Compressed Sensing and $\Sigma\Delta$-Quantization

by

Joe-Mei Feng

Ph.D. Thesis
Georg-August-Universität Göttingen

Advisor: Prof. Dr. Felix Krahmer

December 2017

To my family.

All praise and glory to the Lord.

Acknowledgements

First of all, I would like to thank my advisor Prof. Dr. Felix Krahmer, and co-advisor Prof. Dr. Gerlind Plonka-Hoch for their time and support. Especially thanks to Felix. He is willing to spend his time with us and always ready to help, and is kind to us all the time. (He duzen with us.) I can always expect to broaden my mathematical thoughts in our meetings and discussions, and he is generous to share his ideas and intriguing us by asking questions (Socratic method :)). I also thank our collaborator Prof. Dr. Rayan Saab, for spending time and providing useful suggestions in our meetings. I would also like to thank Dr. Rongrong Wang for her time to answer my questions. And thank Dr. Chia-han Lee from Academia Sinica. It was in his lab, that I first knew about compressed sensing.

As for all my colleagues from TUM, Göttingen and Academia Sinica, especially Christian Kümmerle, Markus Hansen, Dominik Stöger, Sara Krause-Solberg from M15 TUM, and Florian Bossmann from Uni. Göttingen, and Yen-Huan Li from Academia Sinica, I appreciate so much all the interaction with you both on math and life.

Thank my church. Without your help and prayer, it is not possible for me to accomplish my study. Thank my husband Qian, thank my parents-in-law, thank my parents, thank my daughter, Lea. Last but not least, thank you, my Lord, Jesus.

Contents

Chapter 1

Background

1.1 Compressed sensing

Traditionally, given a linear system

$$y = Ax, \tag{1.1}$$

for $y \in \mathbb{C}^m$, $A \in \mathbb{C}^{m \times N}$, it requires the dimension m, N of A satisfy $m \geq N$ to guarantee the uniqueness of the recovery.

From empirical observation one obtains that various types of signals admit sparse representation with respect to certain bases or frames. Which means, comparing to how much information the dimension can be loaded, these signals carry in fact only few information farer away than that. In this situation, can we recover them also from measurements less than the system can carry, i.e., in mathematical expression, $m << N$? Sparse recovery has already long history. This problem is nowadays called compressed sensing (CS).

Candès, Romberg, Tao [12], and Donoho [21] first combined the ideas of linear program, or ℓ_1-minimization with a random choice of compressed sensing matrices [26]. Recovering the sparse signal by solving a linear program is called basis pursuit.

Compressed sensing [12,14,21] deals with reconstructing (approximately) sparse vectors $x \in \mathbb{R}^N$ from significantly few measurements generated linearly from x by the form $(\langle a_i, x \rangle)_{i=1}^m$ with vectors $a_i \in \mathbb{R}^N$ and $m < N$. Exact recovery is theoretical possible, due to the low information carried by the original

signal and its "oversampled" measurements. In contrast to its linear structure between the signal and its measurements, the recovery is done non-linearly by such as a convex optimization problem or a greedy numerical algorithm (e.g., [6, 16, 23, 41]).

Given a measurement matrix A with rows $(A_i)_{i=1}^m \in \mathbb{R}^N$ well-chosen vectors, the measurement noise denoted by $(e_i)_{i=1}^m$, m measurements $(y_i = \langle A_i, x\rangle + e_i)_{i=1}^m$ and $\hat{x}$ recovered from ℓ_1-minimization problem, then the recovery of the standard compressed sensing problem

$$y = Ax + e, \tag{1.2}$$

will have a guaranteed result (e.g., [12, 14, 21], see also [26]) that the solution $\hat{x}$ to the optimization problem

$$\min_z \ \|z\|_1 \text{ subject to } \|Az - y\|_2 < \epsilon, \tag{1.3}$$

can be bounded meaningfully from above, see an example below. Denote the set of sparse signals with unit length in ℓ_2, $D_{s,N} := \{x \in \mathbb{R}^N, \ \text{supp}(x) \leq s, \ \|x\|_2 = 1\}$, define $\sigma_s(x)_p = \min_{v \in D_{s,N}} \|x - v\|_p$ the best s-term approximation error of x in ℓ_p, which is a function that measures how close x is to being s-sparse. [12, 21] show that for a wide class of random matrices the solution $\hat{x}$ to (1.3) satisfies

$$\|x - \hat{x}\|_2 \leq C_1\big(\frac{\epsilon}{\sqrt{m}} + \frac{\sigma_s(x)_1}{\sqrt{s}}\big), \tag{1.4}$$

when $m \geq C_2 s \log(N/k)$, for some positive constants C_1 and C_2. Note that which implies directly that in noise-free scenario, an s-sparse signal x can be uniquely determined.

In the following section, several popular criteria for unique recovery of compressed sensing are introduced.

Sparsity and criteria for reconstruction

We refer the reader to [26] for further information of this chapter. When $m << N$, under certain criteria (for example, null space property), the ℓ_1-minimization problem (1.3) recovers a sparse signal from (1.1) guarantees the reconstruction (when the error is under control). There are three scenarios to

be considered and therefore three varied types of null space property are required respectively to control the reconstruction error. They will be briefly listed below, and will be introduced more precisely in the following sections.

1 If x is sparse and no noise exists, then the null space property guarantees recovery.

2 If x is approximately sparse and no noise exists, then the stable null space space property guarantees approximate recovery.

3 If x is approximately sparse and measurement noise exists, then robust null space property guarantees approximate recovery.

Note that robust null space property implies stable null space property implies robust null space property.

Null Space Property

Definition 1. *[26] A matrix $A \in \mathbb{C}^{m\times N}$ is said to satisfy the null space property relative to a set $S \subset [N]$ if*

$$\|v_S\|_1 < \|v_{\bar{S}}\|_1 \text{ for all } v \in \ker A \setminus \{0\}. \tag{1.5}$$

It is said to satisfy the null space property of order s if it satisfies the null space property relative to any set $S \subset [N]$ with card $S \leq s$.

Theorem 1. *[26] Given a matrix $A \in \mathbb{C}^{m\times N}$, every s-sparse vector is the unique solution of (1.3) with $\epsilon = 0$ if and only if A satisfies the null space property to the set S.*

Proof. Given $v \in \ker A$. Since the theorem is for all y, to verify the theorem, for any support set S with cardinality s the problem (1.3) with $y = Av_S$ and $\epsilon = 0$. Since

$$\begin{aligned} &Av = A(v_S + v_{\bar{S}}) = 0 \\ \Rightarrow &A(v_S + v_{\bar{S}}) = A(v_S) - A(-v_{\bar{S}}) = 0 \\ \Rightarrow &A(v_S) = A(-v_{\bar{S}}) \end{aligned}$$

By assumption v_S is supported on S thus the unique solution of 1.3, thus $\|v_S\|_1 < \|v_{\bar{S}}\|_1$. Conversely, given $x \in \mathbb{C}^N$ a solutions to (1.3), if there is z, which is also a solution to (1.3), (z can be either s-sparse or not.) Denote support set of x, S_x respectively, then,

$$\begin{aligned}
\|x\|_1 &\leq \|x_{S_x} - z_{S_x}\|_1 + \|z_{S_x}\|_1 \\
&< \|(x-z)_{\bar{S_x}}\|_1 + \|z_{S_x}\|_1 \\
&= \|z_{\bar{S_x}}\|_1 + \|z_{S_x}\|_1 \\
&= \|z\|_1,
\end{aligned}$$

which constricts the assumption that both of them are minimizer of 1.3. Therefore the solution to 1.3 is unique. □

Stable Null Space Property

In this chapter a criteria stronger than null space property will be applied for the signal x is now only approximately sparse, then we have

Theorem 2. *[26] For any $1 > p > 0$ and any $x \in \mathbb{C}^N$,*

$$\sigma_s(x)_q \leq \frac{1}{s^{1/p-1/q}} \tag{1.6}$$

Proof. Without loss of generality we can rearrange x_j according to its length (ℓ_1-norm) in nonincreasing order and assume $|x_i| \leq 0$ for all $i = 1, \ldots, N$. Then

$$\begin{aligned}
\sigma_s(x)_q^q &= \sum_{j=s+1}^{N} (|x_j|)^q \\
&\leq (|x_s|)^{q-p} \sum_{j=s+1}^{N} (|x_j|)^p \\
&\leq \big(\frac{1}{s}\sum_{j=1}^{s} \big((|x_j|)^p\big)^{\frac{q-p}{p}}\big)\big(\sum_{j=s+1}^{N} (|x_j|)^p\big) \\
&\leq \big(\frac{1}{s}\|x\|_p^p\big)^{\frac{q-p}{p}}\big)\big(\|x\|_p^p\big) \\
&= \frac{1}{s^{q/p-1}}\|x\|_p^q
\end{aligned}$$

□

A tighter bound to Theorem 2 is:

Theorem 3. *[26] For any $q > p > 0$ and any $x \in \mathbb{C}^N$, the inequality*

$$\sigma_s(x)_q \leq \frac{c_{p,q}}{s^{1/p-1/q}} \|x\|_p \tag{1.7}$$

holds with

$$c_{p,q} := \left[\left(\frac{p}{q}\right)^{p/q} \left(1 - \frac{p}{q}\right)^{1-p/q} \right]^{1/p} \leq 1. \tag{1.8}$$

Proof. Again, following similar steps as in Theorem 2, without loss of generality, given signal $x = (x_j)_{j=1}^N$ nonincreasing rearranged according to length of $x_j's$. □

Robust Null Space Property

If there exists measurement noise, i.e., ϵ in (1.3) is not always 0, then define the criterion robust null space property as following.

Definition 2. *[26] The matrix $A \in \mathbb{C}^{m \times N}$ is said to satisfy the robust null space property (with respect to $\|\cdot\|$) with constants $0 < \rho < 1$ and $\tau > 0$ if for any set $S \subset [N]$ with card$(S) \leq s$ if*

$$\|v_S\|_1 \leq \rho \|v_{\bar{S}}\|_1 + \tau \|Av\| \text{ for all } v \in \mathbb{C}^N. \tag{1.9}$$

Note that in the definition v doesn't need to be in *ker* A.

Theorem 4. *[26] The matrix $A \in \mathbb{C}^{m \times N}$ satisfies the robust null space property with constants $0 < \rho < 1$ and $\tau > 0$ of order s if and only if for any S with $|S| \leq s$*

$$\|z - x\|_1 \leq \frac{1+\rho}{1-\rho}(\|z\|_1 - \|x\|_1 + 2\|x_{\bar{S}}\|_1) + \frac{2\tau}{1-\rho}\|A(z-x)\| \tag{1.10}$$

Further the ℓ_q-robust null space property defined as

Definition 3. *[26] The matrix $A \in \mathbb{C}^{m\times N}$ is said to satisfy the ℓ_q-robust null space property of order s (with respect to $\|\cdot\|$) with constants $0 < \rho < 1$ and $\tau > 0$ if for any set $S \subset [N]$ with card$(S) \leq s$ if*

$$\|v_S\|_q \leq \frac{\rho}{s^{1-1/q}}\|v_{\bar{S}}\|_1 + \tau\|Av\| \text{ for all } v \in \mathbb{C}^N. \tag{1.11}$$

Theorem 5. *[26] Given $1 \leq p \leq q$, suppose that the matrix $A \in \mathbb{C}^{m\times N}$ satisfies the ℓ_q-robust null space property of order s with constants $0 < \rho < 1$ and $\tau > 0$. Then, for any $x, z \in \mathbb{C}^N$,*

$$\|z - x\|_p \leq \frac{C}{s^{1-1/p}}\big(\|z\|_1 - \|x\|_1 + 2\sigma_s(x)_1\big) + Ds^{1/p-1/q}\|A(z-x)\|, \tag{1.12}$$

where $C := (1+\rho)^2/(1-\rho)$ and $D := (3+\rho)\tau/(1-\rho)$.

Restricted Isometry Property

The null space property is not easy to be proved diretly, therefore restricted isometry property (RIP) is used as the most popular criterion in the CS regime since first introduced in [13]. Plenty of papers focus on proving the RIP of different types of matrices such as Gaussian random matrices [2], subgaussian random matrices [26], partial random discrete Fourier matrices [46], In this thesis we will use our new method as another approach to prove the RIP of partial random discrete Fourier matrices.

Definition 4. *[13] The restricted isometry property of order s with constant, called restricted isometry constant, $\delta_s = \delta_s(A)$ of a matrix $A \in \mathbb{C}^{m\times N}$ is the smallest $\delta \geq 0$ such that*

$$(1-\delta)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1+\delta)\|x\|_2^2 \tag{1.13}$$

for all s-sparse vectors $x \in \mathbb{C}^N$.

Checking the restricted isometry property is in general an NP hard problem [53], and deterministic matrices with guaranteed restricted isometry property are known for relative large embedding dimensions (e.g., [20]). Therefore many papers on CS work with random matrices. Random matrices such as subgaussian matrices [2], partial random circulant matrices [39], and partial random Fourier matrices [46] are known to have the restricted isometry property for large enough embedding dimension with high

probability. Examples of subgaussian matrices include Gaussian and Bernoulli. Such matrices are shown to have the restricted isometry property provided $m = \Omega(s \log(eN/s))$ (e.g. [2]). This order of the embedding dimension m is known to be optimal [47].

Define $D_{s,N} := \{x \in \mathbb{R}^N : \ \|x\|_2 = 1 \ and \ |supp(x)| \leq s\}$, equivalently,

$$\delta_s = \sup_{x \in D_{s,N}} \Big| \frac{\|Ax\|_2^2 - \|x\|_2^2}{\|x\|_2^2} \Big| = \sup_{x \in D_{s,N}} \big| \|Ax\|_2^2 - 1 \big|. \tag{1.14}$$

Since ℓ_2-robust null space property implies robust null space property implies stable null space property implies null space property and of purpose of this thesis, only that the restricted isometry property implies robust null space property will be shown. In the following theorem, the restricted isometry property is shown to imply robust null space property.

Theorem 6. *[25] Given compressed sensing matrix $A \in \mathbb{C}^{m \times N}$ having restricted isometry property with constant $\delta_{2s} \leq 1/9$ then the matrix A satisfies the ℓ_2-robust null space property of order s relative to the ℓ_2-norm on $\mathbb{C}^m$ and with constants $0 < \rho < 1$ and $\tau > 0$ depending only on δ_{2s}.*

Proof. Let $v \in \mathbb{C}^m$, and let $S = S_0$ denote an index set of s largest absolute entries of v and further S_1 of next s largest absolute entries, etc. By similar argument as in (2)

$$\|v_{S_k}\|_2 \leq \frac{1}{\sqrt{s}} \|v_{S_{k-1}}\|_1, \ for \ all \ k \geq 1, \tag{1.15}$$

so that a summation gives

$$\sum_{k \geq 1} \|v_{S_k}\|_2 \geq \frac{1}{\sqrt{s}} \|v\|_1. \tag{1.16}$$

By assumption of restricted isometry property

$$
\begin{aligned}
\|v_S\|_2^2 = \|v_{S_0}\|_2^2 &\leq \frac{1}{1-\delta_s}\|Av_{S_0}\|_2^2 \\
&= \frac{1}{1-\delta_s}\langle AV_{S_0}, Av - Av_{\bar{S_0}}\rangle \\
&= \frac{1}{1-\delta_s}\langle Av_{S_0}, Av\rangle - \frac{1}{1-\delta_s}\langle Av_{S_0}, \sum_{k\geq 1} Av_{\bar{S_k}}\rangle \\
&= \frac{1}{1-\delta_s}\|Av_{S_0}\|_2\|Av\|_2 - \frac{1}{1-\delta_s}\sum_{k\geq 1}\langle Av_{S_0}, Av_{S_k}\rangle > \\
&\leq \frac{\sqrt{1+\delta}}{1-\delta_s}\|v_{S_0}\|_2\|Av\|_2 - \frac{1}{1-\delta_s}\sum_{k\geq 1}\langle A(v_{S_0})_{S_0\cup S_k}, A(v_{S_k})_{S_0\cup S_k}\rangle \\
&= \frac{\sqrt{1+\delta_s}}{1-\delta_s}\|v_{S_0}\|_2\|Av\|_2 - \frac{1}{1-\delta_s}\sum_{k\geq 1}(\langle A(v_{S_0})_{S_0\cup S_k}, A(v_{S_k})_{S_0\cup S_k}\rangle - \langle v_{S_0}, v_{S_k}\rangle) \\
&= \frac{\sqrt{1+\delta_s}}{1-\delta_s}\|v_{S_0}\|_2\|Av\|_2 - \frac{1}{1-\delta_s}\sum_{k\geq 1}\langle (A_s^*A_s - Id)(v_{s_0})_{S_0\cup S_k}, (v_{s_k})_{S_0\cup S_k}\rangle \\
&\leq \frac{\sqrt{1+\delta}}{1-\delta_s}\|v_{S_0}\|_2\|Av\|_2 + \frac{1}{1-\delta_s}\sum_{k\geq 1}\|(A_s^*A_s - Id)v_{S_0}\|_2\|v_{S_k}\|_2 \\
&\leq \frac{\sqrt{1+\delta_s}}{1-\delta_s}\|v_{S_0}\|_2\|Av\|_2 + \frac{1}{1-\delta_s}\sum_{k\geq 1}\delta_{2s}\|v_{S_0}\|_2\|v_{S_k}\|_2 \\
&\leq \frac{\sqrt{1+\delta_s}}{1-\delta_s}\|v_{S_0}\|_2\|Av\|_2 + \frac{1}{1-\delta_s}\sum_{k\geq 1}\delta_{2s}\|v_{S_0}\|_2\|v_{S_k}\|_2 \\
&= \frac{\sqrt{1+\delta_s}}{1-\delta_s}\|v_{S_0}\|_2\|Av\|_2 + \frac{\delta_{2s}}{1-\delta_s}\|v_{S_0}\|_2\sum_{k\geq 1}\|v_{S_k}\|_2 \\
&\leq \frac{\sqrt{1+\delta_s}}{1-\delta_s}\|v_{S_0}\|_2\|Av\|_2 + \frac{\delta_{2s}}{1-\delta_s}\|v_{S_0}\|_2\sum_{k\geq 1}\|v_{S_k}\|_1 \\
&\leq \frac{\sqrt{1+\delta_s}}{1-\delta_s}\|v_{S_0}\|_2\|Av\|_2 + \frac{\delta_{2s}}{1-\delta_s}\|v_{S_0}\|_2\|v_{\bar{S}}\|_1,
\end{aligned}
$$

cancel both side by $\|v_{S_0}\|_2$, which equals $\|v_S\|_2$

$$
\|v_S\|_2 \leq \frac{\sqrt{1+\delta_s}}{1-\delta_s}\|Av\|_2 + \frac{\delta_{2s}}{1-\delta_s}\|v_{\bar{S}}\|_1. \tag{1.17}
$$

Since $\delta_s \leq \delta_{2s} \leq \frac{1}{9}$, $\frac{\sqrt{1+\delta_s}}{1-\delta_s} > 0$ and $0 < \frac{\delta_{2s}}{1-\delta_s} < 1$ which ends the proof by setting $\rho = \frac{\delta_{2s}}{1-\delta_s}$, $\tau = \frac{\sqrt{1+\delta_s}}{1-\delta_s}$ in (2). □

1.2 Quantization

Given an analogue signal, one needs to transform the signal into finitely many digits to make the digital transmission and storage possible. How can one represent it by finitely many digits? First thanks to the research result from physicist Nyquist, the Nyquist rate ensures that, with sampling rate twice faster than the largest signal frequency, one can exactly recover the signal without any loss. This means a continuous signal can be represented by these discrete values (called measurements). Since this works for signals, to which the frequency is bounded (otherwise there is no such "largest" frequency). This is an important result of sampling theory in Fourier analysis, and is uniquely determined by these measurements. Precisely, the signal is proved to be able to be represented by a linear expansion of the measurements with respect to a basis formed by sinc functions (this led to research on wavelets).

Sampling indeed discretizes the signal, however these sampled data can be irrational numbers, which cannot be represented by finite digits. Quantization is the technique to represent these data by finitely many digits (rational numbers). This "transform" is called "quantization" (or modulation).

Quantization consists of two steps, sampling and representing by finitely many symbols from a finite alphabet. An alphabet is a finite set of numbers. The most natural and usual choices of alphabets have equispaced elements, as for example we will focus in the so-called mid-rise alphabet with $2L$ levels and step-size Δ, denoted by $\mathcal{A}_L^\Delta$ and given by $\mathcal{A} := \Delta\mathbb{Z} + i\Delta\mathbb{Z}$. The extreme case of such an alphabet is the 1-bit quantization alphabet, which we denote by $\mathcal{A} = \{-1, +1\}$. It is called one-bit quantization, because each element is represented by one-bit digit.

1.2.1 Memoryless scalar quantization

Memoryless scalar quantization quantizes each component independently. Intuitively one might use it to quantize (compressed sensing) measurements.

As an simplest example of MSQ, pulse code modulation (PCM) uses a scalar quantizer

$$\begin{aligned} Q_{\mathcal{A}} : \mathbb{C} &\to \mathcal{A} \\ z &\mapsto \arg\min_{v\in\mathcal{A}} |z - v| \end{aligned} \tag{1.18}$$

to quantize every entry of a vector y independently.

However it has its drawback in context of compressed sensing. This will be discuss further in Chapter 1.3.1, therefore we work in this thesis with another structured quantization method, $\Sigma\Delta$-quantization.

1.2.2 $\Sigma\Delta$-quantization

In [29] Güntürk showed that even with extreme coarse case (one-bit), one-bit well-designed $\Sigma\Delta$-quantization can reach a reconstruction error decays exponentially in $\lambda = \frac{s}{m}$ as $O(2^{-0.07\lambda})$, while the expected bound is $O(2^{-\lambda})$, where $1/\lambda$ is the sampling rate. In [19] Deift et al. designed a family of $\Sigma\Delta$-quantizationby using rth order greedy $\Sigma\Delta$-quantization together with a feedback filter, and then they improved the bound from $O(2^{-0.07\lambda})$ to $O(2^{-0.102\lambda})$. This bound is further improved in [18], in which a near-optimal coefficient comparing to the optimal result in [38]. In [38], the first lower bound for one-bit $\Sigma\Delta$-quantization is provided, which says for K-bit quantization the lower bound is bounded by $O(2^{-K/\lambda})$.

As an introduction to $\Sigma\Delta$-quantization, see an example on the first order greedy $\Sigma\Delta$-quantization, which runs the following iteration:

$$\begin{aligned} q_i &= \mathcal{Q}_{\mathcal{A}}(y_i + u_{i-1}) \\ u_i &= u_{i-1} + y_i - q_i, \end{aligned} \tag{1.19}$$

where $\mathcal{Q}_{\mathcal{A}}$ as defined in (1.18).

Generally, an r^{th}-order $\Sigma\Delta$-quantization with quantization rule $\rho : \mathbb{R}^{r+1} \to \mathbb{R}$ iterates

$$\begin{aligned} q_i &= \mathcal{Q}_{\mathcal{A}}(\rho(y_i, u_{i-1}, u_{i-2}, \ldots, u_{i-r})), \\ u_i &= y_i - q_i - \textstyle\sum_{j=1}^{r} \binom{r}{j} (-1)^j u_{i-j}, \end{aligned} \tag{1.20}$$

for some quantization rule ρ.

We say a $\Sigma\Delta$-quantization is stable, if for all $m \in \mathbb{N}$, and for all $y \in \mathbb{R}^m$ with $\|q\|_\infty$ bounded from above, or equivalently if the quantization rule ρ and property of y in recursion (1.20) imply that u in (1.25) is bounded from above by an absolute constant which depends only on the order r in the form

$$\|u\|_\infty \leq \gamma(r). \tag{1.21}$$

As an example of stable r-th order greedy $\Sigma\Delta$-quantization, which is also the $\Sigma\Delta$-quantization used through this thesis, is introduced below in Chapter 1.2.3.

1.2.3 The r-th order greedy $\Sigma\Delta$-quantization

The rth order greedy $\Sigma\Delta$-quantization is defined as following.

$$\begin{aligned} q_i &= \mathcal{Q}_{\mathcal{A}}(\rho(y_i, u_{i-1}, u_{i-2}, \ldots, u_{i-r})), \\ u_i &= y_i - q_i - \textstyle\sum_{j=1}^{r} \binom{r}{j}(-1)^j u_{i-j}, \end{aligned} \tag{1.22}$$

where

$$\rho(y_i, u_{i-1}, u_{i-2}, \ldots, u_{i-r}) = \sum_{j=1}^{r} \binom{r}{j} u_{n-j} + y_n. \tag{1.23}$$

Using the first-order difference $m \times m$ matrix D with entries given by

$$D_{i,j} := \begin{cases} 1 & \text{if } i = j \\ -1 & \text{if } i = j+1 \\ 0 & \text{otherwise} \end{cases}, \tag{1.24}$$

the relationship between x, u, and q can be concisely written in matrix-vector notation as

$$D^r u \;=\; y - q. \tag{1.25}$$

Since y is bounded, if the alphabet $\mathcal{A} = \Delta\mathbb{Z}$, for some (small) quantity Δ, $|u_i| \leq \frac{\Delta}{2}$, then $\|q\|_\infty$ is bounded, i.e., this quantization is stable. And further, $\|u\|_2 \leq \sqrt{m}\frac{\Delta}{2}$ and

$$\begin{aligned} |y_i - q_i| &= |\sum_{j=1}^{r} \binom{r}{j}(-1)^j u_{i-j} + u_i| \\ &= |\sum_{j=0}^{r} \binom{r}{j}(-1)^j u_{i-j}| \leq |\sum_{j=0}^{r} \binom{r}{j} u_{i-j}| \\ &\leq \frac{\Delta}{2}|\sum_{j=0}^{r} \binom{r}{j}| = \frac{\Delta}{2}2^r = 2^{r-1}\Delta, \end{aligned}$$

q_i is within the range $[\pm(2^{r-1}\Delta + \|y\|_\infty)]$. This can be generalized similarly to an alphabet $\mathcal{A} = \Delta Z + i\Delta Z$.

Daubechies et al. [17] shows that for band-limited functions, reconstruction error from rth order greedy $\Sigma\Delta$ quantized measurements is bounded by $O(\lambda^{-k})$.

1.3 Compressed sensing and quantization

Only concerning about compressed sensing without quantization is actually not practical, since nowadays the process of this technique is done on computer. For transmission and storage of the data, one must represent the data by only finitely many digits.

Signal recovery from quantized compressed sensing measurements is the main topic of this thesis. More precisely, we estimate the error bound of the signal recovery , with a two-stage signal process, first by compressed sensing and then quantization.

There are plenty of works focusing on the first stage of signal process, i.e., the compressed sensing introduced in chapter 1.1. The second stage, the quantization, which was first applied on frame structure was introduced in chapter 1.2.

In this chapter we recap the papers which worked on quantization on compressed sensing.

In the compressed sensing context, quantization is the map that replaces the vector $y = Ax + e \in \mathbb{C}^m$ by a representation that uses a finite number of bits as of the form

$$\mathcal{Q} : \mathbb{C}^m \to \mathcal{A}^m,$$

where $\mathcal{A} \subset \mathbb{C}$ is a finite set, called the quantization alphabet. Both memoryless scalar quantization and $\Sigma\Delta$-quantization discussed in this thesis use this quantization map.

As the reconstruction, Zymnis et al. in [56] provided two decoder based on maximum likelihood and least square respectively. The performance is shown numerically. A more "compressed sensing-like" reconstruction was proposed in [34] by Jacques. Instead of reconstruct the quantized measurements by standard ℓ_1 minimization with ℓ_2-norm constraint on the noise (called Basis Pursuit DeNoise (BPDN)). Jacques et al. proposed Basis Pursuit DeQuantizer of moment p (BPDQ$_p$) as a decoder for general quantization. BPDQ$_p$ is also an ℓ_1-minimization optimization problem, but with ℓ_p-norm constraint on noise. It is showed that the reconstruction error is bounded if the compressed sensing matrix satisfies the

restricted isometry property of ℓ_p-norm. And for Gaussian compressed sensing matrix the reconstruction error outperforms that reconstructed via BPDN by dividing its error bound by $\sqrt{p+1}$. Methods proposed in these two papers are designed for arbitrary quantizer. In contrast to it, [48] provided the decoder designed specifically for rth order greedy $\Sigma\Delta$-quantization. This is also the decoder we used in our research, see Chapter 4.

Considering about the quantizer, in [51], it concerns about designing an optimal quantizer.

In [10], it is the first time that the one bit compressed sensing is considered, and instead of recover the signal by traditional basis pursuit and treating the one-bit quantized Gaussian measurements simply as ± 1, Boufounos et al. treating the measurements as sign constraints and then solving an optimization problem (this is however non-convex) on a unit sphere, and it was shown numerically to outperform the traditional reconstruction stably and robustly.

In contrast to [10], a sub-Gaussian compressed sensing matrix is used in [1]. Also by treating the quantized measurements as signs and reconstruct the signal by an convex optimization problem, one achieves the error bound by $O((s\log(N/s)/m)^{1/4})$.

There can be a variety of choices as the decoder for reconstructing the signal from its one-bit measurements. In [42], Plan et al. showed that the reconstruction for accurately recovering of an s-sparse signal can be achieved by simply solving a linear program. It is in [50], a reconstruction from memoryless one-bit measurements of a structured compressed sensing matrix is analysed. It is shown that with number of measurements $m \sim \epsilon^{-4} s\log(N/s\epsilon)$, any s-sparse can be recovered with error ϵ. This result holds due to the ℓ_1/ℓ_2-restricted isometry property of the circulant Gaussian matrices.

Beside the discussion on one-bit quantization (either memoryless quantization, $\Sigma\Delta$-quantization, or any other quantization schemes), [24, 28, 37, 49], use a more generalized alphabet, with Gaussian or sub-Gaussian compressed sensing matrix. And it is worthwhile to note that in these works the compressed sensing matrices are within the range of sub-Gaussian matrices or directly assuming the restricted isometry property of the compressed sensing matrix. And signal recovery from quanized measurements of a structured compressed sensing matrix is discussed in [32, 50, 55]. No paper prior to [32] analysed quantization on structured compressed sensing matrices, such as discrete Fourier matrix or partial random circulant matrices. More details are provided in Chapter 1.3.2.

1.3.1 MSQ on CS

MSQ is a nice choice to analyse the effect and performance of quantization on compressed sensing, due to its simplicity. The first paper with MSQ on compressed sensing is [10]. It has been however shown that, MSQ has its theoretical limit in the context of compressed sensing. What does it mean? Let us consider how we can do to improve the error bound in practical? First, we can increase the number of the measurements. Or we can use a finer alphabet.

And indeed if treating the measurements after MSQ quantization as noise in compressed sensing problem, and then reconstruct the problem by standard decoder such as (1.3), it is true that one can have smaller noise if a finer alphabet is used, which means by the reconstruction guarantee (1.4) a smaller reconstruction error bound. This is somehow not meaningful for a fixed quantizer, therefore which is outside the discussion about improving the reconstruction error bound in context of quantization on compressed sensing. The meaningful pursuit for better bound is then to increase the number of the measurements.

In fact it has been shown that MSQ is not an efficient quantizer for compressed sensing [9, 27], in [35] that the error in reconstructing sparse signals from 1-bit quantized measurement is bounded by

$$O(s \log(N/s)/m).$$

In fact shown by Goyal, Vetterli, and Thao [27], even if the support set is given, the reconstruction error of MSQ cannot be better than

$$\Omega(\frac{s}{m} \log(N/s)).$$

This bottleneck of MSQ in context of compressed sensing comes from the fact that each measurement is mapped individually, which, however, doesn't benefit from the structure of compressed sensing, i.e., the nicely redundant linear structure. While quantizing each of the measurements, the correlation between measurements is totally ignored.

On the other hand, in $\Sigma\Delta$-quantization schemes, each measurement is quantized by taking the previous quantization steps into account. Although it is not designed based on the correlation of the mea-

surements, one is correcting in each iteration the quantization error from previous iterations, therefore $\Sigma\Delta$-quantization outperforms MSQ. In [51], a quantization scheme is created specifically for compressed sensing measurements. This scheme is however not so commonly used in practice. Therefore in this thesis we still work with $\Sigma\Delta$-quantization.

1.3.2 $\Sigma\Delta$-quantization on CS

A compressed sensing problem has its hidden structure as a finite-frame expansion once the support set is determined.

Besides directly analysing the reconstruction error from $\Sigma\Delta$-quantized compressed sensing measurements, papers [3–5, 7, 8, 15, 36, 37, 43, 55] also work with $\Sigma\Delta$-quantizationon finite-frame expansions. One can view in this context that the sparsity s of x as the dimension of x, i.e., $s = N$ here.

In [4], signal recovery from first order $\Sigma\Delta$ quantized frame expansions under different frames are analysed by Benedetto et al. It is shown that for normalized tight frame and harmonic frames, the reconstruction error is bounded by $O(s/m)$. Soon after that Benedetto et al. provided the result on second order $\Sigma\Delta$ quantized frames expansion [3]. In [3], for unit-norm tight frames and Harmonic frames the reconstruction error is in fact generally bounded again by $O(s/m)$, and only when the dimension of the space is even, the Harmonic frames can reach a bound by $O((s/m)^2)$. In [7], the same bound as in [3] is also proved by Bodmann, and furthermore, the reconstruction error for first order $\Sigma\Delta$ quantized frame expansion is proved to have its maximal error is both lower and upper bounded by $\Omega(s/m)$ and $O(s/m)$. Bodmann et al. in [8] continuing provided the error bound for frame expansion with higher order $\Sigma\Delta$ quantization.

In fact in [8], the result is valid also for any proper quantizer, and the method used involving smooth frame-path. This is also the idea of the Sobolev dual frame. Sobolev duals in frame theory and $\Sigma\Delta$-quantizationwas first studied by Blum et al. in paper [5]. In [5], Blum et al. showed that generally the reconstruction error bound for rth order $\Sigma\Delta$-quantization can achieve $O((s/m)^r)$, while using Sobolev dual frame as the decoder. The above papers are actually recapped historically. And until at this point, the error bound is within polynomial scaling size, while in [9] Boufounos et al. gave an optimal reconstruction error bound for quantization of sparse representation is $O(e^{-(m/s)})$ in year 2007. First

until year 2012 in [36], Krahmer et al. achieved an error bound of $O(e^{-\sqrt{m/s}})$ for higher order $\Sigma\Delta$-quantization.

The frames mention above are somehow deterministic. Actually back in [27] an asymptotic approach was used to demonstrate the tightness of a random frame, and then in [28] random frames came up on stage nonasymptotically as underlying frame on signal recovery problem from quantized frame expansion. This setting is closer to the compressed sensing setting, and is indeed analysed as the first step on the way to compressed sensing. In [28] Güntürk et al. work first with random compressed sensing matrices, more specifically, a Gaussian matrix. In this paper, given a noise-free environment, a two-step method is proposed, in first step, an exactly sparse signal with components appropriate larger than some threshold, it is guaranteed that the support set of the signal can be recovered by solving an ℓ_1-minimization problem. After recovering the support set of the signal, a finite-frame expansion of the signal showed up. As the next step, the Sobolev dual frame is applied to reconstruct the signal from its frame expansion. This method successfully achieves an error bound of $O((m/s)^{(r-1/2)\alpha})$ for any $0 < \alpha < 1$, if $m \geq s(\log N)^{1/(1-\alpha)}$ up to a constant with respect to $\Sigma\Delta$-quantizationorder r with high probability. Since it is actually the background knowledge for our result in Chapter 3, the results is stated in the following up section below.

Recently in [15] another dual frame called Beta duals was designed for recovering the quantized measurements of random (Gaussian) frame expansions, which reaches an error bound of $O(\sqrt{s}L^{-(1-\eta)m/s})$, for L being the quantization levels (how fine the alphabet is used), and η some small quantity.

Two-Step Recovery

Given an s-sparse signal x, and an $m \times N$ compressed sensing matrix Φ, where $m \ll N$, obtaining measurements $y = \Phi x$. Applying an rth order $\Sigma\Delta$- quantization scheme to y, q is obtained. If treat q as perturbed measurements, i.e., $q = y + e = \Phi x + e$, then by [28], the support set can be determined. This is proved by a modified version of Proposition 4.1 in [28] and the reconstruction guarantee in [11].

Proposition 1. *Given $x \in \mathbb{R}^N$ an s-sparse signal, denote e a noise vector with $\|e\|_2 \leq \epsilon$, and let $\Phi \in \mathbb{R}^{N\times m}$ be a compressed sensing matrix. Reconstruct x from $q = \Phi x + e$ via ℓ_1 minimization*

obtaining x', i.e.,

$$\hat{x} = arg\min \|z\|_1 \ \textit{subject to}\ \|\Phi z - q\|_2 \leq \epsilon.$$

If $\frac{1}{\sqrt{m}}\Phi$ has restricted isometry constants such that $\delta_{2s} < \frac{1}{\sqrt{2}}$, then $\|x - \hat{x}\|_2 \leq K\frac{1}{\sqrt{m}}\epsilon$, let $\|T\|_0 = s$, and if $\min_{j\in T} |x_j| \geq K2^{r-\frac{1}{2}}\Delta$, $j \in T$, for some positive constant K, then the index set of largest s components of x' is T.

With criteria of Proposition 1, the support set, T, of x can be identified. Then reconstructing the signal by multiplying a left inverse of Φ_T, say L on the left of the submatrix, Φ_T, consisting of columns of Φ with respect to the support set T, the reconstruction ℓ_2-error is then given by

$$\|x - \hat{x}\|_2 = \|Ly - Lq\| = \|L(y - q)\|_2 = \|L(D^r u)\|_2 \leq \|LD^r\|_{2\to 2}\|u\|_2.$$

The Sobolev dual matrix $L_{sob,r}$, first introduced in [5], is a left inverse of Φ_T defined to minimize $\|LD^r\|_{2\to 2}$, i.e.,

$$L_{sob,r} := arg\min_L \|LD^r\|_{2\to 2} \quad \text{subject to } L\Phi_T = I.$$

The geometric intuition is that the dual frame $L_{sob,r}$ is smoothly varying. Since $L\Phi_T = I$, $LD^rD^{-r}\Phi_T = I$. And $L_{sob,r} := arg\min_L \|LD^r\|_{2\to 2}$ we choose $L_{sob}D^r$ to be the Moore-Penrose pseudoinverse of $D^{-r}\Phi_T$, written as $(D^{-r}\Phi_T)^\dagger$, for which recovers the ℓ_2 minimized solution, as well as in [28], the error bound is then obtained

$$\|x - \hat{x}\|_2 \leq \|(D^{-r}\Phi_T)^\dagger\|_{2\to 2}\|u\|_2 = \frac{1}{\sigma_{min}(D^{-r}\Phi_T)}\|u\|_2. \tag{1.26}$$

Recall that $\|u\|_2 \leq 2^{-1}\Delta\sqrt{m}$, once a bound for $\sigma_{min}(D^{-r}\Phi_T)$ is found from below we can bound $\|x-\hat{x}\|_2$ from above. The bound of this singular value is stated in Proposition 5 proved based on study of Toeplitz matrices, which depends highly on Weyl's inequality [31] (see also for example in [28]).

With the same two-step approach as above from [28], in paper [37] the frame used in [28] was extended to a sub-Gaussian frame expansion, and the error bound was further improved from polynomial to root-exponential $O(e^{-(m/s)})$.

Beyond the quantization on finite-frame expansion, papers [24, 28, 32, 37, 55] provide error bound for

$\Sigma\Delta$-quantizationon compressed sensing. Since this is the main issue of this thesis, we will take a closer look of these papers.

Following the two-step method proposed in [28], the author of this thesis joint with Krahmer in [24] exploit the property of the compressed sensing matrices used in [28,37] to the matrices, of which a certain linear transformation satisfies the restricted isometry property. The results are presented in this thesis, see Chapter 3 for more details.

I would say that this two-step approach is somehow a"frame-like" approach to quantization problem on compressed sensing, and it can analyse the cases without noise and the exactly-sparse signals. Instead of using the two-step recovery approach in [24, 28, 37], in [48] Saab utilized a "compressed sensing"-like approach to estimate the error bound of quantization on compressed sensing, by solving a convex problem, or more precisely, a ℓ_1-minimization problem, which then allowing the analysis of approximately-sparse signals with existence of noise. In contrast to BPDQ$_p$ in [33], which is for general quantization, the decoder here is specifically designed for rth order $\Sigma\Delta$-quantization. This approach is "compressed sensing-like", such that it can get rid of the drawback in the "frame-like" approach to quantization on compressed sensing, that doesn't allow the existence of noise and the signal has to be exact sparse (not robust to noise and not stable of the signal). In [48], even approximately sparse can be handled with noise bounded by ϵ. Specifically, if q results from quantizing compressed sensing measurements y using an r^{th}-order $\Sigma\Delta$ scheme, one approximates x with $\hat{x}$ via

$$(\hat{x}, \hat{e}) := \arg\min_{(z,\nu)} \|z\|_1 \quad \text{subject to } \|D^{-r}(\Phi z + \nu - q)\|_2 \leq \gamma(r)\sqrt{m}$$
$$\text{and} \quad \|\nu\|_2 \leq \epsilon\sqrt{m}, \tag{1.27}$$

where $\gamma(r)$ depends on the quantization scheme used. As the result the reconstruction error decays polynomially in m as $m^{-r+1/2}$, and the approach is shown to be stable and robust.

[48] gave a general form for reconstruction error, however they took only sub-Gaussian matrices for example, we in this thesis (also in paper [32]) applies it to a structured case, i.e., partial random circulant matrices. The main task is to prove the RIP of its transformation.

What is this important transformation of the compressed sensing matrix used in [24, 49]? Such

that once the restricted isometry property of which is satisfied, the reconstruction error is bounded polynomially, as results in [24, 49]. It is in fact the interaction between the compressed sensing matrix and the right unitary matrix, denoted by V^*, of the singular value decomposition of the inverse of rth order difference matrix gone through a projection,denoted by P_ℓ. With these notation the result is stated below.

Theorem 7. *[48] Let Φ be an $m \times N$ matrix, and let $s, l \in \{1, ..., m\}$. Suppose that $\frac{1}{\sqrt{\ell}} P_\ell V^* \Phi$ satisfies the restricted isometry property of order $2s$ and constant $\delta < 1/9$. Denote by $Q^r_{\Sigma\Delta}$ a stable rth order $\Sigma\Delta$ quantizer. Then, for all $x \in \mathbb{R}^N$ with $\|\Phi x\|_\infty \leq \mu < 1$ and all $e \in \mathbb{R}^m$ with $\|e\|_\infty \leq \epsilon < 1 - \mu$ the estimate $\hat{x}$ obtained by solving (1.27) with $q = Q^r_{\Sigma\Delta}(\Phi x + e)$ satisfies*

$$\|\hat{x} - x\|_2 \leq C_1 \left(\frac{m}{\ell}\right)^{-r+1/2} \delta + C_2 \frac{\sigma_s(x)}{\sqrt{k}} + C_3 \sqrt{\frac{m}{\ell}} \epsilon, \tag{1.28}$$

where the constants C_1, C_2, C_3 depend on the quantizer, but not the dimensions of the problem.

Quantization problem on structure random compressed sensing matrices such as discrete Fourier matrix was first analysed in [55]. [55] shows that by using the first order $\Sigma\Delta$-quantization the compressed sensing recovery error decays polynomially as $\mathcal{O}\big(\frac{m}{s^4 \log N}\big)^{-1/2}$, when the compressed sensing matrix is a randomly selected $m \times N$ submatrix of the $N \times N$ discrete Fourier transform matrix, with m scales like k^4, while a linear scaling of m with k (up to log factors) arising in Theorem 7 is expected and a linear scaling of m is also common in compressed sensing without quantization.

As another example of structured random compressed sensing matrix, in Chapter 4 we demonstrate our result in [32] on partial random circulant matrices and which result in the theorem below.

Theorem 8. *Denote by $Q^r_{\Sigma\Delta}$ a stable rth order $\Sigma\Delta$ quantizer. Let Φ be an $m \times N$ partial random circulant matrix associated to a vector with independent L-subgaussian entries with mean 0 and variance 1. Suppose that $N \geq m \geq (C\eta)^{\frac{1}{1-2\alpha}} s \log^{\frac{2}{1-2\alpha}} N \log^{\frac{2}{1-2\alpha}} s$, for some $\eta > 1$ and $\alpha \in [0, 1/2)$. With probability exceeding $1 - e^{-\eta}$, the following holds:*

For all $x \in \mathbb{R}^N$ with $\|\Phi x\|_\infty \leq \mu < 1$ and all $e \in \mathbb{R}^m$ with $\|e\|_\infty \leq \epsilon < 1 - \mu$ the estimate $\hat{x}$ obtained by

solving (1.27) *satisfies*

$$\|\hat{x} - x\|_2 \leq C_1 \left(\frac{m}{\ell}\right)^{-r+1/2} \delta + C_2 \frac{\sigma_k(x)}{\sqrt{k}} + C_3 \sqrt{\frac{m}{\ell}} \epsilon.$$

Here C, C_1, C_2, C_3 *are constants that only depend on* r *and* L.

Still another, it is of great interest to find the error bound of signal recovery from $\Sigma\Delta$-quantization quantized Fourier transformation (DFT) due to its popularity in industry and engineering.

Thus the problem in this chapter is then to bound the reconstruction error of the $\Sigma\Delta$-quantization quantized partial random discrete Fourier transformation.

The difficulty of this problem is that it is still not clear what is exactly the singular value decomposition (svd) of the higher order quantization matrix, i.e. the matrix V^* in svd of $D^r = V^* S U$. A conjecture is proposed as following.

Conjecture 1. *There exists a constant* c *such that for any* r*, the singular vectors* V^* *of* $D^{-r} \in \mathbb{R}^{m \times m}$ *satisfies*

$$\|V\|_{\max} \leq c r^r \sqrt{\frac{1}{m}},$$

where $\|V\|_{\max} := \max_{i,j} |v_{i,j}|$ *is the element-wise norm of* V.

And when trying to apply our method (i.e. Dudley's inequality together with McDiarmid's inequality) in Chapter 4 for finding the restricted isometry property of the product $P_\ell V^* R_\Omega C_x$, another problem is that the DFT doesn't repeat like the circulant matrix, with which the cancellation can happen to reduce the effect of $P_\ell V^*$.

It was hoped that the product $\frac{1}{\sqrt{\ell}} P_\ell V^* R_\Omega F$ satisfies the RIP. In [55], it is shown to achieve the restricted isometry property if F contains no all one column. However if F contains the all one column, there is no satisfactory upper bound for the product.

Theorem 9. *[55] Let* $F \in \mathbb{C}^{N \times s}$ *be an nonnormalized DFT with any* s *out of* N *columns (i.e.* $F^T F = N I_s$*)and assume that* F *contains no all* 1 *column. Then there exists a positive function* c_1 *such that for*

any $c, \epsilon > 0$, as long as ℓ satisfies $m \geq \ell \geq c_1(r,c)s\log^3(m/\epsilon)$, it holds with probability $1-\epsilon$ that

$$\sigma^2_{\min}(P_\ell V^* R_\Omega F) \geq \ell\big(1 - c - 2\sqrt{\frac{s}{\ell}}\log\frac{4s}{\epsilon}\big),$$

and

$$\sigma^2_{\max}(P_\ell V^* R_\Omega F) \geq \ell\big(1 + c + 2\sqrt{\frac{s}{\ell}}\log\frac{4s}{\epsilon}\big),$$

where $c_1(r,c) = c_2 r^{2r}/c^2$ with c_2 being an absolute constant.

In Theorem 9 [55], the error bound is established by controlling the norm of $P_\ell V^* R_\Omega F$ and constraint in the decoder, i.e. ℓ_1-minimization.

Theorem 10. *[55] Let F be an nonnormalized DFT matrix of dimension N, and let $R_\Omega F \in \mathbb{C}^{m\times N}$ be a matrix with randomly selected rows from F with replacement. Assume $x \in \mathbb{R}^N$ is an s-sparse signal. If Conjecture 1 is true, let q be the rth order $\Sigma\Delta$-quantization of the compressed measurements $R_\Omega F$ with the quantization alphabet $\delta\mathbb{Z} + \delta\mathbb{Z}i$ and suppose $\hat{x}$ is the solution to*

$$\min \|z\|_1, \ \ s.t. \ \|D^{-1}(q - R_\Omega F z)\|_\infty \leq \delta/2.$$

Then there exist absolute constants c_1 and c_2 such that for any $\epsilon > 0$,

$$\sup_{x\in D_{s,N}} \|x - \hat{x}\|_2 \leq C(s,N,r)m^{-r+1/2},$$

with probability over $1-\epsilon$ provided that $m > c_2 s^4 \log^3 N/\epsilon$.

1.3.3 What's the goal

An overview of the problem setting can be described as following. Given an original signal x, and a compressed sensing matrix A, collecting the measurement $y = Ax$ randomly and after going through the $\Sigma\Delta$-quantization, recover the signal by the proposed method, say decoder, to get $\hat{x}$.

Fundamentally the problem in this thesis is to find an "good" upper bound for the reconstruction

error in ℓ_2, i.e., $\|\hat{x} - x\|_2$. When the measurement noise exists, the problem is extended by letting the measurement $y = Ax + e$, and for the possibility to recovery the signal nicely, surely the noise cannot be too large therefore set an upper bound for it as conventionally, $\|e\|_2 \leq \epsilon$. And if the signal is not strictly sparse, but approximately the result should show how the nonsparsity would effect the bound, this may be expressed in terms of $\sigma_s(x)_p$.

What is then a "good" upper bound for the reconstruction error? Let us first ask: The upper bound should be a function of which variables?

To be a meaningful bound, this upper bound should reveal what one can do to make the error smaller. And also a good upper bound should be "tight" enough, i.e., should be small.

And if the signal is not strictly sparse, but approximately, the bound should show in which level or how badly this nonsparsity can affect the recovery.

On the perspective to the decoder, since the variables related to the quantizer is fixed once the quantization is decided, Thus the order r, the step size Δ of the quantization can be regarded as constant when the quantizer is decided. The number of the measurements is then the essential variable to see whether the bound is good enough. Literally the optimal reconstruction error of sparse signals decays exponentially in m, $O(e^{-m})$. The dimension of the signal N is decided once the signal is there, which can not be changed by any artificial interference. The sparsity of the signal also cannot be changed. Noise is also not changeable but have to be under control.

To evaluate the results, we first notice that as in the case with $\Sigma\Delta$-quantization in the finite-frames context (e.g., [36]) and in the sub-Gaussian compressed sensing measurements context [37,49], the optimal reconstruction error decays root-exponentially (to the sampling rate m/s) when the original signal is exact sparse and without noise. With this, we may say that polynomial decay (to the sampling rate m/s) is already satisfactory. Second, if with m meaningfully scales linearly up to sparsity s satisfactory? Now comparing to the case of Gaussian or Bernoulli random compressed sensing matrix, it is required to have m linearly scaling to sparsity s. Therefore it is actually quite exciting to reach also a linear scale.

The decoder to be chosen can surely affect the upper bound in this thesis, since a practical, instead of a general, recovery is one of the main issue in the series of signal recovery problem on quantization compressed sensing.

In [28], a two-step recovery is proposed. The support set of a sparse signal is recovered first, and then the problem is reduced to signal recovery for frame expansions. This decoder has the limit that the signals have to be strictly sparse, and the terms should be away from zero. In [48] with robust null space property, or restricted isometry property, the decoder is stable and robust. In this thesis, the decoder proposed in [48] will be applied here.

In the context of quantization compressed sensing, despite the importance of circulant matrices, all of papers in this topic focus on the random subgaussian measurement matrices for its nice properties to be analysed. However since every linear time invariant (LTI) system is represented by convolution to a kernel function, and a convolution is then represented by a circulant matrix. For the generality of a LTI system, it is of importance to analyse the circulant matrix in quantization compressed sensing.

Chapter 2

Review of mathematical tools

2.1 Dudley's inequality

This chapter is modified from [52]. Dudley's inequality, specifically, Dudley's entropy bound is a typical method nowadays to bound the supremum of a random variable over an index set. The technique is based on the chaining argument, which runs along the index set by approximating an aimed index by a series of elements from the index set. To the end, the entropy bound, which contains a covering bound will be evaluated by, here in the thesis, Maurey's method and a volumetric argument.

2.1.1 The generic chaining

The generic chaining is an essential step on the way to Dudley's inequality [52]. To demonstrate it, we need a set of random variables with an index set T, denoted as $X_s,\ s \in T$, which satisfies the tail property.

$$\mathbb{P}(|X_s - X_t| \geq ud(s,t)) \leq 2\exp(\frac{-u^2}{2}),\ \textit{for all } u \in \mathbb{R}. \tag{2.1}$$

As stated above, this index set T is a metric set with distance between $x,\ y$ denoted by $d(x,y)$. Let T_n be a subset of T and for $t \in T$, $\pi_n(t) \in T_n$ which is closest to t.

The chaining argument

Now consider the following argument. If given event $\mathcal{E}$

$$|X_{\pi_n(t)} - X_{\pi_{n-1}(t)}| < ud(\pi_n(t), \pi_{n-1}(t))2^{n/2},\ \forall n \geq 1,\ t \in T, \tag{2.2}$$

then we are ready to go into the chaining step.

$$\begin{aligned}
&\sup_{t\in T}|X_t - X_{t_0}| = \sup_{t\in T}|\sum_{n\geq 1} X_{\pi_n(t)} - X_{\pi_{n-1}(t)}| \\
&\leq \sup_{t\in T}\sum_{n\geq 1}|X_{\pi_n(t)} - X_{\pi_{n-1}(t)}| \\
&\leq \sup_{t\in T}\sum_{n\geq 1} ud(\pi_n(t), \pi_{n-1}(t))2^{n/2} \\
&\leq \sup_{t\in T}\sum_{n\geq 1} u[d(\pi_n(t), t) + d(\pi_{n-1}(t), t)]2^{n/2} \\
&= \sup_{t\in T}\{\sum_{n\geq 1} ud(\pi_n(t), t)2^{n/2} + \sum_{n\geq 1} ud(\pi_{n-1}(t), t)2^{n/2}\} \\
&\leq \sup_{t\in T}\{\sum_{n\geq 0} ud(\pi_n(t), t)2^{n/2} + \sum_{n\geq 0} ud(\pi_n(t), t)2^{n/2}\} \\
&= \sup_{t\in T} 2u\sum_{n\geq 0} d(\pi_n(t), t)2^{n/2},
\end{aligned}$$

which implies that taking infimum of all admitted sequence $(T_n)_{n\geq 0}$ with *card* $T_n \leq N_n$, it holds also

$$\sup_{t\in T}|X_y - X_{t_0}| \leq 2u\sum_{n\geq} \inf_{T_n}\sup_{t\in T} d(T_n, t)2^{n/2}. \tag{2.3}$$

The next thing we want to do is to express the right hand side by the covering number of the set X.

Now define entropy numbers e_n and covering number.

Definition 5. *[52]*

$$e_n = \inf_{t\in T}\sup d(T_n, t), \tag{2.4}$$

where the infimum is taken over all subsets T_n of T.

Definition 6. *[52] Let (X, d) be a metric space and let $\epsilon > 0$. A subset $\mathcal{N}_\epsilon$ of X is called an ϵ-net of X*

if every point $x \in X$ can be approximated to within ϵ by some point $y \in \mathcal{N}_\epsilon$, i.e. so that $d(x,y) \leq \epsilon$. The minimal cardinality of an ϵ-net of X, if finite, is denoted $\mathcal{N}(X,d,\epsilon)$ and is called the covering number of X.

To evaluate the summation in (2.3) by using covering numbers, there should be a connection between Definition (2.1.1) and Definition (2.1.1). This is by setting the cardinality to the sequence T_n such that *card* $T_n \leq N_n = 2^{2^n}$, and than obtaining

$$e_n = \inf\{\epsilon, \mathcal{N}(T,d,\epsilon) \leq 2^{2^n}\}. \tag{2.5}$$

Therefore

$$\begin{aligned}
& \sqrt{\log 2^{2^n}}(e_n - e_{n+1}) \leq \int_{e_{n+1}}^{e_n} \sqrt{\log N(T,d,\epsilon)}d\epsilon \\
\Rightarrow \quad & \sqrt{\log 2}\sum_{n\geq 0} 2^{n/2}(e_n - e_{n+1}) \leq \sum_{n\geq 0}\int_{e_{n+1}}^{e_n} \sqrt{\log N(T,d,\epsilon)}d\epsilon \\
\Rightarrow \quad & \sqrt{\log 2}\sum_{n\geq 0} 2^{n/2}e_n - \sum_{n\geq 1} 2^{\frac{n-1}{2}}e_n \leq \int_0^{e_0} \sqrt{\log N(T,d,\epsilon)}d\epsilon \\
\Rightarrow \quad & \sqrt{\log 2}(1-\frac{1}{\sqrt{2}})\sum_{n\geq 0} 2^{n/2}e_n \leq \int_0^{e_0} \sqrt{\log N(T,d,\epsilon)}d\epsilon.
\end{aligned}$$

Therefore the summation part in (2.3) is bounded by

$$\sum_{n\geq 0} 2^{n/2}e_n \leq \{\sqrt{\log 2}(1-\frac{1}{\sqrt{2}})\}^{-1}\int_0^{e_0} \sqrt{\log N(T,d,\epsilon)}d\epsilon. \tag{2.6}$$

Insert (2.6) to (2.3)obtaining

$$\sup_{t\in T}|X_t - X_{t_0}| \leq 2u\{\sqrt{\log 2}(1-\frac{1}{\sqrt{2}})\}^{-1}\int_0^{e_0} \sqrt{\log N(T,d,\epsilon)}d\epsilon. \tag{2.7}$$

To make the statement clear, summarized this chapter as following:

If

$$|X_{\pi_n(t)} - X_{\pi_{n-1}(t)}| < ud(\pi_n(t),\pi_{n-1}(t))2^{n/2},\ \forall n \geq 1,\ t \in T, \tag{2.8}$$

then

$$\sup_{t\in T}|X_t - X_{t_0}| \leq 2u\{\sqrt{\log 2}(1-\frac{1}{\sqrt{2}})\}^{-1}\int_0^{e_0}\sqrt{\log N(T,d,\epsilon)}d\epsilon. \tag{2.9}$$

2.1.2 Tail bound for $\sup X_t$

Our goal is however to find a upper bound $p(u)$ for $\sup_{t\in T} X_t$, or stronger statement:

$$\mathbb{P}(\sup_{t\in T}|X_t - X_{t_0}| > 2u\{\sqrt{\log 2}(1-\frac{1}{\sqrt{2}})\}^{-1}\int_0^{e_0}\sqrt{\log N(T,d,\epsilon)}d\epsilon) \leq p(u). \tag{2.10}$$

This is an compliment event to event (2.8), therefore

$$\begin{aligned}
&\mathbb{P}(\sup_{t\in T}|X_t - X_{t_0}| > 2u\{\sqrt{\log 2}(1-\frac{1}{\sqrt{2}})\}^{-1}\int_0^{e_0}\sqrt{\log N(T,d,\epsilon)}d\epsilon) \\
\leq&\mathbb{P}(\exists(\pi_n(t),\pi_{n-1}(t)) such\ that\ |X_{\pi_n(t)} - X_{\pi_{n-1}(t)}| > ud(\pi_n(t),\pi_{n-1}(t))2^{n/2}) &(2.11)\\
\leq&2^{2^{n+1}}\mathbb{P}(|X_{\pi_n(t)} - X_{\pi_{n-1}(t)}| > ud(\pi_n(t),\pi_{n-1}(t))2^{n/2}) &(2.12)\\
\leq&2^{2^{n+1}+1}\exp(-u^2/(2^{-n/2+1})) = 2^{2^{n+1}+1}\exp(-u^2 2^{n-1}) &(2.13)\\
\leq&2^{2^{n+1}+1}\exp(-\frac{u^2}{2}-2^{n+1}) \leq 2\exp(\frac{-u^2}{2}) := p(u).
\end{aligned}$$

In (2.13) by argument in Chapter 2 [52],

$$u^2 2^{n-1} \geq \frac{u^2}{2} + u^2 2^{n-2} \geq \frac{u^2}{2} + 2^{n+1}, \tag{2.14}$$

and then applying union bound in (2.11) and Proposition (3) in (2.12)yields the result. For later use, this chapter will be concluded as

Theorem 11 (Dudley's inequality [52])**.**

$$\mathbb{P}\big(\sup_{t\in T}|X_t - X_{t_0}| > u\big) \lesssim \exp\bigg(\frac{-u^2}{\big(\int_0^{e_0}\sqrt{\log N(T,d,\epsilon)}d\epsilon\big)^2}\bigg). \tag{2.15}$$

In the followed up chapter, two classical method for estimating the covering number will be introduced.

2.1.3 Evaluation of covering number

In this chapter Maurey's method and volumetric argument will be demonstrated. And note that two of them can be used together for one same set for volumetric argument is tighter than Maurey's method when ϵ is small.

Volumetric argument

Lemma 1. *[54] Volumetric argument says that the covering number $\mathcal{N}(B(0,1),d,\epsilon)$ of a unit ball $B(0,1) := \{x \in \mathbb{R}^n,\ d(x,0) \leq 1\}$ is less than or equal to $(1+\frac{2}{t})^n$.*

Proof. Let $\epsilon > 0$ and $\mathcal{N}_\epsilon$ be a maximal ϵ-separated subset of $B(0,1)$, i.e., $\forall x \in B(0,1)$ there exists $t \in \mathcal{N}_\epsilon$ such that $d(t,x) \leq \epsilon$, then for any $s \neq t \in \mathcal{N}_\epsilon$ $B(s,\epsilon/2)$ and $B(t,\epsilon/2)$ do not intersect. Which implies that

$$
\begin{aligned}
Vol(|\mathcal{N}_\epsilon| B(t,\epsilon/2)) &\leq B(0,1+\epsilon/2) \\
|\mathcal{N}_\epsilon|\, Vol(B(t,\epsilon/2)) &\leq B(0,1+\epsilon/2) \\
|\mathcal{N}_\epsilon|(\epsilon/2)^n &\leq (1+\epsilon/2)^n \\
|\mathcal{N}_\epsilon| &\leq (1+2/\epsilon)^n.
\end{aligned}
$$

Since the unit ball is covered by ϵ-balls of elements in $\mathcal{N}_\epsilon$, $\mathcal{N}_\epsilon$ is larger then or equal to the covering number. And since they are disjoint, the covering number is greater then or equal to it, and since the covering number is the smallest such number, which is then less then or equal to it. Thus the $\mathcal{N}_\epsilon$ equals the covering number. □

Maurey's method

Maurey's method estimates the covering number of a convex hull consisted from $\mathcal{U} = \{u_j\}_{j=1}^N$ by approaching each element of the convex hull by an average of a summation of random vector taking values from $\mathcal{U}$.

Lemma 2. *[44] Given $\mathcal{U} = \{u_j\}_{j=1}^N$ If $x \in conv(\mathcal{U})$ then $x = \sum_{j=1}^N \theta_j u_j$ with $\theta_j \geq 0$, $\sum_{j=1}^N \theta_j = 1$. Let* Z *be a random vector which takes value u_j. Let M be a number to be determined later, and set $(\mathrm{Z}_j)_{j=1}^M$*

be independent copies of Z*, and treating them for the moment as fixed numbers. If for some norm* $\|\cdot\|_X$ *there is an absolute constant* A *such that* $\mathbb{E}_{\epsilon_k}\|\sum_{k=1}^{M}\epsilon_k \mathrm{Z}_k\|_X \leq A\sqrt{M}$*, then*

$$\log^{\frac{1}{2}}\mathcal{N}(conv(\mathcal{U}),\|\cdot\|_X,\epsilon)\leq(\frac{2A}{\epsilon})\log^{\frac{1}{2}}N. \tag{2.16}$$

Proof. If $x\in conv(\mathcal{U})$ then $x=\sum_{j=1}^{N}\theta_j u_j$ with $\theta_j\geq 0,\ \sum_{j=1}^{N}\theta_j=1$. Let Z be a random vector which takes value u_j with probability θ_j for $j=1,\ldots,N$ and thus $\mathbb{E}\mathrm{Z}=x$. Let M be a number to be determined later, and set $(\mathrm{Z}_j)_{j=1}^{M}$ be independent copies of Z, and $z=\frac{1}{M}\sum_{k=1}^{M}\mathrm{Z_k}$. Given $(\xi_k)_{k=1}^{M}$ a i.i.d. Rademacher random vector applying symmetrization in Lemma 16), and x is approached by z as

$$\mathbb{E}\|z-x\|_X=\mathbb{E}\|\frac{1}{M}\sum_{k=1}^{M}\mathrm{Z}_k-\mathbb{E}\mathrm{Z}_k\|_X\leq\frac{2}{M}\mathbb{E}\|\sum_{k=1}^{M}\xi_k\mathrm{Z}_k\|_X.$$

If for some norm $\|\cdot\|_X$ there is an absolute constant A such that $\mathbb{E}\|\sum_{k=1}^{M}\epsilon_k\mathrm{Z}_k\|_X\leq A\sqrt{M}$, then

$$\mathbb{E}\|z-x\|_x\leq\frac{2A}{\sqrt{M}}. \tag{2.17}$$

Which means there exist at least one such $z^*=\frac{1}{M}\sum_{k=1}^{M}\mathrm{Z}_k$ such that

$$\|z^*-x\|_X\leq\frac{2A}{\sqrt{M}}. \tag{2.18}$$

Now M can be determined since $\|z^*-x\|_X$ is supposed to be less than ϵ. This is achievable if $M=(\frac{2A}{\epsilon})^2$. And since there are N^M candidates which can be the z^*, the covering number is hence bounded by N^M, Hence

$$\log^{\frac{1}{2}}\mathcal{N}(conv(\mathcal{U}),\|\cdot\|_X,\epsilon)\leq(\frac{2A}{\epsilon})\log^{\frac{1}{2}}N. \tag{2.19}$$

□

2.2 Moments and tails

Moments and tails are used commonly to see the behaviour of a random variable. In this section we introduce some properties which will be used in this thesis.

Proposition 2 (Höffding's inequality for Rademacher sums)**.** *[26] Let* $= (b_1, \ldots, b_M) \in \mathbb{R}^M$ *and* $\epsilon = (\epsilon_1, \ldots, \epsilon_M)$ *be an i.i.d. Rademacher sequence, i.e., each* ϵ_i *takes value* ± 1 *with equal probability. Then, for* $u > 0$*,*

$$\mathbb{P}\Big(\sum_{j=1}^{M} |\epsilon_j b_j| \geq \|b\|_2 u\Big) \leq exp(-u^2/2). \tag{2.20}$$

Proof. By Markov's inequality,

$$\begin{aligned}
&\mathbb{P}\Big(\sum_{j=1}^{M} \epsilon_j b_j \geq u\Big) \\
&= \mathbb{P}\Big(\exp(\sum_{j=1}^{M} \lambda \epsilon_j b_j) \geq \exp(\lambda u)\Big) \\
&\leq e^{-\lambda u}\mathbb{E}[\exp(\lambda \sum_{j=1}^{M} \epsilon b_j)] \\
&= e^{-\lambda u}\Pi_{j=1}^{M}\mathbb{E}[\exp(\epsilon_j \lambda b_j)] \\
&= e^{-\lambda u}\Pi_{j=1}^{M}[\frac{1}{2}\exp(-1\lambda b_j) + \frac{1}{2}\exp(1\lambda b_j)] \\
&= e^{-\lambda u}\Pi_{j=1}^{M}\frac{1}{2}[\sum_{k=0}^{\infty}\frac{(-\lambda b_j)^k}{k!} + \sum_{k=0}^{\infty}\frac{(\lambda b_j)^k}{k!}] \\
&= e^{-\lambda u}\Pi_{j=1}^{M}[\sum_{k=0}^{\infty}\frac{(\lambda b_j)^{2k}}{(2k)!}] \\
&\leq e^{-\lambda u}\Pi_{j=1}^{M}[\sum_{k=0}^{\infty}\frac{(\lambda b_j)^{2k}}{(2^k)k!}] \\
&= e^{-\lambda u}\Pi_{j=1}^{M}[e^{(\lambda b_j)^2/2}] \\
&= e^{-\lambda u}[e^{(\lambda^2 \sum_{j=1}^{M} b_j^2)/2)}] \\
&= e^{-\lambda u}[e^{(\lambda^2 \|b\|_2^2/2)}] \\
&= e^{-\lambda u + \lambda^2 \|b\|_2^2/2}.
\end{aligned}$$

Minimizing the exponential by letting $\lambda = \frac{u}{\|b\|_2^2}$, then

$$\mathbb{P}\Big(\sum_{j=1}^{M} \epsilon_j b_j \geq u\Big) \leq e^{\frac{-u^2}{2\|b\|_2^2}}, \tag{2.21}$$

or equivalently,

$$\mathbb{P}\Big(\sum_{j=1}^{M}\epsilon_j b_j \geq \|b\|_2 u\Big) \leq e^{\frac{-u^2}{2}}. \tag{2.22}$$

□

Proposition 3 (Proposition 7.14 [26]). *Let Z be a random variable satisfying*

$$\mathbb{P}(|Z| \geq u) \leq \beta e^{-u^2/2},\ \forall u \geq \sqrt{2},$$

for some constants $\alpha > 0$, $\beta \geq 2$. Then

$$\mathbb{E}|Z| \leq C_\beta \sqrt{\ln(4\beta)},$$

with $C_\beta = \sqrt{2} + 1/(4\sqrt{2}\ln(4\beta)) \leq \sqrt{2} + 1/(4\sqrt{2}\ln(8)) \sim 1.499 < 3/2$.

Proof. For some $\kappa \geq u \geq \sqrt{2}$,

$$\begin{aligned}\mathbb{E}|Z| &= \int_0^\infty \mathbb{P}(|Z| > u)du \\ &= \int_0^\kappa 1du + \int_\kappa^\infty \mathbb{P}(|Z| > u)du \leq \kappa + \int_\kappa^\infty \beta e^{-u^2/2}du,\end{aligned}$$

by Lemma 13 in Appendix, and assumption that $\kappa > \sqrt{2}$,

$$\mathbb{E}|Z| \leq \kappa + \frac{\beta e^{-\kappa^2/2}}{\kappa},$$

choosing $\kappa = \sqrt{2\ln(4\beta)}$ yields

$$\begin{aligned}\mathbb{E}|Z| &\leq \sqrt{2\ln(4\beta)} + \frac{1}{4\sqrt{2\ln(4\beta)}} \\ &= (\sqrt{2} + \frac{1}{4\sqrt{2}\ln(4\beta)})\sqrt{\ln 4\beta}.\end{aligned}$$

□

2.3 McDiarmid's inequality

This chapter is based on [40]. McDiarmid's inequality provides an upperbound to the concentration probability of a function of independent random variables. This is based on measuring how the function changes its value by varying only one random variable in the domain.

Theorem 12 ([40], Theorem 3.7)**.** *Consider a random vector $X = (X_i)_{i=1}^m$, where the X_i are taking values in given sets A_i, $i \in [m]$, and let f be a bounded real-valued function defined on $\prod_{j=1}^m A_j$. For $k \leq m$ and $\vec{x} = (x_i)_{i=1}^m \in \prod_{j=1}^m A_j$, let B_k denote the event that $X_i = x_i$ for all $i = 1, \ldots, k-1$. For $g_k(x) := \mathbb{E}[f(X)|B_k \cup \{X_k = x\}] - \mathbb{E}[f(X)|B_k]$, consider the range $ran(x_1, \ldots, x_{k-1}) := \sup\{|g_k(x) - g_k(y)| : x, y \in A_k\}$. Assume that the sum of squared ranges*

$$R^2(\vec{x}) = \sum_{k=1}^{m} (ran(x_1, \ldots, x_{k-1}))^2,$$

is bounded outside a 'bad' subset B of $\prod_{j=1}^m A_j$, that is, $R^2(\vec{x}) \leq r^2$ for all $\vec{x} \notin B$. Then

$$\mathbb{P}(|f(X) - \mathbb{E}f(X)| \geq t) \leq 2\exp(-2t^2/r^2) + \mathbb{P}(X \in B).$$

McDiarmid's inequality solves a concentration problem, here a generalized version, which states:

Theorem 13 ([40], Theorem 3.14)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(\emptyset, \Omega) = \mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq ... \subseteq \mathcal{F}_m$ a filtration in $\mathcal{F}$. Consider a bounded random variable X, and set $X_k := \mathbb{E}(X|\mathcal{F}_k)$. Define the sum of squared conditional ranges*

$$R^2 = \sum_{k=1}^{m} ran_k^2,$$

where

$$ran_k := \sup(X_k|\mathcal{F}_{k-1}) + \sup(-X_k|\mathcal{F}_{k-1}),$$

and denote its (essential) supremum by

$$\hat{r}^2 := \sup R^2.$$

Then,

$$\mathbb{P}(X - \mathbb{E}(X) \geq t) \leq e^{-2t^2/\hat{r}^2}.$$

Proof. Let $V_i := X_k - X_{k-1}$, and $\mathcal{F}_m = \sigma(X)$, then $X_0 = \mathbb{E}(x)$ and $X_m = X$. Thus

$$\begin{aligned}
&\mathbb{P}(X - \mathbb{E}X \geq \epsilon) \\
&= \mathbb{P}(\sum_{i=1}^{m} V_i \geq \epsilon) \\
&= \mathbb{P}(e^{t\sum_{i=1}^{m} V_i} \geq e^{t\epsilon}) \\
&\textit{by Markov's inequality} \\
&\leq e^{-t\epsilon}\mathbb{E}(e^{\sum_{i=1}^{m} V_i}) \\
&= e^{-t\epsilon}\mathbb{E}\big[\mathbb{E}(e^{t\sum_{i=1}^{m} V_i})|\mathcal{F}_{m-1}\big] \\
&= e^{-t\epsilon}\mathbb{E}\big[e^{t\sum_{i=1}^{m-1} V_i}\mathbb{E}(e^{tV_m}|\mathcal{F}_{m-1})\big].
\end{aligned} \tag{2.23}$$

Since the expectation of random variable $(V_m|\mathcal{F}_{m-1})$

$$\begin{aligned}
\mathbb{E}(V_m|\mathcal{F}_{m-1}) &= \mathbb{E}\big[\mathbb{E}(X|\mathcal{F}_m) - \mathbb{E}(X|\mathcal{F}_{m-1})|\mathcal{F}_{m-1}\big] \\
&= \mathbb{E}\big[\mathbb{E}(X|\mathcal{F}_m)|\mathcal{F}_{m-1}]\big] - \mathbb{E}\big[\mathbb{E}(X|\mathcal{F}_{m-1})|\mathcal{F}_{m-1}\big] \\
&= \mathbb{E}(X|\mathcal{F}_{m-1}) - \mathbb{E}(X|\mathcal{F}_{m-1}) = 0,
\end{aligned}$$

and due to boundedness of X, both $\sup(V_m|\mathcal{F}_{m-1}) = \sup\big[(X_m|\mathcal{F}_{m-1}) - X_{m-1}\big]$, and $\inf(V_m|\mathcal{F}_{m-1}) = \inf\big[(X_m|\mathcal{F}_{m-1}) - X_{m-1}\big]$ are also bounded with range $\sup\big(X_m|\mathcal{F}_{m-1}\big) + \sup\big(-X_m|\mathcal{F}_{m-1}\big) = ran_m$. By Höffding's lemma (see Lemma 17),

$$\mathbb{E}(e^{tV_m}|\mathcal{F}_{m-1}) \leq \exp(\frac{t^2 ran_m^2}{8}). \tag{2.24}$$

Substitute (2.24) to (2.23) and iterating these steps,

$$\mathbb{P}(X - \mathbb{E}X \geq \epsilon) \leq e^{-t\epsilon} \exp(\frac{t^2 \sum_{k=1}^{m} ran_k^2}{8}). \tag{2.25}$$

After minimizing the right hand side of (2.25) with variable $t = \frac{4\epsilon}{\sum_{k-1}^{m} ran_k^2}$,

$$\mathbb{P}(X - \mathbb{E}X \geq \epsilon) \leq \exp(\frac{-2\epsilon^2}{R^2}), \tag{2.26}$$

taking supremum of R^2,

$$\mathbb{P}(X - \mathbb{E}X \geq \epsilon) \leq \exp(\frac{-2\epsilon^2}{\hat{r}^2}). \tag{2.27}$$

□

By definition, $X_k := \mathbb{E}(X|\mathcal{F}_k)$ it is clear that $\mathbb{E}(X_{k+1}|\mathcal{F}_k) = \mathbb{E}(\mathbb{E}(X|\mathcal{F}_{k+1})|\mathcal{F}_k) = \mathbb{E}(X|\mathcal{F}_k) = X_k$. The result of McDiarmid's inequality is thus the consequence of a martingale sequence. The interested readers are refered to [40] for further information.

Chapter 3

RIP approached error bound

The content in this chapter is a joint work with Felix Krahmer. In this chapter the approach to estimating the error of reconstruction from $\Sigma\Delta$ quantized measurements for compressed sensing in [24] is introduced.

Our method is based on the restricted isometry property (RIP) of the matrix $\frac{1}{\sqrt{\ell}}P_\ell V^*\Phi$ of the compressed sensing matrix Φ with projection P_ℓ and the right unitary matrix V^* of singular value decomposition of the rth power to the inverse of difference matrix D, i.e., D^{-r} (1.24). The main application of our result is the error analysis for random subgaussian matrices.

The main result in this chapter is that once we know the restricted isometry property of the interaction matrix $\frac{1}{\sqrt{\ell}}P_\ell V^*\Phi$ of the compressed sensing matrix Φ, then the reconstruction error can be bounded by our result.

In this chapter the rth order greedy $\Sigma\Delta$-quantized measurements with quantization alphabet $\mathcal{Z} = \Delta\mathbb{Z}$ are used as introduced in Chapter 1.2.

Throughout this chapter the rth power of the difference matrix D, i.e., D^r will be used repeatedly, especially its singular value decomposition. Denoting the singular value decomposition of D^{-r} by $D^{-r} = U_{D^{-r}}S_{D^{-r}}V^*_{D^{-r}}$.

Note that the compressed sensing matrix Φ is not normalized, while in almost all the compressed sensing literature, it is usual normalized to have unit-norm columns. Since we here use an alphabet (more precisely, the step size, Δ) which is independent of the size of the measurement, i.e., m, if the columns are however normalized (up to $\frac{1}{\sqrt{m}}$), it is not fair to compare the result when adjusting the

measurement size m. Therefore, in this Chapter 3 as well as in [28] the measurement matrices are not normalized, and assuming that each entry of the measurement matrices has variance one.

In chapter 3.2 its application on Gaussian (generalized to subgaussian) is shown.

We state our main theorem as follows.

Theorem 14. *Given an s-sparse signal $x \in \mathbb{R}^N$, denoted by $\Phi \in \mathbb{R}^{m\times N}$ a compressed sensing matrix, and q the rth order $\Sigma\Delta$-quantized measurements of Φx with step size Δ. Suppose Φ has the restricted isometry property such that the support set T can be determined. Choose L as the Sobolev dual matrix of Φ_T and reconstruct the signal by $\hat{x} = Lq$ (see Chapter 1.3.2 for details), if $\sqrt{\frac{1}{\ell}}P_\ell V_{D^{-r}}^*\Phi$, $\ell \leq m$, has the restricted isometry constant $\delta_s \leq \delta$, where P_ℓ maps a vector to its first ℓ components. then the reconstruction error is bounded above by*

$$\|x - \hat{x}\|_2 \leq \frac{\Delta}{2c_2(r)\sqrt{(1-\delta)}}(\frac{m}{\ell})^{-r+\frac{1}{2}},$$

where $c_2(r) > 0$ is a constant depending only on r.

Note from Theorem 14, the smaller ℓ is the better the bound. However, ℓ has to be large enough such that $\frac{1}{\sqrt{m}}(P_\ell V_{D^{-r}}^*\Phi)$ has the restricted isometry constant $\delta_s \leq \delta$.

This result can be applied to obtain recovery guarantees for various compressed sensing setting such as Gaussian, subgaussian measurements. We will show in later chapter that our result covers which in [28] and [37].

The proof used the two-step analysis. In the first step, the support set is recovered via an ℓ_1-minimization problem. In the second step, the recovery error is estimated by recovering this frame-based problem obtained from step one in two-step recovery by multiplication with the Sobolev dual frame.

3.1 RIP-based error analysis

From Equation (1.26), the main issue to bound the reconstruction error from below of $\sigma_{min}(D^{-r}\Phi_T)$. Finding the infimum of $\sigma_{min}(D^{-r}\Phi_T)$ over all possible support sets T is equivalent to finding the s-sparse vector with the smallest image under $D^{-r}\Phi_T$. In this chapter we show mathematically that how the restricted isometry property of $P_\ell V^*\Phi$ can reveal the reconstruction bound. The reason in words, it is

due to the constraints on both sides of the restricted isometry property, these constraints actually give the bounds for singular values of the matrix $\frac{1}{\sqrt{\ell}}P_\ell V^*\Phi$ over all possible support sets. This gives the connection to the concept of RIP.

In the following proof, how the restricted isometry property can be applied to find this effective smallest singular value is shown.

Proof of Theorem 14. Recall that $D^{-r} = U_{D^{-r}} S_{D^{-r}} V^*_{D^{-r}}$. Then, as S is a diagonal matrix,

$$\begin{aligned}
\sigma_{\min}(D^{-r}\Phi_T) &= \sigma_{\min}(S_{D^{-r}} V^*_{D^{-r}}\Phi_T) \\
&\geq \sigma_{\min}(P_\ell S_{D^{-r}} V^*_{D^{-r}}\Phi_T) \\
&= \sigma_{\min}((P_\ell S_{D^{-r}} P^*_\ell)(P_\ell V^*_{D^{-r}}\Phi_T)_{\ell\times s}) \\
&\geq s_\ell \sigma_{\min}(P_\ell V^*_{D^{-r}}\Phi_T)_{\ell\times s},
\end{aligned}$$

Next, need to bound $\sigma_{\min}(P_\ell V^*_{D^{-r}}\Phi_T)$ uniformly over all support set T.

If $\frac{1}{\sqrt{\ell}}P_\ell V^*_{D^{-r}}\Phi$ has the restricted isometry constant $\delta_s \leq \delta$ then $\sigma_{\min}(P_\ell V^*_{D^{-r}}\Phi_T)_{\ell\times s}$ is uniformly bounded from below by

$$\sqrt{\ell}\sqrt{1-\delta}. \tag{3.1}$$

Theorem 18, Proposition 5 in Appendix A, and (3.1) yields the result that

$$\frac{1}{\sigma_{min}(D^{-r}\Phi_T)}\|u\|_2 \leq \frac{\Delta}{2c_2(r)\sqrt{(1-\delta)}}(\frac{m}{\ell})^{-r+\frac{1}{2}}. \tag{3.2}$$

□

3.2 Gaussian and subgaussian matrices

Given Φ a standard Gaussian random matrix. Since $(P_\ell V^*_{D^{-r}}\Phi)$ is also a standard Gaussian random matrix due to rotation invariance, with $\ell = \Omega(s\log N)$, $\frac{1}{\sqrt{\ell}}P_\ell V^*_{D^{-r}}\Phi$ has the restricted isometry constant

$\delta_s < \delta$ with high probability [26]. Since $s \leq \ell \leq m$,

$$\frac{m}{\ell} \leq (\frac{m}{s})^{\alpha},\ \alpha \in (0,1).$$

Provided that $\ell = \Omega(s \log N)$,

$$\begin{aligned} &\frac{m}{\ell} \lesssim (\frac{m}{s})^{\alpha} \\ \Rightarrow &\frac{m}{s \log N} \lesssim (\frac{m}{s})^{\alpha} \\ \Rightarrow &m \gtrsim s(\log N)^{\frac{1}{\alpha - 1}}. \end{aligned}$$

Applying Theorem 14 directly obtained

$$\|x - \hat{x}\|_2 \lesssim \Delta(\frac{m}{s})^{-\alpha(r-\frac{1}{2})},$$

with high probability. This therefore recovers the result in [28].

By similar steps, this can also be generalized to subgaussian measurement matrices, which recovering the result in [2]. As this argument involves some additional technical steps, the details are omitted.

Chapter 4

Error bound of recovery from $\Sigma\Delta$ quantized partial random circulant measurements

As the first author, the results in this chapter is a joint work with Felix Krahmer and Rayan Saab. This chapter is aimed to estimate the reconstruction error from quantized measurements of a compressed sensing problem.

More specifically, we obtain a quantized measurement vector from an original signal via linear transformation by a randomly subsampled circulant matrix formed of i.i.d. subgaussian vector.

The quantization we use here is the greedy $\Sigma\Delta$-quantization. And note that throughout this chapter we subsample the compressed sensing matrix uniformly at random WITHOUT REPLACEMENT.

4.1 Contributions

In this chapter, we demonstrate our results in [32] of analysis to the reconstruction error on partial random circulant matrices. We show that if the compressed sensing matrix is a randomly (without replacement) subsampled partial random circulant matrix, and the measurements are quantized by the greedy $\Sigma\Delta$-quantization introduced in Chapter 1.2.3, then with the decoder in (1.27) we conclude that:

- The reconstruction error decays polynomially with the number of measurements.

- The recovery is robust (whether or not the noise exists) and stable (whether the signal is exactly or only approximately sparse).

- With number of measurements m meaningfully scales linearly to the sparsity s.

How we can evaluate our results? How do we know that we got a good and suitable one? First notice that as in the case with $\Sigma\Delta$-quantization in the finite-frames context (e.g., [36]) and in the sub-Gaussian compressed sensing measurements context [37, 49], the optimal reconstruction error decays root-exponentially when the original signal is exact sparse and without noise. With this, we may say that polynomial decay with our set-up, i.e., with noise and the original signal is only approximately sparse, is already satisfactory. Second, if with m meaningfully scales linearly up to sparsity s satisfactory? Now comparing to the case of Gaussian or Bernoulli random compressed sensing matrix, it is required to have m linearly scaling to sparsity s. Therefore it is actually quite exciting to reach also a linear scale here. Our analysis relies on proving a restricted isometry property for the product of our compressed sensing compressed sensing matrix and the matrix formed by the left singular vectors of an rth order difference operator in (1.24), denoted in Theorem 7 as $\frac{1}{\sqrt{\ell}}P_\ell V^*\Phi$. To prove the restricted isometry property of this product of matrices, we combine a variation of McDiarmid's inequality [40], Dudley's inequality [22], and recent results on suprema of chaos processes [39].This is worthwhile to point out that comparing to all the previous works on compressed sensing even with quantization context, we are here subsampling the random compressed sensing matrix *without replacement*. The first attention is to prove the restricted isometry property of the matrix $\frac{1}{\sqrt{\ell}}P_\ell V^*\Phi$, and then Theorem 7 yields the reconstruction error. Then in the proof of its restricted isometry property, the tube constraint is divided into three parts. One part of it can be reached by direct computation, another part is a direct application of the results on suprema of chaos process [39], the final part is treated by first solving a concentration problem, and then reaching the suprema by chaining argument and its bound by Dudley's inequality. Binding these three parts together shows that the product of the matrices satisfies the restricted isometry property with high property.

4.2 Notation and basic definitions

Denote by $[N]$ the set $\{1, \dots, N\}$. A vector $x \in \mathbb{R}^N$ is s-sparse if only s of its entries are non-vanishing, that is, its support $T = \operatorname{supp} x = \{j \in [N] : x_j \neq 0\}$ satisfies $|T| = s$. $F = \left(e^{2\pi ijk/N}\right)_{j,k=1}^{N}$ denotes the nonnormalized $N \times N$ discrete Fourier transform matrix, $\bar{F}$ the conjugate of F. That is, $F\bar{F} = \bar{F}F = NId$.

Given a vector $x \in \mathbb{R}$, we denote by $\hat{X} \in \mathbb{R}^{N \times N}$ the diagonal matrix with $\hat{x} := Fx$ on the diagonal. For a matrix A, A_k denotes its kth column.

We write $f \lesssim g$ for two functions f and g if they are defined on the same domain D and there exists an absolute constant C such that $f(y) \leq Cg(y)$ for all $y \in D$, $f \gtrsim g$ is defined analogously.

Given a full-rank matrix $A \in \mathbb{R}^{m \times d}$ with $m > d$, its pseudo-inverse is given by $A^\dagger = (A^*A)^{-1}A^*$.

4.2.1 Subgaussian random variable

Definition 7 (see, e.g., [54]). *A random variable X is called L-subgaussian if*

$$\mathbb{P}(|X| > t) \leq \exp(1 - t^2/L^2). \tag{4.1}$$

Up to absolute multiplicative constants, the subgaussian parameter L is equivalent to the subgaussian norm $\|X\|_{\Psi_2}$ defined as $\|X\|_{\Psi_2} = \sup_{p \geq 1} p^{-1/2}(\mathbb{E}|X|^p)^{1/p}$. Specifically, (4.1) *implies that [54]*

$$\|X\|_{\psi_2} \leq \sqrt{\tfrac{e}{2}}L. \tag{4.2}$$

4.2.2 Partial random circulant matrices

Given a vector $\xi = (\xi_1, \xi_2, \dots, \xi_N) \in \mathbb{R}^N$, the corresponding circulant matrix $C_\xi = C(\xi) \in \mathbb{R}^{N \times N}$ is defined by

$$C_\xi = \begin{bmatrix} \xi_1 & \xi_2 & \xi_3 & \cdots & \xi_N \\ \xi_N & \xi_1 & \xi_2 & \cdots & \xi_{N-1} \\ \vdots & & & & \vdots \\ \xi_2 & \xi_3 & \xi_4 & \cdots & \xi_1 \end{bmatrix}. \tag{4.3}$$

When ξ is a random vector and its entries are L-subgaussian random variables with variance 1 and mean 0, we call the corresponding matrix C_ξ a random circulant matrix. An $m \times N$ partial random circulant matrix with $m < N$ is obtained from an $N \times N$ random circulant matrix by sampling the rows of C_ξ. In this thesis, the rows of our compressed sensing matrix are m rows of a random circulant matrix C_ξ selected randomly (with equal probability), and without replacement.

More precisely, let $\Omega = (\Omega_1, \ldots, \Omega_m)$ be a random vector obtained by sampling from $[N] := \{1, \ldots, N\}$ without replacement, that is, Ω is drawn uniformly at random from the set

$$\Xi := \{\Omega \in [N]^m : \Omega_i \neq \Omega_j \text{ for } i \neq j\}. \tag{4.4}$$

The corresponding subsampling operator is then given by

$$\mathbb{R}^{m \times N} \ni R_\Omega = \sum_{j=1}^{m} e_j e_{\Omega_j}^*,$$

where e_k is the k-th standard basis vector, and we study measurement matrices of the form

$$\Phi = R_\Omega C_\xi.$$

Partial random circulant matrices are important to the practical application of compressed sensing. This is due to the simple observation that a circular convolution of a signal $x \in \mathbb{R}^N$ with $\widetilde{\xi} \in \mathbb{R}^N$ can be represented by the action of a circulant matrix C_ξ, i.e., by $C_\xi x$. And every linear time invariant (LTI) system is represented by convolution to a kernel function, and a convolution is then represented by a circulant matrix. For the generality of a LTI system, it is of importance to analyse the circulant matrix

in quantization compressed sensing. Defining $\xi \in \mathbb{R}^N$ via $\widetilde{\xi}_j = \xi_{N-j+1}$ for $j \in \{1, ..., N\}$ the matrix C_ξ is then as in (4.3). Consequently, partial random circulant matrices model subsampled random convolutions. Since linear time invariant system is so ubiquious in reality, convolutions in signal processing applications, or more precisely partial random circulant matrices (randomly selected convolutions) have played an important role in the development of compressed sensing applications such as radar imaging, Fourier optical imaging, and wireless channel estimation (see, e.g., [30, 45]).

Consequently, as the convolution is commutative, $C_\xi x = C_{\widetilde{x}}\widetilde{\xi}$. In the context in this paper, it is mathematically equivalent to analyse $C_x\xi$ instead of $C_{\widetilde{x}}\widetilde{\xi}$ through the map $\widetilde{\xi} \to \xi$ and $\widetilde{x} \to x$, where

$$C_x = (c_{i,j}) = \begin{bmatrix} x_1 & x_2 & x_3 & \cdots & x_{N-1} & x_N \\ x_2 & x_3 & x_4 & \cdots & x_N & x_1 \\ \vdots & & & & \vdots & \\ x_N & x_1 & x_2 & \cdots & x_{N-2} & x_{N-1} \end{bmatrix}. \tag{4.5}$$

Throughout this chapter, we will repeatedly use $C_x\xi$ rather than $C_\xi x$.

4.3 Probabilistic tools

We will use a number of different probabilistic tools for different parts of our argument. We state them here for convenience. The first one is a variation of McDiarmid's inequality.

Theorem 15 ([40], Theorem 3.14)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(\emptyset, \Omega) = \mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq ... \subseteq \mathcal{F}_m$ a filtration in $\mathcal{F}$. Consider a bounded random variable X, and set $X_k := \mathbb{E}(X|\mathcal{F}_k)$. Define the sum of squared conditional ranges*

$$R^2 = \sum_{k=1}^{m} ran_k^2$$

where

$$ran_k := \sup(X_k|\mathcal{F}_{k-1}) + \sup(-X_k|\mathcal{F}_{k-1}),$$

and denote its (essential) supremum by

$$\hat{r}^2 := \sup R^2.$$

Then,

$$\mathbb{P}(X - \mathbb{E}(X) \geq t) \leq e^{-2t^2/\hat{r}^2}.$$

A second tool that we will be using is Dudley's inequality introduced in Chapter 2.1. Here is a version with subgaussian random variables Z_x and distance $\|\cdot\|_{\Phi_2}$.

Theorem 16 (Dudley's inequality [22])**.** *Let Z_x be a random variable depending on $x \in T$, for some set T and define $d(x,y) = \|Z_x - Z_y\|_{\Psi_2}$, if*

$$\mathbb{P}(|Z_x - Z_y| > t) \lesssim \exp\Big(-t^2/\|Z_x - Z_y\|_{\Psi_2}^2\Big),$$

then for any $x_0 \in T$

$$\mathbb{P}(\sup|Z_x - Z_{x_0}| > t) \lesssim \exp\Big(-t^2/\big(\int_0^{\sup_{x\in D_{s,N}}\|Z_x\|_{\Psi_2}} \sqrt{\log \mathcal{N}(D_{s,N}, \dot{(}x,y), \epsilon)}d\epsilon\big)^2\Big).$$

A third result we will need concerns the subgaussian chaos processes. Whose original version involves the Talagrand γ_2 functional related to the generic chaining [52], which can be bounded in terms of covering numbers via Dudley's inequality (Theorem 16). A combined version in terms of only these upper bounds is stated for easier reading.

Theorem 17 ([39])**.** *Let $\mathcal{C}$ be a set of matrices and consider the complexity parameters*

$$d_F(\mathcal{C}) = \sup_{C\in\mathcal{C}}\|C\|_F, \qquad d_{2\to2}(\mathcal{C}) = \sup_{C\in\mathcal{C}}\|C\|_{2\to2}, \qquad D(\mathcal{C}) = \int_0^{d_{2\to2}(\mathcal{C})} \sqrt{\log \mathcal{N}(\mathcal{C}, \|\cdot\|_{2\to2}, u)}\, du.$$

Let ξ be a random vector whose entries ξ_j are independent, mean-zero, variance 1, L-subgaussian random variables. Then, for $t > 0$, the random variable

$$C_{\mathcal{C}}(\xi) = \sup_{C\in\mathcal{C}} |\|C\xi\|_2^2 - \mathbb{E}_\xi\|C\xi\|_2^2,$$

satisfies

$$\mathbb{P}(C_{\mathcal{C}}(\xi) \geq c_1 E + t) \leq 2\exp(-c_2 \min\{\frac{t^2}{V^2}, \frac{t}{U}\}),$$

where

$$E = D(\mathcal{C})(D(\mathcal{C}) + d_F(\mathcal{C})) + d_F(\mathcal{C})d_{2\to 2}(\mathcal{C}), \qquad V = d_{2\to 2}(\mathcal{C})(D(\mathcal{C}) + d_F(\mathcal{C})), \qquad U = d_{2\to 2}^2,$$

and the constants $c_1,\ c_2$ depend only on L.

4.4 Main results

In this chapter, we prove the following theorem, which is the main result of this chapter.

Theorem 18. *Denote by $Q_{\Sigma\Delta}^r$ a stable rth order $\Sigma\Delta$ quantizer. Let Φ be an $m \times N$ partial random circulant matrix associated to a vector with independent L-subgaussian entries with mean 0 and variance 1. Suppose that $N \geq m \geq (C\eta)^{\frac{1}{1-2\alpha}} s \log^{\frac{2}{1-2\alpha}} N \log^{\frac{2}{1-2\alpha}} s$, for some $\eta > 1$ and $\alpha \in [0, 1/2)$. With probability exceeding $1 - e^{-\eta}$, the following holds:*
For all $x \in \mathbb{R}^N$ with $\|\Phi x\|_\infty \leq \mu < 1$ and all $e \in \mathbb{R}^m$ with $\|e\|_\infty \leq \epsilon < 1 - \mu$ the estimate $\hat{x}$ obtained by solving (1.27) *satisfies*

$$\|\hat{x} - x\|_2 \leq C_1 \left(\frac{m}{\ell}\right)^{-r+1/2} \delta + C_2 \frac{\sigma_k(x)}{\sqrt{k}} + C_3 \sqrt{\frac{m}{\ell}} \epsilon.$$

Here C, C_1, C_2, C_3 are constants that only depend on r and L.

Proof. Theorem 18 can be immediately obtained from Theorem 7, which requires a bound on the restricted isometry constants of $P_\ell V^* R_\Omega C_\xi$ where $\ell = m(\frac{s}{m})^\alpha$, and Proposition 4 below, which provides the required bound. □

Proposition 4. *Consider the same set-up and assumptions as Theorem 18; in particular assume that $m \geq (C\eta)^{\frac{1}{1-2\alpha}} s \log^{\frac{2}{1-2\alpha}} N \log^{\frac{2}{1-2\alpha}} s$, for some $\eta > 1$ and $\alpha \in [0, 1/2)$. Setting $\ell = m(\frac{s}{m})^\alpha$, we have*

$$\mathbb{P}\Big(\sup_x |\|\frac{1}{\sqrt{\ell}} P_\ell V^* R_\Omega C_x \xi\|_2^2 - 1| > \frac{1}{9}\Big) < e^{-\eta},$$

where the supremum is over all s-sparse vectors. In other words, with probability exceeding $1 - e^{-\eta}$*, the matrix* $\frac{1}{\sqrt{\ell}} P_\ell V^* R_\Omega C_\xi$ *satisfies the restricted isometry property of order* s*, with constant* $1/9$*.*

Proof. Note that by the triangle inequality,

$$\begin{aligned}
&\sup_x \Big| \| \frac{1}{\sqrt{\ell}} P_\ell V^* R_\Omega C_x \xi \|_2^2 - 1 \Big| \\
&\leq \sup_x \Big(\Big| \| \frac{1}{\sqrt{\ell}} P_\ell V^* R_\Omega C_x \xi \|_2^2 - \mathbb{E}[\| \frac{1}{\sqrt{\ell}} P_\ell V^* R_\Omega C_x \xi \|_2^2 | \Omega] \Big| + \\
&\quad \Big| \mathbb{E}[\| \frac{1}{\sqrt{\ell}} P_\ell V^* R_\Omega C_x \xi \|_2^2 | \Omega] - \mathbb{E} \| \frac{1}{\sqrt{\ell}} P_\ell V^* R_\Omega C_x \xi \|_2^2 \Big| + \\
&| \mathbb{E} \| \frac{1}{\sqrt{\ell}} P_\ell V^* R_\Omega C_x \xi \|_2^2 - 1 | \Big). \qquad (4.6)
\end{aligned}$$

Thus, the proof of Proposition 4 is divided into controlling these three summands in (4.6). First, Lemma 3 (below) shows that by direct computation the third summand is bounded by $\frac{sm}{\ell N}$, while Lemma 4 and Lemma 5 bound the probability that the remaining summands exceed $\frac{1}{18}$ and $\frac{1}{36}$ respectively. Our bound on m (potentially with an increased value of C) ensures that $\frac{sm}{\ell N} \leq \frac{s}{\ell} = \left(\frac{s}{m}\right)^{1-\alpha} \leq \frac{1}{36}$ and the result follows using a union bound. □

Lemma 3. *Given the same set-up as in Theorem 18 and Proposition 4, one has*

$$| \mathbb{E} \| \frac{1}{\sqrt{\ell}} P_\ell V^* R_\Omega C_x \xi \|_2^2 - 1 | \leq \frac{(s-1)(m-\ell)}{\ell(N-1)} \leq \frac{sm}{\ell N}.$$

Proof. Denoting by $c_{i,j}$ the (i,j)-th entry of C_x and noting that we are sampling without replacement, we observe that for $p \neq q \in [m]$

$$\begin{aligned}
\mathbb{E}(c_{\Omega(p),k} c_{\Omega(q),k}) &= \frac{1}{N(N-1)} \sum_{u \neq v = 1}^{N} c_{u,k} c_{v,k} = \frac{1}{N(N-1)} \Big(\sum_{u,v=1}^{N} c_{u,k} c_{v,k} - \sum_{u=1}^{N} c_{u,k}^2 \Big) \\
&= \frac{1}{N(N-1)} \Big(\sum_{u,v=1}^{N} c_{u,k} c_{v,k} - \sum_{u=1}^{N} x_u^2 \Big) = \frac{1}{N(N-1)} \Big(\big(\sum_{u=1}^{N} x_u \big)^2 - 1 \Big). \qquad (4.7)
\end{aligned}$$

The last two equalities both use the fact that each row of C_x is a shifted copy of x. Furthermore

$$\begin{aligned}\Big|\mathbb{E}\|\frac{1}{\sqrt{\ell}}P_\ell V^* R_\Omega C_x \xi\|_2^2 - 1\Big| &= \Big|\mathbb{E}\|\frac{1}{\sqrt{\ell}}P_\ell V^* R_\Omega C_x\|_F^2 - 1\Big| \\ &= \Big|\frac{1}{\ell}\mathbb{E}\sum_{j=1}^{\ell}\sum_{k=1}^{N}|\sum_{p=1}^{m} v_{jp} c_{\Omega(p),k}|^2 - 1\Big| \\ &= \Big|\frac{1}{\ell}\sum_{j=1}^{\ell}\sum_{k=1}^{N}\big(\sum_{p=1}^{m} v_{jp}^2 \mathbb{E}c_{\Omega(p),k}^2 + \sum_{\substack{p,q=1\\p\neq q}}^{m} v_{jp}v_{jq}\mathbb{E}c_{\Omega(p),k}c_{\Omega(q),k}\big) - 1\Big| \\ &= \Big|\frac{1}{\ell}\sum_{j=1}^{\ell}\big(1 + \frac{(\sum_{i=1}^{N} x_i)^2 - 1}{N-1}\sum_{\substack{p,q=1\\p\neq q}}^{m} v_{jp}v_{jq}\big) - 1\Big|.\end{aligned}$$

where in the last equality we used (4.7) and the fact that the rows of both C_x and V are normalized. Using that x is s-sparse, it follows that

$$\begin{aligned}\Big|\mathbb{E}\|\frac{1}{\sqrt{\ell}}P_\ell V^* R_\Omega C_x \xi\|_2^2 - 1\Big| &\leq \Big|\frac{s-1}{\ell(N-1)}\Big(\sum_{j=1}^{\ell}(\sum_{p=1}^{m} v_{jp})^2 - \sum_{j=1}^{\ell}\sum_{p=1}^{m} v_{jp}^2\Big)\Big| \\ &= \frac{s-1}{\ell(N-1)}\Big|\|V^*(1,\dots,1)^T\|_2^2 - \ell\Big| \\ &\leq \frac{s-1}{\ell(N-1)}\Big|\|V\|_{2\to 2}^2 m - \ell\Big| \\ &= \frac{(s-1)(m-\ell)}{\ell(N-1)}.\end{aligned}$$

□

Lemma 4. *Consider again the set-up of Theorem 18 and Proposition 4 and denote by $D_{N,s}$ the set of all s-sparse vectors in $\mathbb{R}^N$. Then*

$$\mathbb{P}\Big(\sup_{x\in D_{s,N}}\Big|\|\frac{1}{\sqrt{\ell}}P_\ell V^* R_\Omega C_x \xi\|_2^2 - \mathbb{E}_\xi\Big[\|\frac{1}{\sqrt{\ell}}P_\ell V^* R_\Omega C_x \xi\|_2^2\Big|\Omega\Big]\Big| > \frac{1}{18}\Big|\Big) \leq \frac{1}{2}e^{-\eta}.$$

Proof. We will apply Theorem 17 conditionally given Ω with $\mathcal{C} = \{\frac{1}{\sqrt{\ell}}P_\ell V^* R_\Omega C_x : x \in D_{s,N}\}$. This set is almost the same as the one considered in the proof of Theorem 4.1 in [39], the only differences being the additional projection P_ℓ and our normalization factor of $\frac{1}{\sqrt{\ell}}$ (instead of $\frac{1}{\sqrt{m}}$ in [39]). Indeed, since

$\|P_\ell\|_{2\to 2} \le 1$ we can estimate the necessary parameters for applying Theorem 17 exactly as in the proof of Theorem 4.1 in [39]. This yields

$$d_{2\to 2}(\mathcal{C}) \le \sqrt{\frac{s}{\ell}}, \qquad d_F(\mathcal{C}) \le \sqrt{\frac{m}{\ell}}, \qquad D(\mathcal{C}) \le \sqrt{\frac{s}{\ell}} \log N \log s.$$

Consequently for c_1, c_2, and E as in Theorem 17, we have

$$\begin{aligned} E &\le \sqrt{\frac{s}{\ell}} \log N \log s \left(\sqrt{\frac{s}{\ell}} \log N \log s + \sqrt{\frac{m}{\ell}} \right) + \sqrt{\frac{m}{\ell}} \sqrt{\frac{s}{\ell}} \\ &\le \left(\frac{s}{m}\right)^{1-\alpha} \log^2 N \log^2 s + 2 \left(\frac{s}{m}\right)^{1-2\alpha} \log N \log s \le \frac{1}{36 c_1}. \end{aligned}$$

Here, the second inequality follows from our choice of ℓ and the last inequality follows from our assumption on m in Theorem 18 (potentially adjusting the constant C). Again adjusting the constant, we similarly obtain

$$V \le \sqrt{\frac{c_2}{4\eta}} \qquad \text{and} \qquad U \le \frac{c_2}{4\eta}.$$

Hence the probability is bounded by $2e^{-4\eta}$. Finally, as $\eta \ge 1$, $e^{-4\eta} \le \frac{1}{4} e^{-\eta}$ and the result follows by taking the expectation over Ω.

□

Lemma 5. *With the same notation as before, we have*

$$\begin{aligned} &\mathbb{P}\Big(\sup_{x \in D_{s,N}} |\mathbb{E}[\| \frac{1}{\sqrt{\ell}} P_\ell V^* R_\Omega C_x \xi \|_2^2 | \Omega] - \mathbb{E} \| \frac{1}{\sqrt{\ell}} P_\ell V^* R_\Omega C_x \xi \|_2^2 | > \frac{1}{36} \Big) \\ &\le C' \exp(-c/(\frac{\sqrt{sm}}{\ell} \log N \log m)^2) \le \frac{1}{2} e^{-\eta}, \end{aligned}$$

where c, C' are constants that depends only on L.

Proof. The proof is a direct application of Theorem 16 for the random variable

$$Z_x := \mathbb{E}\left[\| \frac{1}{\sqrt{\ell}} P_\ell V^* R_\Omega C_x \xi \|_2^2 - \mathbb{E} \| \frac{1}{\sqrt{\ell}} P_\ell V^* R_\Omega C_x \xi \|_2^2 \middle| \Omega \right] = \| \frac{1}{\sqrt{\ell}} P_\ell V^* R_\Omega C_x \|_F^2 - \mathbb{E} \| \frac{1}{\sqrt{\ell}} P_\ell V^* R_\Omega C_x \|_F^2$$

to find the supremum of the deviation. Since Theorem 16 requires the covering number with respect to the metric $d(x,y) := \|Z_x - Z_y\|_{\Psi_2}$ we need a bound for $d(x,y)$, which we provide in Lemma 7 below. Specifically, the first inequality in Lemma 5 follows from Theorem 16 together with Lemma 3 and Lemma 4 above. Indeed, applying Lemma 7 with $y = 0$ yields

$$\sup_{x,y} \|Z_x\|_{\Psi_2} \leq \frac{\sqrt{m}}{\ell}\|x\|_{\hat{\infty}} \leq \frac{\sqrt{m}}{\ell}\|F(x)\|_\infty \leq \frac{\sqrt{m}}{\ell}\|x\|_1 \leq \frac{\sqrt{sm}}{\ell}\|x\|_2 \leq \frac{\sqrt{sm}}{\ell}. \tag{4.8}$$

To bound the integral in Theorem 16, we note that

$$\mathcal{N}(D_{s,N}, \frac{\sqrt{m}}{\ell}\|\cdot\|_{\hat{\infty}}, \epsilon) = \mathcal{N}(D_{s,N}, \frac{1}{\sqrt{m}}\|\cdot\|_{\hat{\infty}}, \frac{\ell}{m}\epsilon),$$

and hence applying the argument in [39, Section 4] scaled by $\frac{m}{\ell}$, (a detailed calculation stated in Lemma 8)

$$\begin{aligned}&\int_0^{\sup_x \|Z_x\|_{\Psi_2}} \sqrt{\log \mathcal{N}(D_{s,N}, \frac{1}{\sqrt{m}}\|\cdot\|_{\hat{\infty}}, \frac{\ell}{m}\epsilon)}d\epsilon \\ &\lesssim \frac{\sqrt{sm}}{\ell}\log N \log s.\end{aligned}$$

For the second inequality note that by the definition of ℓ and the assumed lower bound on m

$$(\frac{\sqrt{sm}}{\ell}\log N \log s)^2 = \left(\frac{s}{m}\right)^{1-2\alpha}\log^2 N \log^2 s \tag{4.9}$$

$$\leq C^{-1}\eta^{-1}. \tag{4.10}$$

The result follows from the assumption that $\eta \geq 1$ as in the proof of Lemma 4. □

All that remains now is to prove Lemma 7. Before that, we derive a technical bound required for its proof.

Lemma 6. *Let Ω, $\Omega' \in \Xi = \{\Omega \in [N]^m : \Omega_i \neq \Omega_j \text{ for } i \neq j\}$ be such that Ω differs from Ω' in at most two components. Then the function*

$$f(\Omega) := \|\frac{1}{\sqrt{\ell}}P_\ell V^* R_\Omega C_x\|_F^2 - \|\frac{1}{\sqrt{\ell}}P_\ell V^* R_\Omega C_y\|_F^2$$

satisfies

$$|f(\Omega) - f(\Omega')| \leq \frac{24}{\ell}\|x - y\|_{\hat{\infty}},$$

where $\|x\|_{\hat{\infty}} := \|Fx\|_\infty$.

Proof. Note that, as a circulant matrix is diagonalized by the Fourier transform,

$$\begin{aligned} f(\Omega) &= \|\frac{1}{\sqrt{\ell}} P_\ell V^* R_\Omega C_x\|_F^2 - \|\frac{1}{\sqrt{\ell}} P_\ell V^* R_\Omega C_y\|_F^2 \\ &= \|\frac{1}{\sqrt{\ell}} P_\ell V^* R_\Omega F^{-1}\hat{X}F\|_F^2 - \|\frac{1}{\sqrt{\ell}} P_\ell V^* R_\Omega F^{-1}\hat{Y}F\|_F^2 \\ &= \frac{1}{\ell N}\|P_\ell V^* R_\Omega \overline{F}\hat{X}\|_F^2 - \frac{1}{\ell N}\|P_\ell V^* R_\Omega \overline{F}\hat{Y}\|_F^2 \\ &= \frac{1}{\ell N}\sum_{k=1}^{N}\left(|\hat{x}_k|^2 - |\hat{y}_k|^2\right)\|P_\ell V^* R_\Omega \overline{F}_k\|_2^2, \end{aligned} \tag{4.11}$$

where F denotes the non-normalized Fourier transform, F_k^T its k-th row, and $\hat{x} = Fx$.

We first consider the case that Ω and Ω' differ only in one component, say the first (without loss of generality). To bound $|f(\Omega) - f(\Omega')|$ for this case, we note that for V_j^T denoting the j-th row of V, and $\eta = \exp(-\frac{2\pi i}{N})$ an N-th root of unity,

$$\begin{aligned} &\|P_\ell V R_\Omega \overline{F}_k\|_2^2 - \|P_\ell V R_{\Omega'} \overline{F}_k\|_2^2 \\ &= \sum_{p,q=1}^{m} \langle \eta^{-k\Omega_p} P_\ell V_p, \eta^{-k\Omega_q} P_\ell V_q\rangle - \sum_{r,s=1}^{m} \langle \eta^{-k\Omega'_r} P_\ell V_r, \eta^{-k\Omega'_s} P_\ell V_s\rangle \\ &= \sum_{p,q=1}^{m} (\eta^{k(\Omega_p - \Omega_q)} - \eta^{k(\Omega'_p - \Omega'_q)})\langle P_\ell V_p, P_\ell V_q\rangle \\ &= (\eta^{k(\Omega_1 - \Omega_1)} - \eta^{k(\Omega'_1 - \Omega'_1)})\langle P_\ell V_1, P_\ell V_1\rangle + \sum_{q=2}^{m} (\eta^{k(\Omega_1 - \Omega_q)} - \eta^{k(\Omega'_1 - \Omega_q)})\langle P_\ell V_1, P_\ell V_q\rangle \\ &\quad + \sum_{p=2}^{m} (\eta^{k(\Omega_p - \Omega_1)} - \eta^{k(\Omega_p - \Omega'_1)})\langle P_\ell V_p, P_\ell V_1^*\rangle + \sum_{p,q=2}^{m} (\eta^{k(\Omega_p - \Omega_q)} - \eta^{k(\Omega_p - \Omega_q)})\langle P_\ell V_p, P_\ell V_q\rangle \\ &= \sum_{q=2}^{m} (\eta^{k\Omega_1} - \eta^{k\Omega'_1})\eta^{-k\Omega_q}\langle P_\ell V_1, P_\ell V_q\rangle + \sum_{p=2}^{m} (\eta^{-k\Omega_1} - \eta^{-k\Omega'_1})\eta^{k\Omega_p}\langle P_\ell V_p, P_\ell V_1\rangle. \end{aligned}$$

Combining this with (4.11), we obtain

$$f(\Omega) - f(\Omega') = \frac{1}{\ell N} \sum_{k=1}^{N} \big(|\hat{x}_k|^2 - |\hat{y}_k|^2\big)\Big(\sum_{q=2}^{m} (\eta^{k\Omega_1} - \eta^{k\Omega'_1})\eta^{-k\Omega_q} \langle P_\ell V_1, P_\ell V_q\rangle + \sum_{p=2}^{m} (\eta^{-k\Omega_1} - \eta^{-k\Omega'_1})\eta^{k\Omega_p}\langle P_\ell V_p, P_\ell V_1\rangle \Big). \tag{4.12}$$

Observe that the right hand side is a sum of four different rescaled Fourier coefficients of the vector $u \in \mathbb{R}^N$ given by $u_k := |\hat{x}_k|^2 - |\hat{y}_k|^2$, as for example

$$\frac{1}{\ell N} \sum_{p=2}^{m} \langle P_\ell V_1, P_\ell V_p\rangle \sum_{k=1}^{N} (|\hat{x}_k|^2 - |\hat{y}_k|^2)\eta^{k(\Omega_p - \Omega_1)} = \frac{1}{\ell N} \sum_{p=2}^{m} \langle P_\ell V_1, P_\ell V_p\rangle (\overline{F}u)_{\Omega_p - \Omega_1} = V_1^* P_\ell^* P_\ell V^* v,$$

where $v \in \mathbb{R}^m$ is given by $v_1 = 0$ and $v_p = (\overline{F}u)_{\Omega_p - \Omega_1}$ for $2 \le p \le m$. Note that as $\Omega \in \Xi$ and hence the Ω_q are all different, v is a projection of $\overline{F}u$ on a subset of its entries, and so $\|v\|_2 \le \sqrt{N}\|u\|_2$. Note that in this step, it is crucial to sample without replacement, as otherwise, the bound would no longer hold. Consequently, using the Cauchy-Schwartz inequality, and the fact that

$$\begin{aligned} |u_k| = \big||\hat{x}_k| - |\hat{y}_k|\big|(|\hat{x}_k| + |\hat{y}_k|) &\le \big|\hat{x}_k - \hat{y}_k\big|(|\hat{x}_k| + |\hat{y}_k|) \\ &\le \sqrt{2}\|x - y\|_\infty (|\hat{x}_k|^2 + |\hat{y}_k|^2), \end{aligned}$$

therefore

$$\|u\|_2 \le \sqrt{2}\|x - y\|_\infty \sqrt{\|\hat{x}\|_2^2 + \|\hat{y}\|_2^2} \le 2\sqrt{N}\|x - y\|_\infty,$$

and

$$\begin{aligned} &\tfrac{1}{\ell N}\Big| \sum_{p=2}^{m} \langle P_\ell V_1, P_\ell V_p\rangle \sum_{k=1}^{N} \big(|\hat{x}_k|^2 - |\hat{y}_k|^2\big)\eta^{k(\Omega_p - \Omega_1)}\Big| \\ &\le \tfrac{1}{\ell N}\|V\|_{2\to 2}\|P_\ell^* P_\ell V_1\|_2 \|v\|_2 \\ &\le \tfrac{1}{\ell\sqrt{N}}\|u\|_2 \\ &\le \tfrac{2}{\ell}\|x - y\|_\infty. \end{aligned} \tag{4.13}$$

Similar to the other three summands in (4.12), which yields the result for Ω and Ω' differing in only one component with a scale 8 to the result (4.13), i.e.,

$$|f(\Omega) - f(\Omega')| \leq \frac{8}{\ell}\|x - y\|_{\hat{\infty}}.$$

If Ω and Ω' differ in two components (without loss of generality say in first and second component), one can add and minus two middle terms, (say Ω^{HALLO} and Ω^{HELLO}) each of which differs from Ω and Ω' in only one component,

$$\begin{aligned}
\Omega &= (\Omega_1, \Omega_2, ...) \\
\Omega' &= (\Omega'_1, \Omega'_2, ...) \\
\Omega^{HALLO} &= (\Omega'_1, \Omega_2, ...) \\
\Omega^{HELLO} &= (\Omega_1, \Omega'_2, ...),
\end{aligned}$$

then

$$\begin{aligned}
&|f(\Omega) - f(\Omega')| \leq \frac{8}{\ell}\|x - y\|_{\hat{\infty}} \\
&\leq |f(\Omega) - f(\Omega^{HALLO}) + f(\Omega^{HALLO}) - f(\Omega^{HELLO}) + f(\Omega^{HELLO}) - f(\Omega')| \\
&\leq |f(\Omega) - f(\Omega^{HALLO})| + |f(\Omega^{HALLO}) - f(\Omega^{HELLO})| + |f(\Omega^{HELLO}) - f(\Omega')| \\
&\leq \frac{24}{\ell}\|x - y\|_{\hat{\infty}},
\end{aligned}$$

i.e., use triangle inequality we have a scale of 24 (3 times 8). □

We are now ready to bound the distance $d(x, y) = \|x - y\|_{\Psi_2}$.

Lemma 7. *For all $x, y \in \mathbb{R}^N$ it holds that*

$$d(x, y) \leq \frac{24\sqrt{m}}{\ell}\|x - y\|_{\hat{\infty}}.$$

Proof. By (4.2), it suffices to show that for all $t \geq 0$,

$$\mathbb{P}_\Omega(|Z_x - Z_y| > t) \leq \exp\left(1 - t^2/(\frac{24\sqrt{m}}{\ell}\|x-y\|_\infty)^2\right), \tag{4.14}$$

where again with

$$Z_x := \mathbb{E}\left[\|\frac{1}{\sqrt{\ell}}P_\ell V^* R_\Omega C_x \xi\|_2^2 - \mathbb{E}\|\frac{1}{\sqrt{\ell}}P_\ell V^* R_\Omega C_x \xi\|_2^2 \middle| \Omega\right] = \|\frac{1}{\sqrt{\ell}}P_\ell V^* R_\Omega C_x\|_F^2 - \mathbb{E}\|\frac{1}{\sqrt{\ell}}P_\ell V^* R_\Omega C_x\|_F^2.$$

It is proved by applying Theorem 15 with $\mathcal{F}_k$, the σ-algebra generated by $\Omega_1, ..., \Omega_k$ to the function $f(\Omega)$ as defined above

$$f(\Omega) = \|\frac{1}{\sqrt{\ell}}P_\ell V^* R_\Omega C_x\|_F^2 - \|\frac{1}{\sqrt{\ell}}P_\ell V^* R_\Omega C_y\|_F^2.$$

Assuming $(\Omega'_k, ..., \Omega'_m)$ an independent copy of $(\Omega_k, ..., \Omega_m)$, we denote $\Omega' = (\Omega_1, ...\Omega_{k-1}, \Omega'_k, ..., \Omega'_m)$, and $\Omega = (\Omega_m, ..., , \Omega_{k-1}, \Omega_k, ..., \Omega_1)$. And then we need to bound the sum of squared ranges

$$R^2 = \sup \sum_{j=1}^{m} ran_k^2$$

By definition,

$$X_k := \mathbb{E}(X|\mathcal{F}_k) = \mathbb{E}(f(\Omega)|\Omega_k, \ldots, \Omega_1),$$

and

$$
\begin{aligned}
ran_k := & \sup_{\Omega_k \notin \{\Omega_1,\ldots,\Omega_{k-1}\}} \Big(X_k \Big| \Omega_{k-1},\ldots,\Omega_1 \Big) + \sup_{\Omega_k \notin \{\Omega_1,\ldots,\Omega_{k-1}\}} \Big(- X_k \Big| \Omega_{k-1},\ldots,\Omega_1 \Big) \\
= & \sup_{\Omega_k \notin \{\Omega_1,\ldots,\Omega_{k-1}\}} \Big(\mathbb{E}(f(\Omega)|\Omega_k,\ldots,\Omega_1) \Big| \Omega_{k-1},\ldots,\Omega_1 \Big) + \qquad (4.15) \\
& \sup_{\Omega_k \notin \{\Omega_1,\ldots,\Omega_{k-1}\}} \Big(\mathbb{E}(-f(\Omega')|\Omega_k,\ldots,\Omega_1) \Big| \Omega_{k-1},...,\Omega_1 \Big) \\
= & \sup_{\Omega_k,\Omega'_k \notin \{\Omega_1,\ldots,\Omega_{k-1}\}} \Big(\mathbb{E}(f(\Omega)|\Omega_k,\Omega_{k-1}...,\Omega_1) + \mathbb{E}(-f(\Omega')|\Omega'_k,\Omega_{k-1},...,\Omega_1) \Big| \Omega_{k-1},...,\Omega_1 \Big) \qquad (4.16) \\
= & \sup_{\Omega_k,\Omega'_k \notin \{\Omega_1,\ldots,\Omega_{k-1}\}} \Big[\Big(\mathbb{E}(f(\Omega)|\Omega_k,\Omega_{k-1}...,\Omega_1) \Big| \Omega_{k-1},...,\Omega_1 \Big) + \Big(\mathbb{E}(-f(\Omega')|\Omega'_k,\Omega_{k-1},...,\Omega_1) \Big) \Big]. \\
& \qquad (4.17)
\end{aligned}
$$

Now it is essential if we can bound $\mathbb{E}(f(\Omega)|\Omega_k,\Omega_{k-1}...,\Omega_1) + \mathbb{E}(-f(\Omega')|\Omega'_k,\Omega_{k-1},...,\Omega_1)$, conditional on $\Omega_{k-1},...,\Omega_1$ from above. It is expected that we can bound the term by bounding the combination of the two summands by

$$\mathbb{E}[f(\Omega) - f(\Omega')|\Omega_k,\Omega'_k,\Omega_{k-1},...,\Omega_1].$$

However this cannot be done in one glance (at least for me), since we are sampling without replacement, while calculating the expectation over $\Omega'_j s$ for $m \geq j > k$, the space $\{\Omega_{k+1},\ldots,\Omega_m\}$ is different from $\{\Omega'_{k+1},\ldots,\Omega'_m\}$, for there can be some $i > k$, such that $\Omega_i = \Omega'_k$, and vice versa for some $i > k$, $\Omega'_i = \Omega_k$. Therefore $\mathbb{E}(f(\Omega)|\Omega_k,\Omega_{k-1}...,\Omega_1)$ cannot immerses with $\mathbb{E}(f(\Omega')|\Omega'_k,\Omega_{k-1}...,\Omega_1)$ in one step. This is then analysed by dividing the space generated by $(\Omega_i)_{i>k}$ into partition events $(\mathcal{E}_j)_{j=1}^{m-k}$ and $(\mathcal{E}'_j)_{j=1}^{m-k}$ defined in the next paragraph.

Define the events $\mathcal{E}_0 = \{\Omega_j \neq \Omega'_k \ \forall j > k\}$, $\mathcal{E}'_0 = \{\Omega'_j \neq \Omega_k \ \forall j > k\}$, and, for $j \in [m-k]$, $\mathcal{E}_j = \{\Omega_{k+j} = \Omega'_k\}$, $\mathcal{E}'_j = \{\Omega'_{k+j} = \Omega_k\}$ and note that

$$\mathbb{P}[\cup_{j=0}^{m-k} \mathcal{E}_j | \Omega_1,...,\Omega_k,\Omega'_k] = \mathbb{P}[\cup_{j=0}^{m-k} \mathcal{E}'_j | \Omega_1,...,\Omega_k,\Omega'_k] = 1. \qquad (4.18)$$

Which says that events $(\mathcal{E}_j)_{j=1}^{m-k}$ and $(\mathcal{E}'_j)_{j=1}^{m-k}$ are two partitions of the probability space conditional

on $\{\Omega_1, \ldots, \Omega'_k, \Omega_k\}$, and the measure (probability) of each pair of events

$$\big((\mathcal{E}_j|\Omega_1, ..., \Omega_k, \Omega'_k), (\mathcal{E}'_j|\Omega_1, ..., \Omega_k, \Omega'_k)\big)_{j=1}^{m-k}$$

is the same, i.e.,

$$\mathbb{P}[\mathcal{E}_j|\Omega_1, ..., \Omega_k, \Omega'_k] = \mathbb{P}[\mathcal{E}'_j|\Omega_1, ..., \Omega_k, \Omega'_k], \text{ for } j = 1, ..., m-k. \tag{4.19}$$

Now, we can write

$$\mathbb{E}[f(\Omega)|\Omega_1, ..., \Omega_{k-1}, \Omega_k] = \sum_{j=0}^{m-k} \mathbb{E}[f(\Omega)\mathbb{1}_{\mathcal{E}_j}|\Omega_1, ..., \Omega_{k-1}, \Omega_k, \Omega'_k], \tag{4.20}$$

and similarly

$$\mathbb{E}[f(\Omega')|\Omega_1, ..., \Omega_{k-1}, \Omega_k] = \sum_{j=0}^{m-k} \mathbb{E}[f(\Omega')\mathbb{1}_{\mathcal{E}'_j}|\Omega_1, ..., \Omega_{k-1}, \Omega'_k, \Omega_k]. \tag{4.21}$$

Put (4.20), (4.21) together, we have

$$\begin{aligned}&\mathbb{E}[f(\Omega)|\Omega_1, ..., \Omega_{k-1}, \Omega_k] - \mathbb{E}[f(\Omega')|\Omega_1, ..., \Omega_{k-1}, \Omega_k] \\ &\quad = \sum_{j=0}^{m-k} \mathbb{E}[f(\Omega)\mathbb{1}_{\mathcal{E}_j} - f(\Omega')\mathbb{1}_{\mathcal{E}'_j}|\Omega_1, ..., \Omega_{k-1}, \Omega_k, \Omega'_k].\end{aligned} \tag{4.22}$$

It remains to bound the term

$$f(\Omega)\mathbb{1}_{\mathcal{E}_j} - f(\Omega')\mathbb{1}_{\mathcal{E}'_j}|\Omega_1, ..., \Omega_{k-1}, \Omega'_k, \Omega_k.$$

Note that due to the partition of the events, for $j = 0$, i.e., in events $\mathcal{E}_0$, $\mathcal{E}'_0$, Ω and Ω' differ at most in one component, i.e., the kth component,or equivalently Ω_k and Ω'_k can be different. Thus by Lemma 6,

$$f(\Omega)\mathbb{1}_{\mathcal{E}_j} - f(\Omega')\mathbb{1}_{\mathcal{E}'_j}|\Omega_1, ..., \Omega_{k-1}, \Omega'_k, \Omega_k \leq \frac{24}{\ell}\|x - y\|_\infty. \tag{4.23}$$

For $j > 0$, in events $\mathcal{E}_j$, $\mathcal{E}'_j$, Ω and Ω' differ at most in two components, i.e., the kth and the $k+j$th. Thus by Lemma 6,

$$f(\Omega)\mathbb{1}_{\mathcal{E}_j} - f(\Omega')\mathbb{1}_{\mathcal{E}'_j}|\Omega_1, ..., \Omega_{k-1}, \Omega'_k, \Omega_k \leq \frac{24}{\ell}\|x-y\|_{\hat{\infty}}. \tag{4.24}$$

Hence for all $j = 1, ..., m-k$,

$$\mathbb{E}[f(\Omega)\mathbb{1}_{\mathcal{E}_j} - f(\Omega')\mathbb{1}_{\mathcal{E}'_j}|\Omega_1, ..., \Omega_{k-1}, \Omega_k, \Omega'_k] \leq \frac{24}{\ell}\|x-y\|_{\hat{\infty}}\mathbb{P}[\mathcal{E}_j|\Omega_1, ..., \Omega_k, \Omega'_k]. \tag{4.25}$$

The above inequality (4.25) and the fact from space partition (4.18), (4.19), we have

$$\begin{aligned}
ran_k &= \sup_{\Omega_k, \Omega'_k \notin \{\Omega_1, ..., \Omega_{k-1}\}} \Big[\Big(\mathbb{E}(f(\Omega)|\Omega_k, \Omega_{k-1}..., \Omega_1)\Big|\Omega_{k-1}, ..., \Omega_1\Big) + \Big(\mathbb{E}(-f(\Omega')|\Omega'_k, \Omega_{k-1}, ..., \Omega_1)\Big)\Big] \\
&\leq \sup_{\Omega_k, \Omega'_k \notin \{\Omega_1, ..., \Omega_{k-1}\}} \Big[\sum_{j=0}^{m-k} \mathbb{E}[f(\Omega)\mathbb{1}_{\mathcal{E}_j} - f(\Omega')\mathbb{1}_{\mathcal{E}'_j}|\Omega_1, ..., \Omega_{k-1}, \Omega_k, \Omega'_k]\Big] \\
&\leq \sup_{\Omega_k, \Omega'_k \notin \{\Omega_1, ..., \Omega_{k-1}\}} \Big[\sum_{j=0}^{m-k} \frac{24}{\ell}\|x-y\|_{\hat{\infty}}\mathbb{P}[\mathcal{E}_j|\Omega_1, ..., \Omega_k, \Omega'_k]\Big] \leq \frac{24}{\ell}\|x-y\|_{\hat{\infty}}.
\end{aligned}$$

Now applying Theorem 15 with $\hat{r}^2 := \sup R^2 \leq \sum_{k=1}^m ran_k^2 \leq (\frac{24\sqrt{m}}{\ell}\|x-y\|_{\hat{\infty}})^2$, one obtains

$$\mathbb{P}(|Z_x - Z_y| > t) \leq 2\exp(-t^2/(\frac{24\sqrt{m}}{\ell}\|x-y\|_{\hat{\infty}})^2),$$

which implies (4.14). We conclude

$$d(x,y) := \|Z_x - Z_y\|_{\Psi_2} \leq \frac{24\sqrt{m}}{\ell}\|x-y\|_{\hat{\infty}},$$

as desired. □

Now the details for calculating the Dudley's integral is demonstrated below.

Lemma 8.

$$\int_0^{\sup_x \|Z_x\|_{\Psi_2}} \sqrt{\log \mathcal{N}(D_{s,N}, \frac{1}{\sqrt{m}}\|\cdot\|_\infty, \frac{\ell}{m}\epsilon)}d\epsilon$$
$$\lesssim \frac{\sqrt{sm}}{\ell} \log N \log s.$$

Proof. The integral is calculated by three parts. First the Maurey's method for larger integral factor ϵ, secondly the volumetric argument for smaller ϵ, and insert both the above into the Dudley's integration.

In Maurey's method, set $\mathcal{U} = \{\pm\sqrt{2}e_1, \pm\sqrt{2}e_2, \cdots, \pm\sqrt{2}e_N\}$, and $\|\cdot\|_X = \|\cdot\|_\infty = \|\sqrt{N}Fx\|_\infty = \max_{k\in[N]}\langle\sqrt{N}F^k, x\rangle$. Then $B^N(0,1) \subset conv(\mathcal{U})$, and

$$\mathbb{E}\|\sum_{k=1}^M \epsilon_k \mathrm{Z}_k\|_X$$
$$= \mathbb{E}\max_{k=1\ldots N}|\sum_{k=1}^M \epsilon_k\langle\sqrt{N}F^p, Z_k\rangle|.$$

Since by Hölder's inequality $|\langle\sqrt{N}F^p, Z_k\rangle| \le \|\sqrt{N}F^p\|_\infty\|Z_k\|_1 \le 1\cdot\sqrt{2}$. Note that the norm $\|\cdot\|_\infty$ is defined by nonnormalized discrete Fourier matrix because $\|\sqrt{N}F^p\|_\infty$, and it should be a constant independent of N.

$$\|(\langle\sqrt{N}F^p, Z_k\rangle)_{k=1}^M\|_2 \le \sqrt{2M},$$

for $k \in [N]$ By Höffding's inequality, conditional on Z_k,

$$\mathbb{P}_\epsilon(|\sum_{k=1}^M \epsilon_k\langle\sqrt{N}F^p, Z_k\rangle| \ge \sqrt{2M}t) \le 2e^{t^2/2},\ t > 0.$$

By union bound

$$\mathbb{P}_\epsilon(\max_{p\in[N]}|\sum_{k=1}^M \epsilon_k\langle\sqrt{N}F^p, Z_k\rangle| \ge \sqrt{2M}t) \le 2Ne^{t^2/2},\ t > 0.$$

By Proposition 3, we have

$$\mathbb{E}_\epsilon \max_{p\in[N]}|\sum_{k=1}^M \epsilon_k\langle\sqrt{N}F^p, Z_k\rangle| \le 3/2\sqrt{2}\sqrt{\ln 8N}\sqrt{M},$$

then by Fubini's theorem,

$$\mathbb{E}\max_{p\in[N]}|\sum_{k=1}^{M}\epsilon_k\langle\sqrt{N}F^p,Z_k\rangle|\leq 3/2\sqrt{2}\sqrt{\ln 8N}\sqrt{M}.$$

Hence by letting $A=3/2\sqrt{2}\sqrt{\ln 8N}\lesssim\sqrt{\ln N}$, Maurey's method yields

$$\log\mathcal{N}(conv(\mathcal{U}),\|\cdot\|_X,\epsilon)\lesssim(\frac{1}{\epsilon})^2\ln^2 N.$$

Therefore

$$\begin{aligned}
&\log\mathcal{N}(D_{s,N},\|\cdot\|_{\Psi_2},\epsilon)\lesssim\log\mathcal{N}(D_{s,N},\frac{\sqrt{m}}{\ell}\|\cdot\|_{\hat{\infty}},\epsilon)\\
&\leq\log\mathcal{N}(\sqrt{s}B_1^N(0,1),\frac{\sqrt{m}}{\ell}\|\cdot\|_{\hat{\infty}},\epsilon)=\log\mathcal{N}(B_1^N(0,1),\|\cdot\|_{\hat{\infty}},\frac{\ell}{\sqrt{s}\sqrt{m}}\epsilon)\\
&\leq\log\mathcal{N}(conv(\mathcal{U}),\|\cdot\|_{\hat{\infty}},\frac{\ell}{\sqrt{s}\sqrt{m}}\epsilon)\\
&\lesssim(\frac{\sqrt{sm}}{\ell\epsilon})^2\ln^2 N.
\end{aligned}\tag{4.26}$$

Volumetric argument reveals that

$$\begin{aligned}
\log\mathcal{N}(D_{s,N},d(x,x_0),\epsilon)&\leq\log\mathcal{N}(\sqrt{s}D_{s,N}^1,\|\cdot\|_{\hat{\infty}},\epsilon)\\
&\leq\log\binom{N}{s}\mathcal{N}(B_1^s(0,1),\|\cdot\|_\infty,\epsilon)\\
&\leq\log(\frac{eN}{s})^s\mathcal{N}(B_{\hat{\infty}}^s(0,1),\|\cdot\|_{\hat{\infty}},\epsilon)\\
&\leq\log(\frac{eN}{s})^s(1+\frac{2}{\epsilon})^s\\
&=\log(\frac{eN}{s}+\frac{2eN}{s\epsilon})^s\\
&=s\log(\frac{eN\epsilon+2eN}{s\epsilon})\\
&\lesssim s\log(\frac{N}{s\epsilon}).
\end{aligned}\tag{4.27}$$

$B_1^s\subset 1B_{\hat{\infty}}^s$ holds because $\|\cdot\|_{\hat{\infty}}\leq\max_{p\in[N]}\|\sqrt{N}F^p\|_\infty\|\cdot\|_1=1\|\cdot\|_1$.

With the two results 4.26 and 4.27 from above, we are ready to bound the Dudley's integration.

First by definition we have

$$
\begin{aligned}
e_0 &= \sup_{x\in D_{s,N}} \|Z_x\|_{F_2} \le \frac{\sqrt{m}}{\ell}\|x\|_{\hat{\infty}} \\
&= \frac{\sqrt{m}}{\ell}\|\sqrt{N}Fx\|_\infty \le \frac{\sqrt{m}}{\ell}\|x\|_1 \le \frac{\sqrt{sm}}{\ell}\|x\|_2 \le \frac{\sqrt{sm}}{\ell}.
\end{aligned}
$$

Then insert the above to the integration

$$
\begin{aligned}
&\int_0^{e_0} \sqrt{\log N(D_{s,N}, d(x,x_0),\epsilon)}d\epsilon \\
&= \int_0^{\frac{\sqrt{sm}}{\ell}} \sqrt{\log N(D_{s,N}, d(x,x_0),\epsilon)}d\epsilon \\
&= \int_0^{\kappa} \sqrt{s\log\frac{N}{s\epsilon}}d\epsilon + \int_\kappa^{\frac{\sqrt{sm}}{\ell}} \frac{\sqrt{sm}}{\ell\epsilon}\log N\, d\epsilon \\
&= \sqrt{s}\int_0^{\kappa} \sqrt{\log\frac{N}{s\epsilon}}d\epsilon + \frac{\sqrt{sm}}{\ell}\log N \int_\kappa^{\frac{\sqrt{sm}}{\ell}} \frac{1}{\epsilon}d\epsilon,
\end{aligned}
$$

now changing variable with setting $\frac{1}{t} = \frac{N}{s\epsilon}$ one reaches

$$
\begin{aligned}
&\int_0^{e_0} \sqrt{\log N(D_{s,N}, d(x,x_0),\epsilon)}d\epsilon \\
&= \sqrt{s}\frac{N}{s}\int_0^{\frac{s}{N}\kappa} \sqrt{\log\frac{1}{t}}dt + \frac{\sqrt{sm}}{\ell}\log N\log(\frac{\sqrt{sm}}{\kappa\ell}) \\
&\le \frac{N}{\sqrt{s}}\int_0^{\frac{s}{N}\kappa} \sqrt{\log\frac{1}{1+t}}dt + \frac{\sqrt{sm}}{\ell}\log N\log(\frac{\sqrt{sm}}{\kappa\ell}) \\
&\le \frac{N}{\sqrt{s}}(\frac{s\kappa}{N})\sqrt{\log e(1+\frac{N}{s\kappa})} + \frac{\sqrt{sm}}{\ell}\log N\log(\frac{\sqrt{sm}}{\kappa\ell}) \\
&\quad let\ \kappa = \frac{\sqrt{m}}{\ell}, \\
&= \sqrt{s}\frac{\sqrt{m}}{\ell}\log N\log\sqrt{s} + \sqrt{s}\frac{\sqrt{m}}{\ell}\sqrt{\log\frac{N\ell}{s\sqrt{m}}} \\
&\lesssim \frac{\sqrt{sm}}{\ell}\log N\log s \\
&= (\frac{s}{m})^{\frac{1}{2}-\alpha}\log N\log s,
\end{aligned}
$$

where the latest equality stands by writing $\ell = m(\frac{s}{\ell})^\alpha$. Note that, to make the inequality reasonable, i.e., $(\frac{s}{m})^{\frac{1}{2}-\alpha} > 1$, α is naturally restricted as $\alpha \in [0, 1/2)$.

Chapter 5

Restricted Isometry Property of discrete Fourier matrix

Due to popularity of the discrete Fourier matrix, it is important to prove the RIP of it. There are already papers on this topic. We here do not aim to improve the bound rather than that, we apply again our method from Chapter 4 with more details to prove the RIP of the discrete Fourier matrix, and then summarize it to be a quick test for proving RIP in Chapter 5.1.

The discrete Fourier matrix used here is nonnormalized, and the normalization after randomly choosing m rows is then by scale $\frac{1}{\sqrt{m}}$. By using McDiarmid's inequality, the restricted isometry property of partial random discrete Fourier matrix can be shown. Again for clarity the restricted isometry property is stated here below.

$$\begin{aligned}\delta_s &= \sup_{x\in D_{s,N}} \{\|\frac{1}{\sqrt{m}} R_\Omega F x\|_2^2 - 1\} \\ &= \sup_{x\in D_{s,N}} \{\|\frac{1}{\sqrt{m}} R_\Omega F x\|_2^2 - \mathbb{E}\|\frac{1}{\sqrt{m}} R_\Omega F x\|_2^2\}.\end{aligned}$$

The main theorem is thus stated below.

Theorem 19. *For $m \geq \delta^{-2} \ln \epsilon^{-2} s^2 \log^2 s \log^2 N$, F has the restricted isometry property of order s with constant δ with probability larger than $1-\epsilon$.*

The proof of Theorem 19 is derived directly from Lemma 11.

Lemma 9. *Let Ω and Ω' differ at one component. Then the function*

$$f(\Omega) := \|\frac{1}{\sqrt{m}} R_\Omega F x\|_2^2 - \|\frac{1}{\sqrt{m}} R_\Omega F y\|_2^2$$

satisfies

$$|f(\Omega) - f(\Omega')| \leq \frac{2\sqrt{2s}}{m} \|x - y\|_{\hat{\infty}}.$$

Proof. Let complex vector $v := (e^{\frac{-2\pi i}{N}} \cdot 1, e^{\frac{-2\pi i}{N}} \cdot 2, \ldots, e^{\frac{-2\pi i}{N}} \cdot N)$ the first row of the discrete Fourier matrix. Denote $\Omega = (\Omega_1, \Omega_2, \ldots, \Omega_m)$ and $\Omega' = (\Omega_1', \Omega_2, \ldots, \Omega_m)$ the m realizations of the random matrix R_Ω, by assumption and without loss of generality they differ in the first component. Then

$$\begin{aligned}
m\Big|f(\Omega) - f(\Omega')\Big| &= \Bigg| \left\| \begin{matrix} \langle x, v^{\Omega_1}\rangle \\ \langle x, v^{\Omega_2}\rangle \\ \vdots \\ \langle x, v^{\Omega_m}\rangle \end{matrix} \right\|_2^2 - \left\| \begin{matrix} \langle y, v^{\Omega_1}\rangle \\ \langle y, v^{\Omega_2}\rangle \\ \vdots \\ \langle y, v^{\Omega_m}\rangle \end{matrix} \right\|_2^2 - \left\| \begin{matrix} \langle x, v^{\Omega_1'}\rangle \\ \langle x, v^{\Omega_2}\rangle \\ \vdots \\ \langle x, v^{\Omega_m}\rangle \end{matrix} \right\|_2^2 + \left\| \begin{matrix} \langle y, v^{\Omega_1'}\rangle \\ \langle y, v^{\Omega_2}\rangle \\ \vdots \\ \langle y, v^{\Omega_m}\rangle \end{matrix} \right\|_2^2 \Bigg| \\
&= \Big|\langle x, v^{\Omega_1}\rangle^2 - \langle y, v^{\Omega_1}\rangle^2 - \langle x, v^{\Omega_1'}\rangle^2 + \langle y, v^{\Omega_1'}\rangle^2\Big| \\
&= \Big|(\langle x, v^{\Omega_1}\rangle + \langle y, v^{\Omega_1}\rangle)(\langle x, v^{\Omega_1}\rangle - \langle y, v^{\Omega_1}\rangle) - \\
&\quad (\langle x, v^{\Omega_1'}\rangle + \langle y, v^{\Omega_1'}\rangle)(\langle x, v^{\Omega_1'}\rangle - \langle y, v^{\Omega_1'}\rangle)\Big| \\
&\leq \Big|2(\langle x, v^{\Omega_1}\rangle + \langle y, v^{\Omega_1}\rangle)(\langle x, v^{\Omega_1}\rangle - \langle y, v^{\Omega_1}\rangle)\Big| \\
&\leq \Big|2\|x - y\|_{\hat{\infty}}(\langle x + y, v^{\Omega_1'}\rangle)\Big| \\
&\leq 2\sqrt{2s}\|x - y\|_{\hat{\infty}}.
\end{aligned}$$

□

Let $Z_x = \frac{1}{\sqrt{m}} R_\Omega F x\|_2^2 - \mathbb{E}\frac{1}{\sqrt{m}} R_\Omega F x\|_2^2$, and $d(x, y) := \|Z_x - Z_y\|_{\psi_2}$. The distance $d(x, y) := \|Z_x - Z_y\|_{\psi_2}$ is thus bounded in the following lemma.

Lemma 10.

$$d(x,y) \leq \frac{2\sqrt{2s}}{\sqrt{m}}\|x-y\|_{\hat{\infty}}. \tag{5.1}$$

Proof. It is suffice to show that for all $t \geq 0$

$$\mathbb{P}(|Z_x - Z_y| > t) \lesssim \exp\bigg(-t^2/(\frac{2\sqrt{2s}}{\sqrt{m}}\|x-y\|_{\hat{\infty}})^2\bigg).$$

By Lemma 9

$$\begin{aligned}
ran_j &= \sup_{\Omega_j}\Big(\mathbb{E}(f(\Omega)|\Omega_j,...,\Omega_1)\Big|\Omega_{j-1},...,\Omega_1\Big) + \sup_{\Omega_j}\Big(\mathbb{E}(-f(\Omega')|\Omega_j,...,\Omega_1)\Big|\Omega_{j-1},...,\Omega_1\Big) \\
&= \sup_{\Omega_j}\Big(\mathbb{E}(f(\Omega)|\Omega_j,\Omega_{j-1}...,\Omega_1) + \mathbb{E}(-f(\Omega')|\Omega'_j,\Omega_{j-1},...,\Omega_1)\Big|\Omega_{j-1},...,\Omega_1\Big) \\
&= \sup_{\Omega_j,\Omega'_j\notin\{\Omega_1,...,\Omega_{j-1}\}}\Big(\mathbb{E}(f(\Omega)|\Omega_j,\Omega_{j-1}...,\Omega_1) - \mathbb{E}(f(\Omega')|\Omega'_j,\Omega_{j-1},...,\Omega_1)\Big|\Omega_{j-1},...,\Omega_1\Big).
\end{aligned}$$

Since

$$\begin{aligned}
&\mathbb{E}[f(\Omega)|\Omega_1,...,\Omega_{k-1},\Omega_k] - \mathbb{E}[f(\Omega')|\Omega_1,...,\Omega_{k-1},\Omega'_k] \\
&= \sum_{i=j+1}^{m}[f(\Omega)|\Omega_1,...,\Omega_{k-1},\Omega_k]\mathbb{P}(\Pi_{i=j+1}^{m}\Omega_i) - \sum_{i=j+1}^{m}[f(\Omega')|\Omega_1,...,\Omega_{k-1},\Omega_k]\mathbb{P}(\Pi_{i=j+1}^{m}\Omega_i) \\
&= \sum_{i=j+1}^{m}([f(\Omega)|\Omega_1,...,\Omega_{k-1},\Omega_k] - [f(\Omega')|\Omega_1,...,\Omega_{k-1},\Omega_k])\mathbb{P}(\Pi_{i=j+1}^{m}\Omega_i) \\
&\leq \frac{2\sqrt{2s}}{m}\|x-y\|_{\hat{\infty}}\sum_{i=j+1}^{m}\mathbb{P}(\Pi_{i=j+1}^{m}\Omega_i) = \frac{2\sqrt{2s}}{m}\|x-y\|_{\hat{\infty}},
\end{aligned}$$

and $R^2 \leq \frac{2\sqrt{2s}}{\sqrt{m}}\|x-y\|_{\hat{\infty}}$, McDiarmid's inequality yields the result. □

Below volumetric argument together with Maurey's method will be applied to bound the covering number $\mathcal{N}(D_{s,N}, \frac{\sqrt{s}}{\sqrt{m}}\|\cdot\|_{\hat{\infty}}, \epsilon)$.

Dudley's inequality then reveals the property of the restricted isometry property as below.

Lemma 11.

$$\mathbb{P}(\sup_{x\in D_{s,N}}|\|\frac{1}{\sqrt{m}}R_\Omega Fx\|_2^2 - 1| > t) \lesssim \exp\bigg(-t^2/(\frac{s\log N\log s}{\sqrt{m}})^2\bigg).$$

Proof. As setting before $Z_x = \|\frac{1}{\sqrt{m}} R_\Omega F x\|_2^2 - \mathbb{E}\|\frac{1}{\sqrt{m}} R_\Omega F x\|_2^2$, this lemma is proved by bound the supremum of Z_x over $x \in D_{s,N}$ by Dudley's inequality Theorem 11

$$\mathbb{P}(\sup_{x \in D_{s,N}} |Z_x - Z_{x_0}| > u) \lesssim \exp(\frac{-u^2}{\big(\int_0^{e_0} \sqrt{\log N(D_{s,N}, d(x,x_0), \epsilon)} d\epsilon\big)^2}, \tag{5.2}$$

where $d(x, x_0) := \|Z_x - Z_{x_0}\|_{\Psi_2}$ is the (up to a absolute constant) smallest value such that

$$\mathbb{P}(|Z_x - Z_y| > u d(x,y)) \lesssim \exp(\frac{-u^2}{d^2(x,y)}). \tag{5.3}$$

This bound is derived in Lemma 10.

Secondly, we need to bound the integral in denominator, i.e. $\int_0^{e_0} \sqrt{\log N(D_{s,N}, d(x,x_0), \epsilon)} d\epsilon$ by Maurey's method and Volumetric argument. First use Lemma 10 set $y = 0$,

$$e_0 = \sup_{x \in D_{s,N}} \|Z_x\|_{\Psi_2} \leq \frac{2\sqrt{2s}}{\sqrt{m}} \|x\|_{\hat{\infty}} = \frac{2\sqrt{2s}}{\sqrt{m}} \|Fx\|_\infty \leq \frac{2\sqrt{2s}}{\sqrt{m}} \|x\|_1 \leq \frac{2\sqrt{2}s}{\sqrt{m}} \|x\|_2 \leq \frac{2\sqrt{2}s}{\sqrt{m}}. \tag{5.4}$$

Secondly, applying Maurey's method Lemma 2, by setting $\mathcal{U} = \{\pm\sqrt{2}e_1, \pm\sqrt{2}e_2, \cdots, \pm\sqrt{2}e_N\}$, and $\|\cdot\|_X = \|\cdot\|_\infty = \|Fx\|_\infty = \max_{p \in [N]} \langle F^p, x\rangle$. Then $B^N(0,1) \subset conv(\mathcal{U})$, and

$$\mathbb{E}\|\sum_{k=1}^M \epsilon_k Z_k\|_X \tag{5.5}$$

$$= \mathbb{E} \max_{p=1\ldots N} |\sum_{k=1}^M \epsilon_k \langle F^p, Z_k\rangle|. \tag{5.6}$$

Note that the norm $\|\cdot\|_{\hat{\infty}}$ is defined by nonnormalized discrete Fourier matrix because $\|F^p\|_\infty$ should be a constant independent of N. By Hölder's inequality $\langle F^p, Z_k\rangle \leq \|F^p\|_\infty \|Z_k\|_1 \leq 1\sqrt{2}$, we have

$$\|(\langle F^p, Z_k\rangle)_{k=1}^M\|_2 \leq \sqrt{2M}, \tag{5.7}$$

for $k \in [N]$ By Höffding's inequality, conditional on Z_k,

$$\mathbb{P}_\epsilon(|\sum_{k=1}^M \epsilon_k \langle F^p, Z_k\rangle| \geq \sqrt{2M}t) \leq 2e^{t^2/2}, \ t > 0. \tag{5.8}$$

By union bound

$$\mathbb{P}_\epsilon(\max_{p\in[N]}|\sum_{k=1}^{M}\epsilon_k\langle F^p,Z_k\rangle|\geq\sqrt{2M}t)\leq 2Ne^{t^2/2},\ t>0. \tag{5.9}$$

Applying Proposition 3 yields

$$\mathbb{E}_\epsilon\max_{p\in[N]}|\sum_{k=1}^{M}\epsilon_k\langle F^p,Z_k\rangle|\leq 3/2\sqrt{2}\sqrt{\ln 8N}\sqrt{M}, \tag{5.10}$$

and by Fubini's theorem,

$$\mathbb{E}\max_{p\in[N]}|\sum_{k=1}^{M}\epsilon_k\langle F^p,Z_k\rangle|\leq 3/2\sqrt{2}\sqrt{\ln 8N}\sqrt{M}. \tag{5.11}$$

Hence by letting $A=3/2\sqrt{2}\sqrt{\ln 8N}\lesssim\sqrt{\ln N}$, Maurey's method yields

$$\log\mathcal{N}(conv(\mathcal{U}),\|\cdot\|_X,\epsilon)\lesssim(\frac{1}{\epsilon})^2\ln^2 N.$$

Therefore

$$\begin{aligned}
&\log\mathcal{N}(D_{s,N},\|\cdot\|_{\Psi_2},\epsilon)\lesssim\log\mathcal{N}(D_{s,N},\frac{\sqrt{s}}{\sqrt{m}}\|\cdot\|_{\hat{\infty}},\epsilon)\\
&\leq\log\mathcal{N}(\sqrt{s}B_1^N(0,1),\frac{\sqrt{s}}{\sqrt{m}}\|\cdot\|_{\hat{\infty}},\epsilon)=\log\mathcal{N}(B_1^N(0,1),\|\cdot\|_{\hat{\infty}},\frac{\sqrt{m}}{\sqrt{s}\sqrt{s}}\epsilon)\\
&\leq\log\mathcal{N}(conv(\mathcal{U}),\|\cdot\|_{\hat{\infty}},\frac{\sqrt{m}}{\sqrt{s}\sqrt{s}}\epsilon)\\
&\lesssim(\frac{s}{\sqrt{m}\epsilon})^2\ln^2 N.
\end{aligned} \tag{5.12}$$

By a volumetric argument, we have

$$\begin{aligned}
&\log\mathcal{N}(D_{s,N},d(x,x_0),\epsilon)\leq\log\mathcal{N}(B_2^s,\frac{\sqrt{s}}{\sqrt{m}}\|\cdot\|_{\hat{\infty}},\epsilon)\leq\log\binom{N}{s}\mathcal{N}(\sqrt{s}B_1^s,\frac{\sqrt{s}}{\sqrt{m}}\|\cdot\|_{\hat{\infty}},\epsilon)\\
&\leq\log(\frac{eN}{s})^s\mathcal{N}(B_{\hat{\infty}}^s,\|\cdot\|_{\hat{\infty}},\frac{\sqrt{m}}{s}\epsilon)\leq\log(\frac{eN}{s})^s(1+\frac{2s}{\sqrt{m\epsilon}})^s\leq s\log(\frac{eN}{s})(1+\frac{2s}{\sqrt{m\epsilon}})\\
&=s\log(\frac{eN}{s}+\frac{2eN}{\sqrt{m\epsilon}})\lesssim s\log(\frac{N}{s\epsilon}).
\end{aligned} \tag{5.13}$$

$B_1^s \subset 1B_\infty^s$ holds because $\|\cdot\|_\infty \leq \max_{p\in[N]} \|F^p\|_\infty \|\cdot\|_1 = 1\|\cdot\|_1$. Then combining (5.12)(5.13),

$$\int_0^{e_0} \sqrt{\log N(D_{s,N}, d(x,x_0),\epsilon)}d\epsilon = \int_0^{\frac{s}{\sqrt{m}}} \sqrt{\log N(D_{s,N}, d(x,x_0),\epsilon)}d\epsilon$$
$$= \int_0^{\kappa} \sqrt{s\log\frac{N}{s\epsilon}}d\epsilon + \int_\kappa^{\frac{s}{\sqrt{m}}} \frac{s}{\sqrt{m}\epsilon}\log N\, d\epsilon = \sqrt{s}\int_0^{\kappa}\sqrt{\log\frac{N}{s\epsilon}}d\epsilon + \frac{s\log N}{\sqrt{m}}\int_\kappa^{\frac{s}{\sqrt{m}}}\frac{1}{\epsilon}d\epsilon,$$

by changing variable $\frac{1}{t} = \frac{N}{s\epsilon}$,

$$\int_0^{e_0} \sqrt{\log N(D_{s,N}, d(x,x_0),\epsilon)}d\epsilon = \sqrt{s}\frac{N}{s}\int_0^{\frac{s}{N}\kappa}\sqrt{\log\frac{1}{t}}dt + \frac{s}{\sqrt{m}}\log N\log(\frac{s}{\kappa\sqrt{m}})$$
$$\leq \frac{N}{\sqrt{s}}\int_0^{\frac{s}{N}\kappa}\sqrt{\log\frac{1}{1+t}}dt + \frac{s}{\sqrt{m}}\log N\log(\frac{s}{\kappa\sqrt{m}})$$
$$\leq \sqrt{s}\kappa\sqrt{\log e(1+\frac{N}{s\kappa})} + \frac{s}{\sqrt{m}}\log N\log(\frac{s}{\kappa\sqrt{m}}),$$

now let $\kappa = \frac{\sqrt{s}\log N}{\sqrt{m}}$

$$\int_0^{e_0} \sqrt{\log N(D_{s,N}, d(x,x_0),\epsilon)}d\epsilon = \frac{s\log N}{\sqrt{m}}\sqrt{\ln e(1+\frac{N\sqrt{m}}{s\log N})} + \frac{s\log N}{\sqrt{m}}\log\frac{\sqrt{s}}{\log N} \lesssim \frac{s\log N\log s}{\sqrt{m}}.$$

This ends the proof. □

5.1 Quick test of RIP

In this section I summarise the above method of the combination of Dudley's inequality with McDiarmid's inequality to a quick test for proving the RIP of partial random matrices $\frac{1}{\sqrt{m}}R_\Omega A$ (the randomness occurs at drawing the m rows out of N rows), for A any arbitrary matrix $\mathbb{E}\|\frac{1}{\sqrt{m}}R_\Omega Ax\|_2^2 = 1$.
(In case with $\Sigma\Delta$-quantization, multiply a term $(\frac{\sqrt{m}}{\sqrt{\ell}}P_\ell V^*)$ in front of $\frac{1}{\sqrt{m}}R_\Omega A$.)

$$\delta_s = \sup_{x\in D_{s,N}}\{\|\frac{1}{\sqrt{m}}R_\Omega Ax\|_2^2 - 1\}$$
$$= \sup_{x\in D_{s,N}}\{\|\frac{1}{\sqrt{m}}R_\Omega Ax\|_2^2 - \mathbb{E}\|\frac{1}{\sqrt{m}}R_\Omega Ax\|_2^2\}.$$

Lemma 12. *Let Ω and Ω' differ at one component. If the function*

$$f(\Omega) := \|\frac{1}{\sqrt{m}} R_\Omega A x\|_2^2 - \|\frac{1}{\sqrt{m}} R_\Omega A y\|_2^2$$

satisfies

$$|f(\Omega) - f(\Omega')| \lesssim \frac{1}{K(s,\ell,m)} \|x-y\|_\infty,$$

for some function K of variables $s,\ \ell,\ m$, then

$$\mathbb{P}(\delta_s > t) \lesssim \exp(-t^2/(\frac{\sqrt{sm}}{K(s,\ell,m)} \log N \log m)^2).$$

5.2 RIP of random matrices $\frac{1}{\sqrt{\ell}} P_\ell V^* R_\Omega F$

If we can prove the RIP of $\frac{1}{\sqrt{\ell}} P_\ell V^* R_\Omega F$ then we can apply Theorem 7 again as in Chapter 4 to get a reconstruction error bound for the partial random discrete Fourier matrices. The essential step while applying Dudley's inequality together with McDiarmid's inequality to found the restricted isometry property of $\frac{1}{\sqrt{\ell}} P_\ell V^* R_\Omega F$ is to bound the difference similarly in Lemma 12 i.e. $f(\Omega) - f(\Omega')$.

$$\begin{aligned}
&|f(\Omega) - f(\Omega')| := \\
&\left| (\|\frac{1}{\sqrt{\ell}} P_\ell V^* R_\Omega F x\|_2^2 - \|\frac{1}{\sqrt{\ell}} P_\ell V^* R_\Omega F y\|_2^2) - (\|\frac{1}{\sqrt{\ell}} P_\ell V^* R'_\Omega F x\|_2^2 - \|\frac{1}{\sqrt{\ell}} P_\ell V^* R'_\Omega F y\|_2^2) \right| \\
&= \frac{1}{\ell} \Big| \Big\{ \big(< P_\ell V^* R_\Omega F_{1\cdot}, x\rangle^2 + \cdots + < P_\ell V^* R_\Omega F_{\ell\cdot}, x\rangle^2 \big) \\
&- \big(< P_\ell V^* R_\Omega F_{1\cdot}, y\rangle^2 + \cdots + < P_\ell V^* R_\Omega F_{\ell\cdot}, y\rangle^2 \big) \\
&- \big(< P_\ell V^* R_{\Omega'} F_{1\cdot}, x\rangle^2 + \cdots + < P_\ell V^* R_{\Omega'} F_{\ell\cdot}, x\rangle^2 \big) \\
&+ \big(< P_\ell V^* R_{\Omega'} F_{1\cdot}, y\rangle^2 + \cdots + < P_\ell V^* R_{\Omega'} F_{\ell\cdot}, y\rangle^2 \big) \Big\} \Big| \\
&= \frac{1}{\ell} \Big| \Big(P_\ell V^* R_\Omega F_{1\cdot}[x+y] P_\ell V^* R_\Omega F_{\ell\cdot} + \cdots + P_\ell V^* R_\Omega F_{1\cdot}[x+y] P_\ell V^* R_\Omega F_{\ell\cdot} \Big)[x-y]
\end{aligned}$$

$$- \left(P_\ell V^* R_{\Omega'} F_{1\cdot}[x+y] P_\ell V^* R_{\Omega'} F_{\ell\cdot} + \cdots + P_\ell V^* R_{\Omega'} F_{1\cdot}[x+y] P_\ell V^* R_{\Omega'} F_{\ell\cdot} \right) [x-y] \Bigg|.$$

However there is so far no proper upper bound for this difference, and comparing to the result from Theorem 9, I might refer an implication that if the difference is larger than a certain amount there is no RIP. □

Chapter 6

Appendix A

Lemma 13. *[26] For $u > 0$,*

$$\int_u^\infty e^{-t^2/2}dt \leq e^{-u^2/2} \min\left\{\sqrt{\frac{\pi}{2}}, \frac{1}{u}\right\}. \tag{6.1}$$

Proof.

$$\int_u^\infty e^{-t^2/2}dt = \int_0^\infty e^{-(t+u)^2/2}dt = e^{-u^2/2}\int_0^\infty e^{-(tu)}e^{-t^2/2}dt.$$

On one hand: for $t, u > 0$, $e^{-tu} \leq 1$, and then

$$e^{-u^2/2}\int_0^\infty e^{-(tu)}e^{-t^2/2}dt \leq e^{-u^2/2}\int_0^\infty e^{-t^2/2}dt = \sqrt{\frac{\pi}{2}}e^{-u^2/2}. \tag{6.2}$$

On the other hand: for $u, t > 0$, $e^{-t^2/2} \leq 1$, and then

$$e^{-u^2/2}\int_0^\infty e^{-(tu)}e^{-t^2/2}dt \leq e^{-u^2/2}\int_0^\infty e^{-(tu)}dt = \frac{1}{u}e^{-u^2/2}. \tag{6.3}$$

□

Lemma 14. *[26] In $\mathbb{C}^N$, for $0 < p < q$, then*

$$\|x\|_q \leq \|x\|_p \leq N^{(1/p-1/q)}\|x\|_q. \tag{6.4}$$

Proof. See reference for more details. □

Lemma 15. *[26] For $\alpha > 0$ it holds*

$$\int_0^\alpha \sqrt{\ln(1+\frac{1}{t})}dt \leq \alpha\sqrt{\ln(e(1+\frac{1}{\alpha}))}$$

Proof. [26] Apply Cauchy-Schwarz' inequality to obtain

$$\int_0^\alpha \sqrt{\ln(1+t^{-1})} \leq \sqrt{\int_0^\alpha 1dt \int_0^\alpha \ln(1+t^{-1})}. \tag{6.5}$$

Let $u = t^{-1}$, integration by parts yields

$$\begin{aligned}
&\int_0^\alpha \ln(1+t^{-1})dt = \int_{\alpha^{-1}}^\infty u^{-2}\ln(1+u)du \\
&= -u^{-1}\ln(1+u)|_{\alpha^{-1}}^\infty + \int_{\alpha^{-1}}^\infty u^{-1}\frac{1}{1+u}du \leq \alpha\ln(1+\alpha^{-1}) + \int_{\alpha^{-1}}^\infty \frac{1}{u^2}du \\
&= \alpha\ln(1+\alpha^{-1}) + \alpha.
\end{aligned}$$

Substituting this into (6.5) ends the proof.

□

Proposition 5. *[28] Let r be any positive integer and D be as in (1.24). There are positive constants $c_{s_1}(r)$ and $c_{s_2}(r)$, independent of m, such that*

$$c_{s_1}(r)(\frac{m}{j})^r \leq \sigma_j(D^{-r}) \leq c_{s_2}(r)(\frac{m}{j})^r,\ j = 1, \ldots, m. \tag{6.6}$$

Proof. See reference for more details. □

Lemma 16. *[44] Assume that $\xi = (\xi_j)_{j=1}^M$ is a sequence of independent random vector in $\mathbb{C}^n$ equipped*

with a (semi-)norm $\|\cdot\|$*, having expectations* $\mathbf{x_j} = \mathbb{E}\xi_j$*. Then for* $1 \le p \le \infty$

$$\left(\mathbb{E}\|\sum_{j=1}^{M}(\xi_j - \mathbf{x}_j)\|^p\right)^{1/p} \le 2\left(\mathbb{E}\|\sum_{j=1}^{M}\epsilon_j\xi_j\|^p\right)^{1/p},$$

where $\epsilon = (\epsilon_j)_{j=1}^N$ *is a Rademacher sequence independent of* ξ*.*

Proof. Let $\xi' = (\xi'_1, \xi'_2, \ldots, \xi'_M)$ denote an independent copy of the sequence of random vectors $(\xi_1, \xi_2, \ldots, \xi_M)$. Since $\mathbb{E}\xi'_j = \mathbf{x}_j$ an application of Jensen's inequality yields

$$E := \mathbb{E}\|\sum_{j=1}^{M}(\xi_j - \mathbf{x}_j)\|^p = \mathbb{E}\|\sum_{j=1}^{M}(\xi_j - \mathbb{E}\xi'_j)^p\|^p \le \mathbb{E}\|\sum_{j=1}^{M}(\xi_j - \xi'_j)\|^p.$$

Since $(\xi_j - \xi'_j)_{j=1}^M$ is a vector of independent symmetric random variables; thus it has the same distribution as $(\epsilon_j(\xi_j - \xi'_j))_{j=1}^M$. The triangle inequality gives

$$E^{1/p} \le (\mathbb{E}\|\sum_{j=1}^{M}\epsilon_j(\xi_j - \xi'_j)\|^p)^{1/p} \le (\mathbb{E}\|\sum_{j=1}^{M}\epsilon_j\xi_j\|^p)^{1/p} + (\mathbb{E}\|\sum_{j=1}^{M}\epsilon_j\xi'_j\|^p)^{1/p} = 2(\mathbb{E}\|\sum_{j=1}^{M}\epsilon_j\xi_j\|^p)^{1/p}.$$

The last equality is due to the fact that ξ' is an independent copy of ξ. □

Lemma 17. *Let* X *be any real-values random variable with expected value* $\mathbb{E} = 0$ *and such that* $a \le X \le b$ *almost surely. Then for all lambda* $\in \mathbb{R}$*,*

$$\mathbb{E}[e^{\lambda X}] \le \exp\left(\frac{\lambda^2(b-a)^2}{8}\right).$$

Proof. Since $e^{\lambda x}$ is convex

$$e^{\lambda x} \le \frac{b-x}{b-a}e^{\lambda a} + \frac{x-a}{b-a}e^{\lambda b} \quad \forall a \le x.$$

So

$$\mathbb{E}[e^{\lambda x}] \le \frac{b-\mathbb{E}(X)}{b-a}e^{\lambda a} + \frac{\mathbb{E}(X)-a}{b-a}e^{\lambda b} \quad \forall a \le x.$$

Let $h = \lambda(b-a)$, $p = \frac{-a}{b-a}$ and $L(h) = -hp + \ln(1-p+pe^h)$. Then

$$\frac{b-\mathbb{E}(X)}{b-a}e^{\lambda a} + \frac{\mathbb{E}(X)-a}{b-a}e^{\lambda b} = e^{L(h)}$$

since $\mathbb{E}(X) = 0$. Taking derivative of $L(h)$,

$$L(0) = L'(0) = 0 \text{ and } L''(h) \leq \frac{1}{4} \text{ for all } h.$$

By Taylor's expansion,

$$L(h) \leq \frac{1}{8}h^2 = \frac{1}{8}\lambda^2(b-a)^2.$$

Hence

$$\mathbb{E}[e^{\lambda X}] \leq e^{\frac{1}{8}\lambda^2(b-a)^2}.$$

□

Chapter 7

Appendix B

This chapter shows the proof of Theorem 7 [48].

Theorem 20 ([25]). *Given $q \leq 1$, suppose that the matrices $\Phi \in \mathbb{R}^{m \times N}$ satisfy the ℓ_q robust null space property of order s with constants $0 < \rho < 1$ and $\tau > 0$ relative to a norm $\|\cdot\|$ on $\mathbb{R}^m$. Then, for any $1 \leq p \leq q$, the bounds*

$$\|f - g\|_p \leq \frac{C}{s^{1-1/p}}(\|g\|_1 - \|f\|_1 + 2\sigma_s(f)_1) + Ds^{1/p-1/q}\|\Phi(f-g)\| \tag{7.1}$$

hold for all $f \in \mathbb{R}^N$ and $e \in \mathbb{R}^m$ with $\|e\| \leq \eta$. The constants $C, D > 0$ depend only on ρ and τ.

Proof. First note that since for $q \geq 1$, by Young's inequality $\|x_S\|_1 \leq \frac{s^{1/1}}{s^{1/q}}\|x_S\|_q$, the ℓ_q-robust null space property implies ℓ_1-robust null space property for all index set S, with cardinality $|S| \leq s$,

$$\|v_S\|_1 \leq \rho\|v_{\overline{S}}\|_1 + \tau s^{1-1/q}\|\Phi v\|. \tag{7.2}$$

Now write

$$\|f\|_1 = \|f_S\|_1 + \|f_{\overline{S}}\|_1 \leq \|(f-g)_S\|_1 + \|g_S\|_1 + \|f_{\overline{S}}\|_1,$$

$$0 = \|f_S\|_1 + \|f_{\overline{S}}\|_1 \leq \|(f-g)_S\|_1 + \|g_S\|_1 + \|f_{\overline{S}}\|_1 - \|f\|_1, \tag{7.3}$$

$$\tag{7.4}$$

$$\|(f-g)_{\overline{S}}\|_1 \leq \|g_{\overline{S}}\|_1 + \|f_{\overline{S}}\|_1. \tag{7.5}$$

Summing these two up,

$$\|(f-g)_{\overline{S}}\|_1 \leq \|(f-g)_S\|_1 + \|g\|_1 - \|f\|_1 + 2\|f_{\overline{S}}\|_1. \tag{7.6}$$

Inserting (7.2) to substitute $\|(f-g)_S\|_1$,

$$\|(f-g)_{\overline{S}}\|_1 \leq (\rho\|(f-g)_{\overline{S}}\|_1 + \tau s^{1-1/q}\|\Phi(f-g)\|) + 2\|f_{\overline{S}}\|_1 + \|g\|_1 - \|f\|_1 \tag{7.7}$$

$$\Rightarrow \|(f-g)_{\overline{S}}\|_1 \leq \frac{1}{1-\rho}(\|g\|_1 - \|f\|_1 + 2\|f_{\overline{S}}\|_1 + \tau s^{1-1/q}\|\Phi(f-g)\|). \tag{7.8}$$

□

Proof of Theorem 7. Let (z,ν) be a feasible pair to (1.27), and let $\tilde{\gamma} := \gamma(r)/\Delta$. Define $u := D^{-r}(\Phi z + \nu - q)$, and $p := (\frac{1}{\tilde{\gamma}}u, \frac{-\Delta}{\epsilon}\nu)$ and then

$$\|u\|_2 \leq \tilde{\gamma}\Delta\sqrt{m},\ \textit{and}\ \|p\|_2 \leq \Delta\sqrt{2m}. \tag{7.9}$$

By definition, $u,\ p,\ q,\ z$ and ν have the relation

$$\Phi z - q = D^r u - \nu = [\tilde{\gamma}D^r, \frac{\epsilon}{\Delta}I]p. \tag{7.10}$$

Denote $[\tilde{\gamma}D^r, \frac{\epsilon}{\Delta}I]$ by H, and let the singular value decomposition of $H = V\Sigma U^*$, and then the pseudo-inverse of H, denoted by $H^\dagger$.

$$H^\dagger = (HH^*)^{-1}H = (V\Sigma U^*)^*(V\Sigma U^*(V\Sigma U^*)^*)^{-1} = U\Sigma^{-1}V^*. \tag{7.11}$$

Multiplying both side of (7.10),

$$H^\dagger(\Phi z - q) = H^\dagger Hp = UU^*p. \tag{7.12}$$

Since $\|p\|_2 \leq \Delta\sqrt{2m}$, and U is a unitary matrix,

$$\|H^\dagger(\Phi z - q)\|_2 \leq \Delta\sqrt{2m}. \tag{7.13}$$

By triangle inequality,

$$\|H^\dagger\Phi(x - \hat{x})\|_2 \leq \|H^\dagger(\Phi x - q)\|_2 + \|H^\dagger(\Phi\hat{x} - q)\|_2 \leq 2\Delta\sqrt{2m}. \tag{7.14}$$

Turn to another side and see the singular value decomposition of H in terms of singular value decomposition of $D^r = V_D S_D U_D^*$:

$$\begin{aligned}
HH^* &= V\Sigma U^*(V\Sigma U^*)^* = V\Sigma^2 V^* \\
&= [\tilde{\gamma}D^r, \frac{\epsilon}{\Delta}I][\tilde{\gamma}D^r, \frac{\epsilon}{\Delta}I]^* \\
&= [\tilde{\gamma}^2 D^r(D^r)^*) + (\frac{\epsilon}{\Delta})^2 II^*] \\
&= [\tilde{\gamma}^2 V_D S_D U_D^*(V_D S_D U_D^*)^*) + (\frac{\epsilon}{\Delta})^2 I] \\
&= [\tilde{\gamma}^2(V_D S_D^2 V_D^*) + (\frac{\epsilon}{\Delta})^2 I] \\
&= [(V_D(\tilde{\gamma}S_D)^2 V_D^*) + V_D((\frac{\epsilon}{\Delta})^2 I)V_D^*] \\
&= V_D((\tilde{\gamma}S_D)^2 + (\frac{\epsilon}{\Delta})^2 I)V_D^*,
\end{aligned} \tag{7.15}$$

and since

$$H^\dagger = H^*(HH^*)^{-1} = (V\Sigma U^*)^*(V\Sigma U^*(V\Sigma U^*)^*)^{-1} = U\Sigma^{-1}V^*, \tag{7.16}$$

applying Weyl's inequality, the ℓth singular value of $H^\dagger$ is bounded as

$$\sigma_\ell(H^\dagger) = (\tilde{\gamma}^2\sigma_{m-\ell}^2(D^r) + (\frac{\epsilon}{\Delta})^2)^{-1/2} \geq (\tilde{\gamma}^2(\frac{3\pi r\ell}{m})^{2r} + (\frac{\epsilon}{\Delta})^2)^{-1/2}. \tag{7.17}$$

Therefore by denoting $P_\ell \in \mathbb{R}^{\ell \times m}$ a projection to the first ℓ dimension.

$$\|H^\dagger \Phi(x - \hat{x})\|_2 = \|U\Sigma^{-1}V^*\Phi(x - \hat{x})\|_2 \tag{7.18}$$

$$= \|\Sigma^{-1}V^*\Phi(x - \hat{x})\|_2 \geq \|P_\ell \Sigma^{-1}V^*\Phi(x - \hat{x})\|_2 \tag{7.19}$$

$$= \|P_\ell \Sigma^{-1}P_\ell^* P_\ell V^*\Phi(x - \hat{x})\|_2 = \sigma_\ell(H^\dagger)\|P_\ell V^*\Phi(x - \hat{x})\|_2, \tag{7.20}$$

together with (7.14) then

$$2\Delta\sqrt{2m} \geq \sigma_\ell(H^\dagger)\|P_\ell V^*\Phi(x - \hat{x})\|_2 = \sigma_\ell(H^\dagger)\sqrt{\ell}\|\frac{1}{\sqrt{\ell}}P_\ell V^*\Phi(x - \hat{x})\|_2 = \sigma_\ell(H^\dagger)\sqrt{\ell}\|\tilde{\Phi}\Phi(x - \hat{x})\|_2 \tag{7.21}$$

by setting $\tilde{\Phi} := \frac{1}{\sqrt{\ell}}P_\ell V^*\Phi$.

Now by assumption $\tilde{\Phi}$ has restricted isometry property of order $2k$ and constant $\delta < 1/9$, Theorem 6 shows that the ℓ_q-robust null space property is also satisfied, hence Theorem 20 holds here by setting $f = \hat{x}$ and $g = x$.

$$\begin{aligned}
\|x - \hat{x}\|_2 &\leq C_4\|\tilde{\Phi}(x - \tilde{x})\|_2 + C_5\frac{\sigma_k(x)}{\sqrt{s}} \\
&\leq 2\sqrt{2}C_4\Delta\sqrt{\frac{m}{\ell}}\frac{1}{\sigma_\ell(H^d ag)} + C_5\frac{\sigma_k(x)_1}{\sqrt{s}} \\
&\leq 2\sqrt{2}C_4\Delta\sqrt{\frac{m}{\ell}}(\tilde{\gamma}^2(\frac{3\pi r\ell}{m})^{2r} + (\frac{\epsilon}{\Delta})^2)^{1/2} + C_5\frac{\sigma_k(x)_1}{\sqrt{s}} \\
&\leq 2\sqrt{2}C_4\Delta\sqrt{\frac{m}{\ell}}(\tilde{\gamma}(\frac{3\pi r\ell}{m})^{r} + (\frac{\epsilon}{\Delta})) + C_5\frac{\sigma_k(x)_1}{\sqrt{s}} \\
&\leq 2\sqrt{2}C_4\tilde{\gamma}3^r\pi^r r^r(\frac{\ell}{m})^{r-1/2}\Delta + 2\sqrt{2}C_4\sqrt{\frac{m}{\ell}}\epsilon + C_5\frac{\sigma_k(x)_1}{\sqrt{s}}.
\end{aligned}$$

Setting $C_6 = 2\sqrt{2}C_4\tilde{\gamma}3^r\pi^r r^r$, $C_7 = 2\sqrt{2}C_4$, $C_8 = C_5$ finishes the proof. □

Bibliography

[1] A. Ai, A. Lapanowski, Y. Plan, and R. Vershynin. One-bit compressed sensing with non-gaussian measurements. *Linear Algebra and its Applications*, 441(1):222–239, 2014.

[2] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28(3):253–263, 2008.

[3] J.J. Benedetto, A. M. Powell, and Ö. Yımaz. Second-order sigma-delta ($\Sigma\Delta$) quantization of finite frame expansions. *Applied and Computational Harmonic Analysis*, 20(1):126 – 148, 2006.

[4] J.J. Benedetto, A.M. Powell, and Ö. Yımaz. Sigma-delta ($\Sigma\Delta$) quantization and finite frames. *IEEE Transactions on Information Theory*, 52(5):1990–2005, 2006.

[5] J. Blum, M. Lammers, A. M. Powell, and Ö. Yımaz. Sobolev duals in frame theory and sigma-delta quantization. *Journal of Fourier Analysis and Applications*, 16(3):365–381, 2010.

[6] T. Blumensath and M.E. Davies. Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27(3):265–274, 2009.

[7] B. G. Bodmann and V. I. Paulsen. Frame paths and error bounds for sigma–delta quantization. *Appl. Comp. Harmon. Anal.*, 22(2):176–197, 2007.

[8] B.G. Bodmann, V.I. Paulsen, and S.A. Abdulbaki. Smooth frame-path termination for higher order sigma-delta quantization. *J. Fourier Anal. and Appl.*, 13(3):285–307, 2007.

[9] P. T. Boufounos and R. G. Baraniuk. Quantization of sparse representations. In *Rice University ECE Department Technical Report 0701. Summary appears in Proc. Data Compression Conference (DCC)*, Snowbird, UT, March 27-29 2007.

[10] P. T. Boufounos and R. G. Baraniuk. 1-bit compressive sensing. In *Proc. Conf. Inform. Science and Systems (CISS)*, Princeton, NJ, March 19-21 2008.

[11] T. Cai and A. Zhang. Sparse representation of a polytope and recovery of sparse signals and low-rank matrices. *IIEEE Transactions on Information Theory*, 60(1):122–132, 2013.

[12] E. J. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59:1207–1223, 2006.

[13] E. J. Candès and T. Tao. Decoding by linear programming. *Information Theory, IEEE Transactions on*, 51(12):4203–4215, 2005.

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

[15] E. Chou and Güntürk. Distributed noise-shaping quantization: I. beta duals of finite frames and near-optimal quantization of random measurements. *Constructive Approximation*, 44(1):1–22, 2016.

[16] W. Dai and O. Milenkovic. Subspace pursuit for compressive sensing signal reconstruction. *Information Theory, IEEE Transactions on*, 55(5):2230–2249, 2009.

[17] I. Daubechies and R. DeVore. Approximating a bandlimited function using very coarsely quantized data: a family of stable sigma-delta modulators of arbitrary order. *Ann. Math.*, 158(2):679–710, 2003.

[18] I. Daubechies and R. Saab. A deterministic analysis of decimation for sigma-delta quantization of bandlimited functions. *IEEE Signal Proc. Lett.*, 22(11):2093–2096, 2015.

[19] P. Deift, C. S. Güntürk, and F. Krahmer. An optimal family of exponentially accurate one-bit sigma-delta quantization schemes. *Comm. Pure Appl. Math.*, 64(7):883–919, 2011.

[20] R. DeVore. Deterministic constructions of compressed sensing matrices. *J. Complexity*, 23:918–925, 2007.

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

[22] R. M. Dudley. The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *J. Functional Analysis*, 1:290–330, 1967.

[23] Joe-Mei Feng and Chia han Lee. Simultaneous subspace pursuit for signal recovery from multiple measurement vectors. In *IEEE Wireless Commun. and Networking Conf.*, pages 2874–2878, 2013.

[24] Joe-Mei Feng and Felix Krahmer. An RIP-based approach to $\Sigma\Delta$ quantization for compressed sensing. *IEEE Signal Process. Lett.*, 21(11):1351–1355, 2014.

[25] S. Foucart. Stability and robustness of ℓ_1-minimizations with weibull matrices and redundant dictionaries. *Linear Algebra and its Applications, 441:4-21*, 2014.

[26] S. Foucart and H. Rauhut. A mathematical introduction to compressive sensing. *Basel: Birkhäuser, Boston*, 1(3), 2013.

[27] V.K. Goyal, M. Vetterli, and N.T. Thao. Quantized overcomplete expansions in $\mathbb{R}^N$: analysis, synthesis, and algorithms. *IEEE Transactions on Information Theory*, 44(1):16–31, Jan 1998.

[28] C Sinan Güntürk, Mark Lammers, Alexander M Powell, R. Saab, and Ö Yılmaz. Sobolev duals for random frames and $\sigma\delta$ quantization of compressed sensing measurements. *Foundations of Computational mathematics*, 13(1):1–36, 2013.

[29] C.S. Güntürk. One-bit sigma-delta quantization with exponential accuracy. *Comm. Pure Appl. Math.*, 56(11):1608–1630, 2003.

[30] Jarvis Haupt, Waheed U Bajwa, Gil Raz, and Robert Nowak. Toeplitz compressed sensing matrices with applications to sparse channel estimation. *IEEE Transactions on Information Theory*, 56(11):5862–5875, 2010.

[31] R. A. Horn and C. R. Johnson. *Matrix analysis*. Cambridge university press, 2012.

[32] F. Krahmer J. Feng and R. Saab. Quantized compressed sensing for partial random matrices. *arXiv 1702.04711*, 2017.

[33] L. Jacques, D. K. Hammond, and M. J. Fadili. Dequantizing compressed sensing: When oversampling and non-gaussian constraints combine. *IEEE Transactions on Information Theory*, 57(1):559–571, January 2011.

[34] L. Jacques, D.K. Hammond, and J.M. Fadili. Dequantizing compressed sensing: When oversampling and non-gaussian constraints combine. *Information Theory, IEEE Transactions on*, 57(1):559–571, 2011.

[35] L. Jacques, J. N. Laska, P. T. Boufounos, and R. G. Baraniuk. Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors. 2011.

[36] F. Krahmer, R. Saab, and R. Ward. Root-exponential accuracy for coarse quantization of finite frame expansions. *Information Theory, IEEE Transactions on*, 58(2):1069 –1079, February 2012.

[37] F. Krahmer, R. Saab, and Ö. Yılmaz. Sigma-delta quantization of sub-Gaussian frame expansions and its application to compressed sensing. *Information and Inference*, 3(1):40–58, 2014.

[38] F. Krahmer and R. Ward. Lower bounds for the error decay incurred by coarse quantization schemes. *Appl. Comput. Harmonic Anal.*, 32(1):131–138, 2012.

[39] Felix Krahmer, Shahar Mendelson, and Holger Rauhut. Suprema of chaos processes and the restricted isometry property. *Communications on Pure and Applied Mathematics*, 67(11):1877–1904, 2014.

[40] Colin McDiarmid. Concentration. In *Probabilistic methods for algorithmic discrete mathematics*, pages 195–248. Springer, 1998.

[41] D. Needell and J.A. Tropp. Cosamp: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, 26(3):301–321, 2009.

[42] Y. Plan and R. Vershynin. One-bit compressed sensing by linear programming. *Communications on Pure and Applied Mathematics*, 2013.

[43] A. Powell, R. Saab, and Ö Yılmaz. Quantization and finite frames. In P. Casazza and G. Kutinyok, editors, *Finite Frames: Theory and Aplications*, pages 305–328. Springer, 2013.

[44] Holger Rauhut. *Compressive sensing and structured random matrices*, volume 9. 2010.

[45] Justin Romberg. Compressive sensing by random convolution. *SIAM Journal on Imaging Sciences*, 2(4):1098–1128, 2009.

[46] M. Rudelson and R. Vershynin. On sparse reconstruction from fourier and gaussian measurements. *Comm. Pure Appl. Math.*, 61(8):1025–1045, 2008.

[47] H. Rauhut S. Foucart, A. Pajor and T. Ullrich. The gelfand widths of ℓ_p-balls for $0 < p \geq 1$. *J. Complexity*, 2010.

[48] R. Saab, Rongrong Wang, and Ö Yılmaz. Quantization of compressive samples with stable and robust recovery. *Applied and Computational Harmonic Analysis*, 2016.

[49] R. Saab, Rongrong Wang, and Ö. Yılmaz. Quantization of compressive samples with stable and robust recovery. *Applied and Computational Harmonic Analysis*, Accepted, 2016.

[50] Holger Rauhut Sjoerd Dirksen, Hans Christian Jung. One-bit compressed sensing with partial gaussian circulant matrices. *arXiv: 1710.03287*, 2017.

[51] J.Z. Sun and V.K. Goyal. Optimal quantization of random measurements in compressed sensing. In *Information Theory, 2009. ISIT 2009. IEEE International Symposium on*, pages 6–10. IEEE, 2009.

[52] Michel Talagrand. *Upper and lower bounds for stochastic processes: modern methods and classical problems*, volume 60. Springer Science & Business Media, 2014.

[53] A. M. Tillmann and M. E. Pfetsch. The computational complexity of the restricted isometry property, the nullspace property, and related concepts in compressed sensing. *IEEE Transactions on Information Theory*, 60(2):1248–1259, 2014.

[54] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[55] Rongrong Wang. Sigma delta qiantization with harmonic frames and partial fourier ensembles. *arXiv preprint arXiv:1511.05671*, 2016.

[56] A. Zymnis, S. Boyd, and E.J. Candés. Compressed sensing with quantized measurements. *Signal Processing Letters, IEEE*, 17(2):149–152, 2010.

Quantized Compressed Sensing for Partial Random Circulant Matrices

Joe-Mei Feng, Felix Krahmer, Rayan Saab

February 15, 2017

Abstract

We provide the first analysis of a non-trivial quantization scheme for compressed sensing measurements arising from structured measurements. Specifically, our analysis studies compressed sensing matrices consisting of rows selected at random, without replacement, from a circulant matrix generated by a random subgaussian vector. We quantize the measurements using stable, possibly one-bit, Sigma-Delta schemes, and use a reconstruction method based on convex optimization. We show that the part of the reconstruction error due to quantization decays polynomially in the number of measurements. This is in-line with analogous results on Sigma-Delta quantization associated with random Gaussian or subgaussian matrices, and significantly better than results associated with the widely assumed memoryless scalar quantization. Moreover, we prove that our approach is stable and robust; i.e., the reconstruction error degrades gracefully in the presence of non-quantization noise and when the underlying signal is not strictly sparse. The analysis relies on results concerning subgaussian chaos processes as well as a variation of McDiarmid's inequality.

1 Introduction

Compressed sensing [8, 9, 14] deals with accurately reconstructing sparse (or approximately sparse) vectors $x \in \mathbb{R}^N$ from relatively few generalized linear measurements of the form $(\langle a_i, x\rangle)_{i=1}^m$, where $m < N$ and where the vectors $a_i \in \mathbb{R}^N$ are chosen appropriately. Accurate reconstruction is theoretically possible because well chosen compressed sensing measurement maps are injective on the "low-complexity" set of sparse vectors. On the other hand, tractable reconstruction algorithms in the compressed sensing context rely heavily on sophisticated, non-linear techniques including convex optimization and greedy numerical methods (e.g., [3, 10, 31]). Consider the $m \times n$ matrix A whose rows are given by the vectors a_i, and denote the possibly noisy compressed sensing measurements by

$$y = Ax + e, \tag{1}$$

where $e \in \mathbb{R}^m$ represents noise. If $\|e\|_2 \leq \epsilon$, and A is chosen appropriately, then standard compressed sensing results guarantee (e.g., [8, 9, 14], see also [17]) that the solution $\hat{x}$ to the optimization problem

$$\min_z \|z\|_1 \quad \text{subject to} \quad \|Az - y\|_2 \leq \epsilon \tag{2}$$

satisfies

$$\|x - \hat{x}\|_2 \leq C(\|e\|_2 + \frac{\|x - x_s\|_1}{\sqrt{s}}). \tag{3}$$

Above, x_s denotes the best s-sparse approximation to x (i.e., the vector with at most s non-zero entries that best approximates x).

The need for sophisticated non-linear decoders such as (2), which can only be reliably implemented on digital computers, implies that compressed sensing is inextricably linked to a digitization (quantization) step. Through quantization, the measurements are converted from continuous valued quantities to elements from a finite set (e.g., $\{\pm 1\}$), so that they can be stored and manipulated (and ultimately used for reconstruction) via digital computers.

Despite the importance of quantization, and a flurry of recent activity focusing on this subject in the compressed sensing context, its treatment remains rather underdeveloped in at least two ways. First, most of the current literature (e.g., [6, 23, 28, 33, 38, 42]) has focused on the most intuitive approach to quantization, namely memoryless scalar quantization (MSQ). However, MSQ is known to have strong theoretical limitations to its reconstruction error guarantees, which we discuss in Section 2.3. Second, all works on the topic to date have only considered compressed sensing matrices A with subgaussian random entries, both for MSQ and for more sophisticated quantization schemes such as $\Sigma\Delta$ quantization, which have been shown to outperform MSQ (see Section 2.3 below for more details).

1.1 Contributions

In this paper, we address the lack of a non-trivial quantization theory for a practically important class of measurement matrices: partial random circulant matrices. Our main result, Theorem 5 shows that if the compressed sensing measurement matrix is a randomly subsampled partial random circulant matrix, and the measurements are quantized by a stable (even 1-bit) Sigma-Delta quantizer, then with an appropriate tractable decoder (which we specify):

- The reconstruction error due to quantization decays polynomially with the number of measurements.
- The recovery is robust to noise and stable with respect to deviations from the sparsity assumption.

Our analysis relies on proving a restricted isometry property for the product of our compressed sensing measurement matrix and the matrix formed by the left singular vectors of an rth order difference operator, which we provide in Proposition 1. For this, we use a combination of a version of McDiarmid's inequality [29], Dudley's inequality [15], and recent results on suprema of chaos processes [24]. As a notable technical difference to previous works (without quantization) studying measurement systems involving random subsampling, our proof explicitly exploits that we are subsampling without replacement. Let us now introduce the necessary background information, starting with partial random circulant matrices, followed by a brief introduction to quantization and to the concentration of measure techniques we employ.

2 Background and notation

2.1 Notation and basic definitions

We denote by $[N]$ the set $\{1, \dots, N\}$ and by e_k the k-th standard basis vector. A vector $x \in \mathbb{R}^N$ is s-sparse if only s of its entries are non-vanishing, that is, its support $T = \mathrm{supp}(x) = \{j \in [N] : x_j \neq 0\}$ satisfies $|T| = s$. Throughout, the matrix $F = \left(e^{2\pi ijk/N}\right)_{j,k=1}^{N}$ is the unnormalized $N \times N$ discrete Fourier transform matrix, and $\bar{F}$ denotes the complex conjugate of F. That is, $F\bar{F} = \bar{F}F = NId$. We say that a matrix A satisfies the restricted isometry property of order s and constant δ, if for all s-sparse vectors x

$$(1-\delta)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1+\delta)\|x\|_2^2.$$

Given a vector $x \in \mathbb{R}^N$, we denote by $\hat{X} \in \mathbb{R}^{N \times N}$ the diagonal matrix with $\hat{x} := Fx$ on the diagonal. For a matrix A, A_k denotes its k-th column.

We write $f \lesssim g$ for two functions f and g if they are defined on the same domain D and there exists an absolute constant C such that $f(y) \leq Cg(y)$ for all $y \in D$, $f \gtrsim g$ is defined analogously. Given a full-rank matrix $A \in \mathbb{R}^{m \times d}$ with $m > d$, its pseudo-inverse is given by $A^\dagger = (A^*A)^{-1}A^*$.

2.2 Partial random circulant matrices

Given a vector $\xi = (\xi_1, \xi_2, \dots, \xi_N) \in \mathbb{R}^N$, the corresponding circulant matrix $\Phi = \Phi(\xi) \in \mathbb{R}^{N \times N}$ is defined by

$$C_\xi = \begin{bmatrix} \xi_1 & \xi_2 & \xi_3 & \cdots & \xi_N \\ \xi_N & \xi_1 & \xi_2 & \cdots & \xi_{N-1} \\ \vdots & & & & \vdots \\ \xi_2 & \xi_3 & \xi_4 & \cdots & \xi_1 \end{bmatrix}. \tag{4}$$

In this paper we consider random circulant matrices C_ξ arising from random vectors ξ whose entries are independent L-subgaussian random variables with variance 1 and mean 0, in the sense of the following definition.

Definition 1 (see, e.g., [40]). *A random variable X is called L-subgaussian if*

$$\mathbb{P}(|X| > t) \leq \exp(1 - t^2/L^2). \tag{5}$$

Up to absolute multiplicative constants, the subgaussian parameter L is equivalent to the subgaussian norm $\|X\|_{\Psi_2}$ defined as $\|X\|_{\Psi_2} = \sup_{p\geq 1} p^{-1/2}(\mathbb{E}|X|^p)^{1/p}$. Specifically, (5) *implies that [40]*

$$\|X\|_{\psi_2} \leq \sqrt{\tfrac{e}{2}} L. \tag{6}$$

A partial random circulant matrix is obtained from a random circulant matrix by sampling the rows of the latter. In this paper, we consider only sampling without replacement, thus obtaining the following definition.

Definition 2. *Let $\Phi = C_\xi \in \mathbb{R}^{N\times N}$ be a random circulant matrix as in* (4) *and, for $m \leq N$, let $\Omega = (\Omega_1, \ldots, \Omega_m)$ be a random vector obtained by sampling from $[N]$ without replacement. That is, Ω is drawn uniformly at random from the set*

$$\Xi := \{\omega \in [N]^m : \omega_i \neq \omega_j \text{ for } i \neq j\}. \tag{7}$$

Then the associated partial random circulant matrix *is given by*

$$A = R_\Omega \Phi.$$

where R_Ω is the subsampling operator

$$\mathbb{R}^{m\times N} \ni R_\Omega = \sum_{j=1}^{m} e_j e_{\Omega_j}^*.$$

Partial random circulant matrices are important to the practical application of compressed sensing. This is due to the simple observation that a circular convolution of a signal $x \in \mathbb{R}^N$ with a "filter" $\widetilde{\xi} \in \mathbb{R}^N$, as given by the vector $y = x \circledast \tilde{\xi} \in \mathbb{R}^N$ with entries

$$y_j := \sum_{i=1}^{N} x_i \widetilde{\xi}_{j-i \bmod n},$$

can be represented by the action of a circulant matrix. Indeed one has $x \circledast \tilde{\xi} = C_\xi x$, where

$\xi \in \mathbb{R}^N$ is defined via $\xi_{N-j+1} = \widetilde{\xi}_j$ for $j \in \{1, ..., N\}$ and C_ξ is as in (4). Consequently, as the convolution is commutative, one has $C_\xi x = C_x \xi$; we will repeatedly make use of this observation.

Due to the ubiquity of convolutions in signal processing applications, partial random circulant matrices, modeling subsampled random convolutions, have played an important role in the development of compressed sensing applications such as radar imaging, Fourier optical imaging, and wireless channel estimation (see, e.g., [21, 35]). Recovery guarantees for partial circulant matrices have been an active area of research in the last decade, the best known results have recently been proved by Mendelson, Rauhut, and Ward [30].

2.3 Quantization

In the compressed sensing context, quantization is the map that replaces the vector $y = Ax + e \in \mathbb{R}^m$ by a representation that uses a finite number of bits. Most often, practical quantization maps are of the form

$$\mathcal{Q} : \mathbb{R}^m \to \mathcal{A}^m$$

where $\mathcal{A} \subset \mathbb{R}$ is a finite set, called the quantization alphabet. Both memoryless scalar quantization and $\Sigma\Delta$ quantization, which we will discuss in the next paragraphs, execute quantization maps of this form.

The most natural and common choices of alphabets have equispaced elements. As representatives for such alphabets we will focus on the so-called mid-rise alphabet with $2L$ levels and step-size δ, denoted by $\mathcal{A}_L^\delta$ and given by $\mathcal{A}_L^\delta := \big\{\pm(2\ell+1)\delta/2,\ \ell \in \{0,...,L-1\}\big\}$. The minimal instance of such an alphabet is the 1-bit quantization alphabet, which we denote by $\mathcal{A} = \{-1,+1\}$.

The fact that $\mathcal{Q}$ outputs a vector of alphabet elements allows the quantization to be implemented progressively. That is, one can relate each entry of the quantized vector to some measurement and each subsequent measurement can then be quantized in a way that depends on previous measurements. This idea is exploited in $\Sigma\Delta$ schemes.

Memoryless scalar quantization

Memoryless scalar quantization is an intuitive approach to digitizing compressed sensing measurement. It simply uses a scalar quantizer

$$\begin{aligned} Q_{\mathcal{A}} : \mathbb{R} &\to \mathcal{A} \\ z &\mapsto \arg\min_{v\in\mathcal{A}} |z-v| \end{aligned} \tag{8}$$

to quantize every entry of y independently. Using a standard compressed sensing recovery algorithm such as (2), one can use the robustness of standard compressed sensing reconstruction algorithms (3) to bound the reconstruction error. Such results guarantee that the reconstruction error decays as the size of the alphabet increases. However, they do not guarantee error decay as one takes more measurements. One could argue that a better reconstruction algorithm or a sharper analysis would alleviate this issue, but that is hardly the case. Indeed, consider working with a fixed quantization alphabet, as one would do in practice due to fixing the quantization hardware. Then, as shown by Goyal, Vetterli, and Thao [18], the error in reconstructing a k-sparse signal from its m MSQ-quantized measurements cannot decay faster than k/m, even when using an optimal decoder. This means that by linearly increasing the number of measurements, and hence increase the number of bits used, denoted by $\mathcal{R}$ (for rate), one can, at best, only linearly decrease the reconstruction error, denoted $\mathcal{D}_{\mathcal{MSQ}}$ (for distortion). That is, the rate-distortion relationship associated with MSQ satisfies

$$\mathcal{D}_{MSQ}(\mathcal{R}) \geq C\mathcal{R}^{-1}. \tag{9}$$

This lower bound stands in sharp contrast to the rate-distortion relationship that an optimal assignment of bits (for encoding k-sparse vectors in the unit-ball of $\mathbb{R}^N$) yields, namely (see, e.g., [5])

$$\mathcal{D}^*(\mathcal{R}) \leq C\frac{N}{k}e^{-c\mathcal{R}/k}.$$

In this sense MSQ is far from optimal. One factor preventing MSQ from being optimal in general, is that it does not exploit any correlations among the measurements, as it treats each measurement independently of the others.

Sigma-Delta quantization

Sigma-Delta ($\Sigma\Delta$) quantization is an alternative quantization method that, in its simplest form, works by scalar quantizing the sum of the current measurement and a state variable, and then updating the state variable. It is through the state variable that the dependencies between the measurements are accounted for in the quantization. $\Sigma\Delta$ schemes were proposed in the 1960's [22] for quantizing bandlimited functions and have seen widespread use in practice, particularly in audio applications [32]. For almost 40 years, there was no precise understanding of $\Sigma\Delta$ from a mathematical perspective, before recently, following the seminal work of Daubechies and Devore in [11], a number of works analyzed $\Sigma\Delta$ schemes for bandlimited functions from a mathematical perspective [12, 13, 19, 27].

In addition, $\Sigma\Delta$ schemes have recently been shown to be well suited for quantizing finite-frame expansions [1, 2, 4, 25] as well as compressed sensing measurements [20, 26, 36, 37]. We review these results in the following subsection, and we now focus on the relevant details of $\Sigma\Delta$ quantization schemes.

In the simplest $\Sigma\Delta$ scheme, a first order $\Sigma\Delta$ quantizer, the state variable u_i accounts for the accumulated quantization error. That is, the quantizer applies to the measurements y_i the iteration

$$q_i = Q_{\mathcal{A}}(y_i + u_{i-1}) \tag{10}$$

$$u_i = u_{i-1} + y_i - q_i. \tag{11}$$

Here $Q_{\mathcal{A}}$ is the scalar quantizer (8). In an r^{th}-order $\Sigma\Delta$ scheme, the first order finite difference Δ (given by $(\Delta u)_i := u_i - u_{i-1}$, and appearing in (11)) is replaced by an r^{th}-order finite difference Δ^r. Moreover, before applying the scalar quantizer, some quantization rule $\rho : \mathbb{R}^{r+1} \to \mathbb{R}$ is applied.

That is, the quantized measurement vector q with entries $q_i \in \mathcal{A}$ is computed via the recursion

$$q_i = Q_{\mathcal{A}}\left(\rho(y_i, u_{i-1}, u_{i-2}, \ldots, u_{i-r})\right), \tag{12}$$

$$u_i = y_i - q_i - \sum_{j=1}^{r} \binom{r}{j} (-1)^j u_{i-j}. \tag{13}$$

Using the first-order difference matrix D with entries given by

$$D_{i,j} := \begin{cases} 1 & \text{if } i = j \\ -1 & \text{if } i = j+1 \\ 0 & \text{otherwise} \end{cases}, \tag{14}$$

the relationship between x, u, and q can be concisely written in matrix-vector notation as

$$D^r u \;=\; y - q. \tag{15}$$

The inverse D^{-r} will play a crucial role in our analysis, which is why we fix the notation

$$D^{-r} = USV^*,$$

for its singular value decomposition throughout this paper.

Recalling that $(D^{-1}z)_j = \sum_{i=1}^{j} z_i$, in the case of first order schemes (where $r = 1$) the state variable u can be interpreted as an accumulated error, as can be seen by appying D^{-1} to the equation above. It intuitively follows that it is crucial for the sequence of state variable u to be bounded in this case. This intuition can be made precise and generalizes to higher order schemes. For this reason we seek *stable* r^{th}*-order schemes*, i.e., schemes for which (12) and (13) result in

$$\|u\|_\infty \leq C_{\rho,Q}(r)$$

for all $N \in \mathbb{N}$, and $y \in \mathbb{R}^N$ with $\|y\|_\infty \leq 1$. Importantly, we require that $C_{\rho,Q} : \mathbb{N} \mapsto \mathbb{R}^+$ be entirely independent of both N and y. One can show that stable r^{th}-order $\Sigma\Delta$ schemes exist with $C_{\rho,Q}(r) = O((Cr)^r)$ for some constant C [13, 19], even when $\mathcal{A}$ is a 1-bit alphabet, but that there are fundamental lower bounds on C and no better dependence on r can be achieved [7, 27].

2.4 Probabilistic Tools

We will use a number of different probabilistic tools for different parts of our argument. We state them here for convenience. The first one is a variation of McDiarmid's inequality. Note that it closely relates to the Azuma-Hoeffding inequality and the method of bounded differences.

Theorem 1 ([29], Theorem 3.14)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(\emptyset, \Omega) = \mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq ... \subseteq \mathcal{F}_m$ a filtration in $\mathcal{F}$. Consider a bounded random variable X, and set $X_k := \mathbb{E}(X|\mathcal{F}_k)$. Define the sum of squared conditional ranges*

$$R^2 = \sum_{k=1}^{m} ran_k^2$$

where

$$ran_k := \sup(X_k|\mathcal{F}_{k-1}) + \sup(-X_k|\mathcal{F}_{k-1}),$$

and denote its (essential) supremum by

$$\hat{r}^2 := \sup R^2.$$

Then,

$$\mathbb{P}(X - \mathbb{E}(X) \geq t) \leq e^{-2t^2/\hat{r}^2}.$$

A second tool that we will be using is Dudley's inequality. In order to formulate the result, we recall the definitions of the covering number and of subgaussian random variables.

Definition 3. *Let (S,d) be a metric space and $\epsilon > 0$. A subset $\mathcal{N}_\epsilon$ of S is called an ϵ-net if every point in S can be approximated to within ϵ by some point in $\mathcal{N}_\epsilon$, i.e., for all $x \in S$ there exists $y \in \mathcal{N}_\epsilon$ such that $d(x,y) < \epsilon$. The* covering number $\mathcal{N}(S,d,\epsilon)$ *is the minimal cardinality of an ϵ-net of S.*

Theorem 2 (Dudley's inequality [15])**.** *Let Z_x be a random variable depending on $x \in T$, for some set T and define $d(x,y) = \|Z_x - Z_y\|_{\Psi_2}$, if*

$$\mathbb{P}(|Z_x - Z_y| > t) \lesssim \exp\Big(- t^2/\|Z_x - Z_y\|_{\Psi_2}^2\Big),$$

then for any $x_0 \in T$

$$\mathbb{P}(\sup|Z_x - Z_{x_0}| > t) \lesssim \exp\Big(- t^2/\big(\int_0^{\sup_{x\in D_{N,s}} \|Z_x\|_{\Psi_2}} \sqrt{\log \mathcal{N}(D_{N,s}, d(x,y), \epsilon)}d\epsilon\big)^2\Big).$$

A third result that we will be using concerns subgaussian chaos processes. Its original version involves the Talagrand γ_2 functional, an intricate complexity parameter related to the generic chaining [39], which can be bounded in terms of covering numbers via Dudley's inequality (Theorem 2). To avoid discussing the generic chaining methodology in detail, we state a combined version in terms of only these upper bounds.

Theorem 3 ([24])**.** *Let $\mathcal{C}$ be a set of matrices and consider the complexity parameters*

$$d_F(\mathcal{C}) = \sup_{C\in\mathcal{C}} \|C\|_F, \qquad d_{2\to 2}(\mathcal{C}) = \sup_{C\in\mathcal{C}} \|C\|_{2\to 2}, \qquad D(\mathcal{C}) = \int_0^{d_{2\to 2}(\mathcal{C})} \sqrt{\log \mathcal{N}(\mathcal{C}, \|\cdot\|_{2\to 2}, u)}\ du.$$

Let ξ be a random vector whose entries ξ_j are independent, mean-zero, variance 1, L-subgaussian random variables. Then, for $t > 0$, the random variable

$$C_{\mathcal{C}}(\xi) = \sup_{C\in\mathcal{C}} |\|C\xi\|_2^2 - \mathbb{E}_\xi \|C\xi\|_2^2$$

satisfies

$$\mathbb{P}(C_{\mathcal{C}}(\xi) \geq c_1 E + t) \leq 2\exp(-c_2 \min\{\frac{t^2}{V^2}, \frac{t}{U}\}),$$

where

$$E = D(\mathcal{C})(D(\mathcal{C}) + d_F(\mathcal{C})) + d_F(\mathcal{C})d_{2\to 2}(\mathcal{C}), \qquad V = d_{2\to 2}(\mathcal{C})(D(\mathcal{C}) + d_F(\mathcal{C})), \qquad U = d_{2\to 2}^2(\mathcal{C}),$$

and the constants c_1, c_2 depend only on L.

3 Related Work

3.1 $\Sigma\Delta$ quantization of finite-frame expansions

The first paper analyzing $\Sigma\Delta$ quantization of finitely many measurements of finite-dimensional vectors was [1], initiating a series of papers on the subject. For example, the papers [1, 2, 4, 25] all studied $\Sigma\Delta$ quantization when one collects $m > N$ linear measurements $y_i = \langle a_i, x\rangle$ of $x \in \mathbb{R}^N$, where the collection $(a_i)_{i=1}^m$ spans $\mathbb{R}^N$ (and is called a finite-frame). In this *finite-frame* setting, [1] showed that the reconstruction error associated with first order $\Sigma\Delta$ quantization can be made to decay linearly with the number of measurements, hence the bit-rate. With this first order $\Sigma\Delta$ approach, the upper bound on the error already matched the lower bound (9) associated with MSQ. Using higher order $\Sigma\Delta$ schemes, subsequent papers (e.g., [2, 4]) showed that the error can be made polynomial in the number of measurements, significantly outperforming the MSQ lower bound. Importantly, the linear reconstruction scheme proposed in [4] to approximate x from its quantized finite-frame measurements also proved fruitful in the compressed sensing context. Denoting by A the $m \times N$ matrix ($m \geq N$) having a_i as its rows, the r^{th}*-order Sobolev dual* of A is the $N \times m$ matrix

$$B := (D^{-r}A)^\dagger D^{-r},$$

which is easily seen to be a left-inverse of A. The approach of [4] was to estimate x from q via $\hat{x} = Bq$, yielding error rates that decayed like m^{-r} (i.e., polynomially in the number of measurements and bits) provided the rows of A obeyed some smoothness conditions.

3.2 $\Sigma\Delta$ quantization of compressed sensing measurements

The paper [20], soon followed by [16, 26], was first to study $\Sigma\Delta$ quantization of compressed sensing measurements. They focused on the setup where the compressed sensing matrix is subgaussian, the underlying signal is *strictly sparse*, and no noise contaminates the measurements. They analyzed a two-stage approach to signal recovery whereby one uses a standard decoder like (2) to estimate the support of the k-sparse signal, then applies the Sobolev dual of the associated $m \times k$ sub-matrix of A to q. With this approach, the reconstruction error was again shown to decay polynomially in the number of measurements. The proofs in [20, 26] relied on bounding the smallest singular value of a certain anisotropic random matrix, while [16] significantly simplified the analysis by using an approach based on the restricted isometry property. These results showed that frame-theoretic quantization techniques could be extended to the compressed sensing setup. On the other hand, the reliance of [20, 26] on a two-step approach involving support recovery meant that obtaining a result for compressed sensing measurements of *arbitrary* signals in the presence of noise would be difficult.

More recently, in [**?**] a decoder based on convex optimization was proposed (to replace the two-step approach) and analyzed, with the main result being that it could handle both arbitrary signals and measurement noise (bounded by ϵ). Specifically, if q results from quantizing compressed sensing measurements y (as in (1)) using an r^{th}-order $\Sigma\Delta$ scheme, one approximates x with $\hat{x}$ via

$$(\hat{x}, \hat{e}) := \arg\min_{(z,\nu)} \|z\|_1 \quad \text{subject to } \|D^{-r}(Az + \nu - q)\|_2 \leq \gamma(r)\sqrt{m} \\ \text{and} \quad \|\nu\|_2 \leq \epsilon\sqrt{m}, \tag{16}$$

where $\gamma(r)$ depends on the quantization scheme used. The resulting approximation error due to quantization in [**?**] decays as $m^{-r+1/2}$, i.e., polynomially in m, and the approach is shown to be stable and robust. As in [16], a main ingredient in the proofs of [**?**] is an analysis based on the restricted isometry properties of certain matrices arising from the interaction of the difference matrix with the compressed sensing matrix. Indeed, the following result, which we will also use, is proved in [**?**].

Theorem 4. *[?] Let A be an $m \times N$ matrix, and let $k, l \in \{1, ..., m\}$. Suppose that $\frac{1}{\sqrt{\ell}} P_\ell V^* A$ satisfies the restricted isometry property of order $2k$ and constant $\delta < 1/9$. Denote by $Q^r_{\Sigma\Delta}$ a stable rth order $\Sigma\Delta$ quantizer. Then, for all $x \in \mathbb{R}^N$ with $\|Ax\|_\infty \leq \mu < 1$ and all $e \in \mathbb{R}^m$ with $\|e\|_\infty \leq \epsilon < 1 - \mu$ the estimate $\hat{x}$ obtained by solving* (16) *with $q = Q^r_{\Sigma\Delta}(Ax + e)$ satisfies*

$$\|\hat{x} - x\|_2 \leq C_1 \left(\frac{m}{\ell}\right)^{-r+1/2} \delta + C_2 \frac{\sigma_k(x)}{\sqrt{k}} + C_3 \sqrt{\frac{m}{\ell}} \epsilon, \tag{17}$$

where the constants C_1, C_2, C_3 depend on the quantizer, but not the dimensions of the problem.

The combination of stability, robustness, quantization error decay, and practicability make the $\Sigma\Delta$ quantization approach, followed by recovery via (16) amenable to practical applications where one has the freedom to select subgaussian compressed sensing matrices. Nevertheless, the only matrices Φ for which [**?**] proved that the assumptions of Theorem 4 hold are subgaussian. As such, the results of [**?**] do not apply to important practical setups such as system identification, radar, and coded-aperture imaging, where *structured* random matrices such as partial random circulant ones arise naturally in the compressed sensing context (see, e.g., [21, 34]). The only result we are aware of (aside from those of this manuscript) that addresses quantization in the context of structured random measurement matrices is that of [41]. [41] shows that first order $\Sigma\Delta$ quantization coupled with an appropriate decoder yields an error decaying as $\left(\frac{m}{k^4 \log N}\right)^{-1/2}$, when the measurement matrix is a randomly selected $m \times N$ submatrix of the $N \times N$ discrete Fourier transform matrix. Consequently the results are only meaningful when m scales like k^4, which is considerably worse than the linear scaling of m with k (up to log factors) arising in Theroem 4 and commonly in compressed sensing without quantization. One of our main contributions (Theorem 5) is to show that such a linear scaling (up to log factors) also holds for certain structured random measurements, specifically for random circulant matrices.

4 Main results

In this section, we prove the following theorem, which is the main result of this paper.

Theorem 5. *Denote by $Q^r_{\Sigma\Delta}$ a stable rth order $\Sigma\Delta$ quantizer. Let A be an $m \times N$ partial random circulant matrix associated to a vector with independent L-subgaussian entries with mean 0 and variance 1. Suppose that $N \geq m \geq (C\eta)^{\frac{1}{1-2\alpha}} s \log^{\frac{2}{1-2\alpha}} N \log^{\frac{2}{1-2\alpha}} s$, for some $\eta > 1$ and $\alpha \in [0, 1/2)$. With probability exceeding $1 - e^{-\eta}$, the following holds:*
For all $x \in \mathbb{R}^N$ with $\|Ax\|_\infty \leq \mu < 1$ and all $e \in \mathbb{R}^m$ with $\|e\|_\infty \leq \epsilon < 1 - \mu$ the estimate $\hat{x}$ obtained by solving (16) *satisfies*

$$\|\hat{x} - x\|_2 \leq C_1 \left(\frac{m}{\ell}\right)^{-r+1/2} \delta + C_2 \frac{\sigma_k(x)}{\sqrt{k}} + C_3 \sqrt{\frac{m}{\ell}} \epsilon.$$

Here C, C_1, C_2, C_3 are constants that only depend on r and L.

Proof. Theorem 5 can be immediately obtained from Theorem 4, which requires a bound on the restricted isometry constants of $P_\ell V^* R_\Omega C_\xi$ where $\ell = m(\frac{s}{m})^\alpha$, and Proposition 1 below, which provides the required bound. □

Proposition 1. *Consider the same setup and assumptions as Theorem 5; in particular assume that $m \geq (C\eta)^{\frac{1}{1-2\alpha}} s \log^{\frac{2}{1-2\alpha}} N \log^{\frac{2}{1-2\alpha}} s$, for some $\eta > 1$ and $\alpha \in [0, 1/2)$. Setting $\ell = m(\frac{s}{m})^\alpha$, we have*

$$\mathbb{P}\Big(\sup_x |\| \frac{1}{\sqrt{\ell}} P_\ell V^* R_\Omega C_x \xi\|_2^2 - 1| > \frac{1}{9} \Big) < e^{-\eta},$$

where the supremum is over all s-sparse vectors. In other words, with probability exceeding $1 - e^{-\eta}$, the matrix $\frac{1}{\sqrt{\ell}} P_\ell V^* R_\Omega C_\xi$ satisfies the restricted isometry property of order s, with constant $1/9$.

Proof. Note that by the triangle inequality,

$$\begin{aligned}
&\sup_x \Big| \| \frac{1}{\sqrt{\ell}} P_\ell V^* R_\Omega C_x \xi\|_2^2 - 1 \Big| \\
&\leq \sup_x \Big(\Big| \| \frac{1}{\sqrt{\ell}} P_\ell V^* R_\Omega C_x \xi\|_2^2 - \mathbb{E}[\| \frac{1}{\sqrt{\ell}} P_\ell V^* R_\Omega C_x \xi\|_2^2 | \Omega] \Big| + \\
&\quad \Big| \mathbb{E}[\| \frac{1}{\sqrt{\ell}} P_\ell V^* R_\Omega C_x \xi\|_2^2 | \Omega] - \mathbb{E}\| \frac{1}{\sqrt{\ell}} P_\ell V^* R_\Omega C_x \xi\|_2^2 \Big| + \\
&\quad |\mathbb{E}\| \frac{1}{\sqrt{\ell}} P_\ell V^* R_\Omega C_x \xi\|_2^2 - 1| \Big).
\end{aligned} \tag{18}$$

Thus, the proof of Proposition 1 boils down to controlling each of the summands in (18). To that end, Lemma 1 (below) shows that the third summand is bounded by $\frac{sm}{\ell N}$, while Lemma 2 and Lemma 3 bound the probability that the remaining summands exceed $\frac{1}{18}$ and $\frac{1}{36}$ respectively. Our bound on m (potentially with an increased value of C) ensures that $\frac{sm}{\ell N} \leq \frac{s}{\ell} = \left(\frac{s}{m}\right)^{1-\alpha} \leq \frac{1}{36}$ and the result follows using a union bound. □

Lemma 1. *Given the same setup as in Theorem 5 and Proposition 1, one has*

$$|\mathbb{E}\| \frac{1}{\sqrt{\ell}} P_\ell V^* R_\Omega C_x \xi\|_2^2 - 1| \leq \frac{(s-1)(m-\ell)}{\ell(N-1)} \leq \frac{sm}{\ell N}.$$

Proof. Denoting by $c_{i,j}$ the (i,j)-th entry of C_x and noting that we are sampling without replacement, we observe that for $p \neq q \in [m]$

$$\begin{aligned}
\mathbb{E}(c_{\Omega(p),k} c_{\Omega(q),k}) &= \frac{1}{N(N-1)} \sum_{u \neq v = 1}^{N} c_{u,k} c_{v,k} = \frac{1}{N(N-1)} \Big(\sum_{u,v=1}^{N} c_{u,k} c_{v,k} - \sum_{u=1}^{N} c_{u,k}^2 \Big) \\
&= \frac{1}{N(N-1)} \Big(\sum_{u,v=1}^{N} c_{u,k} c_{v,k} - \sum_{u=1}^{N} x_u^2 \Big) = \frac{1}{N(N-1)} \Big(\big(\sum_{u=1}^{N} x_u \big)^2 - 1 \Big).
\end{aligned} \tag{19}$$

The last two equalities both use the fact that each row of C_x is a shifted copy of x. Furthermore

$$\Big| \mathbb{E}\| \frac{1}{\sqrt{\ell}} P_\ell V^* R_\Omega C_x \xi\|_2^2 - 1 \Big| = \Big| \mathbb{E}\| \frac{1}{\sqrt{\ell}} P_\ell V^* R_\Omega C_x\|_F^2 - 1 \Big|$$

$$
\begin{aligned}
&= \Big| \frac{1}{\ell} \mathbb{E} \sum_{j=1}^{\ell} \sum_{k=1}^{N} |\sum_{p=1}^{m} v_{jp} c_{\Omega(p),k}|^2 - 1 \Big| \\
&= \Big| \frac{1}{\ell} \sum_{j=1}^{\ell} \sum_{k=1}^{N} \big(\sum_{p=1}^{m} v_{jp}^2 \mathbb{E} c_{\Omega(p),k}^2 + \sum_{\substack{p,q=1 \\ p \neq q}}^{m} v_{jp} v_{jq} \mathbb{E} c_{\Omega(p),k} c_{\Omega(q),k} \big) - 1 \Big| \\
&= \Big| \frac{1}{\ell} \sum_{j=1}^{\ell} \big(1 + \frac{(\sum_{i=1}^{N} x_i)^2 - 1}{N-1} \sum_{\substack{p,q=1 \\ p \neq q}}^{m} v_{jp} v_{jq} \big) - 1 \Big|
\end{aligned}
$$

where in the last equality we used (19) and the fact that the rows of both C_x and V are normalized. Using that x is s-sparse, it follows that

$$
\begin{aligned}
\Big| \mathbb{E} \| \frac{1}{\sqrt{\ell}} P_\ell V^* R_\Omega C_x \xi \|_2^2 - 1 \Big| &\leq \Big| \frac{s-1}{\ell(N-1)} \Big(\sum_{j=1}^{\ell} (\sum_{p=1}^{m} v_{jp})^2 - \sum_{j=1}^{\ell} \sum_{p=1}^{m} v_{jp}^2 \Big) \Big| \\
&= \frac{s-1}{\ell(N-1)} \Big| \| V^* (1, \dots, 1)^T \|_2^2 - \ell \Big| \\
&\leq \frac{s-1}{\ell(N-1)} \Big| \| V \|_{2\to 2}^2 m - \ell \Big| \\
&= \frac{(s-1)(m-\ell)}{\ell(N-1)}.
\end{aligned}
$$

□

Lemma 2. *Consider again the setup of Theorem 5 and Proposition 1 and denote by $D_{N,s}$ the set of all s-sparse vectors in $\mathbb{R}^N$. Then*

$$
\mathbb{P}\Big(\sup_{x \in D_{N,s}} \Big| \| \frac{1}{\sqrt{\ell}} P_\ell V^* R_\Omega C_x \xi \|_2^2 - \mathbb{E}_\xi \Big[\| \frac{1}{\sqrt{\ell}} P_\ell V^* R_\Omega C_x \xi \|_2^2 \Big| \Omega \Big] \Big| > \frac{1}{18} \Big| \Big) \leq \frac{1}{2} e^{-\eta}.
$$

Proof. We will apply Theorem 3 conditionally given Ω with $\mathcal{C} = \{ \frac{1}{\sqrt{\ell}} P_\ell V^* R_\Omega C_x : x \in D_{N,s} \}$. This set is almost the same as the one considered in the proof of Theorem 4.1 in [24], the only differences being the additional projection P_ℓ and our normalization factor of $\frac{1}{\sqrt{\ell}}$ (instead of $\frac{1}{\sqrt{m}}$ in [24]). Indeed, since $\|P_\ell\|_{2\to 2} \leq 1$ we can estimate the necessary parameters for applying Theorem 3 exactly as in the proof of Theorem 4.1 in [24]. This yields

$$
d_{2\to 2}(\mathcal{C}) \leq \sqrt{\frac{s}{\ell}}, \qquad d_F(\mathcal{C}) \leq \sqrt{\frac{m}{\ell}}, \qquad D(\mathcal{C}) \leq \sqrt{\frac{s}{\ell}} \log N \log s.
$$

Consequently for c_1, c_2, and E as in Theorem 3, we have

$$
\begin{aligned}
E &\leq \sqrt{\frac{s}{\ell}} \log N \log s \left(\sqrt{\frac{s}{\ell}} \log N \log s + \sqrt{\frac{m}{\ell}} \right) + \sqrt{\frac{m}{\ell}} \sqrt{\frac{s}{\ell}} \\
&\leq \Big(\frac{s}{m} \Big)^{1-\alpha} \log^2 N \log^2 s + 2 \Big(\frac{s}{m} \Big)^{1-2\alpha} \log N \log s \leq \frac{1}{36 c_1}.
\end{aligned}
$$

Here, the second inequality follows from our choice of ℓ and the last inequality follows from our assumption on m in Theorem 5 (potentially adjusting the constant C). Again adjusting the constant, we similarly obtain

$$
V \leq \sqrt{\frac{c_2}{4\eta}} \qquad \text{and} \qquad U \leq \frac{c_2}{4\eta}.
$$

Hence the probability is bounded by $2e^{-4\eta}$. Finally, as $\eta \geq 1$, $e^{-4\eta} \leq \frac{1}{4} e^{-\eta}$ and the result follows by taking the expectation over Ω.

□

Lemma 3. *With the same notation as before, we have*

$$\begin{aligned}&\mathbb{P}(\sup_{x\in D_{N,s}} |\mathbb{E}[\|\frac{1}{\sqrt{\ell}}P_\ell V^* R_\Omega C_x\xi\|_2^2|\Omega] - \mathbb{E}\|\frac{1}{\sqrt{\ell}}P_\ell V^* R_\Omega C_x\xi\|_2^2| > \frac{1}{36})\\ &\leq C' \exp(-c/(\frac{\sqrt{sm}}{\ell}\log N \log m)^2) \leq \frac{1}{2}e^{-\eta}\end{aligned}$$

where c, C' are constants that depends only on L.

Proof. The proof is a direct application of Theorem 2 for the random variable

$$Z_x := \mathbb{E}\left[\|\frac{1}{\sqrt{\ell}}P_\ell V^* R_\Omega C_x\xi\|_2^2 - \mathbb{E}\|\frac{1}{\sqrt{\ell}}P_\ell V^* R_\Omega C_x\xi\|_2^2\middle|\Omega\right] = \|\frac{1}{\sqrt{\ell}}P_\ell V^* R_\Omega C_x\|_F^2 - \mathbb{E}\|\frac{1}{\sqrt{\ell}}P_\ell V^* R_\Omega C_x\|_F^2$$

to find the supremum of the deviation. Since Theorem 2 requires the covering number with respect to the metric $d(x,y) := \|Z_x - Z_y\|_{\Psi_2}$ we need a bound for $d(x,y)$, which we provide in Lemma 5 below. Specifically, the first inequality in Lemma 3 follows from Theorem 2 together with Lemma 1 and Lemma 2 above. Indeed, applying Lemma 5 with $y = 0$ yields

$$\sup_{x,y}\|Z_x\|_{\Psi_2} \leq \frac{\sqrt{m}}{\ell}\|x\|_\infty \leq \frac{\sqrt{m}}{\ell}\|F(x)\|_\infty \leq \frac{\sqrt{m}}{\ell}\|x\|_1 \leq \frac{\sqrt{sm}}{\ell}\|x\|_2 \leq \frac{\sqrt{sm}}{\ell}. \tag{20}$$

To bound the integral in Theorem 2, we note that

$$\mathcal{N}(D_{N,s}, \frac{\sqrt{m}}{\ell}\|\cdot\|_\infty, \epsilon) = \mathcal{N}(D_{N,s}, \frac{1}{\sqrt{m}}\|\cdot\|_\infty, \frac{\ell}{m}\epsilon),$$

and hence applying the argument in [24, Section 4] scaled by $\frac{m}{\ell}$,

$$\begin{aligned}&\int_0^{\sup_x \|Z_x\|_{\Psi_2}} \sqrt{\log \mathcal{N}(D_{N,s}, \frac{1}{\sqrt{m}}\|\cdot\|_\infty, \frac{\ell}{m}\epsilon)}d\epsilon\\ &\lesssim \frac{\sqrt{sm}}{\ell}\log N \log s.\end{aligned}$$

For the second inequality note that by the definition of ℓ and the assumed lower bound on m

$$\begin{aligned}(\frac{\sqrt{sm}}{\ell}\log N\log s)^2 &= \left(\frac{s}{m}\right)^{1-2\alpha}\log^2 N \log^2 s &(21)\\ &\leq C^{-1}\eta^{-1}. &(22)\end{aligned}$$

The result follows from the assumption that $\eta \geq 1$ as in the proof of Lemma 2. □

All that remains now is to prove Lemma 5. Before that, we derive a technical bound required for its proof.

Lemma 4. *Let ω, $\omega' \in \Xi = \{\omega \in [N]^m : \omega_i \neq \omega_j \text{ for } i \neq j\}$ be such that ω differs from ω' in at most two components. Then the function*

$$f(\omega) := \|\frac{1}{\sqrt{\ell}}P_\ell V^* R_\omega C_x\|_F^2 - \|\frac{1}{\sqrt{\ell}}P_\ell V^* R_\omega C_y\|_F^2$$

satisfies

$$|f(\omega) - f(\omega')| \leq \frac{12}{\ell}\|x-y\|_\infty,$$

where $\|x\|_\infty := \|Fx\|_\infty$.

Proof. Note that, as a circulant matrix is diagonalized by the Fourier transform,

$$\begin{aligned}f(\omega) &= \|\frac{1}{\sqrt{\ell}}P_\ell V^* R_\omega C_x\|_F^2 - \|\frac{1}{\sqrt{\ell}}P_\ell V^* R_\omega C_y\|_F^2\\ &= \|\frac{1}{\sqrt{\ell}}P_\ell V^* R_\omega F^{-1}\hat{X}F\|_F^2 - \|\frac{1}{\sqrt{\ell}}P_\ell V^* R_\omega F^{-1}\hat{Y}F\|_F^2\end{aligned}$$

$$= \frac{1}{\ell N}\|P_\ell V^* R_\omega \overline{F}\hat{X}\|_F^2 - \frac{1}{\ell N}\|P_\ell V^* R_\omega \overline{F}\hat{Y}\|_F^2$$
$$= \frac{1}{\ell N}\sum_{k=1}^{N}\big(|\hat{x}_k|^2 - |\hat{y}_k|^2\big)\|P_\ell V^* R_\omega \overline{F}_k\|_2^2, \tag{23}$$

where F denotes the non-normalized Fourier transform, F_k^T its k-th row, and $\hat{x} = Fx$.

We first consider the case that ω and ω' differ only in one component, say the first (without loss of generality). To bound $|f(\omega) - f(\omega')|$ for this case, we note that for V_j^T denoting the j-th row of V, and $\eta = \exp(-\frac{2\pi i}{N})$ an N-th root of unity,

$$\begin{aligned}
&\|P_\ell V R_\omega \overline{F}_k\|_2^2 - \|P_\ell V R_{\omega'} \overline{F}_k\|_2^2 \\
&= \sum_{p,q=1}^{m} \langle \eta^{-k\omega_p} P_\ell V_p, \eta^{-k\omega_q} P_\ell V_q\rangle - \sum_{r,s=1}^{m} \langle \eta^{-k\omega'_r} P_\ell V_r, \eta^{-k\omega'_s} P_\ell V_s\rangle \\
&= \sum_{p,q=1}^{m} (\eta^{k(\omega_p-\omega_q)} - \eta^{k(\omega'_p-\omega'_q)})\langle P_\ell V_p, P_\ell V_q\rangle \\
&= (\eta^{k(\omega_1-\omega_1)} - \eta^{k(\omega'_1-\omega'_1)})\langle P_\ell V_1, P_\ell V_1\rangle + \sum_{q=2}^{m} (\eta^{k(\omega_1-\omega_q)} - \eta^{k(\omega'_1-\omega_q)})\langle P_\ell V_1, P_\ell V_q\rangle \\
&\quad + \sum_{p=2}^{m} (\eta^{k(\omega_p-\omega_1)} - \eta^{k(\omega_p-\omega'_1)})\langle P_\ell V_p, P_\ell V_1^*\rangle + \sum_{p,q=2}^{m} (\eta^{k(\omega_p-\omega_q)} - \eta^{k(\omega_p-\omega_q)})\langle P_\ell V_p, P_\ell V_q\rangle \\
&= \sum_{q=2}^{m} (\eta^{k\omega_1} - \eta^{k\omega'_1})\eta^{-k\omega_q}\langle P_\ell V_1, P_\ell V_q\rangle + \sum_{p=2}^{m} (\eta^{-k\omega_1} - \eta^{-k\omega'_1})\eta^{k\omega_p}\langle P_\ell V_p, P_\ell V_1\rangle.
\end{aligned}$$

Combining this with (23), we obtain

$$\begin{aligned}
f(\omega) - f(\omega') = \frac{1}{\ell N}\sum_{k=1}^{N}\big(|\hat{x}_k|^2 - |\hat{y}_k|^2\big)\Big(&\sum_{q=2}^{m} (\eta^{k\omega_1} - \eta^{k\omega'_1})\eta^{-k\omega_q}\langle P_\ell V_1, P_\ell V_q\rangle \\
&+ \sum_{p=2}^{m} (\eta^{-k\omega_1} - \eta^{-k\omega'_1})\eta^{k\omega_p}\langle P_\ell V_p, P_\ell V_1\rangle\Big)
\end{aligned} \tag{24}$$

Observe that the right hand side is a sum of four different rescaled Fourier coefficients of the vector $u \in \mathbb{R}^N$ given by $u_k := |\hat{x}_k|^2 - |\hat{y}_k|^2$, as for example

$$\frac{1}{\ell N}\sum_{p=2}^{m}\langle P_\ell V_1, P_\ell V_p\rangle \sum_{k=1}^{N}(|\hat{x}_k|^2 - |\hat{y}_k|^2)\eta^{k(\omega_p-\omega_1)} = \frac{1}{\ell N}\sum_{p=2}^{m}\langle P_\ell V_1, P_\ell V_p\rangle(\overline{F}u)_{\omega_p-\omega_1} = V_1^* P_\ell^* P_\ell V^* v,$$

where $v \in \mathbb{R}^m$ is given by $v_1 = 0$ and $v_p = (\overline{F}u)_{\omega_p-\omega_1}$ for $2 \le p \le m$. Note that as $\omega \in \Xi$ and hence the ω_q are all different, v is a projection of $\overline{F}u$ on a subset of its entries, and so $\|v\|_2 \le \sqrt{N}\|u\|_2$. Note that in this step, it is crucial to sample without replacement, as otherwise, the bound would no longer hold. Consequently, using the Cauchy-Schwartz inequality,

$$\begin{aligned}
\tfrac{1}{\ell N}\Big|\sum_{p=2}^{m}\langle P_\ell V_1, P_\ell V_p\rangle \sum_{k=1}^{N}\big(|\hat{x}_k|^2 - |\hat{y}_k|^2\big)\eta^{k(\omega_p-\omega_1)}\Big| &\le \tfrac{1}{\ell N}\|V\|_{2\to 2}^2\|P_\ell^* P_\ell V_1\|_2\|v\|_2 \\
&\le \tfrac{1}{\ell\sqrt{N}}\|\overline{F}u\|_2 \le \tfrac{1}{\ell}\|\overline{F}u\|_\infty = \tfrac{1}{\ell}\|x-y\|_\infty.
\end{aligned}$$

Identical bounds for the other three summands in (24) are attained in an analogous way, which yields the result for ω and ω' differing in only one component (with a constant of 4 rather than 12). If they differ in two components, replacing one of these components in both ω and ω' by an entry that appears in neither of them, yields $\omega'', \omega''' \in \Xi$, which differ only in the other one of these components. Thus applying the above bound three times yields the result. □

We are now ready to bound the distance $d(x,y) = \|x-y\|_{\psi_2}$.

Lemma 5. *For all $x, y \in \mathbb{R}^N$ it holds that*

$$d(x,y) \leq \frac{12\sqrt{m}}{\ell}\|x-y\|_\infty.$$

Proof. By (6), it suffices to show that for all $t \geq 0$

$$\mathbb{P}_\Omega(|Z_x - Z_y| > t) \leq \exp\Big(1 - t^2/\big(\frac{12\sqrt{m}}{\ell}\|x-y\|_\infty\big)^2\Big). \tag{25}$$

To prove this, we will apply Theorem 1 for $\mathcal{F}_k$, the σ-algebra generated by $\Omega_1, ..., \Omega_k$. For that, we need to bound the sum of squared ranges

$$R^2 = \sup \sum_{j=1}^m ran_j^2$$

where, for $(\Omega'_j, ..., \Omega'_m)$ an independent copy of $(\Omega_j, ..., \Omega_m)$ and $\Omega' = (\Omega_1, ...\Omega_{j-1}, \Omega'_j, ..., \Omega'_m)$,

$$\begin{aligned} ran_j &= \sup_{\Omega_j \notin \{\Omega_1,...,\Omega_{j-1}\}} \Big(\mathbb{E}(f(\Omega)|\Omega_j, ..., \Omega_1)\Big|\Omega_{j-1}, ..., \Omega_1\Big) + \sup_{\Omega_j \notin \{\Omega_1,...,\Omega_{j-1}\}} \Big(\mathbb{E}(-f(\Omega')|\Omega_j, ..., \Omega_1)\Big|\Omega_{j-1}, ..., \Omega_1\Big) \\ &= \sup_{\Omega_j, \Omega'_j \notin \{\Omega_1,...,\Omega_{j-1}\}} \Big(\mathbb{E}(f(\Omega)|\Omega_j, \Omega_{j-1}..., \Omega_1) + \mathbb{E}(-f(\Omega')|\Omega'_j, \Omega_{j-1}, ..., \Omega_1)\Big|\Omega_{j-1}, ..., \Omega_1\Big). \end{aligned} \tag{26}$$

For that, define the events $\mathcal{E}_0 = \{\Omega_j \neq \Omega'_k \ \forall j > k\}$, $\mathcal{E}'_0 = \{\Omega'_j \neq \Omega_k \ \forall j > k\}$, and, for $j \in [m-k]$, $\mathcal{E}_j = \{\Omega_{k+j} = \Omega'_k\}$, $\mathcal{E}'_j = \{\Omega'_{k+j} = \Omega_k\}$ and note that

$$\mathbb{P}[\cup_{j=0}^m \mathcal{E}_j | \Omega_1, ..., \Omega_k, \Omega'_k] = \mathbb{P}[\cup_{j=0}^m \mathcal{E}'_j | \Omega_1, ..., \Omega_k, \Omega'_k] = 1. \tag{27}$$

Now, we can write

$$\mathbb{E}[f(\Omega)|\Omega_1, ..., \Omega_{k-1}, \Omega_k] - \mathbb{E}[f(\Omega')|\Omega_1, ..., \Omega_{k-1}, \Omega'_k] = \sum_{j=0}^{m-k} \mathbb{E}[f(\Omega)\mathbb{1}_{\mathcal{E}_j} - f(\Omega')\mathbb{1}_{\mathcal{E}'_j}|\Omega_1, ..., \Omega_k, \Omega'_k]. \tag{28}$$

Given $\Omega_1, \ldots, \Omega_k$ and Ω'_k, consider random variables $\Omega''_{k+1}, \ldots \Omega''_m$ drawn subsequently without replacement from $[N] \setminus \{\Omega_1, \ldots, \Omega_k, \Omega'_k\}$ and set

$$\Omega'' = (\Omega_1, \ldots, \Omega_k, \Omega''_{k+1}, \ldots, \Omega''_m), \qquad \Omega''' = (\Omega_1, \ldots, \Omega_{k-1}, \Omega'_k, \Omega''_{k+1}, \ldots, \Omega''_m).$$

Observe that given the event $\mathcal{E}_0$, Ω and Ω'' are conditionally identically distributed, and the same holds for Ω' and Ω''' given the event $\mathcal{E}'_0$. So, using that $\mathcal{E}_0$ and Ω'' as well as $\mathcal{E}'_0$ and Ω''' are conditionally independent given $\Omega_1, \ldots, \Omega_k, \Omega'_k$, the summand in (28) corresponding to $j = 0$ becomes

$$\begin{aligned} &\mathbb{E}[f(\Omega)\mathbb{1}_{\mathcal{E}_0} - f(\Omega')\mathbb{1}_{\mathcal{E}'_0}|\Omega_1, ..., \Omega_k, \Omega'_k] \\ &\quad= \mathbb{E}[f(\Omega'')\mathbb{1}_{\mathcal{E}_0} - f(\Omega''')\mathbb{1}_{\mathcal{E}'_0}|\Omega_1, ..., \Omega_k, \Omega'_k] \\ &\quad= \mathbb{E}[f(\Omega'')|\Omega_1, ..., \Omega_k, \Omega'_k]\mathbb{P}[\mathcal{E}_0|\Omega_1, ..., \Omega_k, \Omega'_k] - \mathbb{E}[f(\Omega''')|\Omega_1, ..., \Omega_k, \Omega'_k]\mathbb{P}[\mathcal{E}'_0|\Omega_1, ..., \Omega_k, \Omega'_k] \\ &\quad= \big(\mathbb{E}[f(\Omega'') - f(\Omega''')|\Omega_1, ..., \Omega_k, \Omega'_k]\big)\mathbb{P}[\mathcal{E}_0|\Omega_1, ..., \Omega_k, \Omega'_k], \\ &\quad\leq \frac{12}{\ell}\|x-y\|_\infty \mathbb{P}[\mathcal{E}_0|\Omega_1, ..., \Omega_k, \Omega'_k] \end{aligned} \tag{29}$$

where the third equality uses that $(\Omega'_k, \ldots, \Omega'_m)$ is an independent copy of $(\Omega_k, \ldots, \Omega_m)$ and so the two probabilities in (29) are equal. The last inequality holds almost surely and follows from Lemma 4.

To bound the summand in (28) for $j > 0$, we proceed in a similar way. Given $\Omega_1, \ldots, \Omega_k$ and Ω'_k, consider random variables $\Omega''_{k+1}, \ldots, \Omega''_{k+j-1}, \Omega''_{k+j+1}, \ldots \Omega''_m$ drawn subsequently without replacement from $[N] \setminus \{\Omega_1, \ldots, \Omega_k, \Omega'_k\}$ and set

$$\begin{aligned} \Omega'' &= (\Omega_1, \ldots, \Omega_k, \Omega''_{k+1}, \ldots, \Omega''_{k+j-1}, \Omega'_k, \Omega''_{k+j-1}, \ldots, \Omega''_m), \\ \Omega''' &= (\Omega_1, \ldots, \Omega'_k, \Omega''_{k+1}, \ldots, \Omega''_{k+j-1}, \Omega_k, \Omega''_{k+j-1}, \ldots, \Omega''_m). \end{aligned}$$

As before, observe that given the event $\mathcal{E}_j$, Ω and Ω'' are conditionally identically distributed, and the same holds for Ω' and Ω''' given the event $\mathcal{E}'_j$. The remainder of the estimate proceeds exactly as for $j = 0$, with the slight difference that Ω'' and Ω''' now differ in two entries, but nevertheless Lemma 4 still applies. Thus we obtain

$$\mathbb{E}[f(\Omega)\mathbb{1}_{\mathcal{E}_j} - f(\Omega')\mathbb{1}_{\mathcal{E}'_j}|\Omega_1, ..., \Omega_k, \Omega'_k] \leq \frac{12}{\ell}\|x-y\|_\infty \mathbb{P}[\mathcal{E}_j|\Omega_1, ..., \Omega_k, \Omega'_k].$$

Consequently, one has almost surely

$$\begin{aligned}\mathbb{E}[f(\Omega)|\Omega_1, ..., \Omega_{k-1}, \Omega_k] - \mathbb{E}[f(\Omega')|\Omega_1, ..., \Omega_{k-1}, \Omega'_k] &\leq \sum_{j=0}^{m-k} \frac{12}{\ell}\|x-y\|_\infty \mathbb{P}[\mathcal{E}_j|\Omega_1, ..., \Omega_k, \Omega'_k] \\ &= \frac{12}{\ell}\|x-y\|_\infty,\end{aligned}$$

where the last equality follows from (27), and hence, by (26), $ran_j \leq \frac{12}{\ell}\|x-y\|_\infty$ and $R^2 \leq \left(\frac{12\sqrt{m}}{\ell}\|x-y\|_\infty\right)^2$.

With this bound, Theorem 1 can be applied. One obtains

$$\mathbb{P}(|Z_x - Z_y| > t) \leq 2\exp(-t^2/(\frac{12\sqrt{m}}{\ell}\|x-y\|_\infty)^2),$$

which implies (25). We conclude

$$d(x,y) := \|Z_x - Z_y\|_{\Psi_2} \leq \frac{12\sqrt{m}}{\ell}\|x-y\|_\infty,$$

as desired. □

Acknowledgements

The three authors acknowledge support by the Hausdorff Institute for Mathematics (HIM), where part of this work was completed in the context of the HIM Trimester Program "Mathematics of Signal Processing", and support by the German Science Foundation in the context of the Emmy Noether Junior Research Group "Randomized Sensing and Quantization of Signals and Images" (KR 4512/1-1). In addition, JF and FK acknowledge support by the German Science Foundation in the context of the Research Training Group 1023 "Identification in Mathematical Models". RS acknowledges support by a Hellman Fellowship and the NSF under DMS-1517204.

References

[1] J. Benedetto, A. Powell, and Ö. Yılmaz. Sigma-delta ($\Sigma\Delta$) quantization and finite frames. *IEEE Trans. Inf. Theory*, 52(5):1990–2005, 2006.

[2] J. Blum, M. Lammers, A.M. Powell, and Ö. Yılmaz. Sobolev duals in frame theory and sigma-delta quantization. *J. Fourier Anal. and Appl.*, 16(3):365–381, 2010.

[3] T. Blumensath and M. Davies. Iterative hard thresholding for compressed sensing. *Appl. Comp. Harmon. Anal.*, 27(3):265–274, 2009.

[4] B. Bodmann, V. Paulsen, and S. Abdulbaki. Smooth frame-path termination for higher order sigma-delta quantization. *J. Fourier Anal. and Appl.*, 13(3):285–307, 2007.

[5] P. Boufounos and R. Baraniuk. Quantization of sparse representations. In *Rice University ECE Department Technical Report 0701. Summary appears in Proc. Data Compression Conference (DCC)*, Snowbird, UT, March 27-29 2007.

[6] P. Boufounos and R. Baraniuk. 1-bit compressive sensing. In *Conf. Inform. Sci. Syst.*, pages 16–21. IEEE, 2008.

[7] R. Calderbank and I. Daubechies. The pros and cons of democracy. *IEEE Trans. Inform. Theory*, 48(6):1721–1725, 2002. Special issue on Shannon theory: perspective, trends, and applications.

[8] E. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory*, 52(2):489–509, 2006.

[9] E. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.*, 59:1207–1223, 2006.

[10] W. Dai and O. Milenkovic. Subspace pursuit for compressive sensing signal reconstruction. *IEEE Trans. Inform. Theory*, 55(5):2230–2249, 2009.

[11] I. Daubechies and R. DeVore. Approximating a bandlimited function using very coarsely quantized data: a family of stable sigma-delta modulators of arbitrary order. *Ann. Math.*, 158(2):679–710, 2003.

[12] I. Daubechies and R. Saab. A deterministic analysis of decimation for sigma-delta quantization of bandlimited functions. *IEEE Signal Proc. Lett.*, 22(11):2093–2096, 2015.

[13] P. Deift, S. Güntürk, and F. Krahmer. An optimal family of exponentially accurate one-bit sigma-delta quantization schemes. *Comm. Pure Appl. Math.*, 64(7):883–919, 2011.

[14] D. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52(4):1289–1306, 2006.

[15] R. Dudley. The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *J. Funct. Anal.*, 1:290–330, 1967.

[16] J. Feng and F. Krahmer. An RIP-based approach to $\Sigma\Delta$ quantization for compressed sensing. *IEEE Signal Process. Lett*, 21(11):1351–1355, 2014.

[17] S. Foucart and H. Rauhut. A mathematical introduction to compressive sensing. *Basel: Birkhäuser, Boston*, 1(3), 2013.

[18] V. Goyal, M. Vetterli, and N. Thao. Quantized overcomplete expansions in $\mathbb{R}^N$: analysis, synthesis, and algorithms. *IEEE Trans. Inform. Theory*, 44(1):16–31, Jan 1998.

[19] S. Güntürk. One-bit sigma-delta quantization with exponential accuracy. *Comm. Pure Appl. Math.*, 56(11):1608–1630, 2003.

[20] S. Güntürk, M. Lammers, A. Powell, R. Saab, and Ö Yılmaz. Sobolev duals for random frames and $\Sigma\Delta$ quantization of compressed sensing measurements. *Found. Comp. Math.*, 13(1):1–36, 2013.

[21] J. Haupt, W. Bajwa, G. Raz, and R. Nowak. Toeplitz compressed sensing matrices with applications to sparse channel estimation. *IEEE Trans. Inform. Theory*, 56(11):5862–5875, 2010.

[22] H. Inose and Y. Yasuda. A unity bit coding method by negative feedback. *Proc. IEEE*, 51(11):1524–1535, 1963.

[23] L. Jacques, D. Hammond, and J. Fadili. Dequantizing compressed sensing: When oversampling and non-gaussian constraints combine. *IEEE Trans. Inform. Theory*, 57(1):559–571, 2011.

[24] F. Krahmer, S. Mendelson, and H. Rauhut. Suprema of chaos processes and the restricted isometry property. *Comm. Pure Appl. Math.*, 67(11):1877–1904, 2014.

[25] F. Krahmer, R. Saab, and R. Ward. Root-exponential accuracy for coarse quantization of finite frame expansions. *IEEE Trans. Inform. Theory*, 58(2):1069 –1079, February 2012.

[26] F. Krahmer, R. Saab, and Ö Yılmaz. Sigma-delta quantization of sub-Gaussian frame expansions and its application to compressed sensing. *Inform. Inf.*, 3(1):40–58, 2014.

[27] F. Krahmer and R. Ward. Lower bounds for the error decay incurred by coarse quantization schemes. *Appl. Comput. Harmonic Anal.*, 32(1):131–138, 2012.

[28] J. Laska, P. Boufounos, M. Davenport, and R. Baraniuk. Democracy in action: Quantization, saturation, and compressive sensing. *Appl. Comp. Harmon. Anal.*, 31(3):429–443, 2011.

[29] C. McDiarmid. Concentration. In *Probabilistic methods for algorithmic discrete mathematics*, pages 195–248. Springer, 1998.

[30] S. Mendelson, H. Rauhut, and R. Ward. Improved bounds for sparse recovery from subsampled random convolutions. preprint, arXiv:1610.04983, 2016.

[31] D. Needell and J. Tropp. Cosamp: Iterative signal recovery from incomplete and inaccurate samples. *Appl. Comp. Harmon. Anal.*, 26(3):301–321, 2009.

[32] S. R. Norsworthy, R. Schreier, and G. C. Temes, editors. *Delta-Sigma-Converters: Theory, Design and Simulation*. Wiley-IEEE, 1996.

[33] Y. Plan and R. Vershynin. One-bit compressed sensing by linear programming. *Comm. Pure Appl. Math.*, 2013.

[34] H. Rauhut, J. Romberg, and J.A. Tropp. Restricted isometries for partial random circulant matrices. *Appl. Comp. Harmon. Anal.*, 32(2):242–254, 2012.

[35] J. Romberg. Compressive sensing by random convolution. *SIAM J. Imaging Sci.*, 2(4):1098–1128, 2009.

[36] R. Saab, R. Wang, and Ö. Yılmaz. From compressed sensing to compressed bit-streams: practical encoders, tractable decoders. preprint, arXiv:1604.00700, 2016.

[37] R. Saab, R. Wang, and Ö. Yılmaz. Quantization of compressive samples with stable and robust recovery. *Appl. Comp. Harmon. Anal.*, 2016.

[38] J.Z. Sun and V.K. Goyal. Optimal quantization of random measurements in compressed sensing. In *Intl. Symp. Inform. Theory*, pages 6–10. IEEE, 2009.

[39] M. Talagrand. *Upper and lower bounds for stochastic processes: modern methods and classical problems*, volume 60. Springer Science & Business Media, 2014.

[40] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Y.C. Eldar and G. Kutyniok, editors, *Compressed Sensing: Theory and Applications*, pages xii+544. Cambridge Univ Press, Cambridge, 2012.

[41] R. Wang. Sigma delta quantization with harmonic frames and partial fourier ensembles. preprint, arXiv:1511.05671, 2015.

[42] A. Zymnis, S. Boyd, and E. Candés. Compressed sensing with quantized measurements. *IEEE Signal Proc. Lett.*, 17(2):149–152, 2010.

An RIP-based approach to $\Sigma\Delta$ quantization for compressed sensing

Joe-Mei Feng and Felix Krahmer

May 1, 2014

Abstract

In this paper, we provide a new approach to estimating the error of reconstruction from $\Sigma\Delta$ quantized compressed sensing measurements. Our method is based on the restricted isometry property (RIP) of a certain projection of the measurement matrix. Our result yields simple proofs and a slight generalization of the best-known reconstruction error bounds for Gaussian and subgaussian measurement matrices.

1 Introduction

1.1 Compressed sensing

Compressed sensing has drawn significant attention since the seminal works by Candès, Romberg, Tao [8], and Donoho [14]. The theory of compressed sensing is based on the observation that various cases of natural signals are approximately sparse with respect to certain bases or frames. The basic idea is to recover such signals from a small number of linear measurements. Hence the problem turns into an underdetermined linear system. Various criteria have been proposed to determine whether such a system has a unique sparse solution. In this paper we will work with the restricted isometry property (RIP) as introduced by Candès et al. [9] in the context of recovery guarantees for ℓ_1 minimization.

Definition 1. A matrix $A \in \mathbb{R}^{m\times N}$ has the restricted isometry property (RIP) of order s if there exists $0 < \delta < 1$ such that for all s-sparse vectors $x \in \mathbb{R}^N$, i.e., vectors that have at most s non-zero components, one has

$$(1-\delta)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1+\delta)\|x\|_2^2.$$

The smallest such δ is called the restricted isometry constant of order s and is denoted by δ_s.

There have been a number of works on recovery guarantees for compressed sensing with RIP measurement matrices. Recovery can be guaranteed for various algorithms. For the original context of ℓ_1 minimization, the most recent

results require the measurement matrix to have a restricted isometry constant of $\delta_{2s} < \frac{1}{\sqrt{2}}$ [7], which is known to be optimal [11].

Finding the restricted isometry constant of a measurement matrix is, in general, an NP hard problem [31]. On the other hand, deterministic matrix constructions with guaranteed RIP are only known for relatively large embedding dimensions (see for example [13]). That is why many papers on the subject work with random matrices.

Examples of random matrices known to have the RIP for large enough embedding dimension with high probability include subgaussian, partial random circulant [23], and partial random Fourier matrices [29]. A subgaussian matrix has independent random entries whose tails are dominated by a Gaussian random variable (cf. Definition 2). Such matrices have been shown to have the RIP provided $m = \Omega(s \log(eN/s))$, see for example [2]. This order of the embedding dimension m is known to be optimal [15]. Examples of subgaussian matrices include Gaussian and Bernoulli matrices.

1.2 Quantization

To allow for digital transmission and storage of compressed sensing measurements, one needs to quantize these measurements. That is, the measurements need to be represented by finitely many symbols from a finite alphabet. In this paper, we only consider alphabets consisting of equispaced real numbers. The extreme case of considering the set of only the two elements $\{-1, 1\}$ is also called 1-bit quantization.

The most intuitive method to quantize the measurements is to map each of them to the closest element from the alphabet. Since this method processes the quantization independently for each measurement, it is also called memoryless scalar quantization (MSQ).

Most of the literature on MSQ compressed sensing up to date considers 1-bit quantization [6, 22, 28, 1], which amounts to considering only the measurement signs. Jacques et al. [22] showed that for Gaussian measurements or measurements drawn uniformly from the unit sphere, a reconstruction error of $O(\frac{s}{m} \log \frac{mN}{s})$ is feasible. However, they did not provide an efficient algorithm that guarantees this accuracy. Later, for Gaussian measurements, Gupta et al. [18] demonstrated that one may tractably recover the support of a signal from $O(s \log N)$ measurements. Plan et al. [28] showed that one can, again for Gaussian measurements, reconstruct the direction of an s-sparse signal via convex optimization, with accuracy $O((\frac{s}{m})^{\frac{1}{5}})$ up to logarithmic factors with high probability. Ai et al. [1] derived similar results for subgaussian measurements under additional assumptions on the size of the signal entries.

On the other hand, in [22] it was shown that the ℓ_2 reconstruction error can never be better than $\Omega(\frac{s}{m})$. To break this bottleneck of MSQ, $\Sigma\Delta$ quantization for compressed sensing has drawn attention recently. $\Sigma\Delta$ quantizes a vector as a whole rather than the components individually, i.e., the quantized values depend on previous quantization steps.

$\Sigma\Delta$ quantization was originally introduced as an efficient quantizer for redundant representation of oversampled band-limited functions [21]. Later on, a rigorous mathematical error analysis was provided by [10] and many follow-up papers. The best known error decay rates are exponential in the oversampling rate, as derived in [16, 12]. This is known to be optimal: in [26], corresponding lower bounds are derived, which also show that the achievable accuracy must depend on the signal amplitude.

In [3], $\Sigma\Delta$ has been extended to frame expansions; this will be also the viewpoint taken in this paper. The first works on $\Sigma\Delta$ schemes for frame quantization, such as [3, 5], required frame constructions with particular smoothness properties to yield reconstruction guarantees. In [27], the authors observed that what is needed is in fact a requirement on the dual frame used for reconstruction rather than the frame itself. Reconstruction guarantees can hence be improved by choosing the dual frame used for reconstruction appropriately. Optimizing the dual frame in this respect led to the definition of Sobolev dual frames [4], cf. Section 2.2.2 below. Combined with the exponential error bounds derived for the corresponding $\Sigma\Delta$ schemes for bandlimited functions [16, 12], Sobolev dual reconstructions yield root-exponential error decay in the oversampling rate. This constitutes the best known accuracy guarantees for coarse frame quantization, both for harmonic frames and Sobolev self-dual frames [24] and for subgaussian random frames [25].

Sobolev dual reconstructions have also been crucial for being able to apply $\Sigma\Delta$ quantization to compressed sensing measurements. Güntürk et al. [17] proved the first recovery guarantees for this setup, showing that for rth order $\Sigma\Delta$ quantization applied to Gaussian compressed sensing measurements, the ℓ_2 reconstruction error is of order $O((\frac{s}{m})^{\alpha(r-\frac{1}{2})})$ with high probability. Here $\alpha \in (0,1)$ is a parameter and the required measurements grows with α, tending to infinity as $\alpha \to 1$. Indeed for r large enough this breaks the MSQ bottleneck. More recently, in [25], this result has been generalized to subgaussian measurements.

1.3 Contributions

The main contribution of this paper is that the restricted isometry property (RIP) is applied to estimate the error bound for $\Sigma\Delta$ quantized compressed sensing. That is, once we know the restricted isometry constant of a modification of the measurement matrix, we can estimate the reconstruction error.

In the following results, we assume that the $\Sigma\Delta$ quantized measurements with quantization alphabet $\mathcal{Z} = \Delta\mathbb{Z}$, $\Delta > 0$, are given. We refer the readers to Section 2.2 for details on the quantization scheme employed. A special role is played by the rth power of the inverse of the finite difference matrix D as introduced in (2) below; denoting the singular value decomposition of D^{-r} by $D^{-r} = U_{D^{-r}} S_{D^{-r}} V^*_{D^{-r}}$, we obtain our main theorem given as follows.

Theorem 1. *Suppose one is given a measurement matrix $\Phi \in \mathbb{R}^{m\times N}$ such that both Φ and $\sqrt{\frac{1}{\ell}} P_\ell V^*_{D^{-r}} \Phi$, $\ell \leq m$ have the restricted isometry constant $\delta_{2s} < \frac{1}{\sqrt{2}}$,*

where P_ℓ maps a vector to its first ℓ components.

Then for an s-sparse signal $x \in \mathbb{R}^N$ satisfying $\min_j |x_j| \geq K2^{r-\frac{1}{2}}\Delta$, for some positive constant K, denote by q the rth order $\Sigma\Delta$ quantized measurements of Φx with step size Δ. Furthermore, denote by T the support set recovered from Φx via ℓ_1 minimization and choose $L_{sob,r}$ to be the Sobolev dual matrix of Φ_T (see Section 2.2.2 for details). Then reconstructing the signal via $\hat{x}_T = L_{sob,r}q$ yields a reconstruction error bounded by

$$\|x - \hat{x}\|_2 \leq C\Delta(\frac{m}{\ell})^{-r+\frac{1}{2}},$$

where $C > 0$ is a constant depending only on r.

Note from Theorem 1 that smaller values of ℓ yield better error bounds. However, ℓ has to be large enough such that $\frac{1}{\sqrt{\ell}}(P_\ell V_{D^{-r}}^* \Phi)$ has the restricted isometry constant $\delta_{2s} \leq \frac{1}{\sqrt{2}}$.

This result can be applied to obtain recovery guarantees for Gaussian and subgaussian measurements (in the sense of Definition 2 below). The resulting bounds for the first two cases agrees with those derived in [17] and [25], as summarized in the following Theorem.

Theorem 2 ([17, 25])**.** *Let Φ be an $m \times N$ matrix whose entries are independent, mean zero, unit variance ρ-subgaussian random variables and suppose that $\lambda := m/k \geq (C\log(eN/k))^{\frac{1}{1-\alpha}}$ where $\alpha \in (0,1)$. With high probability the rth order $\Sigma\Delta$ reconstruction $\hat{x}$ satisfies*

$$\|x - \hat{x}\|_2 \leq C'\lambda^{-\alpha(r-1/2)}\delta,$$

for all $x \in \Sigma_k^N$ for which $\min_{j\in \mathrm{supp}(x)} |x_j| > K'\Delta$. Again, Δ is the step size of the $\Sigma\Delta$ quantization alphabet and C, C', K' are appropriate constants that depend only on r and ρ.

1.4 Organization

The paper is organized as follows. We first introduce in Section 2 some background and previous results on $\Sigma\Delta$ quantization, suprema of chaos processes, and the partial random circulant matrices. In Section 3 we present our main result showing how the RIP is used to estimate the reconstruction error for quantized compressed sensing. In Section 4, we explain how our result recovers the best-known bounds for Gausssian and subgaussian measurement matrices using a simple argument, in Section 5 we slightly generalize these bounds. We conclude in Section 6.

2 Background and previous results

2.1 Notation

Throughout this paper, we use the following notation. The set $D_{s,N} = \{x \in \mathbb{R} | \|x\|_2 \leq 1, \|x\|_0 \leq s\}$ is the set of unit norm s-sparse vectors. The ℓ_0-norm $\|\cdot\|_0$

counts the number of non-zero components of a vector. Given a signal x, the support set of x, in short, supp x, is the index set of the non-zero components. The ℓ_2-operator norm is denoted by $\|A\|_{2\to 2} = \sup_{\|x\|_2=1} \|Ax\|_2$. For a matrix A, $\sigma_i(A)$ and $\sigma_{\min}(A)$ denote the ith largest and the smallest singular value, respectively. Furthermore we write $\gtrsim$ and $\lesssim$ to denote $\geq$ or $\leq$ up to a positive multiplicative constant. The Moore-Penrose pseudoinverse of a matrix A is denoted by $A^\dagger = (A^*A)^{-1}A^*$.

We will mainly study subgaussian random matrices, that is, matrices with independent subgaussian entries in the sense of the following definition.

Definition 2. A random variable X is called ρ-subgaussian if $\mathbb{P}(|X| \geq t) \leq 2\exp(-t^2/2\rho^2)$.

2.2 $\Sigma\Delta$ Quantization

In this paper, we exclusively focus on quantization alphabets $\mathcal{Z}$ such that $\mathcal{Z} = \Delta\mathbb{Z}$, for some $\Delta > 0$. Note that while this is an infinite set, one can show that in fact only a finite range of values are assumed [17, 24]. Hence this setup is in line with requiring a finite alphabet. The idea of rth order $\Sigma\Delta$ quantization is to quantize each component of a vector taking the previous r quantization steps into account. More explicitly, a greedy rth order $\Sigma\Delta$ quantization scheme maps a sequence of inputs (y_j) to elements $q_i \in \mathcal{Z}$ via an internal state variable u_i chosen to satisfy the recurrence relation

$$(\Delta^r u)_i := \sum_{j=0}^{r} \binom{r}{j} (-1)^j u_{i-j} = y_i - q_i, \tag{1}$$

where q_i is chosen such that $|u_i|$ is minimized (Note that only in this equation, Δ denotes the finite difference operator, whereas all other occurences in this paper refer to the quantization step size).

With the initial condition $(u_i)_{i=0}^{-\infty} = 0$, Equation (1) can be expressed as

$$D^r u = y - q,$$

where the finite difference matrix $D \in \mathbb{R}^{m\times m}$ is given by

$$D_{ij} \equiv \begin{cases} 1 & , \text{ if } i = j, \\ -1 & , \text{ if } i = j+1, \\ 0 & , \text{ otherwise.} \end{cases} \tag{2}$$

2.2.1 Support set recovery

Given an s-sparse signal x, and an $m\times N$ measurement matrix Φ, where $m \ll N$, we acquire measurements $y = \Phi x$. Applying an rth order $\Sigma\Delta$ quantization scheme to y, we obtain q. Treating q as perturbed measurements, i.e., $q = y + e = \Phi x + e$, one can determine the support set. This is a consequence of the following observation, which is a modified version of Proposition 4.1 in [17] combined with the reconstruction guarantees in [7].

Proposition 1. *Given $\epsilon > 0$ as well as $x \in \mathbb{R}^N$ an s-sparse signal with $\operatorname{supp} x = T$ and $\min_{j\in T} |x_j| \geq K\frac{\epsilon}{\sqrt{m}}$. Here K is an absolute constant. Let $\Phi \in \mathbb{R}^{N\times m}$ be a measurement matrix such that $\frac{1}{\sqrt{m}}\Phi$ has the RIP with $\delta_{2s} < \frac{1}{\sqrt{2}}$. Denote by $e \in \mathbb{R}^m$ a noise vector with $\|e\|_2 \leq \epsilon$, and let x' be the signal reconstructed from the noisy measurements $q = \Phi x + e$ via ℓ_1 minimization, i.e.,*

$$x' = \arg\min \|z\|_1 \text{ subject to } \|\Phi z - q\|_2 \leq \epsilon.$$

Then the index set of largest s components of x' is T, that is, the support set of x is correctly recovered.

Note that in this result, the measurement matrix Φ is not normalized, while in the compressed sensing literature, it is common to normalize the measurement matrix such that it has unit-norm columns. This is because for normalized matrix columns, each measurement will be of order $\frac{1}{\sqrt{m}}$, so quantizing it with a fixed step size Δ will lead to worse and worse resolution. To allow for a fair comparison when m grows, the measurements should rather be chosen independently of m. Therefore, in this paper as well as in [17] the measurement matrices are not normalized, each entry of the measurement matrices is chosen to have variance one.

To apply Proposition 1 to greedy $\Sigma\Delta$ quantization, one sets $e = q - y$, where q is the quantized measurement vector. Elementary estimates (cf. [17]) yield that $\|q - y\|_2 \leq 2^{r-1}\Delta\sqrt{m}$. Thus one obtains that ℓ_1 minimization recovers the correct support set provided that $\frac{1}{\sqrt{m}}\Phi$ has restricted isometry constant $\delta_{2s} < \frac{1}{\sqrt{2}}$ and $\min_j |x_j| \geq K2^{r-\frac{1}{2}}\Delta$.

2.2.2 Estimating the error and the Sobolev dual

When the support set T has been identified, we solve for x using some left inverse of Φ_T, say L. Then the reconstruction ℓ_2-error is given by

$$\begin{aligned}\|x - \hat{x}\|_2 &= \|Ly - Lq\|_2 = \|L(y - q)\|_2 \\ &= \|L(D^r u)\|_2 \leq \|LD^r\|_{2\to 2}\|u\|_2.\end{aligned}$$

The Sobolev dual matrix $L_{sob,r}$, first introduced in [4], is a left inverse of Φ_T defined to minimize $\|LD^r\|_{2\to 2}$, i.e.,

$$L_{sob,r} = \arg\min_L \|LD^r\|_{2\to 2} \quad \text{subject to } L\Phi_T = I.$$

The geometric intuition is that this dual frame is smoothly varying.

As in [17], the explicit formula $L_{sob,r}D^r = (D^{-r}\Phi_T)^\dagger$ yields the error bound

$$\begin{aligned}\|x - \hat{x}\|_2 &\leq \|(D^{-r}\Phi_T)^\dagger\|_{2\to 2}\|u\|_2 \\ &= \frac{1}{\sigma_{min}(D^{-r}\Phi_T)}\|u\|_2 \leq \frac{\Delta\sqrt{m}}{2\sigma_{min}(D^{-r}\Phi_T)},\end{aligned} \tag{3}$$

where the last inequality is derived in [17].

A key ingredient to bounding $\sigma_{\min}(D^{-r}\Phi_T)$ is the following result from the study of Toeplitz matrices, which depends heavily on Weyl's inequality [20] (see for example [17]).

Proposition 2. *Let r be any positive integer and D be as in (2). There are positive constants $c_{s_1}(r)$ and $c_{s_2}(r)$, independent of m, such that*

$$c_{s_1}(r)(\frac{m}{j})^r \leq \sigma_j(D^{-r}) \leq c_{s_2}(r)(\frac{m}{j})^r,\ j = 1, \ldots, m.$$

3 RIP-based error analysis

In this section we will give the quantized compressed sensing problem a mathematical model, and explain how we approach the reconstruction error via the RIP. In the next two sections we show its applications. From Section 2.2.2, the main issue to estimate the reconstruction error is to estimate $\sigma_{min}(D^{-r}\Phi_T)$. Finding the supremum of this expression over all potential support sets T can be interpreted as finding the supremum of the smallest image under $D^{-r}\Phi$ over all unit norm s-sparse vectors. This motivates the connection to the RIP.

In the following proof we show how the RIP can be applied to find this effective smallest singular value.

Proof of Theorem 1. As the assumptions of the theorem are stronger than those of Proposition 1, we conclude that the support is correctly recovered. Based on this observagtion, we now show the error bound. Recall that $D^{-r} = U_{D^{-r}} S_{D^{-r}} V^*_{D^{-r}}$. Then, as S is a diagonal matrix,

$$\begin{aligned}
\sigma_{\min}(D^{-r}\Phi_T) &= \sigma_{\min}(S_{D^{-r}} V^*_{D^{-r}}\Phi_T) \\
&\geq \sigma_{\min}(P_\ell S_{D^{-r}} V^*_{D^{-r}}\Phi_T) \\
&= \sigma_{\min}((P_\ell S_{D^{-r}} P^*_\ell)(P_\ell V^*_{D^{-r}}\Phi_T)) \\
&\geq s_\ell \sigma_{\min}(P_\ell V^*_{D^{-r}}\Phi_T) \\
&\gtrsim (\frac{m}{\ell})^r \sigma_{\min}(P_\ell V^*_{D^{-r}}\Phi_T), \qquad (4)
\end{aligned}$$

where the final inequality follows from Proposition 2.

Thus we need to bound $\sigma_{\min}(P_\ell V^*_{D^{-r}}\Phi_T)$ uniformly over all possible support sets T. Indeed by the RIP assumption for $\frac{1}{\sqrt{\ell}} P_\ell V^*_{D^{-r}}\Phi$, we obtain that $\sigma_{\min}(P_\ell V^*_{D^{-r}}\Phi_T)$ is uniformly bounded from below by

$$\sqrt{\ell}\sqrt{1 - \tfrac{1}{\sqrt{2}}}. \qquad (5)$$

The theorem follows by combining (3), (4), and (5).

□

4 Gaussian and subgaussian matrices

To illustrate the simplicity of our method, we first present a proof of Theorem 2 for standard Gaussian matrices, i.e. matrices with independent entries $\Phi_{i,j} \sim \mathcal{N}(0,1)$.

Proof of Theorem 2 for Gaussian matrices. Set $\ell := m(\frac{s}{m})^{\alpha}$. As the second factor is less than 1, one always has $1 \leq \ell \leq m$. Since Φ is a standard Gaussian random matrix, due to rotation invariance $\widetilde{\Phi} := (P_\ell V_{D^{-r}}^* \Phi)$ is also a standard Gaussian random matrix. The assumption on λ implies that its embedding dimension satisfies $\ell \geq Cs\log(\frac{eN}{k})$, so standard results (see, e.g., [2]) yield that for C and C' large enough, both $\frac{1}{\sqrt{\ell}}\widetilde{\Phi}$ and $\frac{1}{\sqrt{m}}\Phi$ have the RIP with constant $\delta_{2s} \leq \frac{1}{\sqrt{2}}$. Applying Theorem 1 (choose $K' = K2^{r-\frac{1}{2}}$), we obtain

$$\|x - \hat{x}\|_2 \lesssim \Delta(\frac{m}{s})^{-\alpha(r-\frac{1}{2})},$$

again with high probability, as desired. □

Sketch of proof of Theorem 2 for subgaussian matrices:
The proof proceeds long the same lines as for Gaussian matrices, except that one cannot use the rotation invariance. To bound the RIP constant of $\frac{1}{\sqrt{\ell}}P_\ell V_{D^{-r}}^* \Phi$, we note that for any $x \in D_{s,N}$, $\|P_\ell V_{D^{-r}}^* \Phi x\|_2^2$ is a quadratic form in the "vectorization" of Φ. Hence its tail decay can be estimated via the Hanson-Wright inequality [19, 30]. The RIP then follows via a union bound over an ϵ-net of $D_{s,N}$. This approach is related to certain steps in the original proof in [25].

Remark 1. Note that a complete proof for $\Sigma\Delta$ recovery guarantees needs both support set recovery and fine recovery (cf. [17, 25]) and in this paper we omitted the details of the former. We argue, however, that this coarse recovery step is straightforwardly based on standard compressed sensing results. Hence the core of our error estimate is really just captured in a few lines.

5 Generalization

In contrast to the techniques presented in [25], our method generalizes to certain random matrices with independent subgaussian columns, but no entrywise independence. As an additional criterion, one needs a type of small ball condition for $P_\ell V_{D^{-r}}^*$ applied to one of the random columns of Φ, (which denoted by Φ_j in the following). That is, one needs to exclude that $\|P_\ell V_{D^{-r}} \Phi_j\|_2$ is small with too large probability. If such a condition holds, the necessary RIP bound follows from a modified version of the RIP bound for matrices with independent subgaussian columns [32]. While we do not consider this to be an important generalization (which is why we refrain from presenting the details), we still believe it shows that our method is stronger than previous approaches, so we see the potential to apply it to more relevant, structured measurement scenarios such as partial random Fourier matrices, partial random circulant matrices, etc.

6 Conclusion

In this work we provided a new technique for bounding the reconstruction error arising in $\Sigma\Delta$ quantization for compressed sensing. In addition to greatly simplifying the proofs for the best known recovery guarantees, the new viewpoint hopefully opens the possibility to study broader classes of measurement matries.

Acknowledgment

The authors would like to thank Rayan Saab for providing helpful comments. This work was supported by the DFG Research Training Group 1023.

References

[1] A. Ai, A. Lapanowski, Y. Plan, and R. Vershynin. One-bit compressed sensing with non-Gaussian measurements. *Lin. Alg. Appl.*, 441:222–239, 2014.

[2] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constr. Approx.*, 28(3):253–263, 2008.

[3] J. J. Benedetto, A. M. Powell, and Ö. Yılmaz. Sigma-Delta ($\Sigma\Delta$) quantization and finite frames. *IEEE Trans. Inform. Theory*, 52(5):1990–2005, 2006.

[4] J. Blum, M. Lammers, A. M. Powell, and Ö. Yılmaz. Sobolev duals in frame theory and Sigma-Delta quantization. *J. Fourier Anal. Appl.*, 16(3):365–381, 2010.

[5] B.G. Bodmann, V.I. Paulsen, and S.A. Abdulbaki. Smooth frame-path termination for higher order sigma-delta quantization. *J. Fourier Anal. Appl.*, 13(3):285–307, 2007.

[6] P. T. Boufounos and R. G. Baraniuk. 1-bit compressive sensing. In *Proc. 42nd Annu. Conf. Inf. Sci. Syst. (CISS) , Princeton, NJ*, pages 16–21. IEEE, Mar. 2008.

[7] T. Cai and A. Zhang. Sparse representation of a polytope and recovery of sparse signals and low-rank matrices. *IEEE Trans. Inform. Theory*, to appear.

[8] E. J. Candès, J. K. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.*, 59(8):1207–1223, 2006.

[9] E. J. Candès and T. Tao. Decoding by linear programming. *IEEE Trans. Inform. Theory*, 51(12):4203–4215, 2005.

[10] I. Daubechies and R. DeVore. Approximating a bandlimited function using very coarsely quantized data: A family of stable sigma-delta modulators of arbitrary order. *Ann. Math.*, 158(2):679–710, 2003.

[11] M. E. Davies and R. Gribonval. Restricted isometry constants where ℓ^p sparse recovery can fail for $0 < p \leq 1$. *IEEE Trans. Inform. Theory*, 55(5):2203–2214, 2009.

[12] P. Deift, F. Krahmer, and C. S. Güntürk. An optimal family of exponentially accurate one-bit sigma-delta quantization schemes. *Comm. Pure Appl. Math.*, 64(7):883–919, 2011.

[13] R. DeVore. Deterministic constructions of compressed sensing matrices. *J. Complexity*, 23:918–925, 2007.

[14] D. L. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52(4):1289–1306, 2006.

[15] S. Foucart, A. Pajor, H. Rauhut, and T. Ullrich. The gelfand widths of ℓ_p-balls for $0 < p \geq 1$. *J. Complexity*, 26(6):629–640, 2010.

[16] C. S. Güntürk. One-bit sigma-delta quantization with exponential accuracy. *Comm. Pure Appl. Math.*, 56(11):1608–1630, 2003.

[17] C. S. Güntürk, M. Lammers, A. M. Powell, R. Saab, and Ö. Yılmaz. Sobolev duals for random frames and $\Sigma\Delta$ quantization of compressed sensing measurements. *Found. Comput, Math.*, 13(1):1–36, 2013.

[18] A. Gupta, R. Nowak, and B. Recht. Sample complexity for 1-bit compressed sensing and sparse classification. In *Proc. IEEE Int. Symp. Inf. Theory (ISIT) , Austin, TX*, pages 1553–1557. IEEE, Jun. 2010.

[19] D. L. Hanson and F. T. Wright. A bound on tail probabilities for quadratic forms in independent random variables. *Annals Math. Statistics*, 42(3):1079–1083, 1971.

[20] R. A. Horn and C. R. Johnson. *Matrix analysis*. Cambridge university press, 2012.

[21] H. Inose, Y. Yasuda, and J. Murakami. A telemetering system by code modulation-Δ-Σ modulation. *IRE Trans. Space El. Tel.*, (3):204–209, 1962.

[22] L. Jacques, J. N. Laska, P. T. Boufounos, and R. G. Baraniuk. Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors. *IEEE Trans. Inform. Theory*, 59(4):2082–2102, 2013.

[23] F. Krahmer, S. Mendelson, and H. Rauhut. Suprema of chaos processes and the restricted isometry property. *Comm. Pure Appl. Math.*, to appear.

[24] F. Krahmer, R. Saab, and R. Ward. Root-exponential accuracy for coarse quantization of finite frame expansions. *IEEE Trans. Inform. Theory*, 58(2):1069–1079, 2012.

[25] F. Krahmer, R. Saab, and Ö. Yılmaz. Sigma-Delta quantization of sub-gaussian frame expansions and its application to compressed sensing. *Inform. Inference*, 3(1):40–58, 2014.

[26] F. Krahmer and R. Ward. Lower bounds for the error decay incurred by coarse quantization schemes. *Appl. Comput. Harmonic Anal.*, 32(1):131–138, 2012.

[27] M. Lammers, A.M. Powell, and Ö. Yılmaz. Alternative dual frames for digital-to-analog conversion in sigma–delta quantization. *Adv. Comput. Math.*, pages 1–30, 2008.

[28] Y. Plan and R. Vershynin. One-bit compressed sensing by linear programming. *Comm. Pure Appl. Math.*, 66(8):1163–1333, 2013.

[29] M. Rudelson and R. Vershynin. On sparse reconstruction from fourier and gaussian measurements. *Comm. Pure Appl. Math.*, 61(8):1025–1045, 2008.

[30] M. Rudelson and R. Vershynin. Hanson-wright inequality and sub-gaussian concentration. *Electron. Comm. Probab.*, 18:1–9, 2013.

[31] A. M. Tillmann and M. E. Pfetsch. The computational complexity of the restricted isometry property, the nullspace property, and related concepts in compressed sensing. *IEEE Trans. Inform. Theory*, 2012.

[32] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Y.C. Eldar and G. Kutyniok, editors, *Compressed Sensing: Theory and Applications*, pages 210–268. Cambridge Univ. Press, Cambridge, 2012.

Simultaneous Subspace Pursuit for Signal Recovery from Multiple-Measurement Vectors

Joe-Mei Feng and Chia-han Lee
Research Center for Information Technology Innovation
Academia Sinica
Taipei, Taiwan

***Abstract*—Extension from the single-measurement vector (SMV) problem to the multiple-measurement vectors (MMV) problem is critical for compressed sensing (CS) applications in many fields. A few signal recovery algorithms, such as simultaneous orthogonal matching pursuit (SOMP), have been proposed to recover a jointly-sparse signal from the corresponding multiple-measurement vectors. However, those previously proposed algorithms generally do not have restricted isometry property (RIP) guarantee. In this paper, we propose the simultaneous compressive sampling matching pursuit (SCoSaMP) algorithm, a generalization of the compressive sampling matching pursuit (CoSaMP) algorithm for the MMV problem. We show the RIP guarantee that leads to the convergence of the proposed SCoSaMP algorithm and the uniqueness of the recovered signal. Simulation results confirm that SCoSaMP outperforms SOMP under random sampling matrix setup and with noisy measurements.**

I. Introduction

In a compressed sensing (CS) problem, greedy algorithms such as orthogonal matching pursuit (OMP) [1], [2], regularized OMP (ROMP) [3], subspace pursuit (SP) [4], and compressive sampling matching pursuit (CoSaMp) [5], were proposed to recover sparse signal $x \in \mathbb{R}^N$ from measurement vector $y = \Phi x$, where $y \in \mathbb{R}^n$ and $\Phi \in \mathbb{R}^{n \times N}$ is a sampling matrix. Extension of this single-measurement vector (SMV) problem to a multiple-measurement vectors (MMV) problem [6] in order to recover a signal $\mathbf{X} \in \mathbb{R}^{N \times M}$ from multiple-measurement vectors $\mathbf{Y} = \Phi\mathbf{X}$, where $\mathbf{Y} \in \mathbb{R}^{n \times M}$, is mandatory for applications in many fields, such as magnetoencephalography (MEG), array processing, equalization of sparse communication channels, cognitive radio, and multiband communications [6]. Since algorithms like subspace pursuit can exactly recover signal x from measurement vector $y = \Phi x$ with the restricted isometry property (RIP)-guarantee, a straightforward approach for solving the MMV problem is to apply the SP algorithm on each measurement vector of $\mathbf{Y}$ to recover a k-jointly-sparse signal $\mathbf{X}$ from $\mathbf{Y} = \Phi\mathbf{X}$ if Φ satisfies the exact recovery guarantee of CoSaMP. Since the number of the columns of $\mathbf{Y}$ is M, however, the complexity will scales linearly with M if this approach is taken. A more efficient algorithm is thus required. Several approaches have been proposed for solving the MMV problem, such as optimization-based algorithms [6], [7], greedy pursuit algorithms [6], [7], and the reduction to the SMV problem [6]. These algorithms are basically adapted from the algorithms proposed for the SMV signal recovery problem mentioned above. However, the exact recovery capabilities of these algorithms are either not guaranteed or only assured by such as the coherence-related guarantee and the rank-related guarantee [6], instead of the RIP guarantee. For example, the simultaneous orthogonal matching pursuit (SOMP) algorithm [1], [6], [7], [8], a popular greedy pursuit algorithm developed for the MMV problem, does not provide RIP guarantee.

In this paper, we propose the simultaneous compressive sampling matching pursuit (SCoSaMP) algorithm, an RIP and fusion restricted isometry property (FRIP, which will be explained in Section III)-guaranteed greedy algorithm, for solving the MMV problem. The SCoSaMP algorithm is not only an MMV extension but also a generalization of the subspace pursuit algorithm (a new parameter z is introduced, which will be explained in Section II). We will in this paper show the uniqueness guarantee of the recovery of a k-jointly-sparse signal for the MMV problem and prove the RIP/FRIP guarantee of the SCoSaMP algorithm. Simulation results confirm that (a) SCoSaMP is more efficient compared to applying CoSaMP directly to the MMV problem– unlike CoSaMP which scales linearly with M, SCoSaMP scales sublinearly with M, and (b) SCoSaMP outperforms SOMP under the random sampling matrix setup and the noisy measurement conditions.

The organization of this paper is the following. In Section II we introduce the SCoSaMP algorithm. In Section III we present the uniqueness guarantee of the recovery of a k-jointly-sparse signal for the MMV problem and the exact recovery guarantee of SCoSaMP. Then we evaluate the performance of SCoSaMP in Section IV and conclude the paper in Section V.

II. The SCoSaMP Algorithm

The SCoSaMP algorithm is a generalization of the CoSaMP algorithm in two aspects: (a) The ℓ_2-norm of each row is computed as the representing magnitude and the Frobenious norm (denoted as ℓ_F) replaces the ℓ_2-norm used in CoSaMP, and (b) a parameter z_1 and z_2 is introduced to control the number of columns to be selected (explained later). Before going into the details of the SCoSaMP algorithm, let us define some notations.

Definition 1. Given a matrix $\Phi \in \mathbb{R}^{n \times N}$ and the index set $T \subseteq \{1, \ldots, N\}$, $\Phi_{\cdot,T}$ is a matrix consisting of the T-indexed columns of Φ. For convenience, denote $\Phi_{\cdot,T}$ by Φ_T and denote $\mathtt{span}(\Phi_T)$ as the column space spanned by column vectors of

Φ_T. Also $\mathbf{X}|_T = \begin{cases} \mathbf{X}_{i,\cdot} & \text{, if } i \in T \\ \mathbf{0} & \text{, if } i \notin T \end{cases}$, where $\mathbf{X}_{i,\cdot}$ is the ith row of $\mathbf{X}$.

Definition 2. Let $\mathbf{Y} \in \mathbb{R}^{n\times M}$ and $\Phi_T \in \mathbb{R}^{n\times|T|}$, where $|T|$ is the size of T. Suppose that $\Phi_T^*\Phi_T$ is invertible, where Φ_I^* is the transposition of Φ_T. The projection of $\mathbf{Y}$ onto $\mathtt{span}(\Phi_I)$ is defined as $\mathbf{Y}_p = \mathtt{proj}(\mathbf{Y}, \Phi_T) \equiv \Phi_T\Phi_T^\dagger\mathbf{Y}$, where $\Phi_T^\dagger \equiv (\Phi_T^*\Phi_T)^{-1}\Phi_T^*$ is the pseudo-inverse of the matrix Φ_I. Also, define the residue vector of the projection $\mathbf{Y}_r$ as $\mathbf{Y}_r = \mathtt{resid}(\mathbf{Y}, \Phi_T) \equiv \mathbf{Y} - \mathbf{Y}_p$.

Now let us go through the steps of SCoSaMP, which are summarized in Table I. In SCoSaMP, the inputs are the MMV $\mathbf{Y}$ and the sampling matrix Φ. SCoSaMP in step It2 computes the inner products of each column of Φ with all columns of $\mathbf{Y}_r^{\ell-1}$ and forms the signal proxy. In step It3/It3' SCoSaMP identifies large components from signal proxy. In step It4 SCoSaMP merges supports. Step It5 is called signal estimation by least-squares, in which SCoSaMP projects $\mathbf{Y}$ onto the space spanned by $k+\mathtt{ceil}(k/z_1)$ or $k+\mathtt{ceil}(k/z_2)$ columns with indices selected in step It4. In step It6 SCoSaMP chooses k columns out according to the spanning scalars of the projection of $\mathbf{Y}$ on the spanning space of these $k+\mathtt{ceil}(k/z_1)$ or $k+\mathtt{ceil}(k/z_2)$ columns, and thus in step It6 the index set is pruned. After step It6 we obtain a k-sparse signal approximation $\hat{\mathbf{X}}_{T^\ell} = \Phi_{T^\ell}^\dagger\mathbf{Y}$. In step It7, $\mathbf{Y}$ is projected on the column space spanned by the pruned indexed columns in It6, and the difference between $\mathbf{Y}$ and its projection is the residue, which is fed into the next iteration. The process repeats until the stopping criterion is satisfied, where ℓ is the iteration number and $\|\cdot\|_F$ is the Frobenious norm.

Note that for convenience, throughout this paper we demand $\mathtt{ceil}(k/z_2) \le k+\mathtt{ceil}(k/z_1) \le 2k$. This is because we want to improve the performance at large sparsity, even when the uniqueness is not preserved, and reducing the identification number at step It3 reduces the probability of selecting unwanted columns. In Subsection III-B it will be seen that with the restriction to the identification number, z_2 plays the leading role when deciding the coefficients to ensure the convergence of SCoSaMP.

TABLE I
THE SCOSAMP ALGORITHM.

Input: k, Φ, $\mathbf{Y}$.
Initialization:
1) $T^0 = \emptyset$.
2) $\mathbf{Y}_r^0 = \mathbf{Y}$.
3) $\ell = 0$.
Iteration: At the ℓth iteration, go through the following steps.
It1) $\ell = \ell + 1$.
It2) Compute $\Phi^*\mathbf{Y}_r^{\ell-1}$.
It3) $\Omega = \{\mathtt{ceil}(k/z_2)$ largest ℓ_2-norm of the row vectors of $\Phi^*\mathbf{Y}_r^{\ell-1}\}$.
It3') When $\ell = 1$: $\Omega = \{k + \mathtt{ceil}(k/z_1)$ largest ℓ_2-norm of the row vectors of $\Phi^*\mathbf{Y}_r^{\ell-1}\}$.
It4) $\tilde{T}^\ell = \Omega \bigcup T^{\ell-1}$.
It5) Compute $\Phi_{\tilde{T}^\ell}^\dagger\mathbf{Y}$.
It6) $T^\ell = \{k$ largest ℓ_2-norm of the row vectors of $\Phi_{\tilde{T}^\ell}^\dagger\mathbf{Y}\}$.
It7) $\mathbf{Y}_r^\ell = \mathtt{resid}(\mathbf{Y}, \Phi_{T^\ell})$.
It8) If $\|\mathbf{Y}_r^\ell\|_F$ is small enough quit the iteration.
Output:
1) Index set T^ℓ.
2) Estimated signal $\hat{\mathbf{X}}_{T^\ell} = A_{T^\ell}^\dagger\mathbf{Y}$.

III. RIP GUARANTEE

In the SMV problem, a sampling matrix Φ satisfies RIP of order k with constant δ_k if for any k-sparse x, $(1-\delta_k)\|x\|_2^2 \le \|\Phi x\|_2^2 \le (1+\delta_k)\|x\|_2^2$, where $\|\cdot\|_2$ is the ℓ_2 norm and k is the number of non-zero entries in x. The uniqueness of the solution of the MMV problem can be guaranteed by the multiple-dimension version of RIP, the fusion restricted isometry property. A matrix Φ satisfies FRIP of order k with constant δ_k if there exists a constant δ_k such that for all k-jointly-sparse signals $\mathbf{X}$, $(1-\delta_k)\|\mathbf{X}\|_F^2 \le \|\Phi\mathbf{X}\|_F^2 \le (1+\delta_k)\|\mathbf{X}\|_F^2$, where $\|\cdot\|_F$ is the Frobenious (ℓ_F) norm. A signal $\mathbf{X}$ is k-jointly-sparse if, regardless of the abuse of notation, $\|\mathbf{X}\|_0 = |\{j : \mathbf{X}_{j,\cdot} \neq \mathbf{0}\}| \le k$, where $|\cdot|$ counts the number of a set and $\mathbf{X}_{j,\cdot}$ denotes the jth row of $\mathbf{X}$ [9]. Note that RIP implies FRIP [9], i.e., if for each column of $\mathbf{X}$ and the corresponding column of $\mathbf{Y}$, RIP of order k with constant δ_k holds, then FRIP of order k with constant δ_k also holds. Thus by carefully choosing the coefficients, RIP guarantees the exact recovery of SSP. Several methods have been known to construct a matrix satisfying RIP [10], thus FRIP, of specific order and constant.

In Subsection III-A, a necessary and sufficient condition for the unique recovery of a k-jointly-sparse signal will be stated, and the FRIP-related uniqueness guarantee for a k-jointly-sparse signal will be given. It is easy to see that the unique recovery of a k-jointly-sparse signal can be assured if any $2k$ columns of Φ are linearly independent. Then in Subsection III-B, the exact recovery guarantee of the proposed SCoSaMP algorithm will be demonstrated.

A. Uniqueness of the Recovery of k-jointly-sparse Signals

The non-FRIP-related necessary and sufficient condition of the unique recovery of a k-joint-sparse signal is stated below.

Theorem 1. [8] A necessary and sufficient condition for the measurement vectors $\mathbf{Y} = \Phi\mathbf{X}$ to uniquely determine the k-jointly-sparse matrix $\mathbf{X}$ is that

$$k < \frac{\mathtt{spark}(\Phi) - 1 + \mathtt{rank}(\mathbf{X})}{2}, \tag{1}$$

where $\mathtt{spark}(\Phi)$ is the smallest number of linearly dependent columns of Φ, and $\mathtt{rank}(\mathbf{X})$ is the rank of $\mathbf{X}$.

Proof: See [7], [8]. ■

Now we give an FRIP guarantee of the uniqueness of the recovery of a k-jointly-sparse signal. Denote the support set of $\mathbf{X}$ by $\mathtt{supp}(\mathbf{X})$ as the index set such that for element $j \in \mathtt{supp}(\mathbf{X})$, $\mathbf{X}_{j,\cdot} \neq \mathbf{0}$. Then Theorem 2 below guarantees the uniqueness of the recovery of a k-jointly-sparse signal.

Theorem 2. Suppose that $k \geq 1$ and Φ satisfies FRIP of order $2k$ with constant $\delta_{2k} < 1$, and let $\mathbf{X}$ be a k-jointly-sparse signal and $\mathbf{Y} = \Phi\mathbf{X}$. Then $\mathbf{X}$ can be reconstructed uniquely from the matrix $\mathbf{Y}$.

Proof: If $\mathbf{X} = \mathbf{0}$, then $\mathbf{Y} = \Phi\mathbf{X} = \mathbf{0}$. Given $\mathbf{Y} = \mathbf{0}$, $\mathbf{X} = \mathbf{0}$ is the only solution. If $\mathbf{X} \neq \mathbf{0}$, we prove by contradiction. Suppose there exist two nonzero k-jointly-sparse signals $\mathbf{X}_1$ and $\mathbf{X}_2$ such that $\mathbf{X}_2 \neq \mathbf{X}_1$ and $\mathbf{Y}_2 = \Phi\mathbf{X}_2 = \mathbf{Y}_1$. Let $\mathbf{D} = \mathbf{X}_1 - \mathbf{X}_2$, then $\mathbf{D}$ is a nonzero $2k$-jointly-sparse signal. By applying FRIP, we get

$$\begin{aligned} & (1-\delta_{2k})\|\mathbf{D}\|_F^2 \leq \|\Phi\mathbf{D}\|_F^2 \leq (1+\delta_{2k})\|\mathbf{D}\|_F^2 \\ \Rightarrow\ & (1-\delta_{2k})\|\mathbf{D}\|_F^2 \leq \|\Phi\mathbf{X}_1 - \Phi\mathbf{X}_2\|_F^2 \leq (1+\delta_{2k})\|\mathbf{D}\|_F^2 \\ \Rightarrow\ & (1-\delta_{2k})\|\mathbf{D}\|_F^2 \leq 0 \leq (1+\delta_{2k})\|\mathbf{D}\|_F^2. \end{aligned}$$

While $(1-\delta_{2k})\|\mathbf{D}\|_F^2 > 0$, this shows a contradiction. ∎

Theorem 2 guarantees the exact recovery $\hat{\mathbf{X}}$ of a k-jointly-spare signal $\mathbf{X}$ via the following ℓ_0-norm problem

$$\hat{\mathbf{X}} = \underset{\|\mathbf{X}\|_0 \leq k}{\arg\min} \|\mathbf{X}\|_0, \text{ subject to } \mathbf{Y} = \Phi\mathbf{X}. \tag{2}$$

Thus Theorem 2 implies Theorem 1 and therefore the unique recovery of a k-jointly-sparse signal.

B. RIP Guarantee of SCoSaMP

In order to prove the convergence and the exact recovery guarantee of the SCoSaMP algorithm, we need the following four lemmas.

Lemma 1. (Identification). Under condition of Theorem 3. There exists a constant s_1 such that if for each column of $\mathbf{X}$ and the corresponding column of $\mathbf{Y}$, the sampling matrix Φ satisfies RIP of order $4k$ with constant $\delta_{4k} \leq s_1$, then for each $\ell \geq 0$

$$\|\mathbf{X} - \hat{\mathbf{X}}_{T^\ell}|_{\Omega^c}\|_F \leq 0.2223\|\mathbf{X} - \hat{\mathbf{X}}_{T^\ell}\|_F,$$

where Ω^c is the set complement of Ω.

Proof: Similar to Lemma 4.2 [5] by three changes, and since we are given noiseless MMV, we delete terms including noise e. in [5]. First, $\|\cdot\|_2$ is changed to $\|\cdot\|_F$. Second, since we demand number of Ω, $\texttt{ceil}(k/z_2) \leq k + \texttt{ceil}(k/z_1) \leq 2k$, $\|\Phi^*\mathbf{Y}_r^\ell|_{\texttt{supp}(\mathbf{X}-\hat{\mathbf{X}}_{T^\ell})}\|_F \leq \|\Phi^*\mathbf{Y}_r^\ell|_\Omega\|_F$ is changed to $\|\Phi^*\mathbf{Y}_r^\ell|_{\texttt{supp}(\mathbf{X}-\hat{\mathbf{X}}_{T^\ell})}\|_F \leq 2z_2\|\Phi^*\mathbf{Y}_r^\ell|_\Omega\|_F$. Third, in line 6 of proof Lemma 4.2 [5], we change set $\Omega \setminus \texttt{supp}(\mathbf{X} - \hat{\mathbf{X}}_{T^\ell})$ to set Ω. Then finally it yields that

$$\|\mathbf{X} - \hat{\mathbf{X}}_{T^\ell}|_{\Omega^c}\|_F \leq \frac{\delta_{2k} + 2z_2\delta_{4k}}{1+\delta_{2k}}\|\mathbf{X} - \hat{\mathbf{X}}_{T^\ell}\|_F. \tag{3}$$

Since $\delta_{2k} \leq \delta_{4k}$ and analytical property of equation 3, we can choose $\delta_{4k} \leq s_1$ such that $\frac{\delta_{2k}+2z_2\delta_{4k}}{1+\delta_{2k}} < 0.2223$ ∎

Table II shows the relationship between z_2 and s_1.

Lemma 2. (Support merger). Under condition of Theorem 3.

$$\|\mathbf{X}|_{\tilde{T}^{\ell c}}\|_F \leq \|\mathbf{X} - \hat{\mathbf{X}}_{T^\ell}|_{\Omega^c}\|_F.$$

Proof: Similar to Lemma 4.3 [5] by changing $\|\cdot\|_2$ to $\|\cdot\|_F$. ∎

Lemma 3. (Estimation). Under condition of Theorem 3. If for each column of $\mathbf{X}$ and the corresponding column of $\mathbf{Y}$, the sampling matrix Φ satisfies RIP of order $4k$ with constant $\delta_{4k} \leq 0.1$, then for each $\ell \geq 0$

$$\|\mathbf{X} - \Phi^\dagger_{\tilde{T}^\ell}\mathbf{Y}\|_F \leq 1.112\|\mathbf{X}|_{\tilde{T}^c}\|_F$$

Proof: Similar to Lemma 4.4 [5] by changing $\|\cdot\|_2$ to $\|\cdot\|_F$ and delete terms including noise e. ∎

Lemma 4. (Pruning).

$$\|\mathbf{X} - \Phi^\dagger_{T^\ell}\mathbf{Y}\|_F \leq 2\|\mathbf{X} - \Phi^\dagger_{\tilde{T}^\ell}\mathbf{Y}\|_F$$

Proof: Similar to Lemma 4.5 [5] by changing $\|\cdot\|_2$ to $\|\cdot\|_F$. ∎

Now we are ready to see the main Theorem which shows the convergence of SCoSaMP.

Theorem 3. Demand $\texttt{ceil}(k/z_2) \leq k + \texttt{ceil}(k/z_1) \leq 2k$. Let $\mathbf{X} = (\mathbf{X}_{j,\cdot})_{j=1}^N$ be a k-jointly-sparse signal and $\mathbf{Y} = \Phi\mathbf{X}$ the corresponding noiseless MMV. There exists a constant $s < 1$ such that if for each column of $\mathbf{X}$ and the corresponding column of $\mathbf{Y}$, the sampling matrix Φ satisfies RIP of order $4k$ with constant $\delta_{4k} \leq s$, then for each $\ell \geq 0$, $\|\mathbf{X} - \hat{\mathbf{X}}_{T^{\ell+1}}\|_F \leq 0.5\|\mathbf{X} - \hat{\mathbf{X}}_{T^\ell}\|_F$. Thus SCoSaMP with constant (z_1, z_2) recover $\mathbf{X}$ exactly.

Proof: Similar to Theorem 4.1 in [**?**] by changing $\|\cdot\|_2$ to $\|\cdot\|_F$. Choosing $s = \min\{0.1, s_1\}$, where s_1 is from Lemma 1 and following proof of Theorem 4.1 in [5], then we have $\|\mathbf{X} - \hat{\mathbf{X}}_{T^{\ell+1}}\|_F \leq 0.5\|\mathbf{X} - \hat{\mathbf{X}}_{T^\ell}\|_F$. Since $0.5 < 1$, SCoSaMP converges. From Theorem 2, once SCoSaMP converges, SCoSaMP finds the unique k-jointly-sparse signal, and thus recover exactly. ∎

From the proof we see that by demanding $\texttt{ceil}(k/z_2) \leq k + \texttt{ceil}(k/z_1) \leq 2k$, only z_2 influences the RIP constant.

IV. Performance Evaluation

In this section we will evaluate the performance of the proposed SCoSaMP algorithm. First, the complexity in terms of execution time of SCoSaMP will be compared to that of CoSaMP applying directly to the MMV problem. Then we investigate the performance of SCoSaMP under different constants (z_1, z_2). Finally, we compare the performance between SCoSaMP and SOMP.

In the following simulations, random sampling matrices, instead of RIP-guaranteed sampling matrices, are used. This is because it is generally hard to construct an RIP-guaranteed sampling matrix and furthermore, a random matrix is proved to satisfy RIP criteria with high probability [11], [12]. Thus practical application scenarios favor the use of random matrices and it is critical for a signal recovery algorithm to perform well under this situation.

A. CoSaMP and SCoSaMP

Due to number of identification of CoSaMP, it can only recover one third sparsity out of signal dimension. Thus in SCoSaMP, we reduce the number of identification. Fig.

TABLE II
COEFFICIENTS FOR RIP GUARANTEE OF SCOSAMP.

$z_2 =$	0.5	1	1.5	2	2.5	3	3.5	4	4.5	5	$\cdots$
$s_1 =$	0.0999	0.0689	0.0526	0.0425	0.0357	0.0307	0.0270	0.0240	0.0217	0.0198	$\cdots$

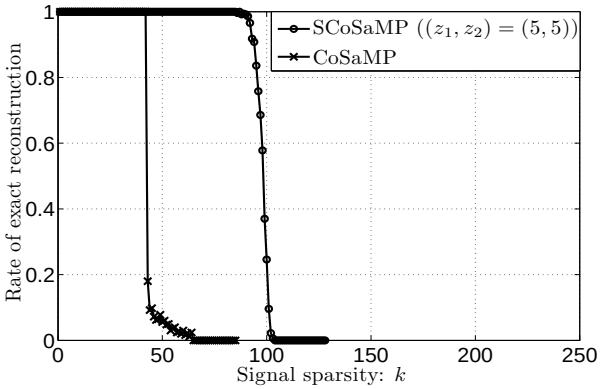

Fig. 1. Comparison of execution time of SCoSaMP and CoSaMP. $k = 5$, $n = 128$, $N = 256$, and the number of realizations is 300.

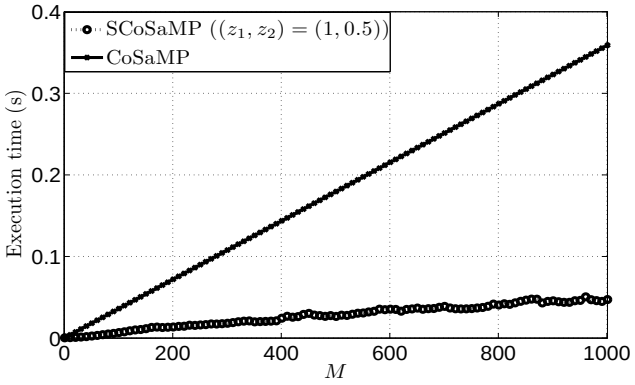

Fig. 2. Comparison of execution time of SCoSaMP and CoSaMP. $k = 5$, $n = 128$, $N = 256$, and the number of realizations is 300.

1 shows the performance of CoSaMP and SCoSaMP with constant $(5, 5)$. The number of realizations is set to 500, $n = 128$, $N = 256$, and $M = 10$, meaning that $\mathbf{X} \in \mathbb{R}^{256\times 10}$, $\Phi \in \mathbb{R}^{128\times 256}$, and $\mathbf{Y} \in \mathbb{R}^{128\times 10}$. Entry values of $\mathbf{X}$ are standard normally distributed and Φ is a random matrix with standard normal distribution.

Fig. 2 compares the execution time of CoSaMP and SCoSaMP with constant $(z_1, z_2) = (1, 0.5)$ when solving the MMV problem (remember that SCoSaMP is a generalization of CoSaMP and when $M = 1$ and $(z_1, z_2) = (1, 0.5)$ SCoSaMP becomes CoSaMP). In this simulation, $k = 5$, $n = 128$, $N = 256$, the number of realizations is 300, and M ranges from 1 to 1001, meaning that $\Phi \in \mathbb{R}^{128\times 256}$, $\mathbf{X} \in \mathbb{R}^{256\times M}$, and $\mathbf{Y} \in \mathbb{R}^{128\times M}$. Also, the entry values of $\mathbf{X}$ are standard normally distributed and Φ is a random matrix with standard normal distribution. The execution time was measured by running the algorithms on MATLAB. It is shown that the execution time of CoSaMP scales linearly with M, but the execution time of SCoSaMP scales sublinearly with M. This shows how much more efficient the SCoSaMP algorithm is when dealing with the MMV problem.

B. SCoSaMP Performance with Different Constant (z_1, z_2)

It is interesting to see how the SCoSaMP constant (z_1, z_2) affects the rate of exact recovery, which is defined as the rate that the estimated signal $\hat{\mathbf{X}}$ is different from $\mathbf{X}$ [1]. In Fig. **??**, we show the rate of exact recovery of SCoSaMP with constant $(z_1, z_2) = (5, 1)$, $(z_1, z_2) = (5, 2)$, $(z_1, z_2) = (5, 5)$ and $(z_1, z_2) = (5, 10)$. In Fig. **??**, we show the rate of exact recovery of SCoSaMP with constant $(z_1, z_2) = (1, 5)$, $(z_1, z_2) = (2, 5)$, $(z_1, z_2) = (5, 5)$ and $(z_1, z_2) = (10, 5)$. The number of realizations is set to 500, $n = 128$, $N = 256$, and $M = 10$, meaning that $\mathbf{X} \in \mathbb{R}^{256\times 10}$, $\Phi \in \mathbb{R}^{128\times 256}$, and $\mathbf{Y} \in \mathbb{R}^{128\times 10}$. Entry values of $\mathbf{X}$ are standard normally distributed and Φ is a random matrix with standard normal distribution. Since that given a 128×256 sampling matrix the best scenario is to recover a signal with sparsity 128, we let the sparsity k range from 1 to 128 (although Theorem 1 does not guarantee the uniqueness of a signal when the sparsity is larger than k).

The results in Fig. 4 show that larger z generally yields better performance. This is because when M is fixed and the sparsity is large, it is highly probable for SCoSaMP to select a column vector which is not linearly independent to other chosen column vectors. Thus increasing the constant z decreases the possibility of selecting column vectors that are dependent to each other. However, when z_1 and z_2 are too large, it also decreases the possibility of selecting the correct spanning column vectors. As a result, properly choosing the value of z leads to the optimal performance of the algorithm.

From the simulations, we also observe that the rate of exact recovery of SCoSaMP with $(z_1, z_2) = (5, 5)$, $(z_1, z_2) = (5, 10)$, and $(z_1, z_2) = (10, 5)$ remain 1 when sparsity is 87, which is larger than 66 that inequality (1) of Theorem 1 guarantees. This demonstrates the robustness of the proposed SCoSaMP algorithm.

C. SCoSaMP v.s. SOMP

Now we compare the performance of SCoSaMP and SOMP. There are extensive discussions about the norm used in the SOMP algorithm [8]. Since ℓ_2-norm is used in SCoSaMP, we also use ℓ_2-norm in SOMP to be fair. We let the entry values of $\mathbf{X}$ be standard normally distributed and we compare the

[1]Due to MATLAB computational precision, in each realization, if the ℓ_2 norm of the difference between $\mathbf{X}$ and $\hat{\mathbf{X}}$ is less than 10^{-10}, then we claim that they are equivalent; otherwise, they are different.

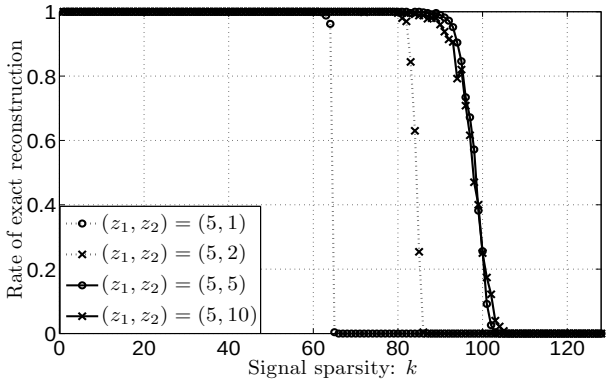

Fig. 3. Rate of exact recovery using SCoSaMP with different constants (z_1, z_2). $n = 128$, $N = 256$, $M = 10$, and the number of realizations is 500.

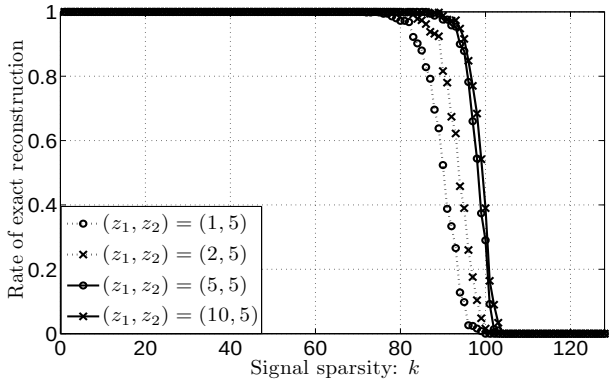

Fig. 4. Rate of exact recovery using SCoSaMP with different constants (z_1, z_2). $n = 128$, $N = 256$, $M = 10$, and the number of realizations is 500.

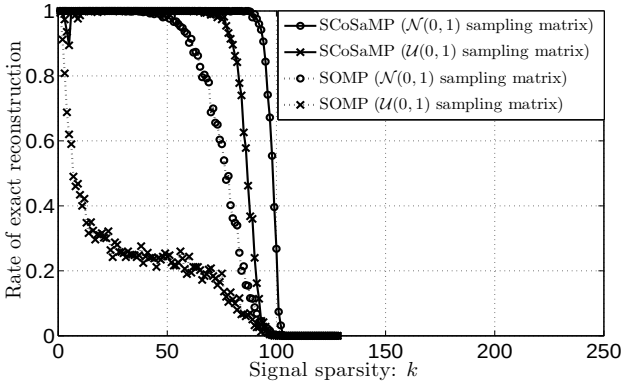

Fig. 5. Rate of exact recovery with differently distributed sampling matrices. $n = 128$, $N = 256$, $M = 10$, the number of realizations is 500, and the SCoSaMP constant is $(5, 5)$.

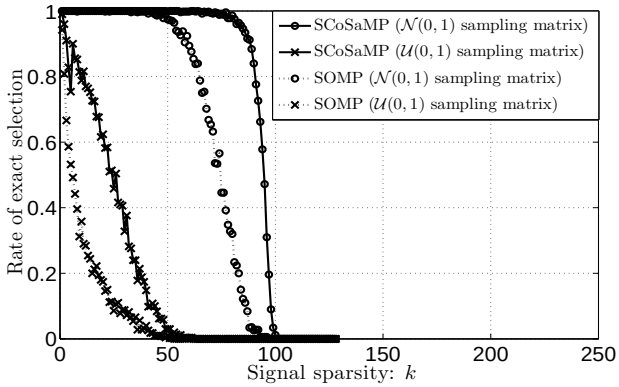

Fig. 6. Rate of exact selection with standard normally distributed noise. $n = 128$, $N = 256$, $M = 10$, the number of realizations is 500, and the SCoSaMP constant is $(5, 5)$.

performance under two types of random sampling matrices: random matrix with standard normal distribution (denoted as $\mathcal{N}(0, 1)$ sampling matrix) and random matrix with uniform distribution (denoted as $\mathcal{U}(0, 1)$ sampling matrix). Let $n = 128$, $N = 256$, $M = 10$, the number of realizations be 500, and the SCoSaMP constant be $(5, 5)$. Fig. 5 clearly shows that SCoSaMP outperforms SOMP in both cases in terms of the rate of exact recovery.

It is also of great interest to know how SCoSaMP and SOMP perform when the measurements are noisy. Let us assume that noise is standard normally distributed. Instead of the rate of exact recovery, we evaluate the performances by the rate of exact selection, which is defined as the rate that $\texttt{supp}(\mathbf{X}) = T^\ell$. Recall that $\texttt{supp}(\mathbf{X}) = \{i : \mathbf{X}_{i,\cdot} \neq \mathbf{0}\}$ and from SCoSaMP we found T^ℓ to be the estimated support set of $\mathbf{X}$, which means $\texttt{span}(\Phi_{T^\ell})$ is the estimated spanning space of $\mathbf{Y}$. Again let $n = 128$, $N = 256$, $M = 10$, the number of realizations be 500, and the SCoSaMP constant be $(5, 5)$. The results in Fig. 6 show that SCoSaMP outperforms SOMP even under noisy measurements. We attribute the stability of SCoSaMP to the pruning part (step 2 and 3 in the iteration) of the algorithm, which allows the algorithm to have better chance to select correct columns.

V. Conclusion

It is natural to extend from the single-measurement vector problem to the multiple-measurement vectors problem in compressed sensing. In this paper, we have proposed an efficient algorithm, the simultaneous subspace pursuit algorithm, for solving the MMV problem. This proposed SCoSaMP algorithm is efficient in the sense that it deals with these multiple-measurement vectors at the same time rather than using the SP algorithm to cope with the single measurement vector multiple times. We have shown the RIP/FRIP guarantees of the proposed algorithm and surprisingly SCoSaMP works well even when the sparsity is larger than what is guaranteed. By properly choosing the constant parameter z, SCoSaMP easily outperforms the popular SOMP algorithm when the sampling matrix is random and under noisy measurements.

ACKNOWLEDGMENT

The support from National Science Council under Grant NSC99-2218-E-001-011-MY2 is gratefully acknowledged.

REFERENCES

[1] J. A. Tropp, "Greedy is good: Algorithmic results for sparse approximation," *IEEE Trans. Info. Theory*, vol. 50, no. 10, pp. 2231–2242, Oct. 2004.

[2] M. A. Davenport and M. B. Wakin, "Analysis of orthogonal matching pursuit using the restricted isometry property," *IEEE Trans. Info. Theory*, vol. 56, no. 9, pp. 4395–9448, Sep. 2010.

[3] D. Needell and R. Vershynin, "Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit," *Found. of Comp. Mathematics*, vol. 9, no. 3, pp. 317–334, Jun. 2009.

[4] W. Dai and O. Milenkovic, "Subspace pursuit for compressive sensing signal reconstruction," *IEEE Trans. Info. Theory*, vol. 55, no. 5, pp. 2230–2249, May 2009.

[5] D. Needell and J. A. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Appl. Comput. Harmon. Anal.*, vol. 26, no. 3, pp. 301–321, May 2009.

[6] M. F. Duarte and Y. C. Eldar, "Structured compressed sensing: From theory to applications," *IEEE Trans. Signal Proc.*, vol. 59, no. 9, pp. 4053–4085, Sep. 2011.

[7] M. E. Davies and Y. C. Eldar, "Rank awareness in joint sparse recovery," Apr. 2010, arxiv preprint arXiv:1004.4529.

[8] J. Chen and X. Huo, "Theoretical results on sparse representations of multiple-measurement vector," *IEEE Trans. Signal Proc.*, vol. 54, no. 12, pp. 4634–4643, Dec. 2006.

[9] P. T. Boufounos, G. Kutyniok, and H. Rauhut, "Sparse recovery from combined fusion frame measurements," *IEEE Trans. Info. Theory*, vol. 6, no. 6, pp. 3864–3876, Jun. 2009.

[10] R. A. DeVore, "Deterministic constructions of compressed sensing matrices," *Journal of Complexity*, vol. 23, no. 4-6, pp. 918–925, Aug.-Dec. 2007.

[11] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, "A simple proof of the restricted isometry property for random matrices," *Constr. Approx.*, vol. 28, no. 3, pp. 253–263, Jan. 2008.

[12] R. Calderbank, S. Howard, and S. Jafarpour, "Construction of a large class of deterministic sensing matrices that satisfy a statistical isometry property," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 2, pp. 358–374, Apr. 2010.

J Glob Optim (2012) 54:275–293
DOI 10.1007/s10898-010-9625-6

Duality and solutions for quadratic programming over single non-homogeneous quadratic constraint

Joe-Mei Feng · Gang-Xuan Lin · Reuy-Lin Sheu · Yong Xia

Received: 10 August 2010 / Accepted: 26 October 2010 / Published online: 17 November 2010

Abstract This paper extends and completes the discussion by Xing et al. (Canonical dual solutions to the quadratic programming over a quadratic constraint, submitted) about the quadratic programming over one quadratic constraint (QP1QC). In particular, we relax the assumption to cover more general cases when the two matrices from the objective and the constraint functions can be simultaneously diagonalizable via congruence. Under such an assumption, the nonconvex (QP1QC) problem can be solved through a dual approach with no duality gap. This is unusual for general nonconvex programming but we can explain by showing that (QP1QC) is indeed equivalent to a linearly constrained convex problem, which happens to be dual of the dual of itself. Another type of hidden convexity can be also found in the boundarification technique developed in Xing et al. (Canonical dual solutions to the quadratic programming over a quadratic constraint, submitted).

Keywords Non-convex quadratic programming · Simultaneously diagonalizable via congruence · Slater's condition · Duality · Hidden convexity

1 Introduction

Let A and B be two $n \times n$ real symmetric matrices, μ be a real number and f, g be two $n \times 1$ vectors. This paper concerns the (nonconvex) quadratic minimization problem (QP1QC)

$$\begin{aligned} P_0 = \min \quad & P(x) = \tfrac{1}{2}x^T Ax - f^T x \\ s.t. \quad & \tfrac{1}{2}x^T Bx - g^T x \le \mu, \end{aligned} \tag{1}$$

J.-M. Feng · G.-X. Lin · R.-L. Sheu (✉)
Department of Mathematics, National Cheng Kung University, Tainan, Taiwan
e-mail: rsheu@mail.ncku.edu.tw

Y. Xia
LMIB of the Ministry of Education, School of Mathematics and System Sciences, Beihang University, Beijing, People's Republic of China

which has a single non-homogeneous quadratic constraint. To make sense the problem, we assume that the problem (QP1QC) has a non-empty feasible domain, and that

- (A1) there exists a $\sigma' \geq 0$ such that $A + \sigma' B \succ 0$,

where the notation "$\succ$" means "positive definiteness". Although the assumption (A1) looks restrictive, it can be shown that our results indeed cover more general cases when A and B are simultaneously diagonalizable via *congruence* (SDC in short). However, we can also prove that cases between (A1) and (SDC) are either unbounded below or transformed to an equivalent unconstrained problem. It is thus clear that the only relevant cases remained for study are those under (A1).

The problem (QP1QC) arises from many optimization algorithms, in particular, the trust region methods. See, e.g., [2,3,7,9,15,17,18,20,21]. Most cases dealed with $B \succ 0$ and $g = 0$. Extensions to an indefinite B but still with $g = 0$ are considered in [1,6,14,16,19,22]. Recently, a general non-convex function $f(x)$ rather than a non-convex quadratic function $P(x)$ over a single quadratic constraint was discussed by Jeyakumar et al. [13]. Stern and Wolkowicz [19] first studied the two-sided nonconvex quadratically constrained problem

$$\begin{array}{l} \min \mu(y) = y^T B y - 2\psi^T y \\ s.t.\ \beta \leq y^T C y \leq \alpha,\ y \in \mathbb{R}^n \end{array} \tag{2}$$

and gave the necessary and sufficient optimality conditions with no duality gap under various assumptions. Ben-Tal and Teboulle [1] further explained the surprising result by proving that, under the assumption (A1), the problem (2) is indeed equivalent to a *convex* minimization problem with linear constraints. Xing et al. [22] relaxed the assumption (A1) to

- (A2) the domain $\mathcal{J} = \{\sigma \geq 0 | A + \sigma B \succeq 0\}$ has a non-empty interior; and
- (A3) the vectors $f, g \in \mathcal{R}(A + \sigma B)$ for some $\sigma \in int(\mathcal{J})$,

where $\mathcal{R}(A + \sigma B)$ denotes the column space of $A + \sigma B$ and $int(\mathcal{J})$ means the interior of $\mathcal{J}$. (Remark: $g = 0$ in [22]). By applying the canonical duality of Gao [8], they formulate and solve the dual problem of (QP1QC) (with $g = 0$) analytically. With a proper selection of the dual optimal solution when there are multiple ones, they used a technique called "boundarification" to construct a primal global optimal solution with no duality gap. Xing et al. [22] also showed that problem (2) can be solved by doing (QP1QC) (with $g = 0$) at most twice. As a result, it is sufficient to consider only (QP1QC).

In spite that all theoretical results were related to the quadratic form, direct applications of (QP1QC) having a nonhomogeneous quadratic constraint can be found in solving an inverse problem via regularization [5,10] and in minimizing the double well potential function [4]. Our purpose in this paper is to provide a complete mathematical treatment for (QP1QC) from the dual side. Following Xing's derivation in [22], it is not difficult to formulate the canonical dual problem (D-QP1QC) (with the appearance of g) as follows.

$$\begin{array}{l} P_0^d = \sup_\sigma \ P^d(\sigma) = -\frac{1}{2}(f + \sigma g)^T (A + \sigma B)^{-1}(f + \sigma g) - \mu\sigma \\ \qquad s.t.\ \sigma \in \mathcal{F} = \{\sigma \geq 0 | A + \sigma B \succ 0\}. \end{array} \tag{3}$$

The Lagrange dual and the canonical dual of (QP1QC) are indeed equivalent in this case. We do not intend to distinguish them in the presentation.

In the next section, we shall discuss and compare various assumptions necessarily for solving (QP1QC), including (A1), (A2)+(A3), Slater conditions, and the (SDC) condition. We will show that the (SDC) condition is the broadest in the sense that it contains all three other assumptions. In fact, the (SDC) condition for more than two quadratic forms is the *twelfth* open question among 14 such ones in nonlinear analysis and optimization raised by

J.-B. Hiriart-Urruty [11]. In the third section, we extend the analysis in Xing et al. [22] to provide a dual approach for solutions of (QP1QC), while focusing on new cases exclusively for $g \neq 0$. In the fourth section, a deeper view for the boundarification technique will be also presented. In the fifth section, we extend the idea in Ben-Tal and Teboulle [1] to derive the dual of the dual for (QP1QC). It is shown that the double dual of (QP1QC) is a linearly constrained convex minimization problem, which is equivalent to (QP1QC) itself under a one-one nonlinear transformation. Finally, we illustrate with some numerical examples.

2 Relaxed assumptions

In a constrained optimization problem, Slater condition requires a strictly feasible solution. The assumption (A1) can be viewed as the dual Slater condition. For the primal problem (1), the primal Slater condition requires an x_0 such that $\frac{1}{2}x_0^T Bx_0 - g^T x_0 < \mu$. Ben-Tal and Teboulle [1] and Xing et al. [22] needed the dual Slater condition (A1) whereas Ye and Zhang's exact semi-definite programming approach [23] did both. Although (A2)+(A3) is weaker than (A1), they can be reduced to (A1) after space reduction. See [22] for details.

Notice that the Lagrange function

$$L(x,\sigma) = \frac{1}{2}x^T(A+\sigma B)x - (f+\sigma g)^T x - \mu\sigma \tag{4}$$

is unbounded below when $A+\sigma B$ is not positive semi-definite. If $\mathcal{J} = \{\sigma \geq 0 | A+\sigma B \succeq 0\} = \emptyset$, the dual problem becomes identically negative infinite and thus useless. Assumptions like (A1) or (A2) are important since they guarantee that a sensible dual information is indeed there. Although $\mathcal{J} = \emptyset$ implies no dual information, we show that those cases satisfying either

- (A4) $\mathcal{F}' = \{(\lambda,\nu) \in \mathbb{R}^2 | \ \lambda A + \nu B \succ 0\} \neq \emptyset$, or more generally
- (A5) (SDC) A and B are simultaneously diagonalizable via congruence

while $\mathcal{J} = \emptyset$ can be reduced to some *trivial* quadratic problems.

Recall that the matrices A and B are said to be simultaneously diagonalizable via congruence if there exists a nonsingular matrix C with

$$C^T AC = \mathfrak{A} := diag(\alpha_1, \cdots, \alpha_n), \quad \alpha_j \in \mathbb{R},\ j \in [1:n];$$

$$C^T BC = \mathfrak{B} := diag(\beta_1, \cdots, \beta_n), \quad \beta_j \in \mathbb{R},\ j \in [1:n].$$

It is different from being simultaneously diagonalizable via *similarity* in the usual sense. It is known that "(A4)⇒(A5)", e.g. in [12]. Using the transformation $x = Cy$, $\eta = C^T f$, and $\varphi = C^T g$, the primal problem (1) can be written as a linear combination of separated squares as follows:

$$\begin{aligned} P_0' = \min\ P'(y) &= \sum_{i=1}^{n} \tfrac{1}{2}\alpha_i y_i^2 - \eta_i y_i \\ s.t.\ S'(y) &= \sum_{i=1}^{n} \tfrac{1}{2}\beta_i y_i^2 - \varphi_i y_i - \mu \leq 0. \end{aligned} \tag{5}$$

Similarly, the dual problem (3) becomes

$$\begin{aligned} P_0^{d'} = \sup_{\sigma\geq 0} P^{d'}(\sigma) &= -\tfrac{1}{2}\sum_{i=1}^{n} \frac{(\eta_i+\sigma\varphi_i)^2}{\alpha_i+\sigma\beta_i} - \mu\sigma \\ s.t.\ \alpha_i + \sigma\beta_i &> 0, \end{aligned} \tag{6}$$

and the condition (A4) becomes $\{(\lambda, \nu) \in \mathbb{R}^2 |\ \lambda\mathfrak{A} + \nu\mathfrak{B} \succ 0\} \neq \emptyset$.

We first call the two matrices $\mathfrak{A}$ and $\mathfrak{B}$ to be of type I if there is some $i \in [1:n]$ for which $\alpha_i \leq 0$ and $\beta_i \leq 0$; and are of type II otherwise. When the two matrices $\mathfrak{A}$ and $\mathfrak{B}$ are of type I, we may assume that it occurs to $i = 1$. Moreover, if (A4) is satisfied, α_1 and β_1 can not be 0 simultaneously. Therefore, when $\alpha_1 = 0$, β_1 must be less than 0. When $\alpha_1 < 0$, β_1 can be non-positive. We now study cases of type I, satisfying (A4) while violating (A1).

(a) Suppose that $\alpha_1 = 0,\ \beta_1 < 0$. Since $\beta_1 < 0$, we observe that any $y \in R^n$ will eventually become feasible if the first component y_1 is increased to infinity or decreased to negative infinity while all others $y_2, y_3, \ldots, y_n$ being fixed. As a result, if $\eta_1 \neq 0$ in the objective function, we can pass y_1 (of any selected y) to $+\infty$ for $\eta_1 > 0$, or to $-\infty$ for $\eta_1 < 0$, both of which imply that the Problem (5) is unbounded below. If $\eta_1 = 0$, the objective function lacks the y_1 term and the constraint becomes redundant. The problem (5) is reduced to the unconstrained quadratic programming problem $\min \sum_{i=2}^{n} \frac{1}{2}\alpha_i y_i^2 - \eta_i y_i$.

(b) Suppose that $\alpha_1 < 0,\ \beta_1 \leq 0$. For any feasible solution y, replace y_1 by $y_1 = \text{sign}(\varphi_1)t$ if $\varphi_1 \neq 0$, and by $y_1 = t$ otherwise. As $t \to +\infty$, the constraint remains satisfied while the objective value approaches to $-\infty$. The problem (5) is unbounded.

Next we assume that the two matrices $\mathfrak{A}$ and $\mathfrak{B}$ are of type II. That is, for all $i \in [1:n]$, either $\alpha_i > 0$ or $\beta_i > 0$. Then we can partition the index set $[1:n] = I_1 \cup I_2 \cup I_3$ where $I_1 = \{i : \beta_i > 0\}$, $I_2 = \{i : \alpha_i > 0, \beta_i < 0\}$, and $I_3 = \{i : \alpha_i > 0, \beta_i = 0\}$. We claim that, if (A1) is violated, $I_1 \neq \emptyset$, $I_2 \neq \emptyset$ and there exist two indices $i \in I_1$, $j \in I_2$ such that $\alpha_i < 0, \beta_i > 0$ and $\alpha_j > 0, \beta_j < 0$.

Suppose $I_1 = \emptyset$. Since $\alpha_i > 0,\ \forall i \in [1:n]$, the assumption (A1) is satisfied by setting $\sigma' = 0$, which is a contradiction and we conclude that $I_1 \neq \emptyset$. Suppose $I_2 = \emptyset$. We can choose σ' sufficiently large to make a contradiction that (A1) is satisfied. Therefore, $I_2 \neq \emptyset$.

Secondly, we must have $\max_{i\in I_1} \frac{-\alpha_i}{\beta_i} \geq \min_{i\in I_2} \frac{-\alpha_i}{\beta_i}$. If this is not true, we can choose $\widetilde{\mu} > 0, \widetilde{\nu} > 0$ such that

$$\max_{i\in I_1} \frac{-\alpha_i}{\beta_i} < \frac{\widetilde{\nu}}{\widetilde{\mu}} < \min_{i\in I_2} \frac{-\alpha_i}{\beta_i}.$$

That is, $\widetilde{\mu}\alpha_i + \widetilde{\nu}\beta_i > 0$ for all $i \in I_1 \cup I_2$. It follows that $\alpha_i + \frac{\widetilde{\nu}}{\widetilde{\mu}}\beta_i > 0$ for all $i \in [1:n]$. It is again a contradiction.

In general, we may assume $1 \in I_1, 2 \in I_2$, $\max_{i\in I_1} \frac{-\alpha_i}{\beta_i} = \frac{-\alpha_1}{\beta_1}$, $\min_{i\in I_2} \frac{-\alpha_i}{\beta_i} = \frac{-\alpha_2}{\beta_2}$, and $\frac{-\alpha_1}{\beta_1} \geq \frac{-\alpha_2}{\beta_2} > 0$. It follows that $\alpha_1 < 0, \beta_1 > 0, \alpha_2 > 0, \beta_2 < 0$, and there are two possibilities:

(c) Suppose $\max_{i\in I_1} \frac{-\alpha_i}{\beta_i} = \min_{i\in I_2} \frac{-\alpha_i}{\beta_i}$, i.e., $\frac{-\alpha_1}{\beta_1} = \frac{-\alpha_2}{\beta_2}$. Since

$$\{(\mu, \nu) \mid \mu\alpha_1 + \nu\beta_1 > 0,\ \mu\alpha_2 + \nu\beta_2 > 0\} = \left\{(\mu, \nu) \mid \mu\frac{\alpha_1}{\beta_1} + \nu > 0,\right.$$
$$\left.\mu\frac{-\alpha_2}{\beta_2} - \nu > 0\right\} = \emptyset,$$

Assumption (A4) is violated.

(d) Suppose $\max_{i\in I_1} \frac{-\alpha_i}{\beta_i} > \min_{i\in I_2} \frac{-\alpha_i}{\beta_i}$, i.e., $\frac{-\alpha_1}{\beta_1} > \frac{-\alpha_2}{\beta_2}$. For any fixed $t \in (\frac{-\beta_1}{\beta_2}, \frac{-\alpha_1}{\alpha_2})$, we have $\alpha_1 + \alpha_2 t < 0$ and $\beta_1 + \beta_2 t < 0$. Consider $y = (y_1, y_2, 0, 0, \ldots, 0)^t$ where y_2 satisfies $y_2^2 = t y_1^2$. Then, y is feasible to (5) for all large enough $|y_1|$, which decreases the objective values indefinitely. In other words, the problem (5) is unbounded below.

Bearing the above discussion in mind, for cases that violate (A4), there is only one case in type I, that is $\alpha_i = \beta_i = 0$ for some $i \in [1:n]$; and only one case in type II, which is case (c). We begin with type I by assuming that $\alpha_1 = \beta_1 = 0$.

(e) Suppose that $\alpha_1 = 0,\ \beta_1 = 0,\ \eta_1 \neq 0,\ \eta_1\varphi_1 \geq 0$. For any feasible solution y, let $\eta_1 y_1 \to +\infty$. The problem (5) is unbounded below.

(f) Suppose that $\alpha_1 = 0,\ \beta_1 = 0,\ \eta_1\varphi_1 < 0$. Rewrite the constraint as $\varphi_1 y_1 \geq (\sum_{i=2}^n \frac{1}{2}\beta_i y_i^2 - \varphi_i y_i) - \mu$. If $\varphi_1 > 0$, we substitute

$$y_1 \geq \frac{1}{\varphi_1}\sum_{i=2}^n \left(\frac{1}{2}\beta_i y_i^2 - \varphi_i y_i\right) - \frac{\mu}{\varphi_1}$$

into the objective and obtain

$$-\eta_1 y_1 + \sum_{i=2}^n \left(\frac{1}{2}\alpha_i y_i^2 - \eta_i y_i\right) \geq \frac{-\eta_1}{\varphi_1}\sum_{i=2}^n \left(\frac{1}{2}\beta_i y_i^2 - \varphi_i y_i\right) + \frac{\eta_1\mu}{\varphi_1} + \sum_{i=2}^n \left(\frac{1}{2}\alpha_i y_i^2 - \eta_i y_i\right) \tag{7}$$

The problem (5) is equivalently reduced to the unconstrained quadratic programming problem: $\min \sum_{i=2}^n (\frac{1}{2}\alpha_i y_i^2 - \eta_i y_i) - \frac{\eta_1}{\varphi_1}(\sum_{i=2}^n (\frac{1}{2}\beta_i y_i^2 - \varphi_i y_i) - \mu)$. For $\varphi_1 < 0$, the same inequality (7) can be also established and the problem (5) is equivalently reduced to the above unconstrained quadratic programming problem.

(g) Suppose that $\alpha_1 = 0,\ \beta_1 = 0,\ \eta_1 = 0,\ \varphi_1 \neq 0$. Since $\lim_{y_1 \to +\infty} -|\varphi_1 y_1| = -\infty$ and the objective function has no y_1 term, the problem (5) is equivalent to the unconstrained quadratic programming problem: $\min \sum_{i=2}^n (\frac{1}{2}\alpha_i y_i^2 - \eta_i y_i)$.

(h) Suppose that $\alpha_1 = 0,\ \beta_1 = 0,\ \eta_1 = 0,\ \varphi_1 = 0$. The variable y_1 can be directly removed from the problem (5).

Next, we discuss the case left in type II, case (c). In this case, $\alpha_1 < 0, \beta_1 > 0, \alpha_2 > 0, \beta_2 < 0$ and $\frac{-\alpha_1}{\beta_1} = \frac{-\alpha_2}{\beta_2}$. We can further assume $\varphi_1 = 0$ and $\varphi_2 = 0$, by introducing the linear transformations $\widetilde{y}_1 = y_1 - \frac{\varphi_1}{\beta_1}$ and $\widetilde{y}_2 = y_2 - \frac{\varphi_2}{\beta_2}$ if necessary. As a result,

$$-\frac{1}{2}\beta_1 y_1^2 - \frac{1}{2}\beta_2 y_2^2 \geq -\mu + \sum_{i=3}^n \left(\frac{1}{2}\beta_i y_i^2 - \varphi_i y_i\right).$$

(i) Suppose $\eta_1 = 0$ and $\eta_2 = 0$. We have

$$\begin{aligned}\frac{1}{2}\alpha_1 y_1^2 + \frac{1}{2}\alpha_2 y_2^2 &= \frac{-\alpha_1}{\beta_1}\left(-\frac{1}{2}\beta_1 y_1^2 - \frac{1}{2}\beta_2 y_2^2\right) \\ &\geq \frac{-\alpha_1}{\beta_1}\left(-\mu + \sum_{i=3}^n \left(\frac{1}{2}\beta_i y_i^2 - \varphi_i y_i\right)\right)\end{aligned} \tag{8}$$

The problem (5) is reduced to an unconstrained quadratic programming problem: $\min \sum_{i=3}^n (\frac{1}{2}(\alpha_i - \frac{\alpha_1}{\beta_1}\beta_i) y_i^2 - (\eta_i - \frac{\alpha_1}{\beta_1}\varphi_i) y_i) + \frac{\alpha_1}{\beta_1}\mu$.

(j) Suppose $\eta_1 \neq 0$ or $\eta_2 \neq 0$. Let us assume $\eta_1 \neq 0$. Set $\widehat{y}_1 = \text{sign}(\eta_1) t$ and if $\eta_2 \neq 0, \widehat{y}_2 = \text{sign}(\eta_2)\sqrt{\frac{-\beta_1}{\beta_2}t^2 + \frac{2\mu}{\beta_2}}$, otherwise $\widehat{y}_2 = \sqrt{\frac{-\beta_1}{\beta_2}t^2 + \frac{2\mu}{\beta_2}}$. Notice that $(\widehat{y}_1, \widehat{y}_2, 0, \ldots, 0)$ is feasible for any large enough t. Furthermore, the corresponding objective function value is $\frac{1}{2}\alpha_1\widehat{y}_1^2 + \frac{1}{2}\alpha_2\widehat{y}_2^2 - \eta_1\widehat{y}_1 - \eta_2\widehat{y}_2 = \frac{\mu\alpha_2}{\beta_2} - \eta_1\widehat{y}_1 - \eta_2\widehat{y}_2 = \frac{\mu\alpha_2}{\beta_2} - |\eta_1| t - |\eta_2|\sqrt{\frac{-\beta_1}{\beta_2}t^2 + \frac{2\mu}{\beta_2}} \to -\infty$ as $t \to +\infty$. That is, the problem (5) is unbounded below.

Therefore, we only have to discuss problem (QP1QC) under the assumption (A1), while in principle we know how to deal with it under the (SDC) condition. Cases between (A1) and (SDC) are either unbounded below or reduced to an unconstrained problem.

In addition, we comment that the condition (A2)+(A3) in in Xing et al. [22] is also a special case of (SDC). They have shown in [22] that, under (A2)+(A3), there exists an orthogonal matrix W such that

$$W^T A W = \begin{bmatrix} 0 & 0 \\ 0 & A_1 \end{bmatrix}, W^T B W = \begin{bmatrix} 0 & 0 \\ 0 & B_1 \end{bmatrix}, \quad W^T (A + \sigma B) W = \begin{bmatrix} 0 & 0 \\ 0 & A_1 + \sigma B_1 \end{bmatrix},$$

and $A_1 + \sigma B_1 \succ 0$. Since $(A1) \Rightarrow (A5)$, there is a nonsingular matrix U such that $U^T A_1 U = \mathrm{diag}(\alpha_1^1, \ldots, \alpha_m^1)$ and $U^T B_1 U = \mathrm{diag}(\beta_1^1, \ldots, \beta_m^1)$. Therefore, A and B can be simultaneously diagonalizable via congruence with a nonsingular matrix $C = W \begin{bmatrix} I & 0 \\ 0 & U \end{bmatrix}$ such that both $C^T A C$ and $C^T B C$ are diagonal matrices, which is (A5).

Finally, we remark that the primal Slater condition is redundant under the assumption (A1). The gap between these two assumptions is the case when $B \succeq 0,\ g \in \mathcal{R}(B),\ \mu = -\frac{1}{2} g^T B^+ g$, where B^+ is the Moore-Penrose generalized inverse of B. This special case can be treated by (case 7) discussed in the next section.

Proposition 1 *Let* $F_0 := \{(B, g, \mu) : \exists x,\ \frac{1}{2} x^T B x - g^T x \le \mu\}$ *and* $F_1 := \{(B, g, \mu) : \exists x,\ \frac{1}{2} x^T B x - g^T x < \mu\}$. *Then it holds that*

$$F_0 \backslash F_1 = \left\{ (B, g, \mu) : B \succeq 0,\ g \in \mathcal{R}(B),\ \mu = -\frac{1}{2} g^T B^+ g \right\}.$$

Proof According to the definitions of F_0 and F_1, $(B, g, \mu) \in F_0 \backslash F_1$ if and only if $\mu = \min_x \{\frac{1}{2} x^T B x - g^T x\} > -\infty$, which is equivalent to $B \succeq 0$ and the linear system $Bx = g$ has a solution. In this case, $\mu = -\frac{1}{2} g^T B^+ g$. □

3 Solutions to (QP1QC) via dual approach

In the general theoretical view of nonlinear programming, the problem (QP1QC) can be written as $\min\{P(x) |\ S(x) \le 0\}$ where $P(x) = \frac{1}{2} x^T A x - f^T x$ and $S(x) = \frac{1}{2} x^T B x - g^T x - \mu$. For the single inequality constrained problem, let $\sigma \ge 0$ be the Lagrange multiplier. Then, the dual problem is formulated as follows:

$$\sup_{\sigma \in \mathfrak{D}} \ (P^d(\sigma) = \inf_{x \in \mathbb{R}^n} \{P(x) + \sigma S(x)\})$$

where a natural dual feasible domain $\mathfrak{D} = \{\sigma \ge 0 | P^d(\sigma) > -\infty\} \subset \{\sigma \ge 0 |\ A + \sigma B \succeq 0\} = \mathcal{J}$. For any $\sigma \in \mathfrak{D}$, suppose $x(\sigma)$ is a global minimizer of the quadratic convex function $P(x) + \sigma S(x)$, we have $P^d(\sigma) = P(x(\sigma)) + \sigma S(x(\sigma))$. Equivalently,

$$P(x(\sigma)) - P^d(\sigma) = -\sigma S(x(\sigma)). \tag{9}$$

Moreover, if $P^d(\sigma)$ and $x(\sigma)$ are smooth with respect to σ, the first derivative of $P^d(\sigma)$ can be computed using the implicit differentiation and the chain rule to get

$$\begin{aligned}\frac{d}{d\sigma}P^d(\sigma) &= \nabla P(x(\sigma))\frac{d}{d\sigma}x(\sigma) + S(x(\sigma)) + \sigma\nabla S(x(\sigma))\frac{d}{d\sigma}x(\sigma)\\ &= [\nabla P(x(\sigma)) + \sigma\nabla S(x(\sigma))]\frac{d}{d\sigma}x(\sigma) + S(x(\sigma))\\ &= S(x(\sigma))\end{aligned} \tag{10}$$

where $\nabla P(x(\sigma)) + \sigma\nabla S(x(\sigma)) = 0$ since $x(\sigma)$ minimizes $P(x) + \sigma S(x)$. Likely, we can compute the second derivative of $P^d(\sigma)$ as

$$\frac{d^2}{d\sigma^2}P^d(\sigma) = -\left[\frac{d}{d\sigma}S(x(\sigma))\right]^T(\nabla^2 P(x) + \sigma\nabla^2 S(x))^{-1}\left[\frac{d}{d\sigma}S(x(\sigma))\right] \tag{11}$$

provided the Lagrangian Hessian $\nabla^2 P(x) + \sigma\nabla^2 S(x)$ is invertible.

Equations (9, 10, 11) are general results for any single constrained smooth nonlinear programming problem. They provide structural insight to many direct computation in Xing et al. [22]. When $P(x)$ and $S(x)$ are quadratic and $\sigma \in \mathcal{F} = \{\sigma \geq 0 | A + \sigma B \succ 0\}$,

$$x(\sigma) = (A + \sigma B)^{-1}(f + \sigma g) \tag{12}$$

is the unique minimum solution to $P(x) + \sigma S(x)$. We thus have

Lemma 1 *For any $\sigma \in int(\mathcal{F})$,*

$$\begin{aligned}\frac{dP^d(\sigma)}{d\sigma} &= S(x(\sigma))\\ &= \frac{1}{2}(f + \sigma g)^T(A + \sigma B)^{-1}B(A + \sigma B)^{-1}(f + \sigma g)\\ &\quad -g^T(A + \sigma B)^{-1}(f + \sigma g) - \mu,\end{aligned} \tag{13}$$

and

$$\frac{d^2P^d(\sigma)}{d\sigma^2} = -u^T(A + \sigma B)^{-1}u \leq 0 \tag{14}$$

where $u = Bx(\sigma) - g = B(A + \sigma B)^{-1}(f + \sigma g) - g$. Moreover,

$$P(x(\sigma)) - P^d(\sigma) = -\sigma\frac{dP^d(\sigma)}{d\sigma}. \tag{15}$$

Observe from Lemma 1 that $P^d(\sigma)$ is a smooth concave function over $int(\mathcal{F})$. Additionally, it has been shown in [22] that $\mathcal{F}$ is an interval. Let σ_0, σ_1 respectively be the left and right boundary of $\mathcal{F}$. According to the various places where the supremum of $P^d(\sigma)$ could possibly attain over $\mathcal{F}$, we can classify into the following cases.

- (case 1) P_0^d is attained at $\sigma^* \in int(\mathcal{F})$. In this case, it is necessary that $\frac{dP^d(\sigma^*)}{d\sigma} = S(x(\sigma^*)) = 0$. By (15), $P(x(\sigma^*)) = P^d(\sigma^*)$ so that$(x(\sigma^*), \sigma^*)$ is the primal-dual optimal pair.
- (case 2) P_0^d is attained at the left boundary σ_0 of $\mathcal{F}$ and $\lim_{\sigma\to\sigma_0^+}\frac{dP^d(\sigma)}{d\sigma} = 0$. In this case, the limit

$$x(\sigma_0) = \lim_{\sigma\to\sigma_0^+}(A + \sigma B)^{-1}(f + \sigma g) \tag{16}$$

exists. By

$$\lim_{\sigma\to\sigma_0^+}\frac{dP^d(\sigma)}{d\sigma}=\lim_{\sigma\to\sigma_0^+}S(x(\sigma))=S(x(\sigma_0))=0,$$

$x(\sigma_0)$ is the primal boundary optimal solution.

- (case 3) P_0^d is attained at σ_0, $\sigma_0=0$ and $\lim_{\sigma\to\sigma_0^+}\frac{dP^d(\sigma)}{d\sigma}<0$. In this case, let $x(\sigma_0)$ be defined as in (16). Since $S(x(\sigma_0))<0$ and $-\sigma_0\frac{dP^d(\sigma_0)}{d\sigma}=0$ in (15), $x(\sigma_0)$ is an interior optimal solution.
- (case 4) P_0^d is attained at σ_0, $\sigma_0>0$ and $\lim_{\sigma\to\sigma_0^+}\frac{dP^d(\sigma)}{d\sigma}<0$. Then, $A+\sigma_0 B\succeq 0$ has at least one zero eigenvalue so that $P^d(\sigma_0)=\inf_{x\in\mathbb{R}^n}\{P(x)+\sigma_0 S(x)\}$ has multiple minimum solutions. That is, the optimal set $\arg\inf\{P(x)+\sigma_0 S(x)\}$ is not a singleton. Since the point $x(\sigma_0)$ in (16) satisfies $(A+\sigma_0 B)x(\sigma_0)=f+\sigma_0 g$, it belongs to the set $\arg\inf\{P(x)+\sigma_0 S(x)\}$ but has a positive duality gap: $P(x(\sigma_0))-P^d(\sigma_0)=-\sigma_0\frac{dP^d(\sigma_0)}{d\sigma}>0$. The boundarification technique developed in [22] chooses from the set $\arg\inf\{P(x)+\sigma_0 S(x)\}$ another more appropriate point than $x(\sigma_0)$ to close the gap.
- (case 5) P_0^d is attained at the right boundary σ_1 of $\mathcal{F}$, $\sigma_1<\infty$ and $\lim_{\sigma\to\sigma_1^-}\frac{dP^d(\sigma)}{d\sigma}=0$. This case is analogous to (case 2). The limit $x(\sigma_1)=\lim_{\sigma\to\sigma_1^-}(A+\sigma B)^{-1}(f+\sigma g)$ is a boundary optimal solution.
- (case 6) P_0^d is attained at σ_1, $\sigma_1<\infty$ and $\lim_{\sigma\to\sigma_1^-}\frac{dP^d(\sigma)}{d\sigma}>0$. This case is similar to (case 4) and requires a boundarification step to move $x(\sigma_1)$ from the exterior $S(x(\sigma_1))>0$ to a boundary optimal solution.
- (case 7) $P^d(\sigma)$ approaches asymptotically to a finite value as $\sigma\to+\infty$. This case can happen only when $B\succeq 0$ and a specific μ is associated.

In the following, we write the above seven cases into a few theorems with proofs which extend from the previous work of Xing et al. [22]. We shall try to reduce the duplicate to a minimum, while keeping it self-contained.

Under (A1), there exists a $\sigma'\geq 0$ such that $A+\sigma' B\succ 0$. The entire analysis relies heavily on the simultaneous decomposition that converts $A+\sigma' B\succ 0$ to an identical matrix I and B to a diagonal matrix H. First decompose $A+\sigma' B=LDL^T$ with a lower triangular matrix L and a diagonal matrix D having only positive diagonal entries. Let $G_1=(LD^{\frac{1}{2}})^{-1}$ so that $G_1(A+\sigma' B)G_1^T=I$. Since $G_1 BG_1^T$ is real and symmetric, there is an orthogonal matrix G_2 such that $G_2(G_1 BG_1^T)G_2^{-1}=H=diag(h_1,h_2,\ldots,h_n)$. Respectively, we have

$$G_1(A+\sigma' B)G_1^T=I,\quad G_2G_1(B)G_1^TG_2^T=H.$$

Lemma 2 (For cases 2–6) *Suppose the dual optimal occurs at $\sigma^*=\sigma_0$ or at $\sigma_1<\infty$. Then the corresponding (right/left) limit*

$$\bar{x}=\lim_{\sigma\to\sigma^*}(A+\sigma B)^{-1}(f+\sigma g)\tag{17}$$

exists. Moreover, $\bar{x}$ minimizes the Lagrange function $P(x)+\sigma^* S(x)$.

Proof Let $G=G_2G_1$. Then, for any $\sigma\in\mathcal{F}$,

$$\begin{aligned}G(A+\sigma B)G^T&=G(A+\sigma' B+\sigma B-\sigma' B)G^T\\&=G(A+\sigma' B)G^T+(\sigma-\sigma')GBG^T\\&=I+(\sigma-\sigma')H\\&=diag(d_1(\sigma),d_2(\sigma),\ldots,d_n(\sigma))\end{aligned}\tag{18}$$

where $d_i(\sigma) = 1 + (\sigma - \sigma')h_i > 0,\ i \in [1:n]$. In addition, we also have

$$P^d(\sigma) = -\frac{1}{2}(f + \sigma g)^T G^T diag(d_1^{-1}(\sigma), d_2^{-1}(\sigma), \ldots, d_n^{-1}(\sigma))G(f + \sigma g) - \mu\sigma.$$

Note that, as a function of $\sigma \in \mathcal{F}$,

$$d_i^{-1}(\sigma) = \frac{1}{1 + (\sigma - \sigma')h_i} > 0$$

is monotone with a lower bound 0. As $\sigma \to \sigma^*$, $d_i^{-1}(\sigma)$ might tend monotonically to a finite positive number or to $+\infty$. In the latter case if $\lim_{\sigma\to\sigma^*} d_i^{-1}(\sigma)$ tends to $+\infty$ for some component i, the corresponding $(G(f+\sigma g))_i$ must be 0 or otherwise P_0^d could have been unbounded below, which is impossible under (A1). In other words, $\lim_{\sigma\to\sigma^*} d_i^{-1}(\sigma)(G(f + \sigma g))_i$ will be either 0 or a finite limit for all $i \in [1:n]$, so the limit

$$\begin{aligned}\bar{x} &= \lim_{\sigma\to\sigma^*} G^T diag(d_1^{-1}(\sigma), d_2^{-1}(\sigma), \ldots, d_n^{-1}(\sigma))G(f + \sigma g)\\ &= \lim_{\sigma\to\sigma^*}(A + \sigma B)^{-1}(f + \sigma g).\end{aligned}$$

exists. Moreover, since $A + \sigma^* B \succeq 0$, the Lagrange function $P(x) + \sigma^* S(x)$ is convex. By

$$(A + \sigma^* B)\bar{x} = (\lim_{\sigma\to\sigma^*}(A + \sigma B))(\lim_{\sigma\to\sigma^*}(A + \sigma B)^{-1}(f + \sigma g)) = f + \sigma^* g, \qquad (19)$$

it is seen that $\bar{x}$ minimizes $P(x) + \sigma^* S(x)$. □

Theorem 1 (Boundarification Technique for cases (4) and (6)) *For any $\tilde{x} \neq 0$ in the null space of $A + \sigma^* B$, there exists at least one real number α_0 satisfying*

$$\alpha^2 \tilde{x}^T B\tilde{x} + 2\alpha(\bar{x}^T B\tilde{x} - g^T\tilde{x}) + \bar{x}^T B\bar{x} - 2g^T\bar{x} - 2\mu = 0$$

such that $x^* = \bar{x} + \alpha_0\tilde{x}$ is a required boundary global optimal solution for (1) with the global minimum value $\frac{1}{2}(x^*)^T Ax^* - f^T x^*$.

Proof Let $\sigma^* = \sigma_0$. The case when $\sigma^* = \sigma_1$ can be done similarly. Since $A + \sigma^* B$ is positive semi-definite but not positive definite, there exists some $\tilde{x} \neq 0$ such that $(A + \sigma^* B)\tilde{x} = 0$. However, $\tilde{x}^T(A + \sigma B)\tilde{x} > 0$ as $(A + \sigma B) \succ 0$ for $\sigma \in \mathcal{F}$. Then,

$$\tilde{x}^T(A + \sigma B)\tilde{x} - \tilde{x}^T(A + \sigma_0 B)\tilde{x} = (\sigma - \sigma^*)\tilde{x}^T B\tilde{x} > 0 \qquad (20)$$

implies that $\tilde{x}^T B\tilde{x} > 0$.

Now we consider the following quadratic function of one variable α:

$$\alpha^2 \tilde{x}^T B\tilde{x} + 2\alpha(\bar{x}^T B\tilde{x} - g^T\tilde{x}) + \bar{x}^T B\bar{x} - 2g^T\bar{x} - 2\mu = 0. \qquad (21)$$

Since $\tilde{x}^T B\tilde{x} > 0$ and $\bar{x}^T B\bar{x} - 2g^T\bar{x} - 2\mu < 0$, we find

$$\triangle = (2\bar{x}^T B\tilde{x} - 2g^T\tilde{x})^2 - 4(\tilde{x}^T B\tilde{x})(\bar{x}^T B\bar{x} - 2g^T\bar{x} - 2\mu) > 0$$

so that a real number α_0 satisfying (21) exists, and

$$\begin{aligned} S(x^*) &= \frac{1}{2}(x^*)^T Bx^* - g^T x^* - \mu \\ &= \frac{1}{2}(\bar{x} + \alpha_0 \tilde{x})^T B(\bar{x} + \alpha_0 \tilde{x}) - g^T(\bar{x} + \alpha_0 \tilde{x}) - \mu \\ &= \frac{1}{2}[\alpha_0^2 \tilde{x}^T B\tilde{x} + 2\alpha_0(\bar{x}^T B\tilde{x} - g^T \tilde{x}) + \bar{x}^T B\bar{x} - 2g^T \bar{x} - 2\mu] \\ &= 0. \end{aligned}$$

Since $(A + \sigma^* B)x^* = (A + \sigma^* B)(\bar{x} + \alpha_0 \tilde{x}) = (f + \sigma^* g)$, the point x^* also minimizes $P(x) + \sigma^* S(x)$. By (9), $P(x^*) - P_0^d = -\sigma^* S(x^*) = 0$ and proves the optimality of x^*. □

Recall that $B \succeq 0$ if and only if $\sigma_1 = +\infty$. For convenience, let the first r diagonal elements of $H = GBG^T$ are 0 and the remaining $n - r$ positive. When $r = 0$, B is positive definite. We denote $Gf = w = (\widetilde{w}^T, \overline{w}^T)^T$; $Gg = s = (\widetilde{s}^T, \overline{s}^T)^T$ where $\widetilde{w}$ means the first r components of w whereas $\overline{w}$ the last $n - r$. Similarly for s, and $\overline{H} \succ 0$ represents $diag(h_{r+1}, h_{r+2}, \ldots, h_n)$.

Theorem 2 (Case 7) *If P_0^d is achieved when σ tends to $+\infty$, then $\mu = -\sum_{h_i>0} \frac{s_i^2}{2h_i}$ and $s_i = 0$ for $h_i = 0$. Otherwise, P_0^d is unbounded above and problem (1) is infeasible. In case $B \succ 0$, the limit*

$$\bar{x} = \lim_{\sigma \to +\infty} (A + \sigma B)^{-1}(f + \sigma g) = B^{-1} g$$

exists and $\bar{x}$ is a boundary optimal solution for problem (1). In case $B \succeq 0$,

$$\bar{x} = G^T(\widetilde{w}^T, (\overline{H}^{-1}\overline{s})^T)^T$$

is the boundary optimal solution for problem (1).

Proof Expand $P^d(\sigma)$ into a rational polynomial of terms $\sigma^2, \sigma, \sigma^0, \sigma^{-1}$ as follows:

$$\begin{aligned} P^d(\sigma) &= -\frac{1}{2}(f + \sigma g)^T (A + \sigma B)^{-1}(f + \sigma g) - \mu\sigma \\ &= -\frac{1}{2}(G(f + \sigma g))^T (I + (\sigma - \sigma')H)^{-1}(G(f + \sigma g)) - \mu\sigma \\ &= -\frac{1}{2}\sum_{i=1}^{n} \frac{(w_i + \sigma s_i)^2}{1 + (\sigma - \sigma')h_i} - \mu\sigma \\ &= -\frac{1}{2}\sum_{h_i=0} (w_i + \sigma s_i)^2 - \frac{1}{2}\sum_{h_i>0} \frac{(w_i + \sigma s_i)^2}{1 + (\sigma - \sigma')h_i} - \mu\sigma \\ &= -\frac{1}{2}\sigma^2 \sum_{h_i=0} s_i^2 - \sigma \sum_{h_i=0} w_i s_i - \frac{1}{2}\sum_{h_i=0} w_i^2 \\ &\quad -\frac{1}{2}\sigma^2 \sum_{h_i>0} \frac{s_i^2}{1 + (\sigma - \sigma')h_i} - \sigma \sum_{h_i>0} \frac{w_i s_i}{1 + (\sigma - \sigma')h_i} \\ &\quad -\frac{1}{2}\sum_{h_i>0} \frac{w_i^2}{1 + (\sigma - \sigma')h_i} - \mu\sigma \\ &= -\frac{1}{2}\sigma^2 \sum_{h_i=0} s_i^2 - \sigma \sum_{h_i=0} w_i s_i - \frac{1}{2}\sum_{h_i=0} w_i^2 \end{aligned}$$

$$
\begin{aligned}
&-\frac{1}{2}\sum_{h_i>0}\left(\frac{\sigma s_i^2}{h_i}-\frac{s_i^2(1-\sigma' h_i)}{h_i^2}+\frac{s_i^2(1-\sigma' h_i)^2}{h_i^2(1+(\sigma-\sigma')h_i)}\right)\\
&-\sum_{h_i>0}\left(\frac{w_i s_i}{h_i}-\frac{w_i s_i(1-\sigma' h_i)}{h_i(1+(\sigma-\sigma')h_i)}\right)-\frac{1}{2}\sum_{h_i>0}\frac{w_i^2}{1+(\sigma-\sigma')h_i}-\mu\sigma\\
=&-\frac{1}{2}\sigma^2\sum_{h_i=0}s_i^2-\sigma\left[\sum_{h_i=0}w_i s_i+\frac{1}{2}\sum_{h_i>0}\frac{s_i^2}{h_i}+\mu\right]\\
&-\left[\frac{1}{2}\sum_{h_i=0}w_i^2-\frac{1}{2}\sum_{h_i>0}\frac{s_i^2(1-\sigma' h_i)}{h_i^2}+\sum_{h_i>0}\frac{w_i s_i}{h_i}\right]\\
&-\left[\sum_{h_i>0}\left(\frac{s_i^2(1-\sigma' h_i)^2}{2h_i^2}-\frac{w_i s_i(1-\sigma' h_i)}{h_i}+\frac{1}{2}w_i^2\right)\left(\frac{1}{1+(\sigma-\sigma')h_i}\right)\right].
\end{aligned}
$$

By the fact that $P_0^d=\lim_{\sigma\to+\infty}\mathrm{P}^d(\sigma)<\infty$, it is necessary that $s_i=0$ if $h_i=0$ and $\mu=-\sum_{h_i>0}\frac{s_i^2}{2h_i}$, or P_0^d is unbounded above, causing an infeasible primal problem.

When $B\succ 0$, $h_i>0$ for all $i\in[1:n]$. Then,

$$
\lim_{\sigma\to+\infty}d_i^{-1}(\sigma)(G(f+\sigma g))_i=\lim_{\sigma\to+\infty}\frac{w_i+\sigma s_i}{1+(\sigma-\sigma')h_i}=\frac{s_i}{h_i}
$$

and

$$
\begin{aligned}
\bar{x}&=\lim_{\sigma\to+\infty}(A+\sigma B)^{-1}(f+\sigma g)\\
&=\lim_{\sigma\to+\infty}G^T(I+(\sigma-\sigma')H)^{-1}G(f+\sigma g)\\
&=G^T\lim_{\sigma\to+\infty}(I+(\sigma-\sigma')H)^{-1}G(f+\sigma g)\\
&=G^T H^{-1}Gg\\
&=B^{-1}g.
\end{aligned}
$$

Consequently,

$$
\begin{aligned}
P(\bar{x})-P_0^d&=\lim_{\sigma\to+\infty}(P(x(\sigma))-P^d(\sigma))\\
&=\lim_{\sigma\to+\infty}-\sigma S(x(\sigma))\\
&\le 0,
\end{aligned}
$$

since, in case 7, $\frac{dP^d(\sigma)}{d\sigma}=S(x(\sigma))\ge 0$ on $int(\mathcal{F})$. This forces $P(\bar{x})=P_0^d$ and $S(\bar{x})=0$ so that $\bar{x}$ is a boundary optimal solution of (1).

In the case that $B\succeq 0$, we have $GAG^T=I-\sigma' H=diag(1,1,\ldots,1,1-\sigma' h_{r+1},1-\sigma' h_{r+2},\ldots,1-\sigma' h_n)$. Let $x=G^T t$, and by the fact that $s_i=0$ if $h_i=0$, problem (1) becomes

$$
\begin{aligned}
\mathrm{P}_0\ =\min\ \mathrm{P}(t)\ &=\ \tfrac{1}{2}t^T\begin{bmatrix}\widetilde{I}&0\\0&\overline{I}-\sigma'\overline{H}\end{bmatrix}t-(\widetilde{w}^T,\overline{w}^T)t\\
s.t.\ \tfrac{1}{2}t^T&\begin{bmatrix}0&0\\0&\overline{H}\end{bmatrix}t-(0_r,\overline{s}^T)t\le\mu,
\end{aligned}
\tag{22}
$$

which can be divided into two separated subproblems, an unconstrained ($\widetilde{\mathrm{P}}$) and a smaller ($\overline{\mathrm{P}}$) with $t = (\widetilde{t}^T, \bar{t}^T)^T$:

$$(\widetilde{\mathrm{P}}) \ : \ \widetilde{P_0} \ = \ \min \ \widetilde{\mathrm{P}}(\widetilde{t}) \ = \ \frac{1}{2}\widetilde{t}^T\widetilde{t} - \widetilde{w}^T\widetilde{t}$$

and

$$\begin{aligned}(\overline{\mathrm{P}}) \ : \ \overline{P_0} \ &= \min \ \overline{\mathrm{P}}(\bar{t}) \ = \ \tfrac{1}{2}\bar{t}^T(\overline{I} - \sigma'\overline{H})\bar{t} - \overline{w}^T\bar{t} \\ &\quad s.t. \ \tfrac{1}{2}\bar{t}^T\overline{H}\bar{t} - \bar{s}^T\bar{t} \le \mu.\end{aligned} \tag{23}$$

The unconstrained convex subproblem ($\widetilde{\mathrm{P}}$) attains the minimum at a $\widetilde{t}^*$ for which $\frac{d\widetilde{P}}{dt}(\widetilde{t}^*) = 0$. Hence $\widetilde{t}^* = \widetilde{w}$ and $\widetilde{P}_0 = -\frac{1}{2}\widetilde{w}^T\widetilde{w}$. As for ($\overline{\mathrm{P}}$), the minimum occurs at $\overline{t}^* = \overline{H}^{-1}\overline{s}$ due to $\overline{H} \succ 0$. Combining both $\widetilde{t}^*$ and $\overline{t}^*$ together, $\bar{x} = G^T(\widetilde{w}^T, (\overline{H}^{-1}\overline{s})^T)^T$ is the boundary optimal solution for problem (1). □

4 Insights into boundarification

To look into the boundarification technique for more insights, we work on the (SDC) form (P') in (5) and ($P^{d'}$) in (6) by thinking of A as $diag(\alpha_1, \ldots, \alpha_n)$; B as $diag(\beta_1, \ldots, \beta_n)$; f as $\eta = (\eta_1, \eta_2, \ldots, \eta_n)^T$; and g as $\varphi = (\varphi_1, \varphi_2, \ldots, \varphi_n)^T$. In (case 4), P_0^d occurs at the left boundary $\sigma_0 > 0$ so that the index set

$$I_0 := \{i | \alpha_i + \sigma_0\beta_i = 0\} \neq \emptyset. \tag{24}$$

We assume that $I_0 = [1 : r], \ r \ge 1$. Hence, $\alpha_i + \sigma_0\beta_i > 0, \ i \in [r+1 : n]$.

Lemma 3 (Case 4) *For all* $i \in I_0$, $\beta_i > 0$ *and* $\alpha_i < 0$. *Moreover,*

$$\sigma_0 = -\frac{\alpha_i}{\beta_i} = -\frac{\eta_i}{\varphi_i} \quad \textit{and thus} \ \frac{\varphi_i}{\beta_i} = \frac{\eta_i}{\alpha_i}. \tag{25}$$

Proof From (24), if $i \in I_0$ and $\beta_i = 0, \alpha_i = 0$. However, from assumption (A1), $\beta_i = 0$ implies $\alpha_i > 0$. The contradiction shows that $\beta_i \neq 0$. On the other hand, suppose $\beta_i < 0$. To make $\alpha_i + \sigma\beta_i > 0$ for $\beta_i < 0$, we need $\sigma < \sigma_0$, which contradicts to the fact that σ_0 is a left boundary of $\mathcal{F}$. Therefore, $\beta_i > 0, \ i \in I_0$ in Case 4. By $\sigma_0 > 0$, we immediately have $\alpha_i < 0$.

From Lemma 2, it is shown that

$$\bar{y} = \lim_{\sigma \to \sigma_0^+} (A + \sigma B)^{-1}(f + \sigma g) = \lim_{\sigma \to \sigma_0^+} \frac{\eta_i + \sigma\varphi_i}{\alpha_i + \sigma\beta_i}$$

exists. Since $\alpha_i + \sigma_0\beta_i = 0$ for $i \in I_0$, it is necessary that $\eta_i + \sigma_0\varphi_i = 0, \ i \in I_0$ and

$$\overline{y}_i = \begin{cases} \dfrac{\eta_i + \sigma_0\varphi_i}{\alpha_i + \sigma_0\beta_i}, & \text{if } i \notin I_0, \\ \dfrac{\varphi_i}{\beta_i}, & \text{if } i \in I_0, \end{cases} \tag{26}$$

which proves (25). □

From the condition of case (4),

$$\lim_{\sigma \to \sigma_0^+} \frac{dP^d(\sigma)}{d\sigma} = S'(\bar{y}) = \sum_{i=1}^{n} \left\{ \frac{\beta_i}{2}\overline{y}_i^2 - \varphi_i\overline{y}_i \right\} - \mu < 0$$

and there is a duality gap $P'(\bar{y}) - P_0^{d'} = -\sigma_0 S'(\bar{y}) > 0$. Since $\beta_i > 0$ for $i \in I_0$, we observe that $\lim_{t\to\pm\infty} \frac{\beta_i}{2}t^2 - \varphi_i t = +\infty$. To improve $S'(\bar{y})$ to 0, it is sufficient to increase (or decrease) any (or all) $\bar{y}_i$ for $i \in I_0$, while keeping all other $\bar{y}_i$ fixed if $i \notin I_0$. This amounts to moving $\bar{y}$ in the null space of $A + \sigma_0 B$ as for any $\widetilde{y}$ in the null space of $A + \sigma_0 B$, it must have

$$\widetilde{y}_i = 0, \quad \text{if } i \notin I_0.$$

This explains why in Eq. (21) there is a real root α_0.

From Theorem 1, we also see that the moving direction in the null space of $A + \sigma_0 B$ is indifferent to global optimality. It has to do with the specialty of case 4 as revealed in Lemma 3. For simplicity, let $r = 2$ and consider the problem:

$$\begin{aligned} &\min \left(\frac{\alpha_1}{2}y_1^2 - \eta_1 y_1\right) + \left(\frac{\alpha_2}{2}y_2^2 - \eta_2 y_2\right) + \sum_{i=3}^{n} \frac{1}{2}\alpha_i y_i^2 - \eta_i y_i \\ &s.t. \left(\frac{\beta_1}{2}y_1^2 - \varphi_1 y_1\right) + \left(\frac{\beta_2}{2}y_2^2 - \varphi_2 y_2\right) + \sum_{i=3}^{n} \frac{1}{2}\beta_i y_i^2 - \varphi_i y_i - \mu \le 0. \end{aligned} \tag{27}$$

By (25), we can assume

$$\frac{\varphi_1}{\beta_1} = \frac{\eta_1}{\alpha_1} = c_1 \quad \text{and} \quad \frac{\varphi_2}{\beta_2} = \frac{\eta_2}{\alpha_2} = c_2$$

so that (27) becomes

$$\begin{aligned} &\min P'(y) = \alpha_1\left(\frac{y_1^2}{2} - c_1 y_1\right) + \alpha_2\left(\frac{y_2^2}{2} - c_2 y_2\right) + \sum_{i=3}^{n} \frac{1}{2}\alpha_i y_i^2 - \eta_i y_i \\ &s.t.\ S'(y) = \beta_1\left(\frac{y_1^2}{2} - c_1 y_1\right) + \beta_2\left(\frac{y_2^2}{2} - c_2 y_2\right) + \sum_{i=3}^{n} \frac{1}{2}\beta_i y_i^2 - \varphi_i y_i - \mu \le 0. \end{aligned} \tag{28}$$

Since $\beta_1(\frac{y_1^2}{2} - c_1 y_1) + \beta_2(\frac{y_2^2}{2} - c_2 y_2) \ge -\frac{\beta_1 c_1^2}{2} - \frac{\beta_2 c_2^2}{2}$, we have

$$\sum_{i=3}^{n} \frac{1}{2}\beta_i y_i^2 - \varphi_i y_i \le \frac{\beta_1 c_1^2}{2} + \frac{\beta_2 c_2^2}{2} + \mu. \tag{29}$$

let $\Omega = \{(y_3, y_4, \ldots, y_n) | \sum_{i=3}^{n} \frac{1}{2}\beta_i y_i^2 - \varphi_i y_i \le \frac{\beta_1 c_1^2}{2} + \frac{\beta_2 c_2^2}{2} + \mu\}$. Rewrite (28) as

$$\min_{(y_3,y_4,\ldots,y_n)\in\Omega} \left\{ \begin{array}{l} \min\limits_{(y_1,y_2)} \alpha_1\left(\frac{y_1^2}{2} - c_1 y_1\right) + \alpha_2\left(\frac{y_2^2}{2} - c_2 y_2\right) + \sum\limits_{i=3}^{n} \frac{1}{2}\alpha_i y_i^2 - \eta_i y_i \\ s.t. \quad \beta_1\left(\frac{y_1^2}{2} - c_1 y_1\right) + \beta_2\left(\frac{y_2^2}{2} - c_2 y_2\right) \le \mu - \sum\limits_{i=3}^{n} \frac{1}{2}\beta_i y_i^2 + \varphi_i y_i \end{array} \right\}. \tag{30}$$

As $-\frac{\alpha_1}{\beta_1} = -\frac{\alpha_2}{\beta_2} = \sigma_0$ by Lemma 3, the inner minimization in (30) becomes

$$\begin{aligned} &\min_{(y_1,y_2)} -\sigma_0\beta_1\left(\frac{y_1^2}{2} - c_1 y_1\right) - \sigma_0\beta_2\left(\frac{y_2^2}{2} - c_2 y_2\right) + \sum_{i=3}^{n} \frac{1}{2}\alpha_i y_i^2 - \eta_i y_i \\ &s.t.\ \beta_1\left(\frac{y_1^2}{2} - c_1 y_1\right) + \beta_2(\frac{y_2^2}{2} - c_2 y_2) \le \mu - \sum_{i=3}^{n} \frac{1}{2}\beta_i y_i^2 + \varphi_i y_i, \end{aligned} \tag{31}$$

which can be easily solved as

$$\sum_{i=3}^{n} \frac{1}{2}(\alpha_i + \sigma_0\beta_i)y_i^2 - (\eta_i + \sigma_0\varphi_i)y_i - \sigma_0\mu \tag{32}$$

so that (30) becomes

$$\min_{(y_3,y_4,\ldots,y_n)\in\Omega} \sum_{i=3}^{n} \frac{1}{2}(\alpha_i + \sigma_0\beta_i)y_i^2 - (\eta_i + \sigma_0\varphi_i)y_i - \sigma_0\mu. \tag{33}$$

Since $\alpha_i + \sigma_0\beta_i > 0$ for $i \notin I_0$, problem (33) has a convex objective function of which the only critical point is $\bar{y}_i = \frac{\eta_i+\sigma_0\varphi_i}{\alpha_i+\sigma_0\beta_i}$, $i \in [3:n]$ as defined in (26). Since $\bar{y}$ is an interior feasible solution in (case 4), $(\bar{y}_3, \bar{y}_4, \ldots, \bar{y}_n)$ is also an interior point in Ω and thus the unique minimizer of (33).

We have just seen that any global minimizer of problem (28) must have its i^{th} component equal to $\bar{y}_i$ for $i \in [3:n]$. The remaining is to determine the optimal y_1 and y_2 from (31) by substituting y_i with $\bar{y}_i$ for $i \in [3;n]$. Then problem (31) has an ellipsoid constraint in variables (y_1, y_2). The optimal (y_1, y_2) happens if and only if it is on the boundary of the ellipsoid. The solution

$$\bar{y}_1 = \lim_{\sigma\to\sigma_0^+} \frac{\eta_1 + \sigma\varphi_1}{\alpha_1 + \sigma\beta_1} = \frac{\varphi_1}{\beta_1} = c_1 \quad \text{and} \quad \bar{y}_2 = \lim_{\sigma\to\sigma_0^+} \frac{\eta_2 + \sigma\varphi_2}{\alpha_2 + \sigma\beta_2} = \frac{\varphi_2}{\beta_2} = c_2$$

is the center of the ellipsoid. The boundarification technique moves it to the boundary and solves case 4.

5 Dual of the dual problem

In this section, we show that the dual of the dual problem reveals the hidden convex nature of the primal problem (1). Notice that if $\beta_i \neq 0$, $(\eta_i + \sigma\varphi_i)^2 = \left(\eta_i - \frac{\alpha_i\varphi_i}{\beta_i} + \frac{\varphi_i}{\beta_i}(\alpha_i + \sigma\beta_i)\right)^2$. Define indices sets

$$\begin{aligned} J_0 &= \{i \in [1:n] : \beta_i = 0\}, \\ J_1 &= \left\{i \in [1:n] : \beta_i \neq 0,\ \eta_i - \frac{\alpha_i\varphi_i}{\beta_i} = 0\right\}, \\ J_2 &= \left\{i \in [1:n] : \beta_i \neq 0,\ \eta_i - \frac{\alpha_i\varphi_i}{\beta_i} \neq 0\right\} \end{aligned}$$

with which we can write the above dual problem (6) as

$$\begin{aligned} P_0^{d'} = \sup_{\sigma\geq 0} \Bigg\{ & -\sigma\mu - \textstyle\sum_{i\in J_2}(\eta_i - \frac{\alpha_i\varphi_i}{\beta_i})\frac{\varphi_i}{\beta_i} - \frac{1}{2}\sum_{i\in J_2} \frac{(\eta_i - \frac{\alpha_i\varphi_i}{\beta_i})^2}{\alpha_i + \sigma\beta_i} \\ & - \frac{1}{2}\sum_{i\in J_1\cup J_2} \frac{\varphi_i^2}{\beta_i^2}(\alpha_i + \sigma\beta_i) - \frac{1}{2}\sum_{i\in J_0} \frac{(\eta_i + \sigma\varphi_i)^2}{\alpha_i} \Bigg\} \\ s.t.\ & \alpha_i + \sigma\beta_i > 0,\ i \in J_1 \cup J_2. \end{aligned} \tag{34}$$

Proposition 2 *A Lagrangian dual of Problem (34) is the following linearly constrained convex minimization problem*

$$
\begin{aligned}
P^{dd'} = & -\sum_{i\in J_2}\left(\eta_i - \frac{\alpha_i\varphi_i}{\beta_i}\right)\frac{\varphi_i}{\beta_i} \\
& + \left\{\inf_{z,w}\left\{\sum_{i\in J_1\cup J_2}\alpha_i z_i - \sum_{i\in J_2}|\eta_i - \frac{\alpha_i\varphi_i}{\beta_i}|\sqrt{2z_i + \frac{\varphi_i^2}{\beta_i^2}} + \sum_{i\in J_0}\left\{\eta_i w_i + \frac{\alpha_i w_i^2}{2}\right\}\right\}\right. \\
& \left. s.t.\ \sum_{i\in J_1\cup J_2}\beta_i z_i + \sum_{i\in J_0}\varphi_i w_i \le \mu,\ z_i + \frac{\varphi_i^2}{2\beta_i^2} \ge 0, i\in J_1\cup J_2;\ w_i\in\mathbb{R}, i\in J_0\right\}.
\end{aligned}
\tag{35}
$$

Proof The dual problem (34) can be equivalently written as

$$
\begin{aligned}
& \sup_{\sigma\ge 0}\left\{-\sigma\mu - \sum_{i\in J_2}(\eta_i - \frac{\alpha_i\varphi_i}{\beta_i})\frac{\varphi_i}{\beta_i} - \frac{1}{2}\sum_{i\in J_2}\frac{(\eta_i - \frac{\alpha_i\varphi_i}{\beta_i})^2}{t_i}\right. \\
& \left. -\frac{1}{2}\sum_{i\in J_1\cup J_2}\frac{\varphi_i^2}{\beta_i^2}t_i - \frac{1}{2}\sum_{i\in J_0}\frac{t_i^2}{\alpha_i}\right\} \\
& s.t.\ \alpha_i + \sigma\beta_i = t_i, \quad i\in J_2, \\
& \qquad \alpha_i + \sigma\beta_i = t_i, \quad i\in J_1, \\
& \qquad \eta_i + \sigma\varphi_i = t_i, \quad i\in J_0, \\
& \qquad t_i > 0, \quad i\in J_1\cup J_2.
\end{aligned}
\tag{36}
$$

Let $z_i\in\mathbb{R}, i\in J_2$; $z_i\in\mathbb{R}, i\in J_1$ and $w_i\in\mathbb{R}, i\in J_0$ be the dual multipliers associated with the first, second and third linear equality constraints in (36), respectively. The dual problem of Problem (34) becomes

$$
\begin{aligned}
& -\sum_{i\in J_2}\left(\eta_i - \frac{\alpha_i\varphi_i}{\beta_i}\right)\frac{\varphi_i}{\beta_i} + \inf_{z,w}\left\{\sum_{i\in J_1\cup J_2}\alpha_i z_i + \sum_{i\in J_0}\eta_i w_i + v(z) + v(w)\right. \\
& \left. + \sup_{t_i>0,\ i\in J_1}\left\{-\frac{\varphi_i^2}{2\beta_i^2}t_i - z_i t_i\right\} + \sup_{\sigma\ge 0}\left\{\sigma\left(\sum_{i\in J_1\cup J_2}\beta_i z_i + \sum_{i\in J_0}\varphi_i w_i - \mu\right)\right\}\right\}
\end{aligned}
\tag{37}
$$

where

$$
\begin{aligned}
v(z) &:= \sum_{i\in J_2}\sup_{t_i>0}\left\{-t_i z_i - \frac{\varphi_i^2}{2\beta_i^2}t_i - \frac{\left(\eta_i - \frac{\alpha_i\varphi_i}{\beta_i}\right)^2}{2t_i}\right\}, \\
v(w) &:= \sum_{i\in J_0}\sup\left\{-\frac{t_i^2}{2\alpha_i} - w_i t_i\right\} = \sum_{i\in J_0}\frac{\alpha_i w_i^2}{2}.
\end{aligned}
\tag{38}
$$

The computation of the first inner maximization in (37) with respect to $t_i > 0,\ i\in J_1$, leads to the optimal value zero with the linear constraint $z_i + \frac{\varphi_i^2}{2\beta_i^2} \ge 0$ and that of the second inner

maximization in (37) with respect to $\sigma \geq 0$ leads to the optimal value zero with the linear constraint $\sum_{i\in J_1\cup J_2}\beta_i z_i + \sum_{i\in J_0}\varphi_i w_i \leq \mu$. Finally, for $i \in J_2$ in (38)

$$\sup_{t_i>0}\left\{-t_i z_i - \frac{\varphi_i^2}{2\beta_i^2}t_i - \frac{(\eta_i - \frac{\alpha_i\varphi_i}{\beta_i})^2}{2t_i}\right\} = \begin{cases} -|\eta_i - \frac{\alpha_i\varphi_i}{\beta_i}|\sqrt{2z_i + \frac{\varphi_i^2}{\beta_i^2}}, & \text{if } z_i + \frac{\varphi_i^2}{2\beta_i^2} \geq 0, \\ +\infty, & \text{if } z_i + \frac{\varphi_i^2}{2\beta_i^2} < 0. \end{cases} \tag{39}$$

Substituting the above computations into (37), we get (35). □

Theorem 3 *Under the assumption* (A1)*, the primal problem* (5) *is equivalent to the convex programming problem* (35)*.*

Proof First we can rewrite the primal problem (5) as follows

$$\begin{aligned} P_0' = \min & \sum_{i\in J_1\cup J_2}\left\{\alpha_i\left(\frac{1}{2}\left(y_i - \frac{\varphi_i}{\beta_i}\right)^2 - \frac{\varphi_i^2}{2\beta_i^2}\right) - \left(\eta_i - \frac{\alpha_i\varphi_i}{\beta_i}\right)y_i\right\} + \sum_{i\in J_0}\left(\frac{1}{2}\alpha_i y_i^2 - \eta_i y_i\right) \\ s.t. & \sum_{i\in J_1\cup J_2}\beta_i\left(\frac{1}{2}\left(y_i - \frac{\varphi_i}{\beta_i}\right)^2 - \frac{\varphi_i^2}{2\beta_i^2}\right) - \sum_{i\in J_0}\varphi_i y_i \leq \mu. \end{aligned}$$

Under the assumption (A1), according to the 7 cases discussed in Sect. 3, the global minimizer of the primal problem (5) exists, denoted by y^*. Let $i_0 \in J_2$ be arbitrary. Construct $\bar{y}$ according to:

$$\overline{y}_i = \begin{cases} 2\frac{\varphi_i}{\beta_i} - y_i^*, & \text{if } i = i_0 \in J_2, \\ y_i^*, & \text{if } i \neq i_0. \end{cases}$$

Then $\bar{y}$ is feasible to (5). Since $\sum_{i=1}^n \frac{\alpha_i}{2}(y_i^*)^2 - \eta_i y_i^* \leq \sum_{i=1}^n \frac{\alpha_i}{2}\overline{y_i}^2 - \eta_i\overline{y_i}$ and $i_0 \in J_2$ is arbitrary, we must have $-(\eta_i - \frac{\alpha_i\varphi_i}{\beta_i})(y_i^* - \frac{\varphi_i}{\beta_i}) \leq 0$ for all $i \in J_2$. In addition, by the definition of J_1, the objective function is homogeneous in terms of $(y_i - \frac{\varphi_i}{\beta_i})$ for $i \in J_1$. That is, we can also restrict $(y_i - \frac{\varphi_i}{\beta_i})$ to be nonnegative for $i \in J_1$. Therefore, the problem (5) is further equivalent to

$$\begin{aligned} P_0' = \min & \sum_{i\in J_1\cup J_2}\left\{\alpha_i\left(\frac{1}{2}\left(y_i - \frac{\varphi_i}{\beta_i}\right)^2 - \frac{\varphi_i^2}{2\beta_i^2}\right) - \left(\eta_i - \frac{\alpha_i\varphi_i}{\beta_i}\right)y_i\right\} + \sum_{i\in J_0}\left(\frac{1}{2}\alpha_i y_i^2 - \eta_i y_i\right) \\ s.t. & \sum_{i\in J_1\cup J_2}\beta_i\left(\frac{1}{2}\left(y_i - \frac{\varphi_i}{\beta_i}\right)^2 - \frac{\varphi_i^2}{2\beta_i^2}\right) - \sum_{i\in J_0}\varphi_i y_i \leq \mu, \\ & y_i - \frac{\varphi_i}{\beta_i} \geq 0, \quad \forall i \in J_2 \text{ and } \eta_i - \frac{\alpha_i\varphi_i}{\beta_i} > 0, \\ & y_i - \frac{\varphi_i}{\beta_i} \leq 0, \quad \forall i \in J_2 \text{ and } \eta_i - \frac{\alpha_i\varphi_i}{\beta_i} < 0, \\ & y_i - \frac{\varphi_i}{\beta_i} \geq 0, \quad \forall i \in J_1. \end{aligned}$$

Introducing $z_i = \frac{1}{2}(y_i - \frac{\varphi_i}{\beta_i})^2 - \frac{\varphi_i{}^2}{2\beta_i{}^2}$ for $i \in J_1 \cup J_2$ (which is now a one-to-one map between z_i and y_i) and $w_i = -y_i$ for $i \in J_0$, we exactly obtain the dual of the dual problem (35).

That is, problem (5) is equivalent to (35) via the following nonlinear transformation

$$y_i = \begin{cases} \dfrac{\varphi_i}{\beta_i} + \dfrac{|\eta_i - \dfrac{\alpha_i\varphi_i}{\beta_i}|}{\eta_i - \dfrac{\alpha_i\varphi_i}{\beta_i}} \sqrt{2z_i + \dfrac{\varphi_i^2}{\beta_i^2}}, & \text{if } i \in J_2, \\ \dfrac{\varphi_i}{\beta_i} + \sqrt{2z_i + \dfrac{\varphi_i^2}{\beta_i^2}}, & \text{if } i \in J_1, \\ -w_i, & \text{if } i \in J_0. \end{cases} \tag{40}$$

Theorem 3 explains why the nonconvex problem (QP1QC) has a strong duality with no duality gap under condition (A1).

6 Examples

A few examples are selected to demonstrate the validity of the dual method for solving (QP1QC).

Example 1 (Illustration for case 4) Let $A = \begin{bmatrix} 2 & -1 \\ -1 & 0 \end{bmatrix}$, $B = \begin{bmatrix} 4 & -2 \\ -2 & 2 \end{bmatrix}$, $f = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$, $g = \begin{bmatrix} 4 \\ -1 \end{bmatrix}$, and $\mu = 5$.

We can calculate that $\mathcal{F} = \{\sigma \geq 0 | A + \sigma B \succ 0\} = (0.5, +\infty)$ with $\sigma_0 = 0.5 > 0$. Since $\lim_{\sigma \to 0.5^+} \frac{d\mathrm{P}^d(\sigma)}{d\sigma} = -7.375 < 0$, this is (case 4). One can verify that $\bar{x} = \lim_{\sigma \to 0.5^+}(A + \sigma B)^{-1}(f + \sigma g) = (1.25, 1)^T$ is an interior feasible point. Applying the boundarification technique, we select a vector $\tilde{x} = (1, 2)^T$ from the null space of $A + 0.5B = \begin{bmatrix} 4 & -2 \\ -2 & 1 \end{bmatrix}$, and then solve the quadratic equation

$$\alpha^2 \tilde{x}^T B \tilde{x} + 2\alpha(\bar{x}^T B \tilde{x} - g^T \tilde{x}) + \bar{x}^T B \bar{x} - 2g^T \bar{x} - 2\mu = 0$$

to obtain two roots: 1.92 and -1.92. If $\alpha_0 = 1.92$, $x_1^* = \bar{x} + \alpha_0 \tilde{x} = (3.17, 4.84)^T$ is the global minimizer for the problem with the optimal value -3.625. If $\alpha_0 = -1.92$, $x_2^* = \bar{x} + \alpha_0 \tilde{x} = (-0.67, -2.84)^T$ is the other global minimizer. Since the null space of $A + 0.5B$ is one-dimensional, x_1^* and x_2^* are the only two global minimizers.

Example 2 (Illustration for case 7) Let $A = \begin{bmatrix} 1 & 3 & 2 \\ 3 & -2 & 0 \\ 2 & 0 & 1 \end{bmatrix}$, $B = \begin{bmatrix} 4 & 1 & -2 \\ 1 & 5 & -2 \\ -2 & -2 & 2 \end{bmatrix}$, $f = \begin{bmatrix} 3 \\ 4 \\ -2 \end{bmatrix}$, $g = \begin{bmatrix} -6 \\ 2 \\ 3 \end{bmatrix}$, and $\mu = -6.95$. Note that B is positive definite and $\mathcal{F} = (1.2673 + \infty)$. Since $\lim_{\sigma \to 1.2673^+} \frac{d\mathrm{P}^d(\sigma)}{d\sigma} = +\infty$ and $\lim_{\sigma \to +\infty} \frac{d\mathrm{P}^d(\sigma)}{d\sigma} = 0$, this is (case 7). By Theorem 2, $\bar{x} = \lim_{\sigma \to +\infty}(A + \sigma B)^{-1}(f + \sigma g) = B^{-1}g = (-0.8, 1.4, 2.1)^T$ is the global minimizer for the problem with the optimal value -5.15.

Example 3 (Another illustration for case 7) Let $A = \begin{bmatrix} 5 & 3 & 2 \\ 3 & 6 & 0 \\ 2 & 0 & 4 \end{bmatrix}$, $B = \begin{bmatrix} 3 & 1 & -2 \\ 1 & 3 & -2 \\ -2 & -2 & 2 \end{bmatrix}$, $f = \begin{bmatrix} 0 \\ 3 \\ 1 \end{bmatrix}$, $g = \begin{bmatrix} 3 \\ 1 \\ -2 \end{bmatrix}$, and $\mu = -1.5$.

Since $A \succ 0$ and $B \succeq 0$, we have $\mathcal{F} = [0, +\infty)$. Since $\lim_{\sigma \to 0^+} \frac{d\mathrm{P}^d(\sigma)}{d\sigma} = 6.05 > 0$ and $\lim_{\sigma \to +\infty} \frac{d\mathrm{P}^d(\sigma)}{d\sigma} = 0$, this is (case 7) with $B \succeq 0$. Choose $\sigma' = 5 \in \mathcal{F}$, by (18), we have

$$G = \begin{bmatrix} -0.156 & -0.156 & -0.312 \\ -0.168 & 0.061 & 0.147 \\ 0.117 & -0.214 & 0.052 \end{bmatrix},$$

with $G(A + \sigma' B)G^T = diag(1, 1, 1)$ and $GBG^T = H = diag(0, 0.182, 0.154)$. Then, compute $w = Gf = (-0.78, 0.33, -0.59)^T$ and $s = Gg = (0, -0.74, 0.03)^T$ to obtain that $\bar{x} = G^T(\widetilde{w}^T, (\overline{H}^{-1}\overline{s})^T)^T = (0.83, -0.17, -0.34)^T$ is the global minimizer for the problem with the optimal value 1.9.

7 Conclusion

Quadratic programming is very important not only because many real world applications from physics, statistics, management sciences, etc. can be formulated in quadratic forms, but also because it can be used as a second-order approximation model for more complex systems. Quadratic problems with multiple quadratic constraints are known to be NP-hard. Nevertheless, this paper provides a thorough and comprehensive study, from the assumptions, solution methods, duality, to the hidden convexity for us to understand better about the single quadratic constrained problem. Hopefully, with the new insights it leads a way to study the global optimization for more general nonconvex programming problems.

Acknowledgments This research was undertaken while Y. Xia visited National Cheng Kung University, Tainan, Taiwan. Sheu's research work was sponsored by Taiwan NSC 98-2115-M-006 -010 -MY2. Xia's research work was supported by the fundamental research funds for the central universities under grant YWF-10-02-021 and by National Natural Science Foundation of China under grant 11001006.

References

1. Ben-Tal, A., Teboulle, M.: Hidden convexity in some nonconvex quadratically constrained quadratic programming. Math. Program. **72**, 51–63 (1996)
2. Fang, S.C., Gao, D.Y., Sheu, R.L., Wu, S.Y.: Canonical dual approach to solve 0-1 quadratic programming problems. J. Ind. Manag. Optim. **4**(1), 125–142 (2008)
3. Fang, S.C., Gao, D.Y., Sheu, R.L., Xing, W.: Global optimization for a class of fractional programming problems. J. Global Optim. **45**(3), 337–353 (2009)
4. Fang, S.C., Lin, G.X., Sheu, R.L., Xing, W.: Canonical dual solutions for the double well potential problem (preprint)
5. Fehmers, G.C., Kamp, L.P.J., Sluijter, F.W.: An algorithm for quadratic optimization with one quadratic constraint and bounds on the variables. Inverse Probl. **14**, 893–901 (1998)
6. Fortin, C., Wolkowicz, H.: The trust region subproblem and semidefinite programming. Optim. Methods Softw. **19**(1), 41–67 (2004)
7. Gao, D.Y.: Canonical dual transformation method and generalized triality theory in nonsmooth global optimization. J. Global Optim. **17**, 127–160 (2000)

8. Gao, D.Y.: Canonical duality theory and solutions to constrainted nonconvex quadratic programming. J. Global Optim. **29**, 377–399 (2004)
9. Gay, D.M.: Computing optimal locally constrained steps. SIAM J. Sci. Stat. Comput. **2**(2), 186–197 (1981)
10. Golub, G.H., Von Matt, U.: Quadratically constrained least squares and quadratic problems. Numer. Math. **59**, 186–197 (1991)
11. Hiriart-Urruty, J.-B.: Potpourri of conjectures and open questions in nonlinear analysis and optimization. SIAM Rev. **49**(2), 255–273 (2007)
12. Horn, R., Johnson, C.R.: Matrix analysis. Cambridge University Press, Cambridge (1985)
13. Jeyakumar, V., Srisatkunarajah, S.: Lagrange multiplier necessary conditions for global optimality for non-convex minimization over a quadratic constraint via S-lemma. Optim. Lett. **3**(1), 23–33 (2009)
14. Martínez, J.M.: Local minimizers of quadratic functions on euclidean balls and spheres. SIAM J. Optim. **4**, 159–176 (1994)
15. More, J.J., Sorensen, D.C.: Computing a trust region step. SIAM J. Sci. Stat. Comput. **4**, 553–572 (1983)
16. Palanthandalam-Madapusi , H.J., Van Pelt, T.H., Bernstein, D.S.: Matrix pencils and existence conditions for quadratic programming with a sign-indefinite quadratic equality constraint. J. Global Optim. **45**(4), 533–549 (2009)
17. Pardalos, P.M., Resende, M.G.C.: Interior point methods for global optimization problems. In: Terlaky, T. (ed.) Interior point methods of mathematical programming, pp. 467–500. Kluwer, Dordrecht (1996)
18. Pardalos, P.M., Resende, M.G.C. (eds.): Handbook of applied optimization. Oxford University Press, Oxford (2002)
19. Stern, R.J., Wolkowicz, H.: Indefinite trust region subproblems and nonsymmetric perturbations. SIAM J. Optim. **5**(2), 286–313 (1995)
20. Sturm, J.F., Zhang, S.: On cones of nonnegtive quadratic functions. Math. Oper. Res. **28**(2), 246–267 (2003)
21. Wang, Z., Fang, S.C., Gao, D.Y., Xing, W.: Global extremal conditions for multi-integer quadratic programming. J. Ind. Manag. Optim. **4**(2), 213–225 (2008)
22. Xing, W., Fang, S.C., Gao, D.Y., Sheu, R.L., Zhang, L.: Canonical dual solutions to the quadratic programming over a quadratic constraint, (submitted)
23. Ye, Y., Zhang, S.: New results on quadratic minimization. SIAM J. Optim. **14**(1), 245–267 (2003)