Bisparse Blind Deconvolution through Hierarchical Sparse Recovery

Axel Flinth
Department of Mathematics and Mathematical Statistics, Umeå University, Umeå, Sweden.
axel.flinth@umu.se

Ingo Roth
Quantum Centre, Technology Innovation Institute, Abu Dhabi, UAE
Dahlem Center for Complex Quantum Systems, Freie Universität Berlin, Berlin, Germany
ingo.roth@tii.ae

Gerhard Wunder
Cybersecurity and AI Group, Freie Universität Berlin, Berlin, Germany
g.wunder@fu-berlin.de

Abstract
The bi-sparse blind deconvolution problem is studied. The latter consists of recovering $h$ and $b$ from the knowledge of $h \ast (Qb)$, where $Q$ is some linear operator, and both $b$ and $h$ are both assumed to be sparse. The approach rests upon lifting the problem to a linear one, and then applying the hierarchical sparsity framework. In particular, the efficient HiHTP algorithm is proposed for performing the recovery. Then, under a random model on the matrix $Q$, it is theoretically shown that an $s$-sparse $h \in \mathbb{K}^\mu$ and $\sigma$-sparse $b \in \mathbb{K}^\nu$ with high probability can be recovered when $\mu \gtrsim s \log(s)^3 \log(\mu \log(\mu n)) + s \sigma \log(n)$.

Index Terms
Blind deconvolution, sparsity, restricted isometry property.

I. INTRODUCTION

We begin with motivating and illustrating the bi-sparse deconvolution problem by a simple communication scenario. Consider two persons, Alice and Bob, communicating with each other. Bob would like to send a message $b \in \mathbb{K}^\nu$, where $\mathbb{K}$ denotes one of the fields $\mathbb{R}$ or $\mathbb{C}$, to Alice. To do so, he first linearly encodes his message in a signal $Qb \in \mathbb{K}^\mu$, and then sends it over a wireless channel to Alice. Due to scattering in the environment, Alice will not receive $Qb$, but rather a superposition of several time- and phase-shifted versions of the signal $Qb$. Put differently, she observes the convolution

$$y = h \ast (Qb)$$

of $Qb$ with a filter $h \in \mathbb{K}^\mu$ modelling the impulse response of the environment. A schematic sketch is provided in Figure 1. The filter $h$ is highly dependent on unknown features of the environment. In traditional communication protocols, Alice determines the filter $h$ prior to the actual communication by receiving a pre-determined pilot signal from Bob through the channel. Interesting alternative approaches are presented by blind communication protocols that omit the separate channel estimation stage. In order to retrieve the message $b$, Alice here needs to solve a blind deconvolution problem, i.e. determine $b$ given $y$ and $Q$ without knowledge of $h$. Since the scattered signal typically reaches the receiver antenna along a few, distinct paths, it is reasonable to assume that the filter $h$ is sparse 1. We can furthermore render the message $b$ sparse by system design. Under these assumptions, we arrive at a bi-sparse blind deconvolution problem. In the following we denote the sparsity of $h$ and $b$ by $s$ and $\sigma$, respectively.

1This paper was presented in part at ICCHA 2022.
This work has been submitted to the IEEE for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible.
A. Related work.

The approaches to solving the sparse blind deconvolution problem can broadly be divided into two categories:

- Minimization of a function of the form
  \[ L : \mathbb{K}^\mu \times \mathbb{K}^n \rightarrow \mathbb{R}, \quad (h, b) \mapsto L(h, b) = \mathcal{L}(y, h \ast (Qb), h, b), \]
  defined in terms of a suitable loss function \( \mathcal{L} \). Let us call such methods direct.
- Using the universal property of tensor spaces, identify the bilinear map \( \gamma(h, b) = h \ast Qb \) on \( \mathbb{K}^\mu \times \mathbb{K}^n \) with a linear map \( C \) on \( \mathbb{K}^\mu \otimes \mathbb{K}^n \), and then minimize
  \[ K : \mathbb{K}^\mu \otimes \mathbb{K}^n \rightarrow \mathbb{R}, \quad w \mapsto K(w) = \mathcal{K}(y, C(w), w). \]

where again \( \mathcal{K} \) refers to some suitable loss function. Effectively, we go over from recovering the pair \((h, b)\) to recovering the tensor \( h \otimes b \). This procedure is generally referred to as lifting, and we will here refer to them as lifted methods.

Let us give a brief overview of the relevant results from the literature about methods from both categories. Throughout the discussion, it is instructive to remember that the optimal sample complexity (i.e. the number of needed observations for injectivity) is given by \( \mu + n \) in the dense setting and \( s + \sigma \) in the sparse setting \cite{2, 3}.

\text{a) Direct approaches: Among direct approaches, alternating minimization is probably the most prominent. It was introduced in \cite{4} as a means for solving the so-called phase-retrieval problem and later adapted to blind deconvolution in \cite{5, 6}. Alternating minimization generally refers to taking turns in solving the following two problems,}

\[ \begin{align*}
\text{minimize } & L(h, b') \quad \text{and} \quad \text{minimize } L(h', b).
\end{align*} \]

Since the convolution is linear in each argument, each subproblem is effectively a classical compressed sensing problem, and can be solved using a number of techniques, e.g. iterative hard thresholding \cite{7} or CoSAMP \cite{8}.

As for the question of success of alternating minimization, the authors of \cite{5, 6} give a recovery guarantee in a special setting: First, they use a generic subspace model, where \( h \) has a representation \( h = Bg \) with some \( s \)-sparse vector \( g \), rather than being sparse itself, and \( B \) is assumed to be a matrix with random Gaussian entries. The measurement matrix \( Q \) is also assumed to be Gaussian. An additional important assumption is that \( h \) and \( x = Qb \) are spectrally flat, i.e. have absolutely approximately constant Fourier transforms. The spectral flatness assumption is not only a regularity assumption on the signals. It is actively exploited in a projection step of the authors’ algorithm. The projection step is hard to perform exactly. The authors resort to heuristics for this step. However, accepting this caveat, the authors prove convergence already when only observing \((s + \sigma) \log(\mu)^5\) of the entries in \( h \ast x \), which is up to log-terms sample optimal.
In [9], Li, Ling, Strohmer and Wei proposed a different non-convex approach for the related problem where $h$ and $b$ are in known subspaces. Concretely, they assume that $h$ is the Fourier transform of a vector $g \in \mathbb{C}^\nu$ supported on its first $s$ entries (rather than being $s$-sparse), and $b = Bc$ for some $c \in \mathbb{C}^\sigma$ and $B$ a Gaussian. Needless to say, this variation of the problem amounts to a significant simplification. For a non-convex, smooth, loss function including regularization terms, together with a careful initialization, they prove global convergence. They also assume a spectral flatness condition. This assumption is again built into the method, now implicitly in that it governs the size of a constant in the loss function. They prove global convergence already when $\mu \gtrsim \max(s, \sigma) \log(\mu)^2$. Note that in our setting, where the supports of the vectors are unknown, this condition would read $\mu \gtrsim \max(\mu, n) \log(\mu)^2$, which would be vacuous.

b) Lifted approaches: As for lifted approaches, it is crucial to understand that the lifted ground truth signal $h_0 \otimes b_0$ or equivalently its matrix form $h_0 b_0^*$, where $b_0^*$ denotes the Hermitian transpose of $b_0$, simultaneously enjoys two structures: it has rank one and is sparse (in a structured manner). Each of these structures can individually be exploited using convex regularizers – if we take the low-rank property as the crucial one, the nuclear norm is the canonical regularizer [10], [11], [12], whereas if the sparsity is taken as the crucial property, the nuclear norm is the canonical regularizer [10], [11], [12]. However, not assuming know supports, that guarantee degenerates to $\mu \gtrsim \max(\mu, n) \log(\mu)^2$. The sparse models ($\ell_1$ and $\ell_{1,2}$) recover the ground truth with high probability when $\mu \gtrsim s \sigma \log(sn) \log(\mu)^2$ under the assumption that the support of $h$ is known [13], [14]. If the support of $h$ is not known, the guarantee again degenerates to $\mu \gtrsim \mu \sigma \log(n) \log(\mu)^2$.

There also is a natural non-convex-way to solve the lifted problem: a gradient descent projected onto the set of (bi)-sparse and low-rank matrices. As is thoroughly discussed in [16], there is however no practical algorithm available to compute this projection. The bi-sparse unit-rank projection is known as the sparse PCA problem, that has been shown to be worst-case NP-hard, as well as under approximation and on average [17], [18], [19]. A canonical way to circumvent this obstacle is to alternate between projections onto the two sets. This approach is for instance investigated in [20]. There, a local convergence guarantee is presented under optimal sample complexity using the considerably simpler measurement model with a Gaussian linear map without the structure of the convolution. Similar results are given in [21] for an alternating minimization approach, this guarantee is however only sample optimal under an additional assumption on the signal, and also under a simpler measurement model.

In this context, [22] should also be mentioned – the authors obtain global convergence in just two alternations steps, however by assuming a nested measurement structure tailor-made for a jointly low-rank and sparse setting, which is often not given.

Finally, in [23], a different but related problem was considered. In the paper, they consider the case of several convolutions $y_i = h \ast (Q_i b_i), i \in [N]$ are independently observed. Importantly, the authors do not assume that the support of $h$ is known, only that it is sparse in a basis $B$ with a certain incoherence. However, they instead assume that the supports of the $b_i$ are known, and that the $Q_i$ are independently Gaussian distributed. Under those assumptions, they achieve recovery (through a nuclear norm minimization approach) of both $h$ and all $b_i$ already when $\mu \gtrsim (\sigma + s \log(s))^2 \log(\mu N)^2$ and $N \gtrsim \log(N \mu)$. If we drop the assumption of known supports of the $b_i$, the guarantee again degenerates, to $\mu \gtrsim (\mu + s \log(s)) \log(\mu s)^2$.

The findings of this literature review can be summarized as follows.

- **Direct methods.** These operate at optimal sample complexity, however only under complicated additional assumptions. Furthermore, they typically rest upon a good initialisation.
- **Lifted methods.** There are convex and non-convex approaches to solve the lifted problems. The convex ones converge globally and are simple to implement, but are computationally costly and cannot recover the signals at optimal sample complexity. Their convergence can only be guaranteed when either the filter or the signal a-priori has known support. The non-convex ones are fast and converge quickly, but only work sample-optimally when computationally intractable projection routines are assumed. When heuristics and approximations of the projection steps are applied instead, only local convergence results are known for practical measurement models.
B. Contribution

In this work, we propose a different, non-convex, approach to the problem, namely to solve the lifted problem as a hierarchically sparse recovery problem. Notice that if both $h$ and $b$ are sparse, the lifted-ground truth signal in its matrix form $w = h_0b^*$ only has $s$ non-zero rows, and each row is a $\sigma$-sparse vector itself. This is an example of an $(s, \sigma)$-hierarchically sparse signal \cite{24, 25, 26, 27, 28}. The signal $w$ has in fact even more structure such as a joint row-support due to the low-rank. At the core of our approach, however, is to only explicitly exploit the hierarchical sparse structure. In \cite{29}, a subset of the authors showed that such signals can be effectively reconstructed from linear measurements using a hard-thresholding algorithm. This type of algorithmic modifications can be regarded as a special case of model based compressed sensing \cite{30}. The algorithm in each step estimates the sparse support from a gradient step to minimize the residual via projection onto the set of hierarchically sparse vectors, and subsequently solves the least squares problem restricted to the support estimate. The resulting algorithm is called HiHTP, Hierarchical Hard Thresholding Pursuit – the hierarchical variant of the HTP of Ref. \cite{7}. A formal definition of the algorithm is given later in the paper (Algorithm 1).

Crucially, the projection step of HiHTP is computationally efficient and can be performed in time-complexity linear in the vector-space dimension $\mu n$ of the lifted signal. The algorithm can further be proven to recover the ground truth in a stable and robust fashion under a suitable restricted isometry condition (RIP). We refer to Ref. \cite{28} for a detailed overview of the framework.

Using HiHTP for bisparse blind deconvolution was proposed and numerically demonstrated to work by a subset of the authors in \cite{31}. In the present paper, we provide a theoretical recovery guarantee for its success. The main result can be informally stated as follows:

**Main Result.** Suppose $y = h \ast (Qb) \in \mathbb{K}^\mu$ with $b \in \mathbb{K}^\sigma$-sparse, $h \in \mathbb{K}^\mu$ s-sparse and Gaussian $Q$. If

$$\mu \gtrsim s \log(s)^2 \log(\mu) \log(\mu n) + s \sigma \log(n),$$

then the HiHTP-algorithm will with high probability recovery the ground truth signal $h \otimes b$ from $y$ and $Q$.

Disregarding logarithmic terms, our bound is of the order $s \sigma$ in the sparsity parameters $s$ and $\sigma$. This is of course not the sample optimal number of $s + \sigma$ measurements. However, it is the optimal sample complexity of the recovery of hierarchically sparse signals, the relaxation of the original problem structure. To explain what the latter means, note that the HiHTP-algorithm would work equally well when the measurements originated from a hierarchically sparse signal $(w_1, \ldots, w_n)$ where the blocks are not equal. Counting the degrees of freedom of the hierarchical sparse vector, thus, suggest an optimal sample complexity of $s \cdot \sigma$.

Compared to the existing guarantees for direct approaches, we do not require a spectral-flatness condition, a generic dictionary for $h$, a heuristic projection step or known subspaces. Thus, although the sample complexity of our proof is not optimal, it solves either a ‘harder’ problem or works under more general conditions than the methods available in the literature. Compared to the existing lifted approaches, our guarantee has an improved sampling complexity in the bi-sparse setting with unknown supports.

**Remark 1.** Compared to previous complexity bounds for HiHTP \cite{29}, the complexity bound differs by the log-terms in $\mu$ and $s$. We believe that these are artefacts of the proof. In Section IV numerically find that the ratio $\mu/(s \log(\mu n) + s \sigma \log(n))$ is a better predictor for the success of the HiHTP algorithm than the ratio that our main result suggests.

A feature of the hierarchical approach to sparse deconvolution problems is that it readily generalizes to more complex settings, e.g. arising in multi-user communications. We here also discuss how to solve a combined deconvolution and demixing problem using the hierarchically sparse framework: Given $M$ measurements of the form

$$y_q = \sum_{p \in [N]} d_{q,p}h_p \ast (Q_pb^p),$$

where $d_{q,p} \in \mathbb{K}$, $q \in [M], p \in [N]$ are ‘mixing factors’, recover the collection of $N$ filter-signal pairs $(h^p, b^p), p \in [N]$. This problem, e.g., appears when users are separated both through their spatial angles and
their multipath delay as in [32]. Under the assumption that only $S$ filters messages are non-zero, we can interpret this again as a hierarchical sparse recovery problem (albeit now in three levels of hierarchy). Using recent results from [33], we show that if each matrix $Q_p$ is constructed as above, HiHTP can recover all signals and all filters from $M \gtrsim S \log(N)$ mixtures.

C. Outline of the paper

In Section II, we give a formal introduction to our approach and our assumptions, and also formally state the main result, Theorem II.1. We also briefly discuss the multi-user setup there. Section III is in its entirety devoted to proving the main result. In Section IV, we perform a small numerical experiment to validate our theory.

a) Notation: Most of the notation is either standard or will be introduced at first use. For a vector $v \in \mathbb{F}^n$, we write $v(i)$ for its $i$:th entry. We use $\| \cdot \|$ to denote the Euclidean norm for both vectors and tensors, $\| \cdot \|_F$ for the Frobenius norm of a matrix, and $\| \cdot \|_{2 \to 2}$ for the induced $\ell_2 \to \ell_2$-operator norm of a matrix. Finally, we denote by $\gtrsim$ ordering up-to a constant factor that is independent on the stated variables.

II. BISPARE BLIND DECONVOLUTION WITH HiHTP

Let us first agree on some notation. As outlined above, we are interested in the blind deconvolution program, i.e. recovering $h \in \mathbb{K}^\mu$ and $x \in \mathbb{K}^\mu$ from the convolution $h \ast x$. We thereby understand the convolution as circular, i.e.

$$(h \ast x)(\ell) = \sum_{k \in \mathbb{Z}} h(\ell - k)x(k),$$

Here, $[\mu]$ is the set of remainder classes modulo $\mu$, i.e. the set of non-negative integers from 0 to $(\mu - 1)$ equipped with the addition modulo $\mu$. In the following also all indices are understood using this cyclic identification.

Let us state the formal assumptions on the filter and message vectors $h$ and $x$.

**Assumption 1.** We assume that
- The filter $h \in \mathbb{K}^\mu$ is $s$-sparse.
- The signal $x \in \mathbb{K}^\mu$ has a $\sigma$-sparse representation in a dictionary $Q \in \mathbb{K}^{\mu,n}$. That is, there exists a $\sigma$-sparse vector $b \in \mathbb{K}^n$ with $x = Qb$.

The map $(h, b) \mapsto h \ast (Qb)$ is bilinear, and can therefore be lifted to a linear map $C : \mathbb{K}^n \otimes \mathbb{K}^\mu \cong (\mathbb{K}^n)^\mu \rightarrow \mathbb{K}^\mu$ by linearly extending

$$h \otimes b \mapsto C(h \otimes b) = h \ast (Qb).$$

This means that we can interpret the blind deconvolution problem as a linear reconstruction problem for the lifted vector $h \otimes b \in \mathbb{R}^n \otimes \mathbb{R}^\mu$. The idea of this paper is to utilize that the sparsity assumptions on $h$ and $b$ implies that the lifted vector has a particular structure: Among others, it is hierarchically sparse [29]. Let $\{e_i\}_{k \in [\mu]}$ denote the standard basis for $\mathbb{K}^\mu$ with entries $e_i(j) = 1$ for $i = j$ and $e_i(j) = 0$ otherwise.

**Definition 1.** Let $\mu, n, s, \sigma \in \mathbb{N}$. A tensor $w = \sum_{k \in [\mu]} e_k \otimes w_k \in \mathbb{K}^n \otimes \mathbb{K}^\mu$

is $(s, \sigma)$-sparse if
- At most $s$ of the blocks $w_i \in \mathbb{K}^n$ are non-zero.
- Each block $w_i$ is $\sigma$-sparse.

We will also simply refer to $w$ as hierarchically sparse.
Algorithm 1 (HiHTP)

Require: Measurement operator \( A \), measurement vector \( y \), block column sparsity \((s, \sigma)\)

1: \( w_0 = 0 \)
2: repeat
3: Calculate the support \( \Omega_{k+1} \) of the best approximation of \( (w_k + A^*(y - A w_k)) \) in the set of \((s, \sigma)\)-sparse vectors.
4: \( w_{k+1} = \arg\min_{z \in \mathbb{R}^m} \{ \| y - A z \|, \supp(z) \subset \Omega_{k+1} \} \).
5: until stopping criterion is met at \( \tilde{k} = k \)

Ensure: \((s, \sigma)\)-sparse vector \( w_{\tilde{k}} \)

Remark 2. We can recursively extend this definition to higher hierarchies of sparsity. For instance, a vector with \( N \) blocks out of which only \( S \) are non-zero, and each non-zero block is \((s, \sigma)\)-sparse, is \((S, s, \sigma)\)-sparse. For details, see [29].

An \((s, \sigma)\)-sparse vector \( w \) can be reconstructed from linear measurements \( A(w) \) with the help of the so-called HiHTP-algorithm [29], see Algorithm (1). In essence, it is a hybrid approach combining projected gradient descent and exact linear inversion [7]. Notably, the projection step can be performed efficiently in the lifted vector space. The main recovery criterion for HiHTP relies on the hierarchical restricted isometry property, or HiRIP, a generalization of the standard RIP [34].

Definition 2. Let \( A : \mathbb{K}^n \otimes \mathbb{K}^\mu \to \mathbb{R}^m \) be a linear operator.

(i) The \( k\)-RIP constant of \( A \) is given by
\[
\delta_k(A) = \min_{w \text{ \( k \)-sparse}} \frac{\| A w \|_2^2 - \| w \|_2^2}{\| w \|_2^2}.
\]

(ii) The \((s, \sigma)\)-HiRIP constant of \( A \) is given by
\[
\delta_{(s,\sigma)}(A) = \min_{w \text{ \((s,\sigma)\)-sparse}} \frac{\| A w \|_2^2 - \| w \|_2^2}{\| w \|_2^2}.
\]

The main recovery guarantee in [29] informally states that if an operator has the HiRIP, HiHTP can stably and robustly recover any hierarchically sparse vector. Concretely, \( \delta_{3s,2\sigma}(A) < \frac{1}{\sqrt{3}} \) ensures that any \((s, \sigma)\) sparse vector will be recovered from the measurements \( A(w) \). Under the same conditions, the algorithm handles small measurement errors and model mismatches (i.e., that the ground truth is not exactly sparse) in a graceful manner [29].

Let us now state assumptions on the linear map \( Q \). These will make it possible to prove that the lifted convolution operator \( C \) has the \((s, \sigma)\)-HiRIP.

Assumption 2. The matrix \( Q \in \mathbb{K}^{\mu, n} \) can be decomposed as follows
\[
Q = UA
\]
where

(i) \( A \in \mathbb{K}^{m, n} \) is a matrix with \( \sigma \)-RIP constant \( \delta_\sigma(A) < 1 \).
(ii) \( U \in \mathbb{K}^{\mu, m} \) is a matrix with columns \((\gamma_j)_{j \in [m]}\), whose entries \( \gamma^{-1/2} \gamma_j(i) \) are
  • centered, i.e. \( \mathbb{E}(\gamma_j(i)) = 0 \).
  • independent (over \( i \) and \( j \)).
  • normalized, i.e. \( \mathbb{E}(\gamma_j(i)^2) = 1 \).
  • sub-Gaussian variables, i.e. that there exists a number \( R \) so that \( \mathbb{E}\left(\exp\left(\frac{|\gamma_j(i)|^2}{R^2}\right)\right) \leq 2 \).
Remark 3.
1. A more common definition of subgaussianity of a random variable $X$ is a tail estimate
   $$\Pr(|X| > t) \geq 2 \exp(-\frac{t^2}{2\sigma^2}).$$
   In fact, this notion is equivalent to the one we use, with $R \sim \tilde{R}$.
2. The infimum of all $R$ for which $\mathbb{E}\left(\exp(|\gamma|^2)| R\right) \leq 2$ is the subgaussian norm of $\gamma$. This is really a norm, so that e.g. $\|\lambda\gamma\|_{\psi_2} = |\lambda| \|\gamma\|_{\psi_2}$ for $\lambda \in \mathbb{K}$.
3. Gaussian variables (in $\mathbb{C}$ and $\mathbb{R}$) are subgaussian, with $\|X\|_{\psi_2} \sim \mathbb{E}\left(|X|^2\right)$. Another example of subgaussian variables are bounded variables - if $|X| \leq \theta$ almost surely, $\|X\|_{\psi_2} \leq \theta^2$.

To understand the intuition behind this construction in our setting, let us rewrite
$$C(h \otimes b) = \sum_{k \in [\mu]} h_k e_k * (Qb) = \sum_{k \in [\mu]} e_k * (UA(h_k b)).$$

Applying $C$ amounts to first applying the operator $Q$ to the vector $b$. Then, we mix the shifts of the resulting signal $e_k * (Qb), k \in [\mu]$ using the scalars $h_k$ as weights. Our standard RIP assumption on $A$ ensures that, if we had access to the measurement $A(h_k b) \in \mathbb{R}^m$, $k \in [\mu]$ (before applying $U$), we could recover $h_k b, k \in [\mu]$ (and thus both $h$ and $b$) using standard compressed sensing methods. We can however only access them in observing their mixture. Being able to recover the different summands of this mixture requires sufficient ‘incoherence’ of the shifts. We achieve this with the operator $U$: it embeds the vectors $A(h_k b)$ into a higher-dimensional space rendering the shifts incoherent.

From the point of view of applications in wireless communication, we can think of $Q$ as applying first a compressive encoding of a sparse message $b \in \mathbb{R}^n$ giving rise to the ‘code word’ $Ab$. The code word is subsequently again encoded using $U$ for transmission via the channel $h$ into a sequence $UAb$.

Remark 4. Note that one can as well ignore the possibility of compression before the transport encoding. Choosing $m = n$ and $A = \text{id}$, $A$ trivially has $\delta_\sigma(A) = 0$ for all $\sigma$. Using a non-trivial matrix $A$, however, typically allows one to considerably reduce the number of parameters needed to describe the measurement map. The product $UA$ is specified using only $\mu \cdot n + m \cdot n$ parameters, whereas without compression, $Q \in \mathbb{K}^{\mu \cdot n}$ has $\mu \cdot n$ degree of freedoms. Since typically $n \gg \sigma$ and the $\sigma$-RIP of $A$ can be guaranteed already when $m \gtrsim \sigma \log(n)$, the former number can be considerably smaller than the latter for small values of $\sigma$. Thus, our low-rank decomposition of $Q$ constitutes a significant de-randomization of the measurement ensembles without considerably complicating the proof. With other more structured standard compressed sensing matrices for $A$ such as sub-sampled Fourier matrices, the complexity of specifying $Q$ and the computational complexity of the algorithm can be further reduced. Finally, it will allow us to slightly improve the sample complexity bound: A term $\log(\mu n)$ can be turned into $\log(\mu n)$.

We are now in a position to formulate the main result.

Theorem II.1. Fix $\delta_0 \in (0,1)$ and let $\epsilon > 0$. Suppose $Q = UA$ fulfills the Assumption 2 and it holds that
$$\mu \gtrsim (s \log(s)^2 \log(\mu n) \log(\mu) + s \sigma \log(n)) \cdot \delta_0^{-2} \cdot \max(1, \log(\epsilon^{-1})).$$

Then, the map $C$ defined in (2) has
$$\delta_{(s,\sigma)}(C) \leq (1 + \delta_\sigma(A))^2 \delta_0 + \delta_\sigma(A)$$
with a probability at least $1 - \epsilon$.

We refer to the introduction for a discussion about the theorem’s implications in terms of sample complexity. Also note that in the case $A = \text{id}, m = n$, the bound on the number of measurements grows as $s \log(s)^2 \log(\mu) \log(\mu n) + s \sigma \log(n)$.
We postpone the technical proof to the next section. Let us instead conclude this section by extending our approach to a combined deconvolution and demixing problem.

A. Deconvolution and demixing.

An attractive feature of the hierarchically sparse approach to the sparse blind deconvolution problem is its generalizability to more complicated problems, where the measurements arise from (potentially multiple) convolutions and further mixing (weighted superpositioning) of the signal. To illustrate this point, we consider the combined blind sparse deconvolution and demixing problem: Recover $N$ filter-signal pairs $(h^p, b^p) \in \mathbb{K}^\mu \times \mathbb{K}^\nu$ from $M$ weighted mixtures of the convolutions $h^p * (Q b^p)$,

$$y_q = \sum_{p \in [N]} d_{q,p} h^p * (Q b^p), \quad q \in [M],$$

with mixing weights $d_{q,p} \in \mathbb{K}$.

a) Motivation: This problem is directly motivated by extending the previous communication model to a multi-user scenario: We imagine $N$ Alices simultaneously transmitting messages $b^p, \; p \in [N]$ over channels with impulse responses $h^p$. They also modulate their transmission, meaning that in timeslot $q$, Alice $p$ sends a message $d_{q,p} b^p$ rather than $b^p$. If the Alices do so over $M$ slots, Bob will receive a signal of exactly the form (5).

Another interesting, and a slightly less direct, route to motivate the model is a MIMO (Multiple-Input-Multiple-Output) scenario. In this setting, Bob has $M$ antennas arranged in an array. The collective response of the antennas is given by a vector $v(\ell) \in \mathbb{K}^M$. Hence, if the scattered transmitted signals $x$ from a user arrives with delays $k$ from directions $\theta_k$, the collective response of the antennas is

$$y_q(\ell) = \sum_{k \in [\mu]} h(k) x(\ell - k) d_{q,k}^p .$$

If several users simultaneously send signals $x^p$ over channels with impulse responses $h^p$, the antenna will measure a superposition of such measurements,

$$y_q(\ell) = \sum_{p \in [N]} \sum_{k \in [\mu]} h(k) x(\ell - k) d_{q,k}^p .$$

Let us assume that the angle for each transmitting user to Bob stays constant for a transmission. Then, the wavefronts of one user are arriving from the same angle, and we have

$$y_q(\ell) = \sum_{p \in [N]} \sum_{k \in [\mu]} h(k) x(\ell - k) d_{q,k}^p = \sum_{p \in [N]} d_{q,p} (h^p * x^p)(\ell) .$$

With $x^p = Q_p b^p$ and $d_{q,p} = d(\theta^p)_q$, we arrive again at the form of (5).

b) Multi-level hierarchical sparsity: The deconvolution and demixing problem can easily be formulated as a hierarchically structured recovery problem. We view the collection $(h^p \otimes b^p)_{p \in [N]} \in (\mathbb{K}^\mu \otimes \mathbb{K}^\nu)^N$ as a three-tensor

$$X = \bigotimes_{p \in [N]} e_p \otimes h^p \otimes b^p \in \mathbb{K}^{N} \otimes \mathbb{K}^\mu \otimes \mathbb{K}^\nu .$$

If we assume that each message $b^p$ is $\sigma$-sparse, each filter is $s$-sparse, and that only $S$ filter-message tensors $h^p \otimes b^p$ are non-zero, the ground truth signal $X$ is $(S, s, \sigma)$-sparse. In the above communication model, such an assumption corresponds to assuming sporadic user activity with only a subset of possible users actively sending at a given instance in time. Such sporadic traffic is well-motivated in today’s and future machine-type communications.

The blind deconvolution and demixing model can hence be treated as a hierarchical sparse recovery problem. As such, we may use the HiHTP to recover it. Defining $d^p = (d_{q,p})_{q \in [M]} \in \mathbb{K}^M$, and $C_p$ as
the lifted convolutions associated to filter-message pair $q$, the collective measurement operator takes the form $\mathcal{M} : \mathbb{K}^N \otimes \mathbb{K}^\mu \otimes \mathbb{K}^n \to \mathbb{K}^M \otimes \mathbb{K}^\mu$,
\begin{equation}
\sum_{p \in [N]} e_p \otimes w_p \mapsto \sum_{p \in [N]} d^p \otimes C_p(w_p). \tag{6}
\end{equation}

These type of operators for arbitrary operators $C_p$ having HiRIP were analyzed by the authors of the article in [33]. The results [33, Theorem 2.1] state that such a hierarchical measurement operator inherits a HiRIP from the RIP of its constituent operators and it holds that
\[ \delta_{(S,s,\sigma)}(\mathcal{M}) \leq \delta_S(D) + \sup_{p} \delta_{(s,\sigma)}(C_p) + \delta_S(D) \cdot \sup_{p} \delta_{(s,\sigma)}(C_p), \]
where $D = [d_p]_{p \in [N]} \in \mathbb{K}^{M,N}$. Hence, if $\delta_S(D)$ and each $\delta_{s,\sigma}(C_p)$ is small, the entire measurement $\mathcal{M}$ will have an $(S, s, \sigma)$-HiRIP. Let us formulate this and its immediate consequences as a corollary.

**Corollary II.2** (Blind Deconvolution and demixing guarantee). (i) Assume that $D \in \mathbb{K}^{M,N}$ has $\delta_S(D) < 1$ and $C_p$ fulfils $\sup_{p \in [N]} \delta_{(s,\sigma)}(C_p) =: \delta_C < 1$. Then, the $\mathcal{M}$ defined in (6) obeys
\[ \delta_{(S,s,\sigma)}(\mathcal{M}) \leq \delta_S(D) + \delta_C + \delta_S(D) \cdot \delta_C. \]

(ii) Assume that the $C_p$ are equal to one operator constructed according to Assumption 2 and $D \in \mathbb{K}^{M,N}$ is Gaussian. Suppose
\[ M \gtrsim S \log(N) \]
\[ \mu \gtrsim s \log(s)^2 \log(\mu m) \log(\mu) + s \sigma \log(n), \]
then with high probability, HiHTP will recover each $(S, s, \sigma)$-sparse ground truth $X$ from the measurements $\mathcal{M}(X)$.

**Remark 5.** The assumption of identical $C_p$ is not essential – it merely simplifies the formulation of the result. Using a union bound together with Theorem II.1, we can also address the case where the $C_p$ are independently chosen and establish a recovery guarantee for
\[ \mu \gtrsim (s \log(s)^2 \log(\mu m) \log(\mu) + s \sigma \log(n)) \log(N). \]
The additional $\log(N)$ factor comes from applying a union bound over the $N$ operators $C_p$. Although this bound does not suggest it, letting the constituent operators $C_p$ be different, and in particular uncorrelated, may in the general hierarchical measurement case assist the recovery. We refer to [33] for a detailed discussion on this matter.

### III. Proof of Theorem II.1

The proof of Theorem II.1 proceeds in five steps. First, we rewrite the problem, partly factoring out the operator $A$. Then, we follow a route taken many times in the literature before, for instance in [36], [37] and in particular presented in the textbook [38]: We identify the HiRIP constant as the supremum a random process as in [36]. Using the results of said paper, we reduce the problem to the estimation of so-called $\gamma_{\alpha}$-functionals of a set of matrices, which in turn boils down to the estimation of covering numbers. The third step consists in providing a simple estimate of the latter, and the fourth in applying Maurey’s empirical method to derive a tighter bound for a certain regime. Both bounds are finally assembled to complete the proof and obtain the sample complexity in a fifth step. Although the proof follows well-known beats, the particular geometry of the set of $(s, \sigma)$-sparse vectors and the structure of our measurement model requires careful and non-trivial modifications of the arguments.
A. Preliminaries

We start out small and provide an explicit form of the lifted convolution operator. To do that, we need some further notation.

**Definition 3.** For $\ell \in [\mu]$, let $S_\ell : \mathbb{K}^\mu \to \mathbb{K}^\mu$ be the shift operator

$$[S_\ell x]_k = x_{k+\ell}.$$  

We define $\mathcal{H} : \mathbb{K}^\mu \otimes \mathbb{K}^n \to \mathbb{K}^\mu$ through

$$w = \sum_{k \in [\mu]} e_k \otimes w_k \mapsto \mathcal{H}(w) = \sum_{k \in [\mu]} S_k Q w_k.$$  

The claim is now that up to reflections, the lifted version of the convolution is given by $\mathcal{H}$.

**Lemma III.1.** Define the reflection operator $R : \mathbb{R}^\mu \to \mathbb{R}^\mu$ through

$$[Rx]_k = x_{-k},$$  

and extend it to $\mathbb{R}^\mu \otimes \mathbb{R}^n$ through linear extension of $R(h \otimes b) = R(h) \otimes b$. Then, for $h \in \mathbb{R}^\mu$ and $b \in \mathbb{R}^n$,

$$C = \mathcal{H} \circ R$$

**Proof.** It suffices to show the claimed equality on elements $h \otimes b$. Hence, it suffices to show that $[h * (Qb)]_\ell = \mathcal{H}(Rh \otimes b)_\ell$ for each $\ell \in [\mu]$. Note that $Rh \otimes b = \sum_{k \in [\mu]} (Rh)_k e_k \otimes b = \sum_{k \in [\mu]} h_{-k} e_k \otimes b$. Hence, for each $\ell \in [\mu]$ we have

$$\mathcal{H}(Rh \otimes b)_\ell = \sum_{k \in [\mu]} [S_k (h_{-k} Qb)]_\ell = \sum_{k \in [\mu]} h_{-k} Qb_{k+\ell} = \sum_{j \in [\mu]} h_{\ell-j} [Qb]_j = [h * (Qb)]_\ell,$$

which was what to be shown. \qed

Since the reflection is an isometry both on $\mathbb{K}^\mu$ and $\mathbb{K}^\mu \otimes \mathbb{K}^n$ and leaves the hierarchical sparsity structure intact, we can concentrate on proving the HiRIP for the $\mathcal{H}$-operator. Let us introduce a further notational simplification.

**Definition 4.** (i) For $w = \sum_{k \in [N]} e_k \otimes w_k \in \mathbb{K}^\mu \otimes \mathbb{K}^n$ and $A \in \mathbb{K}^{m,n}$, we let $Aw$ denote the result of a block-wise application of $A$ to $w$, i.e.

$$Aw := \sum_{k \in [n]} e_k \otimes (Aw_k).$$

(ii) For $s, \sigma$, we define $T_{s,\sigma}$ be the set of $(s, \sigma)$-sparse vectors in $\mathbb{K}^\mu \otimes \mathbb{K}^n$ with norm less than or equal to 1, and $T_{\sigma} \subseteq \mathbb{K}^n$ similarly. We furthermore define $AT_{s,\sigma}$ as the image of $T_{s,\sigma}$ under $A$, i.e.

$$AT_{s,\sigma} = \{ \Sigma_{k \in [n]} e_k \otimes (Aw_k) \mid w = \Sigma_{k \in [n]} e_k \otimes w_k \in T_{s,\sigma} \}.$$  

We now prove a slightly less trivial lemma than the ones above. It shows that we may partially factor out the matrix $A$ from the analysis.

**Lemma III.2.** For $Q = UA$, let $\hat{\mathcal{H}} : \mathbb{K}^\mu \otimes \mathbb{K}^m \to \mathbb{K}^\mu$ denote the operator

$$v = \sum_{k \in [\mu]} e_k \otimes v_k \mapsto \hat{\mathcal{H}}(v) = \sum_{k \in [\mu]} S_k U v_k.$$  

Let further

$$\Delta_{s,\sigma}(\hat{\mathcal{H}}) := \sup_{w \in AT_{s,\sigma}} ||\hat{\mathcal{H}}(w)||^2 - ||w||^2,$$

(7)

We then have

$$\delta_{s,\sigma}(\mathcal{H}) \leq \Delta_{s,\sigma}(\hat{\mathcal{H}}) + \delta_{\sigma}(A) + \Delta_{s,\sigma}(\hat{\mathcal{H}}) \delta_{\sigma}(A).$$
Then, set of matrices in $B$. Suprema of empirical chaos processes stay in the bilevel regime mainly to increase readability. In the same manner, what follows should also be generalizable to the multilevel setting. We have chosen to effort to extend the above results to the case where the signal $\sigma$ (where $\sigma$ hierarchically sparse structure itself. Then, it suffices that the compression matrix $A$ is not only sparse, but hierarchically sparse. The main tool of our argument will be a result from [36]. It uses the so-called Talagrand’s $\gamma_0$-functionals. Since the definition of the functional is rather technical and not particularly insightful, we refer the interested reader to [39, Def. 1.2.5]. For our purposes, we only need to know that for a set $T$ and a metric $d$,
\[ \gamma_0(M, d) \lesssim \int_0^\infty \log(N(T, d, u))^{1/\alpha} du, \]
where $N(T, d, u)$ denotes the covering numbers of the set $T$ [39, 40, App. A], [41], and $\alpha$ is a positive number. The aforementioned result is as follows.

**Theorem III.3.** (36 Th. 3.1, see also Rem 3.4)] Let $\gamma$ be a vector as in assumption [2] and $T$ a symmetric set of matrices in $\mathbb{C}^{n \times n}$. Define
\[ d_{2 \rightarrow 2} = \sup_{T \in T} \|T\|_{2 \rightarrow 2}, \quad d_F = \sup_{T \in T} \|T\|_F \]
and set $E = \gamma_2(T, \| \cdot \|_{2 \rightarrow 2})^2 + d_F \cdot \gamma_2(T, \| \cdot \|_{2 \rightarrow 2})$, $V = d_{2 \rightarrow 2}(T) (\gamma_2(T, \| \cdot \|_{2 \rightarrow 2}) + d_F(T))$ and $U = d_{2 \rightarrow 2}(T)^2$. Then
\[ \mathbb{P} \left( \sup_{T \in T} \|T\gamma\|^2 - \mathbb{E} \left( \|T\gamma\|^2 \right) > CE + t \right) \leq 2 \exp \left( -\min \left\{ \frac{t^2}{4C^2}, \frac{t}{C} \right\} \right) . \]

We now, as in [36], rewrite our measurement $\mathcal{H}(w)$ as a product of a matrix and a random vector.

**Lemma III.4.** Let
\[ \mathcal{A} : \mathbb{C}^\mu \otimes \mathbb{C}^n \rightarrow \mathbb{C}^{\mu \times n}, \quad w \mapsto \left( \frac{1}{\mu \cdot 2^k} w_{j-k}(i) \right)_{k \in [\mu], (j, i) \in [\mu] \times [n]} . \]
Then,
\[ \Delta_{s, \sigma}(\bar{H}) = \sup_{T \in \mathcal{A}(AT_{s, \sigma})} \|T\gamma\|^2 - \mathbb{E} \left( \|T\gamma\|^2 \right) , \]
Proof. We have for each \( k \in [\mu] \)

\[
(\tilde{H}w)(k) = \sum_{\ell \in [\mu]} (S_{\ell}Uw_{\ell})(k) = \sum_{\ell \in [\mu], i \in [n]} (S_{\ell}u_{\ell})(k)w_{\ell}(i) = \sum_{\ell \in [\mu], i \in [n]} u_{i}(\ell + k)w_{\ell}(i)
\]

\[
= \frac{1}{\sqrt{\mu}} \sum_{j \in [\mu], i \in [n]} w_{j-k}(i)\gamma_{i}(j) = (A(w)\gamma)(k),
\]

i.e. \( \tilde{H}(w) = A(w)\gamma \). Now it is only left to note that

\[
E \left( \|A(w)\gamma\|^2 \right) = E \left( \text{tr}(\gamma^* A(w)^* A(w) \gamma) \right) = E \left( \text{tr}(A(w)^* A(w)) \right) = \|A(w)\|^2_F,
\]

where we in the penultimate step used that \( E(\gamma_j \gamma_j^*) = \delta_{ij} \). Now it is only left to note that

\[
\|A(w)\|^2_F = \sum_{k,j \in [\mu], i \in [n]} \frac{1}{\mu} |w_{j-k}(i)|^2 = \|w\|^2.
\]

To bound the constant \( \Delta_{s,\sigma}(H) \), we now need to estimate the parameters \( E, V \) and \( U \) in Theorem III.3.

To start, we have the following Lemma.

Lemma III.5. Define the norm

\[
\|v\|_X = \left( \sum_{i \in [m]} \sup_{\theta \in [0,2\pi]} \left| \sum_{k \in [\mu]} v_k(i)e^{ik\theta} \right|^2 \right)^{1/2}
\]

on \( \mathbb{K}^\mu \otimes \mathbb{K}^m \). Then, for \( u, w \in \mathbb{K}^\mu \otimes \mathbb{K}^m \)

\[
\|A(u - w)\|_{2 \rightarrow 2} \leq \frac{2}{\sqrt{\mu}} \|u - w\|_X.
\]

Proof. Let us for convenience write \( v = u - w \). Let us notice that \( A(v) \) is a matrix consisting of \( m \) Toeplitz blocks

\[
A(v) = [N^0, \ldots, N^m],
\]

with \( N^i = \frac{1}{\mu^{1/2}} (w_{j-k}(i))_{j,k \in [\mu]} \). The operator norm of a Toeplitz matrix \( M = (m_{j-k})_{j,k \in [\mu]} \) is well-known to be upper bounded by (see for example [32])

\[
\|M\|_{2 \rightarrow 2} \leq 2 \sup_{\theta \in [0,2\pi]} \left| \sum_{k \in [\mu]} m_k e^{ik\theta} \right|.
\]

We may therefore estimate

\[
\|N^i\|_{2 \rightarrow 2} \leq 2 \sup_{\theta \in [0,2\pi]} \left| \frac{1}{\sqrt{\mu}} \sum_{k \in [\mu]} v_k(i)e^{ik\theta} \right|,
\]

which together with the bound

\[
\|A(v)\gamma\| \leq \sum_{i \in [m]} \|N^i\| \|\gamma_i\| \leq \left( \sum_{i \in [m]} \|N^i\|^2 \right)^{1/2} \|\gamma\|
\]

gives the claim. \( \square \)

The above lemma shows that \( \mathcal{N}(A(\mathcal{A}T_{s,\sigma}), \| \cdot \|_{2 \rightarrow 2}, t) \leq \mathcal{N}(A(\mathcal{A}T_{s,\sigma}, \frac{2}{\sqrt{\mu}} \| \cdot \|_X, t), \) and similar for the diameters. The rest of the proof will therefore consist in estimating the quantities involving the \( \| \cdot \|_X \)-norm.
C. A simple covering number bound

The estimation of the $\gamma_2$-functional boils down to bounding the covering numbers $\mathcal{N}(AT_{s,\sigma}, \| \cdot \|_{X,t})$. The following lemma gives us a first, simple, estimate of the latter.

**Lemma III.6.** Let $u \in AT_{s,\sigma}$ have a representation $\hat{u} \in T_{s,\sigma}$, i.e., $u = A\hat{u}$. Then

$$\|u\|_{X} \leq \sqrt{s}(1 + \delta_\sigma(A))^{1/2}\|\hat{u}\|.$$ 

In particular,

$$\log(\mathcal{N}(A(\mathcal{A}T_{s,\sigma}), \| \cdot \|_{2\to 2}, t)) \leq s \log(\mu) + s\sigma \log(n) + 2s\sigma \log(1 + \frac{4\sqrt{s}(1 + \delta_\sigma(A))^{1/2}}{t\sqrt{\sigma}}).$$

**Proof.** For each $\theta \in [0, 2\pi]$ and $i \in [m]$, we have

$$\left| \sum_{k \in [\mu]} u_k(i) e^{ik\theta} \right| \leq \sum_{k \in [\mu]} |u_k(i)| \leq \sqrt{s}\|u\|,$$

where we have used reverse norm-ordering $\|v\|_1 \leq \sqrt{s}\|v\|_2$ for an $s$-sparse vector for the $(u_k(i))_{k \in [\mu]}$ with fixed $i$. To prove the first claim, it is now only left to note that since each $\hat{u}_k$ is $\sigma$-sparse and $A$ has the $\sigma$-RIP, we have

$$\|u\|^2 = \sum_{k \in [\mu]} \|A\hat{u}_k\|^2 \leq \sum_{k \in [\mu]} (1 + \delta_\sigma(A))\|\hat{u}_k\|^2 = (1 + \delta_\sigma(A))\|\hat{u}\|^2.$$ 

The above estimate allows us to bound the covering number as follows. Let us set $\kappa = 2\sqrt{\frac{2}{\mu}(1 + \delta_\sigma(A))^{1/2}}$. For each $(s,\sigma)$-sparse support $\Omega$, let $U_\Omega$ denote intersection of the unit ball with the signals supported on $\Omega$. Since the latter spaces in both the real and complex case have a dimension smaller than $2s\sigma$, it is well known that (see for example [38, Appendix 2]) we have for each $t > 0$

$$\mathcal{N}(U_\Omega, \kappa \cdot \| \cdot \|_{s,\sigma}, t) \leq \mathcal{N}(U_\Omega, \| \cdot \|, \frac{\mu}{t^2}) \leq \left(1 + \frac{2\kappa}{\mu}\right)^{2s\sigma}.$$ 

The above means that there exists a $t$-net of $U_\Omega$ w.r.t $\kappa \cdot \| \cdot \|$ of cardinality less than $\left(1 + \frac{2\kappa}{\mu}\right)^{2s\sigma}$. By the inequality we just proved, and Lemma [III.5] such a net is however also a $t$-net of $\mathcal{A}(U_\Omega)$ with respect to $\| \cdot \|_{2\to 2}$. Hence, since there are $\binom{n}{s^2} (\binom{s}{\sigma})^s$ $(s,\sigma)$-sparse supports, we conclude

$$\mathcal{N}(A(\mathcal{A}T_{s,\sigma}), \| \cdot \|_{2\to 2}, t) \leq \binom{n}{s^2} (\binom{s}{\sigma})^s\left(1 + \frac{2\kappa}{\mu}\right)^{2s\sigma},$$

which yields the second claim.

The above estimate of the covering number is a little bit too crude for large values of $t$. To provide a better one, we will have to use more involved tools.

D. The empirical method of Maurey

Now we will use Maurey’s empirical method to provide a stronger bound of the covering numbers $\mathcal{N}(AT_{s,\sigma}, \| \cdot \|_{X,t})$. The idea of this method is to, given $v \in AT_{s,\sigma}$ define random variables $\bar{v}$ that only take on finitely many, say $N$, values, and subsequently show that $\mathbb{E}(\bar{v} - v) \leq t$. This shows in particular that one of the $N$ values $\bar{v}$ can attain is closer to $v$ than $t$ and thus, $\mathcal{N}(AT_{s,\sigma}, \| \cdot \|_{X,t}) \leq N$.

The first step towards defining the random variable $\bar{v}$ is to notice that $T_{s,\sigma}$ is contained in a scaled version of the $\ell_1$-unit ball $B_{1,2}$. Recall that the $\ell_1,2$-norm of an element $\tilde{v} = \sum_{k \in [\mu]} e_k \otimes \tilde{v}_k \in K^\mu \otimes K^n$ is given by $\|\tilde{v}\|_{1,2} = \sum_{k \in [\mu]} \|\tilde{v}_k\|$. Now, if $\hat{v} \in T_{s,\sigma}$, the vector $\nu = (\|\tilde{v}_k\|)_{k \in [\mu]}$ is $s$-sparse. Hence, the reversed norm ordering $\|w\|_1 \leq \sqrt{s}\|w\|_2$ for $s$-sparse $w$ yields

$$\|\hat{v}\|_{1,2} = \sum_{k \in [\mu]} \|\tilde{v}_k\| = \|\nu\|_1 \leq \sqrt{s}\|\nu\|_2 = \sqrt{s}\|\tilde{v}\|.$$
Hence, $\hat{v} \in \sqrt{s}B_{1,2}$, where $B_{1,2}$ is the unit ball with respect to the $\| \cdot \|_{1,2}$-norm. Clearly, $\hat{v}$ is also in the set

$$\Sigma_\sigma'' = \left\{ \sum_{k \in \mu} e_k \otimes \hat{v}_k \mid \hat{v}_k \text{ $\sigma$-sparse.} \right\}$$

so that every $v \in A(T_{s,\sigma})$ must lie in $A(\sqrt{s}B_{1,2} \cap \Sigma_\sigma'')$. This insight leads to the following statement.

**Lemma III.7.**

$$\mathcal{N}(\frac{1}{\sqrt{s}} A(T_{s,\sigma}), \| \cdot \|_{1,2}, t) \leq \mathcal{N}(A(B_{1,2} \cap \Sigma_\sigma''), \| \cdot \|_{1,2}, 14).$$

**Proof.** By the above discussion, $\frac{1}{\sqrt{s}} A(T_{s,\sigma}) \subseteq A(B_{1,2} \cap \Sigma_\sigma'')$. It is only left to utilize a well-known property of covering numbers. For convenience, let us provide the argument here: By what we just argued, a $\frac{t}{2}$-net $\Theta$ for $A(B_{1,2} \cap \Sigma_\sigma'')$ is also a set with the property that every element in $\frac{1}{\sqrt{s}} A(T_{s,\sigma})$ has an element in $\Theta$ at a distance at most $\frac{t}{2}$ from it. Hence, it is almost a $\frac{t}{2}$-net, but not quite: $\Theta$ must not necessarily consist of elements of $\frac{1}{\sqrt{s}} A(T_{s,\sigma})$. But, it is clear that if we define $\Theta'$ as the subset of $\Theta$ that are at most a distance $\frac{t}{2}$ from $\frac{1}{\sqrt{s}} A(T_{s,\sigma})$. By construction $\Theta'$ still covers $\frac{1}{\sqrt{s}} A(T_{s,\sigma})$. Now we can replace all elements of $\Theta'$ with a $\frac{t}{2}$-close point of $\frac{1}{\sqrt{s}} A(T_{s,\sigma})$ and, by triangle inequality, arrive at a $t$-cover net. The statement follows. \(\square\)

Now let us construct the aforementioned random variables $v$ which will help us estimate $\mathcal{N}(A(B_{1,2} \cap \Sigma_\sigma''), \| \cdot \|_{1,2}, t)$. First, take $\Gamma$ be an $(\frac{t}{2}(1+\delta_r(A))^{1/2})$-net for the set $\overline{T_{s,\sigma}} \subseteq \mathbb{R}^n$ with respect to the $\ell_2$ norm. By the same argument as in the proof of Lemma III.6, we know that $\Gamma$ can be chosen so that its cardinality obeys

$$\Gamma \leq \left( \frac{n}{\sigma} \right) (1 + \frac{4(1+\delta_r(A))^{1/2}}{t})^{2\sigma}.$$ 

Now let $v \in U$ be arbitrary, with $v = A\hat{v}$. For each $\hat{v}_k$, choose an $\hat{u}_k \in \Gamma$ with

$$\left\| \frac{\hat{v}_k}{|\hat{v}_k|} - \hat{u}_k \right\| \leq \frac{t}{2(1+\delta_r(A))^{1/2}}.$$ 

This is possible due to the fact that $\frac{\hat{v}_k}{|\hat{v}_k|} \in T_{s,\sigma}$. Notice that we can make sure that $\text{supp} \hat{v}_k = \text{supp} \hat{u}_k$, by the construction of the net of $T_{s,\sigma}$. Now let us define the random variable $Z$ through

$$P(Z = e_k \otimes A\hat{u}_k) = \|\hat{v}_k\|, \quad P(Z = 0) = 1 - \|v\|_{1,2}.$$ 

Letting $Z(r)$, $r \in [M]$ (for some $M$ later to be chosen) be independent copies of $Z$, we now finally define

$$\nu = \frac{1}{M} \sum_{r \in [M]} Z(r).$$ 

Across all possible values of $v$, the variables $\nu$ can only take on a discrete set of values with a certain size. Let us record that for future reference.

**Lemma III.8.** There is a set $S$ with

$$|S| \leq \left( 2\mu \left( \frac{n}{\sigma} \right) (1 + \frac{4(1+\delta_r(A))^{1/2}}{t})^{2\sigma} \right)^M$$

so that

$$P(\forall v \in A(B_{1,2} \cap \Sigma_\sigma'') : \nu \in S) = 1.$$ 

**Proof.** For all $v$ the corresponding random variable $Z$ takes values in the set

$$\{0\} \cap \{e_k \otimes Au \mid u \in \Gamma, k \in [\mu]\},$$

which has a cardinality $1 + \mu \cdot |T_{s,\sigma}| \leq 2\mu \cdot |T_{s,\sigma}|$. This holds true for each copy $Z(r)$. Hence, the array $(Z(0), \ldots, Z(M-1))$ can only $2(\mu \cdot |T_{s,\sigma}|)^M$ different values. This implies the same statement for $\nu = \frac{1}{M} \sum_{r \in [M]} Z(r)$. The bound on $|T_{s,\sigma}|$ now yields the claim. \(\square\)
It is now left to estimate $\mathbb{E} (\|y - v\|)_X$. Let us first show that $\mathbb{E} (y)$ is close to $v$.

**Lemma III.9.**

$$\|\mathbb{E} (y) - v\|_X \leq \frac{t}{2}$$

**Proof.** By the definition of $Z$, we have

$$\mathbb{E} (y) = \mathbb{E} (Z) = \sum_{k \in [\mu]} (e_k \otimes A\tilde{u}_k) \mathbb{P} (Z = e_k \otimes A\tilde{u}_k) = \sum_{k \in [\mu]} e_k \otimes \|\tilde{u}_k\| \|A\tilde{u}_k\|.$$  

Let us write $u_k = A\tilde{u}_k$. The above implies

$$\|\mathbb{E} (Z) - v\|_X^2 = \sum_{i \in [m]} \sup_{\theta \in [0, 2\pi]} \left| \sum_{k \in [\mu]} (v_k(i) - \|\tilde{u}_k\| u_k(i)) e^{ik\theta} \right|^2 \leq \sum_{i \in [m]} \left( \sum_{k \in [\mu]} |v_k(i) - \|\tilde{u}_k\| u_k(i)| \right)^2.$$  

Writing the elements of the inner sums as $\|\tilde{u}_k\|^{1/2} \cdot \|v_k(i) - \|\tilde{u}_k\|^{1/2} u_k(i)\|$ and utilizing the Cauchy-Schwarz inequality yields

$$\left( \sum_{k \in [\mu]} |v_k(i) - \|\tilde{u}_k\| u_k(i)| \right)^2 \leq \left( \sum_{k \in [\mu]} \|\tilde{u}_k\| \right) \cdot \left( \sum_{k \in [\mu]} \left| \frac{v_k(i)}{\|v_k\|^{1/2}} - \|\tilde{u}_k\|^{1/2} u_k(i) \right|^2 \right)$$  

for each $i$. Taking the sum over $i$, and utilizing that $\tilde{u}_k \in B_{1, 2}$, we arrive at

$$\|\mathbb{E} (Z) - v\|_X^2 \leq \sum_{i \in [m]} \sum_{k \in [\mu]} \left| \frac{v_k(i)}{\|v_k\|^{1/2}} - \|\tilde{u}_k\|^{1/2} u_k(i) \right|^2 = \sum_{k \in [\mu]} \|\tilde{u}_k\| \cdot \left| \frac{v_k}{\|v_k\|} - u_k \right|^2.$$  

Now, for each $k$, we have

$$\left| \frac{v_k}{\|v_k\|} - u_k \right| = \left| A \left( \frac{v_k}{\|v_k\|} - \tilde{u}_k \right) \right| \leq (1 + \delta_\sigma (A))^{1/2} \left\| \frac{v_k}{\|v_k\|} - \tilde{u}_k \right\| \leq \frac{1 + \delta_\sigma (A)}{2(1 + \delta_\sigma (A))^{1/2}} \frac{t}{2} = \frac{t}{4}.$$  

We utilized that $\tilde{u}_k$ and $\tilde{u}_k$ are $\sigma$-sparse with the same support, and that $\Gamma$ is a net for $T_{\sigma}$. Consequently,

$$\sum_{k \in [\mu]} \|\tilde{u}_k\| \cdot \left| \frac{v_k}{\|v_k\|} - u_k \right|^2 \leq \frac{t^2}{4} \sum_{k \in [\mu]} \|\tilde{u}_k\| \leq \frac{t^2}{4}.$$  

\[\square\]

The above lemma implies that it suffices to study $\mathbb{E} (\|y - \mathbb{E} (y)\|_X)$ to get the bound we want. This is the purpose of the next lemma.

**Lemma III.10.** Under the assumption that $M \geq \sqrt{2 \log (m \mu_m)}$, we have

$$\mathbb{E} (\|y - \mathbb{E} (y)\|_X) \leq 3\pi \sqrt{\frac{\log (8m \mu_m) (1 + \delta_\sigma (A))}{M}}$$

**Proof.** First, a symmetrization argument (see e.g [38] Lemma 8.4) yields that

$$\mathbb{E} (\|y - \mathbb{E} (y)\|_X) = \mathbb{E} \left( \frac{1}{M} \sum_{r \in [M]} Z^{(r)} - \mathbb{E} \left( Z^{(r)} \right) \right) \leq \mathbb{E} \left( \frac{2}{M} \sum_{r \in [M]} \mathbb{E} \left( \mathbb{E} \left( \epsilon_r Z^{(r)} \right) \right) \right),$$

where $\epsilon_r$ are independent Rademacher random variables. To estimate this expected value, let us for each fixed $i \in [m]$ and $\theta \in [0, 2\pi]$ consider the variable

$$Y_i(\theta) = \frac{2}{M} \sum_{r \in [M]} \sum_{k \in [\mu]} \epsilon_r Z^{(r)}_{k}(i) e^{ik\theta}.$$
\( Y_i(\theta) \) is a random variable of the form \( \sum_{r \in [M]} \beta_r \varepsilon_r \), with
\[
\beta_r = \frac{2}{M} \sum_{k \in [|\mu|]} Z_k^{(r)}(i) e^{ik\theta}.
\]

For each instance of \( Z^{(r)} \), \( Z_k^{(r)} \) is only non-zero for at most one index \( k = k_r \). Consequently,
\[
|\beta_r| = \frac{2}{M} \left| Z_k^{(r)}(i) e^{ik\theta} \right| = \frac{2}{M} |A\tilde{u}_{k_r}(i)|,
\]
where we used the definition of \( Z \) in the final step. (Note that if \( Z^{(r)} = 0 \), we can define \( \tilde{u}_{k_r} = 0 \).) By the Hoeffding inequality for Rademacher sums [38, Lem. 8.8],
\[
P \left( \left| \sum_{r \in [M]} \varepsilon_r \beta_r \right| \geq \alpha \| \beta \| \right) \leq 2e^{-\alpha^2/2},
\]
were \( \beta \in \mathbb{C}^M \) is arbitrary but fixed, we obtain
\[
P_\varepsilon \left( |Y_i(\theta)| \geq \alpha \left( \frac{1}{M} \sum_{r \in [M]} |A\tilde{u}_{k_r}(i)|^2 \right)^{1/2} \right) \leq 2e^{-\alpha^2/2}, \tag{8}
\]
where the \( P_\varepsilon \) means that the probability is with respect to the draw of the \( \varepsilon \), i.e., conditioned on the draw of the \( Z^{(r)} \).

Now let us transform this bound into a bound uniform over \( \theta \in [0, 2\pi] \) and \( i \in [m] \). We have
\[
|Y_i(\theta) - Y_i(\theta')| \leq \frac{2\pi}{M} \sum_{r \in [M]} \sum_{k \in [|\mu|]} |\varepsilon_r| \left| Z_k^{(r)}(i) \right| |e^{ik\theta} - e^{ik\theta'}| \leq \frac{2\pi}{M} \sum_{r \in [M]} \sum_{k \in [|\mu|]} |Z_k^{(r)}(i)| \left| k(\theta - \theta') \right|.
\]
Also, by the same argument as before, the sums over \( k \) actually only consist of one element, and that element is \( |A\tilde{u}_{k_r}(i)||k_r(\theta - \theta')| \). Since \( |k_r(\theta - \theta')| \leq \mu |\theta - \theta'| \), we conclude that the above is smaller than
\[
\frac{2\pi}{M} \sum_{r \in [M]} |A\tilde{u}_{k_r}(i)| \mu |\theta - \theta'|.
\]
Therefore, if \( I_\mu = \left\{ \frac{k \cdot 2\pi}{|\mu|} \mid k \in [|\mu|] \right\} \), we have
\[
\sup_{\theta \in [0,2\pi]} |Y_i(\theta)| \leq \sup_{\theta \in I_\mu} |Y_i(\theta)| + \frac{2\pi}{M} \sum_{r \in [M]} |A\tilde{u}_{k_r}(i)| \leq \sup_{\theta \in I_\mu} |Y_i(\theta)| + \frac{2\pi}{\sqrt{|\mu|}} \left( \sum_{r \in [M]} |A\tilde{u}_{k_r}(i)|^2 \right)^{1/2}. \tag{9}
\]
This equation together with a union bound over \( i \in [m] \) and \( \theta \in I_\mu \) in [38] yields
\[
P_\varepsilon \left( \sup_{i \in [m], \theta \in [0,2\pi]} |Y_i(\theta)| \geq \left( \frac{2\alpha}{M} + \frac{2\pi}{\sqrt{|\mu|}} \right) \left( \sum_{r \in [M]} |A\tilde{u}_{k_r}(i)|^2 \right)^{1/2} \right) \leq 2\mu \alpha e^{-\alpha^2/2},
\]
which for \( \alpha \geq \sqrt{|\mu|} \) yields
\[
P_\varepsilon \left( \sup_{i \in [m], \theta \in [0,2\pi]} |Y_i(\theta)| \geq \frac{2\pi \alpha}{M} \left( \sum_{r \in [M]} |A\tilde{u}_{k_r}(i)|^2 \right)^{1/2} \right) \leq 2\mu \alpha e^{-\alpha^2/2}.
\]

Assuming the counter event, we have
\[
\left\| \frac{1}{M} \sum_{r \in [M]} \varepsilon_r Z^{(r)} \right\|_X = \sum_{i \in [m], \theta \in [0,2\pi]} \left| \sum_{k \in [|\mu|]} \frac{1}{M} \sum_{r \in [M]} \varepsilon_r Z_k^{(r)}(i) e^{ik\theta} \right|^2 = \sum_{i \in [m], \theta \in [0,2\pi]} |Y_i(\theta)|^2 \leq \frac{4\pi}{M} \sum_{r \in [M]} \sum_{i \in [m]} |A\tilde{u}_{k_r}(i)|^2 + \frac{4\pi}{M} \sum_{r \in [M]} \|A_{k_r}\|^2 \leq \frac{4\pi}{M} (1 + \delta_{\alpha}(A)).
\]
where we in the final step utilized that $\tilde{u}_{k_r} \in T_\sigma$. Hence, for $\alpha \geq \sqrt{M}$

$$
P_\varepsilon \left( \left\| \frac{1}{M} \sum_{r \in [M]} \varepsilon_r Z^{(r)} \right\|_X \geq \frac{2\pi \mu}{\sqrt{M}} \left( 1 + \delta_\sigma (A) \right)^{1/2} \right) \leq 2\mu m e^{-\alpha^2/2}.
$$

This tail-bound can be turned into a bound on the expected value by elementary methods. For convenience, let us just invoke [38, Prop 7.14]. It yields, in particular together with the assumption $M \geq \sqrt{2} \log(2\mu m)$,

$$
E_\varepsilon \left( \left\| \frac{2}{M} \sum_{r \in [M]} \varepsilon_r Z^{(r)} \right\|_X \right) \leq 3\pi \sqrt{\frac{\log(8\mu m) (1 + \delta_\sigma (A))}{M}}
$$

where we used that $\tilde{u}_{k_r} \in T_\sigma$ for all $r$. The theorem of Fubini now yields the claim. \hfill \box

**Proposition III.11.** For $t \leq \left( 16 \pi^2 \log(2\mu m) \right)^{1/4}$, we have

$$
\log(N(B_{A(1,2) \cap \Sigma^\alpha_\sigma}), \| \cdot \|_{1, t}) \lesssim \frac{\log(2\mu m) (1 + \delta_\sigma (A))}{t^2} \left( \log(\mu) + \sigma \log(n) + \sigma \log \left( 1 + \frac{4(1 + \delta_\sigma (A))^{1/2}}{t} \right) \right).
$$

**Proof.** Let us choose

$$
M = \left[ \frac{36\pi^2}{t^2} \log(8\mu m) (1 + \delta_\sigma (A)) \right].
$$

Then, due to the assumption $t \leq \left( 16 \pi^2 \log(2\mu m) \right)^{1/4}$, $M \geq \sqrt{2} \log(2\mu m)$. Hence, [III.10] is applicable. Together with Lemma [III.9] and the triangle inequality, we get

$$
E(\|v-u\|) \leq t.
$$

Hence, at least one of the values $v$ can take on must lie closer to $u$ than $t$, so that the set $S$ described in Lemma [III.8] forms a $t$-net for $A(1,2) \cap \Sigma^\alpha$. Consequently

$$
\log(N(U, \| \cdot \|_{1, t}, t)) \leq \log(|S|) \lesssim M \left( \log(2\mu) + \sigma \log(n) + \sigma \log \left( 1 + \frac{4(1 + \delta_\sigma (A))^{1/2}}{t} \right) \right) \lesssim \frac{\log(8\mu m) (1 + \delta_\sigma (A))}{t^2} \left( \log(\mu) + \sigma \log(n) + \sigma \log \left( 1 + \frac{4(1 + \delta_\sigma (A))^{1/2}}{t} \right) \right),
$$

which is the claim. \hfill \box

**E. Estimating the parameters in Theorem [III.3]**

Now we have all the tools we need to bound $\gamma_2(A(T_{s,\sigma}), \| \cdot \|_{2-2})$. For convenience, let us begin by collecting all our previous estimates.

**Corollary III.12.** Define

$$
\varphi_0(t) = s \log \mu + s \sigma \log(n) + 2s \sigma \log(1 + \frac{1}{t})
$$

and

$$
\varphi_1(t) = \frac{\log(8\mu m)}{t^2} \left( \log(\mu) + \sigma \log(n) + \sigma \log \left( 1 + \frac{16}{t} \right) \right).
$$

It holds that

$$
\log(N(A(T_{s,\sigma}), \| \cdot \|_{2-2}, t)) \lesssim \varphi_0 \left( \sqrt{\frac{2}{t}} (1 + \delta_\sigma (A))^{-1/2} \right)
$$

and

$$
\log(N(A(T_{s,\sigma}), \| \cdot \|_{2-2}, t)) \lesssim \varphi_1 \left( \sqrt{\frac{2}{t}} (1 + \delta_\sigma (A))^{-1/2} \right) \quad \text{for } t \leq \frac{24 \sqrt{\pi} \log(2\mu m)^{1/4}}{\sqrt{\varepsilon}}.
$$

**Proof.** The first line is just a paraphrasing of Lemma [III.6] As for the second, we first use elementary properties of covering numbers (see e.g. [38] App. C.2) to argue that

$$
N(A(T_{s,\sigma}), \| \cdot \|_{2-2}, t) = N \left( \frac{1}{\varepsilon} A(T_{s,\sigma}), \| \cdot \|_{2-2}, \frac{\varepsilon}{\sqrt{n}} \right) \leq N \left( \frac{1}{\varepsilon} A(T_{s,\sigma}), \frac{2}{\sqrt{n}} \| \cdot \|_{2-2}, \frac{\varepsilon}{\sqrt{n}} \right) \leq N \left( \frac{1}{\varepsilon} A(T_{s,\sigma}), \| \cdot \|_{2-2}, \frac{\varepsilon}{\sqrt{n}} \right) \leq N \left( \frac{1}{\varepsilon} A(T_{s,\sigma}), \| \cdot \|_{\frac{1}{\sqrt{n}}}, \frac{\varepsilon}{\sqrt{n}} \right).
$$
where we used Lemma [III.7] in the final step. Now we apply Lemma [III.11]. Note that the bound given there is applicable for $\frac{\sqrt{\sigma}}{4\sqrt{\pi}} \leq 6(2\log(2\mu m))^{1/4}$, as assumed in the lemma.

Using the above, it is now straight-forward to bound the parameters relevant for (III.3).

**Corollary III.13.** Suppose that $\delta_\sigma(A) < 1$. Then,

$$d_{2 \rightarrow 2}(A(AT_{s,\sigma}), \| \cdot \|_{2 \rightarrow 2}) \leq \sqrt{\frac{\sigma}{\mu}} (1 + \delta_\sigma(A))^{1/2},$$

$$d_F(A(AT_{s,\sigma}), \| \cdot \|_{2 \rightarrow 2}) \leq (1 + \delta_\sigma(A))^{1/2}$$

$$\gamma_2(A(AT_{s,\sigma}), \| \cdot \|_{2 \rightarrow 2}) \leq \frac{1}{\sqrt{\mu}} (1 + \delta_\sigma(A))^{1/2}\left(s \log(s) \log(\mu m) \log(\mu) + \sigma \log(n)\right)^{1/2}$$

**Proof.** The statement about $d_{2 \rightarrow 2}(A(AT_{s,\sigma}), \| \cdot \|_{2 \rightarrow 2})$ follows directly from Lemma [III.5]. As for the Frobenius norm statement, we have for $w \in AT_{s,\sigma}$ with representation $\tilde{w} \in T_{s,\sigma}$, we have

$$\|A(w)\|_F^2 = \frac{1}{\mu} \sum_{k,j \in [n], i \in [n]} |w_{j-k}(i)|^2 = \|w\|^2 \leq \left(1 + \delta_\sigma(A)\right)\|\tilde{w}\|^2,$$

which implies the bound.

To bound the $\gamma_2$-functional, let us introduce the notation

$$\kappa = \sqrt{\frac{\sigma}{\mu}} (1 + \delta_\sigma(A))^{1/2}.$$

The aforementioned Dudley bound implies

$$\gamma_2(A(AT_{s,\sigma}), \| \cdot \|_{2 \rightarrow 2}) \leq \int_0^\kappa \log(\mathcal{N}(A(AT_{s,\sigma}), \| \cdot \|_{2 \rightarrow 2}, t))^{1/2} dt,$$

where we in particular utilized that $d_{2 \rightarrow 2}(A(AT_{s,\sigma}), \| \cdot \|_{2 \rightarrow 2}) \leq \kappa$. Now, Lemma [III.12] implies that

$$\log(\mathcal{N}(A(AT_{s,\sigma}), \| \cdot \|_{2 \rightarrow 2}, t))^{1/2} \leq \min(\varphi_0(\kappa^{-1} t), \varphi_1(\kappa^{-1} t))^{1/2},$$

for all $t \leq \frac{24\sqrt{\sigma}(2\log(2\mu m))^{1/4}}{\sqrt{\mu}}$. However, since

$$\kappa = \sqrt{\frac{\sigma}{\mu}} (1 + \delta_\sigma(A))^{1/2} \leq \frac{\sqrt{\sigma}}{\sqrt{\mu}} \leq \frac{24\sqrt{\sigma}(2\log(2\mu m))^{1/4}}{\sqrt{\mu}},$$

simply by the fact that $24(2\log(2\mu m))^{1/4} \geq 24(2\log(2))^{1/4} \geq 1$, the bound is valued for all $t \leq \kappa$. Hence,

$$\int_0^\kappa \log(\mathcal{N}(A(AT_{s,\sigma}), \| \cdot \|_{2 \rightarrow 2}, t))^{1/2} dt \leq \int_0^\kappa \min(\varphi_0(\kappa^{-1} t), \varphi_1(\kappa^{-1} t))^{1/2} dt$$

$$= \kappa \int_1^1 \min(\varphi_0(t), \varphi_1(t))^{1/2} dt.$$

Now, let’s split the integral to one from 0 to $\frac{1}{\sqrt{\sigma}}$, and one from $\frac{1}{\sqrt{\sigma}}$ to 1. For the first integral, we have

$$\kappa \int_0^1 \min(\varphi_0(t), \varphi_1(t))^{1/2} dt \leq \kappa \int_0^{1/\sqrt{\sigma}} \varphi_0(t)^{1/2} dt \leq \frac{\sqrt{\sigma}}{\sqrt{\pi}} (s \sigma \log(n) + s \log(\mu))^{1/2}$$

simply because the integral of $\log(1 + \frac{1}{t})^{1/2}$ from 0 to 1 converges. We may furthermore estimate

$$\frac{\sqrt{\sigma}}{\sqrt{\pi}} (s \sigma \log(n) + s \log(\mu))^{1/2} = \frac{1}{\sqrt{\mu}} (1 + \delta_\sigma(A))^{1/2} \left(s \sigma \log(n) + s \log(\mu)\right)^{1/2}$$

$$\leq \frac{1}{\sqrt{\mu}} (1 + \delta_\sigma(A))^{1/2} \left(s \log(s) \log(\mu m) \log(\mu) + \sigma \log(n)\right)^{1/2}$$

As for the second one, note that

$$\varphi_1(t)^{1/2} \leq \frac{1}{t} \left(2 \log(2\mu m) \left(\log(\mu) + \sigma \log(n) + \sigma \log(1 + 8\sqrt{\sigma})\right)\right)^{1/2}$$

for $t \geq \frac{1}{\sqrt{\sigma}}$, and $\int_{1/\sqrt{\sigma}}^1 \frac{1}{t} dt = \log(\sqrt{\pi})$. Hence,

$$\kappa \int_{1/\sqrt{\sigma}}^1 \varphi_1(t)^{1/2} ds \leq \frac{\sqrt{\pi}}{\sqrt{\mu}} \left(1 + \delta_\sigma(A)\right)^{1/2} \left(2 \log(2\mu m) \left(\log(\mu) + \sigma \log(n) + \sigma \log(1 + 8\sqrt{\sigma})\right)\right)^{1/2}$$

$$\leq \frac{1}{\sqrt{\mu}} \left(1 + \delta_\sigma(A)\right)^{1/2} \left(s \log(s) \log(\mu m) \log(\mu) + \sigma \log(n)\right)^{1/2}.$$

The claim has been proven.
F. Conclusion

We now have all the tools we need to prove our main theorem. Let us begin by bounding $\Delta_{s,\sigma}(\mathcal{H})$.

**Theorem III.14.** Under our assumptions,

$$P\left(\Delta_{s,\sigma}(\hat{H}) > (1 + \delta_\sigma(A))\delta_0\right) \leq 2\epsilon.$$

**Proof.** Let us define

$$\lambda = \frac{1}{\mu}(s \log(s)^2 \log(\mu n) \log(\mu) + s\sigma \log(n)).$$

Our assumption on $\mu$ states that

$$\lambda \lesssim \delta_0^2 \min(1, \log(\epsilon^{-1})^{-1}).$$

This together with Corollary III.13 then states that, with the notation used in Theorem III.3,

$$E \lesssim (1 + \delta_\sigma(A))(\lambda + \sqrt{\lambda}) \lesssim (1 + \delta_\sigma(A))\delta_0,$$

$$V \sim (1 + \delta_\sigma(A))(\lambda + \sqrt{\lambda}) \lesssim (1 + \delta_\sigma(A)) \log(\epsilon^{-1})^{-1/2}\delta_0,$$

$$U \lesssim (1 + \delta_\sigma(A))\frac{\lambda}{\mu} \leq (1 + \delta_\sigma(A))\lambda \leq \delta_0 \log(\epsilon^{-1})^{-1}.$$

Therefore, said theorem implies

$$P\left(\Delta_{s,\sigma}(\hat{H}) > (1 + \delta_\sigma(A))\delta_0\right) \leq 2 \exp(-\log(\epsilon^{-1})) = 2\epsilon,$$

which was the claim.

We may now conclude the proof.

**Proof of Theorem II.1.** First, Lemma III.1 and Lemma III.2 imply that

$$\delta_{s,\sigma}(\mathcal{C}) \leq \Delta_{s,\sigma}(\mathcal{H}) \leq \delta_{s,\sigma}(\mathcal{H}) + \delta_\sigma(A) + \Delta_{s,\sigma}(\hat{H})\delta_\sigma(A).$$

Now, the lower bound on the number of measurements II.4 together with Theorem III.14 implies that

$$P\left(\Delta_{(s,\sigma)}(\hat{H}) \geq \delta_0(1 + \delta_\sigma(A))\right) \leq \epsilon.$$

Thus, with a probability larger than $1 - \epsilon$,

$$\delta_{(s,\sigma)}(\mathcal{C}) \leq \delta_0(1 + \delta_\sigma(A)) + \delta_\sigma(A) + \delta_0\delta_\sigma(A)(1 + \delta_\sigma(A)) = (1 + \delta_\sigma(A))^2\delta_0 + \delta_\sigma(A),$$

which was the claim.

We may now conclude the proof.

**IV. Numerics**

Let us make a small numerical experiment to test Theorem II.1. In particular, we want to investigate whether the number of measurements the needed exactly scales as our complexity bound suggests, or if e.g. the $\log(s)$-terms more likely are proof artefacts. Note that such a comparison is inevitably indirect since we numerically observe average case performances while the theoretical results are worst-case bounds over all problem instances.
Fig. 2: Recovery probability plots over $s = 1, 2, \ldots, 7$ and $\mu = 10, 20, \ldots, 250$ for different values of $n$ and $\sigma$.

a) Implementation details: We have implemented our algorithm in the python package CuPy [43], an open source package for running NumPy-based scripts on NVIDIA GPUs. We have chosen to do so to utilize the opportunities for parallelization the HiHTP-algorithm allows in this context: When applying the operator $H$ we need to calculate $S_k(Ww_k)$ for all $k \in [\mu]$. Similarly, when applying applying $H^\ast$  

$$
H^\ast(y) = \sum_{k\in[\mu]} e_k \otimes (Q^\ast S_{-k}y),
$$

$(Q^\ast S_{-k}y)$ needs to be calculated for all $k \in [\mu]$. Both these sets of calculations can be done in parallel.

Finally, the application of the thresholding operation also benefits from parallelization – as discussed in [29], the application of the thresholding operation consists in first projecting each block onto the set of $\sigma$-sparse vectors, and then choosing the $s$ projected blocks with the largest norms. The first step here can again be parallellized over the block dimension.

For solving the least-squares problems in each step of the HiHTP-algorithm, we apply a conjugated gradient algorithm, which we stop once the $\ell_2$-residual is smaller than 1e-4, or 100 iterations have passed. This is justified since in the regime of HiRIP, the restricted least squares problems are expected to be well-conditioned.

b) Experimental setup: In all of the experiments, we choose $A = id$ and in particular $m = n$. $Q = U$ is chosen as a properly renormalized standard Gaussian matrix. We try to solve the blind deconvolution problem for different values of $s, \sigma$ and $\mu$. The values of the sparsity parameters range $\sigma = 5, 10, 15$ and $s = 1, 2, \ldots, 7$. We test three values for $n (n = 50, n = 170$ and $n = 350)$ and let $\mu$ be $10, 20, \ldots, 250$. For each quadruple $(n, \mu, s, \sigma)$ we draw 100 sparse instances of $b$ and $h$ by choosing a sparse support uniformly at random and filling the non-zero entries with independent normally distributed values; we ‘measure’ them with $H$, and try to recover it with the HiHTP algorithm. We halt the algorithm once the difference in Frobenius norm of consecutive approximations drops below 1e-6 or after 25 iterations. We then solve the final restricted least squares problem with a lower residual tolerance (1e-6.5). A success is declared when the final relative error between the approximation and the original value for $h \otimes b$ in Frobenius norm is smaller than 1e-6.

The results are depicted in Figure 2.

c) Results: By a simple visual inspection, we see that the number of measurements needed for successful recovery seem to scale linearly with $s$ across all sparsities and values for $n$. We furthermore see that the dependence on $n$ is relatively mild. Hence, the results suggest that the log($s$)-terms in our complexity bounds are proof artefacts.

Although not crucial – $s\sigma$ will for most parameter values be the dominating term in our bound anyhow – let us also try to test the hypothesis that some terms of our sample complexity bound are spurious a bit
more thoroughly. For a tuple of integers \( a = (a_0, a_1, a_2) \) and constant \( C > 0 \), let us define the parameter
\[
\lambda_{a,C} = \frac{\mu}{s \log(\mu)^{a_0} \log(\mu)^{a_1} \log(s)^{a_2} + Cs\sigma \log(n)}.
\]
The rationale for defining these parameters is clear: \( \lambda_{a,C} \) measures the ‘oversampling factor’ compared to a sample complexity \( \beta_a(s, \sigma, \mu, n) + Cs\sigma \log(n) \), where
\[
\beta_a(s, \sigma, \mu, n) = s \log(\mu)^{a_0} \log(\mu)^{a_1} \log(s)^{a_2}\sigma \log(n) + Cs\sigma \log(n).
\]
Using the \texttt{sklearn} package \cite{scikit-learn}, we now perform logistic regressions of the empirical success probabilities of our data against \( \lambda_{a,C} \) for a range of values of \( a \) and \( C \). The resulting accuracies, when the optimal \( C_a \) for \( C \) is used for each \( a \) are presented in Table I. We see that when using \( a = (0, 1, 0) \), we obtain the highest accuracy. In particular, parameter values with \( a_2 \neq 0 \) lead to worse predictions. Hence, according to our data, the oversampling factor compared to
\[
s \log(\mu n) + C_a s\sigma \log(n)
\]
is the best predictor of the empirical success of the HiHTP algorithm, suggesting that the bound \( \beta_{0,1,0} \) better indicates the regime of success of HiHTP. The differences are small, so one should be cautious read too much into this. However, it is clear that the predictors including additional \( \log(s) \)-terms in the complexity perform worse. This points to them being proof artefacts more than anything else in the bound of Theorem II.1. In Figure 3 the optimal predictor \( \lambda_{(0,1,0)} \) is compared with \( \lambda_{(1,1,2)} \), which is the one suggested by Theorem II.1. The figures clearly suggest that the former fits much better to the data.

\textbf{Acknowledgement}

The authors wish to thank the anonymous reviewers of a previous version of the manuscript for helpful comments and suggestions, leading to a significant improvement of the results. AF acknowledges support from the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation, and CHAIR. GW is supported by the German Science Foundation (DFG) under grants 598/7-1, 598/7-2, 598/8-1, 598/8-2 and the 6G research cluster (6g-ric.de) supported by the German Ministry of Education and Research (BMBF).

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|c|}
\hline
\( \beta_a \) & Accuracy & \( \beta_3 \) & Accuracy & \( \beta_4 \) & Accuracy \\
\hline
\text{ } & \text{ } & \text{ } & \text{ } & \text{ } & \text{ } \\
\text{ } & \text{ } & \text{ } & \text{ } & \text{ } & \text{ } \\
\text{s} & 91.8\% & \text{s log(s)} & 91.2\% & \text{s log(s)}^2 & 91.1\% \\
\text{s log(\mu)} & 91.7\% & \text{s log(\mu) log(s)} & 91.2\% & \text{s log(\mu) log(s)}^2 & 91.1\% \\
\text{s log(\mu n)} & 92.1\% & \text{s log(\mu n) log(s)} & 91.2\% & \text{s log(\mu n) log(s)}^2 & 91.1\% \\
\text{s log(\mu n) log(\mu)} & 91.7\% & \text{s log(\mu n) log(\mu) log(s)} & 90.9\% & \text{s log(\mu n) log(\mu) log(s)}^2 & 90.3\% \\
\hline
\end{tabular}
\caption{Accuracy of the predictions resulting from the logistic regressions for different values of \( a \).}
\end{table}
[34] E. J. Candès, “The restricted isometry property and its implications for compressed sensing,” *Comptes Rendus Mathematique*, vol. 346, no. 9, pp. 589–592, 2008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1631073X08000964

[35] R. Vershynin, *High-Dimensional Probability. An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

[36] F. Krahmer, S. Mendelson, and H. Rauhut, “Suprema of chaos processes and the restricted isometry property,” *Communications on Pure and Applied Mathematics*, vol. 67, pp. 1877–1904, 2014. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.21504

[37] M. Rudelson and R. Vershynin, “On sparse reconstruction from Fourier and Gaussian measurements,” *Communications on Pure and Applied Mathematics*, vol. 61, no. 8, pp. 1025–1045, 2008.

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

[39] M. Talagrand, *The generic chaining. Upper and Lower Bounds of Stochastic Processes*. Springer, 2005.

[40] H. Rauhut, J. Romberg, and J. A. Tropp, “Restricted isometries for partial random circulant matrices.” *Applied and Computational Harmonic Analysis*, vol. 32, pp. 242–254, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1063520311000649

[41] S. Dirksen, “Tail bounds via generic chaining,” *Electronic Journal of Probability*, vol. 20, pp. 1–29, 2015.

[42] A. Böttcher and B. Silbermann, *Introduction to Large Truncated Toeplitz Matrices*. New York, NY: Springer New York, 1999. [Online]. Available: https://doi.org/10.1007/978-1-4612-1426-7_6

[43] R. Okuta, Y. Unno, D. Nishino, S. Hido, and C. Loomis, “Cupy: A numpy-compatible library for nvidia gpu calculations,” in *Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Thirty-first Annual Conference on Neural Information Processing Systems (NIPS)*, 2017. [Online]. Available: http://learingsys.org/nips17/assets/papers/paper_16.pdf

[44] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.