frontiers in Applied Mathematics and Statistics

ORIGINAL RESEARCH
published: 07 June 2021
doi: 10.3389/fams.2021.615573

Efficient Tuning-Free l_1-Regression of Nonnegative Compressible Signals

Hendrik Bernd Petersen[1]*, *Bubacarr Bah*[2,3] *and Peter Jung*[1]

[1]Communications and Information Theory Group, Technische Universtität Berlin, Berlin, Germany, [2]African Institute for Mathematical Sciences, Cape Town, South Africa, [3]Division of Applied Mathematics, Stellenbosch University, Stellenbosch, South Africa

In compressed sensing the goal is to recover a signal from as few as possible noisy, linear measurements with the general assumption that the signal has only a few non-zero entries. The recovery can be performed by multiple different decoders, however most of them rely on some tuning. Given an estimate for the noise level a common convex approach to recover the signal is basis pursuit denoising. If the measurement matrix has the robust null space property with respect to the ℓ_2-norm, basis pursuit denoising obeys stable and robust recovery guarantees. In the case of unknown noise levels, nonnegative least squares recovers non-negative signals if the measurement matrix fulfills an additional property (sometimes called the M^+-criterion). However, if the measurement matrix is the biadjacency matrix of a random left regular bipartite graph it obeys with a high probability the null space property with respect to the ℓ_1-norm with optimal parameters. Therefore, we discuss non-negative least absolute deviation (NNLAD), which is free of tuning parameters. For these measurement matrices, we prove a uniform, stable and robust recovery guarantee. Such guarantees are important, since binary expander matrices are sparse and thus allow for fast sketching and recovery. We will further present a method to solve the NNLAD numerically and show that this is comparable to state of the art methods. Lastly, we explain how the NNLAD can be used for viral detection in the recent COVID-19 crisis.

Keywords: compressed sensing, compressible, sparse, non-negative, regression, tuning-free, expander, uniform

OPEN ACCESS

Edited by:
Qiyu Sun,
University of Central Florida,
United States

Reviewed by:
Guohui Song,
Old Dominion University,
United States
Yunlong Feng,
University at Albany,
United States

***Correspondence:**
Hendrik Bernd Petersen
petersen@tu-berlin.de

Specialty section:
This article was submitted to Mathematics of Computation and Data Science, a section of the journal Frontiers in Applied Mathematics and Statistics

Received: *09 October 2020*
Accepted: *16 April 2021*
Published: *07 June 2021*

Citation:
Petersen HB, Bah B and Jung P (2021) Efficient Tuning-Free l_1-Regression of Nonnegative Compressible Signals. Front. Appl. Math. Stat. 7:615573. doi: 10.3389/fams.2021.615573

1 INTRODUCTION

Since it has been realized that many signals admit a sparse representation in some frames, the question arose whether or not such signals can be recovered from less samples than the dimension of the domain by utilizing the low dimensional structure of the signal. The question was already answered positively in the beginning of the millennium [1, 2]. By now there are multiple different decoders to recover a sparse signal from noisy measurements with robust recovery guarantees. Most of them however rely on some form of tuning, depending on either the signal or the noise.

The basis pursuit denoising requires an upper bound on the norm of the noise ([3], Theorem 4.22), the least shrinkage and selection operator an estimate on the ℓ_1-norm of the signal ([4], Theorem 11.1) and the Lagrangian version of least shrinkage and selection operator allegedly needs to be tuned to the order of the the noise level ([4], Theorem 11.1). The expander iterative hard thresholding needs the sparsity of the signal or an estimate of the order of the expansion property ([3], Theorem 13.15). The order of the expansion property can be calculated from the measurement matrix, however there is no polynomial time method known to do this. Variants of these methods

have similar drawbacks. The non-negative basis pursuit denoising requires the same tuning parameter as the basis pursuit denoising [5]. Other thresholding based decoders like sparse matching pursuit and expander matching pursuit have the same limitations as the expander iterative hard thresholding [6].

If these side information is not known a priori, many decoders yield either no recovery guarantees or, in their imperfectly tuned versions, yield sub-optimal estimation errors ([3], Theorem 11.12). Even though the problem of sparse recovery from under-sampled measurements has been answered long ago, finding tuning free decoders that achieve robust recovery guarantees is still a topic of interest.

The most prominent achievement for that is the non-negative least squares (NNLS) [7–11]. It is completely tuning free [12] and in [13, 14] it was proven that it achieves robust recovery guarantees if the measurement matrix consists of independent biased sub-Gaussian random variables.

1.1 Our Contribution

We will replace the least squares in the NNLS with an arbitrary norm and obtain the non-negative least residual (NNLR). We use the methods of [13] to prove recovery guarantees under similar conditions as the NNLS. In particular, we consider the case where we minimize the ℓ_1-norm of the residual (NNLAD) and give a recovery guarantee if the measurement matrix is a random walk matrix of a uniformly at random drawn D-left regular bipartite graph.

In general, our results state that if the M^+ criterion is fulfilled, the basis pursuit denoising can be replaced by the tuning-less NNLR for non-negative signals. Note that the M^+criterion can be fulfilled by adding only one explicitly chosen measurement if that is possible in the application. Thus, in practice the NNLR does not require more measurements than the BPDN to recover sparse signals. While biased sub-Gaussian measurement matrices rely on a probabilistic argument to verify that such a measurement is present, random walk matrices of left regular graphs naturally have such a measurement. The tuning-less nature gives the NNLR an advantage over other decoders if the noise power can not be estimated, which is for instance the case when we have multiplicative noise or the measurements are Poisson distributed. Note that Laplacian distributed noise or the existence of outliers also favors an ℓ_1 regression approach over an ℓ_2 regression approach and thus motivate to use the NNLAD over the NNLS.

Further, the sparse structure of left regular graphs can reduce the encoding and decoding time to a fraction. Using [15] we can solve the NNLAD with a first order method of a single optimization problem with a sparse measurement matrix. Other state of the art decoders often use non-convex optimization, computationally complex projections or need to solve multiple different optimization problems. For instance, to solve the basis pursuit denoising given a tuning parameter a common approach is to solve a sequence of ℓ_1-constrained least residual[1] problems to approximate where the Pareto curve attains the value of the tuning parameter of basis pursuit denoising [16]. Cross-validation techniques suffer from similar issues [17].

1.2 Relations to Other Works

We build on the theory of [13] that uses the ℓ_2 null space property and the M^+ criterion. These methods have also been used in [12, 14]. To the best of the authors knowledge the M^+ criterion has not been used with an ℓ_1 null space property before. Other works have used adjacency matrices of graphs as measurements matrices including [6, 18–21]. The works [18, 19] did not consider noisy observations. The decoder in [20] is the basis pursuit denoising and thus requires tuning depending on the noise power. [21] proposes two decoders for non-negative signals. The first is the non-negative basis pursuit which could be extended to the non-negative basis pursuit denoising. However, this again needs a tuning parameter depending on the noise power. The second decoder, the Reverse Expansion Recovery algorithm, requires the order of the expansion property, which is not known to be calculatable in a polynomial time. The survey [6] contains multiple decoders including the basis pursuit, which again needs tuning depending on the noise power for robustness, the expander matching pursuit and the sparse matching pursuit, which need the order of the expansion property. Further, [5] considered sparse regression of non-negative signals and also used the non-negative basis pursuit denoising as decoder, which again needs tuning dependent on the noise power. To the best of the authors knowledge, this is the first work that considers tuning-less sparse recovery for random walk matrices of left regular bipartite graphs. The NNLAD has been considered in [22] with a structured sparsity model without the use of the M^+ criterion.

2 PRELIMINARIES

For $K \in \mathbb{N}$ we denote the set of integers from 1 to K by $[K]$. For a set $T \subset [N]$ we denote the number of elements in T by $\#(T)$. Vectors are denoted by lower case bold face symbols, while its corresponding components are denoted by lower case italic letters. Matrices are denoted by upper case bold face symbols, while its corresponding components are denoted by upper case italic letters. For $\mathbf{x} \in \mathbb{R}^N$ we denote its ℓ_p-norms by $||\mathbf{x}||_p$. Given $\mathbf{A} \in \mathbb{R}^{M\times N}$ we denote its operator norm as operator from ℓ_q to ℓ_p by $||\mathbf{A}||_{q\to p} := \sup_{\mathbf{v}\in\mathbb{R}^N, ||\mathbf{v}||_q \le 1} ||\mathbf{A}\mathbf{v}||_p$. By $\mathbb{R}^N_+$ we denote the non-negative orthant. Given a closed convex set $C \subset \mathbb{R}^N$, we denote the projection onto C, i.e., the unique minimizer of $\operatorname{argmin}_{\mathbf{z}\in C} 1/2||\mathbf{z}-\mathbf{v}||_2^2$, by $\mathcal{P}_C(\mathbf{v})$. For a vector $\mathbf{x} \in \mathbb{R}^N$ and a set $T \subset [N]$, $\mathbf{x}|_T$ denotes the vector in $\mathbb{R}^N$, whose nth component is x_n if $n \in T$ and 0 else. Given $N, S \in \mathbb{N}$ we will often need sets $T \subset [N]$ with $\#(T) \le S$ and we abbreviate this by $\#(T) \le S$ if no confusion is possible.

Given a measurement matrix $\mathbf{A} \in \mathbb{R}^{M\times N}$ a decoder is any map $Q_A : \mathbb{R}^M \to \mathbb{R}^N$. We refer to $\mathbf{x} \in \mathbb{R}^N$ as signal. If $\mathbf{x} \in \mathbb{R}^N_+ = \{\mathbf{z} \in \mathbb{R}^N : z_n \ge 0 \text{ for all } n \in [N]\}$, we say the signal is non-negative and write shortly $\mathbf{x} \ge 0$. If additionally $x_n > 0$ for all $n \in [N]$, we write $\mathbf{x} > 0$. An input of a decoder, i.e., any $\mathbf{y} \in \mathbb{R}^M$, is refered to as observation. We allow all possible inputs of the

[1]The ℓ_1-constrained least residual is given by $\operatorname{argmin}_{\|z\|_1 \le \tau} \|Az - y\|$ for some norm $\|\cdot\|$.

decoder as observation, since in general the transmitted codeword $\mathbf{Ax}$ is disturbed by some noise. Thus, given a signal $\mathbf{x}$ and an observation $\mathbf{y}$ we call $\mathbf{e} := \mathbf{y} - \mathbf{Ax}$ the noise. A signal $\mathbf{x}$ is called S-sparse if $||\mathbf{x}||_p := \#(\{n \in [N] : x_n \neq 0\}) \leq S$. We denote the set of S-sparse vectors by

$$\Sigma_S := \{\mathbf{z} \in \mathbb{R}^N : \|\mathbf{z}\|_0 \leq S\}.$$

Given some $S \in [N]$ the compressibility of a signal $\mathbf{x}$ can be measured by $d_1(\mathbf{x}, \Sigma_S) := \inf_{\mathbf{z} \in \Sigma_S} ||\mathbf{x} - \mathbf{z}||_1$.Given N and S, the general non-negative compressed sensing task is to find a measurement matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$ and a decoder $Q_A : \mathbb{R}^M \to \mathbb{R}^N$ with M as small as possible such that the following holds true: There exists a $q \in [1, \infty]$ and a continuous function $C : \mathbb{R} \times \mathbb{R}^M \to \mathbb{R}_+$ with $C(0,0) = 0$ such that

$$\left\|Q_A(\mathbf{y}) - \mathbf{x}\right\|_q \leq C\left(d_1(\mathbf{x}, \Sigma_S), \mathbf{y} - \mathbf{Ax}\right) \text{ for all } \mathbf{x} \in \mathbb{R}_+^N \text{ and } \mathbf{y} \in \mathbb{R}^M$$

holds true. This will ensure that if we can control the compressibility and the noise, we can also control the estimation error and in particular decode every noiseless observation of S-sparse signals exactly.

3 MAIN RESULTS

Given a measurement matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$ and a norm $\|\cdot\|$ on $\mathbb{R}^M$ we define the decoder as follows: Given $\mathbf{y} \in \mathbb{R}^M$ set $Q_A(\mathbf{y})$ as any minimizer of

$$\operatorname*{argmin}_{\mathbf{z} \geq 0} \left\|\mathbf{Az} - \mathbf{y}\right\|.$$

We call this problem non-negative least residual (NNLR). In particular, for $\|\cdot\| = \|\cdot\|_1$ this problem is called non-negative least absolute deviation (NNLAD) and for $\|\cdot\| = \|\cdot\|_2$ this problem is known as the non-negative least squares (NNLS) studied in [13]. In fact, we can translate the proof techniques fairly simple. We just need to introduce the dual norm.

Definition 3.1. Let $\|\cdot\|$ be a norm on $\mathbb{R}^M$. The norm $\|\cdot\|_*$ on $\mathbb{R}^M$ defined by

$$\|\mathbf{v}\|_* := \sup_{\mathbf{u} \leq 1} \langle \mathbf{v}, \mathbf{u} \rangle,$$

is called dual norm to $\|\cdot\|$.

Note that the dual norm is actually a norm. To obtain a recovery guarantee for NNLR we have certain requirements on the measurement matrix $\mathbf{A}$. We use a null space property.

Definition 3.2. Let $S \in [N]$, $q \in [1, \infty)$ and $\|\cdot\|$ be any norm on $\mathbb{R}^M$. Further let $\mathbf{A} \in \mathbb{R}^{M \times N}$. Suppose there exists constants $\rho \in [0, 1)$ and $\tau \in [0, \infty)$ such that

$$\|\mathbf{v}|_T\|_q \leq \rho S^{1/q-1} \|\mathbf{v}|_{T^c}\|_1 + \tau ||\mathbf{Av}|| \text{ for all } \mathbf{v} \in \mathbb{R}^N \text{ and } \#(T) \leq S.$$

Then, we say $\mathbf{A}$ has the ℓ_q-robust null space property of order S with respect to $\|\cdot\|$ or in short $\mathbf{A}$ has the ℓ_q-RNSP of order S with respect to $\|\cdot\|$ with constants ρ and τ. ρ is called stableness constant and τ is called robustness constant.

Note that smaller stableness constants increase the reliability of recovery if many, small, non-zero components are present, while smaller robustness constants increase the reliability if the measurements are noisy. In order to make use of the non-negativity of the signal, we need $\mathbf{A}$ to be biased in a certain way. In [13] this bias was guaranteed with the M^+ criterion.

Definition 3.3. Let $\mathbf{A} \in \mathbb{R}^{M \times N}$. Suppose there exists $\mathbf{t} \in \mathbb{R}^M$ such that $\mathbf{A}^T\mathbf{t} > 0$. Then we say $\mathbf{A}$ obeys the the M^+ criterion with vector $\mathbf{t}$ and constant $\kappa := \max_{n \in [N]} \left|(\mathbf{A}^T\mathbf{t})_n\right| \max_{n \in [N]} \left|(\mathbf{A}^T\mathbf{t})_n^{-1}\right|$.

Note that κ is actually a condition number of the matrix with diagonal $\mathbf{A}^T\mathbf{t}$ and 0 else. Condition number numbers are frequently used in error bounds of numerical linear algebra. The general recovery guarantee is the following and similar results have been obtained in the matrix case in [23].

Theorem 3.4 (NNLR Recovery Guarantee). Let $S \in [N]$, $q \in [1, \infty)$ and $\|\cdot\|$ be any norm on $\mathbb{R}^M$ with dual norm $\|\cdot\|_*$. Further, suppose that $\mathbf{A} \in \mathbb{R}^{M \times N}$ obeys

a) the ℓ_q-RNSP of order S with respect to $\|\cdot\|$ with constants ρ and τ and
b) the M^+ criterion with vector $\mathbf{t}$ and constant κ.

If $\kappa\rho < 1$, the following recovery guarantee holds true: For all $\mathbf{x} \in \mathbb{R}_+^N$ and $\mathbf{y} \in \mathbb{R}^M$ any minimizer $\mathbf{x}^\#$ of

$$\operatorname*{argmin}_{\mathbf{z} \geq 0} \left\|\mathbf{Az} - \mathbf{y}\right\|,$$

obeys the bound

$$\left\|\mathbf{x} - \mathbf{x}^\#\right\|_q \leq 2\frac{(1+\kappa\rho)^2}{1-\kappa\rho} \kappa S^{1/q-1} d_1(\mathbf{x}, \Sigma_S) + 2\left(\frac{(1+\kappa\rho)^2}{1-\kappa\rho} S^{1/q-1} \max_{n \in [N]} \left|(\mathbf{A}^T\mathbf{t})_n^{-1}\right| \|\mathbf{t}\|_* + \frac{3+\kappa\rho}{1-\kappa\rho}\kappa\tau\right) \left\|\mathbf{Ax} - \mathbf{y}\right\|.$$

If $q = 1$, this bound can be improved to

$$\left\|\mathbf{x} - \mathbf{x}^\#\right\|_1 \leq 2\frac{1+\kappa\rho}{1-\kappa\rho} \kappa d_1(\mathbf{x}, \Sigma_S) + 2\left(\frac{1+\kappa\rho}{1-\kappa\rho} \max_{n \in [N]} \left|(\mathbf{A}^T\mathbf{t})_n^{-1}\right| \|\mathbf{t}\|_* + \frac{2}{1-\kappa\rho}\kappa\tau\right) \left\|\mathbf{Ax} - \mathbf{y}\right\|.$$

Proof The proof can be found in **Subsection 6.1**.

Given a matrix with ℓ_q-RNSP we can add a row of ones (or a row consisting of one minus the column sums of the matrix) to fulfill the M^+ criterion with the optimal $\kappa = 1$. Certain random measurement matrices guarantee uniform bounds on κ for fixed vectors $\mathbf{t}$. In ([13], Theorem 12) it was proven that if $A_{m,n}$ are all i.i.d. 0/1 Bernoulli random variables, $\mathbf{A}$ has M^+ criterion with $\mathbf{t} = (1, \ldots, 1)^T \in \mathbb{R}^M$ and $\kappa \leq 3$ with high probability. This is problematic, since if $\kappa > 1$, it might happen that $\kappa\rho < 1$ is not fulfilled anymore. Since the stableness constant $\rho(S')$ as a function of S' is monotonically increasing, the condition $\kappa\rho(S') < 1$ might only hold if $S' < S$. If that is the case, there are vectors $\mathbf{x} \in \Sigma_S$ that are being recovered by basis pursuit denoising but not by NNLS! This is for instance the case for the matrix $\mathbf{A} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}$, which has ℓ_1-robust null space property of order 1 with stableness constant $\rho := 1/2$ and M^+

criterion with $\kappa \geq 2$ for any possible choice of $\mathbf{t}$. In particular, the vector $\mathbf{x} = (0, 0, 1)^T$ is not necessarily being recovered by the NNLAD and the NNLS.

Hence, it is crucial that the vector $\mathbf{t}$ is chosen to minimize κ and ideally obeys the optimal $\kappa = 1$. This motivates us to use random walk matrices of regular graphs since they obey exactly this.

Definition 3.5. Let $\mathbf{A} \in \{0, 1\}^{M \times N}$ and $D \in [M]$. For $T \subset N$ the set

$$\mathrm{Row}(T) := \cup_{n \in T}\{m \in [M] \text{such that } A_{m,n} = 1\}$$

is called the set of right vertices connected to the set of left vertices T. If

$$\#(\mathrm{Row}(\{n\})) = D \text{ for all } n \in [N],$$

then $D^{-1}\mathbf{A} \in \{0, D^{-1}\}^{M \times N}$ is called a random walk matrix of a D-left regular bipartite graph. We also say short that $D^{-1}\mathbf{A}$ is a D-LRBG. If additionally there exists a $\theta \in [0, 1)$ such that

$$\#(\mathrm{Row}(T)) \geq (1 - \theta)D\#(T) \text{ for all } \#(T) \leq S,$$

holds true, then $D^{-1}\mathbf{A}$ is called a random walk matrix of a (S, D, θ)-lossless expander.

Note that we have made a slight abuse of notation. The term D-LRBG as a short form for D-left regular bipartite graph refers in our case to the random walk matrix $\mathbf{A}$ but not the graph itself. We omit this minor technical differentiation, for the sake of shortening the frequently used term random walk matrix of a D-left regular bipartite graph. Lossless expanders are bipartite graphs that have a low number of edges but are still highly connected, see for instance ([24], Chapter 4). As a consequence their random walk matrices have good properties for compressed sensing. It is well known that random walk matrices of a $(2S, D, \theta)$-lossless expanders obey the ℓ_1-RNSP of order S with respect to $\|\cdot\|$, see ([3], Theorem 13.11). The dual norm of $\|\cdot\|_1$ is the norm $\|\cdot\|_\infty$ and the M^+ criterion is easily fulfilled, since the columns sum up to one. From Theorem 3.4 we can thus draw the following corollary.

Corollary 3.6 (Lossless Expander Recovery Guarantee). Let $S \in [N]$, $\theta \in [0, 1/6)$. If $\mathbf{A} \in \{0, D^{-1}\}^{M \times N}$ is a random walk matrix of a $(2S, D, \theta)$-lossless expander, then the following recovery guarantee holds true: For all $\mathbf{x} \in \mathbb{R}_+^N$ and $\mathbf{y} \in \mathbb{R}^M$ any minimizer $\mathbf{x}^\#$ of

$$\underset{\mathbf{z} \geq 0}{\mathrm{argmin}} \left\|\mathbf{A}\mathbf{z} - \mathbf{y}\right\|_1$$

obeys the bound

$$\left\|\mathbf{x} - \mathbf{x}^\#\right\|_1 \leq 2\frac{1 - 2\theta}{1 - 6\theta} d_1(\mathbf{x}, \Sigma_S) + 2\frac{3 - 2\theta}{1 - 6\theta}\left\|\mathbf{A}\mathbf{x} - \mathbf{y}\right\|_1 \quad (1)$$

Proof By ([3], Theorem 13.11) $\mathbf{A}$ has ℓ_1-RNSP with respect to $\|\cdot\|_1$ with constants $\rho = 2\theta/(1 - 4\theta)$ and $\tau = 1/(1 - 4\theta)$. The dual norm of the norm $\|\cdot\|_1$ is $\|\cdot\|_\infty$. If we set $\mathbf{t} := (1, \ldots, 1)^T \in \mathbb{R}^M$, we get

$$\left(\mathbf{A}^T\mathbf{t}\right)_n = \sum_{m \in [M]} A_{m,n} = DD^{-1} = 1 \text{ for all } n \in [N].$$

Hence, $\mathbf{A}$ has the M^+ criterion with vector $\mathbf{t}$ and constant $\kappa = 1$ and the condition $\kappa\rho < 1$ is immediately fulfilled. We obtain $||\mathbf{t}||_* = ||\mathbf{t}||_\infty = 1$ and $\max_{n \in [N]}\left|\left(\mathbf{A}^T\mathbf{t}\right)_n^{-1}\right| = 1$. Applying Theorem 3.4 with improved bound for $q = 1$ and these values yields

$$\left\|\mathbf{x} - \mathbf{x}^\#\right\|_1 \leq 2\frac{1 + \rho}{1 - \rho} d_1(\mathbf{x}, \Sigma_S) + 2\left(\frac{1 + \rho}{1 - \rho} + \frac{2}{1 - \rho}\tau\right)\left\|\mathbf{A}\mathbf{x} - \mathbf{y}\right\|_1.$$

If we additionally substitute the values for ρ and τ we get

$$\begin{aligned}\left\|\mathbf{x} - \mathbf{x}^\#\right\|_1 &\leq 2\frac{1 - 2\theta}{1 - 6\theta} d_1(\mathbf{x}, \Sigma_S) + 2\left(\frac{1 - 2\theta}{1 - 6\theta} + 2\frac{1}{1 - 6\theta}\right)\left\|\mathbf{A}\mathbf{x} - \mathbf{y}\right\|_1 \\ &\leq 2\frac{1 - 2\theta}{1 - 6\theta} d_1(\mathbf{x}, \Sigma_S) + 2\frac{3 - 2\theta}{1 - 6\theta}\left\|\mathbf{A}\mathbf{x} - \mathbf{y}\right\|_1.\end{aligned}$$

This finishes the proof.

Note that ([3], Theorem 13.11) is an adaption of ([20], Lemma 11) to account for robustness and skips proving the ℓ_1 restricted isometry property. If $M \geq 2/\theta \exp(2/\theta)\mathrm{SLn}(eN/S)$ and $D = \lceil 2/\theta \mathrm{Ln}(eN/S)\rceil$, a uniformly at random drawn D-LRBG is a random walk matrix of a $(2S, D, \theta)$-lossless expander with a high probability[[3], Theorem 13.7]. Thus, recovery with the NNLAD is possible in the optimal regime $M \in \mathcal{O}(S\log N/S)$.

3.2 On the Robustness Bound for Lossless Expanders

If $\mathbf{A}$ is a random walk matrix of a $(2S, D, \theta)$-lossless expander with $\theta \in [0, 1/6)$, then we can also draw a recovery guarantee for the NNLS. By ([3], Theorem 13.11) $\mathbf{A}$ has ℓ_1-RNSP with respect to $\|\cdot\|_1$ with constants $\rho = 2\theta/(1 - 4\theta)$ and $\tau = 1/(1 - 4\theta)$ and hence also ℓ_1-RNSP with respect to $\|\cdot\|_2$ with constants $\rho' = \rho$ and $\tau' = \tau M^{1/2}$. Similar to the proof of Corollary 3.6 we can use Theorem 3.4 to deduce that any minimizer $\mathbf{x}^\#$ of

$$\underset{\mathbf{z} \geq 0}{\mathrm{argmin}} \left\|\mathbf{A}\mathbf{z} - \mathbf{y}\right\|_2,$$

obeys the bound

$$\left\|\mathbf{x} - \mathbf{x}^\#\right\|_1 \leq 2\frac{1 - 2\theta}{1 - 6\theta} d_1(\mathbf{x}, \Sigma_S) + 2\frac{3 - 2\theta}{1 - 6\theta} M^{1/2}\left\|\mathbf{A}\mathbf{x} - \mathbf{y}\right\|_2 \quad (2)$$

If the measurement error $\mathbf{e} = \mathbf{y} - \mathbf{A}\mathbf{x}$ is a constant vector, i.e., $\mathbf{e} = \alpha 1$, then $\|\mathbf{e}\|_1 = M^{1/2}\|\mathbf{e}\|_2$. In this case the error bound of the NNLS is just as good as the error bound of the NNLAD. However, if $\mathbf{e}$ is a standard unit vector, then $\|\mathbf{e}\|_1 = \|\mathbf{e}\|_2$. In this case the error bound of the NNLS is significantly worse than the error bound of the NNLAD. Thus, the NNLAD performs better under peaky noise, while the NNLS and NNLAD are tied under noise with evenly distributed mass. We will verify this numerically in **Subsection 5.1**. One can draw a complementary result for matrices with biased sub-Gaussian entries, which obey the ℓ_2-RNSP with respect to $\|\cdot\|_2$ and the M^+ criterion in the optimal regime [13]. **Table 1** states the methods, which have an advantage over the other in each scenario.

4 NON-NEGATIVE LEAST ABSOLUTE DEVIATION USING A PROXIMAL POINT METHOD

In this section we assume that $\|\cdot\| = \|\cdot\|_p$ with some $p \in [1, \infty]$. If $p \in \{1, \infty\}$, the NNLR can be recast as a linear program by

introducing some slack variables. For an arbitrary p the NNLR is a convex optimization problem and the objective function has a simple and globally bounded subdifferential. Thus, the NNLR can directly be solved with a projective subgradient method using a problem independent step size. Such subgradient methods achieve only a convergence rate of $\mathcal{O}(\log(k)k^{-1/2})$ toward the optimal objective value ([25], Section 3.2.3), where k is the number of iterations performed. In the case that the norm is the ℓ_2-norm, we can transfer the problem into a differentiable version, i.e. the NNLS

$$\operatorname*{argmin}_{\mathbf{z}\geq 0} \frac{1}{2}\left\|\mathbf{A}\mathbf{z}-\mathbf{y}\right\|_2^2.$$

Since the gradient of such an objective is Lipschitz, this problem can be solved by a projected gradient method with constant step size, which achieves a convergence rate of $\mathcal{O}(k^{-2})$ toward the optimal objective value [26, 27]. However this does not generalize to the ℓ_1-norm. The proximal point method proposed in [15] can solve the case of the ℓ_1-norm with a convergence rate $\mathcal{O}(k^{-1})$ toward the optimal objective value. Please refer to **Algorithm 1**.

Algorithm 1 is a primal-dual algorithm. Within the loop, lines 7, 8, 1 and 2 calculate the proximal point operator of the Fenchel conjugate of $\left\|\mathbf{A}\cdot-\mathbf{y}\right\|_1$ to update the dual problem, lines 3 and 5 update the primal problem, and lines 4 and 6 perform a momentum step to accelerate convergence. Further, line 8 sets $\tilde{x}$ to $\mathbf{A}\mathbf{x}$ and avoids a third matrix vector multiplication. Note that σ_1 and σ_2 can be replaced by any values that satisfy $\sigma_1\sigma_2 < ||\mathbf{A}||_{2\to 2}^{-2}$. The calculation of σ_1 and σ_2 might be a bottle neck for the computational complexity of the algorithm. If one wants to solve multiple problems with the same matrix, σ_1 and σ_2 should only be calculated once and not in each run of the algorithm. For any $\sigma_1\sigma_2 < \|\mathbf{A}\|_{2\to 2}^{-2}$ the following convergence guarantee can be deduced from ([15], Theorem 1). Let $\mathbf{x}^k$ and $\mathbf{w}^k$ be the values of $\mathbf{x}$ and $\mathbf{w}$ at the end of the kth iteration of the while loop of **Algorithm 1**. Then, the following statements hold true:

(1) The iterates converge: The sequence $(\mathbf{x}^k)_{k\in\mathbb{N}}$ converges to a minimizer of $\operatorname{argmin}_{\mathbf{z}\geq 0}\left\|\mathbf{A}\mathbf{z}-\mathbf{y}\right\|_1$.
(2) The iterates are feasible: We have $\mathbf{x}^k \geq 0$ and $\left\|\mathbf{w}^k\right\|_\infty \leq 1$ for all $k \geq 1$.
(3) There is a stopping criteria for the iterates: $\lim_{k\to\infty}\left\|\mathbf{A}\mathbf{x}^k-\mathbf{y}\right\|_1 + \langle \mathbf{y}, \mathbf{w}^k\rangle = 0$ and $\lim_{k\to\infty}\mathbf{A}^T\mathbf{w}^k \geq 0$. In particular, if $\left\|\mathbf{A}\mathbf{x}^k-\mathbf{y}\right\|_1 + \langle \mathbf{y}, \mathbf{w}^k\rangle \leq 0$ and $\mathbf{A}^T\mathbf{w}^k \geq 0$, then $\mathbf{x}^k$ is a minimizer of $\operatorname{argmin}_{\mathbf{z}\geq 0}\left\|\mathbf{A}\mathbf{z}-\mathbf{y}\right\|_1$.
(5) The averages obey the convergence rate toward the optimal objective value: $\left\|\mathbf{A}1/k\sum_{k'=1}^{k}\mathbf{x}^{k'}-\mathbf{y}\right\|_1 - \left\|\mathbf{A}\mathbf{x}^{\#}-\mathbf{y}\right\|_1 \leq 1/k\,(1/(2\sigma_2))\left\|\mathbf{x}^{\#}-\mathbf{x}^0\right\|_2^2 + 1/(2\sigma_1)(\left\|\mathbf{w}^0\right\|_2^2 + 2\left\|\mathbf{w}^0\right\|_1 + M))$, where $\mathbf{x}^{\#}$ is a minimizer of $\operatorname{argmin}_{z\geq 0}\left\|\mathbf{A}\mathbf{z}-\mathbf{y}\right\|_1$.

ALGORITHM 1 | NNLAD as First Order Method

Data: *measurement* $\mathbf{y}\in\mathbb{R}^M$, *measurement matrix* $\mathbf{A}\in\mathbb{R}^{M\times N}$, *initializations* $\mathbf{x}^0\in\mathbb{R}^N$, $\mathbf{w}^0\in\mathbb{R}^M$, *precision parameters* $\epsilon_1\geq 0, \epsilon_2\geq 0$
Result: *estimator* $\mathbf{x}^{\#}\in\mathbb{R}^N$
$\sigma_1 \leftarrow 0.99\,\|\mathbf{A}\|_{2\to 2}^{-1}$; $\sigma_2 \leftarrow \sigma_1$;
initialize iterates;
$\mathbf{x}\leftarrow\mathbf{x}^0$; $\mathbf{v}\leftarrow\mathbf{x}^0$; $\mathbf{w}\leftarrow\mathbf{w}^0$;
initialize images;
$\tilde{\mathbf{w}}\leftarrow\mathbf{A}^T\mathbf{w}$; $\tilde{\mathbf{x}}\leftarrow\mathbf{A}\mathbf{x}$; $\tilde{\mathbf{v}}\leftarrow\mathbf{A}\mathbf{v}$;
while $\|\tilde{\mathbf{x}}-\mathbf{y}\|_1 + \langle\mathbf{y},\mathbf{w}\rangle > \epsilon_1$ **or** $\min_{n\in[N]}\tilde{\mathbf{w}}_n < -\epsilon_2$ **do**
  calculate iterates;
  $\mathbf{w}\leftarrow\mathbf{w}+\sigma_1(\tilde{\mathbf{v}}-\mathbf{y})$;
  $\mathbf{w}\leftarrow(\min\{1,|w_m|\}\operatorname{sgn}(w_m))_{m\in[M]}$;
  $\tilde{\mathbf{w}}\leftarrow\mathbf{A}^T\mathbf{w}$;
  $\mathbf{v}\leftarrow-\mathbf{x}$;
  $\mathbf{x}\leftarrow(\max\{0, x_n-\sigma_2\tilde{w}_n\})_{n\in[N]}$;
  $\mathbf{v}\leftarrow\mathbf{v}+2\mathbf{x}$;
  $\tilde{\mathbf{v}}\leftarrow\mathbf{A}\mathbf{v}$;
  $\tilde{\mathbf{x}}\leftarrow\frac{1}{2}(\tilde{\mathbf{v}}+\tilde{\mathbf{x}})$;
end
return $\mathbf{x}^{\#}\leftarrow\mathbf{x}$

TABLE 1 | Table of advantages of NNLAD and NNLS over each other.

		Measurement Matrix	
		D-LRBG (ℓ_1)	**Biased sub-Gaussian** (ℓ_2)
Noise	peaky $\|e\|_1 \approx \|e\|_2$	NNLAD	–
	even mass $\|\|e\|\|_1 \approx M^{1/2}\|e\|_2$	–	NNLS
	unknown noise	NNLAD	NNLS

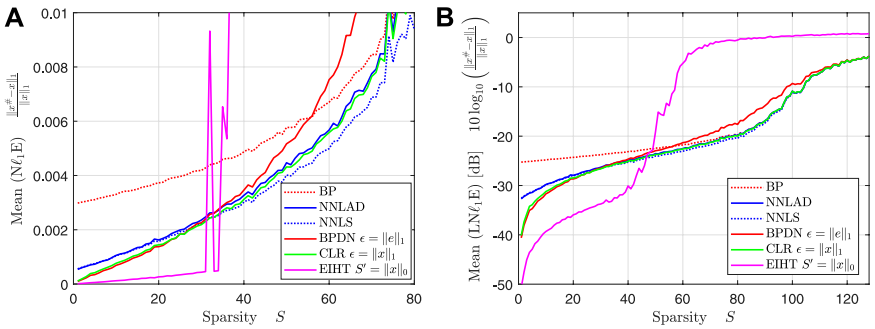

FIGURE 1 | Performance of NNLAD for noise with even mass noise and varying sparsity of the signal. **(A)** NNLAD has almost the same performance as CLR/BPDN. EIHT fails for moderate S. **(B)** NNLAD and NNLS perform roughly the same.

The formal version and proof can be found in [28]. Note that this yields a convergence guarantee for both the iterates and averages, but the convergence rate is only guaranteed for the averages. **Algorithm 1** is optimized in the sense that it uses the least possible number of matrix vector multiplications per iteration, since these govern the computational complexity.

Remark 4.1. Let **A** be D-LRBG. Each iteration of **Algorithm 1** requires at most $4DN + 8N + 16M$ floating point operations and $5N + 4M$ assignments.

4.1 Iterates or Averages

The question arises whether or not it is better to estimate with averages or iterates. Numerical testing suggest that the iterates reach tolerance thresholds significantly faster than the averages. We can only give a heuristically explanation for this phenomenon. The stopping criteria of the iterates yields $\lim_{k\to\infty}\mathbf{A}^T\mathbf{w}^k \geq 0$. In practice we observe that $\mathbf{A}^T\mathbf{w}^k \geq 0$ for all sufficiently large k. However, $\mathbf{A}^T\mathbf{w}^{k+1} \geq 0$ yields $\mathbf{x}^{k+1} \leq \mathbf{x}^k$. This monotonicity promotes the converges of the iterates and gives a clue why the iterates seem to converge better in practice. See **Figures 5**, **6**.

4.2 On the Convergence Rate

As stated the NNLS achieves the convergence rate $\mathcal{O}(k^{-2})$ [27] while the NNLAD only achieves the convergence rate of $\mathcal{O}(k^{-1})$ toward to optimal objective value. However, this should not be considered as weaker, since the objective function of the NNLS is the square of a norm. If $\mathbf{x}^k$ are the iterates of the NNLS implementation of [27], algebraic manipulation yields

$$\left\|\mathbf{A}\mathbf{x}^k - \mathbf{y}\right\|_2 - \left\|\mathbf{A}\mathbf{x}^{\#} - \mathbf{y}\right\|_2 \leq 2^{1/2}\left(\frac{1}{2}\left\|\mathbf{A}\mathbf{x}^k - \mathbf{y}\right\|_2^2 - \frac{1}{2}\left\|\mathbf{A}\mathbf{x}^{\#} - \mathbf{y}\right\|_2^2\right)^{1/2}$$
$$\leq 2^{1/2}\left(Ck^{-2}\right)^{1/2} \leq (2C)^{1/2}k^{-1}.$$

Thus, the ℓ_2-norm of the residual of the NNLS iterates only decays in the same order as the ℓ_1-norm of the residual of the NNLAD averages.

5 NUMERICAL EXPERIMENTS AND APPLICATIONS

In the first part of this section we will compare NNLAD with several state of the art recovery methods in terms of achieved sparsity levels and decoding time. For $p \in [1, \infty]$, we denote $\mathbb{S}_p^{N-1} := \{\mathbf{z} \in \mathbb{R}^N : \|\mathbf{z}\|_p = 1\}$, and $\mathbb{S}_0^{N-1} := \{\mathbf{z} \in \mathbb{R}^N : \|\mathbf{z}\|_0 = 1 = ||\mathbf{z}||_2\} = \Sigma_1 \cap \mathbb{S}_2^{N-1}$.

5.1 Properties of the Non-Negative Least Absolute Deviation Optimizer

We recall that the goal is to recover $\mathbf{x}$ from the noisy linear measurements $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$. To investigate properties of the minimizers of NNLAD we compare it to the minimizers of the well studied problems basis pursuit (BP), optimally tuned basis pursuit denoising (BPDN), optimally tuned ℓ_1-constrained least residual (CLR) and the NNLS, which are given by

$$\begin{aligned}
&\underset{\mathbf{z}:\|\mathbf{Az}-\mathbf{y}\|_1 \le \epsilon}{\operatorname{argmin}} \|\mathbf{z}\|_1 \text{ with } \epsilon = \|\mathbf{e}\|_1 && \text{(BPDN)},\\
&\underset{\mathbf{z}:\|\mathbf{z}\|_1 \le \tau}{\operatorname{argmin}} \left\|\mathbf{Az}-\mathbf{y}\right\|_1 \text{ with } \tau = \|\mathbf{x}\|_1 && \text{(CLR)},\\
&\underset{\mathbf{z}:\mathbf{Az}=\mathbf{y}}{\operatorname{argmin}} \|\mathbf{z}\|_1 && \text{(BP)},\\
&\underset{\mathbf{z}\ge 0}{\operatorname{argmin}} \left\|\mathbf{Az}-\mathbf{y}\right\|_2 && \text{(NNLS)}.
\end{aligned}$$

The NNLAD is designed to recover non-negative signals in general and, as we will see, it is able to recover sparse non-negative signals comparable well to the CLR and BPDN with optimal tuning which are particularly designed for this task. Further, we compare the NNLAD to the expander iterative hard thresholding (EIHT). The EIHT is calculated by stopping the following sequence after a suitable stopping criteria is met:

$$\begin{aligned}
&\mathbf{x}^0 := 0 \text{ and } \mathbf{x}^{k+1} := \mathcal{P}_{\Sigma_{S'}}\left(\mathbf{x}^k + \text{median}\left(\mathrm{y} - \mathbf{A}\mathbf{x}^k\right)\right)\\
&\qquad \text{for all } k \in \mathbb{N}_0 \text{ and with } S' = \|\mathbf{x}\|_0 \quad \text{(EIHT)},
\end{aligned}$$

where $\text{median}\left(\mathrm{z}\right)_n$ is the median of $(\mathbf{z}_m)_{m \in \text{Row}(\{n\})}$ and $\mathcal{P}_{\Sigma_S}(\mathbf{v})$ is a hard thresholding operator, i.e., some minimizer of $\text{argmin}_{z \in \Sigma_S} 1/2\|\mathbf{z}-\mathbf{v}\|_2^2$. There is a whole class of thresholding based decoders for lossless expanders, which all need either the sparsity of the signal or the order of the expansion property as tuning parameter. We choose the EIHT as a represent of this class, since the cluster points of its sequence have robust recovery guarantees ([3], Theorem 13.5). By convex decoders we refer to BPDN, BP, CLR, NNLAD, and NNLS. We choose the optimal tuning $\epsilon = \|\mathbf{e}\|_1$ for the BPDN and $\tau = ||\mathbf{x}||_1$ for the CLR. The optimally tuned BPDN and CLR are representing a best case benchmark. In ([29], Figure 1.1) it was noticed that tuning the BPDN with $\epsilon > ||\mathbf{e}||_p$ often leads to worse estimation errors than tuning with $\epsilon < ||\mathbf{e}||_p$ for $p = 2$. Thus, BP is a version of BPDN with no prior knowledge about the noise and represents a worst case benchmark. At fist we investigate the properties of the estimators. In order to mitigate effects from different implementations we solve all optimization problems with the CVX package of Matlab [30, 31]. For a given $\ell_1 SNR, r, N, M, D, S$ we will do the following experiment multiple times:

Experiment 1

1. Generate a measurement matrix $\mathbf{A} \in \{0, D^{-1}\}^{M \times N}$ uniformly at random among all D-LRBG.
2. Generate a signal $\mathbf{x}$ uniformly at random from $\Sigma_S \cap \mathbb{R}_+^N \cap \mathbb{S}_1^{N-1}$.
3. Generate a noise $\mathbf{e}$ uniformly at random from $\|\mathbf{Ax}\|_1/\ell_1 SNR\, \mathbb{S}_r^{M-1}$.
4. Define the observation $\mathbf{y} := \mathbf{Ax} + \mathbf{e}$.
5. For each decoder Q_A calculate an estimator $\mathbf{x}^\# := Q_A(\mathbf{y})$ and collect the relative estimation error $\left\|\mathbf{x}-\mathbf{x}^\#\right\|_1 = \left\|\mathbf{x}-\mathbf{x}^\#\right\|_1/\|\mathbf{x}\|_1$.

In this experiment we have $\ell_1 SNR = \|\mathbf{Ax}\|_1/\|\mathbf{e}\|_1$ and since $\mathbf{A}$ is a D-LRBG and $\mathbf{x} \ge 0$, we further have $\|\mathbf{Ax}\|_1 = ||\mathbf{x}||_1 = 1$. Note that for $r = 0$ and $r = 1$ we obtain two different noise distributions. If $\mathbf{e}$ is uniformly distributed on $\mathbb{S}_1^{M-1}$, then the absolute value of each component $|e_m|$ is a random variable with density $h \mapsto (M-1)(1-h)^{M-2}$ for $h \in [0,1]$. Thus, $\mathcal{E}[||\mathbf{e}||_2^2] = M2/M(M+1) = 2/(M+1)$. By testing one can observe a concentration around this expected value, in particular that $M^{1/2}\|\mathbf{e}\|_2 \approx \sqrt{2}||\mathbf{e}||_1$ with a high probability. If $\mathbf{e}$ is uniformly distributed on $\mathbb{S}_0^{M-1}$, then $\|\mathbf{e}\|_2 = \|\mathbf{e}\|_1$. Thus, these two noise distributions each represent randomly drawn noise vectors obeying one norm equivalence asymptotically tightly up to a constant. From **Eqs 1, 2** we expect that the NNLS has roughly the same estimation errors as the NNLAD for $r = 1$, i.e. the evenly distributed noise, and significantly worse estimation errors for $r = 0$, i.e., the peaky noise.

5.1.1 Quality of the Estimation Error for Varying Sparsity

We fix the constants $r = 1$, $N = 1024$, $M = 256$, $D = 10$, $\ell_1 SNR = 1000$ and vary the sparsity level $S \in [128]$. For each S we repeat Experiment 1 100 times. We plot the mean of the relative

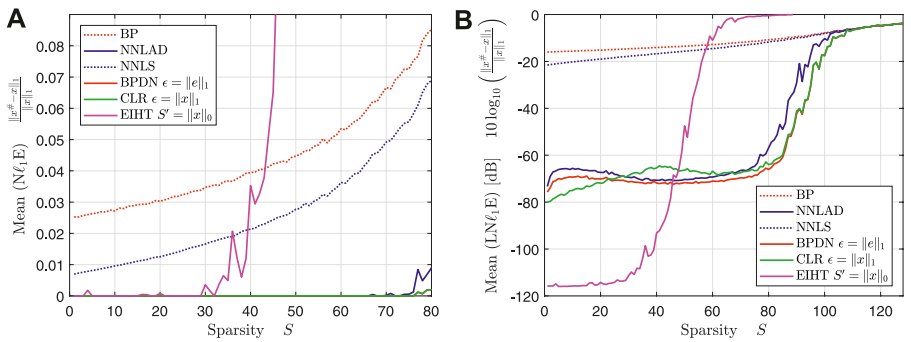

FIGURE 2 | Performance of NNLAD for noise with peaky mass and varying sparsity of the signal. **(A)** The NNLS does not fail, but performs bad. **(B)** The NNLS and NNLAD differ strongly.

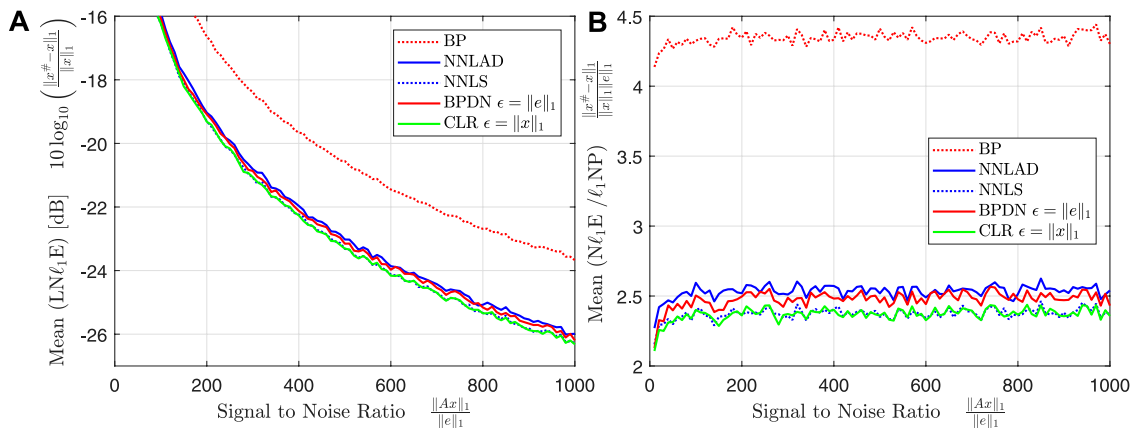

FIGURE 3 | Performance of NNLAD for noise with even mass and varying noise power. **(A)** The NNLAD and NNLS recover reliably for all signal to noise ratios. **(B)** The estimation error scales linearly with the noise power.

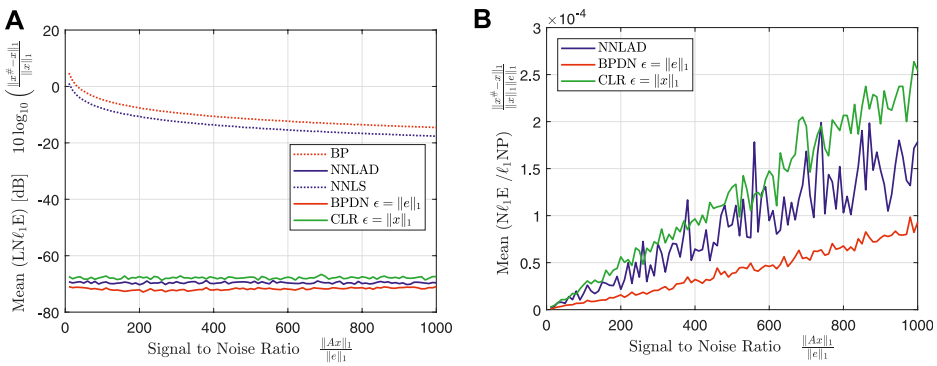

FIGURE 4 | Performance of NNLAD for noise with peaky mass and varying noise power. **(A)** The NNLAD outperforms the NNLS. **(B)** The estimation error does not scale linearly with the noise power.

ℓ_1-estimation error and the mean of the logarithmic relative ℓ_1-estimation error, i.e.,

$$\mathrm{Mean}\,(N\ell_1 E) = \mathrm{Mean}\left(\frac{\left\|\mathbf{x}-\mathbf{x}^{\#}\right\|_1}{\|\mathbf{x}\|_1}\right),$$

$$\mathrm{Mean}\,(\mathrm{LN}\ell_1 E) = \mathrm{Mean}\left(10\log_{10}\left(\frac{\left\|\mathbf{x}-\mathbf{x}^{\#}\right\|_1}{\|\mathbf{x}\|_1}\right)\right),$$

over the sparsity. The result can be found in **Figures 1A,B**.

For $S \geq 30$ the estimation error of the EIHT randomly peaks high. We deduce that the EIHT fails to recover the signal reliably for $S \geq 30$, while the NNLAD and other convex decoders succeed. This is not surprising, since by ([3], Theorem 13.15) the EIHT obeys a robust recovery guartanee for S-sparse signals, whenever $\mathbf{A}$ is the random wak matrix of a $(3S, D, \theta')$-lossless expander with $\theta' < 1/12$. This is significantly stronger than the $(2S, D, \theta)$-lossless expander property with $\theta < 1/6$ required for a null space property. It might also be that the null space property is more likely than the lossless expansion property similar to the gap between ℓ_2-restricted isometry property and null space property [32]. However, if the EIHT recovers a signal, it recovers it significantly better than any convex method. This might be the case, since the originally generated signal is indeed from Σ_S, which is being enforced by the hard thresholding of the EIHT, but not by the convex decoders. This suggests that it might be useful to consider using thresholding on the output of any convex decoder to increase the accuracy if the orignal signal is indeed sparse and not only compressible. For the remainder of this subsection we focus on convex decoders.

Contrary to our expectation the BPDN achieves worse estimation errors than all other convex decoders for $S \geq 60$, even worse than the BP. The authors have no explanation for this phenomenon. Apart from that we observe that the CLR and BP indeed perform as respectively best and worst case benchmark. However, the difference between BP and CLR becomes rather small for high S. We deduce that tuning becomes less important near the optimal sampling rate.

The NNLAD, NNLS and CLR achieve roughly the same estimation errors. However, note that the BPDN and CLR are

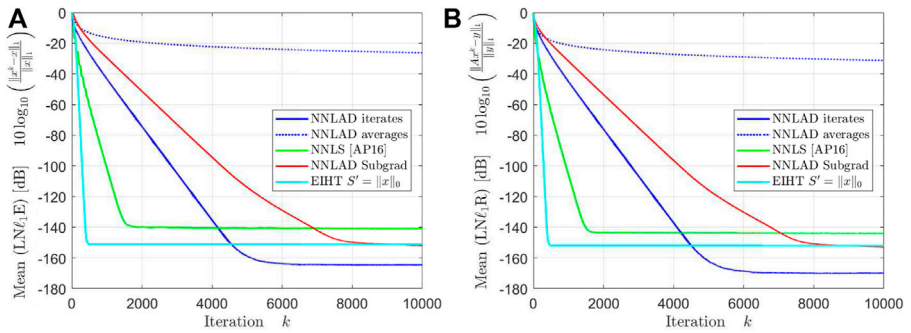

FIGURE 5 | Convergence rates of certain iterated methods with respect to the number of iterations.

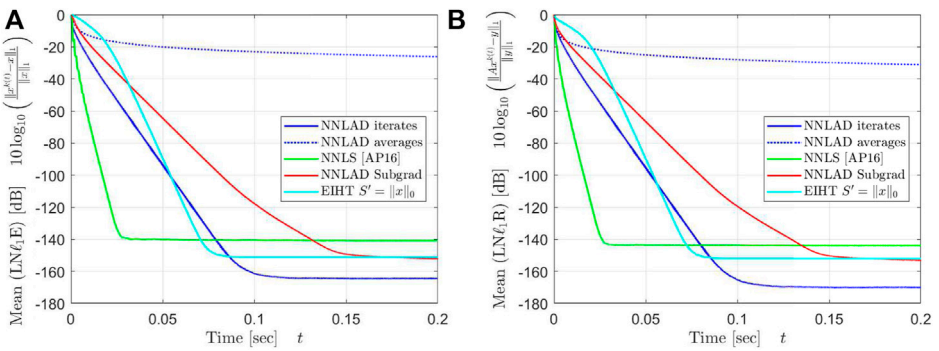

FIGURE 6 | Convergence rates of certain iterated methods with respect to the time.

optimally tuned using unknown prior information unlike the NNLAD and NNLS. As expected the NNLS performs roughly the same as the NNLAD, see **Table 1**. However, this is the result of the noise distribution for $r = 1$. We repeat Experiment 1 with the same constants, but $r = 0$, i.e., $\mathbf{e}$ is a unit vector scaled by $\pm \|\mathbf{Ax}\|_1/\ell_1 SNR$. We plot the mean of the relative ℓ_1-estimation error and the mean of the logarithmic relative ℓ_1-estimation error over the sparsity. The result can be found in **Figures 2A,B**.

We want to note that similarly to **Figure 1A** the EIHT works only unreliably for $S \geq 30$. Even though the mean of the logarithmic relative ℓ_1-estimation error of NNLS is worse than the one of EIHT for $30 \leq S \leq 60$, the NNLS does not fail but only approximates with a weak error bound. As the theory suggests, the NNLS performs significantly worse than the NNLAD, see **Table 1**. It is worth to mention, that the estimaton errors of NNLS seem to be bounded by the estimation errors of BP. This suggests that $\mathbf{A}$ obeys a ℓ_1 quotient property, that bounds the estimation error of any instance optimal decoder, see ([3], Lemma 11.15).

5.1.2 Noise-Blindness

Theorem 3.4 states that the NNLAD has an error bound similarly to the optimally tuned CLR and BPDN. Further, by **Eq. 1** the ratio

$$\frac{\left\|\mathbf{x}-\mathbf{x}^{\#}\right\|_1}{\|\mathbf{e}\|_1\|\mathbf{x}\|_1} = \frac{\left\|\mathbf{x}-\mathbf{x}^{\#}\right\|_1}{\|\mathbf{e}\|_1}$$

should be bounded by some constant. To verify this, we fix the constants $r = 1$, $N = 1024$, $M = 256$, $D = 10$, $S = 32$ and vary the signal to noise ratio $\ell_1 SNR \in 10[100]$. For each $\ell_1 SNR$ we repeat Experiment 1 100 times. We plot the mean of the logarithmic relative ℓ_1-estimation error and the mean of the ratio of relative ℓ_1-estimation error and ℓ_1-noise power, i.e.

$$\text{Mean}\left(\text{LN}\ell_1\text{E}\right) = \text{Mean}\left(10\log_{10}\left(\frac{\left\|\mathbf{x}-\mathbf{x}^{\#}\right\|_1}{\|\mathbf{x}\|_1}\right)\right),$$

$$\text{Mean}\left(\frac{\text{N}\ell_1\text{E}}{\ell_1\text{NP}}\right) = \text{Mean}\left(\frac{\left\|\mathbf{x}-\mathbf{x}^{\#}\right\|_1}{\|\mathbf{e}\|_1\|\mathbf{x}\|_1}\right),$$

over the sparsity. The result can be found in **Figures 3A,B** .

The logarithmic relative ℓ_1-estimation errors of the different decoders stay in a constant relation to each other over the whole range of $\ell_1 SNR$. This relation is roughly the relation we can find in **Figure 1B** for $S = 32$. As expected the the ratio of relative ℓ_1-estimation error and ℓ_1-noise power stays constant independent on the

$\ell_1 SNR$ for all decoders. We deduce that the NNLAD is noise-blind. We repeat the experiment with $r = 0$ and obtain **Figures 4A,B**.

Note that $\left\|\mathbf{x} - \mathbf{x}^{\#}\right\|_1 / \|\mathbf{x}\|_1$ and not $\left\|\mathbf{x} - \mathbf{x}^{\#}\right\|_1 / (\|\mathbf{x}\|_1 \|\mathbf{e}\|_1)$ seems to be constant. Since $\left\|\mathbf{x} - \mathbf{x}^{\#}\right\|_1 / \|\mathbf{x}\|_1 \approx 1.0 \cdot 10^{-7}$ is fairly small, we suspect that this is the result of CVX reaching a tolerance parameter[2] $\sqrt{eps} \approx 1.5 \cdot 10^{-8}$ and terminating, while the actual optimizer might in fact be the original signal. It is remarkable that even with the incredibly small signal to noise ratio of 10 the signal can be recovered by the NNLAD with an estimation error of $1.0 \cdot 10^{-7}$ for this noise distribution.

5.2 Decoding Complexity

5.2.1 Non-Negative Least Absolute Deviation Vs Iterative Methods

To investigate the convergence rates of the NNLAD as proposed in 4, we compare it to different types of decoders when $\mathbf{e} = 0$. There are some sublinear time recovery methods for lossless expander matrices including ([3 Section 13.4, 5]). These are, as the name suggests, significantly faster than the NNLAD. These, as several other greedy methods ([3 Section 13.3, 5, 18, 19, 21]), rely on a strong lossless expansion property. As a representative of all greedy and sublinear time methods we will consider the EIHT, which has a linear convergence rate $\mathcal{O}(c^{-k})$ toward the signal and robust recovery guarantees ([3], Theorem 13.15). The EIHT also represents a best case benchmark. As a direct competitor we consider the NNLS implemented by the methods of [27] [3], which has a convergence rate of $\mathcal{O}(k^{-2})$ toward the optimal objective value. [27] can also be used to calculate the CLR if the residual norm is the ℓ_2-norm. However, calculating the projection onto the ℓ_1-ball in $\mathbb{R}^N$, is computationally slightly more complex than the projection onto $\mathbb{R}^N_+$. Thus, the CLR will be solved slightly slower than the NNLS with [27]. Note that cross-validation techniques would need to solve multiple optimization problems of a similar complexity as the NNLS to estimate a signal. As a consequence such methods have a multiple times higher complexity than the NNLS and are not considered here. As a worst case benchmark we consider a simple projected subgradient implementation of NNLAD using the Polyak step size, i.e.

$$\mathbf{x}^{k+1} := \mathcal{P}_{\mathbb{R}^N_+}\left(\mathbf{x}^k - \frac{\left\|\mathbf{A}\mathbf{x}^k - \mathbf{y}\right\|_1}{\left\|\mathbf{A}^T \operatorname{sgn}\left(\mathbf{A}\mathbf{x}^k - \mathbf{y}\right)\right\|_2^2} \mathbf{A}^T \operatorname{sgn}\left(\mathbf{A}\mathbf{x}^k - \mathbf{y}\right)\right), \quad \text{(NNLAD Subgrad)}$$

which has a convergence rate of $\mathcal{O}(k^{-1/2})$ toward the optimal objective value ([33], Section 7.2.2 & Section 5.3.2). We initialized all iterated methods by 0. The EIHT will always use the parameter $S' = ||\mathbf{x}||_0$, the NNLAD $\sigma_1 = \sigma_2 = 0.99\|\mathbf{A}\|^{-1}_{2\to 2}$ and the NNLS the parameters $s = 0.99\|\mathbf{A}\|^{-2}_{2\to 2}$ and $\alpha = 3.01$, see [27]. Just like the BPDN and CLR, the EIHT needs an oracle to get some unknown prior information, in this case $\|\mathbf{x}\|_0$. Parameters that can be computed from $\mathbf{A}$, will be calculated before the timers start. This includes the adjacency structure of $\mathbf{A}$ for the EIHT, σ_1, σ_2 for NNLAD, s, α for NNLS, since these are considered to be a part of the decoder. We will do the following experiment multiple times:

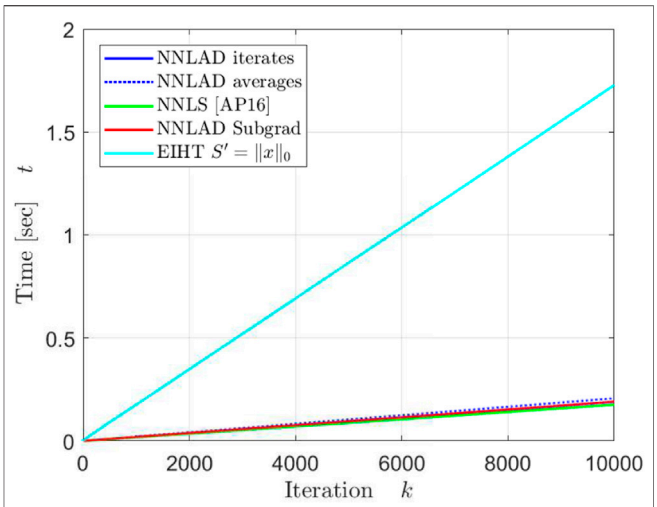

FIGURE 7 | Time required to perform iterations of certain iterated methods.

Experiment 2

1. If $r = 1$, generate a measurement matrix $\mathbf{A} \in \{0, D^{-1}\}^{M\times N}$ uniformly at random among all D-LRBG. If $r = 2$, draw each component $A_{m,n}$ of the measurement matrix independent and uniformly at random from $\{0, 1\}$, i.e., as 0/1 Bernoulli random variables.
2. Generate a signal $\mathbf{x}$ uniformly at random from $\Sigma_S \cap \mathbb{R}^N_+ \cap \mathbb{S}^{N-1}_r$.
3. Define the observation $\mathbf{y} := \mathbf{A}\mathbf{x}$.
4. For each iterative method calculate the sequence of estimators $\mathbf{x}^k$ for all $k \le 20000$ and collect the relative estimation errors $\left\|\mathbf{x}^k - \mathbf{x}\right\|_1 / \|\mathbf{x}\|_1$, the relative norms of the residuals $\left\|\mathbf{A}\mathbf{x}^k - \mathbf{y}\right\|_1 / \left\|\mathbf{y}\right\|_1$ and the time to calculate the first k iterations.

For $r = 2$ this represents a biased sub-Gaussian random ensemble [13] with optimal recovery guarantees for the NNLS.For $r = 1$ this represents a D-LRBG random ensemble with optimal recovery guarantees for the NNLAD. We fix the constants $r = 1$, $N = 1024$, $M = 256$, $S = 16$, $D = 10$ and repeat 2 100 times. We plot the mean of the logarithmic relative ℓ_1-estimation error and the mean of the relative ℓ_1-norm of the residual, i.e.

$$\begin{aligned} \text{Mean}\left(\text{LN}\ell_1\text{E}\right) &= \text{Mean}\left(10\log_{10}\left(\frac{\left\|\mathbf{x}^k - \mathbf{x}\right\|_1}{\left\|\mathbf{x}\right\|_1}\right)\right), \\ \text{Mean}\left(\text{LN}\ell_1\text{R}\right) &= \text{Mean}\left(10\log_{10}\left(\frac{\left\|\mathbf{A}\mathbf{x}^k - \mathbf{y}\right\|_1}{\left\|\mathbf{y}\right\|_1}\right)\right), \end{aligned} \tag{7}$$

[2]The tolerance parameters of CVX are the second and fourth root of the machine precision by default [30, 31].

[3]This was the fastest method found by the authors. Other possibilities would be [15, Algorithm 2], [26].

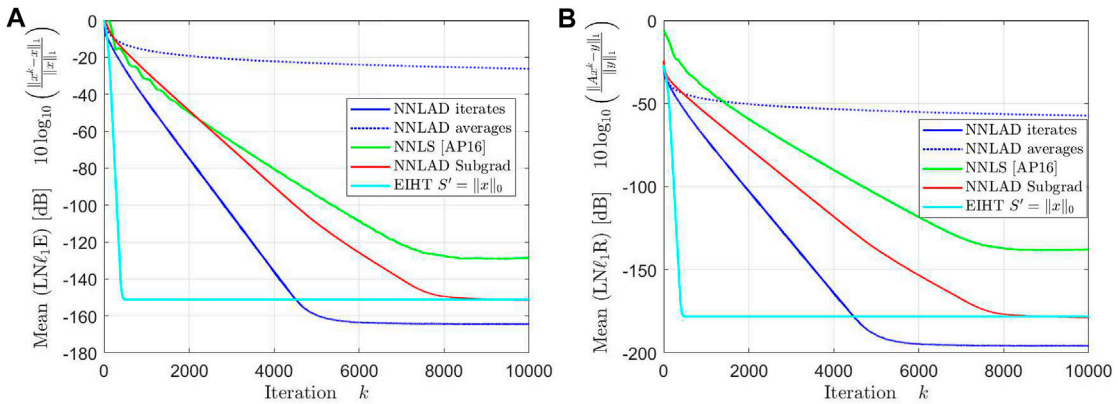

FIGURE 8 | Convergence rates of certain iterated methods with respect to the number of iterations. **A** is Bernoulli for NNLS and D-LRBG for the others.

over the sparsity and the time. The result can be found in **Figures 5**, **6**.

The averages of NNLAD converge significantly slower than the iterates, even though we lack a convergence rate for the iterates. We deduce that one should always use the iterates of NNLAD to recover a signal. Surprisingly, the averages converge even slower than the subgradient method. However, this is not because the averages converge slow, but rather because the subgradient method and all others converges faster than expected. In particular, the NNLAD iterates, EIHT and the NNLS all converge linearly toward the signal. Further, their correspoding objective values also converge linearly toward the optimal objective value. Even the subgradient method converges almost linearly. We deduce that the NNLS is the fastest of these methods if **A** is a D-LRBG.

Apart from a constant the NNLAD iterates, EIHT and NNLS converge in the same order. However, this behavior does not hold if we consider a different distribution for **A** as one can verify by setting each component $A_{m,n}$ as independent 0/1 Bernoulli random variables. While EIHT has better iterations compared to the NNLS, it still takes more time to achieve the same estimation errors and residuals. We plot the mean of the time required to calculate the first k iterations in **Figure 7**.

The EIHT requires roughly 6 times as long as any other method to calculate each iteration. All methods but the EIHT can be implemented with only two matrix vector multiplications, namely once by **A** and once by $\mathbf{A}^T$. Both of these requires roughly $2DN$ floating point operations. Hence, each iteration requires $\mathcal{O}(4DN)$ floating point operations. The EIHT only calculates one matrix vector multiplication, but also the median. This calculation is significantly slower than a matrix vector multiplication. For every $n \in [N]$ we need to order a vector with D elements, which can be performed in $\mathcal{O}(DlogD)$. Hence, each iteration of EIHT requires $\mathcal{O}(DNlogD)$ floating point operations, which explains why the EIHT requires significantly more time for each iteration.

As we have seen the NNLS is able to recover signals faster than any other method, however it also only obeys sub-optimal robustness guarantees for uniformly at random chosen D-LRBG as we have seen in **Figure 4A**. We ask ourself whether or not the NNLS is also faster with a more natural measurement scheme, i.e., if $A_{m,n}$ are independent 0/1 Bernoulli random variables. We repeat 2 100 times with $r = 2$ for the NNLS and $r = 1$ for the other methods. We again plot the mean of the logarithmic relative ℓ_1-estimation error and the mean of the relative ℓ_1-norm of the residual in **Figures 8**, **9**.

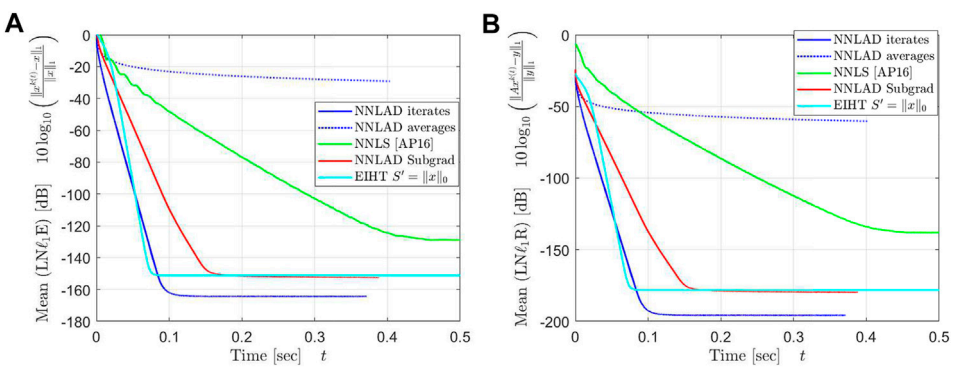

FIGURE 9 | Convergence rates of certain iterated methods with respect to the time. **A** is Bernoulli for NNLS and D-LRBG for the others.

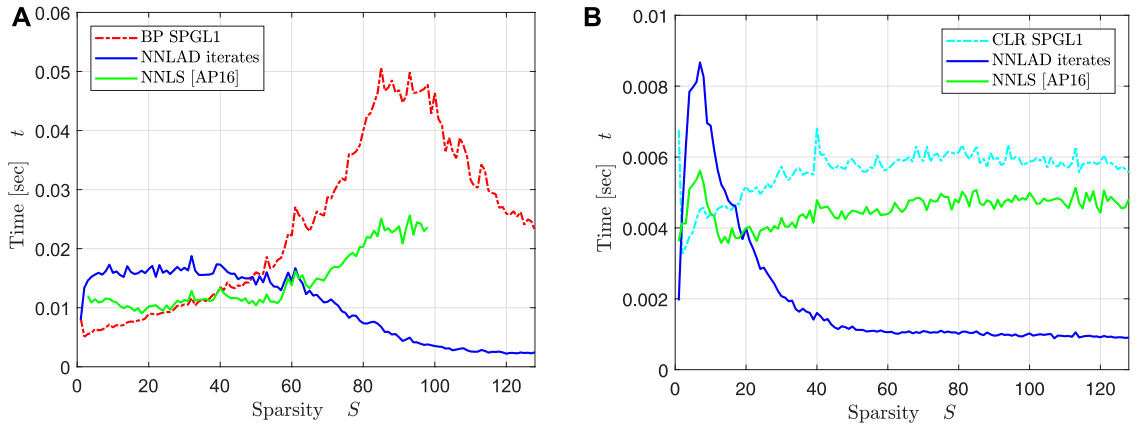

FIGURE 10 | Time of the NNLAD and NNLS to approximate better than SPGL methods. **(A)** The NNLAD is faster than the BP of SPGL1 for high S. **(B)**. The NNLAD is faster than the CLR of SPGL1 for moderate S.

The NNLAD and the EIHT converge to the solution with roughly the same time. Even the subgradient implementation of the NNLAD recovers a signal in less time than the NNLS. Further the convergence of NNLS does not seem to be linear anymore. We deduce that sparse structure of $\mathbf{A}$ has a more significant influence on the decoding time than the smoothness of the data fidelity term. Also we deduce that even the subgradient method is a viable choice to recover a signal.

5.2.2 Non-Negative Least Absolute Deviation Vs SPGL1

As a last test we compare the NNLAD to the SPGL1 [16, 34] toolbox for matlab.

Experiment 3

1. Draw the measurement matrix $\mathbf{A} \in \{0, D^{-1}\}^{M\times N}$ uniformly at random among all D-LRBG.
2. Generate the signal $\mathbf{x}$ uniformly at random from $\Sigma_S \cap \mathbb{R}_+^N \cap \mathbb{S}_r^{N-1}$.
3. Define the observation $\mathbf{y} := \mathbf{A}\mathbf{x}$.
4. Use a benchmark decoder to calculate an estimator $\mathbf{x}^\#$ and collect the relative estimation errors $\left\|\mathbf{x}^\# - \mathbf{x}\right\|_1 / \|\mathbf{x}\|_1, \left\|\mathbf{x}^\# - \mathbf{x}\right\|_2 / \|\mathbf{x}\|_2$ and the time to calculate $\mathbf{x}^\#$.
5. For each iterative method calculate iterations until $\left\|\mathbf{x}^k - \mathbf{x}\right\|_1 / \|\mathbf{x}\|_1 \leq \left\|\mathbf{x}^\# - \mathbf{x}\right\|_1 / ||\mathbf{x}||_1$ and $\left\|\mathbf{x}^k - \mathbf{x}\right\|_2 / \|\mathbf{x}\|_2 \leq \left\|\mathbf{x}^\# - \mathbf{x}\right\|_2 / ||\mathbf{x}||_2$. Collect the time to perform these iterations. If this threshold can not be reached after 10^5 iterations, the recovery failed and the time is set to ∞.

We again fix the dimension $N = 1024$, $M = 256$, $D = 10$ and vary $S \in [128]$. For both the BP implementation of SPGL1 and the CLR implementation of SPGL1 we repeat Experiment 3 100 times for each S. We plot the mean of the time to calculate the estimators and plot these over the sparsity in **Figures 10A,B**.

The NNLAD implementation is slower than both SPGL1 methods for small S. However, if we have the optimal number of measurements $M \in \mathcal{O}(S\log N/S)$, the NNLAD is faster than both SPGL1 methods.

5.2.3 Summary

The implementation of NNLAD as presented in **Algorithm 1** is a reliable recovery method for sparse non-negative signals. There are methods that might be faster, but these either recover a smaller number of coefficients (EIHT, greedy methods) or they obey sub-optimal recovery guarantees (NNLS). The implementation is as fast as the commonly uses SPGL1 toolbox, but has the advantage that it requires no tuning depending on the unknown $\mathbf{x}$ or $\mathbf{e}$. Lastly, the NNLAD can handle peaky noise overwhelmingly good.

5.3 Application for Viral Detection

With the outbreak and rapid spread of the COVID-19 virus we need to test a large amount of people for an infection. Since we can only test a fixed number of persons in a given time, the number of persons tested for the virus grows at most linearly. On the other hand, models suggest that the number of possibly infected persons grows exponentially. At some point, if that is not already the case, we will have a shortage of test kits and we will not be able to test every person. It is thus desirable to test as much persons with as few as possible test kits.

The field group testing develops strategies to test groups of individuals instead of individuals in order to reduce the amount of tests required to identify infected individuals. The first advances in group testing were made in [35]. For a general overview about group testing we refer to [36].

The problem of testing a large group for a virus can be modeled as a compressed sensing problem in the following way: Suppose we want to test N persons, labeled by $[N] = \{1, \ldots, N\}$, to check whether or not they are affected by a virus. We denote by x_n the quantity of viruses in the specimen of the nth person. Suppose we have M test kits, labeled by $[M] = \{1, \ldots, M\}$. By y_m we denote the amount of viruses in the sample of the mth test kit. Let $\mathbf{A} \in [0, 1]^{M\times N}$. For

every n we put a fraction of size $A_{m,n}$ of the specimen of the nth person into the sample for the mth test kit. The sample of the mth test kit will then have the quantity of viruses

$$\sum_{n\in[N]} A_{m,n}x_n + e_m^{con},$$

where e_m^{con} is the amount of viruses in the sample originating from a possible contamination of the sample. A quantitative reverse transcription polymerase chain reaction estimates the quantity of viruses by y_m with a small error $e_m^{pcr} = y_m - \sum_{n\in[N]} A_{m,n}x_n - e_m^{con}$. After all M tests we detect the quantity

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}, \tag{8}$$

where $\mathbf{e} = \mathbf{e}^{con} + \mathbf{e}^{pcr}$. Since contamination of samples happens rarely, $\mathbf{e}^{con}$ is assumed to be peaky in terms of **Table 1**, while $\mathbf{e}^{pcr}$ is assumed to have even mass but a small norm. In total $\mathbf{e}$ is peaky.

Often each specimen is tested separately, meaning that $\mathbf{A}$ is the identity. In particular, we need at least as much test kits as specimens. Further, we estimate the true quantity of viruses x_n by $x_n^{\#} := y_n$, which results in the estimation error $x_n^{\#} - x_n = e_n = e_n^{con} + e_n^{pcr}$. Since the noise vector $\mathbf{e}$ is peaky, some but few tests will be inaccurate and might result in false positives or false negatives.

In general, only a fraction of persons is indeed affected by the virus. Thus, we assume that $\|\mathbf{x}\|_0 \leq S$ for some small S. Since the amount of viruses is a non-negative value, we also have $\mathbf{x} \geq 0$. Hence, we can use the NNLR to estimate $\mathbf{x}$ and in particular we should use the NNLAD due to the noise being peaky. Corollary 3.6 suggests to choose $\mathbf{A}$ as the random walk matrix of a lossless expander or by ([3], Theorem 13.7) to choose $\mathbf{A}$ as a uniformly at random chosen D-LRBG. Such a matrix $\mathbf{A}$ has non-negative entries and the column sums of $\mathbf{A}$ are not greater than one. This is a necessary requirement since each column sum is the total amount of specimen used in the test procedure. Especially, a fraction of D^{-1} of each specimen is used in exactly D test kits.

By Corollary 3.6 and [[3], Theorem 13.7] this allows us to reduce the number of test kits required to $M \approx CS\log eN/S$. As we have seen in **Figures 4A,B** we expect the NNLAD estimator to correct the errors from $\mathbf{e}^{con}$ and the estimation error to be in the order of $||\mathbf{e}^{pcr}||_1$ which is assumed to be small. Hence, the NNLAD estimator with a random walk matrix of a lossless expander might even result in less false positives and false negatives than individual testing.

Note that the lack of knowledge about the noise $\mathbf{e}$ favors the NNLAD recovery method over a (BPDN) approach. Further, since the total sum of viruses in all patients given by $\sum_{n\in[N]} x_n = \|\mathbf{x}\|_1$ is unknown, it is undesirable to use (CLR).

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

BB and PJ proposed the research problem. HP derived the results of the paper, with feedback from BB and PJ. HP wrote the paper also with feedback from BB and PJ.

FUNDING

The work was partially supported by DAAD grant 57417688. PJ has been supported by DFG grant JU 2795/3. BB has been supported by BMBF through the German Research Chair at AIMS, administered by the Humboldt Foundation. We acknowledge support by the German Research Foundation and the Open Access Publication Funds of TU Berlin. This article has appeared as a preprint (27), see https://arxiv.org/abs/2003.13092.

REFERENCES

1. Candes EJ, Romberg J, and Tao T. Robust Uncertainty Principles: Exact Signal Reconstruction from Highly Incomplete Frequency Information. *IEEE Trans Inf Theory* (2006). 52:489–509. doi:10.1109/TIT.2005.862083
2. Donoho DL. Compressed Sensing. *IEEE Trans Inform Theor* (2006). 52: 1289–306. doi:10.1109/TIT.2006.871582
3. Foucart S, and Rauhut H. *A Mathematical Introduction to Compressive Sensing*. Basel, Switzerland: Birkhäuser (2013).
4. Hastie T, Tibshirani R, and Wainwright M. *Statistical Learning with Sparsity: The Lasso and Generalizations*. New York, NY: Chapman and Hall/CRC (2015).
5. Donoho DL, and Tanner J. Sparse Nonnegative Solution of Underdetermined Linear Equations by Linear Programming. *Proc Natl Acad Sci* (2005). 102: 9446–51. doi:10.1073/pnas.0502269102
6. Gilbert A, and Indyk P. Sparse Recovery Using Sparse Matrices. *Proc IEEE* (2010). 98:937–47. doi:10.1109/JPROC.2010.2045092
7. Bruckstein AM, Elad M, and Zibulevsky M. On the Uniqueness of Non-negative Sparse & Redundant Representations. *Proc IEEE Int Conf Acoust Speech Signal Process* (2008). 5145–8. doi:10.1109/ICASSP.2008.4518817
8. Donoho DL, and Tanner J. Counting the Faces of Randomly-Projected Hypercubes and Orthants, with Applications. *Discrete Comput Geom* (2010). 43:522–41. doi:10.1007/s00454-009-9221-z
9. Wang M, Xu W, and Tang A. A Unique "Nonnegative" Solution to an Underdetermined System: From Vectors to Matrices. *IEEE Trans Signal Process* (2011). 59:1007–1016. doi:10.1109/TSP.2010.2089624
10. Slawski M, and Hein M. Sparse Recovery by Thresholded Non-negative Least Squares. *Adv Neural Inf Process Syst* (2011). 24:1926–1934.
11. Slawski M, and Hein M. Non-negative Least Squares for High-Dimensional Linear Models: Consistency and Sparse Recovery without Regularization. *Electron J Stat* (2013). 7:3004–3056. doi:10.1214/13-EJS868
12. Kabanava M, Kueng R, Rauhut H, and Terstiege U. Stable Low-Rank Matrix Recovery via Null Space Properties. *Inf Inference* (2016). 5:405–441. doi:10.1093/imaiai/iaw014
13. Kueng R, and Jung P. Robust Nonnegative Sparse Recovery and the Nullspace Property of 0/1 Measurements. *IEEE Trans Inf Theory* (2018). 64:689–703. doi:10.1109/TIT.2017.2746620
14. Shadmi Y, Jung P, and Caire G. Sparse Non-negative Recovery from Biased Subgaussian Measurements Using NNLS. In: IEEE International Symposium on Information Theory; 2019 July 7–12; Paris, France (2019).

15. Chambolle A, and Pock T. A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging. *J Math Imaging Vis* (2011). 40:120–45. doi:10.1007/s10851-010-0251-1
16. Van Den Berg E, and Friedlander MP. Probing the Pareto Frontier for Basis Pursuit Solutions. *SIAM J Sci Comput* (2009). 31:890–912. doi:10.1137/080714488
17. Bühlmann P, and Van De Geer S. *Statistics for High-Dimensional Data*. Berlin, Heidelberg: Springer (2011).
18. Jafarpour S, Xu W, Hassibi B, and Calderbank R. Efficient and Robust Compressed Sensing Using Optimized Expander Graphs. *IEEE Trans Inf Theory* (2009). 55:4299–4308. doi:10.1109/TIT.2009.2025528
19. Xu W, and Hassibi B. Efficient Compressive Sensing with Deterministic Guarantees Using Expander Graphs. In: IEEE Information Theory Workshop; 2007 September 2–6; Tahoe City, CA, United States (2007). 414–9.
20. Berinde R, Gilbert AC, Indyk P, Karloff H, and Strauss MJ. Combining Geometry and Combinatorics: A Unified Approach to Sparse Signal Recovery. In: 46th Annual Allerton Conference on Communication; 2008 September 23–26; Monticello, IL, United States (2008). 798–805.
21. Khajehnejad MA, Dimakis AG, Xu W, and Hassibi B. Sparse Recovery of Nonnegative Signals with Minimal Expansion. *IEEE Trans Signal Process* (2011). 59:196–208. doi:10.1109/TSP.2010.2082536
22. Morgenshtern VI, and Candès EJ. Super-resolution of Positive Sources: The Discrete Setup. *SIAM J Imaging Sci* (2016). 9:412–44. doi:10.1137/15M1016552
23. Jaensch F, and Jung P. *Robust Recovery of Sparse Nonnegative Weights from Mixtures of Positive-Semidefinite Matrices*(Preprint). (2020). Available from: https://arxiv.org/abs/2003.12005.
24. Vadhan SP. Pseudorandomness. *FNT Theor Comp Sci* (2012). 7:1–336. doi:10.1561/0400000010
25. Nesterov Y *Introductory Lectures on Convex Optimization - A Basic CourseApplied Optimization*. New York, NY: Springer (2004).
26. Beck A, and Teboulle M. A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM J Imaging Sci* (2009). 2: 183–202. doi:10.1137/080716542
27. Attouch H, and Peypouquet J. The Rate of Convergence of Nesterov's Accelerated Forward-Backward Method Is Actually Faster Than $1/k2$. *SIAM J Optim* (2016). 26:1824–1834. doi:10.1137/15M1046095
28. Petersen HB, Bah B, and Jung P. Efficient Tuning-free L1-Regression of Nonnegative Compressible Signals. (2020). Available from: https://arxiv.org/abs/2003.13092. (Accessed May 10, 2021).
29. Kümmerle C. *Understanding and Enhancing Data Recovery Algorithms- From Noise-Blind Sparse Recovery to Reweighted Methods for Low-Rank Matrix Optimization*. [PhD dissertation]. München (Germany): Technical University of Munich (2019).
30. Grant M, and Boyd S. *CVX: Matlab Software for Disciplined Convex Programming* (2014).
31. Grant MC, and BOYd SP. Graph Implementations for Nonsmooth Convex Programs. In: V Blondel, S Boyd, and H Kimura, editors. *Recent Advances in Learning and Control. Lecture Notes in Control and Information Sciences*. New York, NY: Springer-Verlag Limited (2008). p. 95–110.
32. Dirksen S, Lecué G, and Rauhut H. On the Gap between Restricted Isometry Properties and Sparse Recovery Conditions. *IEEE Trans Inf TheorY* (2018). 64: 5478–87. doi:10.1109/TIT.2016.2570244
33. Polyak BT. *Introduction to Optimization (Translations Series in Mathematics and Engineering)*. New York, NY: Optimization Software, Inc (1987).
34. Van Den Berg E, and Friedlander MP. *SPGL1: A Solver for Large-Scale Sparse Reconstruction* (2019).
35. Dorfman R. The Detection of Defective Members of Large Populations. *Ann Math Statist* 14 (1943). 436–40. doi:10.1214/aoms/1177731363
36. Aldridge M, Johnson O, and Scarlett J. Group Testing: An Information Theory Perspective. *FNT Commun Inf Theory* (2019). 15:196–392. doi:10.1561/0100000099
37. Petersen HB, and Jung P. *Robust Instance-Optimal Recovery of Sparse Signals at Unknown Noise Levels*(Preprint) (2020). Available from: https://arxiv.org/abs/2008.08385.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

6 APPENDIX

6.1 Proof of Non-Negative Least Residual Recovery Guarantee

By $\mathbb{1}$ we denote the all ones vector in $\mathbb{R}^N$ or $\mathbb{R}^M$ respectively. The proof is an adaption of the steps used in [13]. As for most convex optimization problems in compressed sensing we use ([3], Theorem 4.25) and [[3], Theorem 4.20] respectively, which require $\mathbf{A}$ to have the RNSP.

Theorem 6.1 ([3], Theorem 4.25) and ([3], Theorem 4.20)). Let $q \in [1,\infty)$ and suppose $\mathbf{A}$ has the ℓ_q-RNSP of order S with respect to $\|\cdot\|$ with constants ρ and τ. Then, it holds that

$$\|\mathbf{x}-\mathbf{z}\|_q \le \frac{(1+\rho)^2}{1-\rho} S^{1/q-1}\left(\|\mathbf{z}\|_1 - \|\mathbf{x}\|_1 + 2d_1(\mathbf{x},\Sigma_S)\right) + \frac{3+\rho}{1-\rho}\tau\|\mathbf{A}(\mathbf{x}-\mathbf{z})\| \text{ for all } \mathbf{x},\mathbf{z}\in\mathbb{R}^n.$$

If $q = 1$, this bound can be improved to

$$\|\mathbf{x}-\mathbf{z}\|_1 \le \frac{1+\rho}{1-\rho}\left(\|\mathbf{z}\|_1 - ||\mathbf{x}||_1 + 2d_1(\mathbf{x},\Sigma_S)\right) + \frac{2}{1-\rho}\tau\|\mathbf{A}(\mathbf{x}-\mathbf{z})\| \text{ for all } \mathbf{x},\mathbf{z}\in\mathbb{R}^n.$$

Note that by a modification of the proof this result also holds for $q = \infty$. The modifications on the proofs of ([3], Theorem 4.25) and ([3], Theorem 4.20) are straight forward, only the modification of ([3], Theorem 2.5) might not be obvious. See also [37]. As a consequence, all our statements also hold for $q = \infty$ with $1/q := 0$. If $\mathbf{W} \in \mathbb{R}^{N\times N}$ is a diagonal matrix, we can calculate some operator norms fairly easy:

$$\|\mathbf{W}\|_{q\to q} := \sup_{||\mathbf{w}||_q\le 1} ||\mathbf{W}\mathbf{w}||_q = \max_{n\in[N]}|W_{n,n}| \text{ for all } q\in[1,\infty].$$

We use this relation at several places throughout this section. Furthermore, we use ([13], Lemma 5) without adaption. For the sake of completeness we add a short proof.

Lemma 6.2 ([13], Lemma 5). Let $q \in [1,\infty)$ and suppose that $\mathbf{A}\in\mathbb{R}^{M\times N}$ has ℓ_q-RNSP of order S with respect to $\|\cdot\|$ with constants ρ and τ. Let $\mathbf{W}\in\mathbb{R}^{N\times N}$ be a diagonal matrix with $W_{n,n} > 0$. If $\rho' = \|\mathbf{W}\|_{q\to q}\left\|\mathbf{W}^{-1}\right\|_{1\to 1}\rho < 1$, then $\mathbf{A}\mathbf{W}^{-1}$ has ℓ_q-RNSP of order S with respect to $\|\cdot\|$ with constants $\rho' = ||\mathbf{W}||_{q\to q}\left\|\mathbf{W}^{-1}\right\|_{p\to p}\rho$ and $\tau' = ||\mathbf{W}||_{q\to q}\tau$.

Proof Let $\mathbf{v}\in\mathbb{R}^N$ and $\#(T)\le S$. If we apply the RNSP of $\mathbf{A}$ for the vector $(\mathbf{W}^{-1}\mathbf{v})|_T$, we get

$$\begin{aligned}||\mathbf{v}|_T||_q &= \left\|\mathbf{W}\mathbf{W}^{-1}(\mathbf{v}|_T)\right\|_q \le \|\mathbf{W}\|_{q\to q}\left\|\mathbf{W}^{-1}(\mathbf{v}|_T)\right\|_q = ||\mathbf{W}||_{q\to q}\left\|(\mathbf{W}^{-1}\mathbf{v})\right|_T\|_q\\ &\le \|\mathbf{W}\|_{q\to q}\left(\rho S^{\frac{1}{q}-1}\left\|\mathbf{W}^{-1}(\mathbf{v}|_{T^c})\right\|_1 + \tau\left\|\mathbf{A}\mathbf{W}^{-1}\mathbf{v}\right\|\right)\\ &= ||\mathbf{W}||_{q\to q}\rho S^{\frac{1}{q}-1}\left\|\mathbf{W}^{-1}(\mathbf{v}|_{T^c})\right\|_1 + ||\mathbf{W}||_{q\to q}\tau\left\|\mathbf{A}\mathbf{W}^{-1}\mathbf{v}\right\|\\ &\le \|\mathbf{W}\|_{q\to q}\left\|\mathbf{W}^{-1}\right\|_{1\to 1}\rho S^{\frac{1}{q}-1}\|\mathbf{v}|_{T^c}\|_1 + \|\mathbf{W}\|_{q\to q}\tau\left\|\mathbf{A}\mathbf{W}^{-1}\mathbf{v}\right\|.\end{aligned}$$

This finishes the proof.

Next we adapt ([13], Theorem 4) to account for arbitrary norms. Further, we obtain a slight improvement in form of the dimensional scaling constant $S^{1/q-1}$. With this, our error bound becomes for $S\to\infty$ asymptotically the error bound of the basis pursuit denoising, whenever $\kappa = 1$ and $q > 1$ [3].

Proposition 6.3 (Similar to ([13], Theorem 4)). Let $q\in[1,\infty)$ and $\|\cdot\|$ be a norm on $\mathbb{R}^M$ with dual norm $\|\cdot\|_*$. Suppose $\mathbf{A}$ has ℓ_q-RNSP of order S with respect to $\|\cdot\|$ with constants ρ and τ. Suppose $\mathbf{A}$ has the M^+ criterion with vector $\mathbf{t}$ and constant κ and that $\kappa\rho < 1$. Then, we have

$$\begin{aligned}\|\mathbf{x}-\mathbf{z}\|_q \le{}& 2\frac{(1+\kappa\rho)^2}{1-\kappa\rho}\kappa S^{1/q-1}d_1(\mathbf{x},\Sigma_S)\\ &+\left(\frac{(1+\kappa\rho)^2}{1-\kappa\rho}S^{1/q-1}\max_{n\in[N]}\left|\left(\mathbf{A}^T\mathbf{t}\right)_n^{-1}\right|\|\mathbf{t}\|_* + \frac{3+\kappa\rho}{1-\kappa\rho}\kappa\tau\right)\\ &\times||\mathbf{A}\mathbf{z}-\mathbf{A}\mathbf{x}|| \text{ for all } \mathbf{x},\mathbf{z}\in\mathbb{R}_+^N.\end{aligned}$$

If $q = 1$, this bound can be improved to

$$\begin{aligned}||\mathbf{x}-\mathbf{z}||_q \le{}& 2\frac{1+\kappa\rho}{1-\kappa\rho}\kappa d_1(\mathbf{x},\Sigma_S)\\ &+\left(\frac{1+\kappa\rho}{1-\kappa\rho}\max_{n\in[N]}\left|\left(\mathbf{A}^T\mathbf{t}\right)_n^{-1}\right||\mathbf{t}||_* + \frac{2}{1-\kappa\rho}\kappa\tau\right)\\ &\times||\mathbf{A}\mathbf{z}-\mathbf{A}\mathbf{x}|| \text{ for all } \mathbf{x},\mathbf{z}\in\mathbb{R}_+^N.\end{aligned}$$

Proof Let $\mathbf{x},\mathbf{z}\ge 0$. In order to apply Lemma 6.2 we set $\mathbf{W}$ as the matrix with diagonal $\mathbf{A}^T\mathbf{t}$ and zero else. It follows that $W_{n,n} > 0$ and $\|\mathbf{W}\|_{q\to q}\left\|\mathbf{W}^{-1}\right\|_{1\to 1}\rho = \kappa\rho < 1$. We can apply Lemma 6.2, which yields that $\mathbf{A}\mathbf{W}^{-1}$ has ℓ_q-RNSP with constants $\rho' = \|\mathbf{W}\|_{q\to q}\left\|\mathbf{W}^{-1}\right\|_{1\to 1}\rho = \kappa\rho$ and $\tau' = \|\mathbf{W}\|_{q\to q}\tau = \max_{n\in[N]}\left|\left(\mathbf{A}^T\mathbf{t}\right)_n\right|\tau$. We apply Theorem 6.1 with the matrix $\mathbf{A}\mathbf{W}^{-1}$, the vectors $\mathbf{W}\mathbf{x}$, $\mathbf{W}\mathbf{z}$ and the constants ρ' and τ' and get

$$\begin{aligned}||\mathbf{W}\mathbf{x}-\mathbf{W}\mathbf{z}||_q &\le \frac{(1+\rho')^2}{1-\rho'}S^{1/q-1}\left(||\mathbf{W}\mathbf{z}||_1 - \|\mathbf{W}\mathbf{x}\|_1\right.\\ &\left.+2d_1(\mathbf{W}\mathbf{x},\Sigma_S)\right) + \frac{3+\rho'}{1-\rho'}\tau'\left\|\mathbf{A}\mathbf{W}^{-1}(\mathbf{W}\mathbf{x}-\mathbf{W}\mathbf{z})\right\|\\ &\le \frac{(1+\rho')^2}{1-\rho'}S^{1/q-1}\left(||\mathbf{W}\mathbf{z}||_1 - ||\mathbf{W}\mathbf{x}||_1 + 2\|\mathbf{W}\|_{1\to 1}d_1(\mathbf{x},\Sigma_S)\right)\\ &+\frac{3+\rho'}{1-\rho'}\tau'||\mathbf{A}\mathbf{x}-\mathbf{A}\mathbf{z}||\\ &= 2\frac{(1+\kappa\rho)^2}{1-\kappa\rho}\max_{n\in[N]}\left|\left(\mathbf{A}^T\mathbf{t}\right)_n\right|S^{1/q-1}d_1(\mathbf{x},\Sigma_S)\\ &+\frac{(1+\kappa\rho)^2}{1-\kappa\rho}S^{1/q-1}\left(\|\mathbf{W}\mathbf{z}\|_1 - ||\mathbf{W}\mathbf{x}||_1\right)\\ &+\frac{3+\kappa\rho}{1-\kappa\rho}\max_{n\in[N]}\left|\left(\mathbf{A}^T\mathbf{t}\right)_n\right|\tau\|\mathbf{A}\mathbf{x}-\mathbf{A}\mathbf{z}\|.\end{aligned}$$

We lower bound the left hand side further to get

$$\begin{aligned}\|\mathbf{x}-\mathbf{z}\|_q &\le \left\|\mathbf{W}^{-1}\right\|_{q\to q}\|\mathbf{W}\mathbf{x}-\mathbf{W}\mathbf{z}\|_q = \max_{n\in[N]}\left|\left(\mathbf{A}^T\mathbf{t}\right)_n^{-1}\right|||\mathbf{W}\mathbf{x}-\mathbf{W}\mathbf{z}||_q\\ &\le 2\frac{(1+\kappa\rho)^2}{1-\kappa\rho}\kappa S^{1/q-1}d_1(\mathbf{x},\Sigma_S) + \frac{(1+\kappa\rho)^2}{1-\kappa\rho}S^{1/q-1}\max_{n\in[N]}\left|\left(\mathbf{A}^T\mathbf{t}\right)_n^{-1}\right|\\ &\left(||\mathbf{W}\mathbf{x}||_1 - \|\mathbf{W}\mathbf{z}\|_1\right) + \frac{3+\kappa\rho}{1-\kappa\rho}\kappa\tau\|\mathbf{A}\mathbf{x}-\mathbf{A}\mathbf{z}\|. \qquad (5)\end{aligned}$$

We want to estimate the term $\|\mathbf{Wx}\|_1 - \|\mathbf{Wz}\|_1$ using the M^+ criterion. Since $\mathbf{z}, \mathbf{x} \geq 0$, $W_{n,n} = (\mathbf{A}^T\mathbf{t})_n > 0$ and $\mathbf{W}$ is a diagonal matrix, we have

$$\begin{aligned}||\mathbf{Wz}||_1 - ||\mathbf{Wx}||_1 &= \langle \mathbb{1}, \mathbf{Wz}\rangle - \langle \mathbb{1}, \mathbf{Wx}\rangle = \langle \mathbf{W}^T\mathbb{1}, \mathbf{z} - \mathbf{x}\rangle \\ &= \langle \mathbf{W}\mathbb{1}, \mathbf{z} - \mathbf{x}\rangle = \langle \mathbf{t}, \mathbf{A}(\mathbf{z} - \mathbf{x})\rangle \leq \|\mathbf{t}\|_* \|\mathbf{Az} - \mathbf{Ax}\|.\end{aligned}$$

Applying this to **Eq. 5** we get

$$\begin{aligned}||\mathbf{x} - \mathbf{z}||_q &\leq 2\frac{(1+\kappa\rho)^2}{1-\kappa\rho}\kappa S^{1/q-1} d_1(\mathbf{x}, \Sigma_S) \\ &+ \left(\frac{(1+\kappa\rho)^2}{1-\kappa\rho} S^{1/q-1} \max_{n\in[N]} \left|(\mathbf{A}^T\mathbf{t})_n^{-1}\right| \|\mathbf{t}\|_* + \frac{3+\kappa\rho}{1-\kappa\rho}\kappa\tau\right) ||\mathbf{Az} - \mathbf{Ax}||.\end{aligned}$$

If $q = 1$ we can repeat the proof with the improved bound of Theorem 6.1.

After these auxiliary statements it remains to prove the main result of **Section 3** about the properties of the NNLR minimizer.

Proof of Theorem 3.4. By applying Proposition 6.3 with $\mathbf{x}$ and $\mathbf{z} := \mathbf{x}^\# \geq 0$ we get

$$\begin{aligned}\left\|\mathbf{x} - \mathbf{x}^\#\right\|_q &\leq 2\frac{(1+\kappa\rho)^2}{1-\kappa\rho}\kappa S^{1/q-1} d_1(\mathbf{x}, \Sigma_S) + \left(\frac{(1+\kappa\rho)^2}{1-\kappa\rho} S^{1/q-1}\right. \\ &\left. \times \max_{n\in[N]} \left|(\mathbf{A}^T\mathbf{t})_n^{-1}\right| ||\mathbf{t}||_* + \frac{3+\kappa\rho}{1-\kappa\rho}\kappa\tau\right) \left|\left|\mathbf{Ax}^\# - \mathbf{Ax}\right|\right| \\ &\leq 2\frac{(1+\kappa\rho)^2}{1-\kappa\rho}\kappa S^{1/q-1} d_1(\mathbf{x}, \Sigma_S) + \left(\frac{(1+\kappa\rho)^2}{1-\kappa\rho} S^{1/q-1}\right. \\ &\left. \times \max_{n\in[N]} \left|(\mathbf{A}^T\mathbf{t})_n^{-1}\right| \|\mathbf{t}\|_* + \frac{3+\kappa\rho}{1-\kappa\rho}\kappa\tau\right) \left(\left\|\mathbf{Ax}^\# - \mathbf{y}\right\| + ||\mathbf{Ax} - \mathbf{y}||\right) \\ &\leq 2\frac{(1+\kappa\rho)^2}{1-\kappa\rho}\kappa S^{1/q-1} d_1(\mathbf{x}, \Sigma_S) \\ &+ 2\left(\frac{(1+\kappa\rho)^2}{1-\kappa\rho} S^{1/q-1} \max_{n\in[N]} \left|(\mathbf{A}^T\mathbf{t})_n^{-1}\right| \|\mathbf{t}\|_* + \frac{3+\kappa\rho}{1-\kappa\rho}\kappa\tau\right) ||\mathbf{Ax} - \mathbf{y}||,\end{aligned}$$

where in the last step we used that $\mathbf{x}^\#$ is a minimizer and $\mathbf{x}$ is feasible. If $q = 1$, we can repeat the proof with the improved bound of Proposition 6.3.