SPARSE RECOVERY BY NON-CONVEX OPTIMIZATION – INSTANCE OPTIMALITY

RAYAN SAAB AND ÖZGÜR YILMAZ

Abstract. In this note, we address the theoretical properties of Δ_p, a class of compressed sensing decoders that rely on ℓ^p minimization with $p \in (0,1)$ to recover estimates of sparse and compressible signals from incomplete and inaccurate measurements. In particular, we extend the results of Candès, Romberg and Tao [3] and Wojtaszczyk [30] regarding the decoder Δ_1, based on ℓ^1 minimization, to Δ_p with $p \in (0,1)$. Our results are two-fold. First, we show that under certain sufficient conditions that are weaker than the analogous sufficient conditions for Δ_1 the decoders Δ_p are robust to noise and stable in the sense that they are $(2,p)$ instance optimal. Second, we extend the results of Wojtaszczyk to show that, like Δ_1, the decoders Δ_p are $(2,2)$ instance optimal in probability provided the measurement matrix is drawn from an appropriate distribution. While the extension of the results of [3] to the setting where $p \in (0,1)$ is straightforward, the extension of the instance optimality in probability result of [30] is non-trivial. In particular, we need to prove that the LQ_1 property, introduced in [30], and shown to hold for Gaussian matrices and matrices whose columns are drawn uniformly from the sphere, generalizes to an LQ_p property for the same classes of matrices. Our proof is based on a result by Gordon and Kalton [18] about the Banach-Mazur distances of p-convex bodies to their convex hulls.

1. Introduction

The sparse recovery problem received a lot of attention lately, both because of its role in transform coding with redundant dictionaries (e.g., [9, 28, 29]), and perhaps more importantly because it inspired compressed sensing [3, 4, 13], a novel method of sensing certain classes of analog signals more efficiently compared to the classical approach based on Nyquist-Shannon sampling theory. Define Σ_S^N to be the set of all S-sparse vectors

$$\Sigma_S^N := \{x \in \mathbb{R}^N; \quad \#\mathrm{supp}(x) \leq S\},$$

and define compressible vectors as vectors that can be well approximated in Σ_S^N. Let $\sigma_S(x)_{\ell^p}$ denote the best S-term approximation error of x in ℓ^p (quasi-)norm where $p > 0$, i.e.,

$$\sigma_S(x)_{\ell^p} := \min_{v \in \Sigma_S^N} \|x - v\|_p.$$

R. Saab was supported in part by a UGF award from the University of British Columbia, and by a Pacific Century Graduate Scholarship from the Province of British Columbia through the Ministry of Advanced Education.

Ö. Yılmaz was supported in part by a Discovery Grant and by a CRD Grant from the Natural Sciences and Engineering Research Council of Canada.

Let A be an $M \times N$ matrix, where $M < N$, and define the associated *encoder* $\mathcal{E}_A : \mathbb{R}^N \mapsto \mathbb{R}^M$ via

$$\mathcal{E}_A(x) = Ax.$$

The transform coding and compressed sensing problems mentioned above require the existence of decoders, say $\Delta : \mathbb{R}^M \mapsto \mathbb{R}^N$, for such encoders with roughly the following properties:

(C1) $\Delta(\mathcal{E}_A(x)) = x$ whenever $x \in \Sigma_S^N$ with sufficiently small S.

(C2) $\|x - \Delta(\mathcal{E}_A(x)+e)\| \lesssim \|e\| + \sigma_S(x)_{\ell^p}$, where the norms are appropriately chosen. Here e denotes measurement error, e.g., thermal and computational noise.

(C3) $\Delta(\mathcal{E}_A(x))$ can be computed efficiently (in some sense).

Below, we denote the (in general noisy) encoding of x by b, i.e., we have

$$b = Ax + e. \tag{1}$$

In general, the problem of constructing decoders with properties (C1)-(C3) is non-trivial (even in the noise-free case) as A is overcomplete, i.e., the linear system of M equations in (1) is underdetermined, and thus admits infinitely many solutions. In order for a decoder to satisfy (C1)-(C3), it must choose the "correct solution" among these infinitely many solutions. Under the assumption that the original signal x is sparse, one can phrase the problem of finding the desired solution as an optimization problem where the objective is to maximize an appropriate "measure" of sparsity while simultaneously satisfying the constraints defined by (1).

In the noise-free case, i.e., when $e = 0$ in (1), under certain conditions on the $M \times N$ matrix A, e.g., if A is in general position, it can be shown that there is a decoder Δ_0 which satisfies $\Delta_0(\mathcal{E}_A(x)) = x$ for all $x \in \Sigma_S^N$ whenever $S < M/2$ [14]. This Δ_0 can be explicitly computed via the optimization problem

$$\Delta_0(b) := \arg\min_y \|y\|_0 \text{ subject to } b = Ay. \tag{2}$$

Here $\|y\|_0$ denotes the number of non-zero components of the vector y, equivalently its so-called ℓ^0-norm. Clearly, the sparsity of y is reflected by its ℓ^0-norm.

1.1. Decoding by ℓ^1 minimization.

As mentioned above, it can be shown that $\Delta_0(Ax) = x$ exactly if x is sufficiently sparse depending on the matrix A. However, the associated optimization problem is combinatorial in nature, thus its complexity grows extremely quickly as N becomes much larger than M. Naturally, one then seeks to modify the optimization problem so that it lends itself to solution methods that are more tractable than combinatorial search. In fact, it has been shown that, in the noise-free setting, the decoder defined by ℓ^1 minimization, given by

$$\Delta_1(b) = \arg\min_x \|x\|_1 \text{ subject to } Ax = b, \tag{3}$$

recovers x exactly if x is sufficiently sparse and the matrix A has certain properties (e.g., [3, 6, 9, 14, 15, 26]). In particular, it has been shown in [3] that if $x \in \Sigma_S^N$ and A satisfies a certain restricted isometry property, e.g., $\delta_{3S} < 1/3$ or more generally $\delta_{(k+1)S} < \frac{k-1}{k+1}$ for some k, $kS \in \mathbb{N}^+$, then $\Delta_1(Ax) = x$. Here δ_S are the S-restricted isometry constants of A, as introduced by Candès, Romberg and Tao (see, e.g., [3]), defined as the smallest constants satisfying

$$(1 - \delta_S)\|c\|_2^2 \le \|Ac\|_2^2 \le (1 + \delta_S)\|c\|_2^2 \tag{4}$$

for every $c \in \Sigma_S^N$. Throughout the paper, using the notation of [30], we say that a matrix satisfies RIP(S, δ) if $\delta_S < \delta$.

Checking whether a given matrix satisfies a certain RIP is computationally intensive, and becomes rapidly untractable as the size of the matrix increases. On the other hand, there are certain classes of random matrices which have favorable RIP. In fact, let A be an $M \times N$ matrix the columns of which are i.i.d. random vectors with any sub-Gaussian distribution. It has been shown that A satisfies RIP (S, δ) with $S \leq c_1 M/log(N/M)$, $\delta < 1$ with probability $> 1 - 2e^{c_2 M}$ (see, e.g., [1], [5]).

In addition to recovering sparse vectors from error-free observations, it is important that the decoder used to obtain the approximation be robust to noise and stable with regards to the "compressibility" of x. In other words, we require that the reconstruction error scale well with the measurement error and with the "non-sparsity" of the signal (i.e., (C2) above). For matrices that satisfy RIP$((k+1)S, \delta)$,with $\delta < \frac{k-1}{k+1}$, it has been shown in [3] that there exists a feasible decoder Δ_1^ϵ for which the approximation error $\|\Delta_1^\epsilon(b) - x\|_2$ scales linearly with the measurement error $\|e\|_2 \leq \epsilon$ and with $\sigma_S(x)_{\ell^1}$. More specifically, define the decoder

$$\Delta_1^\epsilon(b) = \arg\min_x \|x\|_1 \text{ subject to } \|Ax - b\|_2 \leq \epsilon. \tag{5}$$

The following theorem of Candes et al. in [3] provides error guarantees when x is not "exactly" sparse and when the observation is noisy.

Theorem 1.1. [3] *Fix $\epsilon \geq 0$, assume that x is arbitrary, and let $b = Ax + e$ where $\|e\|_2 \leq \epsilon$ (where A is an $M \times N$ matrix with $M < N$). If $\delta_{3S} + 3\delta_{4S} < 2$, then $\Delta_1^\epsilon(b)$ satisfies*

$$\|\Delta_1^\epsilon(b) - x\|_2 \leq C_{1,S}\epsilon + C_{2,S}\frac{\sigma_S(x)_{\ell^1}}{\sqrt{S}}. \tag{6}$$

For reasonable values of δ_{4S}, the constants are well behaved; e.g., $C_{1,S} = 12.04$ and $C_{2,S} = 8.77$ for $\delta_{4S} = 1/5$.

Remark 1.1.1. This means that given $b = Ax + e$, and x is sufficiently sparse, $\Delta_1^\epsilon(b)$ recovers the underlying sparse signal within the noise level. Consequently the recovery is perfect if $\epsilon = 0$.

Remark 1.1.2. By explicitly assuming x to be sparse, Candès et. al. [3] proved a version of the above result with smaller constants, i.e., for $b = Ax + e$ with $x \in \Sigma_S^N$ and $\|e\|_2 \leq \epsilon$,

$$\|\Delta_1^\epsilon(b) - x\|_2 \leq C_S\epsilon, \tag{7}$$

where $C_S < C_{1,S}$.

Remark 1.1.3. Recently, Candès [2] showed that $\delta_{2S} < \sqrt{2} - 1$ is sufficient to guarantee robust and stable recovery in the sense of (6) with slightly better constants.

In the noise free case, i.e., when $\epsilon = 0$, the reconstruction error in Theorem 1.1 is bounded above by $\sigma_S(x)_{\ell^1}/\sqrt{S}$ (see (6)). This upper bound would sharpen if one could replace $\sigma_S(x)_{\ell^1}/\sqrt{S}$ with $\sigma_S(x)_{\ell^2}$ on the right hand side of (6) (note that $\sigma_S(x)_{\ell^1}$ can be large even if all the entries in the reconstruction error are small but nonzero; this follows from the fact that for any vector $y \in \mathbb{R}^N$, $\|y\|_2 \leq \|y\|_1 \leq \sqrt{N}\|y\|_2$, and consequently there are many vectors x for which $\sigma_S(x)_{\ell^1}/\sqrt{S} \gg \sigma_S(x)_{\ell^2}$, especially when N is large). In [10] it was shown that the term $C_{2,S}\sigma_S(x)_{\ell^1}/\sqrt{S}$ on the right hand side of (6) *cannot* be replaced with $C\sigma_S(x)_{\ell^2}$ if one seeks the inequality to hold for all $x \in \mathbb{R}^N$ with a fixed matrix A, unless $M > cN$ for some constant c. This is unsatisfactory since the paradigm of compressed sensing relies on the ability of recovering sparse or compressible vectors x from significantly fewer measurements than its ambient dimension N.

Even though one cannot obtain bounds on the approximation error in terms of $\sigma_S(x)_{\ell^2}$ with constants that are uniform on x (with a fixed matrix A), the situation is significantly better if we relax the uniformity requirement and seek for a version of (6) that holds "with high probability". Indeed, it has been recently shown that for any specific x, $\sigma_S(x)_{\ell^2}$ can be placed in (6) in lieu of $\sigma_S(x)_{\ell^1}/\sqrt{S}$ (albeit with different constants) with high probability on the draw of A if (i) $M > cS\log N$ and (ii) the entries A is drawn independently from a Gaussian distribution or the columns of A are drawn independently from the uniform distribution on the unit sphere in $\mathbb{R}^M$ [30]. In other words, the encoder $\Delta_1 = \Delta_1^0$ is *(2,2) instance optimal in probability*, a property which was discussed extensively in [10].

Following the notation of [30], we say that a decoder is (q,p) instance optimal if

$$\|\Delta(Ax) - x\|_q \leq C\sigma_S(x)_{\ell^p}/S^{1/p-1/q} \tag{8}$$

holds for all $x \in \mathbb{R}^N$. Moreover, a decoder Δ is said to be (q,p) instance optimal in probability if (8) holds for a particular x with high probability on the draw of A. Thus, with this notation the stability results shown by Candès et al. [3] in Theorem 1.1 imply (2,1) instance optimality of the decoder Δ_1 (set $\epsilon = 0$), while the results of Wojtaszczyk in [30] show that Δ_1 is (2,2) instance optimal in probability.

Thus, it is now clear that Δ_1 satisfies conditions (C1) and (C2), and it only remains to note that decoding by Δ_1 amounts to solving an ℓ^1 minimization problem, and is thus tractable. Furthermore, ℓ^1 minimization problems can be solved efficiently with solvers specifically designed for the sparse recovery scenarios (e.g. [27], [16], [11]).

1.2. **Decoding by ℓ^p minimization.**

We have so far seen that the decoder Δ_1^ϵ provides robust and stable recovery for compressible signals even when the measurements are noisy, and that with high probability it is (2,2) instance optimal. The stability and robustness properties are conditioned on an appropriate RIP while the instance optimality property is dependent on the draw of the matrix from an appropriate distribution, in addition to RIP.

Recall that the decoders Δ_1 and Δ_1^ϵ were devised because their action can be computed by solving optimization problems that are convex approximations to the combinatorial ones of (2) required to compute Δ_0. The decoders defined by

$$\Delta_p^\epsilon(b) = \arg\min_x \|x\|_p \text{ s.t. } \|Ax - b\|_2 \leq \epsilon, \tag{9}$$

and

$$\Delta_p = \arg\min \|x\|_p \text{ s.t. } Ax = b, \tag{10}$$

with $0 < p < 1$ are also approximations of Δ_0 the action of which is computed by solving a non-convex optimization problem (which can be solved, at least locally, much faster than (2)). It is natural to ask whether the decoders Δ_p and Δ_p^ϵ possess robustness, stability, and instance optimality properties similar to those of Δ_1, and whether the properties are obtained under weaker conditions than the analogous ones with $p = 1$.

Early work by Gribonval and co-authors [19–22] take some initial steps in answering these questions. In particular, they devise metrics that lead to sufficient conditions for uniqueness of $\Delta_1(b)$ to imply uniqueness of $\Delta_p(b)$ and specifically for having $\Delta_p(b) = \Delta_1(b) = x$. The authors also present stability conditions in terms of various norms that bound the error, and they conclude that the smaller the value of p is, the more non-zero components can be recovered by (9). These conditions, however, are hard to check explicitly and no class of deterministic or random matrices was shown to satisfy them at least with high probability. On the other hand, the authors provide lower bounds for their metrics in terms of generalized mutual coherence. Still, these conditions are pessimistic in the sense that they generally guarantee recovery of only very sparse vectors.

Recently, Chartrand showed that in the noise-free setting, a sufficiently sparse signal can be recovered perfectly with Δ_p, where $p \in (0,1)$, under less restrictive RIP requirements than those needed to guarantee perfect recovery with Δ_1. The following theorem was proved in [7].

Theorem 1.2. [7] *Let $0 < p \leq 1$. Assume that x is S-sparse, $b = Ax$ and suppose that $\delta_{kS} + k^{\frac{2-p}{p}} \delta_{(k+1)S} < k^{\frac{2-p}{p}} - 1$, for some $k > 1$. Then $\Delta_p(b) = x$.*

Note that, for example, when $p = 0.5$ and $k = 3$, the above theorem only requires $\delta_{3S} + 27\delta_{4S} < 26$ to guarantee perfect recovery with $\Delta_{0.5}$, a less restrictive condition than the analogous one needed to guarantee perfect reconstruction with Δ_1, i.e., $\delta_{3S} + 3\delta_{4S} < 2$. Moreover, in [8], Staneva and Chartrand study a modified RIP that is defined by replacing $\|Ax\|_2$ in (4) with $\|Ax\|_p$. They show that under this new definition of δ_S, the same sufficient condition as in Theorem 1.2 guarantees perfect recovery. Steneva and Chartrand also show that if A is an $M \times N$ Gaussian matrix, their sufficient condition is satisfied provided $M > C1(p)S + pC2(p)S\log(N/K)$. In other words, the dependence on N of the required number of measurements M (that guarantees perfect recovery for all $x \in \Sigma_S^N$) disappears as p approaches 0. This result motivates a more detailed study to understand the properties of the decoders Δ_p in terms of stability and robustness, which is the objective of this paper.

1.2.1. *Algorithmic Issues.* Clearly, recovery by ℓ^p minimization poses a non-convex optimization problem with many local minimizers. Moreover, the results presented in this work and in others [7, 19–22, 25] assume that the global minimizer has been found, even though a significant proportion of these results (including all results in this paper continue to hold if we could obtain a solution (feasible point) x^* which satisfies $\|x^*\|_p \leq \|x\|_p$ (where x is the vector to be recovered). It is encouraging that simulation results from recent papers, e.g., [7,25] strongly indicate that simple modifications to known approaches like iterated reweighted least squares algorithms and projected gradient algorithms yield x^* that are the global minimizers of the associated ℓ^p minimization (or approximate the global optimizers very well). Nevertheless, it should be stated that to our knowledge, these algorithms have only been shown to converge to local minima.

1.3. **Paper Outline.** In what follows, we present generalizations of the above results, giving stability and robustness guarantees for ℓ^p minimization. In Section 2.1 we show that the decoders Δ_p and Δ_p^ϵ are robust to noise and (2,p) instance optimal, and in that sense stable. For this section we rely and expand on our note [25]. In Section 2.3 we extend [30] and show that for the same range of dimensions as for decoding by ℓ^1 minimization, i.e., $M > cS\log(N)$, Δ_p is also (2,2) instance optimal in probability provided the measurement matrix is drawn from an appropriate distribution. The generalization is non-trivial and requires a result by Gordon and Kalton [18] to control the Banach-Mazur distance between a p-convex body and its convex hull. In Section 3 we present some numerical results, further illustrating the possible benefits of using ℓ^p minimization and highlighting the behavior of the Δ_p decoder in terms of stability and robustness. Finally, in Section 4 we present the proofs of the main theorems and corollaries.

While writing this paper, we became aware of the work of Foucart and Lai [17] which also shows similar $(2, p)$ instance optimality results for $p \in (0, 1)$ under different sufficient conditions. In essence, one could use the $(2, p)$-results of Foucart and Lai to obtain $(2, 2)$ instance optimality in probability results similar to the ones we present in this paper, albeit with different constants. Since neither the sufficient conditions for $(2, p)$ instance optimality presented in [17] nor the ones in this paper are uniformly weaker, and since neither provides uniformly better constants, we simply use our estimates throughout.

2. Main Results

In this section, we present our main theoretical results pertaining to the ability of ℓ^p minimization to recover sparse and compressible signals in the presence of noise.

2.1. Sparse recovery with Δ_p: stability and robustness.

Here, we present our results on the robustness and stability properties of Δ_p and Δ_p^ϵ. We show that under appropriate sufficient conditions, we obtain deterministic performance guarantees. In fact, it is sufficient that the matrix satisfies an RIP condition that is weaker than the analogous one for recovery by Δ_1 and Δ_1^ϵ. Thus our results are of the same nature as the conditions provided in Section (1.1) for ℓ^1 minimization in the general (noisy and non-sparse) setting while being less restrictive.

We begin with a generalization of Theorem 1.1 where x is arbitrary and $\sigma_S(x)_{\ell^p}$ is its best S-term approximation error measured in ℓ^p-norm. In particular, we are interested in controlling the error $\|\Delta_p^\epsilon(b) - x\|_2^p$.

Theorem 2.1 (General Case)**.** *Let $p \in (0, 1]$. Assume that x is arbitrary and suppose that*

$$\delta_{kS} + k^{\frac{2}{p}-1}\delta_{(k+1)S} < k^{\frac{2}{p}-1} - 1, \tag{11}$$

for some $k > 1$, $kS \in \mathbb{Z}^+$. Let $b = Ax + e$ where $\|e\|_2 \le \epsilon$. Then $\Delta_p^\epsilon(b)$ satisfies

$$\|\Delta_p^\epsilon(b) - x\|_2^p \le C^{(1)}\epsilon^p + C^{(2)}\frac{\sigma_S(x)_{\ell^p}^p}{S^{1-p/2}}, \tag{12}$$

where

$$C^{(1)} = 2^p\frac{1 + k^{p/2-1}(2/p-1)^{-p/2}}{(1-\delta_{(k+1)S})^{p/2} - (1+\delta_{kS})^{p/2}k^{p/2-1}}, \quad \textit{and} \tag{13}$$

$$C^{(2)} = \frac{2(\frac{p}{2-p})^{p/2}}{k^{1-p/2}}\left[1 + \frac{(1+k^{p/2-1})(1+\delta_{kS})^{p/2}}{(1-\delta_{(k+1)S})^{p/2} - \frac{(1+\delta_{kS})^{p/2}}{k^{1-p/2}}}\right]. \tag{14}$$

Remark 2.1.1. By setting $p = 1$ and $k = 3$ in Theorem 2.1, we obtain Theorem 1.1, with precisely the same constants.

Corollary 2.2 ($(2, p)$ instance optimality)**.** *Let $p \in (0, 1]$. Suppose that for some $k > 1$, $kS \in \mathbb{Z}^+$ (11), holds then the decoder Δ_p is $(2, p)$ instance optimal, i.e.,*

$$\|\Delta_p(Ax) - x\|_2 \leq \left(C^{(2)}\right)^{1/p} \frac{\sigma_S(x)_{\ell^p}}{S^{1/p-1/2}},$$

where $C^{(2)}$ is as in (14).

Corollary 2.3 (sparse case)**.** *Assume that x is S-sparse and suppose that for some $k > 1$, $kS \in \mathbb{Z}^+$*

$$\delta_{kS} + k^{\frac{2}{p}-1}\delta_{(k+1)S} < k^{\frac{2}{p}-1} - 1. \tag{15}$$

Let $b = Ax + e$ where $\|e\|_2 \leq \epsilon$. Then $\Delta_p^\epsilon(b)$ satisfies

$$\|\Delta_p^\epsilon(b) - x\|_2 \leq \left(C^{(1)}\right)^{1/p} \epsilon,$$

where $C^{(1)}$ is as in (13).

Remark 2.3.1. In Section 4 we prove Corollary 2.3 independently of Theorem 2.1 which leads to better values of the constants. However, for the sake of clarity of the text, we simply note that Corollary 2.3 can be deduced from the previous theorem by specializing it to the case of sparse signals, i.e., $\sigma_S(x)_{\ell^p} = 0$.

Remark 2.3.2. In [17], Foucart and Lai give different sufficient conditions for exact recovery than those we present. In particular, they show that if

$$\delta_{mS} < g(m) := \frac{(4\sqrt{2}-1)(m/2)^{1/p-1/2} - 1}{(4\sqrt{2}-1)(m/2)^{1/p-1/2} + 1} \tag{16}$$

holds for some $m \geq 2, mS \in \mathbb{N}^+$, then Δ_p will recover signals in Σ_S^N exactly. Note that the sufficient condition we present in this paper, namely (11), holds when

$$\delta_{mS} < f(m) := \frac{(m-1)^{2/p-1} - 1}{(m-1)^{2/p-1} + 1} \tag{17}$$

for some $m \geq 2, mS \in \mathbb{N}^+$. In Figure 1, we compare these different sufficient conditions as a function of m for $p = 0.1, 0.5$, and 0.9 respectively. Figure 1 indicates that neither sufficient condition is weaker than the other for all values of m. In fact, we can deduce that (16) is weaker when m is close to 2, while (17) is weaker when m starts to grow larger. Since both conditions are only sufficient, if either one of them holds for an appropriate m, then Δ_p recovers all signals in Σ_S^N.

Remark 2.3.3. In [12], Davies and Gribonval showed that if one chooses $\delta_{2S} > \delta(p)$ (where $\delta(p)$ can be computed implicitly for $p \in (0, 1]$), then there exist matrices with the prescribed δ_{2S} for which Δ_p fails to recover signals in Σ_S^N. Note that this result does not contradict with the results that we present in this paper: we provide sufficient conditions (e.g., (11)) in terms of $\delta_{(k+1)S}$, where $k > 1$ and $kS \in \mathbb{N}^+$, that guarantee recovery by Δ_p. These conditions are weaker than the corresponding conditions ensuring recovery by Δ_1, which suggests that using Δ_p can be beneficial. Moreover, the numerical examples we provide in Section 3 indicate that by using Δ_p, one can recover signals in $\Sigma_{S_p}^N$, $p \in (0, 1]$ even when Δ_1 fails to recover them (see Figure 2).

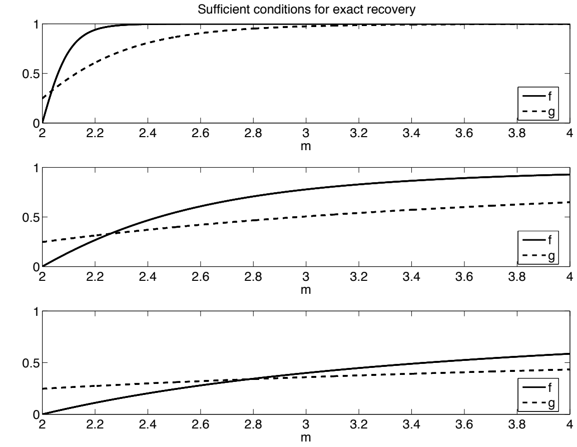

FIGURE 1. A comparison of the sufficient conditions on δ_{mS} in (17) and (16) as a function of m, for $p = 0.1$ (Top), $p = 0.5$ (Center) and $p = 0.9$ (Bottom).

Remark 2.3.4. In summary, Theorem 2.1 states that if (11) is satisfied then we can recover signals in Σ_S^N stably by decoding with Δ_p^ϵ. It is worth mentioning that the sufficient conditions presented here reduce the gap between the conditions for exact recovery with Δ_0 (i.e., $\delta_{2S} < 1$) and with Δ_1, e.g., $\delta_{3S} < 1/3$. For example for $k = 2$ and $p = 0.5$, $\delta_{3S} < 7/9$ is sufficient. In fact, this improvement can be quantified as follows. Let S_p be the largest allowed value of S in (12), which is determined by (11). In the noise free setting, if $x \in \Sigma_S^N$ with $S \leq S_p$, Δ_p recovers x exactly. It is easy to see from (11) that if $0 < p < q \leq 1$, then we have $S_p \geq S_q$. In the next subsection, we compute a lower bound for the ratio S_p/S_q.

2.2. **The relationship between S_1 and S_p.** Let A be an $M \times N$ matrix with RIP(j, δ_j), $j = 1, \ldots, M$. Suppose, in addition, that A is in general position and consequently $\delta_j < 1$ for all $j \leq M$. Let $0 < p < q \leq 1$, and define S_r for the matrix A (with $r \in (0, 1]$) as the largest value of S for which a slightly stronger version of (11) given by

$$\delta_{(k+1)S} < \frac{k^{2/r-1} - 1}{k^{2/r-1} + 1} \tag{18}$$

holds for some k. Let $j = (k + 1)S$, then (18) is equivalent to requiring

$$S < \frac{j}{1 + \left(\frac{1+\delta_j}{1-\delta_j}\right)^{r/(2-r)}}, \tag{19}$$

and consequently this gives

$$S_r = \max_{j \in \{2,\ldots,M\}} \left\lfloor \frac{j}{1 + \left(\frac{1+\delta_j}{1-\delta_j}\right)^{r/(2-r)}} \right\rfloor \text{ subject to } \frac{j}{1 + \left(\frac{1+\delta_j}{1-\delta_j}\right)^{r/(2-r)}} \notin \mathbb{N}. \tag{20}$$

Let j^* be the optimal value of j that yields S_q in (20) and suppose that $S_q \geq 1$. Consequently,

$$\frac{j^*}{1+\left(\frac{1+\delta_{j^*}}{1-\delta_{j^*}}\right)^{p/(2-p)}} \geq \frac{j^*}{1+\left(\frac{1+\delta_{j^*}}{1-\delta_{j^*}}\right)^{q/(2-q)}} \geq 1.$$

Observe that if $a \geq b \geq 1$ then $\frac{\lfloor a \rfloor}{\lceil b \rceil} \geq \lfloor a/b \rfloor$. Thus,

$$S_p/S_q \geq \left\lfloor \frac{1+\left(\frac{1+\delta_{j^*}}{1-\delta_{j^*}}\right)^{q/(2-q)}}{1+\left(\frac{1+\delta_{j^*}}{1-\delta_{j^*}}\right)^{p/(2-p)}} \right\rfloor.$$

In particular, when $q = 1$, we obtain,

$$S_p/S_1 \geq \left\lfloor \frac{1+\left(\frac{1+\delta_{j^*}}{1-\delta_{j^*}}\right)}{1+\left(\frac{1+\delta_{j^*}}{1-\delta_{j^*}}\right)^{p/(2-p)}} \right\rfloor \geq \left\lfloor \frac{1}{(1-\delta_{j^*})^{(2-2p)/(2-p)}} \right\rfloor.$$

In the last inequality we used the fact that $1 + a^r \leq 2^{1-r}(1+a)^r$, when $r \in (0, 1]$.

So, we have proved the following proposition.

Proposition 2.4. *Suppose for some k and S_1, $\delta_{(k+1)S_1} < \frac{k-1}{k+1}$. Then Δ_1 recovers S_1-sparse vectors and Δ_p recovers S_p-sparse vectors with*

$$S_p \geq \left\lfloor \left(\frac{1}{1-\delta_{(k+1)S_1}}\right)^{(2-2p)/(2-p)} \right\rfloor S_1.$$

2.3. **Instance optimality in probability of Δ_p.** In this section, we show that Δ_p is $(2, 2)$ instance optimal in probability. Our approach is based on that of [30], which we summarize now. A matrix A is said to possess the $LQ_1(\alpha)$ property iff

$$A(B_1^N) \supset \alpha B_2^M.$$

In [30], Wojtaszczyk shows that random Gaussian matrices of size $M \times N$, as well as matrices whose columns are drawn uniformly from the sphere posses the $LQ_1(\alpha)$ property, $\alpha = \mu\sqrt{\frac{\log(N/M)}{M}}$ with high probability. Noting that such matrices also satisfy $RIP((k+1)S, \delta)$ with $S < c\frac{M}{log(N/M)}$ with high probability, Wojtaszczyk proves that Δ_1, with these matrices, is (2,2) instance optimal in probability. Our strategy for proving instance optimality for Δ_p, $p \in (0, 1)$, relies on the non-trivial generalization of the LQ_1 property to an $LQ_p(\alpha)$ property with $\alpha = 1/C_p\left(\mu^2\frac{\log(N/M)}{M}\right)^{(1/p-1/2)}$. Specifically, we say that a matrix A satisfies $LQ_p(\alpha)$ iff

$$A(B_p^N) \supset \alpha B_2^M.$$

Once we establish this property, the proof of instance optimality in probability for Δ_p proceeds largely unchanged from Wojtaczszyk's proof with modifications only to account for the non-convexity of the ℓ^p-norm with $p \in (0, 1)$. We present the relevant theorems on instance optimality of the Δ_p decoder, while deferring the proofs to section 4. Note that throughout this section, we will use A_ω to denote matrices whose entries are drawn from a zero mean, normalized column variance

Gaussian distribution and $\tilde{A}_\omega$ to denote matrices drawn uniformly from the sphere. Our main results are as follows.

We start with the main lemma, which shows that the matrices A_ω and $\tilde{A}_\omega$ satisfy the LQ_p property with high probability.

Lemma 2.5. *$\tilde{A}_\omega$ and A_ω satisfy the $LQ_p(\alpha)$ property with $\alpha = 1/C_p\left(\mu^2\frac{\log(N/M)}{M}\right)^{1/p-1/2}$ with probability $\geq 1-e^{-cM}$ on the draw of the matrix. Here, C_p is a constant that depends only on p.*

As we shall see in Section 4, proving Lemma 2.5 is non-trivial and requires a result by [18] also reported in [23] relating the Banach-Mazur distances of p-convex bodies to their convex hulls. On the other hand, this lemma provides the machinery needed to extend the results on (2,2) instance optimality of Δ_1 to Δ_p, where $p \in (0,1]$. In particular, it allows us to show the following theorem, which extends an analogous result of Wojtaszczyk [30].

Theorem 2.6. *Suppose that A satisfies $RIP(S,\delta)$ and $LQ_p\left(1/C_p(\mu^2/S)^{1/p-1/2}\right)$. If (A,Δ) is (2,p) instance optimal, i.e.,*

$$\|\Delta(Ax)-x\|_2 < C_{2,p}\frac{\sigma_S(x)_{\ell^p}}{S^{1/p-1/2}}$$

then for any $x\in\mathbb{R}^N, e\in\mathbb{R}^M$, all of the following hold.

(i) $\|\Delta(Ax+e)-x\|_2 \leq C(\|e\|_2+\frac{\sigma_S(x)_{\ell^p}}{S^{1/p-1/2}})$

(ii) $\|\Delta(Ax)-x\|_2 \leq C(\|Ax_{T_0^c}\|_2+\sigma_S(x)_{\ell^2})$

(iii) $\|\Delta(Ax+e)-x\|_2 \leq C(\|e\|_2+\sigma_S(x)_{\ell^2}+\|Ax_{T_0^c}\|_2)$

Finally, our main theorem on the instance optimality in probability of the Δ_p decoder follows.

Theorem 2.7. *Let $A\in\mathbb{R}^{M\times N}$ be drawn from iid Gaussian random variables or from the uniform distribution on the sphere. $\exists S_0 < cM/\log N$ s.t. $\forall S < S_0$*

(i) *$\exists\Omega_1; P(\Omega_1)\geq 1-e^{cM}$, for any $x\in\mathbb{R}^N, e\in\mathbb{R}^M$:*

$$\|\Delta_p(A_\omega(x)+e)-x\|_2 \leq C(\|e\|_2+\frac{\sigma_S(x)_{\ell^p}}{S^{1/p-1/2}}),$$

(ii) *for any $x\in\mathbb{R}^N, \exists\Omega_1; P(\Omega_1)\geq 1-e^{cM}$, s.t. for any $e\in\mathbb{R}^M$:*

$$\|\Delta_p(A_\omega(x)+e)-x\|_2 \leq C\left(\|e\|_2+\sigma_S(x)_{\ell^2}\right).$$

The statement also holds for $\tilde{A}_\omega$.

Note that the constants above (both denoted by C) rely on the parameters of the particular LQ_p and RIP properties that the matrix satisfies, and are given explicitly in Section 4.

Remark 2.7.1. The above theorem pertains to the decoders Δ_p which, like the analogous theorem for Δ_1 presented in [30], requires no knowledge of the noise level. In other words, Δ_p provides estimates of sparse and compressible signals from limited and noisy observations without having to explicitly account for the noise in the decoding. This provides a practical advantage when estimates of measurement noise levels are absent. This may be especially important in compressed sensing applications since the measurements will generally be Gaussian distributed with zero mean. Assuming that the noise is also Gaussian distributed with unknown variance, it may be hard to estimate its level a posteriori.

3. Numerical Experiments

In this section, we present the results of some numerical experiments to highlight important aspects of sparse reconstruction by decoding using Δ_p, $0 < p \leq 1$. First, we are interested in the sufficient conditions under which decoding with Δ_p can guarantee perfect recovery of signals in Σ_S^N for different values of p and S. We also present numerical results to observe the robustness and instance optimality of the Δ_p decoder. In other words, we want to observe the linear growth of the ℓ^2 reconstruction error $\|\Delta_p(Ax+e) - x\|_2$, as a function of $\sigma_S(x)_{\ell^2}$ and of $\|e\|_2$.

To that end, we generate a 100×300 matrix whose columns are drawn from a Gaussian distribution and estimate its RIP constants δ_S via Monte Carlo (MC) simulations. Under the assumption that the estimated constants are in fact the correct ones (while in fact they are only lower bounds), Figure 2 (left) shows the regions where (11) guarantees recovery for different (S, p)-pairs. On the other hand, Figure 2 (right) shows the empirical recovery rates using the same matrix with fifty different instances of $x \in \Sigma_S^N$, and decoding by Δ_p, where we choose the non-zero coefficients of x randomly from the Gaussian distribution. Moreover, we compute $\Delta_p(Ax)$, as a solution to the ℓ^p optimization problem of (10) by using a projected gradient algorithm on a smoothed version of $\|x\|_p^p$, namely $\sum_i (x_i^2 + \epsilon^2)^{p/2}$, where the solution to each subproblem, starting with a large ϵ is used as an initial estimate for the next subproblem with a smaller ϵ. Note that this approach is similar to the one described in [7]. Clearly, the empirical results show that Δ_p is successful in a wider range of scenarios than those predicted by Theorem 2.1. This can be attributed to the fact that the conditions presented in this paper are only sufficient, or to the fact that in practice what is observed is not necessarily a manifestation of uniform recovery. Rather, the practical results could be interpreted as success of Δ_p with high probability on either x or A.

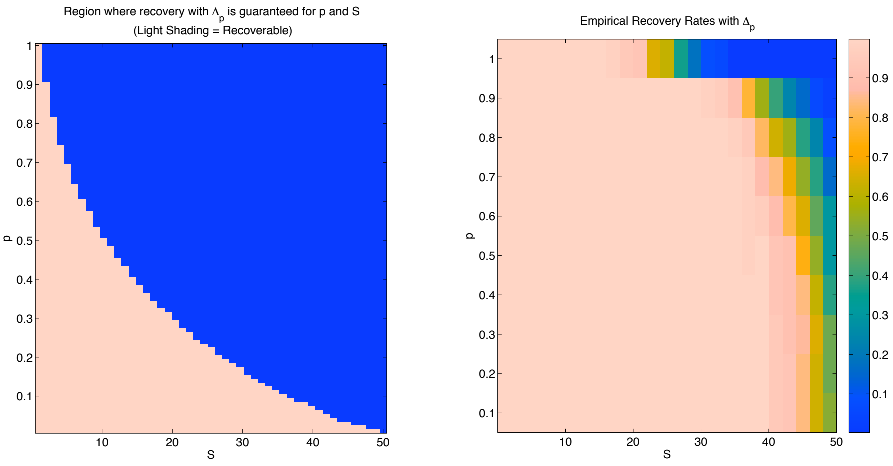

Figure 2. For a Gaussian matrix $A \in \mathbb{R}^{100\times 300}$, whose δ_S values are estimated via MC simulations, we generate the theoretical (left) and practical (right) phase-diagrams for reconstruction via ℓ^p minimization.

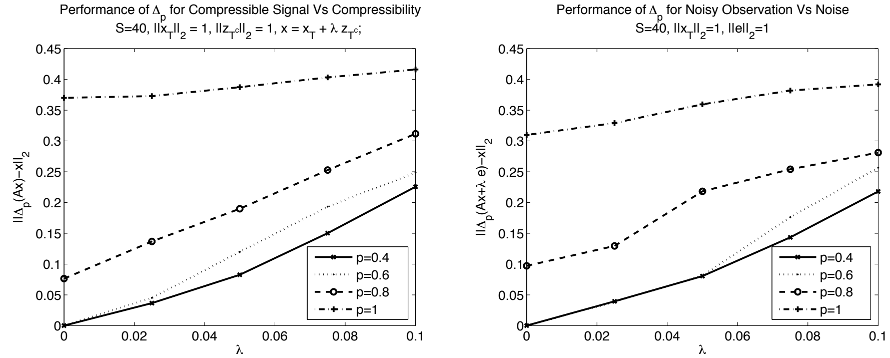

FIGURE 3. Reconstruction error with compressible signals (left), noisy observations (right). Observe the almost linear growth of the error in compressible signals and for different values of p, highlighting the instance optimality of the decoders. The plots were generated by averaging the results of 10 experiments with the same matrix A and randomized locations of the coefficients of x. The dashed line represents the total error $\|\Delta(Ax) - x\|_2$.

Next, we generate scenarios that allude to the conclusions of Theorem 2.7. To that end, we generate a signal composed of $x_T \in \Sigma_{40}^{300}$, supported on an index set T, and a signal z_{T^c} supported on T^c, where all the coefficients are drawn from the Gaussian distribution. Moreover, we normalize x_T and z_{T^c} so that $\|x_T\|_2 = \|z_{T^c}\|_2 = 1$. We now generate $x = x_T + \lambda z_{T^c}$ with increasing values of λ (starting from 0), thereby decreasing the compressibility of the signal x. For this experiment, we choose our measure matrix $A \in \mathbb{R}^{100\times 300}$ by drawing its columns uniformly from the sphere. For each value of λ we measure the reconstruction error $\|\Delta_p(Ax)-x\|_2$, and we repeat the process 10 times while randomizing the index set T but preserving the coefficient values. We report the averaged results in Figure 3 (left) for different values of p. Similarly, we generate noisy observations $Ax_T + \lambda e$, of a sparse signal $x_T \in \Sigma_{40}^{300}$ where $\|x_T\|_2 = \|e\|_2 = 1$ and increase the level of the noise starting from $\lambda = 0$. We then measure $\|\Delta_p(Ax_T + \lambda e) - x_T\|_2$ (again over 10 realizations where we randomize T) and report the averaged results in Figure 3 (right), for different values of p. In both cases, we observe the error increasing linearly. Moreover, when the signal is highly compressible or when the noise level is low, we observe that reconstruction with $\Delta_p, p < 1$ yields lower error than with $p = 1$.

Finally, in Figure 4, we plot the results of an experiment in which we generate signals whose sorted coefficients $x(j)$, decay according to some power law, i.e., $x(j) < Cj^{-1/q}$, where $0 < q < 1$ and with $\|x\|_2 = 1$, for various values of q. We then examine the recovery with Δ_p for different values of $p \in (0, 1)$. The results, obtained by averaging over 50 different experiments with different matrices A, indicate that values of $p \approx q$ provide the lowest reconstruction errors. Note that in Figure 4, we report the results in form of signal to noise ratios defined as

$$SNR = 20 \log_{10}\left(\frac{\|x\|_2}{\|\Delta(Ax) - x\|_2}\right).$$

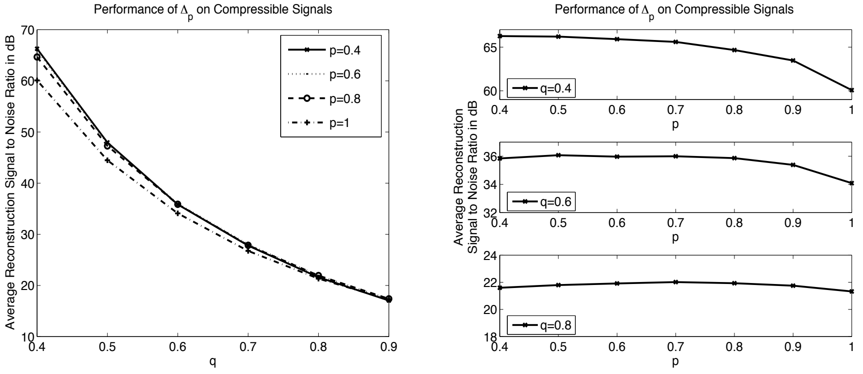

FIGURE 4. Reconstruction signal to noise ratios (in dB) of using Δ_p to recover signals whose sorted coefficients decay according to a power law ($x(j) = cj^{-1/q}, \|x\|_2 = 1$) as a function of q (left) and as a function of p (right). The presented results are averages of 50 experiments performed with different matrices in $\mathbb{R}^{100\times 200}$. Observe that for highly compressible signals, e.g., for $q = 0.4$, there is a 5 dB gain in using $p < 0.6$ as compared to $p = 1$. The performance advantage is about 2 dB for $q = 0.6$. As the signals become much less compressible, i.e., as we increase q to 0.9 the performances are almost identical.

4. PROOFS

Proof of Corollary 2.3. As we stated previously, we will prove Corollary 2.3 independently of Theorem 2.1. Our proof follows the proof by Candès et. al. [3] with modifications to account for the non-convexity of the ℓ^p norms. Let x be the original signal with its S nonzero coefficients supported on T_0 and let $x^* := \Delta_p^\epsilon(b)$. Let $h = x^* - x = h_{T_0} + h_{T_0^c}$ be the difference between the original and recovered signal, divided into two parts h_{T_0} with nonzero coefficients on T_0 and $h_{T_0^c}$ similarly supported on T_0^c. It can easily be shown that $\|h_{T_0^c}\|_p^p \le \|h_{T_0}\|_p^p$.

Divide T_0^c into sets $T_1, T_2, ...$ such that $\cup_{i\ge 1} T_i = T_0^c$, where T_1 supports the kS largest coefficients of $h_{T_0^c}$, T_2 supports the *second* kS largest coefficients of $h_{T_0^c}$, and so on. Let $T_{01} = T_0 \cup T_1$. Note that $Ah = A_{T_{01}} h_{T_{01}} + \sum_{i\ge 2} A_{T_i} h_{T_i}$. Since $\|Ax^* - b\|_2 < \epsilon$ then $\|Ah\|_2 \le 2\epsilon$. This leads to the following inequality

$$(2\epsilon)^p \ge \|Ah\|_2^p \ \ge \|A_{T_{01}} h_{T_{01}}\|_2^p - \sum_{i\ge 2} \|A_{T_i} h_{T_i}\|_2^p. \tag{21}$$

where we define $A_{T_{01}}$ as the sub-matrix of A whose columns correspond to T_{01}. The sub-matrices A_{T_i} are defined analogously. Since $\#(T_{01}) = (k+1)S$ and $\#(T_i) = kS$, then

$$(2\epsilon)^p \ge \left(1 - \delta_{(k+1)S}\right)^{p/2} \|h_{T_{01}}\|_2^p - \left(1 + \delta_{kS}\right)^{p/2} \sum_{i\ge 2} \|h_{T_i}\|_2^p. \tag{22}$$

What remains now is to bound $\sum_{i\ge 2} \|h_{T_i}\|_2^p$ and $\|h_{T_{01}}\|_2^p$ in terms of $\|h\|_2$. Observe that $|h_{T_0^c}|_{(l)}^p \le \frac{\sum_i |h_{T_0^c}|_{(i)}^p}{l} = \frac{\|h_{T_0^c}\|_p^p}{l}$, where $|h_{T_0^c}|_{(l)}$ is the l^{th} largest element of $|h_{T_0^c}|$.

Thus, taking the $\frac{1}{p}th$ power, squaring, and summing over $l \in T_{01}^c$ we get

$$\|h_{T_{01}^c}\|_2^2 \le \frac{\|h_{T_0^c}\|_p^2}{\frac{2-p}{p}(kS)^{2/p-1}} \le \frac{\|h_{T_0}\|_p^2}{\frac{2-p}{p}(kS)^{2/p-1}}. \tag{23}$$

Now, note that $|h_{T_{i+1}(u)}|^p \le \sum_{t\in T_i} |h_{T_i(t)}|^p/(kS) = \|h_{T_j}\|_p^p/(kS)\ \forall u \in T_{i+1}$. Taking the $\frac{1}{p}$th power, squaring, and summing over $u \in T_{i+1}$, we get $\|h_{T_{i+1}}\|_2^2 \le (kS)^{1-2/p}\|h_{T_i}\|_p^2$. Thus,

$$\sum_{i\ge 2} \|h_{T_i}\|_2^p \le (kS)^{p/2-1}\|h_{T_0}\|_p^p. \tag{24}$$

Noting that $\|h_{T_0}\|_p \le S^{1/p-1/2}\|h_{T_0}\|_2$, so $\|h_{T_0}\|_p^p \le S^{1-p/2}\|h_{T_{01}}\|_2^p$ we can now substitute in (22) to get

$$(2\epsilon)^p \ge \big(1-\delta_{(k+1)S}\big)^{p/2}\|h_{T_{01}}\|_2^p - \big(1+\delta_{kS}\big)^{p/2}\frac{\|h_{T_{01}}\|_2^p}{k^{1-p/2}}. \tag{25}$$

Using (23),

$$\|h\|_2^2 = \|h_{T_{01}}\|_2^2 + \|h_{T_{01}^c}\|_2^2 \le \|h_{T_{01}}\|_2^2(1+\frac{1}{k^{2/p-1}(2/p-1)}),$$

which when substituted in (25) yields the desired result.

Proof outline of Theorem 2.1. This proof is similar to the analogous proof in [3] and differs from the previous one by defining T_0 as the support set of the S largest coefficients of x , which is now no longer assumed sparse. This leads to $\|h_{T_0^c}\|_p^p \le \|h_{T_0}\|_p^p + 2\|x_{T_0^c}\|_p^p$. Using this inequality instead of the analogous one from the previous proof, the rest proceeds similarly with minor modifications to lead to the desired result. □

Proof of Lemma 2.5. To prove this Lemma, we will use some results of [30] and [18].

Theorem 4.1 ([30])**.** *Let $0 < \mu < 1/\sqrt{(}2)$ and let $C_1M(\ln(M))^\zeta \le N \le e^{CM}$ for some $\zeta > (1-2\mu^2)^{-1}$ and some constants $C, C_1 > 0$. There exists a constant $c > 0$ such that the set Ω_μ of those ω's where A_ω satisfies $LQ_1(\mu\sqrt{\frac{\ln N/M}{M}})$:*

$$A_\omega(B_1^N) \supset \mu\sqrt{\frac{\ln N/M}{M}}B_2^M \tag{26}$$

has probability $\ge 1-e^{-cM}$. The same is true for $\tilde{A}_\omega$.

We will also use the following adaptation of Lemma 2 from [18]. Note that we use $conv(K)$ to denote the convex-hull of a body K and following [18] and [23] we use $d_0(K,B)$ to denote the Banach-Mazur distance defined as

$$d_0(K,B) := inf_{u:\mathbb{R}^n\to\mathbb{R}^n}\{\lambda;\ K \subset uB \subset \lambda K\}$$

with the infimum taken over all linear operators u.

Lemma 4.2. *Let $0 < p < 1$, and let K be a p-convex body in $\mathbb{R}^n$, B_2^n be the ℓ^2-norm ball in $\mathbb{R}^n$, then*

$$d_0(K, B_2^n) \le C_p d_0(conv(K), B_2^n)^{(2/p-1)},$$

where

$$C_p = \left(2^{1-p} + \frac{(1-p)2^{1-p/2}}{p}\right)^{\frac{2-p}{p^2}} \left(\frac{1}{(1-p)\ln 2}\right)^{\frac{2-2p}{p^2}}.$$

We defer the proof of this lemma to the Appendix. □

Now, note that $\tilde{A}_\omega(B_1^N) \subset B_2^M$, because $(\tilde{A}_\omega(B_1^N) \subset B_2^M) \Leftrightarrow (\forall x \in \mathbb{R}^N$ such that $\|x\|_1 = 1, \|\tilde{A}_\omega x\|_2 \le 1) \Leftrightarrow (\|\tilde{A}_\omega\|_{1\to 2} = 1)$, but $\|\tilde{A}\|_{1\to 2}$ is the largest column norm of $\tilde{A}_\omega$, which is 1 by construction. Therefore we have

$$B_2^M \supset \tilde{A}_\omega(B_1^N) \supset \mu\sqrt{\frac{\ln N/M}{M}} B_2^M.$$

This implies that

$$d_0(\tilde{A}_\omega(B_1^N), B_2^M) \le \left(\mu\sqrt{\frac{\ln N/M}{M}}\right)^{-1}. \tag{27}$$

The next step is to note that $conv(B_p^N) = B_1^N$ and consequently $\tilde{A}_\omega\left(conv(B_p^N)\right) = conv\left(\tilde{A}_\omega(B_p^N)\right) = \tilde{A}_\omega(B_1^N)$. We can now invoke Lemma 4.2 to conclude that

$$\begin{aligned} d_0(\tilde{A}_\omega(B_p^N), B_2^M) &\le C_p d_0(conv(\tilde{A}_\omega(B_p^N)), B_2^M)^{\frac{2-p}{p}} \\ &= C_p d_0(\tilde{A}_\omega(B_1^N), B_2^M)^{\frac{2-p}{p}}. \end{aligned} \tag{28}$$

Finally, by using (27), we find that

$$d_0(\tilde{A}_\omega(B_p^N), B_2^M) \le C_p \left(\mu^2 \frac{\ln N/M}{M}\right)^{1/2-1/p}, \tag{29}$$

and consequently

$$\tilde{A}_\omega(B_p^N) \supset \frac{1}{C_p}\left(\mu^2 \frac{\ln N/M}{M}\right)^{(1/p-1/2)} B_2^M. \tag{30}$$

In other words, the matrix $\tilde{A}_\omega$ satisfies the $LQ_p(\alpha)$ property with $\alpha = 1/C_p\left(\mu^2 \frac{\log(N/M)}{M}\right)^{1/p-1/2}$ provided $\tilde{A}_\omega$ has the LQ_1 property, which we know by [30] is true with high probability. To see that the same is true for A_ω, note that there exists a set Ω with $p(\Omega) > 1 - e^{cM}$ such that $\|A_j(\omega)\|_2 < 2$ for $w \in \Omega$, for every column A_j of A_ω. Using this observation one can trace the above proof with minor modifications. □

Proof of Theorem 2.6. We start with the following lemma, the proof of which for $p < 1$ follows with very little modification from the analogous proof of Lemma 3.1 in [30] and shall be omitted.

Lemma 4.3. *Suppose that A satisfies $RIP(S, \delta)$ and $LQ_p\left(1/C_p(\mu^2/S)^{1/p-1/2}\right)$. Call $\gamma_p = \mu^{2/p-1}/C_p$ then $\forall x \in \mathbb{R}^N, \exists \tilde{x} \in \mathbb{R}^N$ such that:*

(i) $Ax = A\tilde{x}$
(ii) $\|\tilde{x}\|_p \le \frac{S^{1/p-1/2}}{\gamma_p}\|Ax\|_2$
(iii) $\|\tilde{x}\|_2 \le C(\delta, \gamma_p)\|Ax\|_2$

Here, $C(\delta,\gamma_p)=\frac{1}{\gamma_p}+\frac{\gamma_p(1-\delta)+1}{(1-\delta^2)\gamma_p}$.

We now proceed to prove Theorem 2.6. Our proof follows the steps of [30] and differs in the handling of the non-convexity of the ℓ^p norms when $p\in(0,1)$.

Proof of (i). Recall that A satisfies $LQ_p(\frac{\gamma_p}{S^{1/p-1/2}})$, so $\exists z\in\mathbb{R}^N$;
$Az=e$ and $\|z\|_p\le\frac{S^{1/p-1/2}}{\gamma_p}\|Ax\|_2$; $\|z\|_2\le C(\delta,\gamma_p)\|e\|_2=C_1\|e\|_2$.

Now, $A(x+z)=Ax+e$, and Δ is $(2,p)$ instance optimal. Thus,

$$\|\Delta(A(x)+e)-(x+z)\|_2\le C'_{2,p}\frac{\sigma_S(x+z)_{\ell^p}}{S^{1/p-1/2}},$$

and consequently

$$\begin{aligned}
\|\Delta(A(x)+e)-(x)\|_2 &\le \|z\|_2+C'_{2,p}\frac{\sigma_S(x+z)_{\ell^p}}{S^{1/p-1/2}}\\
&\le C_1\|e\|_2+C'_{2,p}\frac{\sigma_S(x+z)_{\ell^p}}{S^{1/p-1/2}}\\
&\le C_1\|e\|_2+2^{1/p-1}C'_{2,p}\frac{\sigma_S(x)_{\ell^p}+\|z\|_p}{S^{1/p-1/2}}\\
&\le C_1\|e\|_2+2^{1/p-1}C'_{2,p}\frac{\sigma_S(x)_{\ell^p}}{S^{1/p-1/2}}+2^{1/p-1}C'_{2,p}\frac{\|Az\|_2}{\gamma_p}\\
\implies \|\Delta(A(x)+e)-(x)\|_2 &\le \left(C_1+2^{1/p-1}C'_{2,p}/\gamma_p\right)\|e\|_2+2^{1/p-1}C'_{2,p}\frac{\sigma_S(x)_{\ell^p}}{S^{1/p-1/2}}.
\end{aligned}$$

Proof of (ii). As in the analogous proof of [30], (ii) can be seen as a special case of (iii), with $e=0$. We therefore turn to proving (iii).

Proof of (iii). Once again, we utilize the LQ_p property to deduce the following preliminary facts,

$$\exists v, Av=e; \|v\|_p\le s^{1/p-1/2}/\gamma_p\|e\|_2, \|v\|_2\le C_1\|e\|_2, \text{and}$$

$$\exists z, Az=Ax_{T_0^c}; \|z\|_p\le s^{1/p-1/2}/\gamma_p\|Ax_{T_0^c}\|_2, \|z\|_2\le C_1\|Ax_{T_0^c}\|_2.$$

Here T_0 supports the S largest coefficients of x. Similar to the previous part we can see that $A(x_{T_0}+z+v)=Ax+e$ and by the hypothesis of $(2,p)$ instance optimality of Δ, we have

$$\|\Delta(Ax+e)-(x_{T_0}+z+v)\|_2\le C'_{2,p}\frac{\sigma_S(x_{T_0}+z+v)_{\ell^p}}{S^{1/p-1/2}}.$$

Consequently by observing that $x_{T_0}=x-x_{T_0^c}$ and using the triangle inequality, we can write

$$\begin{aligned}
\|\Delta(A(x)+e)-(x)\|_2 &\le \|x_{T_0^c}-z-v\|_2+C'_{2,p}\frac{\sigma_S(x_{T_0}+z+v)_{\ell^p}}{S^{1/p-1/2}}\\
&\le \|x_{T_0^c}-z-v\|_2+2^{1/p-1}(C'_{2,p})\left(\frac{\|z\|_p+\|v\|_p}{S^{1/p-1/2}}\right)\\
&\le \sigma_S(x)_{\ell^2}+\|z\|_2+\|v\|_2+2^{1/p-1}C'_{2,p}\left(\frac{\|Ax_{T_0^c}\|_2}{\gamma_p}+\frac{\|e\|_2}{\gamma_p}\right)\\
&\le \sigma_S(x)_{\ell^2}+\left(C_1+2^{1/p-1}\frac{C'_{2,p}}{\gamma_p}\right)(\|e\|_2+\|Ax_{T_0^c}\|_2). \qquad (31)
\end{aligned}$$

This concludes the proof of this theorem. □

Proof of Theorem 2.7. By Theorem 2.1, Δ_p is $(2,p)$ instance optimal if A satisfies an appropriate RIP. But, if $S \approx \alpha M/\log N$, then we have $RIP(S,\delta)$ and $LQ_p(\gamma_p/S^{1/p-1/2})$, with high probability. Therefore we can apply part (i) of Theorem 2.6 to get the first part of this theorem, i.e.,

$$\|\Delta(A(x)+e)-(x)\|_2 \le \left(C_1 + \frac{(2C^{(2)})^{1/p}}{2\gamma_p}\right)\|e\|_2 + (2C^{(2)})^{1/p}\frac{\sigma_S(x)_{\ell^p}}{2S^{1/p-1/2}}.$$

Moreover, note that for any x, $\|Ax_{T_0^c}\|_2 \le 2\|x_{T_0^c}\|_2 = 2\sigma_S(x)_{\ell^p}$ with probability $> 1-e^{-cM}$ on the draw of A. Combined with the part (iii) of Theorem 2.6, we see that the following holds with high probability.

$$\|\Delta(A(x)+r)-(x)\|_2 \le \left(1+2C_1+\frac{(2C^{(2)})^{1/p}}{\gamma_p}\right)\sigma_S(x)_{\ell^2}+\left(C_1+\frac{(2C^{(2)})^{1/p}}{2\gamma_p}\right)\|e\|_2.$$

□

5. Appendix: Proof of Lemma 4.2

In this section we provide the proof of Lemma 4.2. We provide this proof for the sake of completeness and also because we explicitly calculate the optimal constants involved. Thus, let us first introduce some notation used in [18] and [23]. For $q \in (1,2]$ and a body $K \in \mathbb{R}^n$, define the gauge functional $\|x\|_K = \inf\{t>0;\ x \in tK\}$. Also, define $T_q(K)$ as the smallest constant C such that $\forall m$, $x_1,...,x_m \in K$

$$\inf_{\epsilon_i=\pm 1}\left\{\|\sum_{i=1}^{m}\epsilon_i x_i\|_K\right\} \le Cm^{1/q}$$

holds. We call a body K *p-convex* if for any $x,y \in K$ and any $\lambda,\mu \in [0,1], \lambda^p+\mu^p=1$, $\lambda x+\mu y \in K$. Given a p-convex body K, define $\alpha_m = \alpha_m(K) = \sup\{\frac{\|\sum_{i=1}^m x_i\|_K}{m};\ x_i \in K, i \le m\}$. Note that $\alpha_m < m^{-1+1/p}$.

Let $\delta_K = d_0(K, conv(K)) = \inf\{\lambda>0;\ convK \subset \lambda K\} = \sup\alpha_m$, where the last equality is by a result of [24]. We will now prove Lemma 4.2 in its original more general form [18].

Lemma 5.1. *Let $p \in (0,1), q \in (1,2]$, K be a p-convex body and B be a symmetric body with respect to the origin. Define $\phi = \frac{1/p-1/q}{1-1/q}$, then*

$$d_0(K,B) \le C_{p,q}T_q(B)^{\phi-1}d_0(conv(K),B)^{\phi}.$$

Proof. Let $d=d_0(K,B)$ and $T=T_q(B)$. We can assume that $(1/d)B \subset K \subset B$. Let m be a positive integer and let $x_i, i \in 1,2,...,2^m$ be a collection of points in K. Then, $x_i \in B$ and by the definition of T, $\exists$ a choice of signs ϵ_i so that $\|\sum_{i=1}^{2^m}\epsilon_i x_i\|_B \le T2^{m/q}$. Since B is symmetric, we can assume that $D=\{i;\ \epsilon_i=1\}$ has $\#D > 2^{m-1}$. Now we can write

$$\begin{aligned}\|\sum_{i=1}^{2^m} x_i\|_K^p &= \|\sum_{i=1}^{2^m}\epsilon_i x_i + 2\sum_{i\notin D} x_i\|_K^p \le d^p\|\sum_{i=1}^{2^m}\epsilon_i x_i\|_B^p + 2^p\|\sum_{i\notin D} x_i\|_K^p \\ &\le d^pT^p2^{mp/q} + 2^{mp}\alpha_{2^{m-1}}^p\end{aligned}$$

Thus by taking the supremum over all possible x_i's and dividing by 2^{mp}, we obtain, for any m,

$$\alpha_{2^m}^p \leq d^p T^p 2^{mp/q-mp} + \alpha_{2^{m-1}}^p.$$

By applying this inequality for $m-1, m-2, ..., k$, we obtain the following inequality for any $k \leq m$

$$\alpha_{2^m}^p \leq d^p T^p \sum_{i=k+1}^{\infty} 2^{ip(1-1/q)} + \alpha_{2^k}^p \leq d^p T^p \frac{2^{-kp(1-1/q)}}{p(1-1/q)\ln 2} + 2^{k(1-p)}. \tag{32}$$

Note that since $\delta_K = \sup \alpha_m$, we now want to minimize the right hand side in (32) by choosing k appropriately. Since we can freely choose m as large as necessary, we obtain the optimal value of k, say k^*, by taking the derivative with respect to k and setting it to zero. However, k^* is not necessarily an integer. On the other hand, by choosing $k = k^* + 1$, we can bound the right hand side of (32) which is monotonic for $k > k^*$. This yields the following estimate for δ_K.

$$\delta_K \leq (dT)^{\frac{(1-p)}{(1-p/q)}} \left(2^{1-p} + 2^{-p(1-1/q)} \frac{1-p}{p(1-1/q)}\right)^{1/p} \frac{1}{((1-p)\ln 2)^{\frac{1/p-1}{1-p/q}}}. \tag{33}$$

The result follows from the inequality $d_0(K, B) \leq \delta_K d_0(conv(K), B)$ with

$$C_{p,q} = \left(2^{1-p} + 2^{-p(1-1/q)} \frac{1-p}{p(1-1/q)}\right)^{\frac{1-p/q}{p^2(1-1/q)}} \left(\frac{1}{(1-p)\ln 2}\right)^{\frac{1/p-1}{p(1-1/q)}}.$$

□

Finally, to adapt the above proof to obtain Lemma 4.2, observe that in our case $B = B_2^n$, hence by choosing $q = 2$, we have $T_2 = 1$, and

$$C_p = \left(2^{1-p} + \frac{(1-p)2^{1-p/2}}{p}\right)^{\frac{2-p}{p^2}} \left(\frac{1}{(1-p)\ln 2}\right)^{\frac{2-2p}{p^2}}.$$

Acknowledgment

The authors would like to thank Michael Friedlander, Gilles Hennenfent, Felix Herrmann, and Ewout Van Den Berg for many fruitful discussions. This work was finalized during an AIM workshop. We thank the American Institute of Mathematics for its hospitality. R. Saab would also like to thank Rabab Ward for her immense support.

References

[1] R. Baraniuk, M. Davenport, R. DeVore, M. Wakin, A Simple Proof of the Restricted Isometry Property for Random Matrices, Constructive Approximation (To appear).

[2] E. Candès, The restricted isometry property and its implications for compressed sensing, CR Math. Acad. Sci. Paris, Ser. I 346 (2008) 589–592.

[3] E. J. Candès, J. Romberg, T. Tao, Signal recovery from incomplete and inaccurate measurements, Comm. Pure Appl. Math. 59 (8) (2005) 1207–1223.

[4] E. J. Candès, J. Romberg, T. Tao, Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information, IEEE Trans. Inform. Theory 52 (2006) 489–509.

[5] E. J. Candès, T. Tao, Decoding by linear programming., IEEE Trans. Inform. Theory 51 (12) (2005) 489–509.

[6] E. J. Candès, T. Tao, Near-optimal signal recovery from random projections: universal encoding strategies?, IEEE Trans. Inform. Theory 52 (12) (2006) 5406–5425.

[7] R. Chartrand, Exact reconstructions of sparse signals via nonconvex minimization, IEEE Signal Process. Lett. 14 (10) (2007) 707–710.

[8] R. Chartrand, V. Staneva, Restricted isometry properties and nonconvex compressive sensing, Inverse Problems 24 (035020).

[9] S. Chen, D. Donoho, M. Saunders, Atomic decomposition by basis pursuit, SIAM Journal on Scientific Computing 20 (1) (1999) 33–61.
URL `citeseer.ist.psu.edu/chen98atomic.html`

[10] A. Cohen, W. Dahmen, R. DeVore, Compressed sensing and best k-term approximation, Journal of the American Mathematical Society (to appear).
URL `http://www.ams.org/jams/0000-000-00/S0894-0347-08-00610-3/S0894-0347-08-00610-3.pdf`

[11] I. Daubechies, R. DeVore, M. Fornasier, C. Gunturk, Iteratively re-weighted least squares minimization for sparse recovery, Arxiv preprint arXiv:0807.0575.

[12] M. Davies, R. Gribonval, Restricted Isometry Constants where ℓ^p sparse recovery can fail for $0 < p \leq 1$.
URL `http://hal.inria.fr/docs/00/29/90/13/PDF/PI-1899.pdf`

[13] D. Donoho, Compressed sensing., IEEE Transactions on Information Theory 52 (4) (2006) 1289–1306.

[14] D. Donoho, M. Elad, Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization, Proc. Natl. Acad. Sci. USA 100 (5) (2003) 2197–2202.

[15] D. Donoho, X. Huo., Uncertainty principles and ideal atomic decomposition, IEEE Trans. Inf. Theory 47 (2001) 2845–2862.

[16] M. Figueiredo, R. Nowak, S. Wright, Gradient Projection for Sparse Reconstruction: Application to Compressed Sensing and Other Inverse Problems, Selected Topics in Signal Processing, IEEE Journal of 1 (4) (2007) 586–597.

[17] S. Foucart, M. Lai, Sparsest solutions of underdetermined linear systems via ℓ^q-minimization for $0 < q \leq 1$, Submitted to Applied and Computational Harmonic Analysis.

[18] Y. Gordon, N. Kalton, Local structure theory for quasi-normed spaces, Bull. Sci. Math. 118 (1994) 441–453.

[19] R. Gribonval, R. M. Figueras i Ventura, P. Vandergheynst, A simple test to check the optimality of sparse signal approximations, in: Proc. IEEE Intl. Conf. Acoust. Speech Signal Process (ICASSP'05), vol. 5, 2005, pp. V/717 – V/720.

[20] R. Gribonval, R. M. Figueras i Ventura, P. Vandergheynst, A simple test to check the optimality of sparse signal approximations, EURASIP Signal Processing, special issue on Sparse Approximations in Signal and Image Processing 86 (3) (2006) 496–510.

[21] R. Gribonval, M. Nielsen, On the strong uniqueness of highly sparse expansions from redundant dictionaries, in: Proc. Int Conf. Independent Component Analysis (ICA'04), LNCS, Springer-Verlag, Granada, Spain, 2004.

[22] R. Gribonval, M. Nielsen, Highly sparse representations from dictionaries are unique and independent of the sparseness measure, Appl. Comput. Harm. Anal. 22 (3) (2007) 335–355.

[23] O. Guedon, A. Litvak, Euclidean projections of p-convex body, GAFA, Lecture Notes in Math 1745 (2000) 95–108.

[24] N. Peck, Banach-mazur distances and projections on p-convex spaces, Math. Z. 177 (1) (1981) 131–142.

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

[26] J. Tropp, Recovery of short, complex linear combinations via l^1 minimization, IEEE Transactions on Information Theory 51 (4) (2005) 1568–1570.

[27] E. van den Berg, M. Friedlander, In pursuit of a root, UBC Computer Science Technical Report TR-2007-16.
URL `http://www.optimization-online.org/DB_FILE/2007/06/1708.pdf`

[28] P. Vandergheynst, P. Frossard, Efficient image representation by anisotropic refinement inmatching pursuit, Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on 3.

[29] R. Ventura, P. Vandergheynst, P. Frossard, Low rate and scalable image coding with redundant representations, IEEE Transactions on Image Processing.

[30] P. Wojtaszczyk, Stability and instance optimality for gaussian measurements in compressed sensing, Preprint.

Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, B.C. Canada V6T 1Z4
E-mail address: rayans@ece.ubc.ca

Department of Mathematics, University of British Columbia, Vancouver, B.C. Canada V6T 1Z2
E-mail address: oyilmaz@math.ubc.ca