*Pure and Applied Functional Analysis*
Volume 8, Number 2, 2023, 661–691

Yokohama Publishers
ISSN 2189-3764 ONLINE JOURNAL

# A RELAXATION ARGUMENT FOR OPTIMIZATION IN NEURAL NETWORKS AND NON-CONVEX COMPRESSED SENSING

GERRIT WELPER

Abstract. We provide a heuristic argument that views layers in neural networks as a relaxation strategy to aid their optimization. This argument is abstracted to generic non-convex optimization problems and applied to non-convex compressed sensing, where we seek the sparsest solution of a linear system, with comparatively weak assumptions on the measurement matrix for which the usual convex $\ell_1$-minimization is not applicable. We show that the new method allows an exponentially $r^\theta$ increased chance to find global optimizers for $r$ initial guesses of solution sub-blocks of size $\theta$ by solving a convex optimization problem of the non-exponential size $r\theta$.

## 1. Introduction

Neural networks are trained with gradient descent or related methods starting from random initial values. Since the loss function is non-convex, this can result in bad local minima. Indeed, in the worst case, the problem of neural network training is $NP$-hard [6]. Nonetheless, neural networks are successfully trained in a large number of practical applications [26, 28]. Contrary to arbitrary networks in worst case scenarios, practical networks are usually over-parametrized, which has been studied experimentally in e.g. [27, 54]. On the theoretical side, in severely over-parametrized regimes, neural networks can be approximated by the linear neural tangent kernel, which allows a convergence analysis of gradient descent methods [1–4, 14, 20, 21, 29, 34, 40]. Alternative arguments can be found in landscape analysis [19, 25, 39, 47, 50] and Wasserstein gradient flow [13, 36, 42, 44, 46, 48].

In this article, we study the benefits of adding extra weights to neural networks and other non-convex optimization problems like $\ell_p$-minimization with $p < 1$ in compressed sensing, without necessarily reaching the severely over-parametrized regime for which neural network training is understood. To this end, we start with a non-convex reference problem and enlarge it by a relaxation argument that is motivated by the process of widening and deepening a neural network. The enlarged network has sufficient capacity to run several (say $r$), instances of the reference problem in parallel together with an added layer that servers as a *selector variables.* By combining the parallel pieces via a proper choice of the selector variables, the

2020 *Mathematics Subject Classification.* 90C26, 94A12, 68T05.

*Key words and phrases.* Compressed sensing, deep neural networks, relaxation, non-convex, global minima.

enlarged network can be reduced to the reference network in at least $r^\theta$ different ways, where $\theta$ is the width of original network.

Therefore, if we can compute (globally) optimal selector variables, at an $r$-fold increased cost of the larger relaxed optimization problem, we may expect a minimizer comparable to the best of $r^\theta$ numerical optima from random initializations of the reference problem. This would allow us to explore an exponential number of initializations for a non-convex optimization problem in linear time. However due to this "boosting" it is not at all clear if we can indeed find sufficiently good optimizers of the selector. Although this paper does not provide an answer for neural network training, we show that the argument does succeed for other severely non-convex problems from compressed sensing, where we currently have a richer theoretical background.

In compressed sensing, we search for the sparsest solution of an under-determined linear system, i.e. for a measurement matrix $A \in \mathbb{R}^{m\times N}$ and measurements $y \in \mathbb{R}^m$, we are interested in the solution of the optimization problem

$$\min_{x\in\mathbb{R}^N} \|x\|_0, \quad \text{s.t.} \quad Ax = y, \tag{1.1}$$

where the $\ell_0$-norm measures the number of non-zero entries. Since this problem is computationally difficult, it is typically replaced by a $\ell_p$-minimization

$$\min_{x\in\mathbb{R}^N} \|x\|_p^p, \quad \text{s.t.} \quad Ax = y, \tag{1.2}$$

with $0 < p \leqslant 1$. The most common choice is $p = 1$, for which the optimization problem is convex and the restricted isometry property (RIP) or similar conditions on the matrix $A$ guarantee that the solutions of the problems (1.1) and (1.2) coincide, see e.g. [8, 9, 18, 23, 32]. Nonetheless, finding sparse solutions is also of interest in many applications where the RIP is not available. For $p < 1$, the $\ell_p$ norm resembles the $\ell_0$ norm more closely and one may expect better sparse recovery results with less assumptions on the matrix $A$. Such results have been reported by several authors [10, 11, 22, 31, 45, 49].

For $p$ strictly smaller than one, the optimization problem (1.2) is no longer convex making its optimization considerably more difficult. In fact, in the worst case the problem is NP-hard [24, 37]. Nonetheless, there are several iterative algorithms [10, 12, 16, 22, 33, 53], typically variations of reweighted least squares methods, that show promising performance on these problems. Due to the non-convex nature of the problem, the corresponding analysis requires additional assumptions that are hard to validate practically to provide convergence guarantees.

We analyze the "boosting" strategy for a class of simplified discrete non-convex compressed sensing problems, which are still $NP$-hard in general. The main result demonstrates that we can achieves the $r^\theta$ fold increased chance to find global optima by solving a convex optimization problem of size $r\theta$, as described in the motivation above. Contrary to convex compressed sensing, we only require relatively mild assumptions on the measurement matrix $A$, which are considerably weaker than the usual $RIP$.

The presented method in itself is not immediately practical, but merely seeks some insight into highly non-convex minimization problems, for which provably tractable optimization methods are still elusive. Nonetheless, the described relaxation can be used in combination with reweighted least squares methods and other optimizers for non-convex compressed sensing. Alternatively, the follow up paper [52], uses the relaxation argument in combination with a transfer learning argument to obtain some provable efficient algorithms for non-convex compressed sensing.

The paper is organized as follows. In Section 2, we state the generic relaxation method, its application to both compressed sensing and neural network training and provide some estimates of potential success probabilities. In Section 3, we consider the relaxation method for compressed sensing more carefully and prove the main results of the paper.

## 2. A Relaxation Method

In Section 2.1 and 2.2, we describe a simple relaxation method and in Section 2.3 a variant with added structure. A discussion of the optimization problems and success probabilities is given in Sections 2.4 and 2.5, respectively.

2.1. **Simple Relaxation.** We consider the optimization problem

$$\min_{x \in \mathbb{R}^N} g(x), \quad \text{s.t.} \quad h(x) = 0, \tag{2.1}$$

with objective $g$ and constraint $h$ and solve it with a local search method, e.g. gradient descent or variants thereof for neural networks or reweighted least squares for compressed sensing. Since we are particularly interested in non-convex problems, depending on the initial value, this may or may not result in a satisfactory minimizer. Probably the simplest idea to enhance our chance of success is to repeat this optimization for multiple initial values, say $x_k$, $k = 1, \dots, r$ resulting in local (numerical) optima $\bar{x}_k$ from which we select the best one

$$x = \operatorname*{argmin}_{k=1,\dots,r} g(\bar{x}_k). \tag{2.2}$$

For simplicity, we drop the equality constraint during this motivation, but all arguments work with it unchanged. In order to relax this problem to a continuous one, note that with standard unit basis vectors $e_k \in \mathbb{R}^r$ and any splitting $g(x) = \ell(f(x))$, we can equivalently minimize

$$\min_{z \in e_1, \dots, e_r} \ell \left[ \sum_{k=1}^{r} z_k f(\bar{x}_k) \right].$$

The vector $z$ serves as a "*selector*" and picks one guess $f(\bar{x}_k)$ and the split of the objective $g$ into the two components $\ell$ and $f$ allows some flexibility in the placement of the selector. In hope to simplify the problem, we remove the discrete constraint $z \in e_1, \dots, e_r$ in favor of a continuous $z \in \mathbb{R}^r$ and obtain the *relaxed* problem

$$\min_{z \in \mathbb{R}^r} \ell \left[ \sum_{k=1}^{r} z_k f(\bar{x}_k) \right]. \tag{2.3}$$

Similar relaxation strategies are common for many optimization problems, see e.g. [7, 38] in general, [15] for integer programming or [51] for optimal transport. Since the relaxed problem allows a larger choice of selectors $z$, its minimum is at least as small as the un-relaxed one

$$\min_{z\in\mathbb{R}^r} \ell\left[\sum_{k=1}^r z_k f(\bar{x}_k)\right] \leqslant \min_{k=1,\dots,r} \ell[f(\bar{x}_k)].$$

As a last step, we reintroduce the optimization of the $x$ variable and obtain

$$\min_{\substack{z\in\mathbb{R}^r\\ x_1,\dots,x_r\in\mathbb{R}^N}} \ell\left[\sum_{k=1}^r z_k f(x_k)\right]. \tag{2.4}$$

Note that depending on the application, we may be interested in two different quantities: The first one is the smallest possible value of the objective, e.g. in the neural network Example 2.1. Alternatively, for some applications one can show that the optimal selectors $z$ are of the original form $z \in e_1, \dots, e_r$, although optimized over all of $\mathbb{R}^r$. That allows us to recover the optimal $x$, as e.g. in the main results of this paper in Section 3. A more through discussion of the relaxed optimization problem is given in Section 2.4, but first we consider a variant with some additional structure.

2.2. **Interplay of Two Layers.** We apply the relaxation strategy to the optimization of weights $X \in \mathbb{R}^{\theta\times n}$ in one layer of a neural network

$$\min_X \operatorname{loss}[y, g\circ \operatorname{ReLU}[Xh]], \tag{2.5}$$

with hidden layer $h \in \mathbb{R}^n$, downstream network $g : \mathbb{R}^\theta \to \mathbb{R}^d$ and data $y \in \mathbb{R}^d$. Note that $X$ are weights, not the network input to be compliant with the compressed sensing notation later. For simplicity, we ignore the optimization of all other layers, but it can easily be added to the final problem at the end of this section. Typically $g$ is a highly non-convex function, which can render this optimization problem difficult. In this case, we may use the relaxation strategy from the last section by picking the best out of multiple initial guesses $X_k \in \mathbb{R}^{\theta\times n}$ to arrive at

$$\min_{z\in e_1,\dots,e_r} \operatorname{loss}\left[y, g\circ\operatorname{ReLU}\left[\sum_{k=1}^r z_k \otimes I \operatorname{ReLU}(X_k h)\right]\right],$$

where we have used that $\operatorname{ReLU}\circ\operatorname{ReLU} = \operatorname{ReLU}$ and ignored the optimization of $X_k$ for the time being. However, the structure of the problem allows us to explore much bigger search spaces. For example, for each component $i$, or neuron $i$, that is fed into $g$, we can choose a different combination of the initial components $\operatorname{ReLU}(X_k h)_i$, $k = 1, \dots, r$ by

$$\min_{\substack{[z_1,\dots z_r]\in\{0,1\}^{\theta\times r}\\ \|[z_1,\dots z_r]\|_\infty=1}} \operatorname{loss}\left[y, g\circ\operatorname{ReLU}\left[\sum_{k=1}^r \operatorname{diag}(z_k)\operatorname{ReLU}(X_k h)\right]\right],$$

where we optimize over all matrices $[z_1, \dots, z_r]$ whose rows have exactly one non-zero entry with value 1. The discrete search space has exponential $r^\theta$ possible combinations and is therefore untractable. We address the problem by relaxing $[z_1 \otimes I, \dots, z_r \otimes I]$ for the first problem or $[\operatorname{diag}(z_1), \dots, \operatorname{diag}(z_k)] \in \{0,1\}^{\theta\times r\theta}$ for the second to an arbitrary matrix $Z = [Z_1, \dots, Z_r] \in \mathbb{R}^{\theta \times r\theta}$. For both of the above discrete optimization problems, this yields

$$\min_{Z,\mathbb{X}} \operatorname{loss}[y, g \circ \operatorname{ReLU}[Z \operatorname{ReLU}(\mathbb{X}h)], \tag{2.6}$$

where we have reintroduced the optimization of $\mathbb{X} = [X_1, \dots, X_r]^T$ and now have a continuous search space of size $r\theta^2 + r\theta n$, which can easily be handled by gradient descent.

Note in particular, that the latter optimization problem is a standard neural network, now with a wider layer for $\mathbb{X}$ and one additional layer for $Z$. This raises the following question: If the relaxed problem indeed finds a better optimum than the two discrete problems, is should also be better than optimizing the original network (2.5) with an exponential number of initial guesses. This would provide some additional support to the empirical observation that deeper and wider networks often train better.

2.3. **Block Relaxation.** The relaxation strategy from the last section is not confined to neural networks and can easily be applied to other problems as well. To this end, we assume that the optimization problem $\min_x g(x)$ from (2.1) can be split into the block structure

$$\min_{x^1,\dots,x^\theta} \ell\left[f^1(x^1), \dots, f^\theta(x^\theta)\right], \tag{2.7}$$

with $x = [x^1, \dots, x^\theta]$ and each block $f^l(x^l)$ only depending on $x^l$ and not any other $x^j$ with $j \neq l$. The following two examples describe an application to neural networks and compressed sensing.

**Example 2.1.** Denoting the rows of $X$ as $X_{l,\cdot}$, the neural network optimization (2.5) is a special case with

$$\begin{aligned} x^l &= X_{l,\cdot} \\ f^l(x) &= \operatorname{ReLU}(x \cdot h) \\ \ell(\cdot) &= \operatorname{loss}[y, g(\cdot)]. \end{aligned}$$

**Example 2.2.** For $0 < p \leqslant 1$, matrix $A \in \mathbb{R}^{m \times N}$ and vector $y \in \mathbb{R}^m$ consider the $\ell_p$-minimization

$$\min_{x \in \mathbb{R}^N} \|x\|_p^p, \quad \text{s.t.} \quad Ax = y.$$

If we split $x$ into blocks $x = [x^1, \dots x^\theta]$ with $x^l \in \mathbb{R}^n$ and $n\theta = N$ and define

$$f^l(x^l) = \|x^l\|_p^p, \qquad \ell(f^1, \dots, f^\theta) = \sum_{l=1}^{\theta} f^l$$

this problem fits into the general structure (2.7). The additional constraint $h(x) = Ax - y = 0$, can easily be included in the relaxation argument and will be considered more carefully in Section 3.1 below.

Back to the general problem (2.7), in order to exploit the extra block structure, we use the same relaxation argument as before. We optimize $r$ times, typically with different initial values $[x_k^1, \dots, x_k^\theta]$ to find corresponding local minima $[\bar{x}_k^1, \dots, \bar{x}_k^\theta]$ from which we select the best one

$$\min_k \ell \left[ f^1(\bar{x}_k^1), \dots, f^\theta(\bar{x}_k^\theta) \right].$$

However, as in the neural network motivation, we can also explore a much bigger search space

$$\min_{k^1, \dots k^\theta} \ell \left[ f^1(x_{k^1}^1), \dots, f^\theta(x_{k^\theta}^\theta) \right], \tag{2.8}$$

which allows us to make a different selection $k^l$ for every block $f^l(x^l)$ and therefore has a much higher chance to include a good choice. Note that we have not included the result of the optimization $x_k^l \to \bar{x}_k^l$ because for every fixed $k$, it couples the blocks, so that $\bar{x}_k^l$ depends on all other $\bar{x}_k^j$ with $j \neq l$. Instead, we skip this optimization for now, but reintroduce it after relaxation.

The discrete optimization (2.8) has $r^\theta$ possible combinations and is therefore very costly, which we address by a relaxation argument. We first rewrite the selection as a linear combination

$$\min_{\substack{z^l \in \{e_1, \dots, e_r\} \\ l = 1, \dots, \theta}} \ell \left[ \sum_{k=1}^r z_k^1 f^1(x_k^1), \dots, \sum_{k=1}^r z_k^\theta f^\theta(x_k^\theta) \right], \tag{2.9}$$

and then relax it to continuous selectors $z^l$

$$\min_{z^l \in \mathbb{R}^r, \, l = 1, \dots, \theta} \ell \left[ \sum_{k=1}^r z_k^1 f^1(x_k^1), \dots, \sum_{k=1}^r z_k^\theta f^\theta(x_k^\theta) \right]. \tag{2.10}$$

Unlike the block-wise selection (2.8) of guesses, the relaxed variant (2.10) also allows us to replace the optimization of the initial guesses $x_k^l \to \bar{x}_k^l$ with an optimization of all blocks $x_k^l$ in the relaxed optimization

$$\min_{\substack{z^l \in \mathbb{R}^r, \, l = 1, \dots, \theta \\ x_k^l, \, k = 1, \dots, r, \, l = 1, \dots, \theta}} \ell \left[ \sum_{k=1}^r z_k^1 f^1(x_k^1), \dots, \sum_{k=1}^r z_k^\theta f^\theta(x_k^\theta) \right]. \tag{2.11}$$

We may also consider other variables in the optimization such as weights from neural network layers that have been neglected in Example 2.1.

**Remark 2.3.** Applied to the neural network Example 2.1, the relaxed problem is

$$\min_{\substack{Z = [\operatorname{diag}(z_1), \dots, \operatorname{diag}(z_r)] \\ z_1, \dots z_r \in \mathbb{R}^\theta}} \operatorname{loss}[y, g \circ \operatorname{ReLU}[Z \operatorname{ReLU}(\mathbb{X} h)],$$

where the layer $Z$ has some extra block diagonal structure. In (2.6) this has been further relaxed to all full matrices $[Z_1, \dots, Z_r] \in \mathbb{R}^{\theta \times r\theta}$.

2.4. **Notes on the optimization problems.** Both, the simple relaxation from (2.4) or the block relaxation from (2.11) can be written in the form

$$\min_{x_1, \dots, x_r, z} G(x_1, \dots, x_r, z)$$

with different choices of $G$, dimensions of $z$ and copies $x_j$ of the variable $x$ we want to optimize. First note that we can choose special values $z_j$ of the selector $z$ so that $G(x_1, \dots, x_r, z_l) = g(x_l)$. Therefore we directly have

$$\min_{x_1,\dots,x_r,z} G(x_1, \dots, x_r, z) \leqslant g(\bar{x}_j), \qquad j = 1, \dots, r,$$

where as before $\bar{x}_j$ are numerical local optimizers of $g$ with initial values $x_j$. Of course to obtain a fair comparison, we also need a numerical solution of the left hand side. Let $\tilde{x}_j$ and $\tilde{z}$ be such numerical optimizers for initial values $x_j$, which may now be coupled among the indices $j$. What can be said about

$$G(\tilde{x}_1, \dots, \tilde{x}_r, \tilde{z}) \overset{?}{\lesseqqgtr} g(\bar{x}_j), \qquad j = 1, \dots, r? \tag{2.12}$$

The left hand side can be strictly smaller, equal or even larger if we do not find the global minima. Let us first make some simple observations:

(1) In general $\tilde{x}_j \neq \bar{x}_i$ for all $i, j$, or in other words we do not necessarily recover the optimia $\bar{x}_i$ of $g$. The implications depend on the problem at hand. E.g. for neural networks the relaxed problem is a larger neural network so that we can expect a smaller training error, which is desirable as long as this improvement is also reflected in the generalization error.

(2) The relaxed problem computes all $f(x_j)$, $j = 1, \dots, r$ in parallel, but never computes the full outputs $\ell(f(x_j))$ for all $j$. Instead it computes $\ell(\cdot)$ of one mixture of the available $f(x_j)$. Therefore, depending on $\ell$, it is not clear to what extend a gradient descent method can steer the selector variable $z$ to a good choice or balance of the available $f(x_j)$.

(3) Expanding on the last observation, the block relaxation (2.11) never computes $f([x^1_{k_1}, \cdots, x^\theta_{k_\theta}]) = [f^1(x^1_{k_1}), \dots, f^\theta(x^\theta_{k_\theta})]$ for all $r^\theta$ combinations $k_1, \dots, k_\theta$. This is essential for the runtime since $r^\theta$ quickly becomes prohibitively large, but it raises further questions if gradient descent or similar methods can find the right combination or a good balance.

In summary, the relaxation can significantly reduce the optimization time by avoiding to test an exponential number $r^\theta$ of combinations, but we have to answer the question when it can possibly succeed in finding superior optima.

Some hope comes from our original motivation from deep learning, where it has been observed that larger networks often perform better than smaller ones, see e.g. [27, 54]. Also several analytical results [1, 20, 34, 44, 48] show that overparametrization helps neural network training. These papers usually work in a regime where the networks are rich enough to achieve zero training error, and can be accurately approximated by the linear neural tangent kernel. Neither the zero training error nor the linearization are necessarily the case for the relaxation method described above, so that the observations of this paper are targeted to regimes that are much less well understood. However the idea is related: We increase the number of network weights and layers in the hope to enable the optimization algorithms to find better minima.

Although neural networks provide the original motivation for the relaxation idea, we analyze these methods more rigorously for compressed sensing. This area provides non-convex optimization problems as well, but the theoretical background is much better understood. Contrary to (2.12), we consider the simplified problem

$$G(x_1, \dots, x_r, \tilde{z}) \overset{?}{\lesseqgtr} g(x_j), \qquad j = 1, \dots, r, \tag{2.13}$$

where we only optimize the selector $z$. The second and third observation after (2.12) still apply. In particular, this optimization problem never evaluates all possible $r^\theta$ combinations of $f([x^1_{k_1}, \dots, x^\theta_{k_\theta}])$ but instead is a convex problem in the $r\theta$ dimensional variable $z$. Nonetheless, in Section 3 we show that the relaxed problem can find optimal combinations. One may try to incorporate an $x_j$ optimization as well by a perturbation argument, but this is left for future research.

2.5. **Comparison of Probabilities.** In this section, we compare the probabilities to find global optimizers either with $r$ random initial values in (2.2) or with the full block relaxed optimization problem (2.11). The purpose of this discussion is to better understand the prospects of the latter method and therefore, we only consider some informal estimates in a highly idealized scenario. We consider a more rigorous analysis for the compressed sensing in Section 3 below, but for other areas, such as neural networks, it remains unknown to what extend the given estimates are legitimate.

For $r$ random initial values in (2.2) some natural assumptions are

(1) There is an "attractor" $A$ of the global minimum, meaning that for each initial value $x \in A$ our optimization method of choice (e.g. gradient descent) converges to the global optimum $\min_x \ell[f(x)]$.
(2) Each initial guess $x_k$ is sampled from i.i.d random variables $X_k$.

For the block relaxed optimization problem (2.11) we assume:

(1) There are sets $B^l$, $l = 1, \dots, \theta$ such that for each initial choice $x^l \in B^l$ and every initial selectors $z^l_k$, $\ell = 1, \dots, \theta$, $k = 1, \dots, r$, the optimization method of choice (e.g. gradient descent) applied to the block relaxed problem (2.11) converges to the global optimum $\min_x \ell[f(x)]$ with probability $p_{select}$.
(2) Each initial guess $x^l_k$, $\ell = 1, \dots, \theta$, $k = 1, \dots, r$ is sampled from i.i.d random variables $X^l_k$.

The first assumption is quite severe and entails that for any initial selectors $z^l_k$ the optimizer can find an optimal selection of the blocks $x^1_{k^1}, \dots, x^\theta_{k^\theta}$ among all possible combinations. This will be analyzed carefully in the compressed sensing example in Section 3. For neural networks the assumption is unrealistic because the relaxed network likely has a smaller global minimum than the un-relaxed one. Without changing the arguments below, one can alternatively assume that the optimization of the block relaxed problem converges to a minimum that is smaller than $\min_x \ell[f(x)]$. In order to account for the fact that we may not find an optimal balance of the pieces $f^l(x^l_k)$, $k = 1, \dots, r$, we added the extra probability $p_{select}$ to do so successfully.

In the following, we use the abbreviations

$$p := P(X_1 \in A), \qquad p_l := P(X^l_1 \in B^l).$$

Since all guesses $X_k$ are i.i.d., for the optimization of $r$ repeated trials the probability of success is

$$\begin{aligned} P(\text{success } r \text{ trials}) &= P(\exists k \in \{1,\dots,r\} : X_k \in A) \\ &= 1 - P(\forall k \in \{1,\dots,r\} : X_k \notin A) \\ &= 1 - \prod_{k=1}^{r} P(X_k \notin A) \\ &= 1 - P(X_1 \notin A)^r \\ &= 1 - (1-p)^r. \end{aligned} \tag{2.14}$$

With the events $SELECT$ that the block relaxation (2.11) finds the global optimum and $INITIAL := \forall l \in \{1,\dots,\theta\} : \exists k \in \{1,\dots,r\} : X_k^l \in B^l$ of guessing good initial values, the probability that the block-relaxed optimization (2.11) is successful is

$$\begin{aligned} P(\text{success block relaxation}) &= P(SELECT \cap INITIAL) \\ &= P(SELECT|INITIAL)P(INITIAL). \end{aligned}$$

The first probability of the right hand side is $p_{select}$. With the independence of all blocks $l$, the second can be calculated analogously to (2.14), which yields

$$P(INITIAL) = \prod_{l=1}^{\theta} P(\exists k \in \{1,\dots,r\} : X_k^l \in B^l) = \prod_{l=1}^{\theta} 1 - (1-p_l)^r.$$

and thus

$$P(\text{success block relaxation}) = p_{select} \prod_{l=1}^{\theta} 1 - (1-p_l)^r.$$

For easier comparison, let us approximate the success probabilities by some simpler statements. By a first order Taylor expansion for small $q$ we have $1 - q \approx e^{-q}$ and $1 - e^{-qr} \approx qr$ and thus

$$1 - (1-q)^r \approx 1 - (e^{-q})^r = 1 - e^{-qr} \approx qr. \tag{2.15}$$

Applied to the success probabilities and assuming that $p_l$ is independent of $l$, we obtain

$$\begin{aligned} P(\text{success } r \text{ trials}) &\approx pr \\ P(\text{success block relaxation}) &\approx p_{select}(p_l r)^{\theta}. \end{aligned}$$

For the sake of comparing the two methods, we assume that $p \approx p_l^{\theta}$, which can be justified e.g. if $A \approx B^1 \times \cdots \times B^{\theta}$. Then we have

$$\begin{aligned} P(\text{success } r \text{ trials}) &\approx p_l^{\theta} r \\ P(\text{success block relaxation}) &\approx p_l^{\theta}(p_{select} r^{\theta}). \end{aligned} \tag{2.16}$$

In conclusion, for $r$ repeated trials we may achieve an $r$ fold increased chance of success and using $r$ block relaxations, which amounts to the same number of total guessed variables, we can hope for a improvement by a factor of $p_{select} r^{\theta}$. For $p_{select}$ close to one and large $\theta$ this success probability can be significantly larger. However, for the latter result we made quite significant assumptions, which we will only discuss for compressed sensing. In other cases it remains open how much of this potential improvement is realistic.

The above Taylor approximation is a rather crude argument, but in some limiting scenarios the approximations become exact. In order to define the limits properly, first note that the quantities $p$, $r$ and $p_l$ typically depend on some problem parameters such as the dimension of $x$. We denote this parameter by $\gamma$, so that $p = p(\gamma)$ and $r = r(\gamma)$ and $p_l = p_l(\gamma)$.

We now assume that the success probabilities $p_l$ (or $p$) go to zero faster than the number of guesses $r$ goes to infinity, i.e.

$$\lim_{\gamma\to\infty} p_l = 0, \qquad \lim_{\gamma\to\infty} r = \infty, \qquad \lim_{\gamma\to\infty} p_l r = 0.$$

Then, by Lemma A.2 (with $q = 1/r$) in the Appendix A.1, the Taylor approximation (2.15) used in the derivation of the success probabilities (2.16) is accurate in the limit

$$\lim_{\gamma\to\infty} \frac{1 - [1 - p_l]^r}{p_l r} = 1$$

and likewise for $p_l$ replaced by $p$.

## 3. Application to Compressed Sensing

In this section, we consider the block relaxation (2.11) applied to the compressed sensing problem of Example 2.2 in some more detail. In Section 3.1 we describe the method and the main result of this paper. Since the result is quite technical, Section 3.2 provides some more concrete scenarios and connections to the success probabilities in Section 2.5. Finally, Section 3.3 contains the proof of the main result.

### 3.1. **Model Problem.**

3.1.1. *Problem Setup and Relaxation.* For $0 < p \leqslant 1$, measurement matrix $A \in \mathbb{R}^{m\times N}$ and vector $y \in \mathbb{R}^m$ consider the $\ell_p$-minimization

$$\min_{x\in\mathbb{R}^N} \|x\|_p^p, \quad \text{s.t.} \quad Ax = y. \tag{3.1}$$

Upon possibly rescaling the right hand side $y$, we assume without loss of generality that $x \in [-1, 1]^N$, which will simplify our analysis below. In the absence of null-space or restricted isometry properies, we cannot safely replace the non-convex $\ell_p$ by a convex $\ell_1$-minimization. We therefore aim to increase our chances to find a global optimizer by the block relaxation strategy from Section 2.3. To this end, we split $x$ (and for later convenience also the measurement matrix $A$) into $\theta$ blocks

$$\begin{aligned} x &= \begin{pmatrix} x^1 & \cdots & x^\theta \end{pmatrix} \in \mathbb{R}^{n\theta} \\ A &= \begin{pmatrix} A^1 & \cdots & A^\theta \end{pmatrix} \in \mathbb{R}^{m\times n\theta}. \end{aligned} \tag{3.2}$$

and make $r$ initial guesses $x_k^l \in \mathbb{R}^n$, $k = 1, \dots, r$ for each block $l$. In order to find the best possible of $r^\theta$ combination of these initial blocks, we first write the compressed sensing problem in block form

$$\min_{x\in\mathbb{R}^N} \sum_{l=1}^{\theta} \|x^l\|_p^p, \quad \text{s.t.} \quad \sum_{l=1}^{\theta} A^l x^l = y, \tag{3.3}$$

introduce selector variables $z^l \in \{e_1, \dots, e_r\} \in \mathbb{R}^r$ for each block $l$ and optimize the blocks $x^l = \sum_{k=1}^r x_k^l z_k^l$ with respect to the selector

$$\min_{\substack{z^l\in\{e_1,\dots e_r\},\\ l=1,\dots,\theta}} \sum_{l=1}^{\theta} \left( \sum_{k=1}^{r} \|x_k^l\|_p^p |z_k^l| \right) \quad \text{s.t.} \quad \sum_{l=1}^{\theta} A^l \left( \sum_{k=1}^{r} x_k^l z_k^l \right) = y, \tag{3.4}$$

where we have used that for the given standard basis vectors $z^l$ we have $\|x^l\|_p^p = \sum_{k=1}^r \|x_k^l\|_p^p |z_k^l|$. This latter identity differers slightly from Example 2.2 and will yield a convex relaxed optimization problem. Block decompositions similar to (3.3) have been analyzed in [17], with the different purpose to find fast parallel algorithms for convex objective functions.

In order to obtain a more compact notation, let $X^l \in \mathbb{R}^{n\times r}$ be the matrices with columns $x_k^l$, $k = 1, \dots, r$. In addition, we replace the tensor $z \in \mathbb{R}^r \otimes \mathbb{R}^\theta$ with a corresponding block vector in $\mathbb{R}^{n\theta}$ and obtain the block matrix and vector

$$X = \begin{pmatrix} X^1 & & \\ & \ddots & \\ & & X^\theta \end{pmatrix} \in \mathbb{R}^{n\theta\times r\theta}$$

$$z = \begin{pmatrix} z^1 & \cdots & z^\theta \end{pmatrix} \in \mathbb{R}^{r\theta}.$$

Then, relaxing (3.4) to any $z^l \in \mathbb{R}^r$, and setting $R := r\theta$, we obtain

$$\min_{z\in\mathbb{R}^R} \sum_{j=1}^{N} \sum_{k=1}^{R} |X_{jk}|^p |z_k| \quad \text{s.t.} \quad AXz = y. \tag{3.5}$$

We have modified the relaxation argument form Example 2.2 so that the $z$ optimization is convex and we can be sure to find a global minimizer. However it is not immediately clear if this minimization can indeed find the $\ell_p$-minimal combination of the initial blocks $x_k^l$ that satisfy the linearity constraint. This poses a sparse recovery question for the selector and is discussed below.

For simplicity, we ignore the possibility that we can simultaneously optimize the initial guesses $x_k^l$ and thus restrain ourselves to the simplified question (2.13) form the introduction. In order to obtain non-zero success probabilities, we also assume that the correct solutions are discrete. Although this setup is not immediately practical, it is still a worthwhile test problem to gain some insight into the relaxation technique because the problem is still $NP$ hard in general, as shown in Appendix A.6, and therefore not unduly simplified.

3.1.2. *Recovery of the Selectors.* Let us now turn to the question if we can find the best combination of initial blocks $x_k^l$. The argument rests on the following observations:

(1) The relaxed optimization problem (3.5) for the selectors $z$ is a weighted $\ell_1$-minimization [41] with measurement matrix $AX$. This implies that the problem is convex and solvable.

(2) If a discrete selection $z^l \in \{e_1, \dots, e_r\}$ of the initial blocks $x_k^l$ picks a global optimizer of the non-convex problem (3.1), by construction, it is $\theta$-sparse and satisfies the constraints of the relaxed problem.

(3) The initial guesses in $X$ induce some extra randomness into the modified measurement matrix $AX$, so that indeed there is some hope of unique $\theta$-sparse $\ell_1$ recovery, even if $A$ does not satisfy an RIP [30]. In that case, the discrete selection $z^l$ from the previous item is the unique minimizer of the relaxed problem.

Hence, if the global optimizer is contained in the sparse combinations of the initial guesses $x_k^l$ and the matrix $AX$ does allow unique sparse recovery, then we can find the best out of $r^\theta$ combinations by a convex optimization problem of size $r\theta$, giving a positive answer to the questions raised in the introduction.

**Remark 3.1.** Although we assume that $z$ can be uniquely recovered by the relaxed problem (3.5), this does not imply that the solution $x$ of the $\ell_p$ minimization problem (3.1) is unique. If $z$ is unique but $x$ is not, this merely implies that only one solution $x$ can be sparsly combined from the blocks of $X$.

**Remark 3.2.** The argument can be generalized to global optimizers $x$ that are sparse linear combinations of the blocks $x_k^l$ as opposed to a selection of one item per block. A detailed analysis is left for future research.

3.1.3. *Main Result.* Throughout the article, for a matrix $C \in \mathbb{R}^{a,b}$, and a subset $R \subset \{1, \dots, b\}$, the matrix $C_{\cdot,R}$ consists of the columns of $C$ with indices in $R$. In order to state the main result, let us first setup all required assumptions.

First, note that the measurement matrix $AX$ includes the columns $A^l x^l$, $l = 1, \dots, \theta$. Since we only impose rank conditions in $A^l$ below, this allows degenerate cases for unfavorable global optimizers $x = [x^1, \dots, x^\theta]$, e.g. repeated columns $A^l x^l = A^k x^k$, so that unique sparse recovery is impossible. We avoid this problem by considering non-uniform recovery results for random but unknown $x$, recovered from observed right hand sides $y = Ax$.

(A1) Let $x = [x^1, \dots, x^\theta] \in [-1, 1]^N$ be a vector with i.i.d random entries on a deterministic but unknown support $S \subset \{1, \dots, N\}$ with

$$\mathbb{E}[x_j] = 0, \qquad \mathbb{E}[x_j^2] = p_x, \qquad j \in S \tag{3.6}$$

and let $S^l$ be the indices of $S$ in block $l = 1, \dots, \theta$.

Next, we guess the blocks $x_k^l$, or equivalently the matrices $X^l$, given that we know that the blocks $x^l$ of the correct solution are already contained in some guesses with unknown index $k^l$. This is quite unlikely to happen, but the current scheme should merely "boost" our chances to find good initial guesses and not be a complete method in itself, as discussed further in Section 3.2. In addition, the assumption on the sensing matrix $A$ still allow polynomial-time reductions of $NP$-hard problems to $\ell_0$-minimization, although only smaller instances than for general matrices $A$, see

Appendix A.6. Hence, we can expect some gains, but not necessarily a polynmial time algorithm for the global optimizer without any significant assumptions.

(A2) Let $X^l \in [-1,1]^{n\times r}$ be matrices that contain the vectors $x^l$ in unknown columns $k^l$ and with remaining entries i.i.d. random numbers with

$$\mathbb{E}[X^l_{j,k}] = 0, \qquad \mathbb{E}[|X^l_{j,k}|] = \nu, \qquad k \neq k^l, \qquad l = 1,\dots,\theta. \tag{3.7}$$

For abbreviation, let $T = \{k^l | \, l = 1,\dots,\theta\}$.

Note that $\nu$ can be fairly small, e.g. $|S^l|/n$ in the discussion (3.12), which allows products of Bernoulli and Subgaussian entries to generate columns of $X^l$ with sparsity comparable to $x^l$. Finally, we require the following assumptions on the measurement matrix:

(A3) Define the constants

$$F_S(A)^2 := \min_{l=1,\dots,\theta} \|A^l_{\cdot,S^l}\|_F^2, \qquad M(A)^2 := \max_{l=1,\dots,\theta} \|A^l\|^2 \tag{3.8}$$

and

$$\bar{s} := \max\left\{|S^l| : l \in \{1,\dots,\theta\}\right\}. \tag{3.9}$$

The quantity $F_S(A)^2/M(A)^2$ is related to the stable rank $\|A^l\|_F^2/\|A^l\|^2$ of a matrix and is equal to $\min_l |S^l|$ if all blocks $A^l$ have orthonormal columns, see e.g. [30] for more information. Note that all constants that depend on $A$ only do so via individual blocks $A^l$ and are independent of any relation between different blocks $A^l$ and $A^j$, $j \neq l$. Therefore, the given conditions are much weaker than the RIP and allow e.g. repetitions $A^l = U$ of identical unitary matrices. We are now ready to state the main theorem of this article.

**Theorem 3.3.** *Let Assumptions (A1), (A2) and (A3) be satisfied, $y = Ax$ for the unknown $x$ from (A1) and let $z$ be the solution of the block relaxed problem* (3.5), *with right hand side $y$. Then for any $\alpha \geq 0$ and $0 \leq \delta \leq 1$ related by*

$$1 - \delta = \frac{\sqrt{|T|\bar{s}}}{\alpha F_S(A)\sqrt{p_x}} \tag{3.10}$$

*with probability at least*

$$1 - \left[ 2(R - |T|)\exp\left(-\frac{\nu^2 n^2}{n + 2M(A)^2\alpha^2}\right) + 2\left(\frac{12}{\delta}\right)^{|T|} \exp\left(-c\frac{F_S(A)^2}{M(A)^2}\min\left\{\frac{p_x^2\delta^2}{4K^4}, \frac{p_x\delta}{2K^2}\right\}\right)\right] \tag{3.11}$$

*for some positive absolute constants $c$ and $K$, we have $x = Xz$, and hence can recover $x$.*

Since $X$ is known and $z$ is the solution of a convex optimization problem, the theorem allows us to recover the sparse unknown vector $x$ from observations $y$. However, the theorem merely requires that $x$ is $|S|$-sparse, not that it is the global optimizer of the non-convex compressed sensing problem (3.1). We can ensure the latter by requiring that $|S|$-sparse solutions are unique.

(A4) Assume that for all $y$ the system $Ax = y$ has at most one $|S|$-sparse solution.

This is equivalent to all selections of $2|S|$ columns of $A$ being non-singular [23] and considerably weaker than the RIP, which requires tight bounds on the singular values of these sub-matrices.

**Corollary 3.4.** *Let the Assumptions (A1), (A2), (A3) and (A4) be satisfied. Then the unknown vector $x$ is a global optimizer of the $\ell_0$-minimization problem* (1.1) *and with probability at least* (3.11) *the solution $z$ of the relaxed optimization problem* (3.5) *satisfies $x = Xz$.*

With suitable modifications of Assumption (A4) and $\ell_p$-recovery results in [10,11, 22,45,49] one can also obtain analogous statements for the $\ell_p$ minimization problem (3.1).

In conclusion, this result provides a positive answer to the question raised in the introduction: For the non-convex compressed sensing problem (1.1), the relaxation strategy of this section can indeed find the best of $r^\theta$ possible combinations of the initial guess $x_k^l$ by solving a convex optimization problem of non-exponential size $r\theta$.

3.2. **Recovery Probability.** In order to disentangle all requirements and statements in Theorem 3.3, in this section, we consider some more specific scenarios. To this end, in the following let $\gtrsim$, $\lesssim$ and $\sim$ denote greater, smaller and equivalence up to some generic constants independent of the problem dimensions, sparsity and expectations such as $\nu$ or $p_x$.

First, we assume that the support $S$ is equidistributed among the blocks, i.e. that there is some $s$ with $|S| = \theta s$ and

$$|S^l| \sim s.$$

If we choose $S$ uniformly at random, this is satisfied with high probability for sufficiently large $s$. Indeed, $|S^l|$ is distributed by a hypergeometric distribution so that the observation easily follows from Chebyshev's inequality.

Next, assume that $m = n$, so that the blocks $A^l$ are square matrices, and that the columns of $A^l$ are (almost) orthonormal. This implies that

$$\|A_l\| \sim 1, \qquad \|A^l\|_F^2 \sim n, \qquad \|A^l_{\cdot,S^l}\|_F^2 \sim s$$

and therefore

$$F_S(A)^2 \sim s, \qquad \overline{s} \sim s, \qquad M(A) \sim 1, \qquad \frac{F_S(A)^2}{M(A)^2} \sim s.$$

There are no relations between the columns of different blocks, so e.g. it is legitimate to choose all blocks equal, which clearly violates the RIP condition.

Finally, we assume that we have some good a-priory knowledge of the size $|S|$ and choose

$$\nu \sim s/n, \qquad p_x \sim 1. \tag{3.12}$$

The second probability of $p_x$ states that given the information that we are on the support of $x$, the entries are not overly strongly clustered around zero. With these constants, the probability of recovery failure of Theorem (3.3) is at most

$$2\exp\left(-c\frac{s^2}{n+M(A)^2\alpha^2}+\ln(R)\right)+2\exp\left(-cs\min\left\{\frac{p_x^2\delta^2}{4K^4},\frac{p_x\delta}{2K^2}\right\}+|T|\ln\left(\frac{12}{\delta}\right)\right) \tag{3.13}$$

for some generic constant $c$, which may differ from the one in the theorem and change in the calculations below. The algorithm does not depend on the choice of $\alpha$ and $\delta$, which we can now choose to bound this failure probability. To this end, let us choose $\delta \sim (1-\delta) \sim 1$ so that by (3.10) and $|T| = \theta$, we have

$$1 \sim (1-\delta) \sim \frac{\sqrt{|T|\bar{s}}}{\alpha F_S(A)} \sim \frac{\sqrt{\theta s}}{\alpha} \quad \Leftrightarrow \quad \alpha \sim \sqrt{\theta s}.$$

Thus, the failure probability reduces to

$$2\exp\left(-c\frac{s^2}{n+M(A)^2\theta s}+\ln(R)\right)+2\exp\left(-cs+C\theta\right) =: (I)+(II) \tag{3.14}$$

for some new generic constant $C$. Using $M(A) \sim 1$, we have $\frac{s^2}{n+M(A)^2\theta s} \gtrsim \min\left\{\frac{s^2}{n},\frac{s}{\theta}\right\}$, which must be larger than $\ln(R) = \ln(r\theta)$ for the exponent of $(I)$ to be negative. Therefore, we must have

$$r \lesssim \min\left\{\frac{1}{\theta}e^{\frac{s}{\theta}},\frac{1}{\theta}e^{\frac{s^2}{n}}\right\}. \tag{3.15}$$

The first component of the minimum ensures that $s \gtrsim \theta$, which implies that also $(II)$ has a negative exponent.

First note that the condition (3.15) limits the number $r = R/\theta$ of possible trials. Depending on the relative sizes of $s$, $n$ and $\theta$, this number can be exponentially large and the block relaxed scheme is able to correctly select an exponentially large number of pieces $x^l_{k^1}, \dots, x^l_{k_\theta}$ contained in the guesses $X^l$ for a non-linear problem.

Second, the condition (3.15) implies that the sparsity $s$ per block cannot be too small. The reason is that we have very limited assumptions on $A$. In particular the sensing matrix $AX$ contains the columns $A^l x^l$. If $x^l$ is overly sparse, this does not guarantee enough randomness to ensure sparse recovery.

With the probabilities $p_l$ that the blocks of $X^l$ contain the solution blocks $x^l$ and the success probability $p_{select}$ of sparse recovery from (3.14), by the arguments in Section 2.5, we have the probabilities

$$P(\text{success } r \text{ trials}) \approx p_l^\theta r$$

$$P(\text{success block relaxation}) \approx p_l^\theta (p_{select} r^\theta)$$

to recover a $|S|$-sparse vector with constraint $Ax = y$, with very weak conditions on $A$. If this matrix allows unique sparse recovery from $\ell_p$-minimization, it is also the global minimizer of (3.1). Given the conditions in (3.15), we can ensure that $p_{select}$ is close to one so that the block relaxation provides a $r^\theta/r$ enhanced chance to find the solution over $r$ repeated guesses.

In order for the probability $p_l$ to be non-zero, we need to sample $x^l$ and $X^l$ from discrete distributions. Even with this restrictive assumption, $p_l$ is still of negligible size and the resulting success probability of block relaxation is excessively small.

This is not fully unexpected because with the given assumptions on the sensing matrix $A$ and discrete $x$ in e.g. $\{-1, -1/2, 0, 1/2, 1\}$ the $\ell_p$-minimization problem is still $NP$-hard in general, see Appendix A.6. Also recall that for a practical algorithm we would incorporate the blocks $X^l$ into the optimization as in (2.11), which removes the requirement to correctly guess the blocks $x^l$ in one shot and with it the requirement of $x^l$ to be discrete. Alternatively, the followup paper [52], uses the relaxation argument together with a learning algorithm that increase the chance $p_l$ to find good initial guesses, resulting in a tractable scheme to find global optima for some non-convex compressed sensing problems.

3.3. **Proof of Theorem 3.3.** In this section, we prove Theorem 3.3. The proof follows standard lines for sparse recovery results, with some slight twists for the added structure. In Section 3.3.1, we first introduce some notations and setup used throughout the entire section. Then, we show concentration estimates (Section 3.3.2), RIP type results for fixed sparse subsets $S$ (Section 3.3.3) and then finally combine these results for a non-uniform recovery argument in Section 3.3.4.

3.3.1. *Notations and Setup.* Let all assumptions from Theorem 3.3 be satisfied. We calculate the sparse recovery probability given that the blocks $x^l$ are contained in the columns of the respective matrices $X^l$. W.l.o.g, we assume that $x^l$ are always the first columns so that $X$ has the block structure

$$X = \begin{pmatrix} x^1 & \bar{X}^1 & & \\ & \ddots & \ddots & \\ & & x^\theta & \bar{X}^\theta \end{pmatrix}, \qquad x^l \in \mathbb{R}^{n \times 1}, \qquad \bar{X}^l \in \mathbb{R}^{n \times r-1}, \tag{3.16}$$

where $\bar{X}^l$ are i.i.d. random matrices.

**Remark 3.5.** In Theorem 3.3, we show a sparse recovery result only with high probability. Therefore, we must ensure that the matrices $X$ with a given sparsity pattern $S$ in one column $x^l$ are not included in the low probability set where spare recovery fails. Hence, we make this patter explicit in our proof.

By assumptions (3.6) and (3.7) of Theorem 3.3, we have the following expectation, variances and $\psi_2$-norms:

$$\begin{aligned} &\mathbb{E}[x_j] = 0, && \mathbb{E}[x_j^2] = p_x, && \|x_j\|_{\psi_2} \leqslant K, && j \in S \\ &\mathbb{E}[X^l_{jk}] = 0, && \mathbb{E}[(X^l_{jk})^2] = p_X, && \|X^l_{jk}\|_{\psi_2} \leqslant K, && \begin{aligned} j &= 1, \dots, n, \\ k &= 1, \dots r-1 \end{aligned}, \end{aligned} \tag{3.17}$$

for some $p_X \geqslant 0$ constant $K > 0$ and $\psi_2$-norm defined by $\|x\|_{\psi_2} := \sup_{a \geqslant 1} a^{-1/2} (\mathbb{E}[|x|^a])^{1/a}$, see e.g. [43]. Note that all variances and $\psi_2$ norms are bounded because by assumption the entries of $X^l$, including the first column $x^l$, are in the interval $[-1, 1]$.

3.3.2. *Concentration Estimates.* In this section, we state concentration estimates for $\|AXu\|$ for some $u \in \mathbb{R}^R$. To this end, let us split an arbitrary vector $u \in \mathbb{R}^R$ according to the block structure (3.16) of $X$ as

$$u := \begin{pmatrix} v^1 & u^1 & \cdots & v^\theta & u^\theta \end{pmatrix}^T,$$

with $v^l \in \mathbb{R}$ and $u^l \in \mathbb{R}^{r-1}$. Then the concentration inequality is shown with respect to the weighted norm

$$\|u\|_A^2 := \|W_A u\| := \sum_{l=1}^{\theta} \left\{ p_x \|A_{\cdot,S^l}\|_F^2 |v^l|^2 + p_X \|A^l\|_F^2 \|u^l\|^2 \right\}. \tag{3.18}$$

with diagonal weight matrix

$$W_A = \operatorname{diag}(\sqrt{p_x}\|A_{\cdot,S^1}\|_F, \sqrt{p_X}\|A^1\|_F, \dots, \sqrt{p_x}\|A_{\cdot,S^\theta}\|_F, \sqrt{p_X}\|A^\theta\|_F). \tag{3.19}$$

Recall that $S^l$ are the indices in $S$ contained in the block $X^l$ of $X$. For the remainder of this section $c$ denotes a positive absolute constant.

**Proposition 3.6.** *Let $A$ and $X$ be the matrices defined in* (3.2) *and* (3.16) *with independent sub-Gaussian entries satisfying* (3.17)*. Then, for every $u \in \mathbb{R}^R$ and $\epsilon \geqslant 0$,*

$$\Pr\left[\left|\|AXu\|^2 - \|u\|_A^2\right| \geqslant \epsilon F^2\right] \leqslant 2\exp\left(-c\frac{F_S(A)^2}{M(A)^2}\min\left\{\frac{\epsilon^2}{K^4}, \frac{\epsilon}{K^2}\right\}\right)$$

*with*

$$F^2 = \sum_{l=1}^{\theta}\left\{\|A_{\cdot,S^l}\|_F^2\|v^l\|^2 + \|A^l\|_F^2\|u^l\|^2\right\}$$

*and $F_S(A)$ and $M(A)$ defined in* (3.8).

We only need a corollary of this proposition for $u$ restricted to the support set $T$ of the selectors $z$.

**Corollary 3.7.** *Let $A$ and $X$ be the matrices defined in* (3.2) *and* (3.16) *with independent sub-Gaussian entries satisfying* (3.17)*. Then, for every $u \in \mathbb{R}^R$ supported on $T$ and $\epsilon \geqslant 0$*

$$\Pr\left[\left|\|AXu\|^2 - \|u\|_A^2\right| \geqslant \epsilon\|u\|_A^2\right] \leqslant 2\exp\left(-c\frac{F_S(A)^2}{M(A)^2}\min\left\{\frac{p_x^2\epsilon^2}{K^4}, \frac{p_x\epsilon}{K^2}\right\}\right),$$

*with $F_S(A)$ and $M(A)$ defined in* (3.8) *and $p_x$ defined in* (3.6).

*Proof.* Since $u$ is supported on $T$, we have $\|u^l\|_2^2 = 0$ for all $l = 1, \dots, \theta$ and therefore the definition of $F$ in Proposition 3.6 and the definition (3.18) of the $\|\cdot\|_A$-norm yield

$$F^2 = \sum_{i=1}^{\theta}\|A_{\cdot,S^l}\|_F^2\|v^l\|^2, \qquad \|u\|_A^2, = \sum_{i=1}^{\theta}p_x\|A_{\cdot,S^l}\|_F^2\|v^l\|^2.$$

Thus, we have $p_x F^2 = \|u\|_A^2$, which proves the corollary. □

The proof of Proposition 3.6 is similar to [30], and uses the following corollary of the Hanson-Wright inequality, see Appendix A.2 for more details.

**Corollary 3.8.** *Let $v \in \mathbb{R}^d$ be a vector with independent components with $\mathbb{E}[v_i] = 0$ and $\|v_i\|_{\psi_2} \leqslant K$ and $C^TC \in \mathbb{R}^{d\times d}$ be a matrix. Then, for every $t \geqslant 0$,*

$$\Pr\left[\left|v^TC^TCv - \mathbb{E}[v^TC^TCv]\right| \geqslant \epsilon\|C\|_F^2\right] \leqslant 2\exp\left(-c\frac{\|C\|_F^2}{\|C\|^2}\min\left\{\frac{\epsilon^2}{K^4}, \frac{\epsilon}{K^2}\right\}\right).$$

In order to apply the corollary, we construct a vectorization $\hat{X}$ of the matrix $X$ and a matrix $B$ with $\hat{X}^T B^T B \hat{X} = \|AXu\|^2$. Let us first consider this vectorization for a generic matrix $M \in \mathbb{R}^{a \times b}$, vector $w \in \mathbb{R}^c$ and random matrix $R \in \mathbb{R}^{b \times c}$ with i.i.d entries, expectation $\mathbb{E}[r_{ij}] = 0$ and variance $\mathbb{E}[r_{ij}^2] = V$. By Appendix A.3, we identify $R$ with the tensor $\hat{R} \in \mathbb{R}^b \otimes \mathbb{R}^c$ and have

$$MRw = (M \otimes w^T)\hat{R} \tag{3.20}$$

and

$$\mathbb{E}\left[\|MRw\|^2\right] = V\|M\|_F^2 \|w\|^2. \tag{3.21}$$

Let us now construct the vectorization $\hat{X}$ and $B$ with $B\hat{X} = AXu$. Instead of applying the last two identities directly, we are a little more careful with regard to the block structure. $\hat{X}$ is defined by

$$\hat{X} := \begin{pmatrix} \hat{x}^1 & \hat{X}^1 & \cdots & \hat{x}^\theta & \hat{X}^\theta \end{pmatrix} \in \bigtimes_{l=1}^{\theta} \left( \mathbb{R}^{|S^l|} \times \mathbb{R}^{n(r-1)} \right)$$

where the $\hat{x}^l$ and $\hat{X}^l := \hat{\bar{X}}^l$ are the vectorizations of the restriction $x_{S^l}^l$ of $x^l$ to its support $S^l$, and $X^l$, respectively. Likewise, $B$ is defined by

$$B := \begin{pmatrix} A_{\cdot,S^1} \otimes (v^1)^T & A^1 \otimes (u^1)^T & \cdots & A_{\cdot,S^\theta} \otimes (v^\theta)^T & A^\theta \otimes (u^\theta)^T \end{pmatrix}. \tag{3.22}$$

where $v^l$ is considered as a $1 \times 1$ matrix. Then, by (3.20) the vectorization of the product $AXu$ is given by

$$B\hat{X} = \sum_{l=1}^{\theta} (A_{\cdot,S^l} \otimes (v^l)^T)\hat{x}^l + (A^l \otimes (u^l)^T)\hat{X}^l = \sum_{l=1}^{\theta} A_{\cdot,S^l} x_{S^l}^l v^l + A^l X^l u^l = AXu. \tag{3.23}$$

This allows us to prove Proposition 3.6 with Corollary 3.8 of the Hanson-Wright inequality.

*Proof of Proposition 3.6.* From the vectorization (3.23) we have

$$\|AXu\|^2 = \|B\hat{X}\|^2 = \hat{X}^T B^T B \hat{X}.$$

so that we can use Corollary 3.8 of the Hanson-Wright inequality to show concentration inequalities for $\|AXu\|^2$. To this end, in the following we compute all terms in the Corollary. We start with the expectation value:

$$\begin{aligned} \mathbb{E}\left[\|AXu\|^2\right] &= \mathbb{E}\left[ \left\| \sum_{l=1}^{\theta} \left\{ A^l x^l v^l + A^l \bar{X}^l u^l \right\} \right\|^2 \right] \\ &= \sum_{l=1}^{\theta} \left\{ \mathbb{E}\left[ \left\| A^l x^l v^l \right\|^2 \right] + \mathbb{E}\left[ \left\| A^l \bar{X}^l u^l \right\|^2 \right] \right\}, \end{aligned}$$

where we have used that because of independence and zero mean of the entries, all cross terms $\left\langle A^l \bar{X}^l u^l, A^j \bar{X}^j u^j \right\rangle$ and $\left\langle A^l x^l v^l, A^j x^j v^j \right\rangle$ for $l \neq j$ and $\left\langle A^l \bar{X}^l u^l, A^j x^j v^j \right\rangle$ for all $l, j$ vanish.

Using the zero-mean property and the variances defined in (3.17) and applying (3.21) yields

$$\mathbb{E}\left[\left\|A^l x^l v^l\right\|^2\right] = \mathbb{E}\left[\left\|A_{\cdot,S^l} x^l_{S^l} v^l\right\|^2\right] = p_x \|A_{\cdot,S^l}\|_F^2 |v^l|^2$$

$$\mathbb{E}\left[\|A^l \bar{X}^l u^l\|^2\right] = p_X \|A^l\|_F^2 \|u^l\|^2.$$

In conclusion, we have

$$\mathbb{E}\left[\|AXu\|^2\right] = \sum_{i=1}^{m} \left\{ p_x \|A_{\cdot,S^l}\|_F^2 |v^l|^2 + p_X \|A^l\|_F^2 \|u^l\|^2 \right\} = \|u\|_A^2. \tag{3.24}$$

Note that if we could normalize both $\|A_{\cdot,S^l}\|_F$ and $\|A^l\|_F$ to one, the right hand side would reduce to $\|u\|^2$. However that is not possible because $A_{\cdot,S^l}$ is a sub-matrix of $A^l$.

The next quantity in Corollary 3.8 is the Frobenius norm

$$\|B\|_F^2 = \sum_{l=1}^{\theta} \left\{ \|A_{\cdot,S^l}\|_F^2 \|v^l\|^2 + \|A^l\|_F^2 \|u^l\|^2 \right\} = F^2. \tag{3.25}$$

The spectral norm can easily be computed with (A.2) in the appendix, which yields

$$\|B\|^2 \leqslant \sum_{l=1}^{\theta} \|A_{\cdot,S^l}\|^2 \|v^l\|^2 + \|A^l\|^2 \|u^l\|^2.$$

Together with (3.25) and using that $\|A_{\cdot,S^l}\|_F \leqslant \|A^l\|_F$ and $\|A_{\cdot,S^l}\| \leqslant \|A^l\|$ this yields

$$\frac{\|B\|_F^2}{\|B\|^2} \geqslant \frac{\min_{l=1,\dots,\theta} \|A_{\cdot,S^l}\|_F^2}{\max_{l=1,\dots,\theta} \|A^l\|^2} = \frac{F_S(A)^2}{M(A)^2}. \tag{3.26}$$

We have calculated all terms in Corollary 3.8 of the Hanson-Wright inequality, which implies

$$\Pr\left[\left|\|AXu\|^2 - \|u\|_A^2\right| \geqslant \epsilon \|B\|_F^2\right] \leqslant 2 \exp\left( -c \frac{\|B\|_F^2}{\|B\|^2} \min\left\{ \frac{\epsilon^2}{K^4}, \frac{\epsilon}{K^2} \right\} \right),$$

which by (3.25) and (3.26) proves the proposition. □

3.3.3. *RIP Type Estimates.* We show a RIP like estimate, only for one fixed sparse set $T \subset \{1, \dots, R\}$. The result and proof are identical to [5, Lemma 5.1] only with the $\ell_2$-norm replaced by the $\|\cdot\|_A$-norm.

**Lemma 3.9.** *Let all assumptions of Corollary 3.7 be satisfied. Then for the set* $T \subset \{1, \dots, R\}$ *containing the columns* $x^l$ *and* $0 < \delta < 1$*, we have*

$$(1 - \delta)\|u\|_A \leqslant \|AXu\| \leqslant (1 + \delta)\|u\|_A$$

*for all* $u \in \mathbb{R}^R$ *supported on* $T$ *with probability at least*

$$1 - 2\left(\frac{12}{\delta}\right)^{|T|} \exp\left( -c \frac{F_S(A)^2}{M(A)^2} \min\left\{ \frac{p_x^2 \delta^2}{4K^4}, \frac{p_x \delta}{2K^2} \right\} \right). \tag{3.27}$$

*Proof.* Let $U_T \subset \mathbb{R}^R$ be the vectors with support contained in $T$. Then, there is a $\delta/4$ cover $Q_T$ of the unit sphere in $U_T$ with respect to the $\|\cdot\|_A$-norm with $|Q_T| \leqslant (12/\delta)^{|T|}$, see e.g. [23, 35]. From the concentration inequality Corollary 3.7 with $\epsilon = \delta/2$, together with a union bound, we have that

$$\left(1 - \frac{\delta}{2}\right)\|u\|_A^2 \leqslant \|AXu\|^2 \leqslant \left(1 + \frac{\delta}{2}\right)\|u\|_A^2$$

with probability at least (3.27). This is analogous to [5, (5.4)]. Using the remainder of the proof in the reference verbatim, shows that

$$(1-\delta)\|u\|_A \leqslant \|AXu\| \leqslant (1+\delta)\|u\|_A$$

for all $u$ supported on $T$, which completes the proof. □

**Corollary 3.10.** *Let all assumptions of Lemma 3.9 be satisfied and assume that the weight matrix $W_A$ of the $\|\cdot\|_A$-norm defined in* (3.19) *is invertible. Then, with probability at least* (3.27) *the singular values $\sigma_i$ of the matrix $(AXW_A^{-1})_{\cdot,T}$ satisfy*

$$1 - \delta \leqslant \sigma_i \leqslant 1 + \delta.$$

*Proof.* With the definition (3.19) of $W_A$, Lemma 3.9 implies that

$$(1-\delta)\|W_A u\| \leqslant \|AXu\| \leqslant (1+\delta)\|W_A u\|$$

with the given probability (3.27) for all $u$ with support $T$. With $z := W_A u$, this implies

$$(1-\delta)\|z\| \leqslant \|AXW_A^{-1}z\| \leqslant (1+\delta)\|z\|.$$

Choosing right singular vectors of $AXW_A^{-1}$ restricted to columns in $T$ for $z$, directly yields the result. □

3.3.4. *Sparse Recovery.* The remaining proof of the sparse recovery Theorem 3.3, is analogous to non-uniform sparse recovery as in e.g. [23, Theorem 9.16].

By the assumptions of Theorem 3.3, the vectors $x^l$ are contained as columns in the blocks $X^l$. We denote the indices of these columns as $T \subset 1, \dots, R$. In (3.16) above, we have w.l.o.g. assumed that these are the first columns in the respective blocks $X^l$. Note, however, that this choice was only for notational convenience and in general $T$ is unknown, except for some rudimentary properties like $t := |T| = \theta$. In addition, note that the set $T$ coincides with the support of the selector $z$ and the major goal of the sparse recovery problem (3.5) is to find this vector.

In the following, let $W_\ell \in \mathbb{R}^{R\times R}$ be the diagonal matrix with $(W_\ell)_{kk} = \|X_{\cdot,k}\|_p^p$, which constitutes the weights in the weighted $\ell_1$-minimization (3.5) and $W_{\ell T}$ the restriction to the index set $T$. On this special index set, we have

$$(W_\ell)_{kk} = \|X_{\cdot,k}\|_p^p = \|x^l\|_p^p \leqslant |S^l|, \tag{3.28}$$

if $k$ is in the block $l$, where we have used that $x^l$ has entries in the interval $[-1, 1]$ on its support.

In order to simplify notations, for any matrix $C$, let $C^{+*} = (C^*)^+ = (C^+)^*$ be the adjoint of the pseudo inverse.

Before we prove the main result Theorem 3.3, we need two more lemmas.

**Lemma 3.11.** *For any $\alpha \geqslant 0$ and $0 \leqslant \delta \leqslant 1$ satisfying* (3.10) *and for any $u \in \mathbb{R}^R$ with support on $T$, we have*

$$(3.29)\quad P\left(\alpha \leqslant \|(AX_T)^{+*} W_{\ell T} \operatorname{sign}(u_T)\|\right) \leqslant 2\left(\frac{12}{\delta}\right)^{|T|} \exp\left(-c\frac{F_S(A)^2}{M(A)^2} \min\left\{\frac{p_x^2\delta^2}{4K^4}, \frac{p_x\delta}{2K^2}\right\}\right)$$

*for constants $c, K$ from Corollary 3.10.*

*Proof.* Let us use the abbreviations

$$v := (AX_T)^{+*} W_{\ell T} \operatorname{sign}(u_T), \qquad W_{AT} := (W_A)_{\cdot,T}, \qquad F := F_S(A)$$

for the weight matrix $W_A$ of the $\|\cdot\|_A$ norm defined in (3.19). Then, the left hand side of (3.29) becomes $P(\alpha \leqslant \|v\|)$. Before we estimate this probability, we calculate an estimate for $\|v\|$. By the definition (3.8) of $F = F_S(A)$ and the definition of $T$ we have $\|W_{AT}^{-1}\| \leqslant 1/(F\sqrt{p_x})$. Let $\sigma_{min}$ be the smallest singular value of $AX_T W_{AT}^{-1}$. Since $W_{AT}$ is invertible, we have

$$v \in \operatorname{range}[(AX_T)^{+*}] = \ker[(AX_T)^*]^\perp = \ker[W_{AT}^{-1}(AX_T)^*]^\perp = \ker[(AX_T W_{AT}^{-1})^*]^\perp$$

and therefore

$$\|v\| \leqslant \frac{1}{\sigma_{min}}\|(AX_T W_{AT}^{-1})^* v\| = \frac{1}{\sigma_{min}}\|W_{AT}^{-1}(AX_T)^* v\| \leqslant \frac{1}{\sigma_{min} F\sqrt{p_x}}\|(AX_T)^* v\|.$$

Plugging in the definition of $v$ and using that $(AX_T)^*(AX_T)^{+*}$ is an orthogonal projector with matrix norm bounded by one, we conclude that

$$\|v\| \leqslant \frac{1}{\sigma_{min} F\sqrt{p_x}}\|W_{\ell T} \operatorname{sign}(u_T)\| \leqslant \frac{\overline{s}}{\sigma_{min} F\sqrt{p_x}}\|\operatorname{sign}(u_T)\| = \frac{\sqrt{|T|}\overline{s}}{\sigma_{min} F\sqrt{p_x}},$$

where in the second inequality we have used (3.28) and the definition (3.9) of $\overline{s}$.

We now proceed with the estimate of the probability in the left hand side of (3.29). For any $\alpha \geqslant 0$, we have

$$\alpha \leqslant \|v\| \leqslant \frac{\sqrt{|T|}\overline{s}}{\sigma_{min} F\sqrt{p_x}}$$

and thus using the assumption (3.10) in the last identity

$$P\left(\alpha \leqslant \|(AX_T)^{+*} W_{\ell T} \operatorname{sign}(u_T)\|\right) = P(\alpha \leqslant \|v\|) \leqslant P\left(\alpha \leqslant \frac{\sqrt{|T|}\overline{s}}{\sigma_{min} F\sqrt{p_x}}\right) = P\left(\sigma_{min} \leqslant \frac{\sqrt{|T|}\overline{s}}{\alpha F\sqrt{p_x}}\right) = P(\sigma_{min} \leqslant 1 - \delta).$$

The latter probability is smaller, than the probability that there is any singular value that is not contained in the interval $[1-\delta, 1+\delta]$ and thus Corollary 3.10 implies (3.29). □

**Lemma 3.12.** *Let $x \in [-1,1]^d$, $d \geqslant 1$ be a random vector with zero mean and expectation $\mathbb{E}[|x_i|] = \nu$, $i = 1, \dots, d$. Then for any $v \in \mathbb{R}^d$, we have*

$$P\left(\langle x, v\rangle \geqslant \|x\|_p^p\right) \leqslant 2\exp\left(-\frac{\nu^2 d^2}{d + 2\|v\|_2^2}\right).$$

*Proof.* Since $-1 \leqslant x_i \leqslant 1$ and $p \leqslant 1$, we have $|x_i|^p \geqslant |x_i|$ so that $\|x\|_p^p \geqslant \|x\|_1$ and therefore

$$\begin{aligned} P\left(|\langle x, v\rangle| \geqslant \|x\|_p^p\right) &\leqslant P\left(|\langle x, v\rangle| \geqslant \|x\|_1\right) \\ &\leqslant P\left(\langle x, v\rangle \geqslant \|x\|_1\right) + P\left(\langle -x, v\rangle \geqslant \|x\|_1\right). \end{aligned}$$

It suffices to estimate the first summand in the right hand side, the other follows analogously. We have

$$P\left(\langle x, v\rangle \geqslant \|x\|_1\right) = P\left(\sum_{i=1}^{d} [x_i v_i - |x_i| + \nu] \geqslant \nu d\right) =: P\left(\sum_{i=1}^{d} X_i \geqslant \nu d\right),$$

with $X_i := x_i v_i - |x_i| + \nu$. By construction, $X_i$ has zero mean and from $X_i - \nu = |x_i|(\operatorname{sign}(x_i) v_i - 1)$ and $-1 \leqslant x_i \leqslant 1$, we obtain

$$-|v_i| - 1 \leqslant X_i - \nu \leqslant \max\{0, |v_i| - 1\}$$

so that $X_i$ is contained in an interval of length

$$w_i = \max\{1 + |v_i|, 2|v_i|\}.$$

It follows that $w_i^2 \leqslant 2 + 4v_i^2$ and therefore, Hoeffding's inequality implies

$$P\left(\langle x, v\rangle \geqslant \|x\|_1\right) \leqslant \exp\left(-\frac{\nu^2 d^2}{d + 2\|v\|_2^2}\right).$$

Using the same estimate for $P\left(\langle -x, v\rangle \geqslant \|x\|_1\right)$, concludes the proof. □

We are now ready to prove Theorem 3.3.

*Proof of Theorem 3.3.* The proof is a variant of [23, Theorem 9.16]. We start by estimating the probability that the sparse recovery in (3.5) fails. According to the optimality criteria (A.3) for weighted compressed sensing, with the weight $W_\ell$ defined before (3.28) and the complement $\bar{T}$ of $T$, the probability of failure is bounded by

$$\begin{aligned} &P\left(\exists k \in \bar{T} : |\left\langle AX_{\cdot,k}, (AX_{\cdot,T})^{+*} W_{\ell T} \operatorname{sign}(z_T)\right\rangle| \geqslant (W_\ell)_{kk}\right) \\ &\qquad = P\left(\exists k \in \bar{T} : |\left\langle A^{l(k)} X_{\cdot,k}^{l(k)}, (AX_{\cdot,T})^{+*} W_{\ell T} \operatorname{sign}(z_T)\right\rangle| \geqslant (W_\ell)_{kk}\right), \end{aligned}$$

where we have used the block structure of $X$ and $l(k)$ is the number of the block $l$ that contains the index $k \in \{1, \dots, r\theta\}$. With $v := (AX_{\cdot,T})^{+*} W_{\ell T} \operatorname{sign}(z_T)$ we can estimate this by

$$\begin{aligned} &P\left(\exists k \in \bar{T} : |\left\langle X_{\cdot,k}^{l(k)}, (A^{l(k)})^* v\right\rangle| \geqslant (W_\ell)_{kk}\right) \\ &\qquad \leqslant P\left(\exists k \in \bar{T} : |\left\langle X_{\cdot,k}^{l(k)}, (A^{l(k)})^* v\right\rangle| \geqslant (W_\ell)_{kk} \,\middle|\, \|v\| \leqslant \alpha\right) + P(\|v\| \geqslant \alpha). \end{aligned}$$

Note that the columns of $X$ involved in $v$ and $X_{\cdot,k}^{l(k)}$ are mutually exclusive, so that these two objects are independent. Therefore, using $(W_\ell)_{kk} = \|X_{\cdot,k}\|_p^p = \|X_{\cdot,k}^{l(k)}\|_p^p$ and $\mathbb{E}[|X_{j,k}^{l(k)}|] = \nu$ for $k \in \bar{T}$ from assumption (3.7) by Lemma 3.12, we have

$$P\left(|\left\langle X_{\cdot,k}^{l(k)}, (A^{l(k)})^* v\right\rangle| \geqslant (W_\ell)_{kk} \,\middle|\, \|v\| \leqslant \alpha\right) \leqslant 2\exp\left(-\frac{\nu^2 n^2}{n + 2\|(A^{l(k)})^* v\|^2}\right).$$

Since by (3.8) we $\|(A^{l(k)})^*v\| \leqslant M(A)\|v\| \leqslant M(A)\alpha$ and we have $R - |T|$ possible choices for $k$, applying a union bound yields

$$P\left(\exists k \in \bar{T} : |\left\langle X^{l(k)}_{\cdot,k}, (A^{l(k)})^* v\right\rangle| \geqslant (W_\ell)_{kk}\right) \leqslant 2(R-|T|)\exp\left(-\frac{\nu^2 n^2}{n + 2M(A)^2\alpha^2}\right).$$

Finally, estimating $P(\|v\| \geqslant \alpha)$ by Lemma 3.11, we conclude that

$$\begin{aligned} P(\text{recovery fail}) \leqslant{} & 2(R-|T|)\exp\left(-\frac{\nu^2 n^2}{n + 2M(A)^2\alpha^2}\right) \\ & + 2\left(\frac{12}{\delta}\right)^t \exp\left(-c\frac{F_S(A)^2}{M(A)^2}\min\left\{\frac{p_x^2\delta^2}{4K^4}, \frac{p_x\delta}{2K^2}\right\}\right) \end{aligned}$$

where by the assumption (3.10) of Lemma 3.11 the constants are related by

$$1 - \delta = \frac{\sqrt{|T|\bar{s}}}{\alpha F_S(A)\sqrt{p_x}},$$

which completes the proof. □

## Appendix A. Appendix

### A.1. **Probability Limits.**

**Lemma A.1.** *Assume that* $p = p(\gamma) \geqslant 0$, $q = q(\gamma) > 0$ *are two functions and that* $\lim_{\gamma\to\infty} p = \lim_{\gamma\to\infty} q = 0$. *Then*

$$\lim_{\gamma\to\infty}[1-p]^{1/q} = \begin{cases} 1 & \textit{if } \lim_{\gamma\to\infty}\frac{p}{q} = 0 \\ 0 & \textit{if } \lim_{\gamma\to\infty}\frac{p}{q} = \infty \end{cases}.$$

*Proof.* By $[1-p]^{1/q} = \exp\left(\frac{1}{q}\ln(1-p)\right)$ it is sufficient to compute the limit of the exponent. l'Hospital's rule yields:

$$\lim_{\gamma\to\infty}\frac{\ln(1-p)}{q} = \lim_{\gamma\to\infty}\frac{\frac{-1}{1-p}p'}{q'} = \underbrace{\left(\lim_{\gamma\to\infty}\frac{1}{1-p}\right)}_{=1}\left(\lim_{\gamma\to\infty} -\frac{p'}{q'}\right).$$

Applying l'Hospital's rule again to the remaining term on the left hand side, we obtain

$$\lim_{\gamma\to\infty}\frac{\ln(1-p)}{q} = \lim_{\gamma\to\infty} -\frac{p'}{q'} = \lim_{\gamma\to\infty} -\frac{p}{q}.$$

which directly implies the statement of the lemma. □

**Lemma A.2.** *Assume that* $p = p(\gamma) \geqslant 0$, $q = q(\gamma) > 0$ *are two functions and that*

$$\lim_{\gamma\to\infty} p = 0, \qquad \lim_{\gamma\to\infty} q = 0, \qquad \lim_{\gamma\to\infty}\frac{p}{q} = 0.$$

*Then*

$$\lim_{\gamma\to\infty}\frac{1-[1-p]^{1/q}}{p/q} = 1$$

*Proof.* By l'Hospital's rule we have

$$\begin{aligned}\lim_{\gamma\to\infty}\frac{1-[1-p]^{1/q}}{p/q} &= \lim_{\gamma\to\infty} -\frac{[1-p]^{1/q}\left(\frac{\ln(1-p)}{q}\right)'}{(p/q)'}\\ &= \underbrace{\left(\lim_{\gamma\to\infty}[1-p]^{1/q}\right)}_{=1 \text{ by Lemma (A.1)}}\left(\lim_{\gamma\to\infty} -\frac{\left(\frac{p}{q}\frac{\ln(1-p)}{p}\right)'}{(p/q)'}\right)\end{aligned}$$

Since $\lim_{\gamma\to\infty}\frac{\ln(1-p)}{p} = -1$ and the assumption $\lim_{\gamma\to\infty}\frac{p}{q} = 0$, we can apply the $\frac{0}{0}$ case of l'Hospital's rule in reverse to the remaining part on the right hand side and obtain

$$\lim_{\gamma\to\infty}\frac{1-[1-p]^{1/q}}{p/q} = \lim_{\gamma\to\infty} -\frac{\left(\frac{p}{q}\frac{\ln(1-p)}{p}\right)'}{(p/q)'} = \lim_{\gamma\to\infty} -\frac{\frac{p}{q}\frac{\ln(1-p)}{p}}{p/q} = 1.$$

□

A.2. **Hanson-Wright Inequality.** For the Hanson-Wright Inequality, see e.g. [30, 43] and the references therein.

**Theorem A.3** (Hanson-Wright Inequality, [43, Theorem 1.1]). *Let $v \in \mathbb{R}^d$ be a vector with independent components with $\mathbb{E}[v_i] = 0$ and $\|v_i\|_{\psi_2} \leqslant K$ and $M \in \mathbb{R}^{d\times d}$ be a matrix. Then, for every $t \geqslant 0$,*

$$\Pr\left[\left|v^TMv - \mathbb{E}[v^TMv]\right| \geqslant t\right] \leqslant 2\exp\left(-c\min\left\{\frac{t^2}{K^4\|M\|_F^2}, \frac{t}{K^2\|M\|}\right\}\right)$$

*for a positive absolute constant c.*

For convenience, we restate Corollary 3.8.

**Corollary A.4.** *Let all assumptions of Theorem A.3 be true, and let $C^TC \in \mathbb{R}^{d\times d}$. Then, we have*

$$\Pr\left[\left|v^TC^TCv - \mathbb{E}[v^TC^TCv]\right| \geqslant \epsilon\|C\|_F^2\right] \leqslant 2\exp\left(-c\frac{\|C\|_F^2}{\|C\|^2}\min\left\{\frac{\epsilon^2}{K^4}, \frac{\epsilon}{K^2}\right\}\right).$$

*Proof.* Setting $M := C^TC$ and $t = \epsilon\|C\|_F^2$ in the Hanson-Wright inequality, we obtain

$$\begin{aligned}&\Pr\left[\left|v^TC^TCv - \mathbb{E}[v^TC^TCv]\right| \geqslant \epsilon\|C\|_F^2\right]\\ &\qquad\leqslant 2\exp\left(-c\min\left\{\frac{\epsilon^2\|C\|_F^4}{K^4\|C^TC\|_F^2}, \frac{\epsilon\|C\|_F^2}{K^2\|C^TC\|}\right\}\right).\end{aligned}$$

Thus, using that

$$\begin{aligned}\|C^TC\|_F^2 &\leqslant \|C^T\|^2\|C\|_F^2 = \|C\|^2\|C\|_F^2\\ \|C^TC\| &\leqslant \|C\|^2,\end{aligned}$$

we obtain the claimed inequality. □

A.3. **Vectorization.**

**Lemma A.5.** *Let $M \in \mathbb{R}^{a\times b}$ and $R \in \mathbb{R}^{b\times c}$ be matrices and $w \in \mathbb{R}^c$ a vector. Then*

(1) *Identifying the matrix $R \in \mathbb{R}^{b\times c}$ with the tensor $\hat{R} \in \mathbb{R}^b \otimes \mathbb{R}^c$, we have*

$$MRw = (M \otimes w^T)\hat{R}. \tag{A.1}$$

(2) *If in addition $R$ is a random matrix with i.i.d entries and*

$$\mathbb{E}[r_{ij}] = 0, \quad \mathbb{E}[r_{ij}^2] = V$$

*for some $V \geqslant 0$, we have*

$$\mathbb{E}\left[\|MRw\|^2\right] = V\|M\|_F^2\|w\|^2.$$

*Proof.* We first identify the matrix $R \in \mathbb{R}^{b\times c}$ with the tensor product $\hat{R} \in \mathbb{R}^b \otimes \mathbb{R}^c$ via a linear extension of $rs^T \to r \otimes s$. Then, we have

$$(M \otimes w^T)(r \otimes s) = Mr \otimes \underbrace{w^T s}_{\in\mathbb{R}} = (Mr)w^T s = M(rs^T)w,$$

where in the second equality, we have identified $\mathbb{R}^n \otimes \mathbb{R}$ with $\mathbb{R}^n$. By linear extension, we thus have (A.1).

In order to calculate $\mathbb{E}\left[\|MRw\|^2\right]$ note that $\mathbb{E}[\hat{R}\hat{R}^T]_{ij,kl} = \mathbb{E}[r_{ij}r_{kl}] = V\delta_{ik}\delta_{jl}$ so that

$$\mathbb{E}[\hat{R}\hat{R}^T] = V\,Id,$$

with identity matrix $Id \in \mathbb{R}^b \otimes \mathbb{R}^b$. It follows that

$$\begin{aligned}
\mathbb{E}\left[\|MRw\|^2\right] &= \mathbb{E}\left[\|(M \otimes w^T)\hat{R}\|^2\right] = \mathbb{E}\left[\hat{R}^T(M^T \otimes w)(M \otimes w^T)\hat{R}\right] \\
&= \mathbb{E}\left[\operatorname{tr}\left(\hat{R}^T(M^TM \otimes w^Tw)\hat{R}\right)\right] = \mathbb{E}\left[\operatorname{tr}\left(\hat{R}\hat{R}^T(M^TM \otimes w^Tw)\right)\right] \\
&= \operatorname{tr}\left(\mathbb{E}\left[\hat{R}\hat{R}^T\right](M^TM \otimes w^Tw)\right) = V\operatorname{tr}\left(M^TM \otimes w^Tw\right) \\
&= V\|M\|_F^2\|w\|^2.
\end{aligned}$$

□

A.4. **Matrix Norms.** A block matrix $C = \begin{pmatrix} C_1 & \cdots & C_\theta \end{pmatrix}$ has spectral norm

$$\|C\|^2 \leqslant \sum_{l=1}^{\theta} \|C_l\|^2. \tag{A.2}$$

Indeed for any block vector $v = (v_1 \cdots v_\theta)$ we have

$$\|Cv\| = \left\|\sum_{l=1}^{\theta} C_l v_l\right\| \leqslant \sum_{l=1}^{\theta}\|C_l v_l\| \leqslant \sum_{l=1}^{\theta}\|C_l\|\|v_l\| \leqslant \left(\sum_{l=1}^{\theta}\|C_l\|^2\right)^{1/2}\left(\sum_{l=1}^{\theta}\|v_l\|^2\right)^{1/2}.$$

A.5. **Weighted Compressed sensing.** This section provides optimality criteria for weighted compressed sensing, analogous to [23, Theorems 4.26, 4.30, Corollary 4.28] for the unweighted case.

**Lemma A.6.** *Let $A \in \mathbb{R}^{d\times D}$, $x \in \mathbb{R}^D$, $y \in \mathbb{R}^D$ and $W \in \mathbb{R}^{D\times D}$ be a diagonal weight matrix with non-negative diagonal entries. Let $S$ be the support of $x$ and $\bar{S}$ its complement. If $A_{\cdot,S}$ is injective, $Ax = y$ and*

$$|\left\langle A_{\cdot,j}, (A^*_{\cdot,S})^+ W_{\cdot,S} \odot \operatorname{sign}(x_S)\right\rangle| < W_{jj}, \qquad j \in \bar{S}, \tag{A.3}$$

*then $x$ is the unique minimizer of the weighted compressed sensing problem*

$$\min_x \|Wx\|_1 = \sum_{i=1}^D w_i|x_i|, \qquad Ax = y. \tag{A.4}$$

*Proof.* Define $h = (A^*_{\cdot,S})^+ W_{\cdot,S} \odot \operatorname{sign}(x_S)$. Since $A^*_{\cdot,S}$ is surjective, we have $A^*_{\cdot,S}h = W_{\cdot,S}\operatorname{sign}(x_S)$ and (A.3) yields $|\langle A_{\cdot,j}, h\rangle| < W_{jj}$ for $j \in \bar{S}$, so that in summary we have

$$\begin{aligned} h^* A_{\cdot,j} &= W_{jj}\operatorname{sign}(x_j), && j \in S \\ h^* A_{\cdot,j} &\in (-W_{jj}, W_{jj}), && j \in \bar{S}, \end{aligned}$$

which are the KKT conditions for Lagrangian $L = \|Wx\|_1 - h^*(Ax - y)$ of the optimization problem (A.4) with Lagrange multiplier $h$. Since the optimization problem is convex, the KKT conditions are sufficient, and $x$ is a minimizer, see e.g. [38, Theorem 3.1.27].

To show uniqueness, we consider an elementary proof of this statement. Let $g_{\bar{S}} = A^*_{\cdot,\bar{S}}h$. Then for any $z \in \mathbb{R}^D$ satisfying the constraint $Az = y = Ax$ and using the KKT conditions, we have

$$0 = \langle h, A(z-x)\rangle = \langle W_{\cdot,S}\operatorname{sign}(x_S), z_S - x_S\rangle + \langle g_{\bar{S}}, z_{\bar{S}} - x_{\bar{S}}\rangle .$$

The first term can be estimated by

$$\begin{aligned} \langle W_{\cdot,S}\operatorname{sign}(x_S), z_S - x_S\rangle &= \sum_{j\in S} W_{jj}[\operatorname{sign}(x_j)z_j - \operatorname{sign}(x_j)x_j] \\ &\leqslant \sum_{j\in S} W_{jj}[|z_j| - |x_j|] = \|Wz_S\|_1 - \|Wx_S\|_1 \end{aligned}$$

and using $x_{\bar{S}} = 0$ and the KKT condition, the second by

$$\langle g_{\bar{S}}, z_{\bar{S}} - x_{\bar{S}}\rangle = \langle g_{\bar{S}}, z_{\bar{S}}\rangle < \|Wz_{\bar{S}}\|_1 = \|Wz_{\bar{S}}\|_1 - \|Wx_{\bar{S}}\|_1,$$

with a strict inequality for $z_{\bar{S}} \neq 0$. In conclusion, we have

$$0 = \langle h, A(z-x)\rangle \leqslant \|Wz\|_1 - \|Wx\|_1$$

with with a strict inequality if $z_{\bar{S}} \neq 0$. This shows that $x$ is a minimizer. In case $z_{\bar{S}} = 0$, we have $y = Az = A_S z_S + A_{\bar{S}} z_{\bar{S}} = A_S z_S$ and because $A_S$ is injective $z_S = x_S$. This implies that $z = x$, which shows that $x$ is indeed the unique minimizer. □

A.6. $NP$**-hardness.** It is well known that the $\ell_p$-minimization problem (3.1) is $NP$-hard in general. For the results of the paper, we consider extra conditions on the sensing matrix $A$ and some constraints on the solution vector $x$. In this section, we show that these conditions do not generally render the problem tractable.

We consider the following three problems. The first two are known to be $NP$-hard and reduced to the compressed sensing problem with additional constraints used in this paper.

(1) *Exact cover by* 3*-set (*$X3C_{m,\theta}$*):* Given a collection $C^l$, $i = 1, \dots, \theta$ of three element subsets of $\{1, \dots, m\}$ does there exits a sub-collection that is a cover of $\{1, \dots, m\}$? I.e. we want to find indices $J \subset \{1, \dots, \theta\}$ such that $\bigcup_{j \in J} C^l = \{1, \dots, m\}$ and $C^l \cap C^k = \varnothing$ for all $l, k \in J$ with $l \neq k$.
(2) *Partition Problem (*$PP_m$*):* Given: integer or rational numbers $a_1, \dots, a_m$, can one partition $\{1, \dots, m\}$ into two sets $S_1$ and $S_2$ such that $\sum_{i \in S_1} a_i = \sum_{i \in S_2} a_i$?
(3) $\ell_p$*-minimization (*$LP^p_{m,N}$*):* For $0 \leqslant p \leqslant 1$, given a sensing matrix $A \in \mathbb{R}^{m \times N}$ and measurements $y \in \mathbb{R}^m$, find the minimizer
$$\min_{x \in \mathbb{R}^N} \|x\|_p^p, \quad \text{s.t.} \quad Ax = y.$$

For the following discussion, we assume the usual block structure
$$A = \begin{bmatrix} A^1 & \cdots A^\theta \end{bmatrix}, \qquad\qquad A^l \in \mathbb{R}^{m \times n}.$$
with $N = n\theta$.

We first consider the assumptions in the main result Theorem 3.3 on the sensing matrix $A$ or their simplified variants in Section 3.2. Since the theorem states a sparse recovery result instead of directly addressing the $\ell_p$-minimization (3.1), we consider reductions from the covering problem to $\ell_0$-minimization. For general matrix $A$, the covering problem $X3C_{m,N}$ is polynomial-time reducible to $LP^0_{m,N}$. With the given restrictions on $A$ a reduction is still possible, at least for the smaller problem $X3C_{m,\theta}$. Note however that Theorem 3.3 cannot deal with any instance in the following lemma because the solution vector $x$ is contained in the probabilistic part of the statement.

**Lemma A.7.** *For* $n < m - 2$*, there is a polynomial-time reduction from* $X3C_{m,\theta}$ *to* $LP^0_{m+n-1,n\theta}$ *with blocks of size* $A^l \in \mathbb{R}^{m+n-1 \times n}$ *that satisfy*
$$|S^l| \leqslant \|A^l\|^2_{:,S^l} \leqslant 3|S^l|, \qquad\qquad 1 \leqslant \|A^l\| \leqslant \sqrt{3}$$
*for all index sets* $S^l \subset \{1, \dots, n\}$.

*Proof.* Given an instance of $X3C_{m,\theta}$, let us define the vectors $a^l \in \mathbb{R}^m$ such that $a^l_j = 1$ if $j \in C^l$ and $a^l_j = 0$ else, let $U^l \in \mathbb{R}^{n-1 \times n-1}$ be orthogonal matrices and define the sensing matrix blocks
$$A^l = \begin{bmatrix} a^l & \\ & U^l \end{bmatrix} \in \mathbb{R}^{m+n-1 \times n}$$
and measurement vector
$$y = \begin{bmatrix} \bar{y}, \hat{y} \end{bmatrix}, \qquad \bar{y} = \begin{bmatrix} 1 & \cdots & 1 \end{bmatrix}^T \in \mathbb{R}^m, \qquad \hat{y} = \begin{bmatrix} 0 & \cdots & 0 \end{bmatrix}^T \in \mathbb{R}^{n-1}$$

Since all blocks $a^l$ and $U^l$ decouple, the matrix $A$ satisfies all given requirements.

We next show that $X3C_{n,\theta}$ has a solution if and only if the sparsest solution of $Ax = y$ satisfies $\|x\|_0 = m/3$. We first split $x = [x^1, \dots, x^l]$ with $x^l = [v^l, u^l]$ according to the block structure of $A$. This leads to the two decoupled systems

$$\sum_{l=1}^{\theta} a^l v^l = \bar{y}, \qquad\qquad \sum_{l=1}^{\theta} U^l u^l = \hat{y}.$$

It directly follows that $u^l = 0$, $l = 1, \dots, \theta$. The remaining problem is identical to the original proof in [37] or in the book [23]. Since each column $a^l$ has exactly three non-zero components, we must have $\|x\|_0 \geqslant m/3$ to obtain a right hand side $\bar{y}$ with all entries one, with equality if and only if there is a cover $J$ and $v^l = 1$ if $l \in J$ and zero else. □

In this paper, we also consider the case where the solution $x$ comes from a discrete set only. Whereas replacing a continuous variable by a discrete one often makes a problem harder, if we restrict the variables too severely, it might become trivial. With discrete $x$, a reduction from $PP_m$ to $LP^p_{m+1,2m}$ is particularly simple. Unlike Theorem 3.3 this is a $\ell_p$-minimization for $p > 0$ so that a direct connection between the theorem and the following lemma can only be made if $A$ allows sparse recovery by $\ell_p$-minimization. Nonetheless, the result indicates that the discrete sets used in Section 3.2 are not overly simple.

**Lemma A.8.** *For $0 < p < 1$, there is a polynomial-time reduction from $PP_m$ to $LP^p_{m+1,2m}$ with blocks $A^l \in \mathbb{R}^{m,n}$ with*

$$|S^l| \leqslant \|A^l\|^2_{\cdot,S^l} \leqslant 2|S^l|, \qquad\qquad \sqrt{1-\frac{1}{2}} \leqslant \|A^l\| \leqslant \sqrt{1+\frac{1}{2}}$$

*for any $n$ and even $\theta$ with $n\theta = 2m$, for all index sets $S^l \subset \{1, \dots, n\}$ and solution vector $x$ restricted to $\{-1, -1/2, 0, 1/2, 1\}$.*

*Proof.* The proof is identical to [24, equation (9)], we only trace the matrix properties. Given an instance of $PP_m$, define the matrix

$$A = \begin{bmatrix} I & I \\ a^T & -a^T \end{bmatrix}, \qquad\qquad y = \begin{bmatrix} 1 & \cdots & 1 & 0 \end{bmatrix},$$

where in the following $I$ denotes the identity matrix of suitable dimensions. Since $\theta$ is even, it follows that each block $A^l$ has the form

$$A^l = \begin{bmatrix} 0 \\ I \\ 0 \\ \pm b^T \end{bmatrix}$$

for some vector $b$ that consists of suitable components of $a$. Upon possibly rescaling the last row of $A$, the blocks $A^l$ satisfy all requirements of the lemma.

Let $x$ be a $\ell_p$ minimizer with $Ax = y$. We show that the partition problem has a solution if and only if $\|x\|_p^p = m = n\theta/2$. Let us split the solution as $x = [u, v]$ with $u, v \in \mathbb{R}^m$ according to the block structure of $A$. For each component we have $u_i + v_i = 1$ and therefore $|u_i|^p + |v_i|^p \geqslant 1$ with equality if and only if $u_i = 0$ or

$v_i = 0$. Hence we have $\|x\|_p^p \geqslant m$ with equality if and only if $u_i = 0$ and $v_i = 1$ or $u_i = 1$ and $v_i = 0$ for all $i$, which directly implies the equivalence to the partition problem.

The restriction of $x$ to the given discrete set does not change the argument. Note that the equation $Ax = y$ always has at least the solution $x_i = 1/2$ for all $i$. □

## References

[1] Z. Allen-Zhu, Y. Li and Z. Song, *A convergence theory for deep learning via over-parameterization*, in: Proceedings of the 36th International Conference on Machine Learning, K. Chaudhuri and R. Salakhutdinov, eds., vol. 97 of Proceedings of Machine Learning Research, Long Beach, California, USA, 09–15 Jun 2019, PMLR, pp. 242–252.

[2] S. Arora, S. Du, W. Hu, Z. Li and R. Wang, *Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks*, in: Proceedings of the 36th International Conference on Machine Learning, K. Chaudhuri and R. Salakhutdinov, eds., vol. 97 of Proceedings of Machine Learning Research, Long Beach, California, USA, 09–15 Jun 2019, PMLR, pp. 322–332.

[3] S. Arora, S. S. Du, W. Hu, Z. Li, R. R. Salakhutdinov and R. Wang, *On exact computation with an infinitely wide neural net*, in: Advances in Neural Information Processing Systems, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett (eds.), vol. 32, Curran Associates, Inc., 2019.

[4] Y. Bai and J. D. Lee, *Beyond linearization: On quadratic and higher-order approximation of wide neural networks*, in: International Conference on Learning Representations, 2020.

[5] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, *A simple proof of the restricted isometry property for random matrices*, Constructive Approximation **28** (2008), 253–263.

[6] A. Blum and R. L. Rivest, *Training a 3-node neural network is np-complete*, in: Advances in Neural Information Processing Systems 1, D. S. Touretzky, ed., Morgan-Kaufmann, 1989, pp. 494–501.

[7] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, New York, NY, USA, 2004.

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

[9] E. J. Candès, J. K. Romberg and T. Tao, *Stable signal recovery from incomplete and inaccurate measurements*, Communications on Pure and Applied Mathematics **59** (2006).

[10] E. J. Candès, M. B. Wakin and S. P. Boyd, *Enhancing sparsity by reweighted $\ell_1$ minimization*, Journal of Fourier Analysis and Applications **14** (2008), 877–905.

[11] R. Chartrand and V. Staneva, *Restricted isometry properties and nonconvex compressive sensing*, Inverse Problems **24** (2008): 035020.

[12] R. Chartrand and Wotao Yin, *Iteratively reweighted algorithms for compressive sensing*, in: 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, March 2008, pp. 3869–3872.

[13] L. Chizat and F. Bach, *On the global convergence of gradient descent for over-parameterized models using optimal transport*, in: Advances in Neural Information Processing Systems, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), vol. 31, Curran Associates, Inc., 2018.

[14] L. Chizat, E. Oyallon and F. Bach, *On lazy training in differentiable programming*, in: Advances in Neural Information Processing Systems, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett (eds.), vol. 32, Curran Associates, Inc., 2019.

[15] M. Conforti, G. Cornuéols, and G. Zambelli, *Integer Programming*, Springer, 2014.

[16] I. Daubechies, R. DeVore, M. Fornasier, and C. S. Güntürk, *Iteratively reweighted least squares minimization for sparse recovery*, Communications on Pure and Applied Mathematics **63** (2010), 1–38.

[17] W. Deng, M.-J. Lai, Z. Peng and W. Yin, *Parallel Multi-Block ADMM with o(1 / k) Convergence*, Journal of Scientific Computing **71** (2017), 712–736.

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

[19] S. Du and J. Lee, *On the power of over-parametrization in neural networks with quadratic activation*, in: Proceedings of the 35th International Conference on Machine Learning, J. Dy and A. Krause (eds.), vol. 80 of Proceedings of Machine Learning Research, PMLR, 10–15 Jul 2018, pp. 1329–1338.

[20] S. Du, J. Lee, H. Li, L. Wang and X. Zhai, *Gradient descent finds global minima of deep neural networks*, in: Proceedings of the 36th International Conference on Machine Learning, K. Chaudhuri and R. Salakhutdinov (eds.), vol. 97 of Proceedings of Machine Learning Research, Long Beach, California, USA, 09–15 Jun 2019, PMLR, pp. 1675–1685.

[21] S. S. Du, X. Zhai, B. Poczos and A. Singh, *Gradient descent provably optimizes over-parameterized neural networks*, in: International Conference on Learning Representations, 2019.

[22] S. Foucart and M.-J. Lai, *Sparsest solutions of underdetermined linear systems via $\ell_q$-minimization for* $0 < q \leqslant 1$, Applied and Computational Harmonic Analysis **26** (2009), 395–407.

[23] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*, Birkhäuser, 2013.

[24] D. Ge, X. Jiang and Y. Ye, *A note on the complexity of $l_p$ minimization*, Mathematical Programming **129** (2011), 285–299.

[25] R. Ge, J. D. Lee and T. Ma, *Learning one-hidden-layer neural networks with landscape design*, in: International Conference on Learning Representations, 2018.

[26] I. Goodfellow, Y. Bengio and A. Courville, *Deep Learning*, MIT Press, 2016.

[27] I. J. Goodfellow and O. Vinyals, *Qualitatively characterizing neural network optimization problems*, in: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015.

[28] K. He, X. Zhang, S. Ren and J. Sun, *Deep residual learning for image recognition*, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016, pp. 770–778.

[29] A. Jacot, F. Gabriel and C. Hongler, *Neural tangent kernel: Convergence and generalization in neural networks*, in: Advances in Neural Information Processing Systems, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), vol. 31, Curran Associates, Inc., 2018.

[30] S. P. Kasiviswanathan and M. Rudelson, *Restricted isometry property under high correlations*, 2019. https://arxiv.org/abs/1904.05510.

[31] M.-J. Lai and Y. Liu, *The null space property for sparse recovery from multiple measurement vectors*, Applied and Computational Harmonic Analysis **30** (2011), 402–506.

[32] M.-J. Lai and Y. Wang, *Sparse Solutions of Underdetermined Linear Systems and Their Applications*, Society for Industrial and Applied Mathematics, Philadelphia, 2021.

[33] M.-J. Lai, Y. Xu and W. Yin, *Improved iteratively reweighted least squares for unconstrained smoothed $\ell_q$ minimization*, SIAM Journal on Numerical Analysis **51** (2013), 927–957.

[34] Y. Li and Y. Liang, *Learning overparameterized neural networks via stochastic gradient descent on structured data*, in: Advances in Neural Information Processing Systems 31, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), Curran Associates, Inc., 2018, pp. 8157–8166.

[35] G. G. Lorentz, M. v. Golitschek and Y. Makovoz, *Constructive Approximation: Advanced Problems*, Springer-Verlag Berlin Heidelberg, 1996.

[36] S. Mei, A. Montanari and P.-M. Nguyen, *A mean field view of the landscape of two-layer neural networks*, Proceedings of the National Academy of Sciences **115** (2018), E7665–E7671.

[37] B. K. Natarajan, *Sparse approximate solutions to linear systems*, SIAM Journal on Computing **24** (1995), 227–234.

[38] Y. Nesterov, *Lectures on Convex Optimization*, Springer Publishing Company, Incorporated, 2nd ed., 2018.

[39] Q. Nguyen and M. Hein, *The loss surface of deep and wide neural networks*, in: Proceedings of the 34th International Conference on Machine Learning, D. Precup and Y. W. Teh (eds.), vol. 70 of Proceedings of Machine Learning Research, PMLR, 06–11 Aug 2017, pp. 2603–2612.

[40] S. Oymak and M. Soltanolkotabi, *Toward moderate overparameterization: Global convergence guarantees for training shallow neural networks*, IEEE Journal on Selected Areas in Information Theory **1** (2020), 84–105.

[41] H. Rauhut and R. Ward, *Interpolation via weighted $\ell_1$ minimization*, Applied and Computational Harmonic Analysis **40** (2016), 321–351.

[42] G. M. Rotskoff and E. Vanden-Eijnden, *Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error*, CoRR, abs/1805.00915 (2018).

[43] M. Rudelson and R. Vershynin, *Hanson-wright inequality and sub-gaussian concentration*, Electron. Commun. Probab. **18** (2013), p. 9 pp.

[44] I. Safran and O. Shamir, *Spurious local minima are common in two-layer ReLU neural networks*, in: Proceedings of the 35th International Conference on Machine Learning, J. Dy and A. Krause (eds.), vol. 80 of Proceedings of Machine Learning Research, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018, PMLR, pp. 4433–4441.

[45] Y. Shen and S. Li, *Restricted p-isometry property and its application for nonconvex compressive sensing*, Advances in Computational Mathematics **37** (2012), 441–452.

[46] J. Sirignano and K. Spiliopoulos, *Mean field analysis of neural networks: A law of large numbers*, SIAM Journal on Applied Mathematics **80** (2020), 725–752.

[47] M. Soltanolkotabi, A. Javanmard and J. D. Lee, *Theoretical insights into the optimization landscape of over-parameterized shallow neural networks*, IEEE Transactions on Information Theory **65** (2019), 742–769.

[48] D. Soudry and Y. Carmon, *No bad local minima: Data independent training error guarantees for multilayer neural networks*, 2016. https://arxiv.org/abs/1605.08361.

[49] Q. Sun, *Recovery of sparsest signals via $\ell_q$-minimization*, Applied and Computational Harmonic Analysis **32** (2012), 329–341.

[50] L. Venturi, A. S. Bandeira and J. Bruna, *Spurious valleys in one-hidden-layer neural network optimization landscapes*, Journal of Machine Learning Research **20** (2019), 1–34.

[51] C. Villani, *Topics in Optimal Transportation*, American Mathematical Society, 2003.

[52] G. Welper, *Non-convex compressed sensing with training data*, 2021. https://arxiv.org/abs/2101.08310.

[53] J. Woodworth and R. Chartrand, *Compressed sensing recovery via nonconvex shrinkage penalties*, Inverse Problems **32** (2016): 075004.

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

*Manuscript received December 31 2021*
*revised April 27 2022*

G. Welper
Department of Mathematics, University of Central Florida, Orlando, FL 32816, USA
*E-mail address*: `gerrit.welper@ucf.edu`