ITERATIVELY REWEIGHTED LEAST SQUARES FOR ℓ_1-MINIMIZATION WITH GLOBAL LINEAR CONVERGENCE RATE

CHRISTIAN KÜMMERLE*, CLAUDIO MAYRINK VERDUN†, AND DOMINIK STÖGER‡

ABSTRACT. Iteratively Reweighted Least Squares (IRLS), whose history goes back more than 80 years, represents an important family of algorithms for non-smooth optimization as it is able to optimize these problems by solving a sequence of linear systems. In 2010, Daubechies, DeVore, Fornasier, and Güntürk proved that IRLS for ℓ_1-minimization, an optimization program ubiquitous in the field of compressed sensing, *globally converges* to a sparse solution. While this algorithm has been popular in applications in engineering and statistics, fundamental algorithmic questions have remained unanswered. As a matter of fact, existing convergence guarantees only provide global convergence *without any rate*, except for the case that the support of the underlying signal has already been identified. In this paper, we prove that IRLS for ℓ_1-minimization converges to a sparse solution *with a global linear rate*. We support our theory by numerical experiments indicating that our linear rate essentially captures the correct dimension dependence.

1. INTRODUCTION

The field of sparse recovery deals with the problem of recovering an (approximately) sparse vector x from an underdetermined system of linear equations of the form $y = Ax$. One approach to solve this problem is to consider

$$\min_{x \in \mathbb{R}^N} ||x||_0 \qquad \text{subject to } Ax = y, \tag{P_0}$$

where $||x||_0$ denotes the number of nonzero entries of the vector $x \in \mathbb{R}^N$. In general, the problem (P_0) is NP-hard [Nat96, DMA97]. For this reason, (P_0) is often replaced by its convex relaxation

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

which in the literature is referred to as ℓ_1*(-norm) minimization* [TBM79, CM73, DL92] or *basis pursuit* [CD94, CDS01]. Unlike (P_0), the optimization program (P_1) is computationally tractable, and a close relationship of their minimizers has been recognized and well-studied in the theory of compressive sensing [CT06, CRT06, FR13].

In particular, the problem (P_1) can be reformulated as a linear program. As many problems of interest in applications are high-dimensional and therefore challenging for standard linear programming methods, many specialized solvers for (P_1) (or its unconstrained reformulation)

Date: December 21, 2020.

Key words and phrases. Compressed sensing, non-smooth optimization, reweighted algorithms, sparse recovery, global convergence, iteratively reweighted least squares.

*Department of Applied Mathematics & Statistics, Johns Hopkins University, Baltimore, USA (kuemmerle@jhu.edu).

†Department of Mathematics and Department of Electrical and Computer Engineering, Technical University of Munich, Munich, Germany (verdun@ma.tum.de).

‡Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California, California, USA (stoger@usc.edu).

have been proposed, such as the Homotopy Method [DT08], forward-backward algorithms [CW05], Alternating Direction Method of Multipliers [BPC+11], Bregman iterative regularization [YOGD08] and Semismooth Newton Augmented Lagragian Methods [LST18].

Another well-established solver for the ℓ_1-minimization problem (P_1) is Iteratively Reweighted Least Squares (IRLS), which is a method that minimizes non-smooth objectives by solving a sequence of quadratic problem. IRLS algorithms have a long history which dates back to a method proposed by Weiszfeld for the *Fermat-Weber problem* [Wei37, BS15].

In the sparse recovery context, IRLS was first introduced in [RKD99, GR97] and then analyzed by [DDFG10]. To this date, it remains a popular method for sparse recovery problems due its conceptual simplicity and the fact that unlike for many other methods, there is no necessity for step size tuning. A measure for the popularity of IRLS might be the number of Google Scholar citations of the four key papers [GR97, CY08, DDFG10, LYW13], which surpassed 4600 as of the writing of this article. While we do not study that variant in this paper, part of the popularity of IRLS is also due to the fact that the minimization of non-convex ℓ_p-quasinorms with $0 < p < 1$ can be tackled within this framework, where excellent numerical performance was observed. [CY08, DDFG10].

The theoretical analysis in [DDFG10] proved that IRLS converges to an ℓ_1-minimizer with local linear rate, as soon as the support of the true signal has been discovered. Moreover, it was proven in [DDFG10] that IRLS converges globally to an ℓ_1-minimizer. However, since this proof relies on the existence of a convergent subsequence, their proof does not reveal any rates for global convergence. While many extensions and modifications of the IRLS algorithm in [DDFG10] have been proposed (see, e.g., [ABH19, FRW11, FPRW16, MF12]), this following fundamental algorithmic question has remained unanswered:

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

We resolve this question and show that *IRLS converges linearly* to a sparse ground truth, *starting from any initialization*. Analogous to [DDFG10], it is assumed that the measurement matrix A satisfies the so-called null space property [CDD09]. 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 [FRW11].

1.1. **Organization of the paper.** This paper is organized as follows. In Section 2 we review IRLS for sparse recovery as well as the null space property. In Section 3 we state our two main theoretical results about a global linear convergence of IRLS for ℓ_1-minimization, Theorem 3.2 and Theorem 3.3, analyzing the case of sparse and approximately vectors, respectively. In Section 4, we support our theory by numerical experiments and finally, in Section 5, we prove the results of Section 3.

Notation. In this paper, 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 $\mathrm{supp}(x) = \{j \in [N] : x_j \neq 0\}$. 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. Background

2.1. **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 [SBP17], as we will lay out in the following.

IRLS 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

$$\mathcal{J}_\varepsilon(x) := \sum\nolimits_{i=1}^{N} j_\varepsilon(x_i), \tag{1}$$

where j_ε is given by

$$j_\varepsilon(x) := \begin{cases} |x|, & \text{if } |x| > \varepsilon, \\ \frac{1}{2}\left(\frac{x^2}{\varepsilon} + \varepsilon\right), & \text{if } |x| \leq \varepsilon. \end{cases}.$$

The function $\mathcal{J}_\varepsilon$ can be considered as a scaled Huber loss function which is widely used in robust regression analysis [Hub64, Mey20]. From this point of view, we note that the relationship between the ℓ_1-norm and its smoothed version $\mathcal{J}_\varepsilon$ of (1) is very similar to the smoothing achieved by using Huber M-estimators instead of ℓ_1-residuals in robust regression [LS98]. More specifically, 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 $Q(z, x) \geq \mathcal{J}_\varepsilon(z)$ for all $z \in \mathbb{R}^N$. This function is furthermore chosen such that $Q(x, x) = \mathcal{J}_\varepsilon(x)$ holds, which implies that $\min_{z \in \mathbb{R}^n} Q(z, x) \leq \mathcal{J}_\varepsilon(x)$.

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

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

where $\nabla \mathcal{J}_\varepsilon(x) \in \mathbb{R}^N$ is the gradient of $\mathcal{J}_\varepsilon$ at x and the weights $w_\varepsilon(x) \in \mathbb{R}^N$ is a vector of *weights* such that

$$w_\varepsilon(x)_i := [\max(|x_i|, \varepsilon)]^{-1} \quad \text{for } i \in [N].$$

The following lemma shows that $Q(\cdot, \cdot)$ has indeed the above-mentioned properties. We refer to Appendix A 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:*

(1) $\operatorname{diag}(w_\varepsilon(x))x = \nabla \mathcal{J}_\varepsilon(x)$,
(2) $Q_\varepsilon(x, x) = \mathcal{J}_\varepsilon(x)$,
(3) $Q_\varepsilon(z, x) \geq \mathcal{J}_\varepsilon(z)$.

As can be seen from the equality in (2), minimizing of $Q_\varepsilon(\cdot, x)$ corresponds to a minimizing *(re-)weighted least squares objective* $\langle \cdot, \mathrm{diag}(w_\varepsilon(x)) \cdot \rangle$, 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 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: $w_0 = (1, 1, \dots, 1)$).
Set $\varepsilon_0 = \infty$.
for $k = 0, 1, 2, \dots$ **do**

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

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

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

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

The solution $x^{(k+1)}$ of the linearly constrained weighted least squares step (3) be calculated by solving a positive definite linear system, can be tackled by a suitable iterative solver such as a conjugate gradient method.

We note that the smoothing update (4) is slightly different from the ones proposed in [DDFG10, ABH19], we refer to Section 2.3 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 [DDFG10, Theorem 5.3]).

2.2. **Null space property.** As in [DDFG10], the analysis we present is based on the assumption that the measurement matrix A satisfies the so-called null space property [CDD09, GN03], which is a key concept in the compressed sensing literature (see, e.g., [FR13, 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* $|S| \leq s$*,*

$$\|v_S\|_1 \leq \rho_s \|v_{S^c}\|_1$$

for all $v \in \ker(A) \backslash \{0\}$.

In [FR13, 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, as the following theorem shows.

Theorem 2.3 ([FR13, 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., [CCW16]), 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 [RV08, BDDW08]. Furthermore, the ℓ_1-NSP is known to hold for random matrices with i.i.d. entries whose distribution has a logarithmic number of finite moments and fulfills a small-ball condition, which includes a number of more heavy-tailed random matrices [ML17, DLR18].

2.3. **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 [GR97, RKD99]. Asymptotic convergence of FOCUSS to a stationary point from any initialization was claimed in [RKD99], but the proof was not entirely accurate, as pointed out by [CY16]. 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 and changes the algorithm's behavior fundamentally in the non-convex case of $p < 1$.

To mitigate this, [CY08] 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 $p < 1$. More specifically, their smoothing strategy was such that ε_k is initially 1 and then reduced by a factor of 10 each time the relative change $\frac{\|x^k - x^{x-1}\|_2}{\|x^k\|_2}$ from the previous iteration became a number smaller than $\sqrt{\varepsilon_k}/100$.

Further progress for IRLS designed to minimize an ℓ_1-norm was achieved in the seminal paper [DDFG10]. In [DDFG10], 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 [DDFG10, Theorem 6.1] under same conditions.

However, the analysis of [DDFG10] 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 [ABH19] 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, [ABH19, Theorem III.6] establishes convergence with local linear rate similar to [DDFG10] without the gap mentioned above.

The main limitation, however, of the theory of [DDFG10] (which is shared by [ABH19]) 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.

2.4. **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 [Idi01, AIG06] and the *Kačanov method* in numerical PDEs [DFRW20]. Probably the most popular usage of IRLS has been in robust regression [HW77]. For this problem, in particular, convergence results were recently obtained for $p \in [2, \infty)$ [APS19] and a global convergence linear rate was also obtained for $p = 1$ [MGJK19]. We refer to [Bur12] for a survey that also covers applications in approximation theory.

In [ODBP15], 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 [ABS13] to show convergence of the sequence of iterates to a critical point under the Kurdyka-Łojasiewicz property [BDL07]. However, no results about convergence rates were presented.

For the sparse recovery problem, the topic discussed in this paper, the references [LYW13, FPRW16, VD17] 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 [FPRW16] 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 [DDFG10].

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 [FRW11, MF12, KS18, KV20].

3. Main results

As discussed in Section 2.3, the main theoretical advancements for IRLS for the sparse recovery problem were achieved in the work [DDFG10].

Proposition 3.1. *[DDFG10, 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 $\{x^{k_0}, x^{k_0+1}, x^{k_0+2}, \dots\}$ of the IRLS method in [DDFG10] 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.

3.1. **Exactly sparse case.** Our first main result, Theorem 3.2, deals with the scenario that the ground truth vector $x_\star$ is exactly s-sparse.

Theorem 3.2. *Let $x_* \in \mathbb{R}^N$ be an s-sparse vector. Let $A \in \mathbb{R}^{m \times N}$ and $y = Ax_\star$. Assume that A fulfills the ℓ_1-NSP of order s with constant $\rho_s < 1/2$ and the ℓ_1-NSP of order 1 with constant ρ_1.*

Let the IRLS iterates $\{x^k\}_k$ *and* $\{\varepsilon_k\}_k$ *be defined by* (3) *and* (4) *with initialization* x^0*. Then, for all* $k \in \mathbb{N}$*, it holds that*

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

as well as

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

Here $c = 1/768$ *is an absolute constant.*

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 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 $O\left(\rho_1 N_1 \log\left(\frac{\|x_* - x^0\|_1}{\delta}\right)\right)$ iterations.[1]

The key idea in our proof is to use fact that the quadratic functional $Q(\cdot, x^k)$ 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 $Q(x^k + tv^k, x^k) < Q(x^k, x^k)$. 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, see also Proposition 5.3.

In Section 4, we conduct experiments that indeed verify the linear convergence in (9) and (10). Moreover, we study numerically whether one can observe a dependence of the convergence rate on the problem parameters N, s and m. We construct a worst-case example which indicates that the convergence rate indeed may depend on the dimension N in a way as described by (9).

3.2. **Approximately sparse case.** We now generalize the result of Section 3.1 to the scenario where the ground truth x_* is only approximately sparse. Again, this was also studied in [DDFG10], but as in the exactly sparse case only a local convergence result was obtained.
In this case, x_* can only be recovered *approximately* by ℓ_1-minimization. Hence, we expect that also Algorithm 1 will recover x_* only approximately. In our second main result we show that also in this case, we obtain global linear convergence of Algorithm 1. More precisely, Theorem 3.3 says that $\mathcal{J}_{\varepsilon_k}(x^k) - \|x_*\|_1$ decays exponentially fast until a certain accuracy is reached, which scales linearly with $\sigma_s(x_*)_{\ell_1}$.

Theorem 3.3. *Let* $x_* \in \mathbb{R}^N$*. Assume that* A *fulfills the* ℓ_1*-NSP of order* s *with constant* $\rho_s < 1/8$ *and the* ℓ_1*-NSP of order* 1 *with constant* $\rho_1 < 1$*. Assume that* $y = Ax_\star$*. Let the IRLS iterates* $\{x^k\}_k$ *and* $\{\varepsilon_k\}_k$ *be defined by* (3) *and* (4) *with initialization* x^0*. Then the following three statements hold.*

(1) For $k \leq \hat{k} := \min\left\{k \in \mathbb{N} : \sigma_s(x_*)_{\ell_1} > \frac{2}{9}\|\left(x_* - x^k\right)_{S^c}\|_1\right\}$ *it holds that*

$$\mathcal{J}_{\varepsilon_k}(x^k) - \|x_*\|_1 \leq \left(1 - \frac{c}{\rho_1 N}\right)^k \left(\mathcal{J}_{\varepsilon_0}(x_*) - \|x_*\|_1\right), \tag{11}$$

where S *denotes the support of the* s *largest entries of* x_**.*

[1]Inspecting [FR13, p. 142 and Thm. 9.2] we observe that $\rho_1 \lesssim \sqrt{(\log N)/m}$, if A is a Gaussian matrix, which implies that in this case $\rho_1 N$ is at the order of $N\sqrt{\frac{\log N}{m}}$.

(2) For all $1 \leq k \leq \hat{k}$ *it holds that*

$$\|x^k - x_*\|_1 \leq 6\left(1 - \frac{c}{\rho_1 N}\right)^k \|x^0 - x_*\|_1 + 10\sigma_s(x_*)_{\ell_1}. \tag{12}$$

(3) Moreover, for all integers $k \gtrsim \rho_1 N \log\left(\frac{\|x^0 - x_*\|_1}{\sigma_s(x_*)_{\ell_1}}\right)$ *we have that*

$$\|x^k - x_*\|_1 \leq 20\sigma_s(x_*)_{\ell_1}. \tag{13}$$

Here $c = 1/3072$ *is an absolute constant.*

Remark 3.4. *Applying Theorem 3.3 to the special case* $\sigma_s(x_*)_{\ell_1} = 0$*, we observe that inequality (10) yields a seemingly sharper result than inequality (12) in Theorem 3.2, which may seem somewhat counterintuitive. However, note that in Theorem 3.2 we require* $\rho_s < 1/2$*, whereas in Theorem 3.3 we have the stronger assumption* $\rho_s < 1/8$*. Indeed, a closer inspection of the proofs reveals that both the factors* 3 *and* 6 *in the inequalities (10) and (12) can be replaced by the factor* $3\frac{1+\rho_s}{1-\rho_s}$*, reconciling those two results.*

4. Numerical Experiments

In this section, we support our theory with numerical experiments, whose purpose is twofold.

In the first part, 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 [DDFG10, ABH19], corresponding to different linear convergence rate factors. In the second part, 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.

4.1. Local and global convergence phase.

We first note that the local convergence result of [DDFG10, 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

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

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 [ABH19, Theorem III.6, inequality (III.14)] for the IRLS variant considered in [ABH19].

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 Appendix B.

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*.*

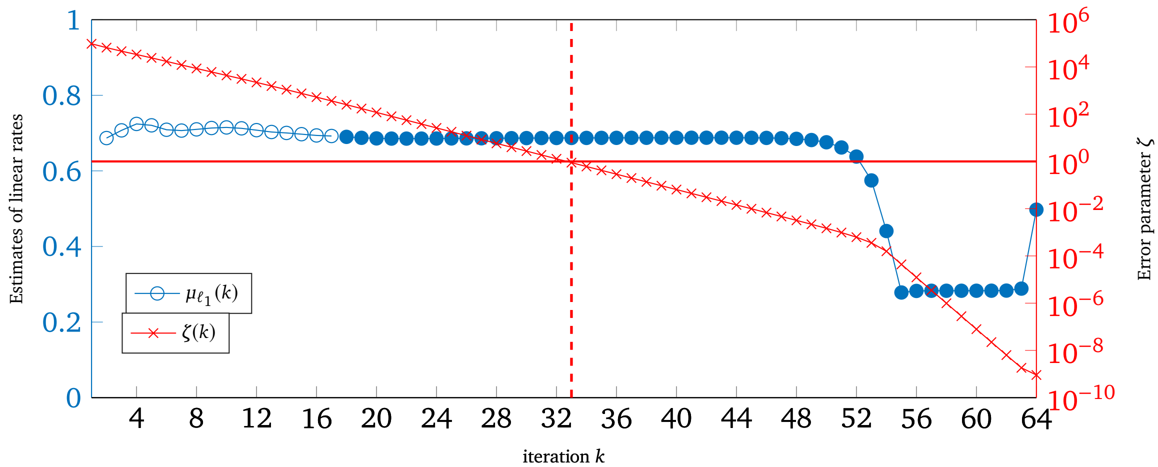


FIGURE 1. Instantaneous linear convergence rates of IRLS for ℓ_1-minimization ($N = 8000$): Linear convergence factors $\mu_{\ell_1}(k) := \|x^k - x_*\|_1/\|x^{k-1} - x_*\|_1$ (in blue), filled blue circle if $S_k = S$ with S_k of Proposition 4.1 (perfect support identification), and error parameter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

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 [DDFG10, Section 8.1]. For this purprose, 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 $\rho_s < 1$ [CT06, FR13].

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.

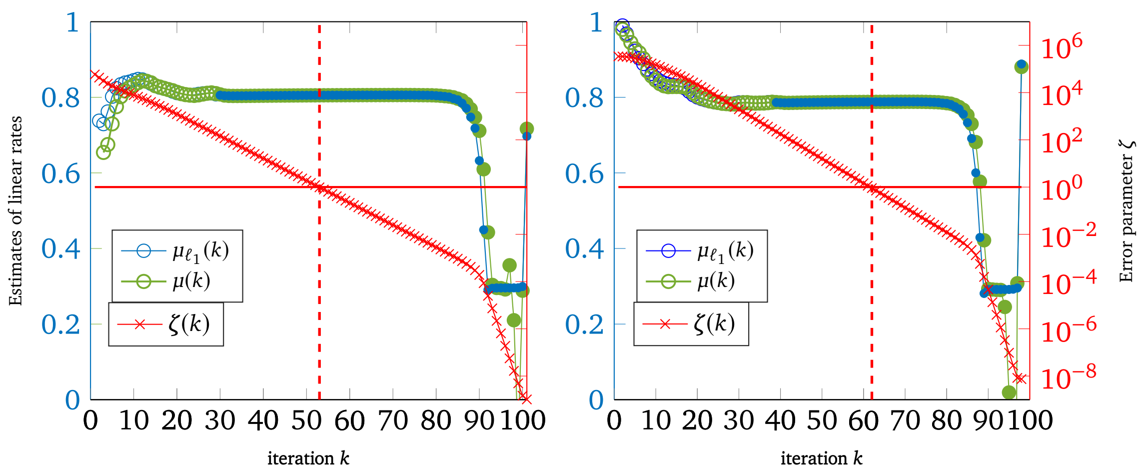
In the first few iterations, $\zeta(k)$ is larger than 1 by several orders of magnitudes, suggesting that the local convergence rate results of [DDFG10, ABH19] do *not* apply until the later stages of the simulation: In fact, we observe that the support S of x_* is already perfectly identified via the 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.[2] 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.

[2]The latter phenomenon cannot be observed for the IRLS algorithm of [DDFG10] as it uses a slightly different objective function than Algorithm 1.

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 ℓ_1-norm objective $\mathcal{J}$, cf. (17).

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 (14)).

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) and (13) 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 $\epsilon_0 := \frac{\sigma_s(x^0)_{\ell_1}}{N}$ and set the first weight vector such that for all $i \in [N]$,

$$(w_0)_i := \frac{1}{\max\left(|x_i^0|, \epsilon_0\right)}, \tag{14}$$

before proceeding with the IRLS steps (3)–(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 [DDFG10, Theorem 6.1] and [ABH19, Theorem III.6]) might not hold in general.

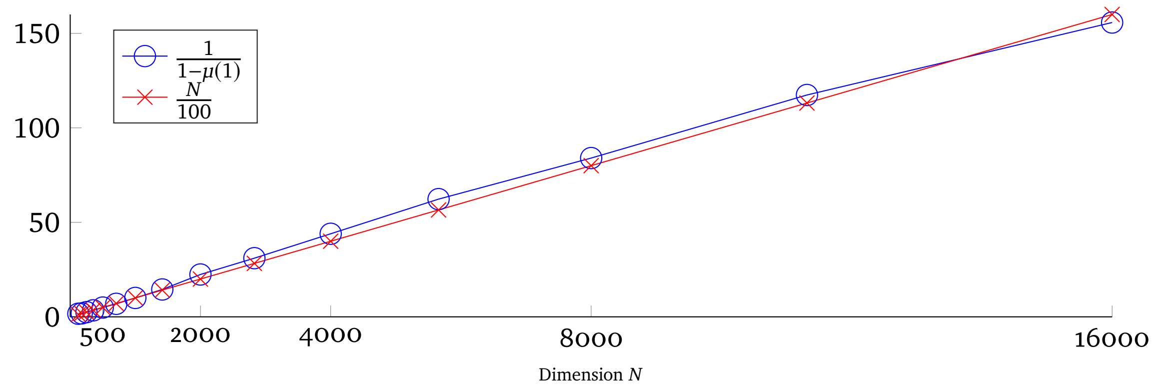


FIGURE 3. Comparison of $\frac{N}{100}$ and $\frac{1}{1-\mu(1)}$ (for which Proposition 5.3 provides un upper bound of $\frac{\rho_1 N}{c}$) for different dimension parameters N, where $\mu(1) = \frac{\mathcal{J}_{\epsilon_1}(x^1)-\|x_*\|_1}{\mathcal{J}_{\epsilon_0}(x^0)-\|x_*\|_1}$ is the linear convergence factor, for IRLS initialized from adversary initialization.

In our next experiment, we further investigate numerically the dimension dependence of the worst-case linear convergence factor $\mu(k) := \frac{\mathcal{J}_{\epsilon_k}(x^k)-\|x_*\|_1}{\mathcal{J}_{\epsilon_{k-1}}(x^{k-1})-\|x_*\|_1}$, which is upper bounded by the result of 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 3.1 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 problem 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 Proposition 5.3 is tight in its dependence on N, and that a dimension-less factor μ cannot be expected in general. In view of this, we believe that it is an interesting open problem to investigate the precise parameter dependence of μ in greater detail.

5. Proofs

In this section, we prove the main results of this paper, Theorem 3.2 and Theorem 3.3. To this end, we first state and prove the following technical lemma, which gives an upper and lower bound for $\mathcal{J}_\varepsilon(x) - \|x_*\|_1$, which is the quantity for which we are going to show linear convergence.

Lemma 5.1. *Let $x_*, x \in \mathbb{R}^N$. Assume that A fulfills the ℓ_1-NSP of order s with constant $\rho_s < 1$. Furthermore, suppose $Ax_* = Ax$ and that $\varepsilon \leq \frac{1}{N}\sigma_s(x)_{\ell_1}$. Then it holds that*

$$\frac{1-\rho_s}{1+\rho_s}\|x - x_*\|_1 - 2\sigma_s(x_*)_{\ell_1} \leq \mathcal{J}_\varepsilon(x) - \|x_*\|_1 \leq 3\sigma_s(x)_{\ell_1}. \tag{15}$$

In order to prove Lemma 5.1 we need the following technical lemma.

Lemma 5.2. *[DDFG10, Lemma 4.3] Assume that the matrix $A \in \mathbb{R}^{m\times N}$ has the ℓ_1-NSP holds for some s and $\rho_s < 1$. Then for all $z, x_* \in \mathbb{R}^N$ such that $Az = Ax_\star$ it holds that*

$$\|z - x_*\|_{\ell_1} \leqslant \frac{1+\rho_s}{1-\rho_s}\left(\|x_*\|_1 - \|z\|_1 + 2\sigma_s(z)_{\ell_1}\right).$$

Proof of Lemma 5.1. We observe that $\mathcal{J}_\varepsilon(x) \geq \|x\|_1$, which follows directly from the definition of $\mathcal{J}_\varepsilon(x)$, see Equation (1). Hence, we obtain that

$$\begin{aligned}
\mathcal{J}_\varepsilon(x) - \|x_*\|_1 &\geq \|x\|_1 - \|x_*\|_1 \\
&= \|x_{S^c}\|_1 + \|x_S\|_1 - \|x_*\|_1 \\
&\geq \|x_{S^c}\|_1 - \|(x - x_*)_S\|_1 - \|(x_*)_{S^c}\|_1 \\
&\geq \|(x - x_*)_{S^c}\|_1 - \|(x - x_*)_S\|_1 - 2\|(x_*)_{S^c}\|_1,
\end{aligned}$$

where in each of the last two inequalities we have applied the reverse triangle inequality. Since $x - x_*$ is contained in the null space of A, it follows from the nullspace property that $\|(x - x_*)_S\|_1 \leq \rho_s\|(x - x_*)_{S^c}\|_1$. Hence, we have shown that

$$\mathcal{J}_\varepsilon(x) - \|x_*\|_1 \geq (1-\rho_s)\|(x - x_*)_{S^c}\|_1 - 2\|(x_*)_{S^c}\|_1.$$

Since it follows from the null space property that $\|(x - x_*)_{S^c}\|_1 \geq \frac{\|x - x_*\|_1}{1+\rho_s}$, this shows the first inequality in (15).

Next, we are going to prove the reverse inequality in (15). For that, set $I := \{i \in [N] : |x_i^k| > \varepsilon_k\}$ and denote by S the set, which contains the s largest entries of x in absolute value. Then we observe that

$$\begin{aligned}\mathcal{J}_\varepsilon(x) - \|x_*\|_1 &= \|x_I\|_1 + \frac{1}{2}\sum_{i\in I^c}\left(\frac{x_i^2}{\varepsilon}+\varepsilon\right) - \|x_*\|_1 \\ &\le \|x_I\|_1 + |I^c|\varepsilon - \|x_*\|_1 \\ &\le \|x_I\|_1 + \sigma_s(x)_{\ell_1} - \|x_*\|_1 \\ &\le \|x\|_1 + \sigma_s(x)_{\ell_1} - \|x_*\|_1.\end{aligned} \tag{16}$$

In the third line we used the assumption $\varepsilon \le \frac{1}{N}\sigma_s(x)_{\ell_1}$. In order to proceed, we first derive an appropriate upper bound for $\|x\|_1 - \|x_*\|_1$. For that, we note

$$\begin{aligned}\left(\frac{1-\rho_s}{1+\rho_s}+1\right)(\|x\|_1 - \|x_*\|_1) &\le \frac{1-\rho_s}{1+\rho_s}\|x-x_*\|_1 - (\|x_*\|_1 - \|x\|_1) \\ &\le \left(\|x_*\|_1 - \|x\|_1 + 2\sigma_s(x)_{\ell_1}\right) - (\|x_*\|_1 - \|x\|_1) \\ &\le 2\sigma_s(x)_{\ell_1},\end{aligned}$$

where in the second line we have used Lemma 5.2. This shows that $\|x\|_1 - \|x_*\|_1 \le \frac{2\sigma_s(x)_{\ell_1}}{1+\frac{1-\rho_s}{1+\rho_s}}$. Combining this with (16), we obtain

$$\mathcal{J}_\varepsilon(x) - \|x_*\|_1 \le 3\sigma_s(x)_{\ell_1},$$

which finishes the proof of inequality (15). □

The next key proposition states that the quantity $\mathcal{J}_{\varepsilon^k}(x^k) - \|x_*\|_1$ decays linearly under appropriate conditions.

Proposition 5.3. *Let $x_* \in \mathbb{R}^N$ be an approximately s-sparse vector with support S. Let $A \in \mathbb{R}^{m\times N}$ and $y = Ax_*$. Assume that A fulfills the ℓ_1-NSP of order s with constant $\rho_s < 3/4$, if $\sigma_s(x_*)_{\ell_1} = 0$, and $\rho_s < 1/4$ otherwise. Moreover, assume that A has the ℓ_1-NSP of order 1 with constant $\rho_1 < 1$. Let the IRLS iterates $\{x^k\}_k$ and $\{\varepsilon_k\}_k$ be defined by (3) and (4) with initialization x^0. Then, for all $k \in \mathbb{N}$, such that $\|(x_*)_{S^c}\|_1 \le \frac{2}{9}\|(x_*)_{S^c} - x^\ell_{S^c}\|_1$ for all $\ell < k$, the following holds*

$$\mathcal{J}_{\varepsilon_k}(x^k) - \|x_*\|_1 \le \left(1 - \frac{c_{\rho_s}}{\rho_1 N}\right)^k\left(\mathcal{J}_{\varepsilon_0}(x^0) - \|x_*\|_1\right). \tag{17}$$

where the constant c_{ρ_s} is defined by

$$c_{\rho_s} := \begin{cases}\frac{(3/4-\rho_s)^2}{48} & \text{if } \sigma_s(x_*)_{\ell_1} = 0 \\ \frac{(1/4-\rho_s)^2}{48} & \text{else}\end{cases}$$

Before proving this statement, let us describe the main ideas of our proof. Recall that $x_\star$ has minimal $\|\cdot\|_1$-norm among all vectors x, which satisfy the constraint $Ax = y$. Hence, setting $v^k = x_\star - x^k$ due to convexity of the ℓ_1-norm we that $\|x^k + tv^k\|_1 < \|x^k\|_1$ for all $0 < t < 1$. Since that the quadratic functional $Q(\cdot, x^k)$ approximates the objective function $\mathcal{J}_\varepsilon$, which is a surrogate for the ℓ_1-norm, in a neighborhood of the current iterate x^k, we also expect that for $t > 0$ sufficiently small we have that $Q(x^k + tv^k, x^k) < Q(x^k, x^k)$. In order to show that the decrease is sufficiently large, we also need to show that t can be chosen large enough. This will guarantee a sufficient decrease of $\mathcal{J}_{\varepsilon^k}(x^k)$ in each iteration.

Proof of Proposition 5.3. In order to show inequality (17) we will prove by induction that for each k, such that $\|(x_*)_{S^c}\|_1 \leq \frac{2}{9}\|(x_*)_{S^c} - x^\ell_{S^c}\|_1$ for all $\ell \leq k$, it holds that

$$\mathcal{J}_{\varepsilon_{k+1}}(x^{k+1}) - \|x_*\|_1 \leq \left(1 - \frac{c_{\rho_s}}{\rho_1 N}\right)\left(\mathcal{J}_{\varepsilon_k}(x^k) - \|x_*\|_1\right).$$

Now choose such a $k \geq 1$ and assume that the statement has been shown for all $k' < k$. Set $v^k = x_* - x^k$. For $t \in \mathbb{R}$, we have, by optimality of x^{k+1} in (3), that

$$\mathcal{J}_{\varepsilon_{k+1}}(x^{k+1}) \leq Q_{\varepsilon_k}(x^{k+1}, x^k) \leq Q_{\varepsilon_k}(x^k + tv^k, x^k). \tag{18}$$

Moreover, by the definition of the quadratic objective $Q_{\varepsilon_k}(\cdot, x^k)$ (see (2)), it holds that

$$Q_{\varepsilon_k}(x^k + tv^k, x^k) - \mathcal{J}_{\varepsilon_k}(x^k) = t\,\langle \nabla \mathcal{J}_{\varepsilon_k}(x^k), v^k\rangle + \frac{t^2}{2}\langle v^k, \mathrm{diag}(w(x^k, \varepsilon_k))v^k\rangle. \tag{19}$$

Our goal is to show that by picking t large enough, we can make $Q_{\varepsilon_k}(x^k + tv^k, x^k) - \mathcal{J}_{\varepsilon_k}(x^k) < 0$ sufficiently small. For that, we now control $\langle \nabla \mathcal{J}_{\varepsilon_k}(x^k), v^k\rangle$ and $\langle v^k, \mathrm{diag}(w_{\varepsilon_k}(x^k))v^k\rangle$ separately.

Part 1: Bounding the linear term $\langle \nabla \mathcal{J}_{\varepsilon_k}(x^k), v^k\rangle$:

Let $I := \{i \in [N] : |x^k_i| > \varepsilon_k\}$ and denote by S the set which contains the s largest entries of $x_\star$ in absolute value. In the case that $x_\star$ is sparse, S is given by the support of x_*, i.e. $S = \mathrm{supp}\,(x_*)$. Consider

$$\langle \nabla \mathcal{J}_{\varepsilon_k}(x^k), v^k\rangle = \sum\nolimits_{i=1}^N \frac{x^k_i}{\max(|x^k_i|, \varepsilon_k)} v^k_i = \sum\nolimits_{i\in S} \frac{x^k_i}{\max(|x^k_i|, \varepsilon_k)} v^k_i + \sum\nolimits_{i\in S^c} \frac{x^k_i}{\max(|x^k_i|, \varepsilon_k)} v^k_i.$$

The first summand can be bounded by

$$\begin{aligned}
\sum\nolimits_{i\in S} \frac{x^k_i}{\max(|x^k_i|, \varepsilon_k)} v^k_i &= \sum\nolimits_{i\in S\cap I} \mathrm{sgn}(x^{(k)}_i) v^k_i + \sum\nolimits_{i\in S\cap I^c} \frac{x^k_i}{\varepsilon_k} v^k_i \\
&\leq \|v^k_{S\cap I}\|_1 + \|v^k_{S\cap I^c}\|_1 \\
&= \|v^k_S\|_1 \\
&\leq \rho_s \|v^k_{S^c}\|_1.
\end{aligned}$$

For the second summand we have that

$$\begin{aligned}
&\sum\nolimits_{i\in S^c} \frac{x^k_i}{\max(|x^k_i|, \varepsilon_k)} v^k_i \\
&= \sum\nolimits_{i\in S^c\cap I} \mathrm{sgn}\left(x^k_i\right) v^k_i + \sum\nolimits_{i\in S^c\cap I^c} \frac{x^k_i v^k_i}{\varepsilon_k} \\
&= \sum\nolimits_{i\in S^c\cap I} \mathrm{sgn}\left(x^k_i\right) (x_*)_i - \sum\nolimits_{i\in S^c\cap I} \mathrm{sgn}\left(x^k_i\right) x^k_i + \sum\nolimits_{i\in S^c\cap I^c} \frac{x^k_i\,(x_*)_i}{\varepsilon_k} - \sum\nolimits_{i\in S^c\cap I^c} \frac{(x^k_i)^2}{\varepsilon_k} \\
&\leq \|(x_*)_{S^c\cap I}\|_1 - \|x^k_{S^c\cap I}\|_1 + \|(x_*)_{S^c\cap I^c}\|_1 - \frac{\|x^k_{S^c\cap I^c}\|_2^2}{\varepsilon_k} \\
&= -\|x^k_{S^c\cap I}\|_1 + \|(x_*)_{S^c}\|_1 - \frac{\|x^k_{S^c\cap I^c}\|_2^2}{\varepsilon_k}
\end{aligned}$$

$$
\begin{aligned}
&=\| (x_*)_{S^c} \|_1 - \|x^k_{S^c}\|_1 + \|x^k_{S^c\cap I^c}\|_1 - \frac{\|x^k_{S^c\cap I^c}\|_2^2}{\varepsilon_k} \\
&\leq 2\| (x_*)_{S^c} \|_1 - \|v^k_{S^c}\|_1 + \|x^k_{S^c\cap I^c}\|_1 - \frac{\|x^k_{S^c\cap I^c}\|_2^2}{\varepsilon_k}.
\end{aligned}
$$

In order to proceed, we note that from the elementary inequality $ab \leq \frac{1}{2}\left(a^2+b^2\right)$ and from $\|x^k_{S^c\cap I^c}\|_1 \leq \sqrt{N}\|x^k_{S^c\cap I^c}\|_2$, it follows that

$$
\|x^k_{S^c\cap I^c}\|_1 \leq \frac{1}{2}\left(\frac{\varepsilon_k\|x^k_{S^c\cap I^c}\|_1^2}{2\|x^k_{S^c\cap I^c}\|_2^2} + 2\frac{\|x^k_{S^c\cap I^c}\|_2^2}{\varepsilon_k}.\right) \leq \frac{\varepsilon_k N}{4} + \frac{\|x^k_{S^c\cap I^c}\|_2^2}{\varepsilon_k}.
$$

Hence, using that $\varepsilon_k \leq \sigma_s(x^k)_{\ell_1}/N$, we have shown that

$$
\begin{aligned}
\sum\nolimits_{i\in S^c} \frac{x_i^k}{\max(|x_i^k|,\varepsilon_k)} v_i^k &\leq 2\| (x_*)_{S^c} \|_1 - \|v^k_{S^c}\|_1 + \frac{\varepsilon_k N}{4} \\
&= 2\| (x_*)_{S^c} \|_1 - \|v^k_{S^c}\|_1 + \frac{\sigma_s(x^k)_{\ell_1}}{4} \\
&\leq 2\| (x_*)_{S^c} \|_1 - \|v^k_{S^c}\|_1 + \frac{\|x^k_{S^c}\|_1}{4} \\
&\leq 2\| (x_*)_{S^c} \|_1 - \|v^k_{S^c}\|_1 + \frac{\|v^k_{S^c}\|_1}{4} + \frac{\|(x_*)_{S^c}\|_1}{4} \\
&= \frac{9}{4}\| (x_*)_{S^c} \|_1 - \frac{3}{4}\|v^k_{S^c}\|_1,
\end{aligned}
$$

where we used the triangular inequality for the vector $v^k = x^k - x_*$ on the set S^c and the fact that $\sigma_s(x^k)_{\ell_1} \leq \|x^k_{S^c}\|_1$. Hence, by adding up terms we obtain that

$$
\langle \nabla \mathcal{J}_{\varepsilon_k}(x^k), v^k\rangle \leq \frac{9}{4}\| (x_*)_{S^c} \|_1 - \left(\frac{3}{4} - \rho_s\right)\|v^k_{S^c}\|_1 \leq -(\beta-\rho_s)\|v^k_{S^c}\|_1.
$$

Here, we have set $\beta = 3/4$ in the case that $\sigma_s\left(x_*\right)_{\ell_1} = 0$ and $\beta = 1/4$ else. Moreover, we used the assumption $\| (x_*)_{S^c} \|_1 \leq \frac{2}{9}\|v^k_{S^c}\|_1$.

Part II: Bounding the quadratic term $\langle v^k, \mathrm{diag}(w_{\varepsilon_k}(x^k))v^k\rangle$
In order bound the quadratic term in (19) we first decompose it into two parts

$$
\langle v^k, \mathrm{diag}(w_{\varepsilon_k}(x^k))v^k\rangle = \sum\nolimits_{i=1}^N \frac{(v_i^k)^2}{\max(|x_i^k|,\varepsilon_k)} = \sum\nolimits_{i\in S} \frac{(v_i^k)^2}{\max(|x_i^k|,\varepsilon_k)} + \sum\nolimits_{i\in S^c} \frac{(v_i^k)^2}{\max(|x_i^k|,\varepsilon_k)}. \tag{20}
$$

For the first summand, we note that

$$
\sum\nolimits_{i\in S} \frac{(v_i^k)^2}{\max(|x_i^k|,\varepsilon_k)} \leq \frac{\|v^k_S\|_1\|v^k_S\|_\infty}{\varepsilon_k} \leq \rho_s \frac{\|v^k_{S^c}\|_1\|v^k\|_\infty}{\varepsilon_k} \leq \frac{\|v^k_{S^c}\|_1\|v^k\|_\infty}{\varepsilon_k}. \tag{21}
$$

For the second summand, it holds that

$$
\sum\nolimits_{i\in S^c} \frac{\left(v_i^k\right)^2}{\max(|x_i^k|,\varepsilon_k)} \leq \frac{\|v^k_{S^c}\|_\infty\|v^k_{S^c}\|_1}{\varepsilon_k} \leq \frac{\|v^k_{S^c}\|_1\|v^k\|_\infty}{\varepsilon_k}, \tag{22}
$$

Hence, by adding (21) and (22) up, it follows that

$$\langle v^k, \operatorname{diag}(w_{\varepsilon_k}(x^k))v^k\rangle \leq 2\frac{\|v^k_{S^c}\|_1 \|v^k\|_\infty}{\varepsilon_k}.$$

Next, we note that

$$\|v^k\|_\infty \leq \rho_1 \|v^k\|_1 \leq \rho_1 (1+\rho_s) \|v^k_{S^c}\|_1 \leq 2\rho_1 \|v^k_{S^c}\|_1.$$

Hence, we have shown that

$$\langle v^k, \operatorname{diag}(w_{\varepsilon_k}(x^k))v^k\rangle \leq 4\rho_1\frac{\|v^k_{S^c}\|_1^2}{\varepsilon_k}.$$

Part III: Combining the bounds to obtain decrease in k-th step:

Inserting the bounds of Part I and Part II into (19) we obtain

$$Q_{\varepsilon_k}(x^k + tv^k, x^k) - \mathcal{J}_{\varepsilon_k}(x^k) \leq -tb + t^2 a =: h(t), \tag{23}$$

where the function $h : \mathbb{R} \to \mathbb{R}$ is a quadratic polynomial with coefficients $b = (\beta - \rho_s) \|v^k_{S^c}\|_1$ and $a = 4\rho_1 \frac{\|v^k_{S^c}\|_1^2}{\varepsilon_k}$. We observe that the minimizer of h is given by $t = \frac{b}{2a}$. Inserting this into h, we obtain that

$$h\left(\frac{b}{2a}\right) = -\frac{b^2}{4a} = -\frac{(\beta-\rho_s)^2 \|v^k_{S^c}\|_1^2 \varepsilon_k}{16\rho_1 \|v^k_{S^c}\|_1^2} = -\frac{(\beta-\rho_s)^2}{16\rho_1}\varepsilon_k. \tag{24}$$

Combining this with (6), we obtain, for $t = \frac{b}{2a}$,

$$\begin{aligned}\mathcal{J}_{\varepsilon_{k+1}}(x^{k+1}) - \mathcal{J}_{\varepsilon_k}(x^k) &\leq Q_{\varepsilon_k}(x^{k+1}, x^k) - \mathcal{J}_{\varepsilon_k}(x^k)\\ &\leq Q_{\varepsilon_k}(x^k + tv^k, x^k) - \mathcal{J}_{\varepsilon_k}(x^k) \leq -\frac{(\beta-\rho_s)^2}{16\rho_1}\varepsilon_k.\end{aligned}$$

Hence, by rearranging terms it follows that

$$\mathcal{J}_{\varepsilon_{k+1}}(x^{k+1}) - \|x_*\|_1 \leq \mathcal{J}_{\varepsilon_k}(x^k) - \|x_*\|_1 - \frac{(\beta-\rho_s)^2}{16\rho_1}\varepsilon_k. \tag{25}$$

In order to proceed, we need to bound ε_k from below. For that, we note that

$$\varepsilon_k = \min\left(\varepsilon_{k-1}, \frac{\sigma_s(x^k)_{\ell_1}}{N}\right) = \frac{\sigma_s\left(x^\ell\right)_{\ell_1}}{N}$$

for some $\ell \leq k$. By Lemma 5.1, we have the following inequality chain

$$N\varepsilon_k = \sigma_s\left(x^\ell\right)_{\ell_1} \geq \frac{1}{3}\left(\mathcal{J}_{\varepsilon^\ell}\left(x^\ell\right) - \|x_*\|_1\right) \geq \frac{1}{3}\left(\mathcal{J}_{\varepsilon^k}\left(x^k\right) - \|x_*\|_1\right),$$

where in the second inequality we have used that, by induction, $\mathcal{J}_{\varepsilon_k}\left(x^k\right) \leq \mathcal{J}_{\varepsilon^\ell}\left(x^\ell\right)$. Plugging this into (25) leads to

$$\mathcal{J}_{\varepsilon_{k+1}}(x^{k+1}) - \|x_*\|_1 \leq \left(1 - \frac{(\beta-\rho_s)^2}{48\rho_1 N}\right)\left(\mathcal{J}_{\varepsilon_k}(x^k) - \|x_*\|_1\right).$$

This finishes the induction step and concludes the proof of Proposition 5.3. □

From Proposition 5.3 we can deduce Theorem 3.2, the first main result of this manuscript.

Proof of Theorem 3.2 . Recall that by Proposition 5.3 we have for all $k \in \mathbb{N}$ that

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

with a constant $c_{\rho_s} = \frac{(3/4-\rho_s)^2}{48}$ and where S denotes the set, which contains the s largest entries of $x_\star$ in absolute value. . By our assumption $\rho_s < 1/2$ it follows that $c_{\rho_s} \geq 1/768$, which implies that inequality (9) holds.

By Lemma 5.1 we have that

$$\mathcal{J}_{\varepsilon_0}(x^0) - \|x_*\|_1 \leq 3\sigma_s \left(x^0\right)_{\ell_1} \leq 3\|x^0 - x_*\|_1.$$

Next, we note that, again by Lemma 5.1, it holds that

$$\frac{1-\rho_s}{1+\rho_s}\|x - x_*\|_1 \leq \mathcal{J}_{\varepsilon_k}(x^k) - \|x_*\|_1.$$

Combining the three inequalities in this proof together with the assumption $\rho_s \leq 1/2$ yields inequality (10), which finishes the proof. □

Next, we are going to prove the second main result in this manuscript, Theorem 3.3, which deals with the approximately sparse case.

Proof of Theorem 3.3. Recall that

$$\hat{k} := \min\left\{k \in \mathbb{N} : \ \sigma_s\left(x_*\right)_{\ell_1} > \frac{2}{9}\|\left(x_*\right)_{S^c} - x^k_{S^c}\|_1\right\}.$$

Recall that by Proposition 5.3 we have for $k \leq \hat{k}$

$$\mathcal{J}_{\varepsilon_k}(x^k) - \|x_*\|_1 \leq \left(1 - \frac{c_{\rho_s}}{\rho_1 N}\right)^k \left(\mathcal{J}_{\varepsilon_0}(x^0) - \|x_*\|_1\right) \tag{26}$$

with a constant $c_{\rho_s} = \frac{(1/4-\rho_s)^2}{48}$. Hence, by our assumption $\rho_s < 1/8$ we obtain $c_{\rho_s} \geq 1/3072$ and inequality (11) follows, which proves the first statement. In order to prove the second statement, let $\tilde{k}$ and k be natural numbers, such that $\tilde{k} \leq \hat{k}$ and $k \geq \tilde{k}$ holds. Then we obtain that

$$\begin{aligned}
\frac{1-\rho_s}{1+\rho_s}\|x^k - x_*\|_1 - 2\sigma_s\left(x_*\right)_{\ell_1} &\leq \mathcal{J}_{\varepsilon_k}(x^k) - \|x_*\|_1 \\
&\leq \mathcal{J}_{\varepsilon_{\tilde{k}}}(x^{\tilde{k}}) - \|x_*\|_1 \\
&\leq \left(1 - \frac{c_{\rho_s}}{\rho_1 N}\right)^{\tilde{k}} \left(\mathcal{J}_{\varepsilon_0}(x^0) - \|x_*\|_1\right) \\
&\leq 3\left(1 - \frac{c_{\rho_s}}{\rho_1 N}\right)^{\tilde{k}} \sigma_s\left(x^0\right)_{\ell_1},
\end{aligned}$$

where in the first inequality we applied Lemma 5.1. In the second inequality we used that the sequence $\left\{\mathcal{J}_{\varepsilon^\ell}\left(x^\ell\right)\right\}_\ell$ is monotonically decreasing and in the third inequality we used inequality

(26). In the fourth inequality we again used Lemma 5.1. By rearranging terms and using the assumption $\rho_s < 1/8$ it follows for all integers $\tilde{k}$ and k such that $\tilde{k} \le \hat{k}$ and $k \ge \tilde{k}$

$$\|x^k - x_*\|_1 \le 6\left(1 - \frac{c_{\rho_s}}{\rho_1 N}\right)^{\tilde{k}} \sigma_s\left(x^0\right)_{\ell_1} + 4\sigma_s\left(x_*\right)_{\ell_1}. \tag{27}$$

In order to proceed, recall that S denotes the support of the s largest entries of $x_\star$. Then we note that

$$\sigma_s\left(x^0\right)_{\ell_1} \le \|x^0_{S^c}\|_1 \le \|\left(x^0 - x_*\right)_{S^c}\|_1 + \|\left(x_*\right)_{S^c}\|_1 \le \|x^0 - x_*\|_1 + \sigma_s\left(x_*\right)_{\ell_1}. \tag{28}$$

Hence, we have shown that for all integers $\tilde{k}$ and k such that $\tilde{k} \le \hat{k}$ and $k \ge \tilde{k}$ it holds that

$$\|x^k - x_*\|_1 \le 6\left(1 - \frac{c_{\rho_s}}{\rho_1 N}\right)^{\tilde{k}} \|x^0 - x_*\|_1 + 10\sigma_s\left(x_*\right)_{\ell_1}. \tag{29}$$

By setting $k = \tilde{k}$, we observe that this implies inequality (12), which shows the second statement. In order to prove the third statement, we will distinguish two cases. For the first case, assume that $\hat{k} \ge \left\lceil \frac{\rho_1 N}{c_{\rho_s}} \log\left(\frac{\|x^0 - x_*\|_1}{\sigma_s(x_*)_{\ell_1}}\right)\right\rceil$. Then for $k \ge \tilde{k} := \left\lceil \frac{\rho_1 N}{c_{\rho_s}} \log\left(\frac{\|x^0 - x_*\|_1}{\sigma_s(x_*)_{\ell_1}}\right)\right\rceil$ it follows from inequality (29) that

$$\|x^k - x_*\|_1 \le 6\left(1 - \frac{c_{\rho_s}}{\rho_1 N}\right)^{\frac{\rho_1 N}{c_{\rho_s}} \log\left(\frac{\|x^0 - x_*\|_1}{\sigma_s(x_*)_{\ell_1}}\right)} \|x^0 - x_*\|_1 + 10\sigma_s\left(x_*\right)_{\ell_1} \le 20\sigma_s\left(x_*\right)_{\ell_1}.$$

where in the second inequality we have used the elementary inequality $\log(1+t) \le t$ for $t > -1$. This shows the third statement in the first case. To prove the second case, assume that $\hat{k} < \left\lceil \frac{\rho_1 N}{c_{\rho_s}} \log\left(\frac{\|x^0 - x_*\|_1}{\sigma_s(x_*)_{\ell_1}}\right)\right\rceil$. Then we can compute that

$$\begin{aligned}
\frac{1-\rho_s}{1+\rho_s}\|x^k - x_*\|_1 - 2\sigma_s\left(x_*\right)_{\ell_1} &\le \mathcal{J}_{\varepsilon_k}(x^k) - \|x_*\|_1 \\
&\le \mathcal{J}_{\varepsilon_{\hat{k}}}(x^{\hat{k}}) - \|x_*\|_1 \\
&\le 3\sigma_s\left(x^{\hat{k}}\right)_{\ell_1} \\
&\le 3\|x^{\hat{k}} - x_*\|_1 + 3\sigma_s\left(x_*\right)_{\ell_1} \\
&\le 3\left(1+\rho_s\right)\|\left(x^{\hat{k}} - x_*\right)_{S^c}\|_1 + 3\sigma_s\left(x_*\right)_{\ell_1} \\
&\le 20\sigma_s\left(x_*\right)_{\ell_1}.
\end{aligned}$$

In the first and third inequality we have used Lemma 5.1. In the second inequality we have used the monotonicity of the sequence $\left\{\mathcal{J}_{\varepsilon_k}\left(x^k\right)\right\}_k$. In the fourth inequality we have argued as in inequality (28) and in the fifth inequality we have used the null space property. In the last inequality we have used that by definition of $\hat{k}$ it holds that $\sigma_s\left(x_*\right)_{\ell_1} > \frac{2}{9}\|\left(x_*\right)_{S^c} - x^{\hat{k}}_{S^c}\|_1$. This shows that also in the second case the third statement holds, which finishes the proof. □

Conclusion

In this paper, we proved that the IRLS algorithm for ℓ_1-minimization converges linearly to the ground truth under a suitable null-space property, i.e., in situations when ℓ_1-minimization is equivalent to the recovery of the sparsest vector. Moreover, we have corroborated our theory with numerical experiments.

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* [ST03, DH18]. We leave it to future work to generalize our analysis to obtain global linear convergence rates for other variants of IRLS, for example in low-rank matrix recovery [FRW11, KS18].

Finally, it would be interesting to extend the techniques of this paper to achieve a better understanding of the IRLS method designed to optimize an ℓ_p-quasinorm (with $0 < p < 1$) [CY08, DDFG10, KS18]. In this non-convex case, it was observed that sparse vectors can be recovered from even fewer measurements, but a global convergence theory has remained elusive so far.

Acknowledgements

The authors want to thank Massimo Fornasier for inspiring discussions.

C.M.V. is supported by the German Science Foundation (DFG) within the Gottfried Wilhelm Leibniz Prize under Grant BO 1734/20-1, under contract number PO-1347/3-2, under Emmy Noether junior research group KR 4512/1-1 and within Germany's Excellence Strategy EXC-2111 390814868. D.S. is supported by the Air Force Office of Scientific Research under award #FA9550-18-1-00788.

Appendix A. Proof of Lemma 2.1

Proof. We prove each of the three statements separately.

(1) Let $x \in \mathbb{R}^N$. Then the i-th coordinate of $\operatorname{diag}(w_\varepsilon(x))x$ is given by

$$(\operatorname{diag}(w_\varepsilon(x))x)_i = \begin{cases} \frac{x_i}{|x_i|} = \operatorname{sgn}(x_i), & \text{if } |x| > \varepsilon, \\ \frac{x_i}{\varepsilon}, & \text{if } |x| \leq \varepsilon. \end{cases} = j'_\varepsilon(x_i) = (\nabla \mathcal{J}_\varepsilon(x))_i\,,$$

where $\mathcal{J}_\varepsilon(x)$ is the gradient of $\mathcal{J}_\varepsilon$ at x.

(2) This follows directly from the definition of $Q_\varepsilon(x,z)$ and by setting $x = z$.

(3) We define $I := \{i \in [N] : |x_i| > \varepsilon\}$ and write the difference $Q_\varepsilon(z,x) - \mathcal{J}_\varepsilon(z)$ as

$$\begin{aligned} Q_\varepsilon(z,x) - \mathcal{J}_\varepsilon(z) &= \frac{1}{2}\left(\langle z, \operatorname{diag}(w_\varepsilon(x))z\rangle - \langle x, \operatorname{diag}(w_\varepsilon(x))x\rangle\right) \\ &= \sum\nolimits_{i\in I}\left(\frac{1}{2}|x_i| + \frac{1}{2}\frac{z_i^2}{|x_i|} - j_\varepsilon(z_i)\right) + \sum\nolimits_{i\in I^c}\left(\frac{1}{2}\varepsilon + \frac{1}{2}\frac{z_i^2}{\varepsilon} - j_\varepsilon(z_i)\right) \end{aligned}$$

and show that each summand of the two sums is non-negative. In particular, if $i \in I$, then assume first that $|z_i| > \varepsilon$. Then

$$\frac{1}{2}|x_i| + \frac{1}{2}\frac{z_i^2}{|x_i|} - j_\varepsilon(z_i) = \frac{1}{2}\left(|x_i| + \frac{z_i^2}{|x_i|}\right) - |z_i| \geq |z_i| - |z_i| = 0$$

due to inequality $a \leq \frac{1}{2}(a^2/b + b)$, which holds for any $b > 0$.

On the other hand, if $|z_i| \leq \varepsilon$, then

$$\begin{aligned} \frac{1}{2}|x_i| + \frac{1}{2}\frac{z_i^2}{|x_i|} - j_\varepsilon(z_i) &= \frac{1}{2}|x_i| + \frac{1}{2}\frac{z_i^2}{|x_i|} - \frac{1}{2}\left(\frac{z_i^2}{\varepsilon} + \varepsilon\right) \\ &= \frac{1}{2}\left(|x_i| - \varepsilon\right) + \frac{1}{2}z_i^2\left(\frac{1}{|x_i|} - \frac{1}{\varepsilon}\right) \\ &\geq \frac{1}{2}\left(|x_i| - \varepsilon\right) + \frac{1}{2}\varepsilon^2\left(\frac{1}{|x_i|} - \frac{1}{\varepsilon}\right) \\ &= \frac{1}{2}\left(|x_i| + \frac{\varepsilon^2}{|x_i|}\right) - \varepsilon \geq \varepsilon - \varepsilon = 0, \end{aligned} \tag{30}$$

where we used that $\frac{1}{|x_i|} - \frac{1}{\varepsilon} < 0$ in the first inequality. In the second inequality, we again used $a \leq \frac{1}{2}(a^2/b + b)$ for any $b > 0$. Now let $i \in I^c$. We again consider the two cases, $|z_i| \leq \varepsilon$ and $|z_i| > \varepsilon$. In the first case we have that $\frac{1}{2}\varepsilon + \frac{1}{2}\frac{z_i^2}{\varepsilon} - j_\varepsilon(z_i) = 0$, and in the second case we have that

$$\frac{1}{2}\varepsilon + \frac{1}{2}\frac{z_i^2}{\varepsilon} - j_\varepsilon(z_i) = \frac{1}{2}\varepsilon + \frac{1}{2}\frac{z_i^2}{\varepsilon} - |z_i| \geq |z_i| - |z_i| = 0,$$

which concludes the proof. □

Appendix B. Proof of Proposition 4.1

Proof. Let $j \in S^c$, where S is the support set of x_*. Then

$$|x_j^k| \leq \sum\nolimits_{i \in S^c} |x_i^k| < \min_{i \in S} |(x_*)_i| - \sum\nolimits_{i \in S} |x_i^k - (x_*)_i|$$

using the assumption that $\sum_{i \in S^c} |(x^k)_i| + \sum_{i \in S} |x_i^k - (x_*)_i| = \|x^k - x_*\|_1 < \min_{i \in S} |(x_*)_i|$.

On the other hand, for $j \in S$, we can estimate that

$$|x_j^k| = |x_j^k - (x_*)_j + (x_*)_j| \geq |(x_*)_j| - |x_j^k - (x_*)_j| \geq \min_{i \in S} |(x_*)_i| - \sum\nolimits_{i \in S} |x_i^k - (x_*)_i|.$$

Taking the previous two inequalities together, we conclude that $\max_{j \in S^c} |x_j^k| < \min_{j \in S} |x_j^k|$, which finishes the proof. □

References

[ABH19] 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.

[ABS13] 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*, 137:91–129, 2013.

[AIG06] M. Allain, J. Idier, and Y. Goussard. On global and local convergence of half-quadratic algorithms. *IEEE Trans. Image Process.*, 15(5):1130–1142, 2006.

[APS19] Deeksha Adil, Richard Peng, and Sushant Sachdeva. Fast, provably convergent irls algorithm for p-norm linear regression. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pages 14189–14200, 2019.

[BDDW08] Richard Baraniuk, Mark Davenport, Ronald DeVore, and Michael Wakin. A simple proof of the restricted isometry property for random matrices. *Constr. Approx.*, 28:253–263, 2008.

[BDL07] Jérôme Bolte, Aris Daniilidis, and Adrian Lewis. The łojasiewicz inequality for non-smooth subanalytic functions with applications to subgradient dynamical systems. *SIAM J. Optim.*, 17(4):1205–1223, 2007.

[BPC+11] 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*, 3:1–122, 2011.

[BS15] A. Beck and S. Sabach. Weiszfeld's method: Old and new results. *J. Optim. Theory Appl.*, 164:1–40, 2015.

[Bur12] C. S. Burrus. Iterative reweighted least squares. *OpenStax CNX: Available online: http://cnx.org/contents/92b90377-2b34-49e4-b26f-7fe572db78a1*, 2012.

[CCW16] Jameson Cahill, Xuemei Chen, and Rongrong Wang. The gap between the null space property and the restricted isometry property. *Linear Algebra Appl.*, 501:363–375, 2016.

[CD94] Shaobing Scott Chen and David Donoho. Basis pursuit. In *Proc. of 1994 28th Asilomar Conference on Signals, Systems and Computers*, pages 41–44, 1994.

[CDD09] Albert Cohen, Wolfgang Dahmen, and Ronald DeVore. Compressed sensing and best k-term approximation. *J. Amer. Math. Soc.*, 22:211–231, 2009.

[CDS01] Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM Rev.*, 43(1):129–159, 2001.

[CM73] J. F. Claerbout and F. Muir. Robust modeling with erratic data. *Geophysics*, 38(5):826–844, 1973.

[CRT06] E. J. Candès, J.K. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.*, 59(8):1207–1223, 2006.

[CT06] E. J. Candès and T. Tao. Near-Optimal Signal Recovery From Random Projections: Universal Encoding Strategies? *IEEE Trans. Inf. Theory*, 52(12):5406–5425, 2006.

[CW05] Patrick L Combettes and Valérie R Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Model. Simul.*, 4(2):1168–1200, 2005.

[CY08] R. Chartrand and W. T. Yin. Iteratively reweighted algorithms for compressive sensing. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3869–3872, 2008.

[CY16] Rick Chartrand and Wotao Yin. Nonconvex sparse regularization and splitting algorithms. In *Splitting Methods in Communication, Imaging, Science, and Engineering*, pages 237–249. Springer International Publishing, 2016.

[DDFG10] I. Daubechies, R. DeVore, M. Fornasier, and C.S. Güntürk. Iteratively reweighted least squares minimization for sparse recovery. *Comm. Pure Appl. Math.*, 63:1–38, 2010.

[DFRW20] L. Diening, M. Fornasier, Tomasi R., and M Wank. A Relaxed Kačanov iteration for the p-poisson problem. *Numer. Math.*, 145(2):1–34, 2020.

[DH18] D. Dadush and S. Huiberts. A friendly smoothed analysis of the simplex method. *SIAM J. Comput.*, 49(5):STOC18–449–STOC18–499, 2018.

[DL92] David L Donoho and Benjamin F Logan. Signal recovery and the large sieve. *SIAM J. Appl. Math.*, 52(2):577–591, 1992.

[DLR18] S. Dirksen, G. Lecué, and H. Rauhut. On the gap between restricted isometry properties and sparse recovery conditions. *IEEE Trans. Inform. Theory*, 64(8):5478–5487, Aug 2018.

[DMA97] G. Davis, S. Mallat, and M. Avellanedas. Adaptive greedy approx-imations. *Constr. Approx.*, 13(1):57–98, 1997.

[DT08] David Donoho and Yaakov Tsaig. Fast solution of ℓ_1-norm minimization problems when the solution may be sparse. *IEEE Trans. Inform. Theory*, 54:4789–4812, 2008.

[FPRW16] Massimo Fornasier, Steffen Peter, Holger Rauhut, and Stephan Worm. Conjugate gradient acceleration of iteratively re-weighted least squares methods. *Comput. Optim. Appl.*, 65(1):205–259, 2016.

[FR13] S. Foucart and H. Rauhut. *A Mathematical Introduction to Compressive Sensing*. Applied and Numerical Harmonic Analysis. Springer New York, 2013.

[FRW11] M. Fornasier, H. Rauhut, and R. Ward. Low-rank matrix recovery via iteratively reweighted least squares minimization. *SIAM J. Optim.*, 21:1614–1640, 2011.

[GN03] Remi Gribonval and Morten Nielsen. Sparse representation in the union of basis. *IEEE Trans. Inform. Theory*, 49:3320–3325, 2003.

[GR97] 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(3):600–616, 1997.

[Hub64] Peter J. Huber. Robust estimation of a location parameter. *Ann. Math. Statist.*, 35(1):73–101, 1964.

[HW77] P. W. Holland and R. E. Welsch. Robust regression using iteratively reweighted least-squares. *Communications in Statistics - Theory and Methods*, 6:813–827, 1977.

[Idi01] J. Idier. Convex half-quadratic criteria and interacting auxiliary variables for image restoration. *IEEE Trans. Image Process.*, 10(7):1001–1009, 2001.

[KS18] C. Kümmerle and J. Sigl. Harmonic mean iteratively reweighted least squares for low-rank matrix recovery. *J. Mach. Learn. Res.*, 19:1815–1863, 2018.

[KV20] C. Kümmerle and C. M. Verdun. Escaping saddle points in ill-conditioned matrix completion with a scalable second order method. In *Workshop on "Beyond first-order methods in ML systems" at the 37th International Conference on Machine Learning, Vienna, Austria*, 2020.

[LS98] W. Li and J. Swetits. The linear l1 estimator and the huber m-estimator. *SIAM J. Optim.*, 8(2):457–475, 1998.

[LST18] X. Li, D. Sun, and K.-C. Toh. A highly efficient semismooth newton augmented lagrangian method for solving lasso problems. *SIAM J. Optim.*, 28(1):433–458, 2018.

[LYW13] M.-J. Lai, X. Yangyang, and Y. Wotao. Improved iteratively reweighted least squares for unconstrained smoothed ℓ_q minimization. *SIAM J. Numer. Anal.*, 2(51):927–957, 2013.

[Mey20] Gregory P. Meyer. An Alternative Probabilistic Interpretation of the Huber Loss. *arXiv preprint arXiv:1911.02088*, 2020.

[MF12] K. Mohan and M. Fazel. Iterative reweighted algorithms for matrix rank minimization. *J. Mach. Learn. Res.*, 13:3441–3473, 2012.

[MGJK19] Bhaskar Mukhoty, Govind Gopakumar, Prateek Jain, and Purushottam Kar. Globally-convergent iteratively reweighted least squares for robust regression problems. In *Proceedings of Machine Learning Research*, volume 89, pages 313–322, 2019.

[ML17] Shahar Mendelson and Guillaume Lecué. Sparse recovery under weak moment assumptions. *J. Eur. Math. Soc.*, 19(3):881–904, 2017.

[Nat96] Balas K. Natarajan. Sparse approximate solutions to linear systems. *SIAM J. Comput.*, 24:227–234, 1996.

[ODBP15] 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(1):331–372, 2015.

[RKD99] B. D. Rao and K. Kreutz-Delgado. An affine scaling methodology for best basis selection. *IEEE Trans. Signal Process.*, 47(1):187–200, 1999.

[RV08] M. Rudelson and R. Vershynin. On sparse reconstruction from fourier and gaussian measurements. *Comm. Pure Appl. Math.*, 61(8):1025–1045, 2008.

[SBP17] Y. Sun, P. Babu, and D. P. Palomar. Majorization-minimization algorithms in signal processing, communications, and machine learning. *IEEE Trans. Signal Process.*, 65(3):794–816, 2017.

[ST03] D. A. Spielman and S.-H. Teng. Smoothed analysis of termination of linear programming algorithms. *Math. Program., Ser. B*, 97:375–404, 2003.

[TBM79] Howard L Taylor, Stephen C Banks, and John F McCoy. Deconvolution with the ℓ_1 norm. *Geophysics*, 44(1):39–52, 1979.

[VD17] Sergey Voronin and Ingrid Daubechies. An iteratively reweighted least squares algorithm for sparse regularization. In *Functional Analysis, Harmonic Analysis, and Image Processing: A Collection of Papers in Honor of Björn Jawerth*, pages 391–441. American Mathematical Society, 2017.

[Wei37] E. Weiszfeld. Sur le point pour lequel la somme des distances de n points donnés est minimum. *Tohoku Mathematical Journal, First Series*, 43:355–386, 1937.

[YOGD08] 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(1):143–168, 2008.