in a neighborhood of $(A^\diamond, y^\diamond)$. Moreover,

$$\mathbf{J}_{(x,\widetilde{\mu})} F(x^\diamond, \widetilde{\mu}^\diamond, A^\diamond, y^\diamond) \cdot \begin{bmatrix} \nabla_y G_1(A^\diamond, y^\diamond) \\ \nabla_y G_2(A^\diamond, y^\diamond) \\ \nabla_y H(A^\diamond, y^\diamond) \end{bmatrix} = -\mathbf{J}_y F(x^\diamond, \widetilde{\mu}^\diamond, A^\diamond, y^\diamond) = \begin{bmatrix} 0 \\ \Gamma_2 \widetilde{\mu}^\diamond A_2^T \\ 0 \end{bmatrix},$$

where the fact that $\Gamma_1 = 0$ is used. Therefore, we have $\nabla_y G_1(A^\diamond, y^\diamond) = 0$, and

$$\begin{pmatrix} \nabla_y G_2(A^\diamond, y^\diamond) \\ \nabla_y H(A^\diamond, y^\diamond) \end{pmatrix} = \begin{bmatrix} \Gamma_2^{-1} + \widetilde{\mu}^\diamond A_2^T A_2 & A_2^T (A^\diamond x^\diamond - y^\diamond) \\ (A^\diamond x^\diamond - y^\diamond)^T A_2 & 0 \end{bmatrix}^{-1} \cdot \begin{pmatrix} \widetilde{\mu}^\diamond A_2^T \\ 0 \end{pmatrix},$$

In what follows, define $b := A_2^T (A^\diamond x^\diamond - y^\diamond) \neq 0$ and $M := \Gamma_2^{-1} + \widetilde{\mu}^\diamond A_2^T A_2$, which is positive definite.

For each $i \in \mathcal{J}^c$, define $q_i(A, y) := (a_i^\diamond)^T (y - A \cdot x^*(A, y))$. It suffices to show that $\nabla_y q_i(A^\diamond, y^\diamond) \neq 0$ for each $i \in \mathcal{J}^c$, where $\nabla_y q_i(A^\diamond, y^\diamond) = (a_i^\diamond)^T (I - A_2 \cdot \nabla_y G_2(A^\diamond, y^\diamond))$. Direct calculations show that

$$I - A_2 \cdot \nabla_y G_2(A^\diamond, y^\diamond) = I - [A_2 \ 0] \cdot \begin{bmatrix} M & b \\ b^T & 0 \end{bmatrix}^{-1} \cdot \begin{pmatrix} \widetilde{\mu}^\diamond A_2^T \\ 0 \end{pmatrix} = I - A_2 M^{-1} \Big( I - \frac{bb^T M^{-1}}{b^T M^{-1} b} \Big) \widetilde{\mu}^\diamond A_2^T$$

$$= I - A_2 M^{-1} \widetilde{\mu}^\diamond A_2^T + \widetilde{\mu}^\diamond \frac{A_2 M^{-1} bb^T M^{-1} A_2^T}{b^T M^{-1} b}.$$

By the definition of $M$ and the Sherman-Morrison-Woodbury formula [21, Section 3.8], we have $I - A_2 M^{-1} \widetilde{\mu}^\diamond A_2^T = (I + \widetilde{\mu}^\diamond A_2 \Gamma_2 A_2^T)^{-1}$, which is positive definite. Hence, $I - A_2 \cdot \nabla_y G_2(A^\diamond, y^\diamond)$ is positive definite and thus invertible. Since $a_i^\diamond \neq 0$, we have $\nabla_y q_i(A^\diamond, y^\diamond) \neq 0$ for each $i \in \mathcal{J}^c$. Consequently, $|\mathrm{supp}(x^*(A, y))| = N$ for almost all $(A, y) \in \mathbb{R}^{m \times N} \times \mathbb{R}^m$. $\qquad \square$

## 5 Extensions and Comparison

This section extends the least sparsity results to constrained measurement vectors, and compares these results for $p > 1$ with those from $\ell_p$ minimization for $0 < p \leq 1$; the complex setting is also considered.

## 5.1 Extensions to Constrained Measurement Vectors and Noisy Case

In the previous sections, we consider general measurement vectors in $\mathbb{R}^m$. However, in many applications such as compressed sensing, a measurement vector $y$ is restricted to a proper subspace of $R(A)$, to which the results in Section 4 are not applicable since this subspace may have dimension less than $m$ so that it has zero measure in $\mathbb{R}^m$. In what follows, we extend the least sparsity results in Section 4 to this scenario. For simplicity, we consider the generalized basis pursuit $\mathrm{BP}_p$ (1) with $p > 1$ only, although its result can be extended to the other generalized optimization problems, e.g., the $\mathrm{BPDN}_p$, $\mathrm{RR}_p$, and $\mathrm{EN}_p$; see Remark 5.2 for the discussions on the generalized ridge regression.

**Theorem 5.1.** *Let $p > 1$, $N \geq 2m - 1$, and $\mathcal{I} \subseteq \{1, \ldots, N\}$ be a nonempty index set. Then there exists a set $S_A \subset \mathbb{R}^{m \times N}$ whose complement has zero measure such that for each fixed $A \in S_A$, the unique optimal solution $x^*$ to the $\mathrm{BP}_p$ (1) satisfies $|supp(x^*)| = N$ for almost all $y \in R(A_{\bullet \mathcal{I}})$.*

*Proof.* For the given $p > 1$, $N, m \in \mathbb{N}$ with $N \geq 2m - 1$ and the index set $\mathcal{I}$, we consider two cases: (i) $|\mathcal{I}| \geq m$; and (ii) $|\mathcal{I}| < m$. For the first case, let $S_A$ be the set of all $A \in \mathbb{R}^{m \times N}$ such that any $m \times m$ submatrix of $A$ is invertible. Clearly, the complement of $S_A$ has zero measure in the space $\mathbb{R}^{m \times N}$. Further, since $|\mathcal{I}| \geq m$, $R(A_{\bullet \mathcal{I}}) = \mathbb{R}^m$ for any $A \in S_A$. Hence, it follows from Corollary 4.1 that the proposition holds.

We then consider the second case where $|\mathcal{I}| < m$. In this case, let $\widetilde{S}_A$ be the set of all $A \in \mathbb{R}^{m \times N}$ satisfying the following condition: for any index set $\mathcal{J}$ with $|\mathcal{J}| = |\mathcal{I}| := r$, the $r \times r$ matrix $(A_{\bullet \mathcal{I}})^T \cdot A_{\bullet \mathcal{J}}$ is invertible. Note that for any index set $\mathcal{J}$ with $|\mathcal{J}| = r$, $\det((A_{\bullet \mathcal{I}})^T A_{\bullet \mathcal{J}}) = 0$ gives rise to a polynomial equation of the elements of $A$. Hence, we deduce that the complement of $\widetilde{S}_A$ has zero measure in $\mathbb{R}^{m \times N}$. Further, for any $A \in \widetilde{S}_A$, the columns of $A_{\bullet \mathcal{I}}$ must be linearly independent. Therefore, for any $y \in R(A_{\bullet \mathcal{I}})$, there exists a unique $z_y \in \mathbb{R}^r$ such that $A_{\bullet \mathcal{I}} \cdot z_y = y$. This shows that the constraint $Ax = y$ in the $\mathrm{BP}_p$ (1) can be equivalently written as

$$\left[(A_{\bullet \mathcal{I}})^T A_{\bullet \mathcal{I}}\right]^{-1} \cdot (A_{\bullet \mathcal{I}})^T \cdot Ax = z_y.$$

Define the $r \times N$ matrix $\widetilde{A} := [(A_{\bullet \mathcal{I}})^T A_{\bullet \mathcal{I}}]^{-1} (A_{\bullet \mathcal{I}})^T A$ for each $A \in \widetilde{S}_A$. It follows from the property of $A \in \widetilde{S}_A$ that any $r \times r$ submatrix of $\widetilde{A}$ is invertible. Hence, for any $A \in \widetilde{S}_A$ and any $y \in R(A_{\bullet \mathcal{I}})$, the original $\mathrm{BP}_p$ (1) is converted to the following equivalent optimization problem: for some $z \in \mathbb{R}^r$,

$$\min_{x \in \mathbb{R}^N} \ \|x\|_p \quad \text{subject to} \quad \widetilde{A}x = z. \tag{19}$$

For a fixed $\widetilde{A}$ obtained from a given $A \in \widetilde{S}_A$, by applying Corollary 4.1 to (19), we deduce that $|\mathrm{supp}(x^*(z))| = N$ for almost all $z \in \mathbb{R}^r$. Since $A_{\bullet \mathcal{I}}$ has full column rank, the same conclusion holds for almost all $y \in R(A_{\bullet \mathcal{I}})$. $\qquad \square$

The following corollary can be easily established with the aid of Theorem 5.1 and the extension of Proposition 3.1 to (19); its proof is thus omitted.

**Corollary 5.1.** *Let $p > 1$, $N \geq 2m - 1$, and $s \in \{1, \ldots, N\}$. The following hold:*

(i) *There exists a set $S_A \subset \mathbb{R}^{m \times N}$ whose complement has zero measure such that for each fixed $A \in S_A$ and any index set $\mathcal{I}$ with $|\mathcal{I}| \leq s$, the unique optimal solution $x^*$ to the $\mathrm{BP}_p$ (1) satisfies $|supp(x^*)| \geq N - m + 1$ for any nonzero $y \in R(A_{\bullet \mathcal{I}})$;*

(ii) *There exists a set $\widehat{S}_A \subset \mathbb{R}^{m \times N}$ whose complement has zero measure such that for each fixed $A \in \widehat{S}_A$ and any index set $\mathcal{I}$ with $|\mathcal{I}| \leq s$, the unique optimal solution $x^*$ to the $\mathrm{BP}_p$ (1) satisfies $|supp(x^*)| = N$ for almost all $y \in R(A_{\bullet \mathcal{I}})$.*

**Remark 5.1.** The least sparsity results can be extended to the case where measurement vectors are polluted by noise or errors. Specifically, consider the measurement vector $y = w + e$, where $w \in R(A)$ and $e \in \mathbb{R}^m$ denotes noise or an error. It follows from Corollary 4.1 that for a given $A \in \mathbb{R}^{m \times N}$ satisfying a suitable condition stated in Corollary 4.1 and any given $w \in R(A)$, the optimal solution $x^*_{(A,y)}$ to the $\text{BP}_p$ (1) has full support for almost all $e \in \mathbb{R}^m$. For comparison, see relevant results on robust sparse recovery using $\ell_1$-norm based basis pursuit denoising [3] and $\ell_p$-minimization with $0 < p < 1$ [25, 29].

## 5.2 Comparison with $p$-norm based Optimization with $0 < p \leq 1$

For a given sparsity level $s$ with $1 \leq s \leq N$ (especially $s \ll N$), we call a vector $x \in \mathbb{R}^N$ $s$-sparse if $|\text{supp}(x)| \leq s$. Furthermore, we call a measurement vector $y$ generated by an $s$-sparse vector if there is an $s$-sparse vector such that $y = Ax$. Using these terminologies, we see that Corollary 5.1 states that when $p > 1$, for almost all $A \in \mathbb{R}^{m \times N}$ with $N \gg \max(m, s)$ and almost all $y$ generated by $s$-sparse vectors, the optimal solution $x^*$ to the $\text{BP}_p$ (1) is far from sparse, i.e., $|\text{supp}(x^*)| \gg s$. Equivalently, it means that when $p > 1$, the $\text{BP}_p$ (1) *might* recover a sparse vector $x$ from $y = Ax$ only for a set of $A$'s of zero measure in $\mathbb{R}^{m \times N}$, no matter how large $N$ and $m$ are. Moreover, an arbitrarily small perturbation to a measurement matrix $A$ in this zero measure set will lead to a least sparse solution. This shows the extremely weak robustness of the $\text{BP}_p$ (1) with $p > 1$ in term of solution sparsity.

For comparison, it is interesting to ask what happens to the $\text{BP}_p$ (1) when $0 < p \leq 1$. We show below that when $0 < p \leq 1$, there exists a non-zero measure set of $A$'s such that the $\text{BP}_p$ (1) recovers any sparse vector $x$ from $y = Ax$. It also demonstrates the strong robustness of the $\text{BP}_p$ (1) for $0 < p \leq 1$. Toward this end, recall that an $m \times N$ matrix $A$ satisfies the *restricted isometry property* (RIP) of order $k$ if there is a constant $\delta_k \in (0, 1)$ such that $(1 - \delta_k)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + \delta_k)\|x\|_2^2$ for all $k$-sparse vectors $x \in \mathbb{R}^N$.

**Proposition 5.1.** *Fix $p \in (0, 1]$, $\gamma \in (0, 1)$ and $s \in \mathbb{N}$. Suppose $m = \lceil \gamma N \rceil$. Then for all $N$ sufficiently large, there exists an open set $U_A \subset \mathbb{R}^{m \times N}$ such that for any $A \in U_A$ and any index set $\mathcal{I}$ with $|\mathcal{I}| \leq s$, the optimal solution $x^*$ to the $\text{BP}_p$ (1) satisfies $|supp(x^*)| = |\mathcal{I}|$ for all $y \in R(A_{\bullet\mathcal{I}})$.*

*Proof.* Consider $p = 1$ first, which corresponds to the $\ell_1$-optimization based basis pursuit [14]. For the given constants $\gamma \in (0, 1)$, $\delta_{3s} \in (0, 1/3)$, and the sparsity level $s$, it is known via random matrix argument that for all $N$ sufficiently large with $s \ll m$, there exists a matrix $A^\diamond \in \mathbb{R}^{m \times N}$ which satisfies the RIP of order $3s$ with constant $\delta_{3s}(A^\diamond) < 1/3$ [1, 4]. Hence, the $\text{BP}_p$ (1) recovers any $s$-sparse vector $x$ exactly from $y = A^\diamond x$ [14, Theorem 6.9] or [4]. Furthermore, in view of $\left| \|Ax\|_2 - \|A^\diamond x\|_2 \right| \leq \|A - A^\diamond\|_2 \cdot \|x\|_2$ for any $A$ and $x$, we see that there exists $\eta > 0$ such that $\delta_{3s}(A) < 1/3$ for all $A$'s with $\|A - A^\diamond\|_2 < \eta$. Let the open set $U_A := \{A \in \mathbb{R}^{m \times N} \mid \|A - A^\diamond\|_2 < \eta\}$. This shows that for any $A \in U_A$, every $s$-sparse vector $x$ can be recovered from $y = Ax$ via the $\text{BP}_p$ (1). Finally, when $0 < p < 1$, it follows from [14, Theorem 4.10] that for any $A \in U_A$, the $\text{BP}_p$ (1) recovers any $s$-sparse vector $x$ from $y = Ax$. $\square$

**Remark 5.2.** Consider the ridge regression $\text{RR}_p$ (4), and we compare the sparsity property for $p > 1$ with that for $0 < p < 1$. It is shown in [9, Theorem 2.1(2)] that when $0 < p < 1$, for any $A \in \mathbb{R}^{m \times N}$, $y \in \mathbb{R}^m$, and $\lambda > 0$, any (local/global) optimal solution $x^*$ to the $\text{RR}_p$ (4) satisfies $|\text{supp}(x^*)| \leq m$. In contrast, Theorem 4.2 shows that when $p > 1$, an optimal solution $x^*$ to (4) has full support, i.e., $|\text{supp}(x^*)| = N$, for almost all $A$ and $y$.

## 5.3 Extension to Complex Measurement Matrices and Vectors

This subsection extends the previous results for the real setting to the complex setting, i.e., $(A, y) \in \mathbb{C}^{m \times N} \times \mathbb{C}^m$. In the latter setting, each of the problems $\text{BP}_p$ (1), $\text{BPDN}_p$ (2) and (3), $\text{RR}_p$ (4), and $\text{EN}_p$ (5) seeks a complex optimal solution $x^* \in \mathbb{C}^N$, which is also unique under the similar conditions stated in

Propositions 2.1 and 3.2. Let $\imath$ denote the imaginary unit. For a complex matrix $A = A_{\mathrm{R}} + \imath A_{\mathrm{I}} \in \mathbb{C}^{m \times N}$ with $A_{\mathrm{R}}, A_{\mathrm{I}} \in \mathbb{R}^{m \times N}$, define the real matrix $\widetilde{A} := [\widetilde{A}_{1:2}, \ldots, \widetilde{A}_{2N-1:2N}] \in \mathbb{R}^{2m \times 2N}$, where

$$\widetilde{A}_{2k-1:2k} := \begin{bmatrix} (A_{\mathrm{R}})_{\bullet k} & -(A_{\mathrm{I}})_{\bullet k} \\ (A_{\mathrm{I}})_{\bullet k} & (A_{\mathrm{R}})_{\bullet k} \end{bmatrix} \in \mathbb{R}^{2m \times 2}, \qquad \forall\, k = 1, \ldots, N.$$

Here $(A_{\mathrm{R}})_{\bullet k}$ and $(A_{\mathrm{I}})_{\bullet k}$ denote the $k$-th columns of $A_{\mathrm{R}}$ and $A_{\mathrm{I}}$ respectively. For a complex $N$-vector $x = u + \imath v$ with $u, v \in \mathbb{R}^N$, define $\widetilde{x} := [\widetilde{x}_{1:2}; \ldots; \widetilde{x}_{2N-1:2N}] \in \mathbb{R}^{2N}$, where $\widetilde{x}_{2k-1:2k} := (u_k, v_k)^T \in \mathbb{R}^2$ for each $k = 1, \ldots, N$. Similarly, we define $\widetilde{y} \in \mathbb{R}^{2m}$ for $y \in \mathbb{C}^m$. Note that $\mathrm{supp}(x) = \{k \mid \widetilde{x}_{2k-1:2k} \neq 0\}$ (but $\mathrm{supp}(x) \neq \mathrm{supp}(\widetilde{x})$). Based on the definitions of $\widetilde{A}, \widetilde{x}$ and $\widetilde{y}$, the following facts can be easily established: (i) $\|Ax - y\|_2^2 = \|\widetilde{A}\widetilde{x} - \widetilde{y}\|_2^2$, and $Ax = y$ if and only if $\widetilde{A}\widetilde{x} = \widetilde{y}$; (ii) $\|x\|_p^p = \sum_{k=1}^N |x_k|^p = \sum_{k=1}^N (u_k^2 + v_k^2)^{p/2} = \sum_{k=1}^N \|\widetilde{x}_{2k-1:2k}\|_2^p$; and (iii) for an index subset $\mathcal{I} \subseteq \{1, \ldots, N\}$, the columns of $A_{\bullet \mathcal{I}}$ are linearly independent (over the complex field $\mathbb{C}$) if and only if the columns of the matrix $[\widetilde{A}_{2k-1:2k}]_{k \in \mathcal{I}} \in \mathbb{R}^{2m \times 2|\mathcal{I}|}$ are linearly independent (over the real field $\mathbb{R}$).

Let $p > 1$. Define the functions $\widetilde{g} : \mathbb{R}^2 \to \mathbb{R}^2$ and $\widetilde{h} : \mathbb{R}^2 \to \mathbb{R}^2$ as: for any $z = (z_1, z_2)^T \in \mathbb{R}^2$,

$$\widetilde{g}(z) := \begin{cases} p \cdot \|z\|_2^{p-2} \cdot z & \text{if } z \neq 0; \\ 0 & \text{if } z = 0 \end{cases}, \qquad \widetilde{h}(z) := \begin{cases} \frac{1}{p^{1/(p-1)}} \cdot \|z\|_2^{\frac{2-p}{p-1}} \cdot z & \text{if } z \neq 0; \\ 0 & \text{if } z = 0 \end{cases}. \tag{20}$$

These functions are analogous to their counterparts defined in (9) in the real setting. It is easy to verify that $\widetilde{h}$ is the inverse function of $\widetilde{g}$ and that $\widetilde{g}$ and $\widetilde{h}$ are positively homogeneous of degrees $p - 1$ and $1/(p-1)$ respectively. Letting $f(\widetilde{x}) := \sum_{k=1}^N \|\widetilde{x}_{2k-1:2k}\|_2^p$ where $\widetilde{x} = [\widetilde{x}_{1:2}; \ldots; \widetilde{x}_{2N-1:2N}] \in \mathbb{R}^{2N}$, then $\nabla f(\widetilde{x}) = [\widetilde{g}(\widetilde{x}_{1:2}); \ldots; \widetilde{g}(\widetilde{x}_{2N-1:2N})]$. Additional properties of $\widetilde{g}$ and $\widetilde{h}$ are given in the following lemma.

**Lemma 5.1.** *When $p \geq 2$, $\widetilde{g}$ is continuously differentiable on $\mathbb{R}^2$; when $p > 2$, its Jacobian $\mathbf{J}\widetilde{g}(z)$ is positive definite at any $z \neq 0$ and $\mathbf{J}\widetilde{g}(0) = 0$. When $1 < p \leq 2$, $\widetilde{h}$ is continuously differentiable on $\mathbb{R}^2$; when $1 < p < 2$, its Jacobian $\mathbf{J}\widetilde{h}(z)$ is positive definite at any $z \neq 0$ and $\mathbf{J}\widetilde{h}(0) = 0$.*

*Proof.* When $p = 2$, $\mathbf{J}\widetilde{g}(z) = 2I$ and $\mathbf{J}\widetilde{h}(z) = I/2$ for all $z \in \mathbb{R}^2$. A straightforward but lengthy computation shows that (i) when $p > 2$, $\mathbf{J}\widetilde{g}(0) = 0$, and for any $z = (z_1, z_2)^T \neq 0$,

$$\mathbf{J}\widetilde{g}(z) = p \cdot \|z\|_2^{p-4} \cdot \begin{bmatrix} (p-1)z_1^2 + z_2^2 & (p-2)z_1 z_2 \\ (p-2)z_1 z_2 & z_1^2 + (p-1)z_2^2 \end{bmatrix} = p \cdot \|z\|_2^{p-2} \cdot U^T \operatorname{diag}(p-1, 1)\, U,$$

where the orthogonal matrix $U := \|z\|_2^{-1} \cdot \begin{bmatrix} z_1 & -z_2 \\ z_2 & z_1 \end{bmatrix}$; and (ii) when $1 < p < 2$, $\mathbf{J}\widetilde{h}(0) = 0$, and

$$\mathbf{J}\widetilde{h}(z) = \frac{1}{p^{1/(p-1)}} \cdot \|z\|_2^{\frac{2-p}{p-1}-2} \cdot \begin{bmatrix} \frac{z_1^2}{p-1} + z_2^2 & \frac{2-p}{p-1} z_1 z_2 \\ \frac{2-p}{p-1} z_1 z_2 & z_1^2 + \frac{z_2^2}{p-1} \end{bmatrix} = \frac{1}{p^{1/(p-1)}} \cdot \|z\|_2^{\frac{2-p}{p-1}} \cdot U^T \operatorname{diag}\left(\frac{1}{p-1}, 1\right) U$$

for any $z = (z_1, z_2)^T \neq 0$. It follows from the above results that $\mathbf{J}\widetilde{g}(z), \mathbf{J}\widetilde{h}(z)$ are positive definite at any $z \neq 0$, and $\widetilde{g}, \widetilde{h}$ are continuously differentiable on $\mathbb{R}^2$. $\qquad \square$

Define the following set in $\mathbb{C}^{m \times N} \times \mathbb{C}^m$ with $N \geq m$ which is analogous to the set $S$ defined in (6):

$$S_{\mathbb{C}} := \left\{ (A, y) \in \mathbb{C}^{m \times N} \times \mathbb{C}^m \mid \text{every } m \times m \text{ submatrix of } A \text{ is invertible, and } y \neq 0 \right\}. \tag{21}$$

In many important applications such as compressed sensing, a complex matrix $A$ satisfying the condition specified in $S_{\mathbb{C}}$ can be obtained by uniformly at random choosing from a $r \times r$ Fourier matrix for a prime number $r$ [11] using row-repetition free Bernoulli selectors or a random subset model [3]. The following result extends Proposition 3.1 to the complex setting.

23

**Proposition 5.2.** *Let $p > 1$. For any $(A, y) \in S_{\mathbb{C}}$, the following hold:*

    (i) *The minimizer $x^* \in \mathbb{C}^N$ of the $BP_p$ (1) satisfies $|supp(x^*)| \geq N - m + 1$;*

    (ii) *If $0 < \varepsilon < \|y\|_2$, then the minimizer $x^* \in \mathbb{C}^N$ of the $BPDN_p$ (2) satisfies $|supp(x^*)| \geq N - m + 1$;*

    (iii) *For any $r > 0$, $\lambda_1 > 0$ and $\lambda_2 \geq 0$, each nonzero minimizer $x^* \in \mathbb{C}^N$ of the $EN_p$ (5) satisfies $|supp(x^*)| \geq N - m + 1$;*

    (iv) *For any $\lambda > 0$, the minimizer $x^* \in \mathbb{C}^N$ of the $RR_p$ (4) satisfies $|supp(x^*)| \geq N - m + 1$;*

    (v) *If $0 < \eta < \min_{Ax=y} \|x\|_p$, then the unique minimizer $x^*$ of (3) satisfies $|supp(x^*)| \geq N - m + 1$.*

*Proof.* For each $(A, y) \in S_{\mathbb{C}}$, let $(\widetilde{A}, \widetilde{y}) \in \mathbb{R}^{2m \times 2N} \times \mathbb{R}^{2m}$ be defined before, where $\widetilde{A} = [\widetilde{A}_{1:2}, \ldots, \widetilde{A}_{2N-1:2N}]$. Hence, $\widetilde{y} \neq 0$, and $[\widetilde{A}_{2k-1:2k}]_{k \in \mathcal{I}}$ is invertible for any index set $\mathcal{I}$ with $|\mathcal{I}| = m$. For any $x \in \mathbb{C}^N$ satisfying $Ax = y$, $\widetilde{x} = [\widetilde{x}_{1:2}; \ldots; \widetilde{x}_{2N-1:2N}] \in \mathbb{R}^{2N}$ satisfies $\widetilde{A}\widetilde{x} = \widetilde{y}$ and $supp(x) = \{k \mid \widetilde{x}_{2k-1:2k} \neq 0\}$. Recall that $f(\widetilde{x}) = \sum_{k=1}^{N} \|\widetilde{x}_{2k-1:2k}\|_2^p = \|x\|_p^p$. We sketch the proofs for (i)-(v) as follows.

    (i) The vector $\widetilde{x}^* \in \mathbb{R}^{2N}$ associated with the unique nonzero minimizer $x^* \in \mathbb{C}^N$ of the $BP_p$ (1) satisfies the KKT condition: $\nabla f(\widetilde{x}^*) - \widetilde{A}^T \nu = 0$, and $\widetilde{A}\widetilde{x}^* = \widetilde{y}$, where $\nu \in \mathbb{R}^{2m}$ is the Lagrange multiplier, and $\nabla f(\widetilde{x}^*) = [\widetilde{g}(\widetilde{x}_{1:2}^*); \ldots; \widetilde{g}(\widetilde{x}_{2N-1:2N}^*)]$ with $\widetilde{g}$ defined in (20). Suppose $x^*$ has at least $m$ nonzero elements. Then there exists an index set $\mathcal{I}$ with $|\mathcal{I}| = m$ such that $\widetilde{x}_{2k-1:2k}^* = 0$ for each $k \in \mathcal{I}$. By the properties of $\widetilde{g}$, we have $\widetilde{g}(\widetilde{x}_{2k-1:2k}^*) = 0$ for all $k \in \mathcal{I}$. Since $[\widetilde{A}_{2k-1:2k}]_{k \in \mathcal{I}}$ is invertible, it follows from the KKT condition that $\nu = 0$ so that $\nabla f(\widetilde{x}^*) = 0$. Therefore, $\widetilde{x}^* = 0$ or equivalently $x^* = 0$, contradiction.

    (ii) Define $\theta(\widetilde{x}) := \|\widetilde{A}\widetilde{x} - \widetilde{y}\|_2^2 - \varepsilon^2$ for any $\widetilde{x} \in \mathbb{R}^{2N}$ associated with $x \in \mathbb{C}^N$. Hence, $\theta(\widetilde{x}) = \|Ax - y\|_2^2 - \varepsilon^2$. By a similar argument for (ii) of Proposition 3.1, we deduce that the vector $\widetilde{x}^* \in \mathbb{R}^{2N}$ associated with the unique nonzero minimizer $x^* \in \mathbb{C}^N$ of the $BPDN_p$ (2) satisfies the KKT condition: $\nabla f(\widetilde{x}^*) + \mu \nabla \theta(\widetilde{x}^*) = 0$ and $0 \leq \mu \perp \theta(\widetilde{x}^*) \geq 0$, where the multiplier $\mu > 0$. It follows from $\nabla \theta(\widetilde{x}^*) = 2\widetilde{A}^T(\widetilde{A}\widetilde{x}^* - \widetilde{y})$ and a similar argument for (ii) of Proposition 3.1 that $|supp(x^*)| \geq N - m + 1$.

    (iii) For an arbitrary $r > 0$, we have $\|x\|_p^r = [f(\widetilde{x})]^{r/p}$. Hence, the $EN_p$ (5) is equivalent to $\min_{\widetilde{x} \in \mathbb{R}^{2N}} \frac{1}{2}\|\widetilde{A}\widetilde{x} - \widetilde{y}\|_2^2 + \lambda_1[f(\widetilde{x})]^{r/p} + \lambda_2\|\widetilde{x}\|_2^2$. For a nonzero minimizer $x^*$, its corresponding $\widetilde{x}^*$ is also nonzero and satisfies the optimality condition: $\widetilde{A}^T(\widetilde{A}\widetilde{x}^* - \widetilde{y}) + \lambda_1(r/p)[f(\widetilde{x}^*)]^{(r-p)/p}\nabla f(\widetilde{x}^*) + 2\lambda_2\widetilde{x}^* = 0$. Applying the similar argument for (iv) of Proposition 3.1 leads to $|supp(x^*)| \geq N - m + 1$.

    (iv)-(v) These proofs are omitted as they resemble those for (ii) and Proposition 3.2 respectively. $\square$

**Theorem 5.2.** *Let $p > 1$. The following hold for almost all $(A, y) \in \mathbb{C}^{m \times N} \times \mathbb{C}^m$:*

    (i) *Let $N \geq 2m - 1$. The unique minimizer $x^* \in \mathbb{C}^N$ of the $BP_p$ (1) satisfies $|supp(x^*)| = N$;*

    (ii) *Let $N \geq m$ and $\lambda > 0$. The unique minimizer $x^* \in \mathbb{C}^N$ of the $RR_p$ (4) satisfies $|supp(x^*)| = N$;*

    (iii) *Let $N \geq m$. If $y \neq 0$ and $0 < \varepsilon < \|y\|_2$, then the unique minimizer $x^* \in \mathbb{C}^N$ of the $BPDN_p$ (2) satisfies $|supp(x^*)| = N$;*

    (iv) *Let $N \geq m$, $r \geq 1$, and $\lambda_1, \lambda_2 > 0$. The unique minimizer $x^* \in \mathbb{C}^N$ of the $EN_p$ (5) satisfies $|supp(x^*)| = N$;*

    (v) *Let $N \geq m$. If $y \in R(A)$ and $0 < \eta < \min_{Ax=y} \|x\|_p$, then the unique minimizer $x^* \in \mathbb{C}^N$ of (3) satisfies $|supp(x^*)| = N$.*

*Proof.* Letting $\widetilde{y} \in \mathbb{R}^{2m}$ be the unique correspondence of $y \in \mathbb{C}^m$ defined above, we define the set $\mathcal{S} := \{(A_{\mathrm{R}}, A_{\mathrm{I}}, \widetilde{y}) \mid (A_{\mathrm{R}} + \imath A_{\mathrm{I}}, y) \in S_{\mathbb{C}}\}$. Clearly, $\mathcal{S}$ is open and its complement has zero measure in $\mathbb{R}^{m \times N} \times \mathbb{R}^{m \times N} \times \mathbb{R}^{2m}$. For any $A = A_{\mathrm{R}} + \imath A_{\mathrm{I}}$ given from $S_{\mathbb{C}}$, we often write $\widetilde{A}(A_{\mathrm{R}}, A_{\mathrm{I}})$ as $\widetilde{A}$ to simplify notation when the context is clear. For any (unique) minimizer $x^* \in \mathbb{C}^N$ associated with $(A, y) \in \mathcal{S}_{\mathbb{C}}$

in each problem, let $\mathcal{J} := \{k \mid x_k^* \neq 0\} = \{k \mid \widetilde{x}_{2k-1:2k}^* \neq 0\} \subseteq \{1, \ldots, N\}$. Partition $\widetilde{A}$ as $\widetilde{A} = [\widetilde{A}_1 \; \widetilde{A}_2]$, where $\widetilde{A}_1 := [\widetilde{A}_{2k-1:2k}]_{k \in \mathcal{J}^c}$ and $\widetilde{A}_2 := [\widetilde{A}_{2k-1:2k}]_{k \in \mathcal{J}}$. Note that $\widetilde{A}$ has full row rank and $|\mathcal{J}^c| \leq m - 1$ by Proposition 5.2. Hence, the columns of $A_{\bullet \mathcal{J}^c}$ are linearly independent, and so are the columns of $\widetilde{A}_1$.

(i) Consider $p > 2$ first. For any $(A_R, A_I, \widetilde{y}) \in \mathcal{S}$, $(\widetilde{x}^*, \nu^*) \in \mathbb{R}^{2N} \times \mathbb{R}^{2m}$ is a unique solution to the equation $F(\widetilde{x}, \nu, A_R, A_I, \widetilde{y}) := [\nabla f(\widetilde{x}) - \widetilde{A}^T \nu; \widetilde{A}\widetilde{x} - \widetilde{y}] = 0$. Let $\Lambda(\widetilde{x}^*) := \text{diag}(\mathbf{J}\widetilde{g}(\widetilde{x}_{1:2}^*), \ldots, \mathbf{J}\widetilde{g}(\widetilde{x}_{2N-1:2N}^*)) = \text{diag}(\Lambda_1, \Lambda_2)$, where $\Lambda_1 := \text{diag}(\mathbf{J}\widetilde{g}(\widetilde{x}_{2k-1:2k}^*))_{k \in \mathcal{J}^c} = 0$, $\Lambda_2 := \text{diag}(\mathbf{J}\widetilde{g}(\widetilde{x}_{2k-1:2k}^*))_{k \in \mathcal{J}}$, and $\mathbf{J}\widetilde{g}(\widetilde{x}_{2k-1:2k}^*)$ is positive definite for any $k \in \mathcal{J}$. Using this result, we can show that $\mathbf{J}_{(\widetilde{x}, \nu)} F(\widetilde{x}^*, \nu^*, A_R, A_I, \widetilde{y})$ is invertible and $\widetilde{x}^*(A_R, A_I, \widetilde{y})$ is a local $C^1$ function. Following a similar argument for Proposition 4.1, it can be further shown that each row of $\nabla_{\widetilde{y}} x_{\mathcal{J}^c}^*(A_R, A_I, \widetilde{y})$ is nonzero. This yields the desired result.

Consider $1 < p \leq 2$. In this case, $\widetilde{x}_{2k-1:2k}^* = \widetilde{h}(\widetilde{A}_{2k-1:2k}^T \nu)$ for each $k = 1, \ldots, N$, where the multiplier $\nu \in \mathbb{R}^{2m}$ satisfies the equation $F(\nu, A_R, A_I, \widetilde{y}) := \sum_{k=1}^{N} \widetilde{A}_{2k-1:2k} \cdot \widetilde{h}(\widetilde{A}_{2k-1:2k}^T \nu) - \widetilde{y} = 0$. Then $Q := \mathbf{J}_\nu F(\nu, A_R, A_I, \widetilde{y}) = \widetilde{A} \Theta \widetilde{A}^T$, where $\Theta := \text{diag}(\mathbf{J}\widetilde{h}(\widetilde{A}_{1:2}^T \nu), \ldots, \mathbf{J}\widetilde{h}(\widetilde{A}_{2N-1:2N}^T \nu))$, and $\widetilde{A}$ has full row rank. When $p = 2$, $\Theta = I/2$ so that $Q$ is positive definite. When $1 < p < 2$, we have $w^T Q w = \sum_{k=1}^{N} (\widetilde{A}_k^T w)^T \cdot \mathbf{J}\widetilde{h}(\widetilde{A}_{2k-1:2k}^T \nu) \cdot (\widetilde{A}_{2k-1:2k}^T w)$ for any $w \in \mathbb{R}^{2m}$. By Lemma 5.1, the property of $\mathcal{S}$, and a similar argument for Proposition 4.2, we see that $Q$ is positive definite for $1 < p < 2$. Since each column of $A$ from the set $S_\mathbb{C}$ is nonzero, each column of $\widetilde{A}$ is also nonzero. This, along with a similar argument for Proposition 4.2, shows that the gradient $\nabla_{\widetilde{y}} \widetilde{A}_{\bullet i}^T \nu(A_R, A_I, \widetilde{y}) = \widetilde{A}_{\bullet i}^T Q^{-1} \neq 0$ for each $i = 1, \ldots, 2N$ at any $(A_R, A_I, \widetilde{y}) \in \mathcal{S}$. In light of $\text{sgn}(\widetilde{x}_i^*) = \text{sgn}(\widetilde{A}_{\bullet i}^T \nu)$ for each $i = 1, \ldots, 2N$, the desired result follows.

(ii) Consider $p > 2$. The vector $\widetilde{x}^* \in \mathbb{R}^{2N}$ associated with the unique minimizer $x^* \in \mathbb{C}^N$ satisfies the equation: $F(\widetilde{x}, A_R, A_I, \widetilde{y}) := \lambda \nabla f(\widetilde{x}) + \widetilde{A}^T(\widetilde{A} - \widetilde{y}) = 0$, where $\mathbf{J}_{\widetilde{x}} F(\widetilde{x}, A_R, A_I, \widetilde{y}) = \lambda D(\widetilde{x}) + \widetilde{A}^T \widetilde{A}$, and $D(\widetilde{x})$ is a block diagonal matrix. Partition $D(\widetilde{x}) = \text{diag}(D_1, D_2)$ as before, where $D_1 = 0$, $D_2$ is positive definite, and $\widetilde{A}_1$ has full column rank. Using these results and a similar argument for Case (i) in Theorem 4.2, we have that $\mathbf{J}_{\widetilde{x}} F(\widetilde{x}^*, A_R, A_I, \widetilde{y})$ is positive definite and each row of $\nabla_{\widetilde{y}} \widetilde{x}_{\mathcal{J}^c}^*(A_R, A_I, \widetilde{y})$ is nonzero at each $(A_R, A_I, \widetilde{y}) \in \mathcal{S}$. The case of $1 < p \leq 2$ can be shown via a similar but lengthy computation and is thus omitted.

(iii) Consider $p \geq 2$ first. For any $(A_R, A_I, \widetilde{y}) \in \mathcal{S}$, $(\widetilde{x}^*, \mu) \in \mathbb{R}^{2N} \times \mathbb{R}$ is a unique solution to the equation $F(\widetilde{x}, \mu, A_R, A_I, \widetilde{y}) := [\nabla f(\widetilde{x}) + 2\mu \widetilde{A}^T(\widetilde{A}\widetilde{x} - \widetilde{y}); \|\widetilde{A}\widetilde{x} - \widetilde{y}\|_2^2 - \varepsilon] = 0$, where $\mu > 0$. It can be shown that $\mathbf{J}_{(\widetilde{x}, \mu)} F(\widetilde{x}, \mu, A_R, A_I, \widetilde{y})$ is invertible and each row of $\nabla_{\widetilde{y}} \widetilde{x}_{\mathcal{J}^c}^*(A_R, A_I, \widetilde{y})$ is nonzero at each $(A_R, A_I, \widetilde{y}) \in \mathcal{S}$. When $1 < p < 2$, it can be shown via a similar argument for Proposition 4.4 that $\widetilde{x}^*(A_R, A_I, \widetilde{y})$ is a local $C^1$ function. Further, using $\widetilde{A}_1^T(\widetilde{A}\widetilde{x}^* - \widetilde{y}) = 0$, $\widetilde{A}_2^T(\widetilde{A}\widetilde{x}^* - \widetilde{y}) \neq 0$ and each $\widetilde{A}_{\bullet i} \neq 0$, it can be shown that for any $i \in \mathcal{I} := \{2k - 1, 2k \mid k \in \mathcal{J}^c\}$, we have $\nabla_{\widetilde{y}} q_i(A_R, A_I, \widetilde{y}) \neq 0$, where $q_i(A_R, A_I, \widetilde{y}) := \widetilde{A}_{\bullet i}^T(\widetilde{y} - \widetilde{A} \cdot \widetilde{x}^*(\widetilde{A}, \widetilde{y}))$. In light of $\text{sgn}(\widetilde{x}_i^*(A_R, A_I, \widetilde{y})) = \text{sgn}(q_i(A_R, A_I, \widetilde{y}))$ for each $i = 1, \ldots, 2N$, the desired result follows.

(iv)-(v) These proofs are omitted as they are similar to those for Theorems 4.3 and 4.5 respectively. $\square$

The extension of Theorem 5.1 and Corollary 5.1 to the complex setting can be made similarly.

# 6 Conclusions

This paper provides an in-depth study of sparse properties of a wide range of $p$-norm based optimization problems with $p > 1$ generalized from sparse optimization and other related areas. By applying optimization and matrix analysis techniques, we show that optimal solutions to these generalized problems are the least sparse for almost all measurement matrices and measurement vectors. We also compare these problems with those when $0 < p \leq 1$. The results of this paper not only give a formal justification of the usage of $\ell_p$-optimization with $0 < p \leq 1$ for sparse optimization but they also offer a quantitative characterization of the adverse sparse properties of $\ell_p$-optimization with $p > 1$. These results will shed light on analysis and computation of general $p$-norm based optimization problems. Future research includes the

compressibility of $\ell_p$ minimization with $p > 1$ and extensions to matrix norm based optimization problems. Our preliminary results show the poor compressibility for $p > 1$; a further study of this property will be reported in a future work.

# 7 Appendix

In what follows, we show that the function $\| \cdot \|_p^p$ with $p > 1$ is strictly convex.

*Proof.* Let $p > 1$. By the Minkowski inequality, we have $\|x + y\|_p \leq \|x\|_p + \|y\|_p$ for all $x, y \in \mathbb{R}^N$, and the equality holds if and only if $y = \mu x$ for $\mu \geq 0$. For any $x, y \in \mathbb{R}^N$ with $x \neq y$ and any $\lambda \in (0, 1)$, consider two cases (i) $y = \mu x$ for some $\mu \geq 0$ with $\mu \neq 1$; (ii) otherwise. For case (i), $\|\lambda x + (1-\lambda) y\|_p^p = \|\lambda x + (1-\lambda)\mu x\|_p^p = [1 \cdot \lambda + \mu \cdot (1-\lambda)]^p \cdot \|x\|_p^p < [\lambda + \mu^p (1-\lambda)] \|x\|_p^p = \lambda \|x\|_p^p + (1-\lambda) \|y\|_p^p$, where we use the fact that $|x|^p$ is strictly convex on $\mathbb{R}_+$. For case (ii), we have $\|\lambda x + (1-\lambda) y\|_p^p < \left( \lambda \|x\|_p + (1-\lambda) \|y\|_p \right)^p \leq \lambda \|x\|_p^p + (1-\lambda) \|y\|_p^p$. This shows that $\| \cdot \|_p^p$ is strictly convex. $\qquad\square$

# References

[1] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, Vol. 28, pp. 253–263, 2008.

[2] A. Beck and Y.C. Eldar. Sparsity constrained nonlinear optimization: Optimality conditions and algorithms. *SIAM Journal on Optimization*, Vol. 23(3), pp. 1480–1509, 2013.

[3] S. Brugiapaglia and B. Adcock. Robustness to unknown error in sparse regularization. To appear in *IEEE Trans. on Information Theory*, 2017.

[4] K. Bryan and T. Leise. Making do more with less: An introduction to compressed sensing. *SIAM Review*, Vol. 55(3), pp. 547–566, 2013.

[5] O.P. Burdakov, C. Kanzow, and A. Schwartz. Mathematical programs with cardinality constraints: Reformulation by complementarity-type conditions and a regularization method. *SIAM Journal on Optimization*, Vol. 26(1), pp. 397–425, 2016.

[6] E.J. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. on Information Theory*, Vol. 52(2), pp. 489–509, 2006.

[7] E.J. Candes and M.B. Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, Vol. 25(2), pp. 21–30, 2008.

[8] S. Chen, D.L. Donoho, and M.A. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, Vol. 43(1), pp. 129–159, 2001.

[9] X. Chen, F. Xu, and Y. Ye. Lower bound theory of nonzero entries in solutions of $\ell_2 - \ell_p$ minimization. *SIAM Journal on Scientific Computing*, Vol. 32(5), pp. 2832–2852, 2010.

[10] M. El Chamie and G. Neglia. Newton's method for constrained norm minimization and its application to weighted graph problems. 2014 American Control Conference, pp. 2983–2988, Portland, OR, 2014.

[11] S. Delvaux and M. Van Barel. Rank-deficient submatrices of Fourier matrices. *Linear Algebra and Its Applications*, Vol. 429(7), pp. 1587–1605, 2008.

[12] D. Donoho. Compressed sensing. *IEEE Trans. Information Theory*, Vol. 52, pp. 1289–1306, 2006.

[13] D. Donoho. For most large underdetermined systems of linear equations, the minimal $\ell_1$-norm solution is also the sparsest solution. *Comm. Pure Appl. Math.*, Vol. 59, pp. 797–829, 2006.

[14] S. Foucart and H. Rauhut. *A Mathematical Introduction to Compressive Sensing.* Springer, New York, 2013.

[15] S. Fourcar and M.-J. Lai. Sparse solutions of under-determined linear systems via $\ell_q$ minimization for $0 < q < 1$. *Applied and Computational Harmonic Analysis*, Vol. 26, pp. 395–407, 2009.

[16] D. Ge, X. Jiang, and Y. Ye. A note on the complexity of $L_p$ minimization. *Mathematical Programming*, Vol. 129(2), pp. 285–299, 2011.

[17] T. Hastie, R. Tibshirani, and J. Friedman. Overview of Supervised Learning. In *The Elements of Statistical Learning*, pp. 9–41, Springer, New York, 2009.

[18] A.E. Hoerl and R.W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), pp. 55–67, 1970.

[19] M.-J. Lai and J. Wang. An unconstrained $\ell_q$ minimization with $0 < q \leq 1$ for sparse solution of underdetermined linear systems. *SIAM Journal on Optimization*, Vol. 21(1), pp. 82–101, 2011.

[20] S. Lefkimmiatis and M. Unser. Poisson image reconstruction with Hessian Schatten-norm regularization. *IEEE Transactions on Image Processing*, Vol. 22(11), pp. 4314–4327, 2013.

[21] C.D. Meyer. *Matrix Analysis and Applied Linear Algebra.* SIAM Press, 2nd Edition, 2000.

[22] J.R. Munkres. *Topology.* 2nd Edition, Prentice-Hall, Englewood Cliffs, NJ, 2000.

[23] H.L. Royden. *Real Analysis.* 3rd Edition, Prentice-Hall, Englewood Cliffs, NJ, 1988.

[24] R. Saab and O. Yilmaz. Sparse recovery by non-convex optimization-instance optimality. *Applied and Computational Harmonic Analysis*, Vol. 29(1), pp. 30–48, 2010.

[25] R. Saab, R. Chartrand, and O. Yilmaz. Stable sparse approximations via nonconvex optimization. In IEEE Conference on Acoustics, Speech and Signal Processing, pp. 3885–3888, March, 2008.

[26] T. Terlaky. On $\ell_p$ programming. *European Journal of Operational Research*, Vol. 22(1), pp. 70–100, 1985.

[27] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B*, Vol. 58(1), pp. 267–288, 1996.

[28] E. Van Den Berg and M.P. Friedlander. Probing the Pareto frontier for basis pursuit solutions. *SIAM Journal on Scientific Computing*, Vol. 31(2), pp. 890–912, 2008.

[29] J. Wen, D. Li, and F. Zhu. Stable recovery of sparse signals via $l_p$-minimization. *Applied and Computational Harmonic Analysis*, Vol. 38(1), pp. 161–176, 2015.

[30] S.J. Wright, R.D. Nowak, and M.A. Figueiredo. Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, Vol. 57(7), pp. 2479–2493, 2009.

[31] H. Xu, C. Caramanis, and S. Mannor. Sparse algorithms are not stable: A no-free-lunch theorem. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 34(1), pp. 187–193, 2012.

[32] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, Vol. 67(2), pp. 301–320, 2005.