More is Less: Inducing Sparsity via Overparameterization

Hung-Hsu Chou,[1], Johannes Maly,[2], and Holger Rauhut[1]

[1]RWTH Aachen University, Germany
[2]Catholic University of Eichstaett-Ingolstadt, Germany

April 1, 2022

Abstract

In deep learning it is common to overparameterize neural networks, that is, to use more parameters than training samples. Quite surprisingly training the neural network via (stochastic) gradient descent leads to models that generalize very well, while classical statistics would suggest overfitting. In order to gain understanding of this implicit bias phenomenon we study the special case of sparse recovery (compressed sensing) which is of interest on its own. More precisely, in order to reconstruct a vector from underdetermined linear measurements, we introduce a corresponding overparameterized square loss functional, where the vector to be reconstructed is deeply factorized into several vectors. We show that, if there exists an exact solution, vanilla gradient flow for the overparameterized loss functional converges to a good approximation of the solution of minimal ℓ_1-norm. The latter is well-known to promote sparse solutions. As a by-product, our results significantly improve the sample complexity for compressed sensing via gradient flow/descent on overparameterized models derived in previous works. The theory accurately predicts the recovery rate in numerical experiments. Our proof relies on a analyzing a certain Bregman divergence of the flow. This bypasses the obstacles caused by non-convexity and should be of independent interest.

Keywords — Overparameterization, ℓ_1-minimization, Gradient Flow, Gradient Descent, Compressed Sensing, Implicit Bias, Deep Factorization, Bregman Divergence

Contents

1 Introduction

Overparameterization is highly successful in learning deep neural networks. While this empirical finding was likely observed by countless practitioners, it was systematically studied in numerical experiments in [18, 19, 29]. Increasing the number of parameters far beyond the number of training samples leads to better generalization properties of the learned networks. This is in stark contrast to classical statistics, which would rather suggest overfitting in this scenario. The loss function typically has infinitely many global minimizers in this setting (there are usually infinitely many networks fitting the training samples exactly in the overparameterized regime [29]), so that the employed optimization algorithm has a significant influence on the computed solution. The commonly used (stochastic) gradient descent and its variants seem to have an implicit bias towards "nice" networks with good generalization properties. It is conjectured that in fact, (stochastic) gradient descent applied to learning deep networks favors solutions of low complexity. Of course, the right notion of "low complexity" needs to be identified and may depend on the precise scenario, i.e., network architecture. In the simplified setting of linear networks, i.e., matrix factorizations, and the problem of matrix recovery or more specifically matrix completion, several works [1, 2, 7, 10, 11, 12, 13, 14, 18, 19, 20, 21, 22, 23, 24] identified the right notion of low complexity to be low rank of the factorized matrix (or tensor, in some these works). Nevertheless, despite initial theoretical results, a fully convincing theory is not yet available even for this simplified setting.

As a contribution to this line of research, this article studies the performance of overparameterized models in the context of the classical compressed sensing problem, where the goal is to recover an unknown high-dimensional signal $\mathbf{x}^\star \in \mathbb{R}^N$ from few linear measurements of the form

$$\mathbf{y} = \mathbf{A}\mathbf{x}^\star \in \mathbb{R}^M, \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{M \times N}$ models the measurement process. Whereas this is in general not possible for $M < N$, the seminal works [5, 6, 8] showed that unique reconstruction of $\mathbf{x}^\star$ from $\mathbf{A}$ and $\mathbf{y}$ becomes feasible if $\mathbf{x}^\star$ is s-sparse, $\mathbf{A}$ satisfies certain conditions (e.g. restricted isometry property), and M scales like $s \log(N/s)$. While it is well-known that $\mathbf{x}^\star$ can be recovered via ℓ_1-minimization, i.e.,

$$\mathbf{x}^\star = \arg\min_{\mathbf{A}\mathbf{z}=\mathbf{y}} \|\mathbf{z}\|_1, \tag{2}$$

also known as basis pursuit, surprisingly enough, we observe similar recovery with gradient descent on overparameterized models, which originate from a very different background. (Here in the following, $\|\cdot\|_p$ denotes the standard ℓ_p-norm for $1 \leq p \leq \infty$.) Our goal is to understand this connection via gradient flow – a continuous version of gradient descent – in this context.

In order to gain a first understanding of the power of overparameterization, we compare the quadratic loss (without overparameterization)

$$\mathcal{L}_{\text{quad}}(\mathbf{x}) := \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 \tag{3}$$

and its overparameterized versions

$$\mathcal{L}_{\text{over}}\big(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(L)}\big) := \frac{1}{2}\Big\|\mathbf{A}\big(\mathbf{x}^{(1)} \odot \dots \odot \mathbf{x}^{(L)}\big) - \mathbf{y}\Big\|_2^2, \tag{4}$$

$$\mathcal{L}_{\text{over}}^{\pm}\big(\mathbf{u}^{(1)}, \dots, \mathbf{u}^{(L)}, \mathbf{v}^{(1)}, \dots, \mathbf{v}^{(L)}\big) := \frac{1}{2}\Big\|\mathbf{A}\Big(\bigodot_{k=1}^{L}\mathbf{u}^{(k)} - \bigodot_{k=1}^{L}\mathbf{v}^{(k)}\Big) - \mathbf{y}\Big\|_2^2, \tag{5}$$

where $\odot$ is the Hadamard (or entry-wise) product. It turns out that the results from minimizing different loss functions via gradient flow/descent are quite different. The quadratic loss $\mathcal{L}_{\text{quad}}$ leads to the least-squares solution

$$\mathbf{x}_\infty := \lim_{t\to\infty}\mathbf{x}(t) = \arg\min_{\mathbf{A}\mathbf{z}=\mathbf{y}}\|\mathbf{z}\|_2 \ ,$$

which is unrelated to the sparse ground-truth $\mathbf{x}^\star$ in general. On the other hand, gradient flow/descent applied to $\mathcal{L}_{\text{over}}$ and $\mathcal{L}_{\text{over}}^{\pm}$ often lead to sparse results, despite the absence of explicit regularization. This can be viewed as a particular instance of implicit bias/regularization. Formally, we define the gradient flow for $\mathcal{L}_{\text{over}}$ as

$$\mathbf{x}^{(k)}(t) = -\nabla_{\mathbf{x}^{(k)}}\mathcal{L}_{\text{over}}(\mathbf{x}^1(t), \dots, \mathbf{x}^{(L)}(t)), \qquad \mathbf{x}^{(k)}(0) = \alpha\mathbf{1}, \quad k = 1, \dots, L,$$

and the gradient flows $\mathbf{u}^{(k)}(t)$ and $\mathbf{v}^{(k)}(t)$ for $\mathcal{L}_{\text{over}}^{\pm}$ are defined similarly. We analyze both $\mathcal{L}_{\text{over}}$ and $\mathcal{L}_{\text{over}}^{\pm}$ because $\mathcal{L}_{\text{over}}$ is simpler to analyzed, but only $\mathcal{L}_{\text{over}}^{\pm}$ can recover both positive and negative entries of a vector.

We show that the product $\tilde{\mathbf{x}}(t) := \bigodot_{k=1}^{L}\mathbf{x}^{(k)}$ (and $\bigodot_{k=1}^{L}\mathbf{u}^{(k)} - \bigodot_{k=1}^{L}\mathbf{v}^{(k)}$) converges to an approximate solution of the ℓ_1-minimization problem (1), provided that the initialization parameter $\alpha > 0$ is sufficiently small. Hence, in situations where ℓ_1-minimization is known to successfully recover sparse solutions also overparameterized gradient flow will succeed. In particular, conditions on the restricted isometry property or the null space property on $\mathbf{A}$, tangent cone conditions and dual certificates on $\mathbf{A}$ and $\mathbf{x}^*$ ensuring recovery of s-sparse vectors transfer to overparameterized gradient flow. For instance, a random Gaussian matrix $\mathbf{A} \in \mathbb{R}^{M\times N}$ ensures recovery of s sparse vectors via overparameterized gradient flow for $M \sim s\log(N/s)$. We refer the interested reader to the monograph [9] for details on compressed sensing.

Our main result improves on previous work: For $L = 2$, it was shown in [16] that gradient descent for $\mathcal{L}^{\pm}$ defined in (34), which is closely related to $\mathcal{L}_{\text{over}}^{\pm}$, converges to an s-sparse $\mathbf{x}^*$ if $\mathbf{A}$ satisfies a certain coherence assumption, which requires at least $M \gtrsim s^2$ measurements. In [25], it was shown for general $L \geq 2$ that $\mathcal{L}^{\pm}$ converges to an s-sparse $\mathbf{x}^*$ if the restricted isometry constant δ_s of $\mathbf{A}$ (see below) essentially satisfies $\delta_s \leq c/\sqrt{s}$. This condition can only be satisfied if $M \gtrsim s^2\log(N/s)$, see [9]. Hence, our result significantly reduces the required number of measurements, in fact, down to the optimal number. However, we note that [16, 25] work with gradient descent, while we use gradient flow. Extending our result to gradient descent is left to future work.

We believe that the method of proof for our main results is of independent interest. In fact, we significantly extend work in [28] and analyze a Bregman divergence of our gradient flow with respect to a suitably chosen function. This allows to characterize the limit of the gradient flow in dependence of the initialization, and thereby derive the relation to ℓ_1-minimizer.

1.1 Main result

We consider vanilla gradient flow on the factorized models (4) and (5). We show that if the solution space is non-empty, gradient flow converges to a solution of (1) whose ℓ_1-norm is ε-close to the minimum among all solutions. We now state our main result.

Theorem 1.1. *Let $L \geq 2$, $\mathbf{A} \in \mathbb{R}^{M \times N}$, and $\mathbf{y} \in \mathbb{R}^M$. For the general overparameterized loss function $\mathcal{L}^{\pm}_{over}$ defined in (5) let $\mathbf{u}^{(k)}(t)$ and $\mathbf{v}^{(k)}(t)$ follow the flow*

$$\begin{aligned} \big(\mathbf{u}^{(k)}\big)'(t) &= -\nabla_{\mathbf{u}^{(k)}} \mathcal{L}^{\pm}_{over}\big(\mathbf{u}^{(1)}, \dots, \mathbf{u}^{(L)}, \mathbf{v}^{(1)}, \dots, \mathbf{v}^{(L)}\big), \qquad \mathbf{u}^{(k)}(0) = \alpha \mathbf{1}, \quad k = 1, \dots, L, \\ \big(\mathbf{v}^{(k)}\big)'(t) &= -\nabla_{\mathbf{v}^{(k)}} \mathcal{L}^{\pm}_{over}\big(\mathbf{u}^{(1)}, \dots, \mathbf{u}^{(L)}, \mathbf{v}^{(1)}, \dots, \mathbf{v}^{(L)}\big), \qquad \mathbf{v}^{(k)}(0) = \alpha \mathbf{1}, \quad k = 1, \dots, L, \end{aligned}$$

for some $\alpha > 0$. Suppose the solution set $S = \{\mathbf{z} \in \mathbb{R}^N : \mathbf{A}\mathbf{z} = \mathbf{y}\}$ is non-empty. Then the limit

$$\tilde{\mathbf{x}}_\infty := \lim_{t \to \infty} \Big(\mathbf{u}^{(1)}(t) \odot \cdots \odot \mathbf{u}^{(L)}(t) - \mathbf{v}^{(1)}(t) \odot \cdots \odot \mathbf{v}^{(L)}(t) \Big) \tag{6}$$

exists and is contained in S.

Further, let $\varepsilon > 0$ and $Q = \min_{\mathbf{z} \in S} \|\mathbf{z}\|_1$, and assume that

$$\alpha \leq h(Q, \varepsilon) := \begin{cases} \min\left(e^{-\frac{1}{2}}, \exp\left(\frac{1}{2} - \frac{Q^2 + 3Ne^{-1}}{2\varepsilon}\right)\right) & \text{if } L = 2, \\ \min\left(1, \left(\frac{2\varepsilon}{L(Q+3N+\varepsilon) - 4N}\right)^{\frac{1}{L-2}}\right) & \text{if } L > 2. \end{cases} \tag{7}$$

Then the ℓ_1-norm of $\tilde{\mathbf{x}}_\infty$ satisfies

$$\|\tilde{\mathbf{x}}_\infty\|_1 - Q \leq \varepsilon. \tag{8}$$

Remark 1.2. For $L \geq 3$ we can find a slightly better but more complicated condition on α than in (7). In fact, it then suffices to have

$$\alpha \leq \widetilde{h}(Q, \varepsilon) = \min\left\{ \varepsilon^{1/L} \min\{N^{-4/L^2} L^{-\frac{1}{L-2}}, (2N(L-2))^{-\frac{1}{L}}\}, (\varepsilon/2)^{\frac{1}{L-2}} N^{-\frac{2}{L^2}} Q^{-\frac{1}{L(L-2)}} \right\}.$$

We refer to the appendix for details.

Theorem 1.1 gives the explicit non-asymptotic scaling (7) between the error ε and the initialization scale α. In particular, the error becomes smaller for smaller initialization. Note, however, that initializing α closer to zero also means longer convergence times in practice, so there is a trade-off between accuracy and computational complexity. Note that (7) scales differently for different L. Whereas the condition for $L = 2$ requires an exponentially small initialization (see also in [27]), the condition for $L > 2$ is far less restrictive. Since empirically smaller α leads to slower convergence, it is desirable to take α as large as possible within the allowed range, i.e., the upper bound. Hence, our theory suggests an advantage of deep factorizations ($L > 2$) over shallow factorization ($L = 2$). However, this statement should be taken with a grain of salt, since we currently have no results about the necessity of (potential variants of) the bounds in (7) for the limit of gradient flow to approximate ℓ_1-mininizers. In fact, our numerical experiments in Section 4 suggest that also for $L = 2$

the scaling of α in ε is not exponential. But they do suggest that larger L improves performance. It seems that our current proof method cannot easily be modified to provide better theoretical scaling for $L = 2$.

We also remark that due to the monotonicity of $Q \mapsto h(Q, \varepsilon)$, in order to use (7), it suffices to find an upper bound for Q instead of exact evaluation of Q. A simple bound is, for instance, $Q \leq \|\mathbf{A}^\dagger \mathbf{y}\|_1$.

1.2 Application to Compressed Sensing

As a consequence of Theorem 1.1, accurately reconstructing (approximately) sparse vectors from incomplete linear measurements (compressed sensing) can provably be achieved via gradient flow on $\mathcal{L}_{\text{over}}$, using the minimal amount of measurements. Hence, gradient flow on overparameterization can be viewed as an alternative algorithm to compressed sensing. To be concrete let us provide example results in this direction. A matrix $\mathbf{A} \in \mathbb{R}^{M\times N}$ is said to satisfy the stable null space property of order s and constant $\rho \in (0,1)$ if for all vector $\mathbf{v} \in \ker(\mathbf{A}) \setminus \{0\}$ and all index sets $S \subset \{1, \dots, N\}$ of cardinality at most s it holds

$$\|\mathbf{v}_S\|_1 \leq \|\mathbf{v}_{S^c}\|_1,$$

where $\mathbf{v}_S$ denotes the restriction of $\mathbf{v}$ to the entries in S and S^c is the complement of S. It is well-known that various types of random matrices satisfy the stable null space property in an (almost) optimal parameter regime with high probability, see for instance [4, 9, 17] and references therein. For instance, for a Gaussian random matrix $\mathbf{A}$ this is true provided that

$$M \gtrsim \rho^{-2} s \log(eN/s). \tag{9}$$

Moreover, let us also introduce the error of best s-term approximation in ℓ_1 of a vector x as

$$\sigma_s(\mathbf{x})_1 = \min_{\mathbf{z}:\|\mathbf{z}\|_0 \leq s} \|\mathbf{x} - \mathbf{z}\|_1,$$

where $\|\mathbf{z}\|_0 = \#\{\ell : z_\ell \neq 0\}$ denotes the sparsity of $\mathbf{z}$. Clearly,$\sigma_s(\mathbf{x})_1 = 0$ if $\mathbf{x}$ is s-sparse.

The following result is a straightforward implication of [9, Theorem 4.14] together with Theorem 1.1. We also refer to Figure 1 and Section 4 for further illustration.

Corollary 1.3. *Let $\mathbf{x}_* \in \mathbb{R}^N$ and $\mathbf{y} = \mathbf{A}\mathbf{x}_*$, where $\mathbf{A} \in \mathbb{R}^{M\times N}$ satisfies the stable null space property of order s with constant $\rho \in (0,1)$. Let $\tilde{\mathbf{x}}_\infty$ be the limit vector of gradient flow as in (6). Let $\varepsilon > 0$ and assume that the initialization parameter satisfies*

$$\alpha \leq h\big(\|\mathbf{x}_*\|_1 + \frac{2(1+\rho)}{1-\rho}\sigma_s(\mathbf{x}_*)_1, \varepsilon\big) \tag{10}$$

with h defined as in (7). Then the reconstruction error satisfies

$$\|\tilde{\mathbf{x}}_\infty - \mathbf{x}_*\|_1 \leq \frac{1+\rho}{1-\rho}\left(2\sigma_s(\mathbf{x}_*)_1 + \varepsilon\right).$$

Clearly, the right hand side equals ε if $\mathbf{x}_*$ is s-sparse.

Note that $\sigma_s(\mathbf{x}_*)_1 \leq \|\mathbf{x}_*\|_1$ and $Q \mapsto h(Q, \varepsilon)$ is monotonically decreasing so that (10) is satisfied under the simpler condition

$$\alpha \leq h\left(\frac{3+\rho}{1-\rho}\|\mathbf{x}_*\|_1, \varepsilon\right).$$

We can show also stability with respect to noise on the measurements assuming a certain quotient property of the measurement matrix [9] in addition to the null space property. To be precise, we say that $\mathbf{A} \in \mathbb{R}^{M \times N}$ satisfies the ℓ_1-quotient property with constant d relative to the ℓ_2-norm if for all $\mathbf{e} \in \mathbb{R}^M$, there exists $\mathbf{u} \in \mathbb{C}^N$ with $\mathbf{Au} = \mathbf{e}$ such that

$$\|\mathbf{u}\|_1 \leq d\sqrt{s_*}\|\mathbf{e}\|_2 \quad \text{ with } s_* = M/(\log(eN/M).$$

Gaussian random matrices satisfy this property for an absolute constant d with high probability, see [9, Chapter 11] for details.

Moreover, we require the following strengthened version of the null space property. A matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$ satisfies the ℓ_2-robust null space property with constants $0 < \rho < 1$ and $\tau > 0$ of order s (with respect to ℓ_2) if, for any subset $S \subset [N]$ of cardinality s,

$$\|\mathbf{v}_S\|_2 \leq \frac{\rho}{\sqrt{s}}\|\mathbf{v}_{S^c}\|_1 + \tau\|\mathbf{Av}\|_2 \quad \text{ for all } \mathbf{v} \in \mathbb{R}^N. \tag{11}$$

Again, Gaussian random matrices satisfy this property with appropriate absolute constants τ and ρ with high probability under (9).

The following theorem establishes robustness under noise for the reconstruction of $\mathbf{x}_*$ via gradient flow on the overparameterized functional $\mathcal{L}^{\pm}_{\text{over}}$ in (5).

Theorem 1.4. *Let $\mathbf{A} \in \mathbb{R}^{M \times N}$ be a matrix satisfying the ℓ_2-robust null space property with constants $0 \leq \rho < 1$ and $\tau > 0$ of order $s = cs_* = cM/\log(eN/M)$ and the ℓ_1-quotient property with respect to the ℓ_2-norm with constant $d > 0$. For $\mathbf{x}_* \in \mathbb{R}^N$ let $\mathbf{y} = \mathbf{Ax}_* + \mathbf{e}$ for some noise vector $\mathbf{e} \in \mathbb{R}^M$. For $\varepsilon > 0$ assume that $\alpha > 0$ satisfies*

$$\alpha \leq h\left(\|\mathbf{x}_*\|_1 + \frac{2(1+\rho)}{1-\rho}\sigma_s(\mathbf{x}_*)_1 + 2d\sqrt{s_*}\|\mathbf{e}\|_2, \varepsilon\right). \tag{12}$$

Let $\tilde{\mathbf{x}}_\infty$ be the limit (6) of the gradient flow for $\mathcal{L}^{\pm}_{over}$ initialized with parameter α. Then the reconstruction error satisfies

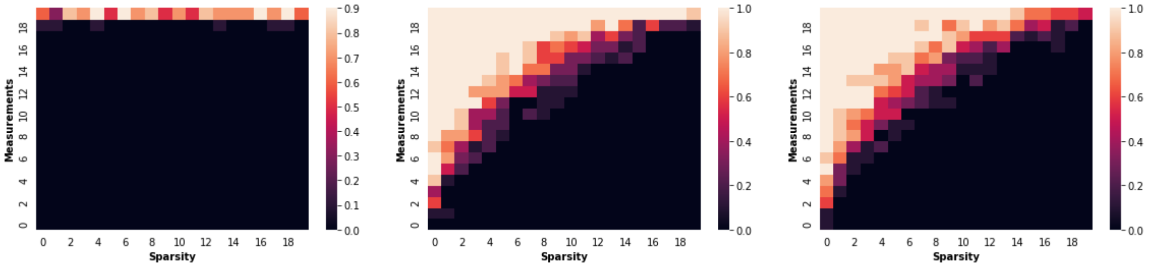
$$\|\tilde{\mathbf{x}}_\infty - \mathbf{x}_*\|_2 \leq \frac{C}{\sqrt{s}}(2\sigma_s(\mathbf{x}_*)_1 + \varepsilon) + C'\|\mathbf{e}\|_2.$$

The constants $C, C' > 0$ only depend on ρ, τ, c, d.

For $M \times N$ Gaussian random matrices $\mathbf{A}$ the assumptions of the theorem are satisfied with high probability if $M \geq Cs\log(eN/s)$ for an appropriate constant C (depending on the other constants ρ, τ, d, c). The error bound suggests that ε should ideally be of the order $\sigma_s(\mathbf{x})_1$ or $\sqrt{s}\|\mathbf{e}\|_2$ (whichever is larger) and α should then satisfy (12). The proof of this theorem is contained in the appendix.

1.3 Related Work

Since training deep neural (linear) networks via gradient descent is closely connected to matrix factorization, many works focus on understanding implicit bias in this setting. The corresponding results on matrix factorization and matrix sensing are of a similar flavor and show an implicit low-rank bias of gradient flow/descent when minimizing factorized quadratic losses [1, 2, 7, 10, 11, 12, 13, 14, 18, 19, 20, 23, 24, 26, 28]. It is noteworthy that the existing matrix sensing results, e.g., [24], require $\mathcal{O}(r^2 n)$ measurements to guarantee reconstruction of rank-r $n \times n$-matrices via gradient descent, i.e., they share the sub-optimal sample complexity of [16, 25]. For comparison, for low-rank matrix reconstruction by conventional methods like nuclear-norm minimization, only $\mathcal{O}(rn)$ measurements



(a) $\mathcal{L}_{\text{quad}}$ minimization (3) via GD

(b) ℓ_1 minimization (2)

(c) $\mathcal{L}^{\pm}_{\text{over}}$ minimization (5) via GD

Figure 1: We compare the recovery probability for different method via heatmaps. The horizontal axis is the sparsity level s, and vertical axis is the number of measurements M. The result (a) shows that we cannot expect any recovery by using the naive quadratic loss. Note that our model (c) achieves similar, and in fact arguably better performance than the well known (b) basis pursuit method.

are needed. For a more detailed discussion of the literature on matrix factorization/sensing via overparametrization, we refer the reader to [7].

Most of the above mentioned works consider the case of $L = 2$ layers, whereas the literature for general $L > 2$ is scarce. In terms of proof method, [28] is most related to ours among the above works. In the setting of matrix sensing, the authors show that mirror flow/descent exhibits an implicit regularization when the mirror map is the spectral (hyper)entropy. Their analysis heavily relies on the concept of Bregman divergence. Furthermore, they draw a connection between gradient descent on symmetric matrix factorization ($L = 2$) and the mirror descent without factorization, cf. [14]. Although having started from a different perspective, we realized that the quantity that we called ***solution entropy*** in the first version of our paper can be viewed as a Bregman divergence. With this observation, we were able to further improve our work and remove technical assumptions on $\mathbf{A}$. Despite some similarity in the proof strategy, the authors of [28] implicitly concentrate on the case $L = 2$ by only considering one specific Bregman divergence. In contrast, our work provides suitable Bregman divergences for all $L \geq 2$ and as such allows to analyze gradient descent on overparamterized loss functions with arbitrary depth. Let us emphasize that the deeper case $L \geq 3$, which leads to better results in our analysis, has not been covered by any existing work to the best of our knowledge.

Due to the strong coupling of weights in factorized matrix sensing, several existing works restrict themselves to the vector case. In [16, 25] the authors derive robust reconstruction guarantees for gradient descent and the compressed sensing model (1). Whereas in [25] the authors consider the case where $L = 2$ and the sensing matrix $\mathbf{A}$ satisfies the restricted isometry property, [16] extends the results to $L \geq 2$ under a coherence assumption for $\mathbf{A}$. In [15] the author shows that solving the LASSO is equivalent to minimizing an ℓ_2-regularized overparameterized functional (for $L = 2$) and uses this to solve LASSO by alternating least-squares methods. Although the author also considers deeper factorization, the presented approach leads to different results since the overparameterized functional is equivalent to $\ell_{\frac{2}{L}}$-norm instead of ℓ_1-norm minimization, for $L > 2$. The subsequent work [30] builds upon those ideas to perform sparse recovery with gradient descent by assuming a restricted isometry property of $\mathbf{A}$. Nevertheless, the presented results share the sub-optimal sample complexity discussed above.

Instead of specific initialization, in [27] the authors examine the limits of gradient flow when initialized by $\alpha \mathbf{w}_0$, for any $\mathbf{w}_0$. They show that for large $\alpha > 0$ the limit of gradient flow approximates the least-square solution, whereas for $\alpha > 0$ small it approximates an ℓ_1-norm minimizer. While the authors discuss more general types of initialization, their proof strategy is fundamentally different from ours and has certain shortcomings. They need to *assume* convergence of the gradient flow and obtain only for $L = 2$ non-asymptotic bounds on the initialization magnitude required for implicit ℓ_1-regularization. In contrast, we actually *show* convergence of gradient flow and provide non-asymptotic bounds for all L leading to less restrictive assumptions on the initialization magnitude.

1.4 Outline

Sections 2 and 3 are dedicated to proving Theorem 1.1. Section 2 illustrates the proof strategy in a simplified setting of positive solutions, whereas Section 3 extends the proof to full generality. Finally, we present in Section 4 numerical evidence supporting our claims and conclude with a brief summary/outlook on future research directions in Section 5.

1.5 Notation

For $N \in \mathbb{N}$, we denote $[N] = \{1, 2, \dots, N\}$. Boldface lower-case letters like $\mathbf{x}$ represent vectors with entries x_n, while boldface upper-case letters like $\mathbf{A}$ represent matrices with entries A_{mn}. For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$, $\mathbf{x} \geq \mathbf{y}$ means that $x_n \geq y_n$, for all $n \in [N]$. We use $\odot$ to denote the Hadamard product, i.e., the vectors $\mathbf{x} \odot \mathbf{y}$ and $\mathbf{x}^{\odot p}$ have entries $(\mathbf{x} \odot \mathbf{y})_n = x_n y_n$ and $(\mathbf{x}^{\odot p})_n = x_n^p$, respectively. We abbreviate $\tilde{\mathbf{x}} := \bigodot_{k \in [L]} \mathbf{x}^{(k)} = \mathbf{x}^{(1)} \odot \cdots \odot \mathbf{x}^{(L)}$. The logarithm is applied entry-wise to positive vectors, i.e., $\log(\mathbf{x}) \in \mathbb{R}^N$ with $\log(\mathbf{x})_n = \log(x_n)$. For convenience we denote $\mathbb{R}_+^N = \{\mathbf{x} \in \mathbb{R}^N : x_n \geq 0 \, \forall n \in [N]\}$.

2 The Positive Case

To show Theorem 1.1, we are going to prove in this section the following simplified version, Theorem 2.1, that treats the model in (4) and is restricted to the positive orthant (recall that gradient flow applied to (4) preserves the entry-wise sign of the iterates). To this end, we define the set of non-negative solutions

$$S_+ = \{\mathbf{z} \geq 0 : \mathbf{A}\mathbf{z} = \mathbf{y}\}. \tag{13}$$

In fact, the proof strategy for Theorem 1.1 is then a straight-forward adaption of the proof of Theorem 2.1 and will be discussed in Section 3. Note that Theorem 2.1 can be easily adapted to other orthants by changing the signs of the initialization vector.

Theorem 2.1 (Equivalence to ℓ_1-minimization, positive case)**.** *Let $L \geq 2$, $\mathbf{A} \in \mathbb{R}^{M \times N}$ and $\mathbf{y} \in \mathbb{R}^M$ and assume that S_+ defined in (13) is non-empty. With the overparameterized loss function $\mathcal{L}_{over}$ in (4), define the $\mathbf{x}^{(k)}(t)$ via*

$$\left(\mathbf{x}^{(k)}\right)'(t) = -\nabla_{\mathbf{x}^{(k)}} \mathcal{L}_{over}(\mathbf{x}^{(1)}(t), \dots, \mathbf{x}^{(L)}(t)), \qquad \mathbf{x}^{(k)}(0) = \alpha \mathbf{1}, \quad k = 1, \dots, L, \tag{14}$$

for some $\alpha > 0$. Then the limit $\tilde{\mathbf{x}}_\infty := \lim_{t \to \infty} \bigodot_{k \in [L]} \mathbf{x}^{(k)}(t) = \lim_{t \to \infty} \tilde{\mathbf{x}}(t)$ exists and $\tilde{\mathbf{x}}_\infty \in S_+$.

Further, let $\varepsilon > 0$ and $Q_+ = \min_{\mathbf{z} \in S_+} \|\mathbf{z}\|_1$. Assume that

$$\alpha \leq h_+(Q_+, \varepsilon) := \begin{cases} \min\left(e^{-\frac{1}{2}}, \exp\left(1 - \frac{Q_+^2 + Ne^{-1}}{2\varepsilon}\right)\right) & \textit{if } L = 2, \\ \left(\frac{2\varepsilon}{L(Q_+ + N + \varepsilon)}\right)^{\frac{1}{L-2}} & \textit{if } L > 2. \end{cases} \tag{15}$$

Then the ℓ_1-norm of $\tilde{\mathbf{x}}_\infty$ satisfies

$$\|\tilde{\mathbf{x}}_\infty\|_1 - Q_+ \leq \varepsilon. \tag{16}$$

Remark 2.2. The condition (15) can be slightly improved to a weaker, but more complicated bound, with a longer proof. In fact, for the theorem to hold in the case $L \geq 3$ if suffices that

$$\alpha \leq \widehat{h}_+(Q_+, \varepsilon) = N^{-\frac{2}{L^2}} L^{-\frac{1}{L-2}} \min\left\{(2\varepsilon)^{1/L}, (\varepsilon/L)^{\frac{1}{L-2}} Q_+^{-\frac{2}{L(L-2)}}\right\}. \tag{17}$$

We refer to the appendix for details.

2.1 Reduced Factorized Loss

To analyze the dynamics $\left(\mathbf{x}^{(k)}\right)'(t) = -\nabla_{\mathbf{x}^{(k)}} \mathcal{L}_{\text{over}}(\mathbf{x}^1, \dots, \mathbf{x}^{(L)})$, we first derive a compact expression for $\nabla \mathcal{L}_{\text{over}}$. We further simplify this expression by assuming identical initialization, i.e., $\mathbf{x}^{(k)}(0) = \alpha \mathbf{1}$ for all k, and arrive at what we call the *Reduced Factorized Loss* $\mathcal{L}$, which will be used in the proofs later on.

Lemma 2.3. *For $L \geq 2$, let $\tilde{\mathbf{x}} := \bigodot_{\ell \in [L]} \mathbf{x}^{(\ell)}$ and $\tilde{\mathbf{x}}_{k^c} := \bigodot_{\ell \in [L] \setminus \{k\}} \mathbf{x}^{(\ell)}$ for $k \in [L]$. Then*

$$\nabla_{\mathbf{x}^{(k)}} \mathcal{L}_{over}\big(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(L)}\big) = [\mathbf{A}^T(\mathbf{A}\tilde{\mathbf{x}} - \mathbf{y})] \odot \tilde{\mathbf{x}}_{k^c} \ . \tag{18}$$

Proof. By the chain rule we have, for any $n \in [N]$, that

$$\begin{aligned} \nabla_{x_n^{(k)}} \mathcal{L}_{\text{over}}\big(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(L)}\big) &= \frac{1}{2} \sum_{m \in [M]} \nabla_{x_n^{(k)}} \left(\mathbf{A}\tilde{\mathbf{x}} - \mathbf{y}\right)_m^2 \\ &= \sum_{m \in [M]} (\mathbf{A}\tilde{\mathbf{x}} - \mathbf{y})_m (\mathbf{A})_{mn} (\tilde{\mathbf{x}}_{k^c})_n = [\mathbf{A}^{\mathrm{T}}(\mathbf{A}\tilde{\mathbf{x}} - \mathbf{y})]_n (\tilde{\mathbf{x}}_{k^c})_n \ . \end{aligned}$$

This completes the proof. □

Using Lemma 2.3, we now show that the dynamics of the factors $\mathbf{x}^{(k)}$ can be simplified if all factors are identically initialized.

Lemma 2.4 (Identical Initialization)**.** *Suppose $\mathbf{x}^{(k)}(t)$ follows the negative gradient flow*

$$\left(\mathbf{x}^{(k)}\right)'(t) = -\nabla_{\mathbf{x}^{(k)}} \mathcal{L}_{over}\big(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(L)}\big) \ .$$

If all initialization vectors are identical, i.e., $\mathbf{x}^{(k)}(0) = \mathbf{x}^{(k')}(0)$ for all $k, k' \in [L]$, then $\mathbf{x}^{(k)}(t) = \mathbf{x}^{(k')}(t)$ for all $t \geq 0$ and all $k, k' \in [L]$. Moreover, with $\mathbf{x}(t) := \mathbf{x}^{(1)}(t) = \dots = \mathbf{x}^{(L)}(t)$ and $\mathcal{L}(\mathbf{x}) = \frac{1}{2}\|\mathbf{A}\mathbf{x}^{\odot L} - \mathbf{y}\|_2^2$ the dynamics is given by

$$\mathbf{x}'(t) = -\nabla \mathcal{L}(\mathbf{x}) \ .$$

Proof. It suffices to show that if $\mathbf{x}^{(k)} = \mathbf{x}^{(k')}$ for all $k, k' \in [L]$, then $\nabla_{\mathbf{x}^{(k)}} \mathcal{L}_{\text{over}} = \nabla_{\mathbf{x}^{(k')}} \mathcal{L}_{\text{over}}$ for all $k, k' \in [L]$. By Lemma 2.3, the only dependence of $\nabla_{\mathbf{x}^{(k)}} \mathcal{L}_{\text{over}}$ on k is through $\tilde{\mathbf{x}}_{k^c}$. Since $\mathbf{x}^{(k)}$ and $\mathbf{x}^{(k')}$ are identical,

$$\tilde{\mathbf{x}}_{k^c} = \bigodot_{\ell \in [L] \setminus \{k\}} \mathbf{x}^{(\ell)} = \bigodot_{\ell \in [L] \setminus \{k'\}} \mathbf{x}^{(\ell)} = \mathbf{x}_{(k')^c}$$

and hence $\nabla_{\mathbf{x}^{(k)}}\mathcal{L}_{\text{over}} = \nabla_{\mathbf{x}^{(k')}}\mathcal{L}_{\text{over}}$. Since by assumption all $\mathbf{x}^{(k)}$ are identical, we can replace them with $\mathbf{x}$. Hence $\tilde{\mathbf{x}} = \mathbf{x}^{\odot L}$ and $\tilde{\mathbf{x}}_{k^c} = \mathbf{x}^{\odot(L-1)}$. Plugging this into (18), we get that $\mathbf{x}'(t) = -\left[\mathbf{A}^{\mathrm{T}}(\mathbf{A}\mathbf{x}^{\odot L} - \mathbf{y})\right] \odot \mathbf{x}^{\odot(L-1)}$, which is exactly $\nabla\mathcal{L}(\mathbf{x})$. □

At first sight, the reduction of the number of parameters in Lemma 2.4 may seem counter-intuitive to the idea of overparameterization. However, as opposed to the standard loss function $\mathcal{L}_{\text{quad}}(\mathbf{x}) = \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2$, we arrive at an alternative loss function $\mathcal{L}(\mathbf{x}) = \frac{1}{2}\|\mathbf{A}\mathbf{x}^{\odot L} - \mathbf{y}\|_2^2$ induced by the overparameterization and having a different optimization landscape than $\mathcal{L}_{\text{quad}}$. Motivated by Lemma 2.4, we will thus consider the loss $\mathcal{L}$ for the rest of the paper.

Definition 2.5 (Reduced Factorized Loss). *Let $L \in \mathbb{N}$, $L \geq 2$. For $\mathbf{A} \in \mathbb{R}^{M\times N}$ and $\mathbf{y} \in \mathbb{R}^M$, the* reduced factorized loss function *is defined as*

$$\mathcal{L}\colon \mathbb{R}^N \to [0,\infty), \qquad \mathcal{L}(\mathbf{x}) := \frac{1}{2}\|\mathbf{A}\mathbf{x}^{\odot L} - \mathbf{y}\|_2^2. \tag{19}$$

Note that the gradient of $\mathcal{L}$ is given by $\nabla\mathcal{L}(\mathbf{x}) = L\left[\mathbf{A}^{\mathrm{T}}(\mathbf{A}\mathbf{x}^{\odot L} - \mathbf{y})\right] \odot \mathbf{x}^{\odot(L-1)}$.

2.2 Bregman Divergence

As crucial ingredient for Theorem 2.1 we show convergence of $\mathbf{x}(t)^{\odot L}$ and characterize its limit. In order to state the corresponding result, we introduce the function

$$g_{\mathbf{x}}(\mathbf{z}) = \begin{cases} \langle \mathbf{z}, \log(\mathbf{z}) - \mathbf{1} - \log(\mathbf{x})\rangle & \text{if } L = 2, \\ 2\|\mathbf{z}\|_1 - L\langle \mathbf{z}^{\odot\frac{2}{L}}, \mathbf{x}^{\odot(1-\frac{2}{L})}\rangle & \text{if } L > 2. \end{cases}$$

Below $\mathbf{x}$ will take the role of the initialization $\tilde{\mathbf{x}}(0)$, and $g_{\tilde{\mathbf{x}}(0)}$ encodes the dependence of the limit $\lim_{t\to\infty}\tilde{\mathbf{x}}(t)$.

Theorem 2.6. *For $L \geq 2$, let $\tilde{\mathbf{x}}(t) = \mathbf{x}(t)^{\odot L}$ and*

$$\mathbf{x}'(t) = -\nabla\mathcal{L}(\mathbf{x}(t)) = -L\left[\mathbf{A}^{T}(\mathbf{A}\mathbf{x}^{\odot L}(t) - \mathbf{y})\right] \odot \mathbf{x}^{\odot L-1}(t) \tag{20}$$

with $\mathbf{x}(0) \geq 0$. Assume that S_+ in (13) is non-empty. Then $\tilde{\mathbf{x}}_\infty := \lim_{t\to\infty}\tilde{\mathbf{x}}(t)$ exists and

$$\tilde{\mathbf{x}}_\infty = \operatorname*{arg\,min}_{\mathbf{z}\in S_+} g_{\tilde{\mathbf{x}}(0)}(\mathbf{z}). \tag{21}$$

The proof of this theorem is based on the Bregman divergence defined as follows.

Definition 2.7 (Bregman Divergence). *Let $F : \Omega \to \mathbb{R}$ be a continuously-differentiable, strictly convex function defined on a closed convex set Ω. The Bregman divergence associated with F for points $p, q \in \Omega$ is defined as*

$$D_F(p,q) = F(p) - F(q) - \langle \nabla F(q), p - q\rangle. \tag{22}$$

By strict convexity of F it is straight-forward to verify the following.

Lemma 2.8 ([3]). *The Bregman divergence D_F is non-negative and, for any $q \in \Omega$, the function $p \mapsto D_F(p,q)$ is strictly convex.*

To prove Theorem 2.6, we use the Bregman divergence with respect to function $F : \mathbb{R}_+^N \to \mathbb{R}$ with

$$F(\mathbf{x}) = \left\{ \begin{matrix} \frac{1}{2}\langle \mathbf{x} \odot \log(\mathbf{x}) - \mathbf{x}, \mathbf{1}\rangle & \text{if } L = 2 \\ \frac{L}{2(2-L)}\langle \mathbf{x}^{\odot \frac{2}{L}}, \mathbf{1}\rangle & \text{if } L > 2 \end{matrix} \right\} = \begin{cases} \frac{1}{2}\sum_{n=1}^N x_n \log(x_n) - x_n & \text{if } L = 2 \\ \frac{L}{2(2-L)}\sum_{n=1}^N x_n^{\frac{2}{L}} & \text{if } L > 2, \end{cases} \tag{23}$$

with the understanding that $z\log(z) = 0$ for $z = 0$. Note that F is strictly convex because its Hessian $\mathbf{H}_F(\mathbf{x})$ is diagonal with positive diagonal entries $\frac{1}{L}x_n^{-2+\frac{2}{L}}$, for $\mathbf{x}$ in the interior of $\mathbb{R}_+^N$, i.e., if $x_n > 0$ for all n. Thus the Bregman divergence D_F is well-defined. It has the following property.

Lemma 2.9. *Let F be the function defined in (23), $\mathbf{z} \geq \mathbf{0}$ be fixed, and $\mathbf{x}(t)\colon \mathbb{R}_+ \to \mathbb{R}_+^N$ be a continuous function with $\mathbf{x}(0) > \mathbf{0}$ and $\|\mathbf{x}(t)\|_2 \to \infty$. Then $D_F(\mathbf{z}, \mathbf{x}(t)) \to \infty$.*

Proof. Since

$$\nabla F(\mathbf{x}) = \begin{cases} \frac{1}{2}\log(\mathbf{x}) & \text{if } L = 2 \\ \frac{1}{(2-L)}\mathbf{x}^{\odot(\frac{2}{L}-1)} & \text{if } L > 2 \end{cases},$$

we obtain

$$\begin{aligned} D_F(\mathbf{z}, \mathbf{x}(t)) &= \begin{cases} \frac{1}{2}\langle \mathbf{z} \odot \log(\mathbf{z}) - \mathbf{z} - \mathbf{x}(t) \odot \log(\mathbf{x}(t)) + \mathbf{x}(t), \mathbf{1}\rangle - \frac{1}{2}\langle \log(\mathbf{x}(t)), \mathbf{z} - \mathbf{x}(t)\rangle & \text{if } L = 2, \\ \frac{L}{2(2-L)}\langle \mathbf{z}^{\odot\frac{2}{L}} - \mathbf{x}(t)^{\odot\frac{2}{L}}, \mathbf{1}\rangle - \frac{1}{2-L}\langle \mathbf{x}(t)^{\odot(\frac{2}{L}-1)}, \mathbf{z} - \mathbf{x}(t)\rangle & \text{if } L > 2, \end{cases} \\ &= \begin{cases} \frac{1}{2}\big(\langle \mathbf{x}(t) - \mathbf{z} \odot \log(\mathbf{x}(t)), \mathbf{1}\rangle + \langle \mathbf{z} \odot \log(\mathbf{z}) - \mathbf{z}, \mathbf{1}\rangle\big) & \text{if } L = 2, \\ \frac{1}{L-2}\big(\langle (\frac{L}{2} - 1)\mathbf{x}(t)^{\odot\frac{2}{L}} + \mathbf{z} \odot \mathbf{x}(t)^{\odot(\frac{2}{L}-1)}, \mathbf{1}\rangle - \frac{L}{2}\langle \mathbf{z}^{\odot\frac{2}{L}}, \mathbf{1}\rangle\big) & \text{if } L > 2. \end{cases} \end{aligned}$$

If $\|\mathbf{x}(t)\|_2 \to \infty$, we know that $x_n(t) \to \infty$, for some $n \in [N]$. For $z \geq 0$, both the function $x \mapsto x - z\log(x)$ (for $L = 2$) and the function $x \mapsto (\frac{L}{2} - 1)x^{\frac{2}{L}} + zx^{\frac{2}{L}-1}$ (for $L > 2$) are bounded from below and grow to infinity for $x \to \infty$. □

Proof of Theorem 2.6. Let us begin with a brief outline of the proof. We will first compute the time derivative of $D_F(\mathbf{z}, \tilde{\mathbf{x}}(t))$, for $\mathbf{z} \in S_+$. Using this, we can show that $\lim_{t\to\infty} \mathcal{L}(\tilde{\mathbf{x}}(t)^{\odot\frac{1}{L}}) = 0$, which implies that $\mathbf{A}\tilde{\mathbf{x}}(t) \to \mathbf{y}$. Employing Lemma 2.9 we will also deduce convergence of $\tilde{\mathbf{x}}(t)$ and characterize the limit $\tilde{\mathbf{x}}_\infty$ as the unique $\mathbf{z} \in S_+$ satisfying $\lim_{t\to\infty} D_F(\tilde{\mathbf{x}}_\infty, \tilde{\mathbf{x}}(t)) = 0$. Finally, we use this characterization of $\tilde{\mathbf{x}}_\infty$ to obtain (21).

We start by computing $\partial_t D_F(\mathbf{z}, \tilde{\mathbf{x}}(t))$. Note that due to continuity $\mathbf{x}(t) \geq 0$, for all t, and hence $\mathbf{x}$ remains non-negative (if any entry $x(t)_i$ vanishes at $t_* > 0$, then $x(t)_i = 0$ for all $t \geq t_*$ by the shape of (20)). Suppose $\mathbf{z} \in S_+$. Then, we have

$$\begin{aligned} \partial_t D_F(\mathbf{z}, \tilde{\mathbf{x}}(t)) &= \partial_t \left[F(\mathbf{z}) - F(\tilde{\mathbf{x}}(t)) - \langle \nabla F(\tilde{\mathbf{x}}(t)), \mathbf{z} - \tilde{\mathbf{x}}(t)\rangle\right] \\ &= 0 - \langle \nabla F(\tilde{\mathbf{x}}(t)), \tilde{\mathbf{x}}'(t)\rangle - \langle \partial_t \nabla F(\tilde{\mathbf{x}}(t)), \mathbf{z} - \tilde{\mathbf{x}}(t)\rangle + \langle \nabla F(\tilde{\mathbf{x}}(t)), \tilde{\mathbf{x}}'(t)\rangle \\ &= -\langle \partial_t \nabla F(\tilde{\mathbf{x}}(t)), \mathbf{z} - \tilde{\mathbf{x}}(t)\rangle. \end{aligned}$$

By the chain rule and the diagonal shape of $\mathbf{H}_F$, we know that

$$\begin{aligned} \partial_t \nabla F(\tilde{\mathbf{x}}(t)) &= \mathbf{H}_F(\tilde{\mathbf{x}}(t)) \cdot \tilde{\mathbf{x}}'(t) = \frac{1}{L}\tilde{\mathbf{x}}(t)^{\odot(-2+\frac{2}{L})} \odot \tilde{\mathbf{x}}'(t) = \frac{1}{L}\mathbf{x}(t)^{\odot(-2L+2)} \odot \big(L\mathbf{x}(t)^{\odot(L-1)} \odot \mathbf{x}'(t)\big) \\ &= \mathbf{x}(t)^{\odot(-2L+2)} \odot \mathbf{x}(t)^{\odot(L-1)} \odot \big(- L\left[\mathbf{A}^{\mathrm{T}}(\mathbf{A}\mathbf{x}^{\odot L}(t) - \mathbf{y})\right] \odot \mathbf{x}^{\odot L-1}(t)\big) \\ &= -L\left[\mathbf{A}^{\mathrm{T}}(\mathbf{A}\tilde{\mathbf{x}}(t) - \mathbf{y})\right]. \end{aligned}$$

Therefore,

$$\begin{aligned}\partial_t D_F(\mathbf{z}, \tilde{\mathbf{x}}(t)) &= L\langle \mathbf{A}^{\mathrm{T}}(\mathbf{A}\tilde{\mathbf{x}}(t) - \mathbf{y}), \mathbf{z} - \tilde{\mathbf{x}}(t)\rangle = -L\langle \mathbf{A}\tilde{\mathbf{x}}(t) - \mathbf{y}, \mathbf{A}\tilde{\mathbf{x}}(t) - \mathbf{A}\mathbf{z}\rangle = -L\|\mathbf{A}\tilde{\mathbf{x}}(t) - \mathbf{y}\|_2^2 \\ &= -2L\mathcal{L}(\tilde{\mathbf{x}}(t)^{\odot\frac{1}{L}}). \end{aligned} \tag{24}$$

We now show that $\lim_{t\to\infty} \mathcal{L}(\tilde{\mathbf{x}}(t)^{\odot\frac{1}{L}}) = 0$. Since $D_F(\mathbf{z}, \tilde{\mathbf{x}}(t)) \geq 0$ and $\partial_t D_F(\mathbf{z}, \tilde{\mathbf{x}}(t)) \leq 0$, $D_F(\mathbf{z}, \tilde{\mathbf{x}}(t))$ must converge for any fixed $\mathbf{z} \in S_+$. Let us assume that $\lim_{t\to\infty} \mathcal{L}(\tilde{\mathbf{x}}(t)^{\odot\frac{1}{L}}) \neq 0$. Since $\mathcal{L}(\tilde{\mathbf{x}}(t)^{\odot\frac{1}{L}})$ is non-negative and decreasing, this is equivalent to the existence of $\varepsilon > 0$ with $\mathcal{L}(\tilde{\mathbf{x}}(t)^{\odot\frac{1}{L}}) \geq \varepsilon$, for all $t \geq 0$. But then, for all $T > 0$,

$$D_F(\mathbf{z}, \tilde{\mathbf{x}}(T)) - D_F(\mathbf{z}, \tilde{\mathbf{x}}(0)) = \int_0^T \partial_t D_F(\mathbf{z}, \tilde{\mathbf{x}}(t))dt = -2L\int_0^T \mathcal{L}(\tilde{\mathbf{x}}(t)^{\odot\frac{1}{L}})dt \leq -2\varepsilon LT,$$

which contradicts convergence of $D_F(\mathbf{z}, \tilde{\mathbf{x}}(t))$ for $t \to \infty$. We conclude that $\lim_{t\to\infty} \mathcal{L}(\tilde{\mathbf{x}}(t)^{\odot\frac{1}{L}}) = 0$ and thus $\lim_{t\to\infty} \mathbf{A}\tilde{\mathbf{x}}(t) = \mathbf{y}$.

Furthermore, we can deduce that $\|\tilde{\mathbf{x}}(t)\|_2$ is bounded. To see this, fix any $\bar{\mathbf{z}} \in S_+$ and notice that $D_F(\bar{\mathbf{z}}, \tilde{\mathbf{x}}(t)) \to \infty$ if $\|\tilde{\mathbf{x}}(t)\|_2 \to \infty$ which contradicts the fact that $0 \leq D_F(\bar{\mathbf{z}}, \tilde{\mathbf{x}}(t)) \leq D_F(\bar{\mathbf{z}}, \tilde{\mathbf{x}}(0))$, cf. Lemma 2.9. Let us denote by B a sufficiently large compact ball around the origin such that $B \cap S_+ \neq \emptyset$ and $\tilde{\mathbf{x}}(t) \in B$, for all $t \geq 0$.

Now assume that there exists no $\mathbf{z} \in S_+ \cap B$ such that $\lim_{t\to\infty} D_F(\mathbf{z}, \tilde{\mathbf{x}}(t)) = 0$, i.e., by compactness of $S_+ \cap B$ there exists $\varepsilon > 0$ such that $\lim_{t\to\infty} D_F(\mathbf{z}, \tilde{\mathbf{x}}(t)) > \varepsilon$, for all $\mathbf{z} \in S_+ \cap B$. Then $\mathcal{L}(\tilde{\mathbf{x}}(t)^{\odot\frac{1}{L}})$ cannot converge to zero because $\tilde{\mathbf{x}}$ is bounded away from the solution set S_+ on the region of interest B. This contradicts $\lim_{t\to\infty} \mathbf{A}\tilde{\mathbf{x}}(t) = \mathbf{y}$ and shows the existence of $\mathbf{z} \in S_+$ such that $\lim_{t\to\infty} D_F(\mathbf{z}, \tilde{\mathbf{x}}(t)) = 0$.

For any such $\mathbf{z}$, let us assume that $\tilde{\mathbf{x}}(t) \nrightarrow \mathbf{z}$. Then there exists $\varepsilon > 0$ and a sequence of times $t_0, t_1, \dots$ with $\|\mathbf{z} - \tilde{\mathbf{x}}(t_k)\|_2 \geq \varepsilon$ and $\lim_{k\to\infty} D_F(\mathbf{z}, \tilde{\mathbf{x}}(t_k)) = 0$. Since $\tilde{\mathbf{x}}(t_k)$ is a bounded sequence, a (not relabeled) subsequence converges to some $\bar{\mathbf{x}}$ with $\|\mathbf{z} - \bar{\mathbf{x}}\|_2 \geq \varepsilon$ and $D_F(\mathbf{z}, \bar{\mathbf{x}}) = 0$. Since $D_F(\bar{\mathbf{x}}, \bar{\mathbf{x}}) = 0$ and D_F is non-negative, this is a contradiction to the strict convexity of $D_F(\cdot, \bar{\mathbf{x}})$. Hence, $\tilde{\mathbf{x}}_\infty = \lim_{t\to\infty} \tilde{\mathbf{x}}(t) \in S_+$ exists and is the unique solution satisfying $\lim_{t\to\infty}\ D_F\ (\tilde{\mathbf{x}}_\infty, \tilde{\mathbf{x}}(t)) = 0$.

Because $\partial_t D_F(\mathbf{z}, \tilde{\mathbf{x}}(t))$ is identical for all $\mathbf{z} \in S_+$, the difference

$$\Delta_{\mathbf{z}} = D_F(\mathbf{z}, \tilde{\mathbf{x}}(0)) - D_F(\mathbf{z}, \tilde{\mathbf{x}}_\infty) \tag{25}$$

is constant in $\mathbf{z} \in S_+$. By non-negativity of D_F,

$$D_F(\mathbf{z}, \tilde{\mathbf{x}}(0)) \geq \Delta_{\mathbf{z}} = \Delta_{\tilde{\mathbf{x}}_\infty} = D_F(\tilde{\mathbf{x}}_\infty, \tilde{\mathbf{x}}(0)). \tag{26}$$

Thus

$$\begin{aligned}\tilde{\mathbf{x}}_\infty \in \arg\min_{\mathbf{z}\in S_+} D_F(\mathbf{z}, \tilde{\mathbf{x}}(0)) &= \arg\min_{\mathbf{z}\in S_+} F(\mathbf{z}) - F(\tilde{\mathbf{x}}(0)) - \langle \nabla F(\tilde{\mathbf{x}}(0)), \mathbf{z} - \tilde{\mathbf{x}}(0)\rangle \\ &= \arg\min_{\mathbf{z}\in S_+} F(\mathbf{z}) - \langle \nabla F(\tilde{\mathbf{x}}(0)), \mathbf{z}\rangle \\ &= \arg\min_{\mathbf{z}\in S_+} \begin{cases} \sum_{n=1}^N z_n \log(z_n) - z_n - \log(\tilde{x}_n(0))z_n & \text{if } L = 2, \\ \sum_{n=1}^N -z_n^{\frac{2}{L}} + \frac{2}{L}\tilde{x}_n(0)^{-1+\frac{2}{L}} z_n & \text{if } L > 2, \end{cases} \\ &= \arg\min_{\mathbf{z}\in S_+} \begin{cases} \langle \mathbf{z}, \log(\mathbf{z}) - \mathbf{1} - \log(\tilde{\mathbf{x}}(0))\rangle & \text{if } L = 2, \\ 2\|\mathbf{z}\|_1 - L\langle \mathbf{z}^{\frac{2}{L}}, \tilde{\mathbf{x}}(0)^{1-\frac{2}{L}}\rangle & \text{if } L > 2. \end{cases}\end{aligned}$$

This completes the proof. □

Proof of Theorem 2.1. By Lemma 2.4, the assumption that $\mathbf{x}^{(k)}(0) = \mathbf{x}^{(k')}(0)$, for all $k, k' \in [L]$, implies that $\mathbf{x}^{(k)}(t) = \mathbf{x}^{(k')}(t)$, for all $t \geq 0$. Furthermore, each $\mathbf{x}^{(k)}(t)$ equals $\mathbf{x}(t)$ defined via $\mathbf{x}(0) = \mathbf{x}^{(k)}(0)$ and $\mathbf{x}'(t) = -\nabla\mathcal{L}(\mathbf{x})$. By Theorem 2.6, the limit $\mathbf{x}_\infty := \lim_{t\to\infty} \mathbf{x}(t)$ exists and $\tilde{\mathbf{x}}_\infty = \mathbf{x}_\infty^{\odot L}$ lies in S_+. The remainder of the proof uses (21) to deduce (16).

Let $\mathbf{z} \in S_+$. Due to non-negativity, $\|\tilde{\mathbf{x}}_\infty\|_1 = \sum_{n\in[N]} (\tilde{x}_\infty)_n$ and $\|\mathbf{z}\|_1 = \sum_{n\in[N]} z_n$. Again, we will separate the case $L = 2$ from the case $L > 2$. Assume $0 < \alpha < e^{-\frac{1}{2}}$ for $L = 2$ and $\alpha > 0$ for $L > 2$. Define the quantity

$$0 < \delta_{L,\alpha} := \begin{cases} -\frac{1}{2\log(\alpha)+1} & \text{if } L = 2, \\ \frac{L}{2}\alpha^{L-2} & \text{if } L > 2\,. \end{cases} \tag{27}$$

which goes to zero as α goes to zero. For $L = 2$, by Theorem 2.6 the optimality of $\tilde{\mathbf{x}}_\infty$ implies that

$$\langle \log(\tilde{\mathbf{x}}_\infty), \tilde{\mathbf{x}}_\infty\rangle + \delta_{L,\alpha}^{-1}\langle \mathbf{1}, \tilde{\mathbf{x}}_\infty\rangle \leq \langle \log(\mathbf{z}), \mathbf{z}\rangle + \delta_{L,\alpha}^{-1}\langle \mathbf{1}, \mathbf{z}\rangle \tag{28}$$

and hence (recall that one has $\langle \mathbf{z}, \mathbf{1}\rangle = \|\mathbf{z}\|_1$, for $\mathbf{z} \geq \mathbf{0}$)

$$\|\tilde{\mathbf{x}}_\infty\|_1 - \|\mathbf{z}\|_1 \leq \delta_{L,\alpha}(\langle \log(\mathbf{z}), \mathbf{z}\rangle - \langle \log(\tilde{\mathbf{x}}_\infty), \tilde{\mathbf{x}}_\infty\rangle).$$

Since $\xi^2 \geq \xi\log(\xi) \geq -e^{-1}$ for $\xi \geq 0$,

$$\|\tilde{\mathbf{x}}_\infty\|_1 - \|\mathbf{z}\|_1 \leq \delta_{L,\alpha}(\langle \mathbf{z}, \mathbf{z}\rangle - \langle -e^{-1}\mathbf{1}, \mathbf{1}\rangle) = \delta_{L,\alpha}(\|\mathbf{z}\|_2^2 + Ne^{-1}) \leq \delta_{L,\alpha}(\|\mathbf{z}\|_1^2 + Ne^{-1}).$$

For $L > 2$, by Theorem 2.6 the optimality of $\tilde{\mathbf{x}}_\infty$ implies that

$$2\langle \mathbf{1}, \tilde{\mathbf{x}}_\infty\rangle - 2\delta_{\alpha,L}\langle \mathbf{1}, \tilde{\mathbf{x}}_\infty^{\odot\frac{2}{L}}\rangle \leq 2\langle \mathbf{1}, \mathbf{z}\rangle - 2\delta_{\alpha,L}\langle \mathbf{1}, \mathbf{z}^{\odot\frac{2}{L}}\rangle \tag{29}$$

and hence

$$\begin{aligned} \|\tilde{\mathbf{x}}_\infty\|_1 - \|\mathbf{z}\|_1 &\leq \delta_{L,\alpha}(\langle \mathbf{1}, \tilde{\mathbf{x}}_\infty^{\odot\frac{2}{L}}\rangle - \langle \mathbf{1}, \mathbf{z}^{\odot\frac{2}{L}}\rangle) \\ &\leq \delta_{L,\alpha}\langle \mathbf{1}, \tilde{\mathbf{x}}_\infty^{\odot\frac{2}{L}}\rangle \leq \delta_{L,\alpha}\langle \mathbf{1}, \tilde{\mathbf{x}}_\infty + \mathbf{1}\rangle = \delta_{L,\alpha}(\|\tilde{\mathbf{x}}_\infty\|_1 + N). \end{aligned} \tag{30}$$

The above shows that for all $\mathbf{z} \in S_+$

$$\|\tilde{\mathbf{x}}_\infty\|_1 \leq \begin{cases} \|\mathbf{z}\|_1 + \delta_{L,\alpha}(\|\mathbf{z}\|_1^2 + Ne^{-1}) & \text{if } L = 2, \\ \frac{\|\mathbf{z}\|_1}{1-\delta_{L,\alpha}} + \frac{N\delta_{L,\alpha}}{1-\delta_{L,\alpha}} & \text{if } L > 2. \end{cases}$$

Taking the minimum over $\mathbf{z} \in S_+$ and rearranging the terms we get that

$$\begin{aligned} \|\tilde{\mathbf{x}}_\infty\|_1 - Q_+ &\leq \delta_{L,\alpha} \cdot \begin{cases} Q_+^2 + Ne^{-1} & \text{if } L = 2, \\ \frac{Q_+ + N}{1-\delta_{L,\alpha}} & \text{if } L > 2, \end{cases} \\ &= \begin{cases} -\frac{1}{2\log(\alpha)+1}(Q_+^2 + Ne^{-1}) & \text{if } L = 2, \\ \frac{\frac{L}{2}\alpha^{L-2}(Q_+ + N)}{1-\frac{L}{2}\alpha^{L-2}} & \text{if } L > 2, \end{cases} \end{aligned}$$

where $Q_+ = \min_{\mathbf{z}\in S_+} \|\mathbf{z}\|_1$. The right hand side is bounded by $\varepsilon > 0$ if and only if

$$\alpha \leq \begin{cases} \min\left(e^{-\frac{1}{2}}, \exp\left(\frac{1}{2} - \frac{Q_+^2 + Ne^{-1}}{2\varepsilon}\right)\right) & \text{if } L = 2, \\ \left(\frac{2\varepsilon}{L(Q_+ + N + \varepsilon)}\right)^{\frac{1}{L-2}} & \text{if } L > 2, \end{cases}$$

and we arrive at the conclusion of the theorem. □

3 General Case

This section is dedicated to the proof of Theorem 1.1. Since the proof strategy is very similar to the one of Theorem 2.1, we will not replicate all arguments, but only highlight the key ideas. We use the following additional notation in this section. Let

$$S = \{\mathbf{z} \in \mathbb{R}^N : \mathbf{A}\mathbf{z} = \mathbf{y}\} \tag{31}$$

For $\mathbf{z} \in \mathbb{R}^N$, we denote $I_{\mathbf{z},+} = \{n : z_n > 0\}$ and $I_{\mathbf{z},-} = \{n : z_n < 0\}$ the index sets corresponding to positive and negative entries of $\mathbf{z}$. We decompose $\mathbf{z} = \mathbf{z}_+ - \mathbf{z}_-$ with

$$(z_+)_n = \begin{cases} z_n & \text{if } z_n > 0, \\ 0 & \text{otherwise,} \end{cases} \qquad \text{and} \qquad (z_-)_n = \begin{cases} -z_n & \text{if } z_n < 0, \\ 0 & \text{otherwise.} \end{cases}$$

We furthermore define

$$S_\pm = \{(\mathbf{z}_+, \mathbf{z}_-) : \mathbf{z}_+, \mathbf{z}_- \geq 0, \mathbf{A}(\mathbf{z}_+ - \mathbf{z}_-) = \mathbf{y}\}. \tag{32}$$

as an alternative representation of the solution set S.

3.1 Generalization of Reduced Factorized Loss

It is straight-forward to check that the negative flow of the general overparameterized loss function

$$\mathcal{L}^\pm_{\text{over}}(\mathbf{u}^{(1)}, \dots, \mathbf{u}^{(L)}, \mathbf{v}^{(1)}, \dots, \mathbf{v}^{(L)}) = \Big\| \mathbf{A}\big(\bigodot_{k=1}^{L} \mathbf{u}^{(k)} - \bigodot_{k=1}^{L} \mathbf{v}^{(k)} \big) - \mathbf{y} \Big\|_2^2 \tag{33}$$

is equivalent to the flow of the following general reduced factorized loss if all $\mathbf{u}^{(k)}$ and $\mathbf{v}^{(k)}$ are identically initialized.

Definition 3.1 (General Reduced Factorized Loss). *Let $L \in \mathbb{N}$, $L \geq 2$. For $\mathbf{A} \in \mathbb{R}^{M \times N}$ and $\mathbf{y} \in \mathbb{R}^M$, the* general reduced factorized loss function *is defined as*

$$\mathcal{L}^\pm : \mathbb{R}^N \times \mathbb{R}^N \to [0, \infty), \qquad \mathcal{L}^\pm(\mathbf{u}, \mathbf{v}) = \frac{1}{2} \| \mathbf{A}(\mathbf{u}^{\odot L} - \mathbf{v}^{\odot L}) - \mathbf{y} \|_2^2. \tag{34}$$

Note that the partial gradients of $\mathcal{L}^\pm$ are given by

$$\begin{aligned} &\nabla_{\mathbf{u}} \mathcal{L}^\pm(\mathbf{u}, \mathbf{v}) = L \left[\mathbf{A}^{\mathrm{T}}(\mathbf{A}(\mathbf{u}^{\odot L} - \mathbf{v}^{\odot L}) - \mathbf{y}) \right] \odot \mathbf{u}^{\odot L-1} \\ \text{and} \quad &\nabla_{\mathbf{v}} \mathcal{L}^\pm(\mathbf{u}, \mathbf{v}) = -L \left[\mathbf{A}^{\mathrm{T}}(\mathbf{A}(\mathbf{u}^{\odot L} - \mathbf{v}^{\odot L}) - \mathbf{y}) \right] \odot \mathbf{v}^{\odot L-1}. \end{aligned}$$

3.2 General Bregman Divergence

In order to prove that $\mathbf{u}^{\odot L} - \mathbf{v}^{\odot L}$ converges to an element in S, we will again work with the Bregman Divergence. The main difference is that, instead of F as defined in (23), we use the function $F^\pm : \mathbb{R}^N_+ \times \mathbb{R}^N_+ \to \mathbb{R}$ with

$$F^\pm(\mathbf{u}, \mathbf{v}) = F(\mathbf{u}) + F(\mathbf{v}). \tag{35}$$

Note that $F^\pm$ is strictly convex because F is strictly convex and hence the Bregman divergence $D_{F^\pm} : (\mathbb{R}^N_+ \times \mathbb{R}^N_+)^2 \to \mathbb{R}$ is well-defined.

Theorem 3.2. *Let* $(\tilde{\mathbf{u}}(t),\, \tilde{\mathbf{v}}(t)) = (\mathbf{u}(t)^{\odot L},\, \mathbf{v}(t)^{\odot L})$ *and*

$$\mathbf{u}'(t) = -\nabla_{\mathbf{u}}\mathcal{L}^{\pm}(\mathbf{u},\, \mathbf{v}), \quad \mathbf{v}'(t) = -\nabla_{\mathbf{v}}\mathcal{L}^{\pm}(\mathbf{u},\, \mathbf{v}) \tag{36}$$

with $\mathbf{u}(0), \mathbf{v}(0) \geq 0$*. Then the limit* $(\tilde{\mathbf{u}}_\infty, \tilde{\mathbf{v}}_\infty) := \lim_{t\to\infty}(\tilde{\mathbf{u}}(t),\, \tilde{\mathbf{v}}(t))$ *exists and*

$$(\tilde{\mathbf{u}}_\infty, \tilde{\mathbf{v}}_\infty) \in \operatorname*{arg\,min}_{(\mathbf{z}_+,\, \mathbf{z}_-)\in S_\pm} \; g_{\tilde{\mathbf{u}}(0)}(\mathbf{z}_+) + g_{\tilde{\mathbf{v}}(0)}(\mathbf{z}_-) \tag{37}$$

where g *and* $S_\pm$ *are defined in* (21) *and* (32)*.*

Proof. The proof follows the same steps as the proof of Theorem 2.6, i.e., we start with computing $\partial_t D_{F^\pm}((\mathbf{z}_+,\, \mathbf{z}_-),\, (\tilde{\mathbf{u}}(t),\, \tilde{\mathbf{v}}(t)))$. As before, note that due to continuity $\mathbf{u}(t)$ and $\mathbf{v}(t)$ remain non-negative for all $t \geq 0$. Let $\tilde{\mathbf{u}} = \mathbf{u}^{\odot L}$ and $\tilde{\mathbf{v}} = \mathbf{v}^{\odot L}$. Suppose $\mathbf{Az} = \mathbf{y}$. Decompose $\mathbf{z}$ into $\mathbf{z}_+ - \mathbf{z}_-$ where $\mathbf{z}_+ \geq \mathbf{0}$ and $\mathbf{z}_- \geq \mathbf{0}$ contain the positive resp. negative entries of $\mathbf{z}$ in absolute value and are zero elsewhere. We now compute the time derivative of $D_{F^\pm}((\mathbf{z}_+,\, \mathbf{z}_-), (\mathbf{u}(t),\, \mathbf{v}(t)))$. By linearity

$$\begin{aligned} D_{F^\pm}((\mathbf{z}_+,\, \mathbf{z}_-),\, (\tilde{\mathbf{u}}(t),\, \tilde{\mathbf{v}}(t))) &= F(\mathbf{z}_+) + F(\mathbf{z}_-) - F(\tilde{\mathbf{u}}(t)) - F(\tilde{\mathbf{v}}(t)) \\ &\quad - [\langle \nabla_{\tilde{\mathbf{u}}} F(\tilde{\mathbf{u}}(t)),\, \mathbf{z}_+ - \tilde{\mathbf{u}}(t)\rangle + \langle \nabla_{\tilde{\mathbf{v}}} F(\tilde{\mathbf{v}}(t)),\, \mathbf{z}_- - \tilde{\mathbf{v}}(t)\rangle] \\ &= D_F(\mathbf{z}_+,\, \tilde{\mathbf{u}}(t)) + D_F(\mathbf{z}_-,\, \tilde{\mathbf{v}}(t)). \end{aligned}$$

By a similar calculation as in the proof of Theorem 2.6, we have

$$\begin{aligned} &\frac{1}{L}\partial_t D_{F^\pm}((\mathbf{z}_+,\, \mathbf{z}_-),\, (\tilde{\mathbf{u}}(t),\, \tilde{\mathbf{v}}(t))) \\ &\quad = \langle \mathbf{A}^{\mathrm{T}}(\mathbf{A}(\tilde{\mathbf{u}}(t) - \tilde{\mathbf{v}}(t)) - \mathbf{y}), \mathbf{z}_+ - \tilde{\mathbf{u}}(t)\rangle - \langle \mathbf{A}^{\mathrm{T}}(\mathbf{A}(\tilde{\mathbf{u}}(t) - \tilde{\mathbf{v}}(t)) - \mathbf{y}), \mathbf{z}_- - \tilde{\mathbf{v}}(t)\rangle \\ &\quad = \langle \mathbf{A}(\tilde{\mathbf{u}}(t) - \tilde{\mathbf{v}}(t)) - \mathbf{y}, \mathbf{A}(\mathbf{z}_+ - \tilde{\mathbf{u}}(t) - \mathbf{z}_- + \tilde{\mathbf{v}}(t))\rangle = -\|\mathbf{A}(\tilde{\mathbf{u}}(t) - \tilde{\mathbf{v}}(t)) - \mathbf{y}\|_2^2 \\ &\quad = -2\mathcal{L}^{\pm}(\tilde{\mathbf{u}}(t)^{\odot\frac{1}{L}}, \tilde{\mathbf{v}}(t)^{\odot\frac{1}{L}}). \end{aligned}$$

The same line of argument as in the proof of Theorem 2.6 yields $\lim_{t\to\infty}\mathcal{L}^{\pm}(\tilde{\mathbf{u}}(t)^{\odot\frac{1}{L}},\, \tilde{\mathbf{v}}(t)^{\odot\frac{1}{L}}) = 0$, that $\lim_{t\to\infty}\mathbf{A}(\tilde{\mathbf{u}}(t) - \tilde{\mathbf{v}}(t)) = \mathbf{y}$, the limit $(\tilde{\mathbf{u}}_\infty, \tilde{\mathbf{v}}_\infty) := \lim_{t\to\infty}(\tilde{\mathbf{u}}(t),\, \tilde{\mathbf{v}}(t))$ exists and both components are non-negative. Since the time derivative $\partial_t D_{F^\pm}$ is identical for all $(\mathbf{z}_+,\, \mathbf{z}_-) \in S_\pm$, the difference

$$\Delta_{\mathbf{z}_+,\, \mathbf{z}_-} = D_{F^\pm}((\mathbf{z}_+,\, \mathbf{z}_-), (\tilde{\mathbf{u}}(0),\, \tilde{\mathbf{v}}(0))) - D_{F^\pm}((\mathbf{z}_+,\, \mathbf{z}_-), (\tilde{\mathbf{u}}_\infty,\, \tilde{\mathbf{v}}_\infty))$$

is also identical for all $(\mathbf{z}_+,\, \mathbf{z}_-) \in S_\pm$. By non-negativity of $D_{F^\pm}$,

$$D_{F^\pm}((\mathbf{z}_+,\, \mathbf{z}_-),\, (\tilde{\mathbf{u}}(0),\, \tilde{\mathbf{v}}(0))) \geq \Delta_{\mathbf{z}_+,\, \mathbf{z}_-} = \Delta_{\tilde{\mathbf{u}}_\infty,\, \tilde{\mathbf{v}}_\infty} = D_{F^\pm}((\tilde{\mathbf{u}}_\infty,\, \tilde{\mathbf{v}}_\infty),\, (\tilde{\mathbf{u}}(0),\, \tilde{\mathbf{v}}(0))).$$

Expanding the expression we obtain

$$\begin{aligned} (\tilde{\mathbf{u}}_\infty, \tilde{\mathbf{v}}_\infty) &\in \operatorname*{arg\,min}_{(\mathbf{z}_+,\, \mathbf{z}_-)\in S_\pm} \; D_{F^\pm}((\mathbf{z}_+,\, \mathbf{z}_-),\, (\tilde{\mathbf{u}}(0),\, \tilde{\mathbf{v}}(0))) = \operatorname*{arg\,min}_{(\mathbf{z}_+,\, \mathbf{z}_-)\in S_\pm} \; D_F(\mathbf{z}_+,\, \tilde{\mathbf{u}}(0)) + D_F(\mathbf{z}_-,\, \tilde{\mathbf{v}}(0)) \\ &= \operatorname*{arg\,min}_{(\mathbf{z}_+,\, \mathbf{z}_-)\in S_\pm} \; g_{\tilde{\mathbf{u}}(0)}(\mathbf{z}_+) + g_{\tilde{\mathbf{v}}(0)}(\mathbf{z}_-) \end{aligned}$$

by the same calculation as in Theorem 2.6. □

Proof of Theorem 1.1. Suppose $\mathbf{u}(0) = \mathbf{v}(0) = \alpha\mathbf{1}$ with $\alpha > 0$. By Theorem 3.2, $(\tilde{\mathbf{u}}_\infty, \tilde{\mathbf{v}}_\infty)$ exists and satisfies (37). We will use (37) in order to deduce (8) in the remainder. The proof works similarly as the one of Theorem 2.1. It is, however, slightly more involved on a technical level since the supports of $\tilde{\mathbf{u}}_\infty$ and $\tilde{\mathbf{v}}_\infty$ might intersect, i.e., $(\tilde{\mathbf{u}}_\infty)_i > 0$ and $(\tilde{\mathbf{v}}_\infty)_i > 0$ may occur for some i.

Let $\tilde{\mathbf{x}}_\infty = \tilde{\mathbf{u}}_\infty - \tilde{\mathbf{v}}_\infty$ and fix $\mathbf{z} \in S$. We consider the unique decomposition $\mathbf{z} = \mathbf{z}_+ - \mathbf{z}_-$ where $(\mathbf{z}_+)_i = z_i$ and $(\mathbf{z}_-)_i = 0$ if $z_i \geq 0$ and $(\mathbf{z}_+)_i = 0$ and $(\mathbf{z}_-)_i = -z_i$ otherwise. Clearly, $\mathbf{z}_+$ and $\mathbf{z}_-$ have disjoint support. We consider the cases $L = 2$ and $L > 2$ separately. Recall the definition of $\delta_{L,\alpha} > 0$ from (27), i.e.,

$$0 < \delta_{L,\alpha} := \begin{cases} -\frac{1}{2\log(\alpha)+1} & \text{if } L = 2 \\ \frac{L}{2}\alpha^{L-2} & \text{if } L > 2\,. \end{cases}$$

Since $|\tilde{\mathbf{x}}_\infty| = |\tilde{\mathbf{u}}_\infty - \tilde{\mathbf{v}}_\infty|$ (where $|\cdot|$ is applied entry-wise) a componentwise application of Lemma A.2 with $\rho = \delta_{L,\alpha}^{-1}$ gives

$$\begin{aligned}
&\langle \log(|\tilde{\mathbf{x}}_\infty|), |\tilde{\mathbf{x}}_\infty|\rangle + \delta_{L,\alpha}^{-1}\langle \mathbf{1}, |\tilde{\mathbf{x}}_\infty|\rangle \\
&\quad = \langle \log(|\tilde{\mathbf{u}}_\infty - \tilde{\mathbf{v}}_\infty|), |\tilde{\mathbf{u}}_\infty - \tilde{\mathbf{v}}_\infty|\rangle + \delta_{L,\alpha}^{-1}\langle \mathbf{1}, |\tilde{\mathbf{u}}_\infty - \tilde{\mathbf{v}}_\infty|\rangle \\
&\quad \leq \langle \log(\tilde{\mathbf{u}}_\infty), \tilde{\mathbf{u}}_\infty\rangle + \delta_{L,\alpha}^{-1}\langle \mathbf{1}, \tilde{\mathbf{u}}_\infty\rangle + \langle \log(\tilde{\mathbf{v}}_\infty), \tilde{\mathbf{v}}_\infty\rangle + \delta_{L,\alpha}^{-1}\langle \mathbf{1}, \tilde{\mathbf{v}}_\infty\rangle + 2Ne^{-1-\delta_{L,\alpha}^{-1}}.
\end{aligned}$$

By the optimality of $(\tilde{\mathbf{u}}_\infty, \tilde{\mathbf{v}}_\infty)$ in Theorem 3.2 (recall the definition of $\delta_{L,\alpha}^{-1}$, for $L = 2$), we have

$$\begin{aligned}
&\langle \log(\tilde{\mathbf{u}}_\infty), \tilde{\mathbf{u}}_\infty\rangle + \delta_{L,\alpha}^{-1}\langle \mathbf{1}, \tilde{\mathbf{u}}_\infty\rangle + \langle \log(\tilde{\mathbf{v}}_\infty), \tilde{\mathbf{v}}_\infty\rangle + \delta_{L,\alpha}^{-1}\langle \mathbf{1}, \tilde{\mathbf{v}}_\infty\rangle \\
&\quad \leq \langle \log(\mathbf{z}_+), \mathbf{z}_+\rangle + \delta_{L,\alpha}^{-1}\langle \mathbf{1}, \mathbf{z}_+\rangle + \langle \log(\mathbf{z}_-), \mathbf{z}_-\rangle + \delta_{L,\alpha}^{-1}\langle \mathbf{1}, \mathbf{z}_-\rangle = \langle \log|\mathbf{z}|, |\mathbf{z}|\rangle + \delta_{L,\alpha}^{-1}\langle \mathbf{1}, |\mathbf{z}|\rangle,
\end{aligned}$$

where the last equality holds because $\mathbf{z}_+$ and $\mathbf{z}_-$ have disjoint support. Consequently,

$$\langle \log(|\tilde{\mathbf{x}}_\infty|), |\tilde{\mathbf{x}}_\infty|\rangle + \delta_{L,\alpha}^{-1}\langle \mathbf{1}, |\tilde{\mathbf{x}}_\infty|\rangle \leq \langle \log|\mathbf{z}|, |\mathbf{z}|\rangle + \delta_{L,\alpha}^{-1}\langle \mathbf{1}, |\mathbf{z}|\rangle + 2Ne^{-1-\delta_{L,\alpha}^{-1}}.$$

By following the steps in the proof of Theorem 2.1, we find that

$$\|\tilde{\mathbf{x}}_\infty\|_1 - \|\mathbf{z}\|_1 \leq \delta_{L,\alpha}(\|\mathbf{z}\|_1^2 + Ne^{-1}(1 + 2e^{-\delta_{L,\alpha}^{-1}})).$$

In the case $L \geq 3$, a similar calculation exploiting Lemma A.2 with $\rho = \delta_{L,\alpha}$ leads to

$$\langle \mathbf{1}, |\tilde{\mathbf{x}}_\infty|\rangle - \delta_{\alpha,L}\langle \mathbf{1}, |\tilde{\mathbf{x}}_\infty|^{\odot\frac{2}{L}}\rangle \leq \langle \mathbf{1}, |\mathbf{z}|\rangle - \delta_{\alpha,L}\langle \mathbf{1}, |\mathbf{z}|^{\odot\frac{2}{L}}\rangle + N(L-2)\left(\frac{2\delta_{L,\alpha}}{L}\right)^{\frac{L}{L-2}}. \tag{38}$$

Again, following the same steps as in the proof of Theorem 2.1, we find that

$$\|\tilde{\mathbf{x}}_\infty\|_1 - \|\mathbf{z}\|_1 \leq \delta_{L,\alpha}(\|\tilde{\mathbf{x}}_\infty\|_1 + N) + N(L-2)\left(\frac{2\delta_{L,\alpha}}{L}\right)^{\frac{L}{L-2}}.$$

We conclude that

$$\|\tilde{\mathbf{x}}_\infty\|_1 \leq \begin{cases} \|\mathbf{z}\|_1 + \delta_{L,\alpha}(\|\mathbf{z}\|_1^2 + Ne^{-1}(1 + 2e^{-\delta_{L,\alpha}^{-1}}) & \text{if } L = 2, \\ \frac{\|\mathbf{z}\|_1}{1-\delta_{L,\alpha}} + \frac{N}{1-\delta_{L,\alpha}}\left(\delta_{L,\alpha} + (L-2)\left(\frac{2\delta_{L,\alpha}}{L}\right)^{\frac{L}{L-2}}\right) & \text{if } L > 2. \end{cases}$$

Since the above holds for all $\mathbf{z} \in S$, we can take the minimum over $\mathbf{z}$. Setting $Q = \min_{\mathbf{z}\in S} \|\mathbf{z}\|_1$ and rearranging the terms yields

$$
\begin{aligned}
\|\tilde{\mathbf{x}}_\infty\|_1 - Q &\le \begin{cases} \delta_{L,\alpha}(c^2 + Ne^{-1}(1 + 2e^{-\delta_{L,\alpha}^{-1}})) & \text{if } L = 2, \\ \frac{1}{1-\delta_{L,\alpha}}(\delta_{L,\alpha}(Q+N) + N(L-2)\left(\frac{2\delta_{L,\alpha}}{L}\right)^{\frac{L}{L-2}}) & \text{if } L > 2, \end{cases} \\
&= \begin{cases} -\frac{1}{2\log(\alpha)+1}(c^2 + Ne^{-1}(1 + 2e^{2\log(\alpha)+1})) & \text{if } L = 2, \\ \frac{1}{1-\frac{L}{2}\alpha^{L-2}}(\frac{L}{2}\alpha^{L-2}(Q+N) + N(L-2)(\alpha^{L-2})^{\frac{L}{L-2}}) & \text{if } L > 2, \end{cases} \\
&= \begin{cases} -\frac{1}{2\log(\alpha)+1}(c^2 + Ne^{-1}(1 + 2e\alpha^2)) & \text{if } L = 2, \\ \frac{1}{2-L\alpha^{L-2}}(L\alpha^{L-2}(Q+N) + 2N(L-2)\alpha^L) & \text{if } L > 2. \end{cases}
\end{aligned}
$$

If we further assume that $\alpha \le e^{-\frac{1}{2}}$ for $L = 2$ and $\alpha \le 1$ for $L > 2$, then

$$
\|\tilde{\mathbf{x}}_\infty\|_1 - Q \le \begin{cases} -\frac{1}{2\log(\alpha)+1}(c^2 + 3Ne^{-1}) & \text{if } L = 2, \\ \frac{\alpha^{L-2}}{2-L\alpha^{L-2}}(L(Q+N) + 2N(L-2)) & \text{if } L > 2. \end{cases}
$$

Hence, in order to upper bound the right hand side by ε, it is sufficient to require

$$
\alpha \le \begin{cases} \min\left(e^{-\frac{1}{2}}, \exp\left(\frac{1}{2} - \frac{Q^2+3Ne^{-1}}{2\varepsilon}\right)\right) & \text{if } L = 2, \\ \min\left(1, \left(\frac{2\varepsilon}{L(Q+3N+\varepsilon)-4N}\right)^{\frac{1}{L-2}}\right) & \text{if } L > 2. \end{cases}
$$

The proof is completed. □

4 Numerical Experiment

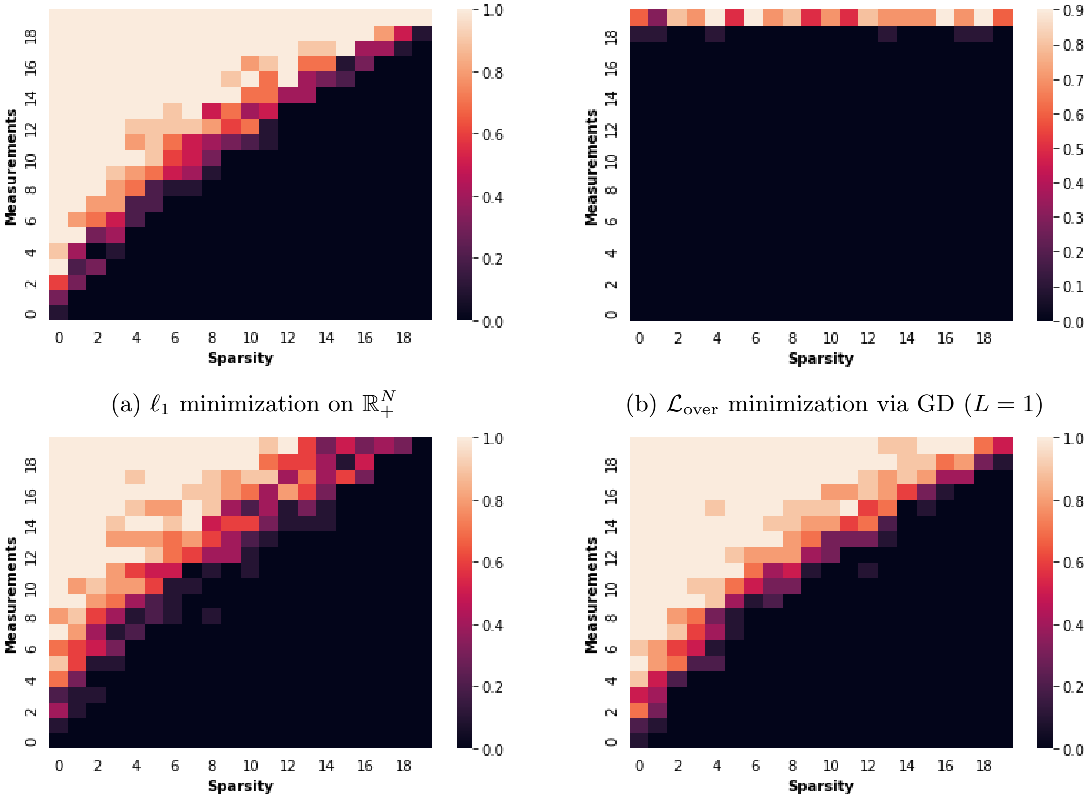
To evaluate our theoretical results, we conduct several numerical experiments. As outlined below they show recovery of sparse vectors via gradient descent (as approximation of gradient flow) on the overparameterized loss function (or equivalently on the reduced factorized loss function) from the minimal number of Gaussian random measurements as predicted by compressed sensing theory for ℓ_1-minimization, and thereby confirm experiments in in previous works [16, 25].

4.1 Positive Solution

In the first experiment, cf. Figure 2, we consider the simplified setting of Theorem 2.1, where we replace gradient flow with gradient descent in order to have an implementable method. We fix the ambient dimension $N = 20$, vary the number of measurements M and the sparsity level s of the non-negative ground truth vector $\mathbf{x}_*$ that we wish to recover. The color in the plots represents the probability of successful recovery. Hence, the boundary curves encode the empirically required number of samples in dependence of the sparsity for each algorithm. In order to be fair, we compare to ℓ_1-minimization with a positivity constraint on the solution, since we can only recover positive solutions via gradient descent in the setting of Theorem 2.1.

The measurement process $\mathbf{A} \in \mathbb{R}^{M\times N}$ and the ground truth $\mathbf{x}_*$ are given by

$$
\mathbf{A} = \frac{1}{\sqrt{M}}\mathbf{W}, \quad \mathbf{x}_* = \frac{|\mathbf{b}_S|}{\|\mathbf{b}_S\|},
$$



(a) ℓ_1 minimization on $\mathbb{R}_+^N$

(b) $\mathcal{L}_{\text{over}}$ minimization via GD ($L = 1$)

(c) $\mathcal{L}_{\text{over}}$ minimization via GD ($L = 2$)

(d) $\mathcal{L}_{\text{over}}$ minimization via GD ($L = 3$)

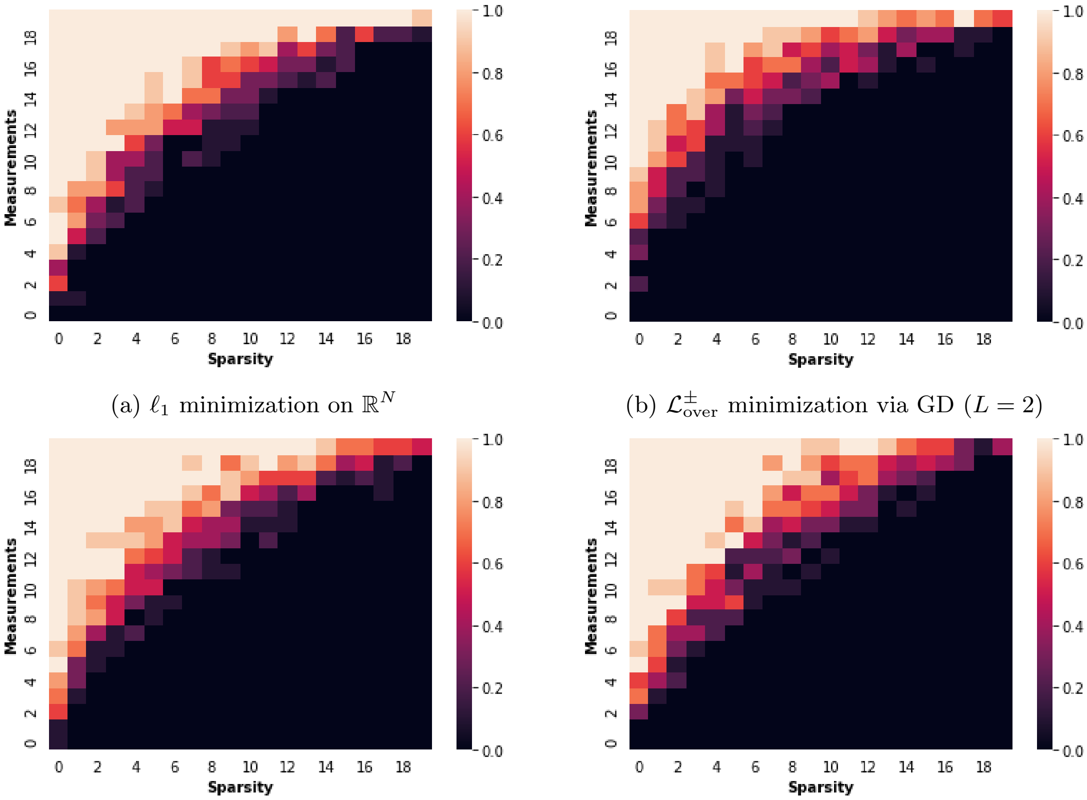
Figure 2: $\mathbf{x}_* \in \mathbb{R}_+^N$: Comparison of recovery probability between different optimization method.

where $\mathbf{W}$ is a standard Gaussian matrix, S is a random subset of $\{1, \dots, N\}$ of size s, which represents the sparsity level, and $\mathbf{b}_S$ is a standard Gaussian vector supported on S. (The index in Figure 2 and 3 starts from 0 (instead of 1) purely due to coding convention.) We take the absolute value of $\mathbf{b}$ to ensure that the assumption of Theorem 2.1 is satisfied. The parameters are set as the following: initialization parameter is set to $\alpha = 10^{-\frac{6}{L}}$, step size of gradient descent is set to $\eta = 10^{-2}$ and number of iterations equals $T = 10^7$. The recovery is considered successful if the resulting error is less than 1%, i.e., $\|\tilde{\mathbf{x}}_\infty - \mathbf{x}_*\|_2 \leq 0.01 \|\mathbf{x}_*\|_2$.

We observe that ℓ_1-minimization (with positivity constraint) and gradient descent on the reduced factorized loss (or equivalently, on the overparameterized loss when initializing identically) behave similarly, whereas gradient descent on the regular quadratic loss barely recovers until $M = N$, i.e., when the linear system is fully determined.

4.2 General Solution

Our second experiment resembles the previous one from Section 4.1, but uses gradient descent on the generalized overparameterized loss function (5) in order to reconstruct general ground truth vectors $\mathbf{x}_*$, which do not need to be non-negative, i.e. $\mathbf{x}_* = \frac{\mathbf{b}_S}{\|\mathbf{b}_S\|}$ instead of $\frac{|\mathbf{b}_S|}{\|\mathbf{b}_S\|}$. We compare with standard ℓ_1-minimization, i.e., without positivity constraints. As Figure 3 illustrates, the outcome



(a) ℓ_1 minimization on $\mathbb{R}^N$

(b) $\mathcal{L}^{\pm}_{\text{over}}$ minimization via GD ($L = 2$)

(c) $\mathcal{L}^{\pm}_{\text{over}}$ minimization via GD ($L = 3$)

(d) $\mathcal{L}^{\pm}_{\text{over}}$ minimization via GD ($L = 4$)

Figure 3: $\mathbf{x}_* \in \mathbb{R}^N$: Comparison of recovery probability between different optimization method.

is comparable to the one of the previous experiment, in the sense that GD applied on factorizations performs again similarly to ℓ_1-minimization.

4.3 Scaling of Initialization

Both the theory and our experiments show that the reconstruction accuracy improves as the initialization α decreases. However, setting $\alpha > 0$ small also requires more iterations for convergence. Thus there is a trade-off between accuracy and efficiency in the choice of α. We also observe that, in terms of accuracy, gradient descent on the reduced factorized loss improves when $L > 2$, which is consistent with the scaling of (7) described in Theorem 1.1.

In this final experiment we compare various α and see how the scaling of the initialization influences the reconstruction. The results in Figure 4 confirm our prediction in (7). The linear behavior the log-log plot suggests that the relation between error and initialization is polynomial rather than exponential, which is consistent with the scaling (7) in Theorem 1.1 for $L > 2$. On the other hand, the sufficient upper bound suggested by Theorem 1.1 seems pessimistic for $L = 2$ because empirically it also suffices to have initialization with polynomial scaling. The key difference is the slope, which empirically suggests that for $L > 2$ different choices of L behave similarly and out-perform the case of $L = 2$.

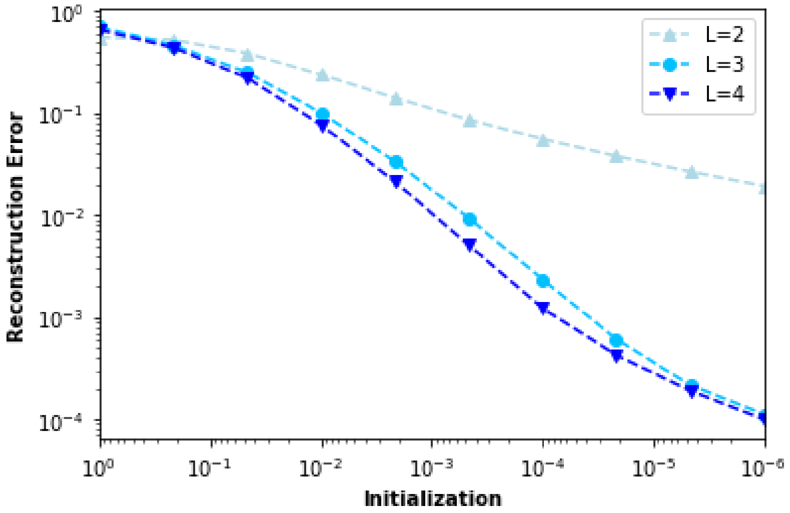


Figure 4: The reconstruction error decreases as the initialization decreases. Here we choose $N = 20$, $M = 14$, $L = 3$ and $\mathbf{x}_*$ has sparsity level $s = 3$. The initialization on the horizontal axes refers to the parameter $\rho = \alpha^L$ in the initialization $\tilde{\mathbf{x}}(0) = (\alpha\mathbf{1})^{\odot L} = \alpha^L\mathbf{1}$.

5 Summary and Future Directions

In the present work, we considered overparametrized square losses for sparse vector recovery from linear measurements. We showed that in these settings vanilla gradient flow exhibits an implicit bias towards solutions that minimize the ℓ_1-norm among all possible solutions. This lead to near-optimal sampling rates. Several intriguing research questions remain for the future.

First, our theory focuses on gradient flow for minimizing the factorized square loss. However, all conducted experiments use the corresponding gradient descent algorithm instead. A natural next step is thus to extend our results from gradient flow to gradient descent.

Second, we are also interested in low rank matrix/tensor recovery via overparameterization, which works well empirically but is not yet fully understood. Although sparsity and low-rankness are closely related, one of the significant differences between the Hadamard and the matrix product is that the later is non-commutative in general. In fact, many works [2, 7] show that low rankness can be deduced if commutativity is assumed. We believe that extending our proof concepts to the non-commutative matrix case might lead to near-optimal matrix sensing results in this more challenging setting.

Acknowledgement

This research was supported through the program ”Research in Pairs” by the Mathematisches Forschungsinstitut Oberwolfach (MFO) in 2021. We wish to thank the MFO at this point for their hospitality. HHC acknowledges funding by the DAAD through the project *Understanding stochastic gradient descent in deep learning* (project no. 57417829).

References

[1] S. Arora, N. Cohen, and E. Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, pages 244–253, 2018.

[2] S. Arora, N. Cohen, W. Hu, and Y. Luo. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems*, pages 7413–7424, 2019.

[3] L. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217, 1967.

[4] S. Brugiapaglia, S. Dirksen, H. C. Jung, and H. Rauhut. Sparse recovery in bounded riesz systems with applications to numerical methods for pdes. *Applied and Computational Harmonic Analysis*, 53:231–269, 2021.

[5] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.

[6] E. J. Candès, J. K. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 59(8):1207–1223, 2006.

[7] H. Chou, C. Gieshoff, J. Maly, and H. Rauhut. Gradient descent for deep matrix factorization: Dynamics and implicit bias towards low rank. *arXiv preprint: 2011.13772*, 2020.

[8] D. L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[9] S. Foucart and H. Rauhut. *A Mathematical Introduction to Compressive Sensing*. Birkhäuser, New York, NY, 2013.

[10] K. Geyer, A. Kyrillidis, and A. Kalev. Low-rank regularization and solution uniqueness in overparameterized matrix sensing. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, pages 930–940, 2020.

[11] G. Gidel, F. Bach, and S. Lacoste-Julien. Implicit regularization of discrete gradient dynamics in linear neural networks. In *Advances in Neural Information Processing Systems*, pages 3202–3211, 2019.

[12] D. Gissin, S. Shalev-Shwartz, and A. Daniely. The implicit bias of depth: How incremental learning drives generalization. *International Conference on Learning Representations (ICLR).*, 2020.

[13] S. Gunasekar, J. D. Lee, D. Soudry, and N. Srebro. Implicit bias of gradient descent on linear convolutional networks. In *Advances in Neural Information Processing Systems*, pages 9461–9471, 2018.

[14] S. Gunasekar, B. E. Woodworth, S. Bhojanapalli, B. Neyshabur, and N. Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, pages 6151–6159, 2017.

[15] P. D. Hoff. Lasso, fractional norm and structured sparse estimation using a Hadamard product parametrization. *Computational Statistics & Data Analysis*, 115:186–198, 2017.

[16] J. Li, T. Nguyen, C. Hegde, and K. W. Wong. Implicit sparse regularization: The impact of depth and early stopping. In *Advances in Neural Information Processing Systems*, 2021.

[17] S. Mendelson, H. Rauhut, and R. Ward. Improved bounds for sparse recovery from subsampled random convolutions. *The Annals of Applied Probability*, 28(6):3491–3527, 2018.

[18] B. Neyshabur, R. Tomioka, R. Salakhutdinov, and N. Srebro. Geometry of optimization and implicit regularization in deep learning. *arXiv preprint: 1705.03071*, 2017.

[19] B. Neyshabur, R. Tomioka, and N. Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. In *International Conference on Learning Representations*, 2015.

[20] N. Razin and N. Cohen. Implicit regularization in deep learning may not be explainable by norms. In *Advances in Neural Information Processing Systems*, pages 21174–21187, 2020.

[21] N. Razin, A. Maman, and N. Cohen. Implicit regularization in tensor factorization. *arXiv preprint: 2102.09972*, 2021.

[22] N. Razin, A. Maman, and N. Cohen. Implicit regularization in hierarchical tensor factorization and deep convolutional neural networks. *CoRR*, abs/2201.11729, 2022.

[23] D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

[24] D. Stöger and M. Soltanolkotabi. Small random initialization is akin to spectral learning: Optimization and generalization guarantees for overparameterized low-rank matrix reconstruction. *arXiv preprint: 2106.15013*, 2021.

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

[26] Y. Wang, M. Chen, T. Zhao, and M. Tao. Large learning rate tames homogeneity: Convergence and balancing effect. In *International Conference on Learning Representations*, 2022.

[27] 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 *Proceedings of Thirty Third Conference on Learning Theory*, pages 3635–3673, 2020.

[28] F. Wu and P. Rebeschini. Implicit regularization in matrix sensing via mirror descent. *Advances in Neural Information Processing Systems*, 34, 2021.

[29] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

[30] P. Zhao, Y. Yang, and Q.-C. He. Implicit regularization via hadamard product overparametrization in high-dimensional linear regression. *arXiv preprint: 1903.09367*, 2019.

A Appendix

A.1 Improved condition on initialization

In order to show the slightly relaxed conditions on α for $L \geq 3$ in Remarks 1.2 and 2.2 we will use the following lemma.

Lemma A.1. *For vectors $\mathbf{v}, \mathbf{z} \in \mathbb{R}^N$, constants $\rho, \kappa > 0$ and $0 < p < 1$ assume that*

$$\|\mathbf{v}\|_1 - \|\mathbf{z}\|_1 \leq \rho(\|\mathbf{v}\|_p^p - \|\mathbf{z}\|_p^p) + N\kappa.$$

Then

$$\|\mathbf{v}\|_1 - \|\mathbf{z}\|_1 \leq \max\left\{(2\rho)^{\frac{1}{1-p}} N^p, 2\rho N^{p-p^2}\|\mathbf{z}\|_1^p + 2N\kappa\right\}. \tag{39}$$

If $\kappa = 0$ then we have the improved estimate

$$\|\mathbf{v}\|_1 - \|\mathbf{z}\|_1 \leq \max\left\{2^{\frac{p}{1-p}} \rho^{\frac{1}{1-p}} N^p, 2^{1/p}\rho N^{p-p^2}\|\mathbf{z}\|_1^p\right\}. \tag{40}$$

Proof. Using that $\|\mathbf{w}\|_1 \leq \|\mathbf{w}\|_p \leq N^{1-p}\|\mathbf{w}\|_1$, for any vector $\mathbf{w} \in \mathbb{R}^N$, we have

$$\eta := \|\mathbf{v}\|_1 - \|\mathbf{z}\|_1 \leq \rho(N^{1-p}\|\mathbf{v}\|_1)^p + N\kappa$$

Convexity of $x \mapsto x^{1/p}$ on $[0, \infty)$ yields

$$(a+b)^{1/p} = 2^{1/p}((a+b)/2)^{1/p} \leq 2^{1/p}\frac{1}{2}(a^{1/p} + b^{1/p}) \quad \text{for all } a, b \geq 0 \tag{41}$$

so that

$$\begin{aligned} \eta^{1/p} &\leq 2^{1/p-1}\left(\rho^{1/p}N^{1-p}\|\mathbf{v}\|_1 + N^{1/p}\kappa^{1/p}\right) \\ &= 2^{1/p-1}\left(\rho^{1/p}N^{1-p}\eta + \rho^{1/p}N^{1-p}\|\mathbf{z}\|_1 + N^{1/p}\kappa^{1/p}\right). \end{aligned} \tag{42}$$

If $\rho^{1/p}N^{1-p}\eta \leq \rho^{1/p}N^{1-p}\|\mathbf{z}\|_1 + N^{1/p}\kappa^{1/p}$ this implies $\eta^{1/p} \leq 2^{1/p-1} \cdot 2 \cdot (\rho^{1/p}N^{1-p}\|\mathbf{z}\|_1 + N^{1/p}\kappa^{1/p})$, which is equivalent to

$$\eta \leq 2\left(\rho^{1/p}N^{1-p}\|\mathbf{z}\|_1 + N^{1/p}\kappa^{1/p}\right)^p \leq 2\rho N^{p-p^2}\|\mathbf{z}\|_1^p + 2N\kappa.$$

If $\rho^{1/p}N^{1-p}\eta > \rho^{1/p}N^{1-p}\|\mathbf{z}\|_1 + N^{1/p}\kappa^{1/p}$ then (42) implies $\eta^{1/p} \leq 2^{1/p}\rho^{1/p}N^{1-p}\eta$, or equivalently

$$\eta \leq (2\rho)^{\frac{1}{1-p}} N^p.$$

Combining both cases gives (39). If $\kappa = 0$ then (41) holds without the factor $2^{1/p}\frac{1}{2} = 2^{1/p-1}$. Cancelling this factor in all the subsequent steps leads to (40). □

Let us now show the sufficiency of the improved condition (17) for the positive case. The bound (30) in the proof of Theorem 2.1 can be written as

$$\|\tilde{\mathbf{x}}_\infty\|_1 - \|\mathbf{z}\|_1 \leq \delta_{L,\alpha}\left(\|\tilde{\mathbf{x}}_\infty\|_{2/L}^{2/L} - \|\mathbf{z}\|_{2/L}^{2/L}\right).$$

An application of Lemma A.1 with $\rho = \delta_{L,\alpha}$, $\kappa = 0$, $p = 2/L \in (0,1)$ and $\mathbf{z}$ being the ℓ_1-minimizer (recalling that $Q_+ = \min_{\mathbf{z}\in S_+} \|\mathbf{z}\|_1$) gives

$$\|\tilde{\mathbf{x}}_\infty\|_1 - Q_+ \leq \max\left\{2^{\frac{2}{L-2}}\delta_{L,\alpha}^{\frac{L}{L-2}}N^{\frac{2}{L}}, 2^{L/2}\delta_{L,\alpha}N^{\frac{2(L-2)}{L^2}}Q_+^{\frac{2}{L}}\right\}. \tag{43}$$

Recalling the definition $\delta_{L,\alpha} = \frac{L}{2}\alpha^{L-2}$, the first term in the maximum above is upper bounded by ε if

$$\alpha \leq (2\varepsilon)^{1/L} N^{-2/L^2} L^{-\frac{1}{L-2}}.$$

The second term in the maximum in (43) is bounded by ε if

$$\alpha \leq \left(\frac{\varepsilon}{L}\right)^{\frac{1}{L-2}} 2^{-\frac{L}{L-2}} N^{-\frac{2}{L^2}} Q^{-\frac{2}{L(L-2)}}.$$

Altogether we obtain that $\|\tilde{\mathbf{x}}_\infty\|_1 - Q \leq \varepsilon$ provided that

$$\alpha \leq N^{-\frac{2}{L^2}} L^{-\frac{1}{L-2}} \min\left\{(2\varepsilon)^{1/L}, (\varepsilon/L)^{\frac{1}{L-2}} Q^{\frac{2}{L(L-2)}}\right\}.$$

This is the claim (17) of Remark 2.2.

For the improved bound for the general (not necessarily positive) case claimed in Remark 1.2 we observe that inequality (38) in the proof of Theorem 1.1 can be stated as

$$\|\tilde{\mathbf{x}}_\infty\|_1 - \|\mathbf{z}\|_1 \leq \delta_{L,\alpha}\left(\|\tilde{\mathbf{x}}_\infty\|_{2/L}^{2/L} - \|\mathbf{z}\|_{2/L}^{2/L}\right) + N(L-2)\left(\frac{2\delta_{L,\alpha}}{L}\right)^{\frac{L}{L-2}}.$$

Lemma A.1 applied with $\rho = \delta_{L,\alpha}$, $\kappa = (L-2)\left(\frac{2\delta_{L,\alpha}}{L}\right)^{\frac{L}{L-2}}$, $p = 2/L \in (0,1)$ and $\mathbf{z}$ being the ℓ_1-minimizer yields

$$\|\tilde{\mathbf{x}}_\infty\|_1 - Q \leq \max\left\{(2\delta_{L,\alpha})^{\frac{L}{L-2}}N^{\frac{2}{L}}, 2\delta_{L,\alpha}N^{\frac{2(L-2}{L^2}}Q^{\frac{2}{L}} + 2N(L-2)\left(\frac{2\delta_{L,\alpha}}{L}\right)^{\frac{L}{L-2}}\right\}.$$

Again using that $\delta_{L,\alpha} = \frac{L}{2}\alpha^{L-2}$, the first term in the maximum is upper bounded by ε if

$$\alpha \leq N^{-\frac{4}{L^2}} L^{-\frac{1}{L-2}} \varepsilon^{1/L}$$

The second term in the maximum equals

$$P := \alpha^{L-2}\left(N^{\frac{2(L-2)}{L^2}}Q^{\frac{2}{L}}L + N(L-2)\alpha^2\right)$$

If $N(L-2)\alpha^2 \leq N^{\frac{2(L-2)}{L^2}}Q^{\frac{2}{L}}L$ then $P \leq \alpha^{L-2}\cdot 2 \cdot N^{\frac{2(L-2)}{L^2}}Q^{\frac{2}{L}}L$, so that $P \leq \varepsilon$ if

$$\alpha \leq (\varepsilon/2)^{\frac{1}{L-2}} N^{-\frac{2}{L^2}} Q^{-\frac{1}{L(L-2)}}.$$

In the case that $N(L-2)\alpha^2 \geq N^{\frac{2(L-2)}{L^2}}Q^{\frac{2}{L}}L$ we have $P \leq \alpha^L \cdot 2N(L-2)$ so that $P \leq \varepsilon$ provided that

$$\alpha \leq \left(\frac{\varepsilon}{2N(L-2)}\right)^{\frac{1}{L}}.$$

Combining all cases, we conclude that

$$\|\tilde{\mathbf{x}}_\infty\|_1 - Q \leq \varepsilon$$

if

$$\alpha \leq \min\left\{\varepsilon^{1/L}\min\{N^{-4/L^2}L^{-\frac{1}{L-2}}, (2N(L-2))^{-\frac{1}{L}}\}, (\varepsilon/2)^{\frac{1}{L-2}}N^{-\frac{2}{L^2}}Q^{-\frac{1}{L(L-2)}}\right\}.$$

This is the statement of Remark 1.2.

A.2 Auxiliary bound

The proof of Theorem 1.1 requires the following technical statement.

Lemma A.2. *For $L \geq 2$ and a constant $\rho > 0$ let $h : \mathbb{R}_+^2 \to \mathbb{R}$ be defined, for $u, v \in \mathbb{R}_+$, as*

$$h(u,v) = \begin{cases} u\log(u) + v\log(v) - |u-v|\log|u-v| + \rho(u+v-|u-v|) & \text{if } L = 2, \\ u + v - |u-v| + \rho(|u-v|^{\frac{2}{L}} - u^{\frac{2}{L}} - v^{\frac{2}{L}}) & \text{if } L \geq 3. \end{cases}$$

The function h satisfies the lower bound

$$h(u,v) \geq \begin{cases} -2e^{-1-\rho} & \text{if } L = 2, \\ -(L-2)\left(\frac{2\rho}{L}\right)^{\frac{L}{L-2}} & \text{if } L \geq 3, \end{cases}$$

for all $u, v \in \mathbb{R}_+$.

Note that it is understood that $z\log(z) = 0$ for $z = 0$ in the definition of h for $L = 2$ so that, in particular, $h(u,u) = 2u\log(u) + 2\rho u$, $u \in \mathbb{R}_+$.

Proof. By symmetry of h we may assume $u \geq v$ (the case $u \leq v$ then follows by interchanging the roles of u and v).

Let us now assume $L = 2$. Then we have that

$$\nabla h(u,v) = \begin{pmatrix} \log(u) - \log(u-v) \\ \log(v) + \log(u-v) + 2 + 2\rho \end{pmatrix}.$$

Note that the gradient of h is not vanishing in the interior of the domain $\{(u,v) : u \geq v \geq 0\}$ because $\log(u) > \log(u-v)$ for $u > v > 0$. Hence h can attain a minimum only on the boundary $\{(u,v) : u \geq 0, v = 0\} \cup \{(u,v) : u = v \geq 0\}$. A direct calculation gives $h(u,0) = 0$ and

$$h(u,u) = 2u(\log(u) + \rho) \geq 2e^{-1-\rho}(\log(e^{-1-\rho}) + \rho) = -2e^{-1-\rho},$$

where the inequality follows since the function $\xi \mapsto 2\xi(\log(\xi) + \rho$ is minimized for $\xi = e^{-1-\rho}$. Hence, $h(u,v) \geq -2e^{-1-\rho}$ for all $u, v \geq 0$.

For $L \geq 3$, the gradient of h is given by

$$\nabla h(u,v) = \begin{pmatrix} \frac{2\rho}{L}((u-v)^{-1+\frac{2}{L}} - u^{-1+\frac{2}{L}}) \\ 2 - \frac{2\rho}{L}((u-v)^{-1+\frac{2}{L}} + v^{-1+\frac{2}{L}}) \end{pmatrix}.$$

Again, ∇h does not vanish in the interior of the domain $\{(u,v) : u \geq v \geq 0\}$. Hence, h can attain a minimum only on the boundary $\{(u,v) : u \geq 0, v = 0\} \cup \{(u,v) : u = v \geq 0\}$. A direct calculation gives $h(u,0) = 0$ and

$$\begin{aligned} h(u,u) = 2(u - \rho u^{\frac{2}{L}}) &\geq 2\left(\left(\frac{L}{2\rho}\right)^{\frac{L}{2-L}} - \rho\left(\left(\frac{L}{2\rho}\right)^{\frac{L}{2-L}}\right)^{\frac{2}{L}}\right) \\ &= 2\left(\left(\frac{L}{2\rho}\right)^{\frac{L}{2-L}} - \rho\left(\frac{L}{2\rho}\right)^{\frac{2}{2-L}}\right) = 2\left(\frac{L}{2\rho}\right)^{\frac{L}{2-L}}\left(1 - \frac{L}{2}\right) = -(L-2)\left(\frac{2\rho}{L}\right)^{\frac{L}{L-2}}, \end{aligned}$$

where the inequality is due to the fact that the function $\xi \mapsto 2(\xi - \rho\xi^{\frac{2}{L}})$ attains its minimum at $\xi = (2\rho/L)^{\frac{L}{L-2}}$. Hence, h is globally bounded from below by $-(L-2)\left(\frac{2\rho}{L}\right)^{\frac{L}{L-2}}$. □

A.3 Noise robust compressed sensing

For completion we provide a proof of Theorem 1.4.

Proof of Theorem 1.4. The ℓ_2-robust null space property of order cs_* with respect to ℓ_2 with constants $\rho \in (0,1)$ and τ as in (11) together with the ℓ_1-quotient property implies the so-called simultaneous (ℓ_2, ℓ_1)-quotient property by [9, Lemma 11.15], i.e., for any $\mathbf{e} \in \mathbb{R}^M$ there exists $\mathbf{u} \in \mathbb{R}^M$ such that $\mathbf{Ae} = \mathbf{u}$ and both

$$\|\mathbf{u}\|_2 \leq d' \|\mathbf{e}\|_2 \quad \text{and} \quad \|\mathbf{u}\|_1 \leq d\sqrt{s_*}\|\mathbf{e}\|_2, \tag{44}$$

where the constant d' only depends on d, c, ρ, τ. Let $\Delta : \mathbb{R}^M \to \mathbb{R}^N$ the reconstruction map represented by gradient flow, i.e., $\Delta(\mathbf{y}')$ is the limit of gradient flow (initialized with parameter α) for the functional $\mathcal{L}^{\pm}_{\text{over}}$, the latter depending on $\mathbf{y}'$. We know from Theorem 1.1 that $\mathbf{A}(\Delta(\mathbf{y})) = \mathbf{y}$ for any $\mathbf{y}$ (the limit $\tilde{\mathbf{x}}_\infty$ is contained in S). By [9, Theorem 4.25] the ℓ_2-robust null space property of $\mathbf{A}$ implies that

$$\|\mathbf{v} - \mathbf{w}\|_2 \leq \frac{C}{\sqrt{s}} \left(\|\mathbf{v}\|_1 - \|\mathbf{w}\|_1 + 2\sigma_s(\mathbf{w})_1 \right) + D\|\mathbf{A}(\mathbf{v} - \mathbf{w})\|_2 \quad \text{for all } \mathbf{v}, \mathbf{w} \in \mathbb{R}^N. \tag{45}$$

For the noise vector $\mathbf{e}$ of the theorem, let $\mathbf{u}$ be the vector such that $\mathbf{Au} = \mathbf{e}$ and (44) holds. Note that $\mathbf{y} = \mathbf{Ax}_* + \mathbf{e} = \mathbf{A}(\mathbf{x}_+ + \mathbf{u})$. Further let, $\widehat{\mathbf{z}}$ be a minimizer of $\min_{\mathbf{z}:\mathbf{Az}=\mathbf{A}(\mathbf{x}_*+\mathbf{u})} \|\mathbf{z}\|_1$. By Theorem 1.1 we have

$$\|\mathbf{w}\|_1 - \|\widehat{\mathbf{z}}\|_1 \leq \varepsilon \tag{46}$$

provided that the initialization parameter satisfies

$$\alpha \leq h(\|\widehat{\mathbf{z}}\|_1, \varepsilon). \tag{47}$$

Applying inequality (45) for $\mathbf{v} = \mathbf{x}_* + \mathbf{u}$ and $\mathbf{w} = \Delta(\mathbf{A}(\mathbf{x}_* + \mathbf{u}))$ below (noting that $\mathbf{Av} = \mathbf{Aw}$ as argued above) yields

$$\begin{aligned}
\|\mathbf{x}_* - \Delta(\mathbf{y})\|_2 = \|\mathbf{x}_* - \Delta(\mathbf{A}(\mathbf{x}_* + \mathbf{u})\|_2 &\leq \|\mathbf{x}_* + \mathbf{u} - \Delta(\mathbf{A}(\mathbf{x}_+ + \mathbf{u}))\|_2 + \|\mathbf{u}\|_2 \\
&\leq \frac{C}{\sqrt{s}} \left(\|\mathbf{w}\|_1 - \|\mathbf{x}_* + \mathbf{u}\|_1 + 2\sigma_s(\mathbf{x}_* + \mathbf{u})_1 \right) + \|\mathbf{u}\|_2 \\
&\leq \frac{C}{\sqrt{s}} \left(\|\mathbf{w}\|_1 - \|\widehat{\mathbf{z}}\|_1 + \|\widehat{\mathbf{z}}\|_1 - \|\mathbf{x}_* + \mathbf{u}\|_1 + 2\sigma_s(\mathbf{x}_*)_1 + \|\mathbf{u}\|_1 \right) + \|\mathbf{u}\|_2.
\end{aligned}$$

Since $\widehat{\mathbf{x}}$ minimizes the ℓ_1-norm among all vectors satisfying $\mathbf{Az} = \mathbf{A}(\mathbf{x}_* + \mathbf{u})$ we have $\|\mathbf{x}_* + \mathbf{u}\|_1 \geq \|\widehat{\mathbf{z}}\|_1$. Using (44) and (46) and that $s = cs_*$ gives

$$\|\mathbf{x}_* - \Delta(\mathbf{y})\|_2 \leq \frac{C}{\sqrt{s}} \left(2\sigma_s(\mathbf{x}_*)_1 + \varepsilon \right) + \frac{Cd\sqrt{s_*}}{\sqrt{s}} \|\mathbf{e}\|_2 + d'\|\mathbf{e}\|_2) = \frac{C}{\sqrt{s}} \left(2\sigma_s(\mathbf{x}_*)_1 + \varepsilon \right) + C'\|\mathbf{e}\|_2$$

with $C' = Cd/\sqrt{c} + d'$.

Furthermore, the ℓ_2-robust null space property with constants $\rho \in (0,1)$ and $\tau > 0$ of order s implies the stable null space property of order s, see [9, Chapter 4]. Therefore, [9, Theorem 4.12] together with the ℓ_1-quotient property shows that

$$\begin{aligned}
\|\widehat{\mathbf{z}}\|_1 \leq \|\widehat{\mathbf{z}} - (\mathbf{x}_* + \mathbf{u})\|_1 + \|\mathbf{x}_* + \mathbf{u}\|_1 &\leq \frac{2(1+\rho)}{1-\rho} \sigma_s(\mathbf{x}_* + \mathbf{u})_1 + \|\mathbf{x}_*\|_1 + \|\mathbf{u}\|_1 \\
\leq \frac{2(1+\rho)}{1-\rho} \sigma_s(\mathbf{x}_*)_1 + \|\mathbf{x}_*\|_1 + 2\|\mathbf{u}\|_1 &\leq \frac{2(1+\rho)}{1-\rho} \sigma_s(\mathbf{x}_*)_1 + \|\mathbf{x}_*\|_1 + 2d\sqrt{s_*}\|\mathbf{e}\|_2.
\end{aligned}$$

Since $Q \mapsto h(Q, \varepsilon)$ is monotonically decreasing, it is sufficient for (47) to require

$$\alpha \leq h\left(\|\mathbf{x}_*\|_1 + \frac{2(1+\rho)}{1-\rho}\sigma_s(\mathbf{x}_*)_1 + 2d\sqrt{s_*}\|\mathbf{e}\|_2, \varepsilon\right).$$

This completes the proof of Theorem 1.4. □