Global Linear and Local Superlinear Convergence of IRLS for Non-Smooth Robust Regression

Anonymous Author(s)
Affiliation
Address
email

Abstract

We advance both the theory and practice of robust ℓ_p-quasinorm regression for $p \in (0, 1]$ by using novel variants of *iteratively reweighted least-squares* (IRLS) to solve the underlying non-smooth problem. In the convex case, $p = 1$, we prove that this IRLS variant converges globally at a linear rate under a mild, deterministic condition on the feature matrix called the *stable range space property*. In the non-convex case, $p \in (0, 1)$, we prove that under a similar condition, IRLS converges locally to the global minimizer at a superlinear rate of order $2 - p$; the rate becomes quadratic as $p \to 0$. We showcase the proposed methods in three applications: real phase retrieval, regression without correspondences, and robust face restoration. The results show that (1) IRLS can handle a larger number of outliers than other methods, (2) it is faster than competing methods at the same level of accuracy, (3) it restores a sparsely corrupted face image with satisfactory visual quality.

1 Introduction

Given a feature matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ with $m \gg n$ and a response vector $\boldsymbol{y} \in \mathbb{R}^m$, the problem

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} \left\| \boldsymbol{A}\boldsymbol{x} - \boldsymbol{y} \right\|_p \tag{1}$$

of ℓ_p-regression has a variety of different applications depending on the choice of $\|\boldsymbol{v}\|_p = \left(\sum_i |\boldsymbol{v}_i|^p\right)^{1/p}$. While $p = 2$ corresponds to standard linear regression, the choice of $p > 2$ arises naturally in semi-supervised learning on graphs [1, 2, 3], and a lot of activity has been dedicated recently to the computational complexity analysis for the case $p > 1$ [4, 5, 6].

In this paper, we assume that there is a coefficient vector $\boldsymbol{x}^* \in \mathbb{R}^n$ such that the residual $\boldsymbol{r}^* := \boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{y}$ is k-sparse, in which case a choice of $p \in (0, 1]$ is of interest. Indeed, for the convex and non-smooth case $p = 1$, (1) is known as *least absolute deviation*, which dates back to the time of Boscovich around the turn of the 18th century. Since then, it has been known intuitively that (1) ($p = 1$) is robust to *large but few measurement errors* [7], i.e., that (1) is robust to outliers.

The algorithmic and theoretical understanding of (1) for $p = 1$ has been of long-standing interest to statisticians, which has led to a vast literature spanning from the classic 1964 paper of Huber [8] to recent contributions [9, 10, 11, 12, 13, 14, 15, 16, 17, 18]. That being said, for $p = 1$, there is a close, but somewhat underexplored connection between (1) and the *basis pursuit* problem [19], and more specifically, the theoretical [20, 21, 22] and algorithmic aspects [23, 24, 25] of compressed sensing [26] (see also Section 2.1). For $p \in (0, 1)$, problem (1) is non-smooth and non-convex, and generally less well understood. To the best of our knowledge, the only paper that considered (1) with

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

Algorithm 1: IRLS for ℓ_p-Regression ($\texttt{IRLS}_p$)

Input: $\boldsymbol{A} = [\boldsymbol{a}_1, \dots, \boldsymbol{a}_m]^\top \in \mathbb{R}^{m\times n}$, $\boldsymbol{y} = [y_1, \dots, y_m]^\top \in \mathbb{R}^m$, $p \in (0, 1]$;

Weight initialization $\boldsymbol{w}^{(0)} \leftarrow [w_1^{(0)}; \dots; w_m^{(0)}] \in \mathbb{R}^m$; // Instead, one can initialize some vector $\boldsymbol{x}^{(1)}$.

For $t \leftarrow 0, 1, \dots$:

$$\boldsymbol{x}^{(t+1)} \leftarrow \operatorname*{argmin}_{\boldsymbol{x}\in\mathbb{R}^n} \sum\nolimits_{i=1}^m w_i^{(t)} (\boldsymbol{a}_i^\top \boldsymbol{x} - \boldsymbol{y})^2 \tag{2}$$

// Update $\epsilon^{(t+1)}$ suitably as per certain rules; see Section 2.2 for details.

$$w_i^{(t+1)} \leftarrow \max\left\{|\boldsymbol{a}_i^\top \boldsymbol{x}^{(t+1)} - y_i|, \epsilon^{(t+1)}\right\}^{p-2} \quad \forall i = 1, \dots, m \tag{3}$$

$p \in (0, 1)$ is [27], where a condition (based on *restricted isometry constants*) that guarantees exact recovery of $\boldsymbol{x}^*$ from (1) was presented. Other related works come either from the compressed sensing literature, where an ℓ_p (and noisy) version of basis pursuit has been considered as a sparsity-promoting formulation [28, 29, 30, 24, 31, 32, 33, 34], or from the literature of matrix recovery/completion, where the *Schatten-p* norm has come into play as a non-convex surrogate for rank minimization [35, 36, 37, 38, 39]. A key message there is that ℓ_p (Schatten-p) minimization with $p \in (0, 1)$ offers better information-theoretic properties (e.g., requires fewer samples for exact recovery) than $p = 1$.

Here, we study an *iteratively reweighted least-squares* method ($\texttt{IRLS}_p$) to solve (1) with $p \in (0, 1]$. As listed in Algorithm 1, $\texttt{IRLS}_p$ alternates between solving a weighted least-squares problem (2) and updating the weights (3); see Section 2.2 for more elaboration on IRLS. The simplicity of this idea (with its impressive performance) justifies its popularity in many machine learning [40, 41, 42, 43] and computer vision [44, 45, 46, 47] applications. Based on recent advances on IRLS [24, 25], we make several contributions for understanding (1) and $\texttt{IRLS}_p$. We state our contributions next.

The Stable Range Space Property (Section 3.1). We put forward the use of the (stable) *range space property* (RSP) for studying the robust regression problem (1). In the case $p = 1$, the stable RSP was proposed in a different context by [48] to analyze a compressed sensing algorithm solving a weighted basis pursuit problem at each iteration. In analogy to the *null space property* in compressed sensing [22, 26], we show that (1) the RSP is a necessary and sufficient condition for guaranteeing that the ℓ_p-regression problem in (1) admits $\boldsymbol{x}^*$ as its unique solution (Proposition 2), and (2) if $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ has i.i.d. $\mathcal{N}(0, 1)$ entries and if m is large enough, then the (stable) RSP holds with high probability (Proposition 3). This justifies the use of the stable RSP as our core assumption for analysis.

Global Linear Convergence (Section 3.2). We prove in Theorem 1 that, under a stable RSP assumption and with a suitable update rule for $\epsilon^{(t)}$, the $\texttt{IRLS}_p$ Algorithm 1 with $p = 1$ converges linearly to the ground-truth $\boldsymbol{x}^*$ from any initial weight $\boldsymbol{w}^{(0)}$ (or equivalently from any initial point $\boldsymbol{x}^{(1)}$). Note that (accelerated) first-order methods (e.g., (sub-)gradient descent and proximal algorithms) can also solve the convex and non-smooth problem (1) with $p = 1$, but in general they exhibit (global) sub-linear rates at best [49, 50]. To our knowledge, [15] is the only paper that claims global linear convergence of a different IRLS variant for (1) with $p = 1$; we will compare with the results of [15] in Section 3.2. On the other hand, Theorem 1 is inspired by the IRLS algorithm of [25] for basis pursuit; based on [25], we suitably modify their proof strategy, and thus obtain a faster linear rate.

Local Superlinear Convergence (Section 3.3). We prove in Theorem 2 that, under a stable RSP assumption and with $\epsilon^{(t)}$ suitably updated, the $\texttt{IRLS}_p$ Algorithm 1 with $p \in (0, 1)$ converges to $\boldsymbol{x}^*$superlinearly, provided that $\boldsymbol{x}^{(1)}$ falls into a certain neighborhood of $\boldsymbol{x}^*$. To the best of our knowledge, no similar result exists for ℓ_p-regression (1). While Theorem 2 is inspired by the IRLS algorithm of [24] for ℓ_p-basis pursuit, their IRLS method does not work well for small p (Section 2.2), and unlike in our work, their radius of local convergence diminishes greatly as $p \to 0$ or $m \to \infty$.

Applications (Section 4). We illustrate the performance of $\texttt{IRLS}_p$ for *real phase retrieval* [51, 52], *linear regression without correspondences* [53, 54, 55, 56], and *face restoration from sparsely corrupted measurements* [57, 58, 59]. For real phase retrieval (Section 4.1), we show that $\texttt{IRLS}_{0.1}$

needs only $m = 2n - 1$ measurements to recover $\boldsymbol{x}^*$ up to sign, with an additional assumption. Theoretically, even brute-force can fail to identify $\pm\boldsymbol{x}^*$ for fewer than $2n - 1$ measurements [60]. Empirically, many methods, including *Kaczmarz* [61, 62, 52], *PhaseLamp* [63], *truncated Wirtinger flow* [64], and *coordinate descent* [65], failed with $m = 2n - 1$ Gaussian measurements (Figure 2a). For linear regression without correspondences (Section 4.2), we show in Figures 2b-2c that, $\texttt{IRLS}_{0.1}$ is uniformly faster (20-100x) and more accurate than *PDHG* [66] (merged into Google or-tools) and the commercial solver *Gurobi* [67], both of which solve (1) with $p = 1$ as a linear program, and also than *subgradient descent* implemented by Beck & Guttmann-Beck [68]. For face restoration (Section 4.3), we present both quantitative and qualitative results on the Extended Yale B dataset [69].

2 Background

2.1 Connection of Robust Regression and Compressed Sensing

The ℓ_p-regression problem (1) has a natural correspondence to the following sparse recovery problem:

$$\min_{\boldsymbol{r}\in\mathbb{R}^m} \|\boldsymbol{r}\|_p \qquad \text{s.t.} \qquad \boldsymbol{D}\boldsymbol{r} = \boldsymbol{z} \tag{4}$$

In (4), $\boldsymbol{D}$ is a $(m-n)\times m$ matrix, and $\boldsymbol{z} \in \mathbb{R}^{m-n}$. For $p = 1$, (4) is also called *basis pursuit* [70]. Problems (1) and (4) are known to be related in the following sense (as implicitly stated in [19, 27]):

Proposition 1. *Suppose that the range space of $\boldsymbol{A}$ is equal to the null space of $\boldsymbol{D}$, and that $\boldsymbol{z} = -\boldsymbol{D}\boldsymbol{y}$. If $\boldsymbol{x}_1$ globally minimizes (1), then $\boldsymbol{r}_1 := \boldsymbol{A}\boldsymbol{x}_1 - \boldsymbol{y}$ globally minimizes (4). On the other hand, if $\boldsymbol{r}_2$ globally minimizes (4), then there exists some $\boldsymbol{x}_2$ with $\boldsymbol{r}_2 = \boldsymbol{A}\boldsymbol{x}_2 - \boldsymbol{y}$ that globally minimizes (1).*

Proposition 1 sheds light on how we "transfer", with new insights, from the analysis of [24, 25] for (4) to results for (1). In what follows, we treat [24, 25] in the context of robust regression and highlight the contribution of our work relative to [24, 25] and other existing results, whenever possible.

2.2 Iteratively Reweighted Least-Squares: The Basics and New Insights

We first discuss two different updating rules for the *smoothing parameter* $\epsilon^{(t)}$ of Algorithm 1: the *fixed* rule and the *dynamic* rule. We then consider the weight updating strategy (3). Along the way we use experimental results to illustrate ideas (for synthetic data generation, see the appendix). In doing so, we intend to provide a review of the state-of-the-art on the variants of IRLS.

Fixed Smoothing Parameter. Instead of updating $\epsilon^{(t)}$ at each iteration, most works on IRLS use a fixed and small positive number, e.g., $\epsilon := \epsilon^{(t)} = 0.0001$ for each t [40, 71, 41, 43, 45, 47, 46, 72]. The intuition is to avoid division by the potentially very small residual $|\boldsymbol{a}_i^\top \boldsymbol{x}^{(t)} - y_i|$. However, this common practice of fixing ϵ comes with at least three issues. First, IRLS with a fixed $\epsilon > 0$ converges only to an "ϵ-approximate" point, not exactly the global minimizer ([73, 74, 40, 71], Figure 1a). Second, even if setting ϵ small (e.g., $\epsilon = 10^{-15}$) could lead to an accurate enough solution for $p = 1$ (Figure 1a), it can fail for $p < 1$ (Figure 1b). Finally, it makes obtaining a global linear convergence rate guarantee difficult: We have not seen any theory about a global linear rate of IRLS with fixed ϵ.

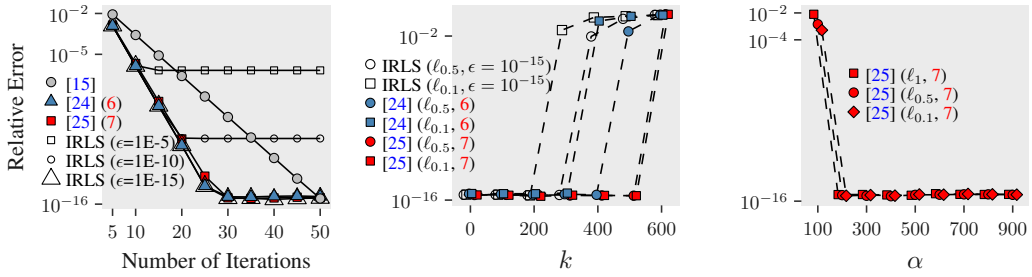
Dynamic Smoothing Parameter. Researchers have used different insights to reach the consensus of dynamically updating $\epsilon^{(t)}$ [24, 15, 75, 25, 76]. The first insight is that convergence to global minimizers ensues if $\epsilon^{(t)}$ is suitably decreased to 0 [24]. With $\beta \in (0, 1)$, one such decreasing rule is

$$\epsilon^{(t+1)} \leftarrow \beta\epsilon^{(t)} \text{ if certain conditions are satisfied, or keep } \epsilon^{(t+1)} \leftarrow \epsilon^{(t)} \text{ otherwise [15, 76].} \tag{5}$$

Figure 1a depicts the performance of [15] with an arbitrary choice $\beta = 0.5$ and $\epsilon^{(0)} = 1$. Also shown in Figure 1a are the IRLS methods with update rules of [24] and [25], which we discuss next.

The basis pursuit paper [24] also proposed a dynamic updating rule for $\epsilon^{(t)}$. In our context (1), it sets

$$\epsilon^{(0)} \leftarrow \infty, \quad \epsilon^{(t+1)} \leftarrow \min\left\{\epsilon^{(t)}, [\boldsymbol{r}^{(t+1)}]_{\alpha+1}/m\right\}, \qquad \boldsymbol{r}^{(t+1)} := \boldsymbol{A}\boldsymbol{x}^{(t+1)} - \boldsymbol{y} \in \mathbb{R}^m, \tag{6}$$



(a) $p = 1$ (b) $p < 1$ (c) Sensitivity to α ($k = 200$)

Figure 1: Figure 1a: The relative error $\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\|_2 / \|\boldsymbol{x}^*\|_2$ at each iteration t of different IRLS variants for ℓ_1-robust regression with p varying ($k = 200$). Figure 1b: The relative error of $\texttt{IRLS}_p$ Algorithm 1 and that of [24] for ℓ_p-regression (50 iterations). Figure 1c: The sensitivity of (6) and (7) to mis-specification of α (50 iterations). In Figure 1, we use $m = 1000, n = 10$, and 20 trials.

where the hyper-parameter[1] α is a non-negative integer, and $[\boldsymbol{r}^{(t+1)}]_{\alpha+1}$ is the $(\alpha+1)$-th largest element of the residual $\boldsymbol{r}^{(t+1)}$. Clearly, (6) creates a non-increasing sequence of smoothing parameters $\epsilon^{(t)}$, and the decay rate of $\epsilon^{(t)}$ is adaptive to the data (in this case to the residual $\boldsymbol{r}^{(t+1)}$). While at first glance it is not clear how fast $\epsilon^{(t)}$ decays, [24] showed that the decay rate of $\epsilon^{(t)}$ in (6) is locally linear for $p = 1, \alpha = k$ under mild conditions, in accordance with the decay of the objective values.

The final dynamic update rule for the parameter $\epsilon^{(t)}$ is from [25][1] and improves upon (6) [24] via:

$$\begin{aligned} \epsilon^{(0)} \leftarrow \infty, \qquad \epsilon^{(t+1)} \leftarrow \min\left\{\epsilon^{(t)}, \sigma^{(t+1)}/m\right\} \\ \text{where } \sigma^{(t+1)} \leftarrow \min\left\{\left\|\boldsymbol{r}^{(t+1)} - \boldsymbol{z}\right\|_1 : \boldsymbol{z} \in \mathbb{R}^m \text{ is } \alpha\text{-sparse}\right\} \end{aligned} \tag{7}$$

Observe that (7) computes the best α-term approximation $\sigma^{(t+1)}$ of the residual $\boldsymbol{r}^{(t+1)}$, which is in general larger than $[\boldsymbol{r}^{(t+1)}]_{\alpha+1}$ of (6), while both update rules (6) and (7) rely on the hyper-parameter α. Ideally, if $\alpha = k$, and if IRLS with (6) or (7) converges to $\boldsymbol{x}^*$, then both $[\boldsymbol{r}^{(t+1)}]_{\alpha+1}$ and $\sigma^{(t+1)}$ will approach 0 for large enough t. While update rules (6) and (7) perform similarly for ℓ_1-regression (Figure 1a), (6) fails more easily for small p than (7) (Figure 1b). Finally, we note that overestimating the sparsity level k by setting $\alpha > k$ deteriorates the performance of (7) only slightly (Figure 1c).

Summary. In Section 2.2 we delivered two take-away messages: (1) IRLS with a fixed $\epsilon^{(t)}$ finds some approximate solution, sometimes good enough, (2) dynamically updating $\epsilon^{(t)}$ as per (5)-(7) leads to local [24] or global [15, 25] linear convergence guarantees. In view of Figure 1, we will next consider the $\texttt{IRLS}_p$ Algorithm 1 with update rule (7) for $\epsilon^{(t)}$, and analyze its convergence rates.

3 Convergence Theory of IRLS for Robust Regression

In Section 3.1 we introduce the stable range space property and justify it as our core assumption. Under this assumption, we prove the global linear convergence (Theorem 1) and local superlinear convergence (Theorem 2) of the $\texttt{IRLS}_p$ Algorithm 1 in Sections 3.2 and 3.3, respectively.

3.1 The Stable Range Space Property

Definition 1 (Stable Range Space Property)**.** A matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ is said to satisfy the *range space property (RSP)* of order k with constant $p \in (0, 1]$, or the (p, k)-RSP for short, if the following holds

[1] In [24] and [25], α is the sparsity level k, but k might be unknown in practice, so we treat it as a hyper-parameter. That being said, in the experiments we set $\alpha = k$ by default, for simplicity and in light of Figure 1c.

for any vector $\boldsymbol{d}$ in the range space of $\boldsymbol{A}$ and any set $S \subset \{1, \ldots, m\}$ of cardinality at most k:

$$\sum_{i \in S} |d_i|^p < \sum_{i \in S^c} |d_i|^p. \tag{8}$$

The *stable RSP* of $\boldsymbol{A}$ is defined in the same way as the RSP except that we now require

$$\sum_{i \in S} |d_i|^p \leq \eta \sum_{i \in S^c} |d_i|^p \tag{9}$$

for some $\eta \in (0, 1)$. We write "(p, k, η)-stable RSP" to emphasize the parameters p, k, and η.

By definition, the stable RSP implies the RSP. Note that the (p, k, η)-stable RSP with $p = 1$ was proposed in [48] to analyze a reweighted ℓ_1-minimization algorithm for compressed sensing. With the notation of Proposition 1, we can see that checking (8) for all $\boldsymbol{d}$ in the range space of $\boldsymbol{A}$ is equivalent to checking it for all $\boldsymbol{d}$ in the nullspace of $\boldsymbol{D}$, the latter being the well-known *null space property* (NSP) [22, 26]. In other words, the (stable) RSP of $\boldsymbol{A}$ is equivalent to the (stable) NSP of $\boldsymbol{D}$.

In Sections 3.2 and 3.3, we will use the (p, k, η)-stable RSP as an assumption for analysis. Arguably, this is a weak assumption to make as the (stable) RSP is very close to a sufficient and necessary condition for exact recovery of the coefficient vector via ℓ_p-robust regression (1):

Proposition 2 (Exact Recovery $\Leftrightarrow$ RSP)**.** *Let $p \in (0, 1]$. For all $\boldsymbol{x}^*$ and $\boldsymbol{y}$ such that $\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{y}$ is k-sparse $\boldsymbol{x}^*$ is the unique solution to (1) if and only if $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ satisfies the (p, k)-RSP.*

While it is difficult to check the RSP computationally, we show that random Gaussian matrices of size $m \times n$ with sufficiently large m satisfy the stable RSP with high probability, which further justifies the usage of the stable RSP as an assumption in our analysis of $\texttt{IRLS}_p$:

Proposition 3 (Gaussian $\Rightarrow$ RSP)**.** *Suppose $m - n \geq 2k$. Let $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ be a matrix with i.i.d. $\mathcal{N}(0, 1)$ entries. Let $\delta \in (0, 1)$ and $\eta \in (0, 1]$ be fixed constants. If it holds that*

$$\frac{(m-n)^2}{m-n+1} \geq 2k \ln(em/k) \cdot \left(1.67 + \eta^{-1} + \frac{\sqrt{18 \ln(2.5\delta^{-1})}}{\sqrt{2k \ln(em/k)}}\right)^2, \tag{10}$$

then $\boldsymbol{A}$ satisfies the stable (p, k, η)-RSP for any $p \in (0, 1]$ with probability at least $1 - \delta$.

The assumption $m - n \geq 2k$ of Proposition 3 is necessary, as it is not hard to prove that, if $m - n < 2k$ and $\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{y}$ is k-sparse, then the ℓ_0-minimization problem $\min_{x \in \mathbb{R}^n} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_0$ has multiple solutions. With this assumption, (10) becomes $m \geq ck \log(em/k)$ for some constant c if m is large enough and n is fixed. Also, observe that (10) holds for large enough m, with k and n fixed.

Proposition 2 follows directly from [26, Theorem 4.9], so we omit its proof. The reader might also find Proposition 3 corresponds to [26, Corollary 9.34]. This corollary assumes $\boldsymbol{D}$ of (4) has i.i.d. $\mathcal{N}(0, 1)$ entries, from which is not immediate what the distribution of $\boldsymbol{A}$ is, thus it does not imply Proposition 3 directly; this is why we provide a complete proof for Proposition 3 in the appendix.

3.2 Global Linear Convergence of $\texttt{IRLS}_1$

With the background provided in Section 2, we are now ready to state the first main result:

Theorem 1 (Global Linear Convergence)**.** *Suppose $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ obeys the $(1, k, \eta)$-stable RSP with $\eta \in (0, 3/4)$ (cf. Definition 1). Let $\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{y}$ be k-sparse. If the smoothing parameter $\epsilon^{(t)}$ is updated as per (7) with $\alpha = k$ during the execution of the $\texttt{IRLS}_1$ Algorithm 1, then it holds for any $t \geq 1$ that*

$$\left\|\boldsymbol{A}\boldsymbol{x}^{(t+1)} - \boldsymbol{y}\right\|_1 - \left\|\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{y}\right\|_1 \leq \left(1 - \frac{(3-4\eta)^2}{294\eta m}\right)^t \cdot 3 \cdot \left\|\boldsymbol{A}\boldsymbol{x}^{(1)} - \boldsymbol{y}\right\|_1, \tag{11}$$

meaning that the $\texttt{IRLS}_1$ Algorithm 1 converges linearly and globally in objective value.

Discussion. The condition $\eta \in (0, 3/4)$ is better understood via Proposition 3, which asserts that this condition holds with high probability if $\boldsymbol{A}$ has i.i.d. Gaussian entries and if m is large enough

(compared to n, k). Moreover, if the $(1, k, \eta)$-RSP holds, Proposition 2 implies that $\boldsymbol{x}^*$ is a global minimizer of (1) with $p = 1$. The strength of Theorem 1 is in that it guarantees linear convergence for *any initialization, starting at the first iteration*. However, the price to pay is that the convergence rate seems conservative with constant $1/294$, and depends on the number m of samples: (11) predicts that $\texttt{IRLS}_1$ converges with accuracy δ in $O(m \log(1/\delta))$ iterations and m can be large. We argue that such dependency is not an artifact of our analysis, because similar dependencies have been established for IRLS variants for ℓ_p-regression (1) with $p > 2$ [2, Theorem 3.1], for interior point methods in convex optimization with m inequality constraints ([77, Chapter 11], [50, Chapter 5]), and for SGD in smooth & strongly convex optimization with a sum of m function components [78, Theorem 2.1]). Moreover, an adversarial initialization for which the rate empirically depends on m was pointed out for IRLS for basis pursuit in [25]. That being said, we conjecture that the linear dependency on m in (11) can be improved to $\sqrt{m} \log m$ for some different choice of $\epsilon^{(t)}$; see [2] and [77, Sections 11.5 and 11.8.2] for why this conjecture makes sense. Furthermore, Figure 1a empirically depicts that, with the least-squares initialization, i.e., with $w_i^{(0)} = 1$ for all i, $\texttt{IRLS}_1$ typically converges to $\boldsymbol{x}^*$ in 30 iterations. This hints to the possibility of alleviating the dependency on m by analyzing the least-squares initialization, a challenging task which we leave to future work.

Comparison to [15]. To our knowledge, [15] is the only paper that claimed the global linear convergence of an IRLS variant with $\epsilon^{(t)}$ updated as per (5) and with $p = 1$. In particular, [15] sets $\epsilon^{(t+1)} \leftarrow \beta\epsilon^{(t)}$ whenever $\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t)}\|_2 \leq 2\beta\epsilon^{(t)}$; see Figure 1a for the convergence of IRLS under this rule. The update rule $\epsilon^{(t+1)} \leftarrow \beta\epsilon^{(t)}$ of [15] has the advantage of being faster to compute than (7). In spite of this, and of their beautiful proof idea, their result has several disadvantages compared to Theorem 1. First, their proof involves sophisticated probabilistic arguments, and is tailored towards a feature matrix that satisfies strong concentration properties. Our analysis, on the other hand, uses the $(1, k, \eta)$-RSP, which is close to a necessary and sufficient condition for the success of ℓ_1-regression (cf. Proposition 2) and expected to hold for a much larger range of matrices than just (sub-)Gaussian matrices (see [79, 80] for related results in the context of basis pursuit). Second, some statements in their proof are inaccurate (e.g., their Lemma 5 does not hold for $k < n$, the last paragraph in their proving Lemma 8 is not rigorous). Third, their update rule incurs two hyper-parameters, $\epsilon^{(0)}$ and β, while Theorem 1 is based on the decreasing rule (7) that is adaptive to data, and the only hyper-parameter α is easier to set (Figure 1c).

Connection to [25]. A related global linear convergence was established for the IRLS method for basis pursuit in [25, Theorem 3.2]. Our proof is inspired by [25], but also improves on its strategy. For example, their Theorem 3.2 involves two parameters for the *stable nullspace property*, while we only have a single RSP parameter η, simplifying matters. Also, we obtain a constant of $1/294$ in Theorem 1, which is better than the value of $1/384$ for the respective constant in [25].[2]

While both [15] and [25] are limited to the $p = 1$ case, we present our Theorem 2 for $p \in (0, 1)$ next.

3.3 Local Superlinear Convergence of $\texttt{IRLS}_p$

The ℓ_p-regression problem (1) with $p \in (0, 1)$ is more challenging than the case $p = 1$ due to the lack of convexity. However, here we show that at least locally, $\texttt{IRLS}_p$ converges with a superlinear rate of order $2 - p$. Moreover, since it is valid to run $\texttt{IRLS}_p$ with $p = 0$, we carefully design the proof such that the following result holds for any $p \in [0, 1]$, not just for $p \in (0, 1)$:

Theorem 2 (Local Superlinear Convergence)**.** *Run $\texttt{IRLS}_p$ with $p \in [0, 1]$ and update $\epsilon^{(t)}$ by (7) with $\alpha = k$. Assume $\boldsymbol{A}$ satisfies the $(1, k, \eta)$-RSP (Definition 1). Let $c \in (0, 1)$ be a sufficiently small constant such that $2c^{1-p}\eta(\eta + 1) < (1 - c)^{2-p}$. Let $\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{y}$ be k-sparse with support S^*. Define*

$$\mu := 2\eta(\eta + 1)(1 - c)^{p-2} \cdot \min_{i \in S^*} |\boldsymbol{a}_i^\top \boldsymbol{x}^* - y_i|^{p-1}. \tag{12}$$

If the initialization $\boldsymbol{x}^{(1)}$ is in a neighborhood of $\boldsymbol{x}^*$, in the sense that

$$\left\|\boldsymbol{A}\boldsymbol{x}^{(1)} - \boldsymbol{A}\boldsymbol{x}^*\right\|_1 \leq c \cdot \min_{i \in S^*} |\boldsymbol{a}_i^\top \boldsymbol{x}^* - y_i|, \tag{13}$$

[2] In Theorem 3.2 of [25] the constant was $1/768$; we find that this is a miscalculation in their proof.

then $\texttt{IRLS}_p$ *achieves the following superlinear convergence rate of order* $2-p$ *for every* $t \geq 1$*:*

$$\left\|\boldsymbol{A}\boldsymbol{x}^{(t+1)} - \boldsymbol{A}\boldsymbol{x}^*\right\|_1 \leq \mu \cdot \left(\left\|\boldsymbol{A}\boldsymbol{x}^{(t)} - \boldsymbol{A}\boldsymbol{x}^*\right\|_1\right)^{2-p} < \left\|\boldsymbol{A}\boldsymbol{x}^{(t)} - \boldsymbol{A}\boldsymbol{x}^*\right\|_1. \tag{14}$$

In particular, for $p \in [0,1)$*, the superlinear rate* (14) *implies:*

$$\left\|\boldsymbol{A}\boldsymbol{x}^{(t+1)} - \boldsymbol{A}\boldsymbol{x}^*\right\|_1 \leq \left(\mu^{1/(1-p)} \cdot \left\|\boldsymbol{A}\boldsymbol{x}^{(1)} - \boldsymbol{A}\boldsymbol{x}^*\right\|_1\right)^{(2-p)^t - 1} \cdot \left\|\boldsymbol{A}\boldsymbol{x}^{(1)} - \boldsymbol{A}\boldsymbol{x}^*\right\|_1. \tag{15}$$

Example 1 (*Quadratic versus linear rates*)**.** We illustrate the practical benefit of a superlinear (quadratic) convergence rate with the following simple calculation. Suppose that the inequality in (13) is barely fulfilled such that in the case of $p = 0$, $\boldsymbol{x}^{(1)}$ satisfies $\mu \cdot \|\boldsymbol{A}\boldsymbol{x}^{(1)} - \boldsymbol{A}\boldsymbol{x}^*\|_1 \leq 0.9999999$. Then (15) implies that after 30 iterations, the residual error is far below numerical precision already since $\|\boldsymbol{A}\boldsymbol{x}^{(31)} - \boldsymbol{A}\boldsymbol{x}^*\|_1 \leq 0.9999999^{2^{30}-1} \cdot \|\boldsymbol{A}\boldsymbol{x}^{(1)} - \boldsymbol{A}\boldsymbol{x}^*\|_1 \approx 10^{-47} \cdot \|\boldsymbol{A}\boldsymbol{x}^{(1)} - \boldsymbol{A}\boldsymbol{x}^*\|_1$. On the other hand, for $p = 1$, a linear convergence factor μ of $\mu \leq 0.9999999$ is only able to guarantee a decay of the order $\mu^{30} \leq 0.9999999^{30} \approx 0.999997$ after 30 iterations.

Discussion. Even though $\texttt{IRLS}_p$ with $p \in (0,1)$ is tailored to ℓ_p-regression, we assume the $(1,k,\eta)$-RSP, which is stronger than the (p,k,η)-RSP. Also, we state the convergence in terms of the ℓ_1 norm, not ℓ_p. Here we prefer the assumption of the $(1,k,\eta)$-RSP for two reasons. First, the $(1,k,\eta)$-RSP is justified by Proposition 3. Second, using the $(1,k,\eta)$-RSP leads to simpler proofs, cleaner statements, and stronger results, and this is the key behind our improvement over [24] (see the next paragraph). Next, one might ask how to obtain such an initialization $\boldsymbol{x}^{(1)}$ that satisfies (13). One could run $\texttt{IRLS}_1$ to produce such an initialization, or alternatively, one could run $\texttt{IRLS}_p$ with a least-square initialization and count on empirical global (super)linear convergence of $\texttt{IRLS}_p$; the latter is what we did in the experiments and is what we recommend. A final and important remark is that ℓ_p-regression (1) with $p \in (0,1)$ is in general NP-hard [26, Exercise 2.10], but this does not contradict the theoretical local superlinear convergence of Theorem 2 and the empirical global convergence in Figures 1b and 2b, and this does not mean that $\texttt{IRLS}_p$ solves an NP-hard problem in polynomial time. The catch is that we operate under the assumption that $\boldsymbol{x}^*$ leads to a k-sparse residual and is a unique global minimizer of (1) (cf. Proposition 2). With this assumption and small enough k/m, ℓ_p-regression is tractable and can be solved via $\texttt{IRLS}_p$ to high accuracy (Figure 1b).

Connection to [24]. [24, Theorem 7.9] proves the local superlinear convergence of IRLS with update rule (6) for ℓ_p-basis pursuit (4), which motivates Theorem 2. However, [24, Theorem 7.9] requires c of (13) to satisfy $c = O(1/m^{2/p-1})$, which means that, as $p \to 0$ or $m \to \infty$, the eligible value of c quickly becomes vanishingly small, and thus the radius of local convergence (13) diminishes greatly. In contrast, in our condition $2c^{1-p}\eta(\eta+1) < (1-c)^{2-p}$, c has no direct dependency on m and is well-behaved even if $p \to 0$. Finally, [24, Theorem 7.9] does not hold for $p = 0$, for which Theorem 2 still holds and suggests a local quadratic rate. (Note that $\texttt{IRLS}_p$ with $p = 0$ can be interpreted rather as minimization of a smoothed sum-of-log objective than as a direct minimization of (1) with $p = 0$; we discuss this in more detail in the supplementary material.)

4 Applications

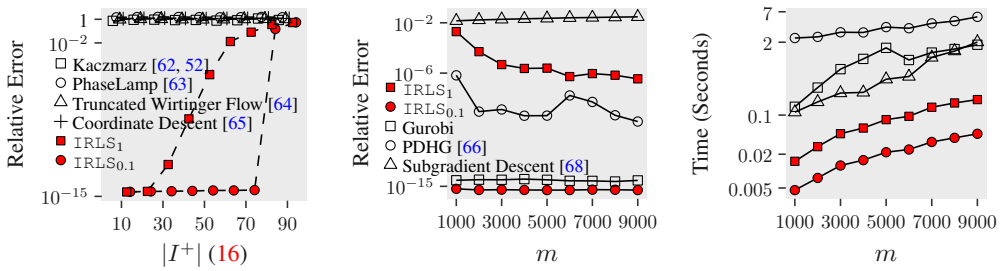
We now explore the performance of $\texttt{IRLS}_p$ in three applications, *real phase retrieval* (Section 4.1), *linear regression without correspondences* (Section 4.2), and *face restoration* (Section 4.3).

4.1 Real Phase Retrieval

In real phase retrieval [51, 52], we are given a measurement matrix $\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m]^\top$ and $\boldsymbol{y} = [y_1, \ldots, y_m]^\top$, where $y_i := |\boldsymbol{a}_i^\top \boldsymbol{x}^*|$, and we need to find either $\boldsymbol{x}^*$ or $-\boldsymbol{x}^*$. This problem is a relative of the complex phase retrieval problem [60], where all data $\boldsymbol{y}$ and $\boldsymbol{A}$, as well as the ground-truth $\boldsymbol{x}^*$, are complex-valued, and which has applications in X-ray crystallography [81].

Here we show that real phase retrieval can be solved via ℓ_p-regression (1). Consider the index sets

$$I^+ := \{i : \boldsymbol{a}_i^\top \boldsymbol{x}^* > 0\}, \qquad I^- := \{i : \boldsymbol{a}_i^\top \boldsymbol{x}^* < 0\}. \tag{16}$$



(a) $\frac{n}{m} = \frac{200}{399}$ (b) $n = 50$, $k/m = 0.5$ (c) $n = 50$, $k/m = 0.5$

Figure 2: Figure 2a: Relative error $\min\{\|\hat{\boldsymbol{x}} - \boldsymbol{x}^*\|_2, \|\hat{\boldsymbol{x}} + \boldsymbol{x}^*\|_2\}/\|\boldsymbol{x}^*\|_2$ of the methods that produce estimates $\hat{\boldsymbol{x}}$ for real phase retrieval. Figures 2b-2c: Relative errors $\|\hat{\boldsymbol{x}} - \boldsymbol{x}^*\|_2/\|\boldsymbol{x}^*\|_2$ and running times for linear regression without correspondences. $\texttt{IRLS}_p$ run for at most 50 iterations. 50 trials.

We might assume $\boldsymbol{a}_i^\top \boldsymbol{x}^* \neq 0$ for any i without loss of generality. Then we know that either $\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{y}$ has its ℓ_0 norm equal to $m - |I^-|$, or $\boldsymbol{A}(-\boldsymbol{x}^*) - \boldsymbol{y}$ has its ℓ_0 norm equal to $m - |I^+|$. With such sparsity patterns on these two residuals, we can minimize (1), which serves as a (non-convex) relaxation of ℓ_0-minimization, to recover $\boldsymbol{x}^*$ or $-\boldsymbol{x}^*$, whichever corresponds to a sparser residual. Here, we essentially treat one of the two clusters defined by (16) as inliers, and the other as outliers. This robust regression point of view on real phase retrieval seems to be known by experts [82, Section 1.2], but we have not found a paper that actually proposes to solve (1) for real phase retrieval.

Here we argue with Figure 2a that, solving (1) (by $\texttt{IRLS}_{0.1}$) for real phase retrieval can be beneficial. Indeed, theoretically, $\texttt{IRLS}_{0.1}$ converges to $\pm\boldsymbol{x}^*$ locally but superlinearly (as a corollary of Theorem 2). Empirically, Figure 2a shows that $\texttt{IRLS}_{0.1}$ succeeds in recovering $\pm\boldsymbol{x}^*$ by using only $m = 2n - 1$ samples, which is exactly the theoretical minimum for real phase retrieval [60, Proposition 2.5]. Figure 2a also shows that multiple other state-of-the-art methods[3] [52, 62, 63, 64, 65] fail to recover $\pm\boldsymbol{x}^*$ in this extreme situation, even though some of them have *nearly optimal* sample complexity, which requires roughly $O(n)$ samples up to a logarithmic factor to succeed (*caution*: nearly optimal $\neq$ optimal). As suggested by Proposition 3, for $|I^+|$ fixed, minimizing (1) with a Gaussian matrix $\boldsymbol{A}$ enjoys $m = O(n)$ sample complexity, up to also a logarithmic factor.

Despite these advantages, we also emphasize that solving (1) (via $\texttt{IRLS}_{0.1}$) for real phase retrieval has an inherent limit: The performance depends on the number $|I^+|$ of positive signs and the number m of samples; we might call $|I^+|/m$ the inlier rate or outlier rate. At $n/m = 200/399$, we see that $\texttt{IRLS}_{0.1}$ succeeds if $|I^+| \leq 70$, and if $|I^+| \geq 399 - 70$ by symmetry, but it fails if $|I^+|$ and $|I^-|$ get closer (Figure 2a). For $\texttt{IRLS}_{0.1}$ to successfully handle the case $|I^+| = |I^-|$, we need more samples, empirically, say $m \geq 5n$. It is an interesting future direction to design an algorithm that can handle the case of minimum samples ($m = 2n - 1$) and balanced data ($|I^+| = |I^-|$).

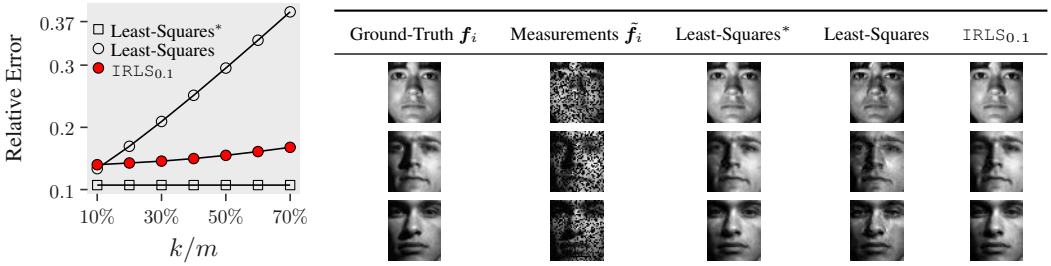
4.2 Linear Regression without Correspondences

Compared to (real) phase retrieval, the problem of *linear regression without correspondences* is an also important, but less developed subject; see [53, 84, 54, 85, 86, 55, 56, 87, 88, 89] for recent advances. In this problem we are given $\boldsymbol{y} \in \mathbb{R}^m$ and $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, and we need to solve the equations

$$\boldsymbol{y} = \boldsymbol{\Pi}\boldsymbol{A}\boldsymbol{x} \tag{17}$$

for some unknown permutation matrix $\boldsymbol{\Pi} \in \mathbb{R}^{m \times m}$ and vector $\boldsymbol{x} \in \mathbb{R}^n$, under the assumption that there exists a solution $(\boldsymbol{\Pi}^*, \boldsymbol{x}^*)$. Solving (17) is in general NP-hard for $n > 1$ [84, 90]. However, if we further assume that $\boldsymbol{\Pi}^*$ permutes at most k rows of $\boldsymbol{A}$, then we know that the residual $\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{y}$ is k-sparse. This is an insight of [54], where it was proposed to estimate $\boldsymbol{x}^*$ by solving (1) with $p = 1$.

[3] Reviewing these algorithms is beyond the scope of this paper. Some of them are designed for complex phase retrieval but is applicable to the real case. We used the implementations from PhasePack [83].



(a) Quantitative results

(b) Qualitative results (randomly selected images)

Figure 3: Recover a face image of the Extended Yale B dataset [69] that is corrupted by *salt & pepper* noise [58]. Figure 3a: relative error as a function of the amount of random salt & pepper noise or sparsity level, averaged over 20 trials, all faces, and all individuals. Figure 3b: qualitative results.

We show that $\texttt{IRLS}_p$ with $p \in (0,1)$ is more suitable than several baselines for solving this robust regression problem reduced from linear regression without correspondences. As baselines, we use the PDHG [66] and Gurobi [67] solvers to solve (1) with $p = 1$ as a linear program, and we also use a subgradient descent method implemented in the FOM toolbox of Beck & Guttmann-Beck [68]. Figure 2b shows that $\texttt{IRLS}_{0.1}$ is the most accurate, and Figure 2c shows that $\texttt{IRLS}_{0.1}$ is 30 times faster than Gurobi and subgradient descent, and is more than 100 times faster than PDHG. Interestingly, $\texttt{IRLS}_1$ is not very competitive in terms accuracy (Figure 2b). Moreover, $\texttt{IRLS}_1$ is slower than $\texttt{IRLS}_{0.1}$, as $\texttt{IRLS}_{0.1}$ converges faster and thus terminates earlier than $\texttt{IRLS}_1$; we terminate $\texttt{IRLS}_p$ whenever the objective stops decreasing (up to a tolerance 10^{-15}). Finally, see also [41, 44, 45, 46, 47] where their IRLS variants outperform a different set of baselines (for different problems).

4.3 Face Restoration from Sparsely Corrupted Measurements

Here we consider a simple face restoration experiment on the Extended Yale B dataset [69], downsampled as per [91]. This dataset contains the face images of 38 individuals under about $n \approx 60$ different illuminations. Let $\boldsymbol{F} = [\boldsymbol{f}_1, \ldots, \boldsymbol{f}_{n+1}]$ be the matrix of faces of the same individual, where each $\boldsymbol{f}_i \in \mathbb{R}^m$ is a (vectorized) face image with $m = 2,016$ pixels. We assume that $\boldsymbol{F}$ is approximately low-rank (which is true under the *Lambertian reflectance* assumption [92]), thus each $\boldsymbol{f}_i$ can be approximately represented as a linear combination of other faces of the individual, i.e., $\boldsymbol{f}_i \approx \boldsymbol{F}_i \boldsymbol{x}_i^*$ for some $\boldsymbol{x}_i^* \in \mathbb{R}^n$, where $\boldsymbol{F}_i \in \mathbb{R}^{m \times n}$ is the same as $\boldsymbol{F}$, except with the i-th column excluded.

In our experiments of Figure 3, we corrupt $\boldsymbol{f}_i$ by random salt & pepper noise [58] and obtain the noisy face $\tilde{\boldsymbol{f}}_i$ as our measurement. Thus, $\boldsymbol{F}_i \boldsymbol{x}_i^* - \tilde{\boldsymbol{f}}_i$ is approximately sparse. We consider three methods to estimate $\boldsymbol{x}_i^*$ and thus to find an estimate $\hat{\boldsymbol{f}}_i$ of $\boldsymbol{f}_i$: (1) Least-Squares that solves (1) with $\boldsymbol{A} = \boldsymbol{F}_i$, $\boldsymbol{y} = \tilde{\boldsymbol{f}}_i$, and $p = 2$, (2) Least-Squares* that solves (1) with $\boldsymbol{A} = \boldsymbol{F}_i$, $\boldsymbol{y} = \boldsymbol{f}_i$, and $p = 2$, used as a golden baseline, and (3) $\texttt{IRLS}_{0.1}$ Algorithm 1 which solves (1) with $\boldsymbol{A} = \boldsymbol{F}_i$ and $\boldsymbol{y} = \tilde{\boldsymbol{f}}_i$. We see from Figure 3a that $\texttt{IRLS}_{0.1}$ is more robust to salt & pepper noise, whose amount is k/m, than Least-Squares, and has lower relative error $\|\hat{\boldsymbol{f}}_i - \boldsymbol{f}_i\|_2 / \|\boldsymbol{f}_i\|_2$ (recall though that Least-Squares is statistically optimal for Gaussian noise). In Figure 3b we visualize some images randomly chosen from the dataset, and observe that Least-Squares tends to "average" the illumination in all images $\boldsymbol{F}_i$, while $\texttt{IRLS}_{0.1}$ delivers faithful restoration. Finally, see the appendix for more experiments.

Discussion & Future Work. While we established novel global linear & local superlinear rates of $\texttt{IRLS}_p$ for ℓ_p-regression (1) and explored several applications, many different directions deserve further investigation. Theoretically, the Gaussian assumption of Proposition 3 might be weakened [79, 80], and it is tempting to carry out a global rate analysis of $\texttt{IRLS}_p$ with $p \in (0,1)$ for ℓ_p-regression. Algorithmically, designing inexact solvers for the inner (weighted) least-squares problem (2) is expected to further accelerate our current IRLS implementation [93, 94]. Finally, our empirical experiments suggest that $\texttt{IRLS}_p$ (in particular $p \in (0,1)$) is worth being adopted and applied to many other outlier-robust estimation tasks beyond robust regression [41, 44, 46].

References

[1] A. El Alaoui, X. Cheng, A. Ramdas, M. J. Wainwright, and M. I. Jordan, "Asymptotic behavior of ℓ_p-based laplacian regularization in semi-supervised learning," in *Conference on Learning Theory*, pp. 879–906, 2016.

[2] D. Adil, R. Peng, and S. Sachdeva, "Fast, provably convergent irls algorithm for p-norm linear regression," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[3] D. Slepcev and M. Thorpe, "Analysis of p-laplacian regularization in semisupervised learning," *SIAM Journal on Mathematical Analysis*, vol. 51, no. 3, pp. 2085–2120, 2019.

[4] S. Bubeck, M. B. Cohen, Y. T. Lee, and Y. Li, "An homotopy method for lp regression provably beyond self-concordance and in input-sparsity time," in *ACM Symposium on Theory of Computing*, pp. 1130–1137, 2018.

[5] D. Adil, R. Kyng, R. Peng, and S. Sachdeva, "Iterative refinement for ℓ_p-norm regression," in *ACM-SIAM Symposium on Discrete Algorithms*, pp. 1405–1424, 2019.

[6] A. Jambulapati, Y. P. Liu, and A. Sidford, "Improved iteration complexities for overconstrained p-norm regression," tech. rep., arXiv:2111.01848v2 [cs.DS], 2021.

[7] J. Wright and Y. Ma, *High-Dimensional Data Analysis with Low-Dimensional Models: Principles, Computation, and Applications*. Cambridge University Press, 2020.

[8] P. J. Huber, "Robust estimation of a location parameter," *The Annals of Mathematical Statistics*, vol. 35, no. 1, pp. 73–101, 1964.

[9] E. Candes, M. Rudelson, T. Tao, and R. Vershynin, "Error correction via linear programming," in *IEEE Symposium on Foundations of Computer Science*, pp. 668–681, 2005.

[10] K. Bhatia, P. Jain, and P. Kar, "Robust regression via hard thresholding," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[11] K. Bhatia, P. Jain, P. Kamalaruban, and P. Kar, "Consistent robust regression," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[12] J. Yang, Y.-L. Chow, C. Ré, and M. W. Mahoney, "Weighted sgd for lp regression with randomized preconditioning," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 7811–7853, 2017.

[13] D. Durfee, K. A. Lai, and S. Sawlani, "l1 regression using lewis weights preconditioning and stochastic gradient descent," in *Conference On Learning Theory*, pp. 1626–1656, 2018.

[14] A. S. Suggala, K. Bhatia, P. Ravikumar, and P. Jain, "Adaptive hard thresholding for near-optimal consistent robust regression," in *Conference on Learning Theory*, pp. 2892–2897, 2019.

[15] B. Mukhoty, G. Gopakumar, P. Jain, and P. Kar, "Globally-convergent iteratively reweighted least squares for robust regression problems," in *International Conference on Artificial Intelligence and Statistics*, pp. 313–322, 2019.

[16] S. Pesme and N. Flammarion, "Online robust regression via sgd on the l1 loss," *Advances in Neural Information Processing Systems*, vol. 33, pp. 2540–2552, 2020.

[17] X. Chen and M. Derezinski, "Query complexity of least absolute deviation regression via robust uniform convergence," in *Conference on Learning Theory*, pp. 1144–1179, 2021.

[18] A. Parulekar, A. Parulekar, and E. Price, "L1 regression with lewis weights subsampling," tech. rep., arXiv:2105.09433 [cs.LG], 2021.

[19] E. J. Candes and T. Tao, "Decoding by linear programming," *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[20] E. J. Candes, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006.

[21] D. L. Donoho, "For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution," *Communications on Pure and Applied Mathematics*, vol. 59, no. 6, pp. 797–829, 2006.

[22] A. Cohen, W. Dahmen, and R. DeVore, "Compressed sensing and best k-term approximation," *Journal of the American mathematical society*, vol. 22, no. 1, pp. 211–231, 2009.

[23] T. Blumensath and M. E. Davies, "Iterative hard thresholding for compressed sensing," *Applied and Computational Harmonic Analysis*, vol. 27, no. 3, pp. 265–274, 2009.

[24] I. Daubechies, R. DeVore, M. Fornasier, and C. S. Güntürk, "Iteratively reweighted least squares minimization for sparse recovery," *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, vol. 63, no. 1, pp. 1–38, 2010.

[25] C. Kümmerle, C. Mayrink Verdun, and D. Stöger, "Iteratively reweighted least squares for basis pursuit with global linear convergence rate," *Advances in Neural Information Processing Systems*, 2021.

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

[27] R. Chartrand, "Nonconvex compressed sensing and error correction," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 3, pp. III–889, 2007.

[28] R. Chartrand, "Exact reconstruction of sparse signals via nonconvex minimization," *IEEE Signal Processing Letters*, vol. 14, no. 10, pp. 707–710, 2007.

[29] R. Chartrand and V. Staneva, "Restricted isometry properties and nonconvex compressive sensing," *Inverse Problems*, vol. 24, no. 3, p. 035020, 2008.

[30] S. Foucart and M.-J. Lai, "Sparsest solutions of underdetermined linear systems via ell_q-minimization for $0 < q \leq 1$," *Applied and Computational Harmonic Analysis*, vol. 26, no. 3, pp. 395–407, 2009.

[31] M. Wang, W. Xu, and A. Tang, "On the performance of sparse recovery via ℓ_p-minimization ($0 \leq p \leq 1$)," *IEEE Transactions on Information Theory*, vol. 57, no. 11, pp. 7255–7278, 2011.

[32] Q. Sun, "Recovery of sparsest signals via ℓ_q-minimization," *Applied and Computational Harmonic Analysis*, vol. 32, no. 3, pp. 329–341, 2012.

[33] D. Ba, B. Babadi, P. L. Purdon, and E. N. Brown, "Convergence and stability of iteratively re-weighted least squares algorithms," *IEEE Transactions on Signal Processing*, vol. 62, no. 1, pp. 183–195, 2013.

[34] L. Zheng, A. Maleki, H. Weng, X. Wang, and T. Long, "Does ℓ_p-minimization outperform ℓ_1-minimization?," *IEEE Transactions on Information Theory*, vol. 63, no. 11, pp. 6896–6935, 2017.

[35] G. Marjanovic and V. Solo, "On ℓ_q optimization and matrix completion," *IEEE Transactions on Signal Processing*, vol. 60, no. 11, pp. 5714–5724, 2012.

[36] F. Nie, H. Huang, and C. Ding, "Low-rank matrix recovery via efficient schatten p-norm minimization," in *AAAI Conference on Artificial Intelligence*, 2012.

[37] C. Kümmerle and J. Sigl, "Harmonic mean iteratively reweighted least squares for low-rank matrix recovery," *The Journal of Machine Learning Research*, vol. 19, no. 1, pp. 1815–1863, 2018.

[38] P. Giampouras, R. Vidal, A. Rontogiannis, and B. Haeffele, "A novel variational form of the schatten-p quasi-norm," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21453–21463, 2020.

[39] C. Kümmerle and C. M. Verdun, "A scalable second order method for ill-conditioned matrix completion from few samples," in *International Conference on Machine Learning*, pp. 5872–5883, 2021.

[40] G. Lerman, M. B. McCoy, J. A. Tropp, and T. Zhang, "Robust computation of linear models by convex relaxation," *Foundations of Computational Mathematics*, vol. 15, no. 2, pp. 363–410, 2015.

[41] M. C. Tsakiris and R. Vidal, "Dual principal component pursuit," *Journal of Machine Learning Research*, vol. 19, no. 18, pp. 1–50, 2018.

[42] Z. Zhu, Y. Wang, D. Robinson, D. Naiman, R. Vidal, and M. C. Tsakiris, "Dual principal component pursuit: Improved analysis and efficient algorithms," in *Advances in Neural Information Processing Systems*, 2018.

[43] T. Ding, Z. Zhu, T. Ding, Y. Yang, R. Vidal, M. C. Tsakiris, and D. Robinson, "Noisy dual principal component pursuit," in *International Conference on Machine Learning*, pp. 1617–1625, 2019.

[44] K. Aftab and R. Hartley, "Convergence of iteratively re-weighted least squares to robust m-estimators," in *IEEE Winter Conference on Applications of Computer Vision*, pp. 480–487, 2015.

[45] W. Dong, X.-j. Wu, and J. Kittler, "Sparse subspace clustering via smoothed ℓ_p minimization," *Pattern Recognition Letters*, vol. 125, pp. 206–211, 2019.

[46] D. Iwata, M. Waechter, W.-Y. Lin, and Y. Matsushita, "An analysis of sketched irls for accelerated sparse residual regression," in *European Conference on Computer Vision*, pp. 609–626, 2020.

[47] T. Ding, Y. Yang, Z. Zhu, D. P. Robinson, R. Vidal, L. Kneip, and M. C. Tsakiris, "Robust homography estimation via dual principal component pursuit," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6080–6089, 2020.

[48] Y.-B. Zhao and D. Li, "Reweighted ℓ_1-minimization for sparse solutions to underdetermined linear systems," *SIAM Journal on Optimization*, vol. 22, no. 3, pp. 1065–1088, 2012.

[49] A. Beck, *First-Order Methods in Optimization*. Society for Industrial and Applied Mathematics, 2017.

[50] Y. Nesterov, *Lectures on Convex Optimization*. Springer, 2018.

[51] Y. Chen, Y. Chi, J. Fan, and C. Ma, "Gradient descent with random initialization: fast global convergence for nonconvex phase retrieval," *Mathematical Programming*, vol. 176, no. 1, pp. 5–37, 2019.

[52] Y. S. Tan and R. Vershynin, "Phase retrieval via randomized kaczmarz: theoretical guarantees," *Information and Inference: A Journal of the IMA*, vol. 8, no. 1, pp. 97–123, 2019.

[53] J. Unnikrishnan, S. Haghighatshoar, and M. Vetterli, "Unlabeled sensing with random linear measurements," *IEEE Transactions on Information Theory*, vol. 64, no. 5, pp. 3237–3253, 2018.

[54] M. Slawski and E. Ben-David, "Linear regression with sparsely permuted data," *Electronic Journal of Statistics*, vol. 13, no. 1, pp. 1–36, 2019.

[55] M. C. Tsakiris, L. Peng, A. Conca, L. Kneip, Y. Shi, and H. Choi, "An algebraic-geometric approach for linear regression without correspondences," *IEEE Transactions on Information Theory*, vol. 66, no. 8, pp. 5130–5144, 2020.

[56] L. Peng and M. C. Tsakiris, "Linear regression without correspondences via concave minimization," *IEEE Signal Processing Letters*, vol. 27, pp. 1580–1584, 2020.

[57] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 2, pp. 210–227, 2008.

[58] C. Solomon and T. Breckon, *Fundamentals of Digital Image Processing: A practical approach with examples in Matlab*. John Wiley & Sons, 2011.

[59] Y. Yao, L. Peng, and M. Tsakiris, "Unlabeled principal component analysis," *Advances in Neural Information Processing Systems*, 2021.

[60] R. Balan, P. Casazza, and D. Edidin, "On signal reconstruction without phase," *Applied and Computational Harmonic Analysis*, vol. 20, no. 3, pp. 345 – 356, 2006.

[61] T. Strohmer and R. Vershynin, "A randomized kaczmarz algorithm with exponential convergence," *Journal of Fourier Analysis and Applications*, vol. 15, no. 2, pp. 262–278, 2009.

[62] K. Wei, "Solving systems of phaseless equations via kaczmarz methods: A proof of concept study," *Inverse Problems*, vol. 31, no. 12, p. 125008, 2015.

[63] O. Dhifallah, C. Thrampoulidis, and Y. M. Lu, "Phase retrieval via linear programming: Fundamental limits and algorithmic improvements," in *Allerton Conference on Communication, Control, and Computing*, pp. 1071–1077, 2017.

[64] Y. Chen and E. Candes, "Solving random quadratic systems of equations is nearly as easy as solving linear systems," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[65] W.-J. Zeng and H.-C. So, "Coordinate descent algorithms for phase retrieval," tech. rep., arXiv:1706.03474 [cs.IT], 2017.

[66] D. Applegate, M. Díaz, O. Hinder, H. Lu, M. Lubin, B. O'Donoghue, and W. Schudy, "Practical large-scale linear programming using primal-dual hybrid gradient," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[67] "Gurobi 9.5.1." https://www.gurobi.com/, Date Accessed: May 10, 2022.

[68] A. Beck and N. Guttmann-Beck, "Fom – a matlab toolbox of first-order methods for solving convex optimization problems," *Optimization Methods and Software*, vol. 34, no. 1, pp. 172–193, 2019.

[69] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 643–660, 2001.

[70] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM review*, vol. 43, no. 1, pp. 129–159, 2001.

[71] G. Lerman and T. Maunu, "Fast, robust and non-convex subspace recovery," *Information and Inference: A Journal of the IMA*, vol. 7, no. 2, pp. 277–336, 2018.

[72] Q. Qu, Z. Zhu, X. Li, M. C. Tsakiris, J. Wright, and R. Vidal, "Finding the sparsest vectors in a subspace: Theory, algorithms, and applications," tech. rep., arXiv:2001.06970 [cs.LG], 2020.

[73] T. F. Chan and P. Mulet, "On the convergence of the lagged diffusivity fixed point method in total variation image restoration," *SIAM Journal on Numerical Analysis*, vol. 36, no. 2, pp. 354–367, 1999.

[74] A. Beck, "On the convergence of alternating minimization for convex programming with applications to iteratively reweighted least squares and decomposition schemes," *SIAM Journal on Optimization*, vol. 25, no. 1, pp. 185–209, 2015.

[75] A. Y. Aravkin, J. V. Burke, and D. He, "IRLS for sparse recovery revisited: Examples of failure and a remedy," tech. rep., arXiv:1910.07095 [math.ST], 2019.

[76] X. Yang, J. Wang, and H. Wang, "Towards an efficient approach for the nonconvex ℓ_p-ball projection: algorithm and analysis," tech. rep., arXiv:2101.01350v5 [math.OC], 2021.

[77] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[78] Z. Allen-Zhu, "Katyusha: The first direct acceleration of stochastic gradient methods," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 8194–8244, 2017.

[79] G. Lecué and S. Mendelson, "Sparse recovery under weak moment assumptions," *Journal of the European Mathematical Society*, vol. 19, no. 3, pp. 881–904, 2017.

[80] S. Dirksen, G. Lecué, and H. Rauhut, "On the gap between restricted isometry properties and sparse recovery conditions," *IEEE Transactions on Information Theory*, vol. 64, no. 8, pp. 5478–5487, 2016.

[81] P. Grohs, S. Koppensteiner, and M. Rathmair, "Phase retrieval: Uniqueness and stability," *SIAM Review*, vol. 62, no. 2, pp. 301–350, 2020.

[82] M. Huang and Y. Wang, "Linear convergence of randomized kaczmarz method for solving complex-valued phaseless equations," tech. rep., arXiv:2109.11811 [math.NA], 2021.

[83] R. Chandra, T. Goldstein, and C. Studer, "Phasepack: A phase retrieval library," in *International conference on Sampling Theory and Applications*, pp. 1–5, 2019.

[84] A. Pananjady, M. J. Wainwright, and T. A. Courtade, "Linear regression with shuffled data: Statistical and computational limits of permutation recovery," *IEEE Transactions on Information Theory*, vol. 64, no. 5, pp. 3286–3300, 2018.

[85] M. C. Tsakiris and L. Peng, "Homomorphic sensing," in *International Conference on Machine Learning*, 2019.

[86] L. Peng, B. Wang, and M. Tsakiris, "Homomorphic sensing: Sparsity and noise," in *International Conference on Machine Learning*, pp. 8464–8475, 2021.

[87] M. Slawski, G. Diao, and E. Ben-David, "A pseudo-likelihood approach to linear regression with partially shuffled data," *Journal of Computational and Graphical Statistics*, vol. 0, no. 0, pp. 1–31, 2021.

[88] Y. Xie, Y. Mao, S. Zuo, H. Xu, X. Ye, T. Zhao, and H. Zha, "A hypergradient approach to robust regression without correspondence," in *International Conference on Learning Representations*, 2021.

[89] F. Li, K. Fujiwara, F. Okura, and Y. Matsushita, "Generalized shuffled linear regression," in *IEEE/CVF International Conference on Computer Vision*, pp. 6474–6483, 2021.

[90] D. Hsu, K. Shi, and X. Sun, "Linear regression without correspondence," in *Advances in Neural Information Processing Systems*, 2017.

[91] C. You, D. Robinson, and R. Vidal, "Scalable sparse subspace clustering by orthogonal matching pursuit," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3918–3927, 2016.

[92] R. Basri and D. W. Jacobs, "Lambertian reflectance and linear subspaces," *IEEE transactions on pattern analysis and machine intelligence*, vol. 25, no. 2, pp. 218–233, 2003.

[93] M. Fornasier, S. Peter, H. Rauhut, and S. Worm, "Conjugate gradient acceleration of iteratively re-weighted least squares methods," *Computational Optimization and Applications*, vol. 65, no. 1, pp. 205–259, 2016.

[94] S. Gazzola, C. Meng, and J. G. Nagy, "Krylov methods for low-rank regularization," *SIAM Journal on Matrix Analysis and Applications*, vol. 41, no. 4, pp. 1477–1504, 2020.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes]
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...

(a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] Furthermore, we made a lot of efforts in comparing the theoretical results of prior works and position our contribution in the literature.

(b) Did you describe the limitations of your work? [Yes] See for example at Lines 268-274 on real phase retrieval. We also discussed several directions for improvements or future research at the end of the paper.

(c) Did you discuss any potential negative societal impacts of your work? [No]

(d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...

(a) Did you state the full set of assumptions of all theoretical results? [Yes] In particular, we used an entire subsection (Section 3.1) to justify the major assumption, the stable range space property.

(b) Did you include complete proofs of all theoretical results? [Yes] We omitted the proofs of Proposition 1 and 2 as they follow directly from existing results; see Line 83 and Line 155. All other theoretical results are proved in the appendix.

3. If you ran experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We attached a copy of our code in the supplemental material.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] In the appendix, we use an entire section to describe our experimental stup. We did not do that in the main paper in interest of space.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] We did not report error bars for two reasons. The main reason is that some of our figures are quite dense, and adding error bars on them makes it harder to read. A slightly subjective reason is that we believe our current figures are informative enough, correctly deliver the message, and are fair for other methods.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes]

(b) Did you mention the license of the assets? [No] In the main paper we used the Extended Yale B face dataset, which is free for research purposes.

(c) Did you include any new assets either in the supplemental material or as a URL? [No]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]