Iteratively Reweighted Least Squares for Basis Pursuit with Global Linear Convergence Rate

Anonymous Author(s)
Affiliation
Address
email

Abstract

The recovery of sparse data is at the core of many applications in machine learning and signal processing. While such problems can be tackled using ℓ_1-regularization as in the LASSO estimator and in the Basis Pursuit approach, specialized algorithms are typically required to solve the corresponding high-dimensional non-smooth optimization for large instances. Iteratively Reweighted Least Squares (IRLS) is a widely used algorithm for this purpose due its excellent numerical performance. However, while existing theory is able to guarantee convergence of this algorithm to the minimizer, it does not provide a global convergence rate. In this paper, we prove that a variant of IRLS converges *with a global linear rate* to a sparse solution, i.e., with a linear error decrease occurring immediately from any initialization, if the measurements fulfill the usual null space property assumption. We support our theory by numerical experiments showing that our linear rate captures the correct dimension dependence. We anticipate that our theoretical findings will lead to new insights for many other use cases of the IRLS algorithm, such as in low-rank matrix recovery.

1 Introduction

The field of sparse recovery deals with the problem of recovering an (approximately) sparse vector x from a few linear measurements, presented by an underdetermined system of linear equations of the form $y = Ax$. One approach to solve this problem is to consider the ℓ_0-minimization under linear constraints, which is NP-hard in general [23, 50]. For computational reasons, instead of ℓ_0-minimization, it is common practice to consider its convex relaxation

$$\min_{x \in \mathbb{R}^N} ||x||_1 \qquad \text{subject to } Ax = y, \tag{P_1}$$

where $A \in \mathbb{R}^{m \times N}$, $b \in \mathbb{R}^m$ are given, which in the literature is referred to as ℓ_1*(-norm) minimization* [18, 27, 58] or *basis pursuit* [16, 17]. Unlike ℓ_0-minimization, the optimization program (P_1) is computationally tractable in general, and a close relationship of their minimizers has been recognized and well-studied in the theory of compressive sensing [11, 12, 25, 31]. In statistics and machine learning, an unconstrained variant of (P_1), often called *LASSO*, amounts to the most well-studied tractable estimator for variable selection in high-dimensional inference [36, 45, 59]. ℓ_1-minimization has many other applications and it was even called the *modern least squares* [13].

The tractability of (P_1) becomes evident from the fact that it can be reformulated as a linear program [60]. However, as many problems of interest in applications are high-dimensional and therefore challenging for standard linear programming methods, many specialized solvers for (P_1) have been proposed, such as the Homotopy Method [26], forward-backward algorithms [20], Alternating Direction Method of Multipliers [8], Bregman iterative regularization [64] and Semismooth Newton

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

Augmented Lagragian Methods [44] and Iteratively Reweighted Least Squares (IRLS), the latter of which is in the focus of this paper.

Iteratively Reweighted Least Squares corresponds to a family of algorithms that minimizes non-smooth objective functions by solving a sequence of quadratic problems, with its idea going back to a method proposed by Weiszfeld for the *Fermat-Weber problem* [6, 62]. A variety of different problems such as robust regression in statistics [37, 48], total variation regularization in image processing [2, 32, 51], joint learning of neural networks [65], robust subspace recovery [43] and the recovery of low-rank matrices [29, 40, 41, 47] can be solved efficiently by IRLS in practice, as it relies on simple linear algebra to solve the linear systems arising from the quadratic problems at each iteration, without the need of a careful initialization or intricate parameter tuning. On the other hand, the analysis of IRLS methods is typically challenging: General convergence results are often weak, and stronger convergence results are only available in particular cases; see Section 2.3 for more details.

IRLS for sparse recovery. In the sparse recovery context, the first variants of IRLS were introduced in [33, 53] for the ℓ_p-quasinorm minimization problem (P_p) with $0 < p \leq 1$ that is similar to (P_1), but with $\|x\|_p$ instead of $\|x\|_1$ as an objective. In [14], modifications of the method of [33, 53] using specific smoothing parameter update rules were observed to exhibit excellent numerical performance for solving (P_p), retrieving the underlying sparse vector when most of the methods fail. A useful fact is that IRLS is one of the few methods (ADMM being the other one [8]) that provides a framework to solve both constrained and unconstrained formulations of ℓ_p-minimization problems.

A major step forward in the theoretical understanding of IRLS was achieved in [22], where the authors showed that a variant of IRLS for (P_1) converges globally to an ℓ_1-minimizer if the measurement operator A fulfills the null space property of sufficient order, which essentially ensures that an ℓ_1-minimizer is actually sparse. However, since this proof relies on a compactness argument, their proof does not reveal any rate for *global* convergence. The analysis of [22] provides, furthermore, a *locally* linear convergence rate, but this local linear rate has the drawback that it only applies *if the support of the true signal has been discovered*, which is arguably the difficult part of ℓ_0-minimization—cf. Proposition 3.1 below and Section 4.1.

While several extensions and modifications of the IRLS algorithm in [22] have been proposed (see, e.g., [3, 30]), this following fundamental algorithmic question has remained unanswered:

What is the global convergence rate of the IRLS algorithm for ℓ_1-minimization?

Our contribution. We resolve this question, formally stated in [56], and present a new IRLS algorithm that *converges linearly* to a sparse ground truth, *starting from any initialization*, as stated in Theorem 3.2. Our algorithm returns a feasible solution with δ-accuracy, i.e., $\|x_* - x^k\|_1 \leq \delta$, where x_* is the underlying s-sparse vector, in $k = O(N\sqrt{(\log N)/m}\log(1/\delta))$ iterations. Analogous to [22], it is assumed that the measurement matrix A satisfies the so-called null space property [19]. We also provide a similar result for approximately sparse vectors. Our proof relies on a novel quantification of the descent of a carefully chosen objective function in the direction of the ground truth. Additionally, we support the theoretical claims by numerical simulations indicating that we capture the correct dimension dependence. We believe that the new analysis techniques in this paper are of independent interest and will pave the way for establishing global convergence rates for other variants of IRLS such as in low-rank matrix recovery [29].

Notation. We denote the cardinality of a set I by $|I|$ and the support of a vector $x \in \mathbb{R}^N$, i.e., the index set of its nonzero entries, by $\operatorname{supp}(x) = \{j \in [N] : x_j \neq 0\}$. We call a vector s-sparse if at most s of its entries are nonzero. We denote by x_I the restriction of x onto the coordinates indexed by I, and use the notation $I^c := \mathbb{R}^N \setminus I$ to denote the complement of a set I. Furthermore, $\sigma_s(x)_{\ell_1}$ denotes the *ℓ_1-error of the best s-term approximation* of a vector $x \in \mathbb{R}^N$, i.e., $\sigma_s(x)_{\ell_1} = \inf\{\|x - z\|_1 : z \in \mathbb{R}^N \text{ is } s\text{-sparse}\}$.

2 IRLS for sparse recovery

We now present a simple derivation of the Iteratively Reweighted Least Squares (IRLS) algorithm for ℓ_1-minimization which is studied in this paper. IRLS algorithms can be interpreted as a variant of a Majorize-Minimize (MM) algorithm [57], as we will lay out in the following. It mitigates the non-smoothness of the $\|\cdot\|_1$-norm by using the smoothed objective function $\mathcal{J}_\varepsilon : \mathbb{R}^N \to \mathbb{R}$, which is

defined, for a given $\varepsilon > 0$, by

$$\mathcal{J}_\varepsilon(x) := \sum_{i=1}^{N} j_\varepsilon(x_i) \quad \text{with} \quad j_\varepsilon(x) := \begin{cases} |x|, & \text{if } |x| > \varepsilon, \\ \frac{1}{2}\left(\frac{x^2}{\varepsilon} + \varepsilon\right), & \text{if } |x| \leq \varepsilon. \end{cases} \tag{1}$$

The function $\mathcal{J}_\varepsilon$ can be considered as a scaled Huber loss function which is widely used in robust regression analysis [38, 46]. Moreover, the function $\mathcal{J}_\varepsilon$ is continuously differentiable and fulfills $|x| \leq j_\varepsilon(x) \leq |x| + \varepsilon$ for each $x \in \mathbb{R}$. Instead of minimizing the function $\mathcal{J}_\varepsilon$ directly, the idea of IRLS is to minimize instead a suitable chosen quadratic function $Q_\varepsilon(\cdot, x)$, which majorizes $\mathcal{J}_\varepsilon$ such that $Q_\varepsilon(z, x) \geq \mathcal{J}_\varepsilon(z)$ for all $z \in \mathbb{R}^N$. This function is furthermore chosen such that $Q_\varepsilon(x, x) = \mathcal{J}_\varepsilon(x)$ holds, which implies that $\min_{z \in \mathbb{R}^n} Q_\varepsilon(z, x) \leq \mathcal{J}_\varepsilon(x)$. The latter inequality implies that by minimizing $Q_\varepsilon(\cdot, x)$, IRLS actually achieves an improvement in the value of $\mathcal{J}_\varepsilon$ as well. More specifically, $Q_\varepsilon(\cdot, x)$ is defined by

$$\begin{aligned} Q_\varepsilon(z, x) &:= \mathcal{J}_\varepsilon(x) + \langle \nabla \mathcal{J}_\varepsilon(x), z - x \rangle + \frac{1}{2}\langle (z - x), \operatorname{diag}(w_\varepsilon(x))(z - x) \rangle \\ &= \mathcal{J}_\varepsilon(x) + \frac{1}{2}\langle z, \operatorname{diag}(w_\varepsilon(x)) z \rangle - \frac{1}{2}\langle x, \operatorname{diag}(w_\varepsilon(x)) x \rangle, \end{aligned} \tag{2}$$

where $\nabla \mathcal{J}_\varepsilon(x) = \left(\begin{cases} \frac{x_i}{|x_i|}, & \text{if } |x_i| > \varepsilon \\ \frac{x_i}{\varepsilon}, & \text{if } |x_i| \leq \varepsilon \end{cases} \right)_{i=1}^{N}$ is the gradient of $\mathcal{J}_\varepsilon$ at x and the weight vector $w_\varepsilon(x) \in \mathbb{R}^N$ is a vector of *weights* such that $w_\varepsilon(x)_i := [\max(|x_i|, \varepsilon)]^{-1}$ for $i \in [N]$. The following lemma shows that $Q_\varepsilon(\cdot, \cdot)$ has indeed the above-mentioned properties. We refer to the supplementary material for a proof.

Lemma 2.1. *Let $\varepsilon > 0$, let $\mathcal{J}_\varepsilon : \mathbb{R}^N \to \mathbb{R}$ be defined as in* (1) *and $Q_\varepsilon : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ as defined in* (2)*. Then, for any $z, x \in \mathbb{R}^N$, the following affirmations hold:*

i. $\operatorname{diag}(w_\varepsilon(x))x = \nabla \mathcal{J}_\varepsilon(x)$, *ii.* $Q_\varepsilon(x, x) = \mathcal{J}_\varepsilon(x)$, *iii.* $Q_\varepsilon(z, x) \geq \mathcal{J}_\varepsilon(z)$.

As can be seen from the equality in (2), minimizing of $Q_\varepsilon(\cdot, x)$ corresponds to a minimizing *(re-)weighted least squares objective* $\langle \cdot, \operatorname{diag}(w_\varepsilon(x)) \cdot \rangle$, which lends its name to the method. Note that unlike a classical MM approach, however, IRLS comes with an *update* step *of the smoothing parameter* ε at each iteration. We provide an outline of the method in Algorithm 1.

Algorithm 1 Iteratively Reweighted Least Squares for ℓ_1-minimization

Input: Measurement matrix $A \in \mathbb{R}^{m \times N}$, data vector $y \in \mathbb{R}^m$,
initial weight vector $w_0 \in \mathbb{R}^N$ (default: $w_0 = (1, 1, \ldots, 1)$).
Set $\varepsilon_0 = \infty$.
for $k = 0, 1, 2, \ldots$ **do**

$$x^{k+1} := \operatorname*{arg\,min}_{z \in \mathbb{R}^N} \langle z, \operatorname{diag}(w_k) z \rangle \ \text{ subject to } \ Az = y, \tag{3}$$

$$\varepsilon_{k+1} := \min\left(\varepsilon_k, \frac{\sigma_s(x^{k+1})_{\ell_1}}{N}\right), \tag{4}$$

$$(w_{k+1})_i := \frac{1}{\max\left(|x_i^{k+1}|, \varepsilon_{k+1}\right)} \qquad \text{for each } i \in [N], \tag{5}$$

end for
return Sequence $(x^k)_{k \geq 1}$.

The weighted least squares update (3) can be computed such that $x^{k+1} = W_k^{-1} A^* (A W_k^{-1} A^*)^{-1}(y)$ with $W_k = \operatorname{diag}(w_k)$, with the solution of the $(m \times m)$ linear system $(A W_k^{-1} A^*) z = y$ as a main computational step. This linear system is positive definite and suitable for the use of iterative solvers. In [30], an analysis of how accurately the linear system of a similar IRLS method needs to be solved to ensure overall convergence. We note that for small ε_k, the Sherman-Woodbury formula [63] can be used so that the calculation of x^{k+1} boils down to solving a smaller linear system that is well-conditioned, c.f. the supplementary material for details. This numerically advantageous property is not shared by the methods of [3, 22, 30], as our smoothing update (4) is slightly different from the ones proposed in these papers. We refer to Section 2.2 for a discussion.

A consequence of Lemma 2.1, step (4) and the fact that $\varepsilon \to \mathcal{J}_\varepsilon(z)$ is monotonously non-increasing is that the iterates x^k, x^{k+1} of Algorithm 1 fulfill

$$\mathcal{J}_{\varepsilon_{k+1}}(x^{k+1}) \leq \mathcal{J}_{\varepsilon_k}(x^{k+1}) \leq Q_{\varepsilon_k}(x^{k+1}, x^k) \leq Q_{\varepsilon_k}(x^k, x^k) = \mathcal{J}_{\varepsilon_k}(x^k). \tag{6}$$

This shows in particular that the sequence $\left\{\mathcal{J}_{\varepsilon_k}\left(x^k\right)\right\}_{k=0}^{\infty}$ is non-increasing. For this reason, it can be shown that each accumulation point of the sequence of iterates $(x^k)_{k\geq 0}$ is a (first-order) stationary point of the smoothed ℓ_1-objective $J_{\overline{\varepsilon}}(\cdot)$ subject to the measurement constraint imposed by A and y, where $\overline{\varepsilon} = \lim_{k\to\infty} \varepsilon_k$ (see [22, Theorem 5.3]).

2.1 Null space property

As in [22], the analysis we present is based on the assumption that the measurement matrix A satisfies the so-called null space property [19, 35], which is a key concept in the compressed sensing literature (see, e.g., [31, Chapter 4] for an overview).

Definition 2.2. *A matrix $A \in \mathbb{R}^{m\times N}$ is said to satisfy the ℓ_1-*null space property *(ℓ_1-NSP) of order $s \in \mathbb{N}$ with constant $0 < \rho_s < 1$ if for any set $S \subset [N]$ of cardinality $|S| \leq s$, it holds that $\|v_S\|_1 \leq \rho_s \|v_{S^c}\|_1$, for all $v \in \ker(A)\backslash\{0\}$.*

In [31, Chapter 4], the property of Definition 2.2 was called *stable* null space property. The importance of the null space property is due to the fact that it gives a necessary and sufficient criterion for the success of basis pursuit for sparse recovery, as the following theorem shows.

Theorem 2.3 ([31, Theorem 4.5])**.** *Given a matrix $A \in \mathbb{R}^{m\times N}$, every vector $x \in \mathbb{R}^N$ such that $||x||_0 \leq s$ is the unique solution of (P_1) with $Ax = y$ if and only if A satisfies the null space property of order s for some $0 < \rho_s < 1$.*

The ℓ_1-NSP is implied by the restricted isometry property (see, e.g., [10]), which is fulfilled by a large class of random matrices with high probability. For example, this includes matrices with (sub-)gaussian entries and random partial Fourier matrices [5, 54].

2.2 Existing theory

A predecessor of IRLS for the sparse recovery problem (P_1), and more generally, for ℓ_p-quasinorm minimization with $0 < p \leq 1$, is the *FOCal Underdetermined System Solver* (FOCUSS) as proposed by Gorodnitsky, Rao and Kreutz-Delgado [33, 53]. Asymptotic convergence of FOCUSS to a stationary point from any initialization was claimed in [53], but the proof was not entirely accurate, as pointed out by [15]. One limitation of FOCUSS is that, unlike in IRLS as presented in Algorithm 1, no smoothing parameter ε is used, which leads to ill-conditioned linear systems.

To mitigate this, [14] proposed an IRLS method that uses smoothing parameters ε (such as used in Q_ε defined above) that are updated iteratively. It was observed that this leads to a better condition number for the linear systems to be solved in each step of IRLS and, furthermore, that this smoothing strategy has the advantage of finding sparser vectors if the weights of IRLS are chosen to minimize a non-convex ℓ_p-quasinorm for $p < 1$.

Further progress for IRLS designed to minimize an ℓ_1-norm was achieved in the seminal paper [22]. In [22], it was shown that if the measurement operator fulfills a suitable ℓ_1-null space property as in Definition 2.2, an IRLS method with iteratively updated smoothing converges to an s-sparse solution, coinciding with the ℓ_1-minimizer, if there exists one that is compatible with the measurements. This method uses not exactly the update rule of (4), but rather updates the smoothing parameter such that $\varepsilon_{k+1} = \min(\varepsilon_k, R(x^{k+1})_{s+1}/N)$, where $R(x^{k+1})_{s+1}$ is the $(s+1)^{\text{st}}$-largest element of the set $\{|x_j^{k+1}|, j \in [N]\}$. Furthermore, a *local linear convergence rate* of IRLS was established [22, Theorem 6.1] under same conditions.

However, the analysis of [22] has its limitations: First, there is a gap in the assumption of their convergence results between the sparsity s of a vector to be recovered and the order $\widehat{s}$ of the NSP of the measurement operator. Recently, this gap was circumvented in [3] with an IRLS algorithm that uses a smoothing update rule based on an ℓ_1-norm, namely, $\varepsilon_{k+1} = \min(\varepsilon_k, \eta(1 - \rho_s)\sigma_s(x^{k+1})_{\ell_1}/N)$, where $\eta \in (0, 1)$, and ρ_s is the NSP constant of the order s of the NSP fulfilled by the measurement matrix A—this rule is quite similar to the rule (4) that we use in Algorithm 1. In

particular, [3, Theorem III.6] establishes convergence with local linear rate similar to [22] without the gap mentioned above. The main limitation, however, of the theory of [22] (which is shared by [3]) is that the linear convergence rate only holds *locally*, i.e., in a situation where the support of the sparse vector has already been identified, see also Section 3 and Section 4.1 for a discussion.

We finally mention two relevant recent papers for the theoretical understanding of IRLS. [56] explores the relationship of IRLS for ℓ_1-minimization and a slime mold dynamics, interpreting both as an instance of the same meta-algorithm. Without requiring any connection between sparse recovery and ℓ_1-minimization, [28] shows that an IRLS-like algorithm for (P_1), requires $O(N^{1/3}\log(\log(N)) + N^{1/3}\log(1/\delta)/\delta^{2/3} + \log(N)/\delta^2)$ iterations to obtain an multiplicative error of $1+\delta$ on the minimizer $||x||_1$. Unlike our result Theorem 3.2, this corresponds not to a linear, but to a sublinear convergence rate.

2.3 Related work

As mentioned in the introduction, IRLS has a long history and has appeared under different names within different communities, e.g., similar algorithms are usually called *half-quadratic algorithms* in image processing [2, 39] and the *Kačanov method* in numerical PDEs [24]. Probably the most common usage of IRLS has been in robust regression [34, 37], c.f. [9] for a survey that also covers applications in approximation theory. For p-norm regression, [1] proposed a version of IRLS for which convergence results for $p \in [2, \infty)$ were established, solving a problem that was open for over thirty years. Also, for robust regression, by using an ℓ_1-objective on the residual, [49] showed recently global convergence of IRLS with a linear rate, with high probability for sub-Gaussian data. We note that our proof strategy is different from the one of [49] due to a structural difference of (P_1) from robust regression.

In [52], the authors provide a general framework for formulating IRLS algorithms for the optimization of a quite general class of non-convex and non-smooth functions, however, without updated smoothing. They use techniques developed in [4] to show convergence of the sequence of iterates to a critical point under the Kurdyka-Łojasiewicz property [7]. However, no results about convergence rates were presented.

For the sparse recovery problem, the topic discussed in this paper, the references [30, 42, 61] analyzed IRLS for an unconstrained version of (P_1), which is usually a preferable formulation if the measurements are corrupted by noise. Additionally, the work [30] addressed the question of how to solve the successive quadratic optimization problems. The authors developed a theory that shows, under the NSP, how accurately the quadratic subproblems need to be solved via the conjugate gradient method in order to preserve the convergence results established in [22].

Finally, for the related problems of low-rank matrix recovery and completion, IRLS strategies have emerged as one of the most successful methods in terms of data-efficiency and scalability [29, 40, 41, 47].

3 IRLS for Basis Pursuit with Global Linear Rate

As discussed in Section 2.2, the main theoretical advancements for IRLS for the sparse recovery problem were achieved in the work [22].

Proposition 3.1. [22, Theorem 6.1] *Assume that $A \in \mathbb{R}^{m\times N}$ satisfies the NSP of order $\widehat{s} > s$ with constant $\rho_{\widehat{s}}$ such that $0 < \rho_{\widehat{s}} < 1 - \frac{2}{\widehat{s}+2}$ and $\hat{s} > s + \frac{2\rho_{\widehat{s}}}{1-\rho_{\widehat{s}}}$ hold. Let $x_* \in \mathbb{R}^N$ be an s-sparse vector and set $y = Ax_*$. Assume that there exists an integer $k_0 \geq 1$ and a positive number $\xi > 0$ such that*

$$\xi := \frac{\|x^{k_0} - x_*\|_1}{\min_{i\in S} |(x_*)_i|} < 1. \tag{7}$$

Then the iterates $\{x^{k_0}, x^{k_0+1}, x^{k_0+2}, \ldots\}$ of the IRLS method in [22] *converge* linearly *to x_*, i.e., for all $k \geq k_0$, the kth iteration of IRLS satisfies*

$$\|x^{k+1} - x_*\|_1 \leq \frac{\rho_{\widehat{s}}(1+\rho_{\widehat{s}})}{1-\xi}\left(1 + \frac{1}{\widehat{s}-1-s}\right)\|x^k - x_*\|_1. \tag{8}$$

The main contribution of this paper is that we overcome a local assumption such as (7) and show that IRLS as defined by Algorithm 1 *exhibits a global linear convergence rate*, i.e., there is a linear convergence rate starting from any initialization, as early as in the first iteration.

Exactly sparse case. Our first main result, Theorem 3.2, deals with the scenario that the ground truth vector x_* is exactly s-sparse. Our second result, presented in the supplementary material, generalizes the first one to the more realistic situation of approximately sparse vectors.

Theorem 3.2. *Consider the problem of recovering an unknown s-sparse vector $x_* \in \mathbb{R}^N$ from known measurements of the form $y = Ax_*$. Assume that the measurement matrix $A \in \mathbb{R}^{m\times N}$ fulfills the ℓ_1-NSP of order s with constant $\rho_s < 1/2$. Let the IRLS iterates $\left\{x^k\right\}_k$ and $\{\varepsilon_k\}_k$ be defined by the IRLS algorithm* (3) *and* (4) *with initialization x^0. Then, for all $k \in \mathbb{N}$, it holds that*

$$\mathcal{J}_{\varepsilon_k}(x^k) - \|x_*\|_1 \leq \left(1 - \frac{c}{\rho_1 N}\right)^k \left(\mathcal{J}_{\varepsilon_0}(x^0) - \|x_*\|_1\right) \tag{9}$$

as well as

$$\|x^k - x_*\|_1 \leq 9\left(1 - \frac{c}{\rho_1 N}\right)^k \|x^0 - x_*\|_1. \tag{10}$$

Here $c = 1/768$ is an absolute constant and $\rho_1 < 1/2$ denotes the ℓ_1-NSP constant of order 1.

Inequality (9) says that the difference $\mathcal{J}_{\varepsilon_k}(x^k) - \|x_*\|_1$ converges linearly with a uniform upper bound of $1 - \frac{c}{\rho_1 N}$ on the linear convergence factor. As our proof, which is detailed in the supplementary material, shows, this implies inequality (10), which implies that also $\|x_* - x^k\|_1$ exhibits linear convergence in the number of iterations k. In particular, this means that for some error tolerance $\delta > 0$, we obtain $\|x_* - x^k\|_1 \leq \delta$ after $O\left(\rho_1 N \log\left(\frac{\|x_* - x^0\|_1}{\delta}\right)\right)$ iterations.

Remark 3.3. *Note that it follows directly from Definition 2.2 that the constant ρ_1 of the ℓ_1-NSP of order* 1 *satisfies $\rho_1 \leq \rho_s \leq 1$, which implies that δ-accuracy is obtained after $O\left(N \log\left(\frac{\|x_* - x^0\|_1}{\delta}\right)\right)$ iterations. This bound can be improved in many scenarios where one can obtain more explicit bounds on ρ_1, for example, when A is a Gaussian matrix. Namely, inspecting* [31, p. 142 and Thm. 9.2]*, we observe in this scenario that $\rho_1 \lesssim \sqrt{(\log N)/m}$ with high probability. Hence, in this scenario, at most $O\left(N\sqrt{\frac{\log N}{m}} \log\left(\frac{\|x_* - x^0\|_1}{\delta}\right)\right)$ iterations are needed to achieve δ-accuracy.*

The key idea in our proof is to use fact that the quadratic functional $Q_{\varepsilon_k}(\cdot, x^k)$ approximates the ℓ_1-norm in a neighborhood of the current iterate x^k. For this reason, we also expect that for $t > 0$ sufficiently small, we have that $Q_{\varepsilon_k}(x^k + tv^k, x^k) < Q_{\varepsilon_k}(x^k, x^k)$ if $v^k = x_* - x^k$ is the vector between x^k and the ground truth x_*. Then by choosing t properly, we can guarantee a sufficient decrease of the functional $\mathcal{J}_{\varepsilon^k}\left(x^k\right)$ in each iteration.

4 Numerical experiments

In this section, we support our theory with numerical experiments. First, we examine whether IRLS indeed exhibits two distinct convergence phases, a "global" one, as described in this paper, and a local one, as described in [3, 22], corresponding to different linear convergence rate factors. Second, we explore to which extent the dimension dependence in the convergence rates (9) and (10) indicated by Theorem 3.2 is necessary, or if we rather can expect a dimension-free linear convergence rate factor. All experiments are conducted on an iMac computer with a 4 GHz Quad-Core Intel Core i7 CPU, using MATLAB R2020b.

4.1 Local and global convergence phase

We first note that the local convergence result of [22, Theorem 6.1] depends on the locality condition $\xi(k) := \frac{\|x^k - x_*\|_1}{\min_{i\in S}|(x_*)_i|} < 1$, cf. (7). Under this condition (and an appropriate null space condition), Proposition 3.1 stated above implies that

$$\|x^{k+1} - x_*\|_1 \leq \mu \|x^k - x_*\|_1$$

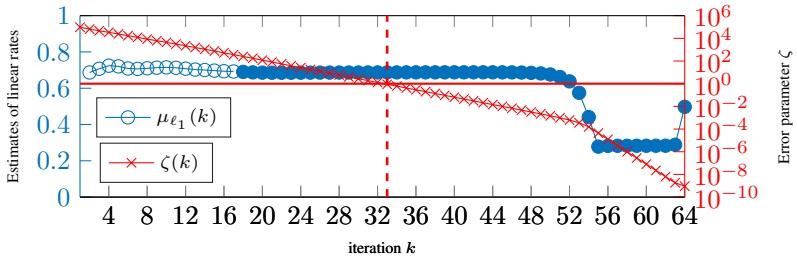


Figure 1: Instantaneous linear convergence rates of IRLS for ℓ_1-minimization ($N = 8000$): Linear conv. factors $\mu_{\ell_1}(k) := \|x^k - x_*\|_1 / \|x^{k-1} - x_*\|_1$ (in blue), filled blue circle if $S_k = S$ with S_k of Proposition 4.1 (perfect support identification), and error parameter $\zeta(k) := \|x^k - x_*\|_1 / \min_{i \in S} |(x_*)_i|$ (in red). Horizontal (red) line: Threshold $\zeta = 1$. Vertical (red) line: First iterate k with $\zeta(k) < 1$.

with an absolute constant $\mu < 1$ which, in particular, does *not* depend on the dimension N, m, and s. This corresponds to a locally linear rate for IRLS. A very similar condition to (7) is required by the comparable and more recent local convergence statement [3, Theorem III.6, inequality (III.14)] for the IRLS variant considered in [3].

However, a closer look at the locality condition (7) reveal that its *basin of attraction* is very restrictive: This condition means that the *support identification* problem underlying the sparse recovery *has already been solved*, as can be seen from the following proposition, whose proof we provide in the supplementary material.

Proposition 4.1. *Let $x^k, x_* \in \mathbb{R}^N$, let $S \subset [N]$ be the support set of x_* of size $|S| = s$. If (7) holds, i.e., if $\|x^k - x_*\|_1 < \min_{i \in S} |(x_*)_i|$, then the set $S_k \subset [N]$ of the s largest coordinates of x^k coincides with S.*

We now explore the behavior of the IRLS algorithm for ℓ_1-minimization, Algorithm 1, and the sharpness of Proposition 4.1 in experiments that build on those of [22, Section 8.1]. For this purpose, for $N = 8000$, we sample independently a 200-sparse vector $x_* \in \mathbb{R}^N$ with random support $S \subset [N]$, $s = 200 = |S|$, chosen uniformly at random such that $(x_*)_S$ is chosen according the Haar measure on the sphere of a 200-dimensional unit ℓ_2-ball, and a measurement matrix $A \in \mathbb{R}^{m \times N}$ with i.i.d. Gaussian entries such that $A_{ij} \sim \mathcal{N}(0, 1/m)$, while setting $m = \lfloor 2s \log(N/s) \rfloor$. Such a matrix is known to fulfill with high probability the ℓ_1-null space property of order s with constant $\rho_s < 1$ [12, 31].

In Figure 1, we track the decay of the ℓ_1-error $\|x^k - x_*\|_1$ of the iterates x^k returned by Algorithm 1 via the values of $\zeta(k) := \|x^k - x_*\|_1 / \min_{i \in S} |(x_*)_i|$, depicted in red, and the behavior of the factor $\mu_{\ell_1}(k) := \|x^k - x_*\|_1 / \|x^{k-1} - x_*\|_1$, depicted in blue. We observe that the condition (7) for local convergence with the fast, dimension-less linear rate (8) is satisfied after $k = 33$ iterations, as indicated by the vertical dashed red line.

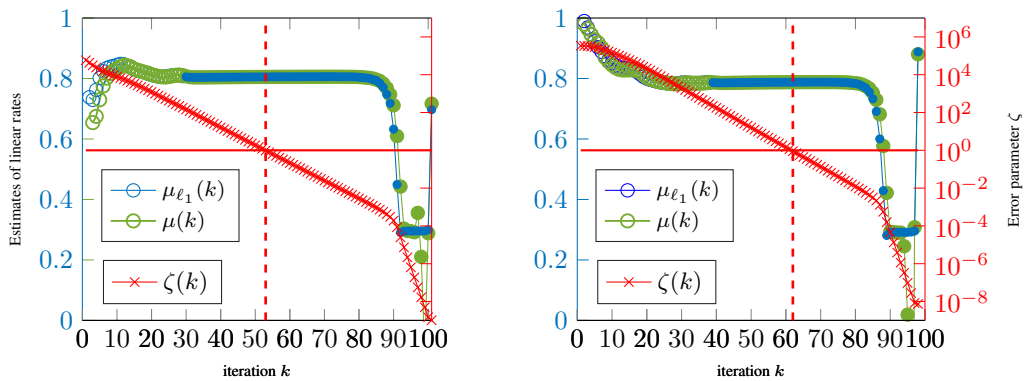
In the first few iterations, $\zeta(k)$ is larger than 1 by several orders of magnitudes, suggesting that the local convergence rate results of [3, 22] do *not* apply until the later stages of the simulation: In fact, we observe that the support S of x_* is already perfectly identified via the s largest coordinates of x^k as soon as $k \geq 18$. For iterations $18 \leq k \leq 50$, the linear rate $\mu_{\ell_1}(k)$ remains very stably around ≈ 0.7, after which an accelerated linear rate can be observed.[1] Before $k = 18$, for this example, the rate $\mu(k)$ hovers around 0.7 with slight variations. For all iterations k, $\mu(k)$ is smaller than 1, in line with the global linear convergence rate implied by Theorem 3.2.

Repeating a similar experiment for a larger ambient space dimension $N = 16000$ and a smaller measurement-to-sparsity ratio such that $m = \lfloor 1.75 s \log(N/s) \rfloor$ results in a qualitatively similar situation, as seen in Figure 2(a): In Figure 2(a), we add also a plot of the linear convergence factor $\mu(k) := \frac{\mathcal{J}_{\varepsilon_k}(x^k) - \|x_*\|_1}{\mathcal{J}_{\varepsilon_{k-1}}(x^{k-1}) - \|x_*\|_1}$ that tracks the behavior of the linear convergences in the smoothed

[1] The latter phenomenon cannot be observed for the IRLS algorithm of [22] as it uses a slightly different objective function than Algorithm 1.

ℓ_1-norm objective $\mathcal{J}$. In addition to what have been observed in Figure 1, we see that $\mu(k)$ and $\mu_{\ell_1}(k)$ exhibit a very similar behavior for this example.

Hence, these experiments indicate that we can distinguish two phases. In the first, global phase linear convergence already sets in, but the instantaneous linear convergence rate has not yet stabilized. In the second one, when the support identification problem has been solved, the instantaneous linear convergence stabilizes.



(a) Standard initialization (uniform weights $(w_0)_i = 1$ for all i).

(b) Adversary initialization (weights $(w_0)_i$ as in (11)).

Figure 2: Instantaneous linear convergence rates of IRLS for ℓ_1-minimization ($N = 16000$): Linear convergence factors $\mu_{\ell_1}(k) := \frac{\|x^k - x_*\|_1}{\|x^{k-1} - x_*\|_1}$ (in blue) and $\mu(k) := \frac{\mathcal{J}_{\varepsilon_k}(x^k) - \|x_*\|_1}{\mathcal{J}_{\varepsilon_{k-1}}(x^{k-1}) - \|x_*\|_1}$ (in green), filled circles if $S_k = S$ (perfect support identification), and error parameter $\zeta(k) := \|x^k - x_*\|_1 / \min_{i \in S} |(x_*)_i|$ (in red), horizontal and vertical red lines as in Figure 1.

4.2 Global convergence rate and its dimension dependence

In this section, we explore to which extent the dependence on N in the convergence rates (10) is necessary or if we can rather expect a dimension-free linear convergence rate factor. To this end, we run a variation of IRLS that initializes the weight vector $w_0 \in \mathbb{R}^N$ not uniformly as in Algorithm 1, but based on an *adversary initialization*, here denoted by z^{adv}. More specifically, we first compute a minimizer

$$z^{\text{adv}} \in \underset{z \in \mathbb{R}^{S^c} : A_{S^c} z = y}{\arg\min} \|z\|_1$$

of the ℓ_1-minimization problem restricted to the off-support coordinates of x_* indexed by S^c and set then $x^0 \in \mathbb{R}^N$ such that $x^0_{S^c} := z^{\text{adv}}$ and $x^0_S = 0$. Based on this *initialization* x^0, we compute $\varepsilon_0 := \frac{\sigma_s(x^0)_{\ell_1}}{N}$ and set the first weight vector such that for all $i \in [N]$,

$$(w_0)_i := \frac{1}{\max\left(|x^0_i|, \varepsilon_0\right)}, \tag{11}$$

before proceeding with the IRLS steps (3), (4) and (5) until convergence.

We observe in Figure 2(b) that this initialization, which is *adversary* as it sets very large initial weights on the coordinates of S that correspond to the true support of x_*, eventually results in the same behavior of Algorithm 1 as for the standard initialization by uniform weights, identifying the true support at iteration $k = 39$ compared to $k = 30$. However, in the first few iterations, we see that the instantaneous linear convergence factor $\mu(k)$ is close to 1 with $\mu(1) = 0.980$, decreasing only slowly before stabilizing around 0.79 after around $k = 30$.

While this is just one example, this already indicates that in general, a linear rate such as (8), i.e., without dependence on the dimension N (which has been proven locally in [22, Theorem 6.1] and [3, Theorem III.6]) might not hold in general.

In our next experiment, we further investigate numerically the dimension dependence of the worst-case linear convergence factor $\mu(k) := \frac{\mathcal{J}_{\varepsilon_k}(x^k) - \|x_*\|_1}{\mathcal{J}_{\varepsilon_{k-1}}(x^{k-1}) - \|x_*\|_1}$, which is upper bounded by the result of

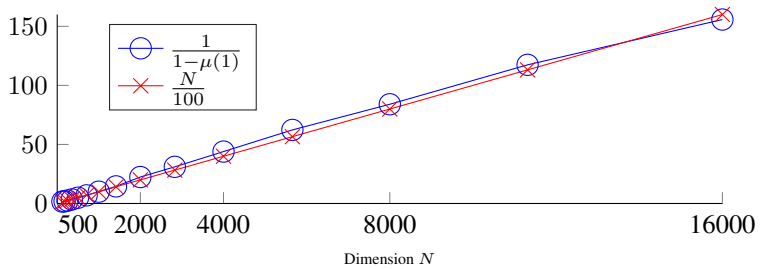


Figure 3: Comparison of $\frac{N}{100}$ and $\frac{1}{1-\mu(1)}$ (for which Theorem 3.2 provides un upper bound of $\frac{\rho_1 N}{c}$) for different dimension parameters N, where $\mu(1) = \frac{\mathcal{J}_{\varepsilon_1}(x^1)-\|x_*\|_1}{\mathcal{J}_{\varepsilon_0}(x^0)-\|x_*\|_1}$ is the linear convergence factor, for IRLS initialized from adversary initialization.

Theorem 3.2. We saw that in the experiment using the adversary initialization mentioned above and depicted in Figure 2(b), the maximal value was attained in the first iteration, i.e., for $\mu(1)$, as the effect of the adversary initialization is most eminent for $k = 1$.

We now run IRLS starting from the adversary initialization for different ambient dimensions $N = 125 \cdot 2^{\ell/2}$ for $\ell = 0, 1, \ldots, 14$. For each of the values of N, we sample vectors $x_* \in \mathbb{R}^N$ of sparsity $s = 40$ from the same random model as above, and scale the number of i.i.d. Gaussian measurements with $m = \lfloor 2s \log(N/s) \rfloor$. We average the resulting values for $\mu(1)$ across 500 independent realizations of the experiment.

In Figure 3, we see that dependence on N of linear convergence factor $\mu(1)$ that is observed for this experiment is quite well described by the upper bound (9) provided by our main result Theorem 3.2, as $\frac{1}{1-\mu(1)}$ scales almost linearly with N. As a footnote in Section 3 indicates, the constant ρ_1 of the null space property of order 1 scales with $\sqrt{\frac{\log N}{m}}$, and therefore a precise dependence on all the parameters such as m and s might be more complicated than what can be observed in this experiment.

Nevertheless, we interpret Figure 3 as strong evidence that the linear convergence rate factor of **??** is tight in its dependence on N, and that a dimension-less factor μ cannot be expected in general.

Conclusion

In this paper, we solved an open problem in the algorithmic theory for sparse recovery, in particular, we established a new variant of the IRLS algorithm for Basis Pursuit or ℓ_1-minimization for which we show a global linear convergence rate under a suitable and sharp assumption, namely, the null-space property. Moreover, we have corroborated our theory with numerical experiments that, first, discussed the difference between the local and global convergence phase and, second, that elucidated the optimality of the dimension dependence of convergence rate given by our main theorem.

We think that the results in this paper give rise to a number of interesting research directions for follow-up work. While the numerical experiments in Section 4 substantiate the hypothesis that the dependence of the convergence rate on N and ρ_1 in our theory is not an artifact of our proof, we also observed in this section that for a *generic initialization* no such dependence can be observed. In view of this, it is interesting to investigate whether a dimension-independent global convergence rate is possible, for example via a *smoothed analysis* [21, 55]. Furthermore, there are currently no convergence rates available for IRLS optimizing a nuclear norm-type objective, which is of great interest for low-rank matrix recovery [29, 40, 47], and we expect that our analysis may be generalizable to this setting as well.

Finally, it was observed that sparse vectors can be recovered from even fewer measurements via the optimization of a non-convex ℓ_p-quasinorm (with $0 < p < 1$), and that IRLS exhibits excellent performance in this case [14, 22]. While a thorough understanding has remained elusive so far for this non-convex case, we consider our results as a first step towards a global convergence theory for IRLS for the optimization of ℓ_p-quasinorms or similar non-convex surrogate objectives.

References

[1] Deeksha Adil, Richard Peng, and Sushant Sachdeva, *Fast, provably convergent irls algorithm for p-norm linear regression*, Advances in Neural Information Processing Systems (NeurIPS), 2019, pp. 14189–14200.

[2] M. Allain, J. Idier, and Y. Goussard, *On global and local convergence of half-quadratic algorithms*, IEEE Trans. Image Process. **15** (2006), no. 5, 1130–1142.

[3] Aleksandr Aravkin, James V Burke, and Daiwei He, *IRLS for Sparse Recovery Revisited: Examples of Failure and a Remedy*, arXiv preprint arXiv:1910.07095 (2019).

[4] H. Attouch, J. Bolte, and B. Svaiter, *Convergence of descent methods for semi-algebraic and tame problems: Proximal algorithms, forward-backward splitting, and regularized gauss-seidel methods*, Math. Programming **137** (2013), 91–129.

[5] Richard Baraniuk, Mark Davenport, Ronald DeVore, and Michael Wakin, *A simple proof of the restricted isometry property for random matrices*, Constr. Approx. **28** (2008), 253–263.

[6] A. Beck and S. Sabach, *Weiszfeld's method: Old and new results*, J. Optim. Theory Appl. **164** (2015), 1–40.

[7] Jérôme Bolte, Aris Daniilidis, and Adrian Lewis, *The łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems*, SIAM J. Optim. **17** (2007), no. 4, 1205–1223.

[8] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, *Distributed optimization and statistical learning via the alternating direction method of multipliers*, Foundations and Trends® in Machine Learning **3** (2011), 1–122.

[9] C. S. Burrus, *Iterative reweighted least squares*, OpenStax CNX: Available online: http://cnx.org/contents/92b90377-2b34-49e4-b26f-7fe572db78a1 (2012).

[10] Jameson Cahill, Xuemei Chen, and Rongrong Wang, *The gap between the null space property and the restricted isometry property*, Linear Algebra Appl. **501** (2016), 363–375.

[11] E. J. Candès, J.K. Romberg, and T. Tao, *Stable signal recovery from incomplete and inaccurate measurements*, Comm. Pure Appl. Math. **59** (2006), no. 8, 1207–1223.

[12] E. J. Candès and T. Tao, *Near-Optimal Signal Recovery From Random Projections: Universal Encoding Strategies?*, IEEE Trans. Inf. Theory **52** (2006), no. 12, 5406–5425.

[13] Emmanuel J Candes, Michael B Wakin, and Stephen P Boyd, *Enhancing sparsity by reweighted ℓ_1 minimization*, Journal of Fourier analysis and applications **14** (2008), no. 5-6, 877–905.

[14] R. Chartrand and W. T. Yin, *Iteratively reweighted algorithms for compressive sensing*, Ieee international conference on acoustics, speech and signal processing (icassp), 2008, pp. 3869–3872.

[15] Rick Chartrand and Wotao Yin, *Nonconvex sparse regularization and splitting algorithms*, Splitting methods in communication, imaging, science, and engineering, 2016, pp. 237–249.

[16] Scott Shaobing Chen, David L Donoho, and Michael A Saunders, *Atomic decomposition by basis pursuit*, SIAM Rev. **43** (2001), no. 1, 129–159.

[17] Shaobing Scott Chen and David Donoho, *Basis pursuit*, Proc. of 1994 28th asilomar conference on signals, systems and computers, 1994, pp. 41–44.

[18] J. F. Claerbout and F. Muir, *Robust modeling with erratic data*, Geophysics **38** (1973), no. 5, 826–844.

[19] Albert Cohen, Wolfgang Dahmen, and Ronald DeVore, *Compressed sensing and best k-term approximation*, J. Amer. Math. Soc. **22** (2009), 211–231.

[20] Patrick L Combettes and Valérie R Wajs, *Signal recovery by proximal forward-backward splitting*, Multiscale Model. Simul. **4** (2005), no. 2, 1168–1200.

[21] D. Dadush and S. Huiberts, *A friendly smoothed analysis of the simplex method*, SIAM J. Comput. **49** (2018), no. 5, STOC18–449–STOC18–499.

[22] I. Daubechies, R. DeVore, M. Fornasier, and C.S. Güntürk, *Iteratively reweighted least squares minimization for sparse recovery*, Comm. Pure Appl. Math. **63** (2010), 1–38.

[23] G. Davis, S. Mallat, and M. Avellanedas, *Adaptive greedy approx-imations*, Constr. Approx. **13** (1997), no. 1, 57–98.

[24] L. Diening, M. Fornasier, Tomasi R., and M Wank, *A Relaxed Kačanov iteration for the p-poisson problem*, Numer. Math. **145** (2020), no. 2, 1–34.

[25] D. L. Donoho, *Compressed sensing*, IEEE Trans. Inf. Theor. **52** (2006), no. 4, 1289–1306.

[26] David Donoho and Yaakov Tsaig, *Fast solution of ℓ_1-norm minimization problems when the solution may be sparse*, IEEE Trans. Inform. Theory **54** (2008), 4789–4812.

[27] David L Donoho and Benjamin F Logan, *Signal recovery and the large sieve*, SIAM J. Appl. Math. **52** (1992), no. 2, 577–591.

[28] Alina Ene and Adrian Vladu, *Improved convergence for ℓ_1 and ℓ_∞ regression via iteratively reweighted least squares*, International Conference on Machine Learning, 2019, pp. 1794–1801.

[29] M. Fornasier, H. Rauhut, and R. Ward, *Low-rank matrix recovery via iteratively reweighted least squares minimization*, SIAM J. Optim. **21** (2011), 1614–1640.

[30] Massimo Fornasier, Steffen Peter, Holger Rauhut, and Stephan Worm, *Conjugate gradient acceleration of iteratively re-weighted least squares methods*, Comput. Optim. Appl. **65** (2016), no. 1, 205–259.

[31] S. Foucart and H. Rauhut, *A mathematical introduction to compressive sensing*, Applied and Numerical Harmonic Analysis, Springer New York, 2013.

[32] D. Geman and G. Reynolds, *Constrained Restoration and the Recovery of Discontinuities*, IEEE Trans. Pattern Anal. Mach. Intell. **14** (1992), no. 03, 367–383.

[33] I. F. Gorodnitsky and B. D. Rao, *Sparse signal reconstruction from limited data using focuss: A re-weighted minimum norm algorithm*, IEEE Trans. Signal Process. **45** (1997), no. 3, 600–616.

[34] Peter J Green, *Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives*, Journal of the Royal Statistical Society: Series B (Methodological) **46** (1984), no. 2, 149–170.

[35] Remi Gribonval and Morten Nielsen, *Sparse representation in the union of basis*, IEEE Trans. Inform. Theory **49** (2003), 3320–3325.

[36] Trevor Hastie, Robert Tibshirani, and Martin Wainwright, *Statistical learning with sparsity: The lasso and generalizations*, Chapman and Hall/CRC, 2019.

[37] P. W. Holland and R. E. Welsch, *Robust regression using iteratively reweighted least-squares*, Communications in Statistics - Theory and Methods **6** (1977), 813–827.

[38] Peter J. Huber, *Robust estimation of a location parameter*, Ann. Math. Statist. **35** (1964), no. 1, 73–101.

[39] J. Idier, *Convex half-quadratic criteria and interacting auxiliary variables for image restoration*, IEEE Trans. Image Process. **10** (2001), no. 7, 1001–1009.

[40] C. Kümmerle and J. Sigl, *Harmonic mean iteratively reweighted least squares for low-rank matrix recovery*, J. Mach. Learn. Res. **19** (2018), 1815–1863.

[41] Christian Kümmerle and Claudio Mayrink Verdun, *A scalable second order method for ill-conditioned matrix completion from few samples*, 38th international conference on machine learning, vienna, austria, 2021.

[42] M.-J. Lai, X. Yangyang, and Y. Wotao, *Improved iteratively reweighted least squares for unconstrained smoothed ℓ_q minimization*, SIAM J. Numer. Anal. **2** (2013), no. 51, 927–957.

[43] Gilad Lerman and Tyler Maunu, *Fast, robust and non-convex subspace recovery*, Information and Inference: A Journal of the IMA **7** (2018), no. 2, 277–336.

[44] X. Li, D. Sun, and K.-C. Toh, *A highly efficient semismooth newton augmented lagrangian method for solving lasso problems*, SIAM J. Optim. **28** (2018), no. 1, 433–458.

[45] Nicolai Meinshausen, Peter Bühlmann, et al., *High-dimensional graphs and variable selection with the lasso*, Annals of statistics **34** (2006), no. 3, 1436–1462.

[46] Gregory P. Meyer, *An Alternative Probabilistic Interpretation of the Huber Loss*, arXiv preprint arXiv:1911.02088 (2020).

[47] K. Mohan and M. Fazel, *Iterative reweighted algorithms for matrix rank minimization*, J. Mach. Learn. Res. **13** (2012), 3441–3473.

[48] B. Mukhoty, G. Gopakumar, P. Jain, and P. Kar, *Globally-convergent iteratively reweighted least squares for robust regression problems*, ArXiv e-prints (2020), available at 2006.14211.

[49] Bhaskar Mukhoty, Govind Gopakumar, Prateek Jain, and Purushottam Kar, *Globally-convergent iteratively reweighted least squares for robust regression problems*, Proceedings of machine learning research, 2019, pp. 313–322.

[50] Balas K. Natarajan, *Sparse approximate solutions to linear systems*, SIAM J. Comput. **24** (1996), 227–234.

[51] M. Nikolova and M. K. Ng, *Analysis of half-quadratic minimization methods for signal and image recovery*, SIAM J. Sci. Comput. **27** (2005), no. 3, 937–966.

[52] Peter Ochs, Alexey Dosovitskiy, Thomas Brox, and Thomas Pock, *On iteratively reweighted algorithms for nonsmooth nonconvex optimization in computer vision*, SIAM J. Imaging Sci. **8** (2015), no. 1, 331–372.

[53] B. D. Rao and K. Kreutz-Delgado, *An affine scaling methodology for best basis selection*, IEEE Trans. Signal Process. **47** (1999), no. 1, 187–200.

[54] M. Rudelson and R. Vershynin, *On sparse reconstruction from fourier and gaussian measurements*, Comm. Pure Appl. Math. **61** (2008), no. 8, 1025–1045.

[55] D. A. Spielman and S.-H. Teng, *Smoothed analysis of termination of linear programming algorithms*, Math. Program., Ser. B **97** (2003), 375–404.

[56] Damian Straszak and Nisheeth K Vishnoi, *Iteratively reweighted least squares and slime mold dynamics: connection and convergence*, Mathematical Programming (2021), 1–33.

[57] Y. Sun, P. Babu, and D. P. Palomar, *Majorization-minimization algorithms in signal processing, communications, and machine learning*, IEEE Trans. Signal Process. **65** (2017), no. 3, 794–816.

[58] Howard L Taylor, Stephen C Banks, and John F McCoy, *Deconvolution with the ℓ_1 norm*, Geophysics **44** (1979), no. 1, 39–52.

[59] Robert Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society: Series B (Methodological) **58** (1996), no. 1, 267–288.

[60] Andreas M Tillmann, *Equivalence of linear programming and basis pursuit*, PAMM **15** (2015), no. 1, 735–738.

[61] Sergey Voronin and Ingrid Daubechies, *An iteratively reweighted least squares algorithm for sparse regularization*, Functional analysis, harmonic analysis, and image processing: A collection of papers in honor of björn jawerth, 2017, pp. 391–441.

[62] E. Weiszfeld, *Sur le point pour lequel la somme des distances de n points donnés est minimum*, Tohoku Mathematical Journal, First Series **43** (1937), 355–386.

[63] Max A. Woodbury, *Inverting modified matrices*, Memorandum report **42** (1950), no. 106, 336.

[64] Wotao Yin, Stanley Osher, Donald Goldfarb, and Jerome Darbons, *Bregman iterative algorithms for ℓ_1-minimization with applications to compressed sensing*, SIAM J. Imaging Sci **1** (2008), no. 1, 143–168.

[65] Zaiwei Zhang, Xiangru Huang, Qixing Huang, Xiao Zhang, and Yuan Li, *Joint learning of neural networks via iterative reweighted least squares.*, Cvpr workshops, 2019, pp. 18–26.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 4, where we discuss the sharpness of our result and whether our convergence rate is optimal.
   (c) Did you discuss any potential negative societal impacts of your work? [No] This is a theoretical contribution. We expect it to be insightful for the theory community and lead to improved algorithms, but we do not see any immediate societal impact of our work.

(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] See Section 3.
   (b) Did you include complete proofs of all theoretical results? [Yes] Due to space limitations, we provide all proofs in the supplemental material.

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 provide MATLAB code to reproduce the experiments of Section 4 in the supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] In the experiment of Section 4.2, cf. Figure 3, we averaged the results across 500 independent realizations, as indicated. We did not report error bars as the observed variance was small and does not really affect the conclusion of the experiment.
   (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)? [Yes] Yes, we unly used the CPU of a personal computer, see the beginning of Section 4.

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? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

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]