Implicit ℓ^1-regularization of positively quadratically reparameterized linear regression: precise upper and lower bounds

Hannes Matt
Mathematical Institute for
Data Science and Machine Learning
Catholic University Eichstätt-Ingolstadt
Auf der Schanz 49, 85049 Ingolstadt, Germany
hannes.matt@ku.de

Dominik Stöger
Mathematical Institute for
Data Science and Machine Learning
Catholic University Eichstätt-Ingolstadt
Auf der Schanz 49, 85049 Ingolstadt, Germany
dominik.stoeger@ku.de

***Abstract*—Modern neural networks are often trained in a setting where the number of parameters vastly exceeds the number of training samples. While statistical folklore might suggest overfitting due to the huge capacity of these models, they show remarkable performance in practice, even if no regularization is applied at all. To explain this phenomenon, it has been conjectured that the training algorithm itself is biased towards models of low capacity by implicitly regularizing the model. While such an explanation remains elusive for deep neural networks, significant progress has been made for simpler models.**

In order to understand the implicit regularization of gradient flow, diagonal linear neural networks have been studied extensively. It was observed that for a sufficiently small initialization, gradient flow converges towards the model with almost smallest ℓ^1-norm among all models that perfectly interpolate the training data.

In this work, we study positive diagonal linear neural networks of depth $D = 2$ in a regression task (a.k.a. quadratically reparameterized linear regression). We analyze the approximation error between the limit of the gradient flow and the solution of the ℓ^1-minimization problem. We derive precise upper and lower bounds on the approximation error in dependence of the scale of initialization α: the error decays with rate $\alpha^{1-\varrho}$, where $\varrho < 1$. Furthermore, ϱ can be explicitly characterized and is closely related to quantities prominent in the field of compressive sensing. Our upper bounds improve on previous work in the literature, and, to the best of our knowledge, no lower bounds were available before.

I. Introduction

Neural networks are trained in an overparameterized setting, where the number of parameters is much larger than the number of data points in the training dataset. However, it is not fully understood why these networks generalize so well [1]. Namely, by the abundance of degrees of freedom, there exist an infinitude of models that perfectly fit the training data, some of which generalize poorly. Yet, it has been observed that gradient descent finds a model that also generalizes well, even when the loss is not explicitly regularized. This has led to the conjecture that the optimization algorithm itself induces this regularization and thus favors models of low complexity. This hypothesis is commonly referred to as *implicit regularization*.

While the implicit regularization phenomenon has not yet been understood for general neural networks, significant progress has been made in several simpler models, such as diagonal linear neural networks [2] and low-rank matrix sensing [3] [4] [5]. In the case of linear neural networks, it was shown that gradient flow starting at a sufficiently small initialization is implicitly biased towards a minimizer of the training loss that has minimal ℓ^1-norm [6].

In this work, we investigate positive diagonal linear neural networks of depth 2 in the regression setting. We consider the linear regression problem

$$\mathcal{L}(x) = \sum_{i=1}^{N} \left(\langle a_i, x\rangle - y_i\right)^2 = \|Ax - y\|_{\ell^2}^2 ,$$

where $A \in \mathbb{R}^{N\times d}$ is a given design matrix with rows $(a_i)_{i=1}^N$ and $y \in \mathbb{R}^N$ is a target vector. The model parameter x is then reparameterized as $x(u) := u^{\odot 2}$, where we denote by $\odot$ the Hadamard product $(v \odot w)_i := v_i w_i$. By design, the effective linear model $x(u)$ is non-negative. We notice that while the model is linear in the input a, the reparameterized loss

$$\widetilde{\mathcal{L}}(u) := \mathcal{L}(u^{\odot 2}) = \left\|Au^{\odot 2} - y\right\|_{\ell^2}^2$$

is non-linear and non-convex in the model parameter u.

In the case of linear regression, it is well-known that gradient flow of $\mathcal{L}$ initialized at 0 converges to a solution of $Ax = y$ with minimal ℓ^2-norm. In contrast, for diagonal linear neural networks of depth at least two, it was shown that for sufficiently small initialization, gradient flow is biased towards a solution with minimal ℓ^1-norm: when training $\hat{x}(u(t))$ using gradient flow for the reparameterized loss function $\widetilde{\mathcal{L}}$, the effective linear model $\hat{x}(u(t))$ is biased towards

$$x^* \in \operatorname*{arg\,min}_{x\geq 0: Ax=y} \|x\|_{\ell^1} .$$

This implicit bias can be made rigorous using the notion of *Bregman divergence*. For a strictly convex function $\Phi\colon \mathbb{R}^d \to \mathbb{R}$, the Bregman divergence is defined as

$$B_\Phi(x, x_0) := \Phi(x) - \Phi(x_0) - \langle \nabla\Phi(x_0), x - x_0\rangle.$$

A point $\tilde{x}$ is said to be the *Bregman projection* of x_0 onto a set Ω if $\tilde{x} \in \arg\min_{x\in\Omega} B_\Phi(x, x_0)$.

The implicit bias can now be derived in two steps. Firstly, it is shown that for every fixed initialization, the gradient flow converges to the Bregman projection of the initialization onto the set of solutions [7]. Secondly, it is shown that as the initialization vanishes, the Bregman projection converges to the ℓ^1 minimizer x^*.

To make things rigorous, for a *scale of initialization* $\alpha > 0$, denote by $(u_t(\alpha))_{t\geq 0}$ the gradient flow of $\tilde{\mathcal{L}}$ starting at $u_0(\alpha) := \alpha\mathbb{1}$. Here $\mathbb{1}$ is the vector where every entry is equal to one. Furthermore, let $x_t(\alpha) := u_t(\alpha)^{\odot 2}$ denote the effective linear model and let $x^\infty(\alpha) := \lim_{t\to\infty} x_t(\alpha)$.

Regarding the first step, it was shown in [7] [8, Theorem 2.2] that

$$x^\infty(\alpha) = \operatorname*{arg\,min}_{x\geq 0: Ax=y} B_E(x, \mathbb{1}\alpha),$$

where the Bregman potential is the entropy functional

$$E(x) := \sum_{i=1}^d x_i \log(x_i) - x_i,$$

with $0\log(0) := 0$.

For the second step, it was shown that

$$\|x^\infty(\alpha)\|_{\ell^1} - \|x^*\|_{\ell^1} \tag{I.1}$$

or

$$\|x^* - x^\infty(\alpha)\|_{\ell^1} \tag{I.2}$$

converge to 0 as $\alpha \to 0$. To achieve this, suitable upper bounds on these two quantities were derived. In [6] [8], it was shown that (I.1) converges to 0 with rate $\mathcal{O}\big(\log(1/\alpha)\big)$. This has been improved in [9], who show that (I.2) converges to 0 with a rate $\mathcal{O}\big(\alpha^p\big)$, where $p \in (0,1)$ is an undetermined constant.

In this paper, we derive upper and lower bounds for (I.2). These results show that the quantity (I.2) converges to 0 with rate $\alpha^{1-\varrho}$, where $\varrho \in (0,1)$ is an explicitly specified constant, which is closely related to null-space constants in the theory of compressive sensing [10]. Our upper bound improves on existing results. To the best of our knowledge, no lower bounds were known so far.

In a forthcoming work, [11], we generalize these upper and lower bounds to the reparameterization $x(u,v) := u^{\odot D} - v^{\odot D}$ for any $D \geq 2$. Moreover, we construct explicit examples that show that the upper and lower bounds are asymptotically sharp.

II. Main Results

To state our main results, we make the following assumptions. For $x \in \mathbb{R}^d$, we write $x \geq 0$ if $x_i \geq 0$ for all $i \in [d]$.

Assumption II.1. Let $A \in \mathbb{R}^{N\times d}$ and $y \in \mathbb{R}^N$. We assume that

(a) $y \neq 0$,
(b) there exist $x, x' \geq 0$ with $x \neq x'$ and $Ax = Ax' = y$,
(c) and there is a unique minimizer x^* of the minimization problem $\min_{x\geq 0: Ax=y} \|x\|_{\ell^1}$.

Notice that if assumption (b) does not hold, then $\mathcal{L}$ only has one positive global minimizer. In that case, the question of implicit bias is meaningless. We note that regarding assumption (c), that our theory can be generalized to the case of a non-unique minimizer. However, the theoretical result becomes more technical. For a full discussion, we refer to our forthcoming paper [11].

In order to state our main results, we need to quantify some properties of x^* and $\ker(A)$. Let $\mathcal{S} := \{i \in [d] : x_i^* \neq 0\}$ denote the support of x^* and let $\mathcal{S}^c := [d] \setminus \mathcal{S}$. Furthermore, let

$$\kappa_* := \frac{\max_{i\in\mathcal{S}} x_i^*}{\min_{i\in\mathcal{S}} x_i^*}.$$

Moreover, for $n \in \mathbb{R}^d$ and $T \subset [d]$, write $n_T := (n_i)_{i\in T}$ and define

$$\mathcal{N} := \big\{n \in \ker(A) : n_{\mathcal{S}^c} \geq 0\big\}$$

and let

$$\varrho := \sup_{0\neq n\in\mathcal{N}} \frac{-\sum_{i\in\mathcal{S}} n_i}{\|n_{\mathcal{S}^c}\|_{\ell^1}} \tag{II.1}$$

and

$$\tilde{\varrho} := \sup_{0\neq n\in\mathcal{N}} \frac{\|n_{\mathcal{S}}\|_{\ell^1}}{\|n_{\mathcal{S}^c}\|_{\ell^1}}, \quad \varrho^- := \sup_{0\neq n\in\mathcal{N}} \frac{\sum_{i\in\mathcal{S}: n_i<0} |n_i|}{\|n_{\mathcal{S}^c}\|_{\ell^1}}, \tag{II.2}$$

whenever they exist. Lemma III.1 below ensures that the quantities introduced above make sense and that $\varrho < 1$.

With these definitions in place, we can state our main result.

Theorem II.2. *Let A, y and x^* as in Assumption II.1. Let $\alpha > 0$ and let*

$$x^\infty(\alpha) \in \operatorname*{arg\,min}_{x\geq 0: Ax=y} B_E(x, \alpha\mathbb{1}).$$

*(a) **Upper bound.** If*

$$0 < \alpha < \min_{i\in\mathcal{S}} |x_i^*|,$$

then

$$\frac{\|x^\infty(\alpha) - x^*\|_{\ell^1}}{\alpha^{1-\varrho}} \leq (1+\tilde{\varrho})\,|\mathcal{S}^c|\,\kappa_*^{\varrho^-}\,(\min_{i\in\mathcal{S}} x_i^*)^\varrho. \tag{II.3}$$

*(b) **Lower bound**. If*

$$\Big(\frac{\alpha}{\min_{i\in\mathcal{S}} x_i^*}\Big)^{1-\varrho} \leq \frac{1}{2(1+\tilde{\varrho})\,|\mathcal{S}^c|\,\kappa_*^{\varrho^-}},$$

then

$$\frac{\|x^*\|_{\ell^\infty}^\varrho}{\kappa_*^{\varrho^-}} \cdot (1 - C\cdot\varepsilon^{1-\varrho}) \leq \frac{\|x^\infty(\alpha)_{\mathcal{S}^c} - x^*_{\mathcal{S}^c}\|_{\ell^\infty}}{\alpha^{1-\varrho}}, \tag{II.4}$$

where $\varepsilon := \frac{\alpha}{\min_{i\in\mathcal{S}} x_i^*}$ and $C := 2(1+\tilde{\varrho})\,|\mathcal{S}^c|\,\kappa_*^{\varrho^-}$.

Before we proceed to the proof, let us make some remarks.

Remark. Note that by the equivalence of the ℓ^1-norm and the ℓ^∞-norm, our result Theorem II.2 implies that if A and y are fixed, then we have

$$c \leq \frac{\|x^\infty(\alpha) - g^*\|_{\ell^1}}{\alpha^{1-\varrho}} \leq C \quad \text{as } \alpha \downarrow 0$$

for some constants

$$c := \frac{\|x^*\|_{\ell^\infty}^{\varrho}}{\kappa_*^{\varrho^-}} \quad \text{and} \quad C := (1+\tilde{\varrho})\,|\mathcal{S}^c|\,\kappa_*^{\varrho^-}\,(\min_{i\in\mathcal{S}} x_i^*)^{\varrho}$$

that only depend on A and y. In particular, the convergence rate is proportional to $\alpha^{1-\varrho}$ and is completely determined by the null-space parameter ϱ. Moreover, we expect that the constants c and C are optimal in an asymptotic sense. For a more detailed discussion of this topic, we refer to our forthcoming work [11].

Remark. In practical applications, the dimension d is often rather high. For example, in sparse MRI, the dimension d may be around 10^5, whereas the sparsity s may be around 10^3, see, e.g., [12]. We refer to [10] for some more references to applications.

Remark. In our upcoming work [11], we conducted numerical experiments with synthetic data, where A is a random Gaussian matrix. For highly sparse signals x^* and measurements without noise, ϱ is often substantially smaller than 1. In fact, one can show that $\varrho \lesssim \sqrt{\frac{s\log(N/s)}{N}}$, see, e.g., [10]. In the presence of noise, however, we observed in the experiments that ϱ is close to 1. In summary, our experiments show that in the noiseless scenario, as the scale of initialization converges to 0, the convergence of $x^\infty(\alpha)$ towards the ℓ^1-minimizer is much faster than in the noisy scenario.

III. Proofs

A. Technical preliminaries

The following Lemma III.1 is proved in Section III-C.

Lemma III.1. *Assume that A, y fulfill Assumption II.1.*

(i) We have $\mathcal{S} \neq \emptyset$, $\mathcal{S}^c \neq \emptyset$, and $\mathcal{N} \neq \{0\}$. Furthermore, for every $\tilde{n} \in \mathcal{N} \setminus \{0\}$ we have $\tilde{n}_{\mathcal{S}^c} \neq 0$. In particular, the null-space constants in (II.1) *and* (II.2) *are well-defined.*

(ii) We have $-\infty < \varrho < 1$ and $0 \leq \tilde{\varrho}, \varrho^- < \infty$. Furthermore, the suprema in (II.1) *and* (II.2) *are attained.*

The Bregman minimizer x^∞ has maximal support in the sense of the following Lemma III.2. This is proved in [9].

Lemma III.2. *For all $\alpha > 0$, all $\tilde{n} \in \mathcal{N}$ and all $i \in [d]$ we have: if $\tilde{n}_i \neq 0$, then $x_i^\infty(\alpha) \neq 0$.*

B. Proof of the main theorem

Let $n(\alpha) := x^\infty(\alpha) - x^*$. In the following, we will conceal the dependency on α and simply write x^∞ and n. The proofs of the upper bound for $\|n\|_{\ell^1}$ and the lower bound for $\|n\|_{\ell^\infty}$ in Theorem II.2 are both based on the first order optimality condition for the Bregman divergence. The following Lemma III.3 distills this condition down to the relevant equality.

Lemma III.3. *(i) We have $n \in \mathcal{N}$ and $n_{\mathcal{S}^c} \neq 0$.*
(ii) For all $\tilde{n} \in \mathcal{N}$, we have

$$-\sum_{i\in\mathcal{S}} \tilde{n}_i \log\Big(\frac{x_i^* + n_i}{\alpha}\Big) = \sum_{i\in\mathcal{S}^c} \tilde{n}_i \log\Big(\frac{n_i}{\alpha}\Big).$$

Proof. (i) By Lemma III.1(i) there exists $\tilde{n} \in \mathcal{N} \setminus \{0\}$, and, in addition, we have $\tilde{n}_{\mathcal{S}^c} \neq 0$. It follows from Lemma III.2 that $x^\infty_{\mathcal{S}^c} \neq 0$ and so $n_{\mathcal{S}^c} = (x^\infty - x^*)_{\mathcal{S}^c} = x^\infty_{\mathcal{S}^c} \neq 0$.

(ii) Let $\mathscr{L}_+ := \{x \geq 0 : Ax = y\}$. By Lemma III.2 we have $x_i^\infty > 0$ for all i such that $\tilde{n}_i \neq 0$. Therefore, the map $t \mapsto D_E(x^\infty + t\tilde{n}, \alpha\mathbb{1})$ is differentiable at $t = 0$. Furthermore, $x^\infty + t\tilde{n} \in \mathscr{L}_+$ for all $t \in \mathbb{R}$ with $|t|$ sufficiently small. The optimality of x^∞ implies that

$$\begin{aligned} 0 &= \frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} B_E(x^\infty + t\tilde{n}, \alpha\mathbb{1}) \\ &= \langle \nabla_x\big|_{x=x^\infty} B_E(x, \alpha\mathbb{1}), \tilde{n}\rangle \\ &= \langle \nabla E(x^\infty) - \nabla E(\alpha\mathbb{1}), \tilde{n}\rangle \\ &= \sum_{i=1}^d \tilde{n}_i \log\Big(\frac{x_i^\infty}{\alpha}\Big), \end{aligned}$$

where we use the convention $0 \cdot \log(0) = 0$. Splitting up the indices into $\mathcal{S}$ and $\mathcal{S}^c$, inserting $x^\infty = x^* + n$, and using $x^*_{\mathcal{S}^c} = 0$, we deduce the claim. □

With this Lemma III.3 in place, we can prove the main result. We first show the upper bound.

Proof of Theorem II.2(a). Since $n \in \mathcal{N} \setminus \{0\}$ we have by definition of $\tilde{\varrho}$ that

$$\begin{aligned} \|x^\infty - x^*\|_{\ell^1} = \|n\|_{\ell^1} &= \|n_{\mathcal{S}^c}\|_{\ell^1} + \|n_{\mathcal{S}}\|_{\ell^1} \\ &\leq (1+\tilde{\varrho}) \cdot \|n_{\mathcal{S}^c}\|_{\ell^1}\,. \end{aligned} \tag{III.1}$$

In the remainder of the proof, we will derive a suitable upper bound for $\|n_{\mathcal{S}^c}\|_{\ell^1}$.

Invoking Lemma III.3, with $\tilde{n} := n$, we obtain

$$\sum_{i\in\mathcal{S}^c} n_i \log\Big(\frac{n_i}{\alpha}\Big) = \sum_{i\in\mathcal{S}} (-n_i) \log\Big(\frac{x_i^* + n_i}{\alpha}\Big). \tag{III.2}$$

In the following, we will bound the left-hand side and the right-hand side of (III.2) individually.

For the right-hand side of (III.2), we use the monotonicity of $t \mapsto \log\left(\frac{x_i^*+t}{\alpha}\right)$ to obtain

$$\sum_{i\in\mathcal{S}} (-n_i) \log\Big(\frac{x_i^* + n_i}{\alpha}\Big) \leq \sum_{i\in\mathcal{S}} (-n_i) \log\Big(\frac{x_i^*}{\alpha}\Big).$$

Now let $\lambda := \min_{i\in\mathcal{S}} x_i^*$. Using $\log(\frac{x_i^*}{\lambda}) \geq 0$ at (i), and the definitions of the null-space constants at (ii), we infer that

$$\begin{aligned} &\sum_{i\in\mathcal{S}} (-n_i) \log\Big(\frac{x_i^*}{\alpha}\Big) \\ &= \sum_{i\in\mathcal{S}} (-n_i) \log\Big(\frac{\lambda}{\alpha}\Big) + \sum_{i\in\mathcal{S}} (-n_i) \log\Big(\frac{x_i^*}{\lambda}\Big) \\ &\overset{(i)}{\leq} \log\Big(\frac{\lambda}{\alpha}\Big)\Big(-\sum_{i\in\mathcal{S}} n_i\Big) + \sum_{i\in\mathcal{S}:n_i<0} |n_i| \log\Big(\frac{x_i^*}{\lambda}\Big) \\ &\overset{(ii)}{\leq} \|n_{\mathcal{S}^c}\|_{\ell^1} \Big[\log\Big(\frac{\lambda}{\alpha}\Big) \cdot \varrho + \log\Big(\frac{\sup_{i\in\mathcal{S}} x_i^*}{\lambda}\Big) \cdot \varrho^-\Big]. \end{aligned} \tag{III.3}$$

For the term on the right-hand side of (III.2), we invoke the log sum inequality, see [13, Theorem 2.7.1], to deduce that

$$\begin{aligned}\sum_{i\in\mathcal{S}^c} n_i \log\Big(\frac{n_i}{\alpha}\Big) &\geq \Big(\sum_{i\in\mathcal{S}^c} n_i\Big)\cdot\log\Big(\frac{1}{\alpha\,|\mathcal{S}^c|}\sum_{i\in\mathcal{S}^c} n_i\Big)\\ &= \|n_{\mathcal{S}^c}\|_{\ell^1}\cdot\log\Big(\frac{\|n_{\mathcal{S}^c}\|_{\ell^1}}{\alpha\,|\mathcal{S}^c|}\Big).\end{aligned} \tag{III.4}$$

Inserting (III.3) and (III.4) into (III.2), and dividing both sides by $\|n_{\mathcal{S}^c}\|_{\ell^1}$, which by Lemma III.3 is non-zero, we obtain

$$\log\Big(\frac{\|n_{\mathcal{S}^c}\|_{\ell^1}}{\alpha\,|\mathcal{S}^c|}\Big) \leq \varrho\log\Big(\frac{\lambda}{\alpha}\Big) + \varrho^-\log\Big(\frac{\sup_{i\in\mathcal{S}} x_i^*}{\lambda}\Big).$$

Applying the exponential function, we deduce that

$$\|n_{\mathcal{S}^c}\|_{\ell^1} \leq \alpha\,|\mathcal{S}^c|\cdot\Big(\frac{\lambda}{\alpha}\Big)^{\varrho}\cdot\Big(\frac{\sup_{i\in\mathcal{S}} x_i^*}{\lambda}\Big)^{\varrho^-}.$$

Since $\lambda = \min_{i\in\mathcal{S}} x_i^*$, it follows that

$$\|n_{\mathcal{S}^c}\|_{\ell^1} \leq \alpha^{1-\varrho}\,|\mathcal{S}^c|\cdot\big(\min_{i\in\mathcal{S}} x_i^*\big)^{\varrho}\cdot\kappa_*^{\varrho^-}. \tag{III.5}$$

Combining (III.1) and (III.5), we deduce (II.3). □

Next, we prove the lower bound.

Proof of Theorem II.2(b). By Lemma III.1(ii), there exists $m\in\ker(A)$ with $m_{\mathcal{S}^c}\geq 0$, $m_{\mathcal{S}^c}\neq 0$, and

$$-\sum_{i\in\mathcal{S}} m_i = \varrho\,\|m_{\mathcal{S}^c}\|_{\ell^1}. \tag{III.6}$$

By Lemma III.3 with $\tilde{n} := m$, we have

$$-\sum_{i\in\mathcal{S}} m_i\log\Big(\frac{x_i^*+n_i}{\alpha}\Big) = \sum_{i\in\mathcal{S}^c} m_i\log\Big(\frac{n_i}{\alpha}\Big). \tag{III.7}$$

We will first look at the right-hand side of (III.7). We have $m_i\geq 0$ for all $i\in\mathcal{S}^c$. Hence $\sum_{i\in\mathcal{S}^c} m_i = \|m_{\mathcal{S}^c}\|_{\ell^1}$. Using the concavity of the log function at (i) and its monotonicity at (ii), we infer that

$$\begin{aligned}\sum_{i\in\mathcal{S}^c}\log\Big(\frac{n_i}{\alpha}\Big)m_i &\overset{(i)}{\leq} \|m_{\mathcal{S}^c}\|_{\ell^1}\log\Big(\sum_{i\in\mathcal{S}^c}\frac{m_i}{\|m_{\mathcal{S}^c}\|_{\ell^1}}\frac{n_i}{\alpha}\Big)\\ &\overset{(ii)}{\leq} \|m_{\mathcal{S}^c}\|_{\ell^1}\log\Big(\frac{\|n_{\mathcal{S}^c}\|_{\ell^\infty}}{\alpha}\Big).\end{aligned} \tag{III.8}$$

Next we turn to the left-hand side of (III.7). Recall that we defined $\varepsilon = \frac{\alpha}{\min_{i\in\mathcal{S}} x_i^*}$ and $C = 2(1+\tilde{\varrho})\,|\mathcal{S}^c|\,\kappa_*^{\varrho^-}$. Using Theorem II.2(a) at (i) and our assumption on α at (ii), we obtain

$$\frac{\|n\|_{\ell^\infty}}{\min_{i\in\mathcal{S}} x_i^*} \overset{(i)}{\leq} \frac{C}{2}\cdot\varepsilon^{1-\varrho} \overset{(ii)}{\leq} \frac{1}{2}. \tag{III.9}$$

In particular, we have

$$x_i^* + tn_i \geq \min_{i\in\mathcal{S}} x_i^* - \|n\|_{\ell^\infty} \geq \frac{1}{2}\min_{i\in\mathcal{S}} x_i^* \tag{III.10}$$

for all $t\in[0,1]$ and all $i\in\mathcal{S}$. Using the fundamental theorem of calculus at (i), inequality (III.10) at (ii), and (III.9) at (iii), we obtain

$$\begin{aligned}&\left|\sum_{i\in S}(-m_i)\log\Big(\frac{x_i^*+n_i}{\alpha}\Big) - \sum_{i\in S}(-m_i)\log\Big(\frac{x_i^*}{\alpha}\Big)\right|\\ &\overset{(i)}{=} \left|\sum_{i\in S}\int_0^1 \frac{m_i n_i}{x_i^*+tn_i}\mathrm{d}t\right| \overset{(ii)}{\leq} 2\frac{\|m_{\mathcal{S}}\|_{\ell^1}\,\|n_{\mathcal{S}}\|_{\ell^\infty}}{\min_{i\in\mathcal{S}} x_i^*}\\ &\overset{(iii)}{\leq} \|m_{\mathcal{S}}\|_{\ell^1}\cdot C\cdot\varepsilon^{1-\varrho}.\end{aligned}$$

Let $\lambda := \|x^*\|_{\ell^\infty}$. Using (III.6), the definition of ϱ^-, and κ_*, we obtain

$$\begin{aligned}&\sum_{i\in\mathcal{S}}(-m_i)\log\Big(\frac{x_i^*}{\alpha}\Big)\\ &= \sum_{i\in\mathcal{S}}(-m_i)\log\Big(\frac{\lambda}{\alpha}\Big) + \sum_{i\in\mathcal{S}}(-m_i)\log\Big(\frac{x_i^*}{\lambda}\Big)\\ &\geq \sum_{i\in\mathcal{S}}(-m_i)\log\Big(\frac{\lambda}{\alpha}\Big) + \sum_{i\in\mathcal{S}:m_i<0}(-m_i)\log\Big(\frac{x_i^*}{\lambda}\Big)\\ &\geq \varrho\,\|m_{\mathcal{S}^c}\|_{\ell^1}\log\Big(\frac{\lambda}{\alpha}\Big) - \varrho^-\,\|m_{\mathcal{S}^c}\|_{\ell^1}\log(\kappa_*).\end{aligned}$$

Combining the previous inequalities, we infer that

$$\frac{\text{l.h.s.(III.7)}}{\|m_{\mathcal{S}^c}\|_{\ell^1}} \geq \varrho\log\Big(\frac{\|x^*\|_{\ell^\infty}}{\alpha}\Big) - \varrho^-\log(\kappa_*) - C\varepsilon^{1-\varrho}. \tag{III.11}$$

Finally, inserting (III.8) and (III.11) into (III.7), we deduce that

$$\varrho\log\Big(\frac{\|x^*\|_{\ell^\infty}}{\alpha}\Big) - \varrho^-\log(\kappa_*) - C\varepsilon^{1-\varrho} \leq \log\Big(\frac{\|n_{\mathcal{S}^c}\|_{\ell^\infty}}{\alpha}\Big)$$

Applying the exponential function to both sides and using $\exp(-t)\geq 1-t$, we deduce (II.4).

□

C. Proof of the null-space properties

Proof of Lemma III.1. (i) Since $y\neq 0$ we have $x^*\neq 0$ and so $\mathcal{S}\neq\emptyset$. By assumption, there exists $x\geq 0$ with $Ax = y$ and $x^*\neq x$. It follows that $x - x^*\in\mathcal{N}\setminus\{0\}$.

Now let $\tilde{n}\in\mathcal{N}\setminus\{0\}$. Since x^* is the unique $\|\cdot\|_{\ell^1}$-minimizer and $x^*\geq 0$, we have

$$0 < \frac{\|x^*+\varepsilon\tilde{n}\|_{\ell^1} - \|x^*\|_{\ell^1}}{\varepsilon} = \sum_{i\in\mathcal{S}}\tilde{n}_i + \sum_{i\in\mathcal{S}^c}|\tilde{n}_i| \tag{III.12}$$

for all sufficiently small $\varepsilon>0$. If $\mathcal{S}^c=\emptyset$ or if $\tilde{n}_{\mathcal{S}^c}=0$, then (III.12) is also true for $-\tilde{n}$. We obtain

$$0 < \sum_{i\in\mathcal{S}}\tilde{n}_i, \quad\text{and}\quad 0 < -\sum_{i\in\mathcal{S}}\tilde{n}_i,$$

which is impossible. Hence $\mathcal{S}^c\neq\emptyset$ and $\tilde{n}_{\mathcal{S}^c}\neq 0$.

(ii) For $\tilde{n}\in\mathcal{N}\setminus\{0\}$ let

$$\varrho(\tilde{n}) := \frac{-\sum_{i\in\mathcal{S}}\tilde{n}_i}{\|\tilde{n}_{\mathcal{S}^c}\|_{\ell^1}}.$$

Then (III.12) implies that $\varrho(\tilde{n})<1$. Let $\mathcal{N}_1 := \mathcal{N}\cap\partial B_1(0)$. Since $\mathcal{N}\setminus\{0\}$ is a cone we also have $\mathcal{N}_1\neq\emptyset$. Since $\varrho(t\tilde{n}) = \varrho(\tilde{n})$ for all $\tilde{n}\in\mathcal{N}\setminus\{0\}$ and all $t>0$, we have

$$\varrho = \sup_{\tilde{n}\in\mathcal{N}_1}\varrho(\tilde{n}).$$

The claim for ϱ now follows by compactness of $\mathcal{N}_1$ and continuity of $\varrho(\cdot)$. The remaining claims are proved analogously. □

References

[1] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization," *Communications of the ACM*, vol. 64, no. 3, pp. 107–115, 2021.

[2] T. Vaskevicius, V. Kanade, and P. Rebeschini, "Implicit regularization for optimal sparse recovery," in *Advances in Neural Information Processing Systems*, 2019, pp. 2972–2983.

[3] S. Gunasekar, B. Woodworth, S. Bhojanapalli, B. Neyshabur, and N. Srebro, *Implicit regularization in matrix factorization*, 2017. arXiv: 1705.09280.

[4] D. Stöger and M. Soltanolkotabi, "Small random initialization is akin to spectral learning: Optimization and generalization guarantees for overparameterized low-rank matrix reconstruction," *Advances in Neural Information Processing Systems*, vol. 34, pp. 23 831–23 843, 2021.

[5] Y. Li, T. Ma, and H. Zhang, "Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations," in *Conference On Learning Theory*, PMLR, 2018, pp. 2–47.

[6] B. Woodworth, S. Gunasekar, J. D. Lee, E. Moroshko, P. Savarese, I. Golan, D. Soudry, and N. Srebro, "Kernel and rich regimes in overparametrized models," in *Conference on Learning Theory*, PMLR, 2020, pp. 3635–3673.

[7] S. Gunasekar, J. Lee, D. Soudry, and N. Srebro, "Characterizing implicit bias in terms of optimization geometry," in *International Conference on Machine Learning*, PMLR, 2018, pp. 1832–1841.

[8] H.-H. Chou, J. Maly, and H. Rauhut, *More is less: Inducing sparsity via overparameterization*, 2023. arXiv: 2112.11027.

[9] J. S. Wind, V. Antun, and A. C. Hansen, *Implicit regularization in AI meets generalized hardness of approximation in optimization – Sharp results for diagonal linear networks*, 2023. arXiv: 2307.07410.

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

[11] H. Matt and D. Stöger, *Linear regression with overparameterized linear neural networks: Tight upper and lower bounds for implicit ℓ^1-regularization*, To appear, 2025.

[12] F. Hoppe, F. Krahmer, C. Mayrink Verdun, M. I. Menzel, and H. Rauhut, "High-dimensional confidence regions in sparse mri," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5. DOI: 10.1109/icassp49357.2023.10096320.

[13] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken N.J.: Wiley-Interscience, 2006.