On Compressive Orthonormal Sensing

Yi Zhou, Huishuai Zhang, and Yingbin Liang,

***Abstract*—The Compressive Sensing (CS) approach for recovering sparse signal with orthonormal measurements has been studied under various notions of coherence. However, existing notions of coherence either do not exploit the structure of the underlying signal, or are too complicated to provide an explicit sampling scheme for all orthonormal basis sets. Consequently, there is lack of understanding of key factors that guide the sampling of CS with orthonormal measurements and achieve as low sample complexity as possible. In this paper, we introduce a new notion of π-coherence that exploits both the sparsity structure of the signal and the local coherence. Based on π-coherence, we propose a sampling scheme that is adapted to the underlying true signal and is applicable for CS under all orthonormal basis. Our scheme outperforms (within a constant factor) existing sampling schemes for orthonormal measurements, and achieves a near-optimal sample complexity (within some logarithm factors) for several popular choices of orthonormal basis. Furthermore, we characterize the necessary sampling scheme for CS with orthonormal measurements. We then propose a practical multi-phase implementation of our sampling scheme, and verify its advantage over existing sampling schemes via application to Magnetic Resonance Imaging (MRI) in Medical Science.**

***Index Terms*—Compressive Sensing, Orthonormal, Coherence.**

I. Introduction

With ever growing data size in signal processing applications, taking full number of linear measurements of the signal suffers from both high cost and low efficiency. In the past decade, numerous work focus on recovering signals with reduced number of linear measurements by exploiting special structures of the signals. Among them, the sparsity of the signal is found to be the key factor to reduce the sample complexity [1], [2], [3], [8]. This leads to the development of the so called Compressive Sensing (CS) technique, which has been successfully applied to various applications such as sparse MRI [4], face recognition [5], background subtraction [6] and photo-acoustic tomography [7], just to name a few.

We consider an orthonormal setting of CS, which consists of a pair of orthonormal basis $\mathbf{\Psi}, \mathbf{\Phi} \in \mathbb{C}^{n\times n}$ with basis vectors $\mathbf{\Psi} := [\boldsymbol{\psi}_1^T; ...; \boldsymbol{\psi}_n^T]$, $\mathbf{\Phi} := [\boldsymbol{\phi}_1^T; ...; \boldsymbol{\phi}_n^T]$, and an underlying true signal $\mathbf{x}_0 \in \mathbb{C}^n$. The basis $\mathbf{\Psi}$ provides a sparse representation of the true signal, i.e., $\boldsymbol{\theta}_0 = \mathbf{\Psi}\mathbf{x}_0$ is assumed to be s-sparse. The basis $\mathbf{\Phi}$ corresponds to the space over which the orthonormal measurements $\{\langle \boldsymbol{\phi}_j, \mathbf{x}_0\rangle\}_{j=1}^n$ are taken. Compressive sensing aims to recover the true signal $\mathbf{x}_0$ by taking a few number of measurements. To be specific, denote $\Omega \subset \{1, \ldots, n\}$ as the index set of the sampled orthonormal measurements, and denote $\mathbf{\Phi}_\Omega$ as $\mathbf{\Phi}$ restricted on the rows with index in Ω. Then CS aims to recover $\mathbf{x}_0$ via

$$\textbf{(P)} \quad \min_{\mathbf{x}\in\mathbb{C}^n} \|\mathbf{\Psi}\mathbf{x}\|_1 \quad \text{s.t.} \quad \mathbf{\Phi}_\Omega\mathbf{x} = \mathbf{\Phi}_\Omega\mathbf{x}_0.$$

A fundamental problem is to characterize the number of samples that guarantees problem **(P)** uniquely recover the underlying signal. It has been well understood that sample complexity is highly related to sampling schemes, and three types of sampling schemes have been considered so far.

(a) ***Uniform*** sampling is considered in [9], in which each orthonormal measurement $\phi_j, j = 1, \cdots, n$ is equally likely to be taken. It has been shown via a RIPless theory that the sample complexity depends on the following notion of coherence

$$\text{(Mutual Coherence)} \quad \mu(\mathbf{\Psi}, \mathbf{\Phi}) := \max_{1\le i,j\le n} |\langle \boldsymbol{\psi}_i, \boldsymbol{\phi}_j\rangle|^2,$$

and is of the order $\mathcal{O}(\mu(\mathbf{\Psi}, \mathbf{\Phi}) sn \log n)$. In the low mutual coherence regime (i.e. $\mu(\mathbf{\Psi}, \mathbf{\Phi}) = \mathcal{O}(1/n)$) this sample complexity is order-wise optimal up to a logarithm factor. However, the sample complexity scales as large as the trivial order $\mathcal{O}(n \log n)$ if any pair of the basis vectors are exactly aligned (i.e. maximally coherent) with each other.

(b) ***Measurement-adaptive*** sampling scheme was studied in [10], [11], in which each orthonormal measurement $\boldsymbol{\phi}_j$ is sampled with probability being proportional to the so-called *local coherence*, i.e.,

$$\text{(Local Coherence)} \quad \mu(\mathbf{\Psi}, \boldsymbol{\phi}_j) := \max_{1\le i\le n} |\langle \boldsymbol{\psi}_i, \boldsymbol{\phi}_j\rangle|^2. \qquad (1)$$

The local coherence is determined by the basis pair and can be viewed as mutual coherence localized at the j-th measurement. It was shown in [10] that the sample complexity is of the order $\mathcal{O}(\sum_j \mu(\mathbf{\Psi}, \boldsymbol{\phi}_j) s \log^3 s \log n))$ via a RIP argument, and was further improved to $\mathcal{O}(\sum_j \mu(\mathbf{\Psi}, \boldsymbol{\phi}_j) s \log s \log n))$ via a RIPless argument in [11].

(c) ***Measurement-and-signal adaptive*** sampling scheme was proposed in [11]. There, the sampling of the orthonormal measurements is based on some complicated notions of coherence [1], which are related to both the support information of the underlying true signal and the basis pair. Only under $\mathbf{\Psi}$ being the bivariate Haar wavelet basis and $\mathbf{\Phi}$ being the discrete Fourier basis, the sampling scheme is given in explicit form, and sample complexity is shown to be on the order $\mathcal{O}(s \log s \log n)$.

The focus of this paper is to resolve remaining issues for the third type of sampling scheme. The notion of coherence introduced in [11] does not yield a universal sampling scheme for recovering the true signal in all orthonormal basis pairs, and may not be the real underlying quantity that governs

Yi Zhou, Huishuai Zhang and Yingbin Liang are with the Department of Electrical Engineering and Computer Science, Syracuse University, Syracuse, NY, e-mail: yzhou35@syr.edu.

Manuscript received April 19, 2015; revised August 26, 2015.

[1]For simplicity we do not present them here.

the sample complexity. Hence, our goal is to find out the fundamental notion of coherence to guide the design of the sampling scheme, which 1) is adapted to the underlying true signal; 2) is generally applicable to all basis pairs; and 3) yields better sample complexity.

A. Main Contributions

We propose the new notion of π-coherence (see Definition 1) to capture how an orthonormal measurement $\boldsymbol{\phi}_j$ is coherent with the subspace where the true signal lies. Based on the notion of π-coherence, we propose a Bernoulli sampling scheme that is generally applicable for all pairs of orthonormal basis.

Our sampling scheme reveals the relationship among sample complexity, sparsity structure of the true signal and coherence pattern of the basis pair. We further show that our scheme achieves lower sample complexity (within a constant factor) than uniform sampling [8], measurement-adaptive sampling [10] and measurement-and-signal adaptive sampling [11]. For several popular choices of orthonormal basis pairs, we show that our sampling scheme achieves a near-optimal sample complexity (within some logarithm factors). Furthermore, we characterize the fundamental necessary sampling scheme for CS with orthonormal measurements.

Our technical proof introduces a weighted infinity norm to control the concentration bounds in a tighter way. Consequently, these new concentration bounds avoid involving complicated notions of coherence as those in [11], and lead to our notion of coherence that has an intuitive physical meaning on the sparsity structure of the signal and the coherence pattern of the orthonormal basis pair.

For practical applications, we propose a multi-phase version of our sampling scheme, which iteratively learns the subspace information of the true signal and and updates the sampling scheme in the next phase.

B. Related Work

Various kinds of sampling schemes have been proposed for CS with orthonormal measurements. Some schemes are empirically oriented [12], [13], [14], [15], where the schemes were demonstrated to be useful in specific applications. Closely to our work are schemes theoretically oriented such as uniform sampling [8], [9], mixture of full sampling in finite regime and uniform sampling in asymptotic regime [16], measurement-adaptive sampling [10], [11], and measurement-and-signal adaptive sampling [11]. Our work falls into the last type of sampling scheme that is adapted to both the underlying true signal and the basis pair. Differently from [11] which explicitly derives sampling scheme only for a specific basis pair, our scheme is generally applicable to all orthonormal basis pairs and yields a better sample complexity.

There is another series of work on adaptive compressive sensing [17], [18], [19]. There, the focus was on recovering the support of the signal via adapting the distribution from which the random measurements are generated, and do not involve any notion of coherence. In contrast, we are interested in recovering the signal by sampling deterministic orthonormal measurements based on our notion of coherence.

Adaptive sampling has also been considered in matrix-related problems by exploiting the leverage score [20], [21], [22], which is related to the projections of canonical unit vectors onto row/column spaces of the matrix. In particular, [22] proposed a sampling scheme that is adapted to the leverage scores of each entry of the matrix for matrix completion problems.

C. Organization and Notations

This paper proceeds as follows. Section II introduces the signal model and our notion of coherence; In Section III we present our sampling scheme and analyze the performance guarantee of the sampling scheme; In Section IV we propose a multi-phase algorithm to illustrate how to implement our sampling scheme in practice; In Section V we present various numerical experiments to demonstrate our theoretical characterization, and finally we conclude the paper in Section VI.

We adopt the following notations in the paper
$\mathbf{x}_0$: the underlying signal to recover;
$\boldsymbol{\Psi}, \boldsymbol{\Phi}$: the representation basis and measurement basis;
Ω: the index set of sampled measurements;
$\boldsymbol{\theta}_0$: the representation of $\mathbf{x}_0$ in basis $\boldsymbol{\Psi}$;
S: support set of $\boldsymbol{\theta}_0$;
$\boldsymbol{\Pi}$: the subspace of $\boldsymbol{\Psi}$ where $\mathbf{x}_0$ lies in.
$\|\cdot\|_p$: Euclidean l_p norm;
$\|\cdot\|$: spectral norm of an operator;
$\mathcal{I}, \mathcal{H}, \mathcal{F}$: the canonical basis, bivariate Haar wavelet basis and discrete Fourier basis, respectively.

II. Signal Model and π-Coherence

Suppose that the representation of signal $\mathbf{x}_0$ in basis $\boldsymbol{\Psi}$, i.e. $\boldsymbol{\theta}_0 = \boldsymbol{\Psi}\mathbf{x}_0$, is s-sparse. Denote S as the support set of $\boldsymbol{\theta}_0$. Then the following two complementary subspaces are well defined

$$\boldsymbol{\Pi} : \mathrm{span}\{\boldsymbol{\psi}_j \mid j \in S\}, \quad \boldsymbol{\Pi}^{\perp} : \mathrm{span}\{\boldsymbol{\psi}_j \mid j \in S^c\}. \tag{2}$$

It follows that $\mathbf{x}_0 \in \boldsymbol{\Pi}$, and $\dim(\boldsymbol{\Pi}) = s$. Following [9], [23], we consider the following random sign model of the underlying signal.

Assumption 1. *The sign of the non-zero entries of $\boldsymbol{\theta}_0$ is distributed independently from their locations as*

$$\forall j \in S, \quad \mathrm{sgn}(\boldsymbol{\theta}_0)_j = \begin{cases} 1, & \textit{w.p. } 1/2 \\ -1, & \textit{w.p. } 1/2 \end{cases}$$

We also denote $\mathrm{P}_{\boldsymbol{\Pi}}, \mathrm{P}_{\boldsymbol{\Pi}^{\perp}}$ as the projection operators onto $\boldsymbol{\Pi}, \boldsymbol{\Pi}^{\perp}$, respectively. Now we propose the following notion of coherence, which plays a fundamental role in our design of the sampling scheme and characterization of its performance guarantee.

Definition 1. *(π-coherence) The π-coherence of measurement $\boldsymbol{\phi}_j$ w.r.t. space $\boldsymbol{\Pi}$ is defined as*

$$\pi(\boldsymbol{\Pi}, \boldsymbol{\phi}_j) := \max\{\|\mathrm{P}_{\boldsymbol{\Pi}}\boldsymbol{\phi}_j\|_2^2, \|\mathrm{P}_{\boldsymbol{\Pi}}\boldsymbol{\phi}_j\|_2 \|\boldsymbol{\Psi}\mathrm{P}_{\boldsymbol{\Pi}^{\perp}}\boldsymbol{\phi}_j\|_{\infty}\}.$$

In the above definition of π-coherence, the first term (i.e. l_2^2) corresponds to the "energy" of each measurement $\boldsymbol{\phi}_j$ onto $\boldsymbol{\Pi}$. Intuitively, a larger value of this term implies that measurement $\boldsymbol{\phi}_j$ is more aligned with $\boldsymbol{\Pi}$ where the underlying signal lies in, and hence can retrieve more information of the signal. Moreover, the l_2^2 structure of this term also reveals the sparsity structure of the signal.

To be specific, consider *arbitrary* orthogonal decomposition of $\boldsymbol{\Pi}$ into p subspaces $\{\boldsymbol{\Pi}_l\}_{l=1}^p$ with corresponding dimensions $\{s_l\}_{l=1}^p$. Then for all j, one has

$$\boldsymbol{\Pi} = \bigoplus_{l=1}^p \boldsymbol{\Pi}_l, \quad \|\mathrm{P}_{\boldsymbol{\Pi}}\boldsymbol{\phi}_j\|_2^2 = \sum_{l=1}^p \|\mathrm{P}_{\boldsymbol{\Pi}_l}\boldsymbol{\phi}_j\|_2^2. \tag{3}$$

With regard to the full support $\dim(\boldsymbol{\Pi}) = s$ of the signal, $\{\dim(\boldsymbol{\Pi}_l) = s_l\}_{l=1}^p$ can be viewed as a sparsity structure into subspace decompositions. Correspondingly, each term $\|\mathrm{P}_{\boldsymbol{\Pi}_l}\boldsymbol{\phi}_j\|_2^2$ measures the alignment of $\boldsymbol{\phi}_j$ with subspace $\boldsymbol{\Pi}_l$, and their summation yields the full coherence.

The second term (i.e. $l_2 \cdot l_\infty$) further contains an l_∞ component, which can be interpreted as (square root of) the local coherence of $\boldsymbol{\phi}_j$ with regard to $\boldsymbol{\Pi}^\perp$, i.e.

$$\|\boldsymbol{\Psi}\mathrm{P}_{\boldsymbol{\Pi}^\perp}\boldsymbol{\phi}_j\|_\infty = \sqrt{\mu(\boldsymbol{\Pi}^\perp, \boldsymbol{\phi}_j)} := \max_{i\in S^c} \ |\langle \boldsymbol{\psi}_i, \boldsymbol{\phi}_j\rangle|. \tag{4}$$

Intuitively, such local coherence is involved because one need to recover zeros on $\boldsymbol{\Pi}^\perp$ (recall that $\mathbf{x}_0 \in \boldsymbol{\Pi}$) to guarantee fully correct recovery of the signal.

In summary, our notion of π-coherence incorporates the general sparsity structure of the signal as in eq. (3) and the local coherence as in eq. (4). This is different from the mutual coherence [8], [9], the asymptotic coherence [16], and the local coherence [10], [11], none of which exploits the sparsity structure of the signal. Moreover, it has a simpler form than the notions of coherence introduced in [11], and is more general in subspace decomposition as eq. (3). In the next section, we demonstrate that our notion of coherence serves more fundamental purpose in characterizing sufficient and necessary conditions on sample complexity for performance guarantee.

III. Main Results

In this section, we present our main results with proofs provided in the appendix.

A. Sampling Scheme and Performance Guarantee

We propose the following Bernoulli sampling, i.e., measurement $\boldsymbol{\phi}_j$ is taken with probability given by

$$\mathbb{P}(j \in \Omega) \sim \mathrm{Bernoulli}(p_j), \quad \text{for } j = 1, \ldots, n. \tag{5}$$

The sampling probability is set based on the π-coherence as we state in the following theorem.

Theorem 1. *Let Assumption 1 hold and fix any pair of orthonormal basis $\boldsymbol{\Psi}, \boldsymbol{\Phi}$. Then $\mathbf{x}_0$ is the unique minimizer of **(P)** with probability at least $1 - s^{-\sqrt{C_0}}$, provided that*

$$p_j \geq \min\{C_0 \pi(\boldsymbol{\Pi}, \boldsymbol{\phi}_j) \log s \log n, \ 1\} \tag{6}$$

$$p_j \geq 1/s^{20}. \tag{7}$$

where $C_0 > 1$ is a universal constant.

In fact, the constraint in eq. (7) is to avoid singularity in the proof, and the power of s can be raised up further by slightly increasing C_0. Effectively, eq. (6) determines the sampling scheme, which requires to sample the $\boldsymbol{\phi}_j$ measurement with probability proportional to the corresponding π-coherence $\pi(\boldsymbol{\Pi}, \boldsymbol{\phi}_j)$. Thus, inheriting from π-coherence, our sampling scheme exploits the sparsity structure of the signal in eq. (3) and the local coherence in eq. (4). Intuitively, if measurement $\boldsymbol{\phi}_j$ is more aligned with the signal space $\boldsymbol{\Pi}$, then such measurement is sampled with higher probability.

We further note that our sampling scheme is generally applicable to all pairs of orthonormal basis. This is in contrast to the same type of alternative sampling scheme in [11] that also exploit both signal and basis information based on two complicated notions of coherence. There, the sampling scheme in general do not have closed form. Only for the MRI example (with Haar wavelet basis and Fourier basis), an explicit scheme is obtained that samples uniformly among the Fourier measurements in different dyadic levels based on the sparsity of the signal in different levels of wavelet basis. In contrast, our sampling scheme exploits the fully decomposable sparsity structure in eq. (3) for all orthonormal basis.

Theorem 1 also characterizes the expected number of measurements, i.e., $\sum_{j=1}^n p_j$, that guarantees correct recovery of the signal. Clearly, the π-coherence plays a central role similar to other notions of coherence in determining sample complexity. Thus, in the following theorem, we compare π-coherence with other notions of coherence, which thus yields comparison of sample complexity among the corresponding sampling schemes.

Theorem 2. *For $j = 1, \ldots, n$, the following inequality that compares various notions of coherence holds:*

$$\pi(\boldsymbol{\Pi},\boldsymbol{\phi}_j) \leq \|\boldsymbol{\Psi}\mathrm{P}_{\boldsymbol{\Pi}}\boldsymbol{\phi}_j\|_1 \sqrt{\mu(\boldsymbol{\Psi},\boldsymbol{\phi}_j)} \leq s\mu(\boldsymbol{\Psi},\boldsymbol{\phi}_j) \leq s\mu(\boldsymbol{\Psi},\boldsymbol{\Phi}).$$

Remark 1. *Theorem 2 suggests that our sampling scheme based on π-coherence has the lowest sample complexity (within a constant factor) in comparison to the basis-and-signal adaptive sampling schemes in [11] (based on coherences lower bounded by the second term above), the basis adaptive sampling scheme [10] (based on local coherence in the third term above), and the uniform sampling [9] (with sample complexity captured by the maximum coherence in the last term above).*

The proof of Theorem 1 is based on the convex duality argument and the so called golfing scheme originated from matrix completion literature [24]. The notion of coherence and the corresponding sampling scheme naturally arise to control the concentration bounds. Specially, novelty of our proof lies in introduction of the following weighted infinity norm of a vector $\mathbf{w}$

$$\|\mathbf{w}\|_{\boldsymbol{\Phi},\infty} := \max_j \frac{|\langle \mathbf{w}, \boldsymbol{\phi}_j\rangle|}{\|\mathrm{P}_{\boldsymbol{\Pi}}\boldsymbol{\phi}_j\|_2}, \tag{8}$$

which together with the π-coherence allows to control the following concentration bounds in Lemma 5 and Lemma 6

with high probability.

$$\|\mathbf{\Psi}_{S^c}(R_{\Omega_k} - I)\mathbf{w}\|_\infty \leq \|\mathbf{w}\|_{\mathbf{\Phi},\infty}/\sqrt{C_0}, 1 \leq k \leq k_0,$$
$$\|(\mathrm{P}_{\mathbf{\Pi}}\mathrm{R}_{\Omega_k}\mathrm{P}_{\mathbf{\Pi}} - \mathrm{P}_{\mathbf{\Pi}})\mathbf{w}\|_{\mathbf{\Phi},\infty} \leq \|\mathbf{w}\|_{\mathbf{\Phi},\infty}/2, 1 \leq k \leq k_0.$$

The above $\{\Omega_k\}_{k=1}^{k_0}$ are decomposed Bernoulli random models of Ω, and the second bound is further tightened in the detailed proof. We refer to the appendix for more details of the parameters and the way to apply the bounds. Consequently, these new concentration bounds help to prove the high probability guarantee and avoid involving complicated notions of coherence. In fact, the idea of using weighted norm to control the concentration bounds has also been explored in low-rank matrix completion problems [25], [26]. However, their weighted norms are different, which depend on the incoherence property of the row and column spaces of low rank matrices.

B. Complexity, Sparsity and Coherence Pattern

In this subsection, we understand further the sample complexity of our sampling scheme. Due to eq. (6), the sample complexity, i.e. $\sum_{j=1}^{n} p_j$, is upper bounded by the summation of the two terms involved in π-coherence. The summation of the first l_2^2 term provides a clear view of the sparsity structure of the signal. Specifically, consider the general orthogonal decomposition in eq. (3), one has

$$\sum_{j=1}^{n} \|\mathrm{P}_{\mathbf{\Pi}}\boldsymbol{\phi}_j\|_2^2 = \sum_{l=1}^{p}\sum_{j=1}^{n} \|\mathrm{P}_{\mathbf{\Pi}_l}\boldsymbol{\phi}_j\|_2^2 = \sum_{l=1}^{p} s_l = s. \quad (9)$$

That is, the summation of the first term is a collection of the sparsity of the signal in each subspace $\mathbf{\Pi}_l$, and in total contributes the whole sparsity of the signal to the sample complexity. The second $l_2 \cdot l_\infty$ term is related to both the sparsity structure and the local coherence, and thus its summation depends on both the sparsity structure and the pattern of the local coherence.

We next demonstrate that the above characterization of the sample complexity for all basis pairs $\mathbf{\Psi}$ and $\mathbf{\Phi}$ reduces to the state-of-the-art results in specific examples. In particular, we wish to understand more explicitly how interaction of the two terms in π-coherence yields the sample complexity via these specific examples. In the sequel, $\mathcal{I}, \mathcal{H}, \mathcal{F}$ denote the canonical basis, bivariate Haar wavelet basis and discrete Fourier basis, respectively.

Example 1. *(maximally incoherent)*
$\mathbf{\Psi} = \mathcal{I}, \mathbf{\Phi} = \mathcal{F}$, and $\mathcal{I}\mathbf{x}_0$ is s-sparse.

This example arises in applications that aims to recover sparse signals via a number of Fourier measurements [27]. It is well known that this pair of basis have an incoherent local coherence pattern, i.e. for all $i, j \in \{1, \cdots, n\}$

$$\mu(\boldsymbol{\psi}_i, \boldsymbol{\phi}_j) = 1/n. \quad (10)$$

Then a simple calculation yields that $\forall j, \pi(\mathbf{\Pi}, \boldsymbol{\phi}_j) \equiv s/n$, and consequently our sampling scheme becomes the optimal uniform sampling scheme for this example with sample complexity being of the order $\mathcal{O}(s \log s \log n)$.

Example 2. *(highly coherent)*
$\mathbf{\Psi} = \mathcal{H}, \mathbf{\Phi} = \mathcal{F}$, and $\mathcal{H}\mathbf{x}_0$ is s-sparse.

This example arises in the popular Magnetic Resonance Imaging (MRI) application [28], [11], [10]. There, $\mathbf{x}_0$ corresponds to medical image that can be sparsely represented in the Haar wavelet basis, and the physical device takes Fourier measurements of the signal. While the mutual coherence is as high as $\mu(\mathcal{H}, \mathcal{F}) = 1$, this pair of basis have a well behaved local coherence pattern. To be specific, we assume without of loss of generality that $n = 2^p$ and introduce a dyadic partition of the index set $\{1, \cdots, n\}$ into $p+1$ levels, i.e. $L_0 = \{1\}, L_1 = \{2\}, L_2 = \{3, 4\}, \cdots, L_p = \{n/2+1, \cdots, n\}$. For any $j \in \{1, \cdots, n\}$ we also define $k(j)$ as the dyadic level that j belongs to, i.e., $j \in L_{k(j)}$. Then for all $i, j \in \{1, \cdots, n\}$, the following pattern of local coherence follows from Lemma D.1 in [11]

$$\mu(\boldsymbol{\psi}_i, \boldsymbol{\phi}_j) \leq 2^{-k(j)} 2^{-|k(j)-k(i)|}. \quad (11)$$

Intuitively, the pair of basis are more incoherent in the asymptotic region (i.e., large i, j), and this observation motivates the introduction of the asymptotic coherence in [16]. By exploiting the coherence pattern in eq. (11), we can control the summation of the second term in π-coherence, and characterize the following sample complexity for our sampling scheme with the proof given in Section C.

Lemma 1. *With the coherence pattern in eq.* (11)*, one yields that*

$$\textstyle\sum_{j=1}^{n} \|\mathrm{P}_{\mathbf{\Pi}}\boldsymbol{\phi}_j\|_2 \|\mathbf{\Psi}\mathrm{P}_{\mathbf{\Pi}^\perp}\boldsymbol{\phi}_j\|_\infty \lesssim \sum_{l=0}^{p} \sqrt{s_l}, \quad (12)$$

where $\{s_l\}_{l=0}^{p}$ are sparsity of the signal in different levels of wavelet basis. Consequently, the sample complexity of eq. (6) *for Example 2 is of the order $\mathcal{O}(s \log s \log n)$.*

In comparison, our sampling scheme in eq. (6) achieves better sample complexity than [10] for Example 2 (which is on the order $\mathcal{O}(s \log^3 s \log^2 n)$), and recovers the same order-level sample complexity as [11] for Example 2. However, the first inequality in Theorem 2 implies that our sample complexity is in fact lower than that in [11], although they are on the same order. One possible reason is due to the fact that sampling scheme in [11] samples uniformly among the Fourier measurements over the same dyadic levels, leading to a block structure of the sampled Fourier frequencies which does not fit the spectrum of natural images. The advantage of our sampling scheme turns out to be more prominent in experiments in Section V.

Example 3. *(maximally coherent)*
$\mathbf{\Psi} = \mathbf{\Phi}$, and $\mathbf{\Psi}\mathbf{x}_0$ is s-sparse.

That is, the basis pair are exactly aligned with

$$\mu(\boldsymbol{\psi}_i, \boldsymbol{\phi}_j) = \mathbb{I}\{i = j\}, \quad (13)$$

where $\mathbb{I}\{\cdot\}$ is the binary indicator. Thus, π-coherence satisfies $\pi(\mathbf{\Pi}, \boldsymbol{\phi}_j) = \mathbb{I}\{\boldsymbol{\phi}_j \in \mathbf{\Pi}\}$, and Theorem 1 implies that with $C_0 \log s \log n > 1$, $p_j = 1$ if $\boldsymbol{\phi}_j \in \mathbf{\Pi}$, and $p_j = 0$ otherwise. Namely, sample only the measurements that are in $\mathbf{\Pi}$. This is intuitive because only if each basis vector in the signal space

$\mathbf{\Pi}$ is sampled, the signal component over that basis vector can be recovered. Clearly, the sample complexity is $\mathcal{O}(s)$, which is in contrast to the basis adaptive sampling scheme in [10] that samples all n measurements for this example. This justifies the advantage of incorporating support structure of the signal into our sampling scheme.

The above three examples represent different types of coherence patterns. Our sampling scheme based on π-coherence for general basis pairs achieves near optimal sample complexity when specialized to these cases.

C. Lower bound for orthonormal sensing

In this subsection, we present a lower bound for orthonormal sensing and further justify the fundamental role that π-coherence plays. The following theorem states the result.

Theorem 3. *Fix and pair of orthonormal basis Ψ, Φ. If for all $j = 1, \cdots, n$*

$$p_j \leq \|\mathrm{P}_{\mathbf{\Pi}} \boldsymbol{\phi}_j\|_2^2, \tag{14}$$

*Then there are infinitely many solutions other than $\mathbf{x}_0$ that achieves the same objective value of problem **(P)** with high probability.*

An intuitive explanation for the above converse is that the number of measurements sampled by eq. (14) can be less than its expectation s, hence problem (P) becomes under-determinant for uniquely identifying an s-sparse signal. Clearly, one can see that our sampling scheme in Theorem 1 has a small gap, i.e., the additional $l_2 l_\infty$ term in the $\boldsymbol{\pi}$-coherence, compared to the above necessary sampling scheme (except for the logarithm terms). This might be an artifact of the duality argument and the use of weighted norm in the proof, but fortunately it does not affect the order of sample complexity of the popular examples discussed above. It is unclear how to fix this small gap and we leave it for future study.

IV. Multi-Phase Sampling Scheme

Our adaptive sampling scheme requires the knowledge of the signal space $\mathbf{\Pi}$, which is usually unknown a priori. We now propose a multi-phase implementation of our sampling scheme, where the signal space is estimated via previous phases, and is then exploited in subsequent phases for adaptive sampling. More specifically, suppose the multi-phase sampling scheme consists of P phases, and each phase takes m measurements. Clearly, we require $mP \leq n$. Then the main steps in each phase $p = 1, \cdots, P$ are described in Algorithm 1.

We note that one can of course use other sampling schemes in the initial phase than the uniform sampling scheme. We use uniform sampling scheme to emphasize the advantage of our adaptive sampling scheme in subsequent phases.

(Trim the $\mathbf{\Pi}$ space): In the multi-phase sampling scheme, estimation of $\mathbf{\Pi}, \mathbf{\Pi}^\perp$ from the recovered signal $\mathbf{x}_p$ is based on the definition in eq. (2). However, in practice, all entries of the representation vector $\boldsymbol{\theta}_p := \mathbf{\Psi}\mathbf{x}_p$ can be strictly non-zero with majority components having very small magnitudes. Such phenomenon yields $\mathbf{\Pi}$ that span over the entire space, and the sampling scheme in eq. (6) reduces to the uniform sampling scheme and hence not signal adaptive. To overcome this issue, we relax the criterion in eq. (2) by removing the entries with small magnitudes. To be specific, we set a threshold of the magnitude $\tilde{\boldsymbol{\theta}} > 0$ with which $\mathbf{\Pi}$ is set to span over $\mathrm{span}\{\boldsymbol{\psi}_j, |(\boldsymbol{\theta}_p)_j| > \tilde{\boldsymbol{\theta}}\}$. The threshold is searched among all magnitudes of entries to satisfy the following criterion

$$\|\mathrm{P}_{\mathbf{\Pi}} \mathbf{x}_p\|_2^2 = 0.95 \|\mathbf{x}_p\|_2^2. \tag{15}$$

That is, $\tilde{\boldsymbol{\theta}}$ thresholds out those entries with small magnitudes and keeps 95% of the energy of $\mathbf{x}_p$. The resulting $\mathbf{\Pi}$ is then spanned by the small amount of basis vectors with coefficients of high magnitudes, providing accurate approximation with a much smaller dimension. We note that one can of course change 95% to other quantities, and it controls the approximation error via $\|\mathrm{P}_{\mathbf{\Pi}} \mathbf{x}_p - \mathbf{x}_p\|_2^2 = 5\%$.

Algorithm 1 Multi-phase Sampling for CS

Initial phase $p = 1$:

1). Uniform sampling scheme $\overset{\text{get}}{\rightarrow} \Omega_1$ with m measurements.

2). Solve problem **(P)** with $\Omega \overset{\text{set}}{=} \Omega_1$ for the solution $\mathbf{x}_1$, and evaluate its $\mathbf{\Pi}, \mathbf{\Pi}^\perp$ according to eq. (2).

Phase $p = 2, \cdots, P$:

1). Evaluate p_j in eq. (6) with the $\mathbf{\Pi}, \mathbf{\Pi}^\perp$ evaluated in previous phase for unsampled $j \notin \cup_{k=1}^{p-1} \Omega_k$; Set $p_j = 0$ for the sampled measurements.

2). Normalize $\{p_j\}$ to a probability distribution, according to which sample m measurements sequentially to get Ω_p. After each sample j we set $p_j = 0$ and renormalize $\{p_j\}$.

3). Solve problem **(P)** with $\Omega \overset{\text{set}}{=} \cup_{k=1}^{p} \Omega_k$ for the solution $\mathbf{x}_p$, and evaluate its $\mathbf{\Pi}, \mathbf{\Pi}^\perp$ according to eq. (2).

Output $\mathbf{x}_P$.

V. Numerical Experiments

In this section, we study the performance of our basis-and-signal adaptive sampling scheme defined in eq. (6), and compare it with the uniform sampling [8], basis-adaptive sampling [10], and uniform by level sampling (a specific basis-and-signal adaptive scheme) [11]. Since our scheme and the uniform by level sampling [11] are signal-dependent, we implement them by the multi-phase scheme described in Algorithm 1. Since some of the above sampling schemes are proposed only for special case, we use the common Example 2 as the test example. All experiments are repeated five times and the average is reported as the final result.

A. Total Budget v.s. Relative Error

We first consider the noisefree setting, and apply the sampling schemes of interest to recover a 256×256 MRI brain image. The budget of measurements is set to be $20\%, 25\%, \cdots, 70\%$ of the total dimensions, respectively. The number P of phases for the multi-phase sampling schemes is set to be two, which is justified in our subsequent experiment in Section V-B. Denote $\mathbf{x}^*$ as the original image and $\hat{\mathbf{x}}$ as the recovered image. We report the relative error $\frac{\|\mathbf{x}^* - \hat{\mathbf{x}}\|_2}{\|\mathbf{x}^*\|_2}$.

Figure 1 plots how the relative error reduces as the total budget increases for four sampling schemes. It can be seen that the uniform sampling [8] and the uniform by level scheme [11] suffer from high relative error and instability, while the basis-adaptive scheme in [10] and our scheme provide more accurate and stable recovery. In particular, our scheme outperforms all other three schemes. It is superior over the uniform and basis adaptive sampling due to the fact that it adapts to both the local coherence of basis and the signal space $\mathbf{\Pi}$. Although the uniform by level sampling also adapts to the signal information, adaption is limited only across levels, not over the entire basis vectors as in our scheme.

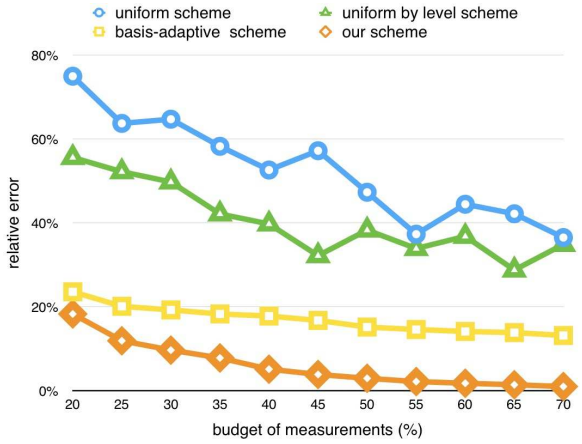

Fig. 1. Comparison of relative errors of four sampling schemes to recover a medical image as the number of measurements increases.

To provide a better illustration, we present in Figure 2 the recovered images and their corresponding sampling masks of the sampled Fourier frequencies for the setting with 30% budget of total dimensions.

One can see that the image recovered by our sampling scheme has the best quality. Note that we have shifted the Fourier basis to center the spectrum and the sampling masks. One can see that the sampling mask of the uniform by level scheme in [11] has a block structure, which does not fit the spectrum of the original signal well. The sampling mask of the basis-adaptive scheme in [10] is a good approximation of the spectrum of practical smooth images, but is generated by a fixed probability distribution that is not adapted to the underlying signal. Our sampling mask is the mixture of a uniform mask (in the initial phase) and a mask generated by adaptive sampling scheme (in the second phase). Despite its uniform part of the mask, one can see that the mask of our scheme is well adapted to the spectrum of the original signal.

B. Number of Phases

A key parameter that affects multi-phase sampling schemes is the number P of phases. Intuitively, larger P allows more measurements to be sampled adaptively, but requires more computation for estimating $\mathbf{\Pi}, \mathbf{\Pi}^{\perp}$ and $\{p_j\}$ in each phase. Hence, a good choice of P can well balance the recovery error and computational load. We apply our multi-phase sampling scheme with 30% measurements of the total dimensions to recover an MRI image, and set the number of phases to be $P = 2, \cdots, 5$, respectively. Table I shows the relationship between the number of phases and the corresponding relative error. One can see that for more than 2 phases the relative error stays at the same level. This suggests that two phases are sufficiently good to achieve a low relative error for our experiments.

Number of phase	2	3	4	5
Relative Error	0.095	0.085	0.089	0.082

TABLE I
COMPARISON OF THE PERFORMANCE OF OUR MULTI-PHASE SAMPLING SCHEME UNDER DIFFERENT NUMBERS OF PHASES

VI. Conclusion

We propose a Bernoulli sampling scheme for CS with orthonormal basis, which is applicable to all orthonormal basis pairs. The scheme is based on π-coherence that exploits both the support structure of the signal and the local coherence. Our sampling scheme reveals the relationship between sample complexity and sparsity of the signal as well as coherence pattern of the basis, and achieves lower sample complexity (within a constant factor) than that of existing basis and signal adaptive sampling schemes. Furthermore, we characterize the neccessary sampling scheme for orthonormal measurements. We also propose a practical multi-phase implementation of our sampling scheme, and demonstrate its advantage over other existing sampling schemes via experiments. We anticipate that such signal-dependent adaptive sampling together with multi-phase implementation can be useful for other high dimensional problems with limited budget of measurements.

References

[1] D. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, pp. 1289–1306, 2006.
[2] E. Candès and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?" *IEEE Transactions on Information Theory*, vol. 52, no. 12, pp. 5406–5425, 2006.
[3] E. Candès, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006.
[4] M. Lustig, D. Donoho, J. Santos, and J. Pauly, "Compressed sensing mri," in *IEEE Signal Processing Magazine*, 2007.
[5] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.
[6] V. Cevher, A. Sankaranarayanan, M. Duarte, D. Reddy, and R. Baraniuk, "Compressive sensing for background subtraction," in *European Conference on Computer Vision*, 2008, pp. 155–168.
[7] J. Provost and F. Lesage, "The application of compressed sensing for photo-acoustic tomography," *IEEE Transactions on Medical Imaging*, vol. 28, no. 4, pp. 585–594, 2009.
[8] E. Candès and Y. Plan, "A probabilistic and ripless theory of compressed sensing," *IEEE Transactions on Information Theory*, vol. 57, no. 11, pp. 7235–7254, 2011.
[9] E. Candès and J. Romberg, "Sparsity and incoherence in compressive sampling," *Inverse Problems*, vol. 23, no. 3, p. 969, 2007.
[10] F. Krahmer and R. Ward, "Stable and robust sampling strategies for compressive imaging," *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 612–622, 2014.
[11] C. Boyer, J. Bigot, and P. Weiss, "Compressed sensing with structured sparsity and structured acquisition," *arxiv*, 2015.
[12] Z. Wang and G. Arce, "Variable density compressed image sampling," *IEEE Transactions on Image Processing*, vol. 19, no. 1, pp. 264–270, 2010.

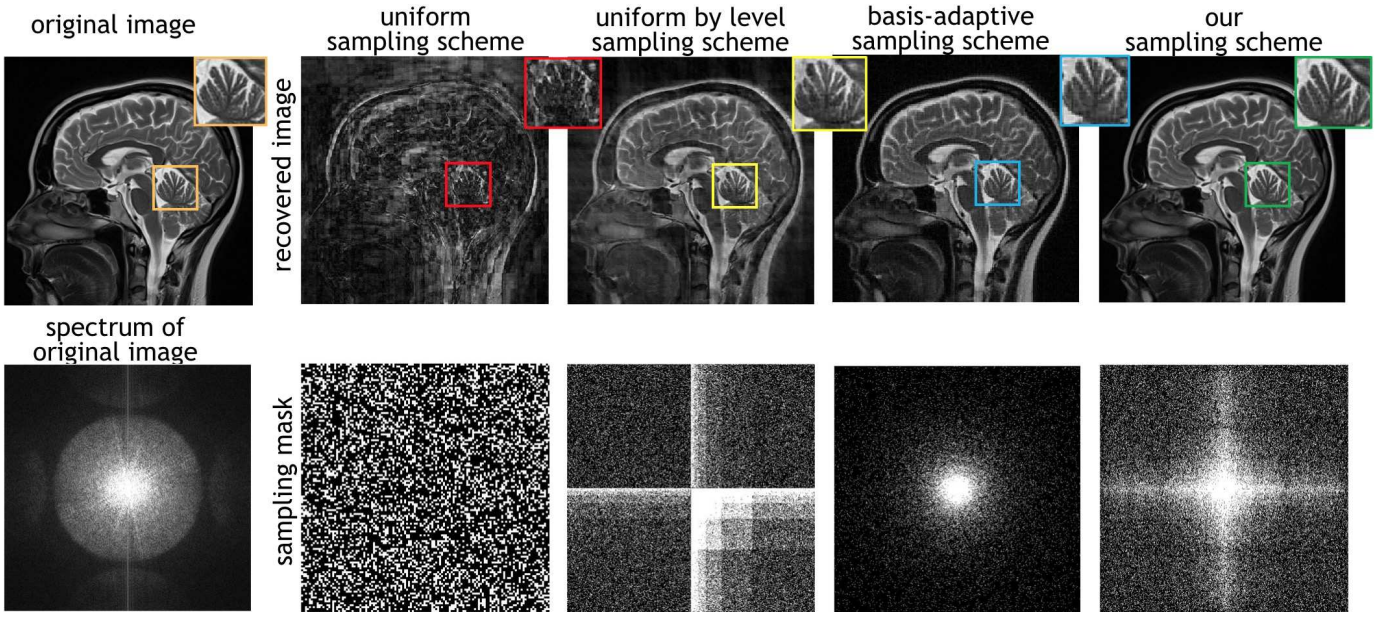

Fig. 2. Comparison of recovered images and sampling masks of four sampling schemes with 30% measurements of total dimensions.

[13] C. Schroder, P. Bornert, and B. Aldefeld, "Spatial excitation using variable-density spiral trajectories," *Journal of Magnetic Resonance Imaging*, vol. 18, no. 1, pp. 136–141, 2003.

[14] G. Puy, J. Marques, R. Gruetter, J. Thiran, D. Ville, P. Vandergheynst, and Y. Wiaux, "Spread spectrum magnetic resonance imaging," *arxiv*, 2012.

[15] N. Chauffert, P. Ciuciu, and P. Weiss, "Variable Density Compressed Sensing In MRI. Theoretical vs Heuristic Sampling Strategies," in *International Symposium on Biomedical Imaging*, 2013.

[16] B. Adcock, P. Univ, A. Hansen, C. Poon, and B. Roman, "Breaking the coherence barrier: asymptotic incoherence and asymptotic sparsity in compressed sensing," Tech. Rep., 2013.

[17] L. Malloy and D. Nowak, "Near-optimal adaptive compressed sensing," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 4001–4012, 2013.

[18] M. Iwen and A. H. Tewfik, "Adaptive compressed sensing for sparse signals in noise," in *Conference on Circuits, Systems & Computers*, 2011, pp. 1240–1244.

[19] J. Haupt, R. Baraniuk, R. Castro, and R. Nowak, "Sequentially designed compressed sensing," in *IEEE Statistical Signal Processing Workshop, 2012*, 2012, pp. 401–404.

[20] C. Boutsidis, M. Mahoney, and P. Drineas, "An improved approximation algorithm for the column subset selection problem," *Computing Research Repository*, vol. abs/0812.4, 2008.

[21] M. Mahoney, "Randomized algorithms for matrices and data," *Computing Research Repository*, 2011.

[22] Y. Chen, S. Bhojanapalli, S. Sanghavi, and R. Ward, "Coherent matrix completion," in *Proceedings of the 31st International Conference on Machine Learning*, 2014, pp. 674–682.

[23] H. Rauhut, "Compressive sensing and structured random matrices," *Radon Series Comp. Appl. Math*, 2010.

[24] D. Gross, "Recovering low-rank matrices from few coefficients in any basis," *IEEE Transactions on Information Theory*, vol. 57, no. 3, pp. 1548–1566, 2011.

[25] Y. Chen, "Completing any low-rank matrix, provably," *arxiv*, 2013.

[26] H. Zhang, Y. Liang, and Y. Zhou, "Analysis of robust pca via local incoherence," *NIPS*, 2015.

[27] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 489–509, Feb 2006.

[28] M. Guerquin-Kern, M. Haberlin, K. Pruessmann, and M. Unser, "A fast wavelet-based reconstruction method for magnetic resonance imaging," *IEEE Transactions on Medical Imaging*, vol. 30, no. 9, pp. 1649–1660, Sept 2011.

APPENDIX A
PROOF OF THEOREM 1

We first characterize the sub-differential of $\|\mathbf{\Psi x}_0\|_1$ via the following proposition for convenience.

Proposition 1. *The sub-differential of* $\|\mathbf{\Psi x}_0\|_1$ *w.r.t.* $\mathbf{x}_0$ *can be expressed as* $\partial_{\mathbf{x}_0}\|\mathbf{\Psi x}_0\|_1 = \boldsymbol{\gamma}_S + \boldsymbol{\gamma}_{S^c}$*, where*

$$\boldsymbol{\gamma}_S = \sum_{j\in S}\mathrm{sgn}(\boldsymbol{\theta})_j\boldsymbol{\psi}_j, \quad \boldsymbol{\gamma}_{S^c} := \left\{\sum_{j\in S^c}\mathbf{f}_j\boldsymbol{\psi}_j \mid \|\mathbf{f}\|_\infty \le 1\right\}.$$

Moreover, $\boldsymbol{\gamma}_S \in \mathbf{\Pi}$ *and* $\boldsymbol{\gamma}_{S^c} \in \mathbf{\Pi}^\perp$.

Proof. Consider arbitrary vector $\mathbf{x}$ with its support set denoted as S, it is well known that

$$\partial\|\mathbf{x}\|_1 := \{\mathrm{sgn}(\mathbf{x}) + \mathbf{f} \mid \mathrm{supp}(\mathbf{f}) = S^c, \|\mathbf{f}\|_\infty \le 1.\}. \quad (16)$$

Then for the vector $\boldsymbol{\theta} = \mathbf{\Psi x}_0$ with support set S, the chain rule gives

$$\begin{aligned}\partial_{\mathbf{x}_0}\|\mathbf{\Psi x}_0\|_1 &= \mathbf{\Psi}^T\partial\|\mathbf{\Psi x}_0\|_1 && (17)\\ &\overset{\text{(i)}}{=} \mathbf{\Psi}^T(\mathrm{sgn}(\boldsymbol{\theta}) + \mathbf{f}) && (18)\\ &\overset{\text{(ii)}}{=} \underbrace{\sum_{j\in S}\mathrm{sgn}(\boldsymbol{\theta})_j\boldsymbol{\psi}_j}_{\boldsymbol{\gamma}_S} + \underbrace{\sum_{j\in S^c}\mathbf{f}_j\boldsymbol{\psi}_j}_{\boldsymbol{\gamma}_{S^c}}, && (19)\end{aligned}$$

where (i) follows from the sub-differential in eq. (16), and (ii) follows from the fact that $\mathrm{supp}(\boldsymbol{\theta}) = S, \mathrm{supp}(\mathbf{f}) = S^c$. Clearly, $\boldsymbol{\gamma}_S \in \mathbf{\Pi}, \boldsymbol{\gamma}_{S^c} \in \mathbf{\Pi}^\perp$. □

Next we introduce the random variable $\delta_j \sim \mathrm{Bernoulli}(p_j)$, and define the weighted sampling operators

$$\mathrm{R}_\Omega(\mathbf{x}) := \sum_{j=1}^n \frac{\delta_j}{p_j}\langle\mathbf{x},\ \boldsymbol{\phi}_j\rangle\boldsymbol{\phi}_j, \mathrm{R}_\Omega^{1/2}(\mathbf{x}) = \sum_{j=1}^n \frac{\delta_j}{\sqrt{p_j}}\langle\mathbf{x}, \boldsymbol{\phi}_j\rangle\boldsymbol{\phi}_j.$$

Then we present below a set of conditions that guarantee $\mathbf{x}_0$ be the unique minimizer of problem **(P)**.

Lemma 2. *If there exists a vector $\boldsymbol{\nu}$ with the following conditions hold*

$$\boldsymbol{\Phi}_{\Omega^c}\boldsymbol{\nu} = 0; \tag{20}$$
$$\|\boldsymbol{\Psi}_{S^c}\mathrm{P}_{\boldsymbol{\Pi}^\perp}\boldsymbol{\nu}\|_\infty \le 1/4; \tag{21}$$
$$\|\mathrm{P}_{\boldsymbol{\Pi}}(\boldsymbol{\gamma}_S - \boldsymbol{\nu})\|_2 \le 1/2s^{10}; \tag{22}$$
$$\|\mathrm{P}_{\boldsymbol{\Pi}}\mathrm{R}_\Omega\mathrm{P}_{\boldsymbol{\Pi}} - \mathrm{P}_{\boldsymbol{\Pi}}\| \le 1/2, \tag{23}$$

*then problem **(P)** has the unique minimizer $\mathbf{x}_0$.*

Proof. Our goal is to show that any feasible perturbation of the signal, i.e. $\mathbf{x}_0 + \Delta\mathbf{x}$ with $\boldsymbol{\Phi}_\Omega\Delta\mathbf{x} = 0$, strictly increases the objective function of problem **(P)** with the above set of conditions. By convexity we have

$$\begin{aligned}
&\|\boldsymbol{\Psi}(\mathbf{x}_0 + \Delta\mathbf{x})\|_1 \\
&\quad\ge \|\boldsymbol{\Psi}\mathbf{x}_0\|_1 + \langle\Delta\mathbf{x}, \partial_{\mathbf{x}_0}\|\boldsymbol{\Psi}\mathbf{x}_0\|_1\rangle \\
&\quad\overset{\text{(i)}}{=} \|\boldsymbol{\Psi}\mathbf{x}_0\|_1 + \langle\Delta\mathbf{x}, \boldsymbol{\gamma}_S + \boldsymbol{\gamma}_{S^c}\rangle \\
&\quad= \|\boldsymbol{\Psi}\mathbf{x}_0\|_1 + \langle\Delta\mathbf{x}, \boldsymbol{\gamma}_S - \boldsymbol{\nu}\rangle + \langle\Delta\mathbf{x}, \boldsymbol{\nu}\rangle + \langle\Delta\mathbf{x}, \boldsymbol{\gamma}_{S^c}\rangle \\
&\quad\overset{\text{(ii)}}{=} \|\boldsymbol{\Psi}\mathbf{x}_0\|_1 + \langle\Delta\mathbf{x}, \boldsymbol{\gamma}_S - \boldsymbol{\nu}\rangle + \langle\Delta\mathbf{x}, \boldsymbol{\gamma}_{S^c}\rangle \\
&\quad\overset{\text{(iii)}}{=} \|\boldsymbol{\Psi}\mathbf{x}_0\|_1 + \langle\mathrm{P}_{\boldsymbol{\Pi}}\Delta\mathbf{x}, \mathrm{P}_{\boldsymbol{\Pi}}(\boldsymbol{\gamma}_S - \boldsymbol{\nu})\rangle \\
&\qquad + \langle\mathrm{P}_{\boldsymbol{\Pi}^\perp}\Delta\mathbf{x}, -\mathrm{P}_{\boldsymbol{\Pi}^\perp}\boldsymbol{\nu}\rangle + \langle\mathrm{P}_{\boldsymbol{\Pi}^\perp}\Delta\mathbf{x}, \boldsymbol{\gamma}_{S^c}\rangle \\
&\quad\overset{\text{(iv)}}{=} \|\boldsymbol{\Psi}\mathbf{x}_0\|_1 + \langle\mathrm{P}_{\boldsymbol{\Pi}}\Delta\mathbf{x}, \mathrm{P}_{\boldsymbol{\Pi}}(\boldsymbol{\gamma}_S - \boldsymbol{\nu})\rangle \\
&\qquad + \langle\boldsymbol{\Psi}\mathrm{P}_{\boldsymbol{\Pi}^\perp}\Delta\mathbf{x}, -\boldsymbol{\Psi}\mathrm{P}_{\boldsymbol{\Pi}^\perp}\boldsymbol{\nu}\rangle + \langle\boldsymbol{\Psi}\mathrm{P}_{\boldsymbol{\Pi}^\perp}\Delta\mathbf{x}, \mathbf{f}\rangle \\
&\quad\overset{\text{(v)}}{=} \|\boldsymbol{\Psi}\mathbf{x}_0\|_1 + \langle\mathrm{P}_{\boldsymbol{\Pi}}\Delta\mathbf{x}, \mathrm{P}_{\boldsymbol{\Pi}}(\boldsymbol{\gamma}_S - \boldsymbol{\nu})\rangle \\
&\qquad + \langle\boldsymbol{\Psi}\mathrm{P}_{\boldsymbol{\Pi}^\perp}\Delta\mathbf{x}, -\boldsymbol{\Psi}_{S^c}\mathrm{P}_{\boldsymbol{\Pi}^\perp}\boldsymbol{\nu}\rangle + \langle\boldsymbol{\Psi}\mathrm{P}_{\boldsymbol{\Pi}^\perp}\Delta\mathbf{x}, \mathbf{f}\rangle.
\end{aligned}$$

Above, (i) follows from Proposition 1, (ii) is due to eq. (20) and the fact that $\boldsymbol{\Phi}_\Omega\Delta\mathbf{x} = 0$, (iii) follows from the facts that $\boldsymbol{\gamma}_S \in \boldsymbol{\Pi}, \boldsymbol{\gamma}_{S^c} \in \boldsymbol{\Pi}^\perp$, (iv) follows from the facts that $\boldsymbol{\Psi}\boldsymbol{\gamma}_{S^c} = \boldsymbol{\Psi}\boldsymbol{\Psi}^T\mathbf{f} = \mathbf{f}$, and (v) follows from $\boldsymbol{\Psi}\mathrm{P}_{\boldsymbol{\Pi}^\perp} = \boldsymbol{\Psi}_{S^c}\mathrm{P}_{\boldsymbol{\Pi}^\perp}$.

Since $\|\mathbf{f}\|_\infty \le 1$, the duality between l_1 and l_∞ guarantees the existence of an $\mathbf{f}$ such that $\langle\boldsymbol{\Psi}\mathrm{P}_{\boldsymbol{\Pi}^\perp}\Delta\mathbf{x}, \mathbf{f}\rangle = \|\boldsymbol{\Psi}\mathrm{P}_{\boldsymbol{\Pi}^\perp}\Delta\mathbf{x}\|_1$. Then the last equation above further gives

$$\begin{aligned}
&\|\boldsymbol{\Psi}(\mathbf{x}_0 + \Delta\mathbf{x})\|_1 \\
&\quad\ge \|\boldsymbol{\Psi}\mathbf{x}_0\|_1 + \langle\mathrm{P}_{\boldsymbol{\Pi}}\Delta\mathbf{x}, \mathrm{P}_{\boldsymbol{\Pi}}(\boldsymbol{\gamma}_S - \boldsymbol{\nu})\rangle \\
&\qquad + \langle\boldsymbol{\Psi}\mathrm{P}_{\boldsymbol{\Pi}^\perp}\Delta\mathbf{x}, -\boldsymbol{\Psi}_{S^c}\mathrm{P}_{\boldsymbol{\Pi}^\perp}\boldsymbol{\nu}\rangle + \|\boldsymbol{\Psi}\mathrm{P}_{\boldsymbol{\Pi}^\perp}\Delta\mathbf{x}\|_1 \\
&\quad\ge \|\boldsymbol{\Psi}\mathbf{x}_0\|_1 - \|\mathrm{P}_{\boldsymbol{\Pi}}\Delta\mathbf{x}\|_2\|\mathrm{P}_{\boldsymbol{\Pi}}(\boldsymbol{\gamma}_S - \boldsymbol{\nu})\|_2 \\
&\qquad - \|\boldsymbol{\Psi}\mathrm{P}_{\boldsymbol{\Pi}^\perp}\Delta\mathbf{x}\|_1\|\boldsymbol{\Psi}_{S^c}\mathrm{P}_{\boldsymbol{\Pi}^\perp}\boldsymbol{\nu}\|_\infty + \|\boldsymbol{\Psi}\mathrm{P}_{\boldsymbol{\Pi}^\perp}\Delta\mathbf{x}\|_1 \\
&\quad\overset{\text{(i)}}{\ge} \|\boldsymbol{\Psi}\mathbf{x}_0\|_1 - \|\mathrm{P}_{\boldsymbol{\Pi}}\Delta\mathbf{x}\|_2\|\mathrm{P}_{\boldsymbol{\Pi}}(\boldsymbol{\gamma}_S - \boldsymbol{\nu})\|_2 \\
&\qquad + \|\mathrm{P}_{\boldsymbol{\Pi}^\perp}\Delta\mathbf{x}\|_2(1 - \|\boldsymbol{\Psi}_{S^c}\mathrm{P}_{\boldsymbol{\Pi}^\perp}\boldsymbol{\nu}\|_\infty) \\
&\quad\overset{\text{(ii)}}{\ge} \|\boldsymbol{\Psi}\mathbf{x}_0\|_1 - \frac{1}{2s^{10}}\|\mathrm{P}_{\boldsymbol{\Pi}}\Delta\mathbf{x}\|_2 + \frac{3}{4}\|\mathrm{P}_{\boldsymbol{\Pi}^\perp}\Delta\mathbf{x}\|_2,
\end{aligned} \tag{24}$$

where (i) uses the fact that $\|\boldsymbol{\Psi}\mathrm{P}_{\boldsymbol{\Pi}^\perp}\Delta\mathbf{x}\|_1 \ge \|\boldsymbol{\Psi}\mathrm{P}_{\boldsymbol{\Pi}^\perp}\Delta\mathbf{x}\|_2 = \|\mathrm{P}_{\boldsymbol{\Pi}^\perp}\Delta\mathbf{x}\|_2$, and (ii) follows from eq. (21) and eq. (22). Notice that the feasible perturbation condition $\boldsymbol{\Phi}_\Omega\Delta\mathbf{x} = 0$ implies that $\|\mathrm{R}_\Omega^{1/2}\Delta\mathbf{x}\|_2 = 0$, which further leads to

$$\|\mathrm{R}_\Omega^{1/2}\mathrm{P}_{\boldsymbol{\Pi}}\Delta\mathbf{x}\|_2 = \|\mathrm{R}_\Omega^{1/2}\mathrm{P}_{\boldsymbol{\Pi}^\perp}\Delta\mathbf{x}\|_2. \tag{25}$$

On the other hand, the quantity $\|\mathrm{R}_\Omega^{1/2}\mathrm{P}_{\boldsymbol{\Pi}}\Delta\mathbf{x}\|_2^2$ satisfies

$$\begin{aligned}
\|\mathrm{R}_\Omega^{1/2}\mathrm{P}_{\boldsymbol{\Pi}}\Delta\mathbf{x}\|_2^2 &= \langle\mathrm{R}_\Omega^{1/2}\mathrm{P}_{\boldsymbol{\Pi}}\Delta\mathbf{x}, \mathrm{R}_\Omega^{1/2}\mathrm{P}_{\boldsymbol{\Pi}}\Delta\mathbf{x}\rangle \\
&\overset{\text{(i)}}{=} \langle\mathrm{P}_{\boldsymbol{\Pi}}\Delta\mathbf{x}, \mathrm{P}_{\boldsymbol{\Pi}}\mathrm{R}_\Omega\mathrm{P}_{\boldsymbol{\Pi}}\Delta\mathbf{x}\rangle \\
&= \langle\mathrm{P}_{\boldsymbol{\Pi}}\Delta\mathbf{x}, (\mathrm{P}_{\boldsymbol{\Pi}}\mathrm{R}_\Omega\mathrm{P}_{\boldsymbol{\Pi}} - \mathrm{P}_{\boldsymbol{\Pi}})\Delta\mathbf{x}\rangle + \|\mathrm{P}_{\boldsymbol{\Pi}}\Delta\mathbf{x}\|_2^2 \\
&\ge -\|\mathrm{P}_{\boldsymbol{\Pi}}\mathrm{R}_\Omega\mathrm{P}_{\boldsymbol{\Pi}} - \mathrm{P}_{\boldsymbol{\Pi}}\|\|\mathrm{P}_{\boldsymbol{\Pi}}\Delta\mathbf{x}\|_2^2 + \|\mathrm{P}_{\boldsymbol{\Pi}}\Delta\mathbf{x}\|_2^2 \\
&\overset{\text{(ii)}}{\ge} \frac{1}{2}\|\mathrm{P}_{\boldsymbol{\Pi}}\Delta\mathbf{x}\|_2^2,
\end{aligned} \tag{26}$$

where (i) follows from the fact that $\mathrm{P}_{\boldsymbol{\Pi}}$, $\mathrm{R}_\Omega^{1/2}$ are self adjoint, and (ii) is by eq. (23). Then combine eq. (26) with eq. (25) we obtain $\|\mathrm{R}_\Omega^{1/2}\mathrm{P}_{\boldsymbol{\Pi}^\perp}\Delta\mathbf{x}\|_2^2 \ge \frac{1}{2}\|\mathrm{P}_{\boldsymbol{\Pi}}\Delta\mathbf{x}\|_2^2$. Also observe that $\|\mathrm{R}_\Omega^{1/2}\mathrm{P}_{\boldsymbol{\Pi}^\perp}\Delta\mathbf{x}\|_2^2 \le \|\mathrm{R}_\Omega^{1/2}\|^2\|\mathrm{P}_{\boldsymbol{\Pi}^\perp}\Delta\mathbf{x}\|_2^2 = \frac{1}{\min_j p_j}\|\mathrm{P}_{\boldsymbol{\Pi}^\perp}\Delta\mathbf{x}\|_2^2$. We then further obtain

$$\|\mathrm{P}_{\boldsymbol{\Pi}^\perp}\Delta\mathbf{x}\|_2^2 \ge \frac{\min_j p_j}{2}\|\mathrm{P}_{\boldsymbol{\Pi}}\Delta\mathbf{x}\|_2^2 \overset{\text{eq. (7)}}{\ge} 1/2s^{20}\|\mathrm{P}_{\boldsymbol{\Pi}}\Delta\mathbf{x}\|_2^2. \tag{27}$$

With the above inequality, eq. (24) further becomes

$$\begin{aligned}
&\|\boldsymbol{\Psi}(\mathbf{x}_0 + \Delta\mathbf{x})\|_1 \\
&\quad\ge \|\boldsymbol{\Psi}\mathbf{x}_0\|_1 - \frac{\sqrt{2}s^{10}}{2s^{10}}\|\mathrm{P}_{\boldsymbol{\Pi}^\perp}\Delta\mathbf{x}\|_2 + \frac{3}{4}\|\mathrm{P}_{\boldsymbol{\Pi}^\perp}\Delta\mathbf{x}\|_2, \\
&\quad\ge \|\boldsymbol{\Psi}\mathbf{x}_0\|_1 + (\frac{3}{4} - \frac{\sqrt{2}}{2})\|\mathrm{P}_{\boldsymbol{\Pi}^\perp}\Delta\mathbf{x}\|_2.
\end{aligned} \tag{28}$$

Now we argue by contradiction to prove that $\|\mathrm{P}_{\boldsymbol{\Pi}^\perp}\Delta x\|_2 > 0$. Assume this fails to hold, i.e. $\mathrm{P}_{\boldsymbol{\Pi}}\Delta\mathbf{x} = \Delta\mathbf{x}$. Then feasibility condition $\boldsymbol{\Phi}_\Omega\Delta\mathbf{x} = 0$ implies that $\mathrm{R}_\Omega\mathrm{P}_{\boldsymbol{\Pi}}\Delta\mathbf{x} = 0$. Putting all these together we obtain $\|(\mathrm{P}_{\boldsymbol{\Pi}}\mathrm{R}_\Omega\mathrm{P}_{\boldsymbol{\Pi}} - \mathrm{P}_{\boldsymbol{\Pi}})\Delta\mathbf{x}\|_2 = \|\Delta\mathbf{x}\|_2$, which contradicts the fact that $\|\mathrm{P}_{\boldsymbol{\Pi}}\mathrm{R}_\Omega\mathrm{P}_{\boldsymbol{\Pi}} - \mathrm{P}_{\boldsymbol{\Pi}}\| \le 1/2$. Hence, we have shown that any feasible perturbation of the signal strictly increases the objective function. □

All left is to provide a high probability guarantee for the set of conditions based on our sampling scheme. We first present the bound in eq. (23), and the proof is in Section E.

Lemma 3. *With the sampling scheme in Theorem 1, one has*

$$\|\mathrm{P}_{\boldsymbol{\Pi}}\mathrm{R}_\Omega\mathrm{P}_{\boldsymbol{\Pi}} - \mathrm{P}_{\boldsymbol{\Pi}}\| \le 1/2 \tag{29}$$

with probability at least $1 - n^{-C_0}$.

Now we construct the vector $\boldsymbol{\nu}$ that satisfies eq. (20), eq. (21) and eq. (22) via the clever 'golfing scheme'. First we decompose the Bernoulli sampling model $\mathrm{Bernoulli}(p_j)$ into k_0 independent Bernoulli schemes , i.e. $\{\mathrm{Bernoulli}(q_j^k)\}_{k=1}^{k_0}$. Then to be equivalent we must have $(1-p_j) = \Pi_{k=1}^{k_0}(1-q_j^k)$. We also define the corresponding weighted sampling operator $\{\mathrm{R}_{\Omega_k}\}_{k=1}^{k_0}$ of each decomposed Bernoulli sampling scheme. More specifically, we set $k_0 = 11\log s + 2$, and for all j we set $q_j^1 = q_j^2 = p_j/6; q_j^k \equiv q_j, k \ge 3$. Then a non-overlapping argument shows that $q_j \ge p_j/(k_0 - 2) \ge C_0\boldsymbol{\pi}(\boldsymbol{\Pi}, \boldsymbol{\phi}_j)\log n$. Now we start with $\boldsymbol{\nu}_0 = 0$, and construct $\boldsymbol{\nu} := \boldsymbol{\nu}_{k_0}$ via the following iterative scheme

$$\boldsymbol{\nu}_{k+1} = \boldsymbol{\nu}_k + R_{\Omega_{k+1}}(\mathrm{P}_{\boldsymbol{\Pi}}\boldsymbol{\nu}_k - \boldsymbol{\gamma}_S), k = 0, 1, ..., k_0 - 1.$$

By introducing the variables $\mathbf{w}_0 = -\boldsymbol{\gamma}_S$, $\mathbf{w}_k = \mathrm{P}_{\boldsymbol{\Pi}}\boldsymbol{\nu}_k - \gamma_S$, the above iterative scheme can be expressed as

$$\boldsymbol{\nu}_{k_0} = \sum_{k=1}^{k_0} R_{\Omega_k}\mathbf{w}_{k-1}, \tag{30}$$

$$\mathbf{w}_k = (\mathrm{P}_{\boldsymbol{\Pi}}\mathrm{R}_{\Omega_k}\mathrm{P}_{\boldsymbol{\Pi}} - \mathrm{P}_{\boldsymbol{\Pi}})\mathbf{w}_{k-1}, k = 1, ..., k_0, \tag{31}$$

which brings more convenience for subsequent analysis. Before proving the conditions on the constructed $\boldsymbol{\nu}_{k_0}$, we first collect the following concentration bound for convenience. The proof is in Section F.

Lemma 4. *For any vector* $\mathbf{w}$*, one has*

$$\|(\mathrm{P}_{\boldsymbol{\Pi}}\mathrm{R}_{\Omega_k}\mathrm{P}_{\boldsymbol{\Pi}} - \mathrm{P}_{\boldsymbol{\Pi}})\mathbf{w}\|_2 \le \frac{1}{2}\|\mathbf{w}\|_2, k = 1, 2..., k_0 \tag{32}$$

with probability at least $1 - n^{-C_0}$.

Also, note that the infinity norm induced by basis $\boldsymbol{\Phi}$ is $\max_j |[\boldsymbol{\Phi}\mathbf{w}]_j| = \max_j |\langle \mathbf{w}, \boldsymbol{\phi}_j\rangle|$. By scaling this norm with the scalar $\|\mathrm{P}_{\boldsymbol{\Pi}}\boldsymbol{\phi}_j\|_2$, we then define the following weighted norm $\|\cdot\|_{\boldsymbol{\Phi},\infty}$

$$\|\mathbf{w}\|_{\boldsymbol{\Phi},\infty} := \max_j \frac{|\langle \mathbf{w}, \boldsymbol{\phi}_j\rangle|}{\|\mathrm{P}_{\boldsymbol{\Pi}}\boldsymbol{\phi}_j\|_2}.$$

Equipped with this weighted norm, the following bounds holds. The proof are in Section G, Section H, and Section I, respectively.

Lemma 5. *For any vector* $\mathbf{w}$*, one has*

$$\|\boldsymbol{\Psi}_{S^c}(R_{\Omega_k} - I)\mathbf{w}\|_\infty \le \frac{1}{\sqrt{C_0}}\|\mathbf{w}\|_{\boldsymbol{\Phi},\infty}, \quad k = 1, 2...k_0$$

with probability at least $1 - n^{-\sqrt{C_0}}$.

Lemma 6. *For any vector* $\mathbf{w} \in \boldsymbol{\Pi}$*, one has*

$$\|(\mathrm{P}_{\boldsymbol{\Pi}}\mathrm{R}_{\Omega_k}\mathrm{P}_{\boldsymbol{\Pi}} - \mathrm{P}_{\boldsymbol{\Pi}})\mathbf{w}\|_{\boldsymbol{\Phi},\infty} \le \begin{cases} \|\mathbf{w}\|_{\boldsymbol{\Phi},\infty}/\sqrt{\log s}, & k \in \{1, 2\} \\ \|\mathbf{w}\|_{\boldsymbol{\Phi},\infty}/2, & k = 3, 4..., k_0 \end{cases}$$

with probability at least $1 - n^{-C_0}$.

Lemma 7. *With probability at least* $1 - s^{-\sqrt{C_0}/4}$*, one has*

$$\|\boldsymbol{\gamma}_S\|_{\boldsymbol{\Phi},\infty} \le \frac{\sqrt{C_0}\log s}{4}. \tag{33}$$

Now equipped with all technical lemmas above, we show that the constructed $\boldsymbol{\nu}_{k_0}$ satisfies the set of conditions. First, by the decomposition of the Bernoulli sampling model, eq. (30) directly implies eq. (21). Next we verify eq. (22) via

$$\begin{aligned} \|\mathrm{P}_{\boldsymbol{\Pi}}(\boldsymbol{\gamma}_S - \boldsymbol{\nu}_{k_0})\|_2 &= \|\mathbf{w}_{k_0}\|_2 \\ \text{(eq. (31))} &= \|(\mathrm{P}_{\boldsymbol{\Pi}}\mathrm{R}_{\Omega_{k_0}}\mathrm{P}_{\boldsymbol{\Pi}} - \mathrm{P}_{\boldsymbol{\Pi}})\mathbf{w}_{k_0-1}\|_2 \\ \text{(Lemma 4)} &\le \frac{1}{2}\|\mathbf{w}_{k_0-1}\|_2 \le ... \le \frac{1}{2^{11\log s+2}}\|\boldsymbol{\gamma}_S\|_2 \\ &\le \frac{\sqrt{s}}{4s^{11}} \le \frac{1}{2s^{10}}. \end{aligned}$$

Finally, we verify eq. (21) via

$$\begin{aligned} \|\boldsymbol{\Psi}_{S^c}\mathrm{P}_{\boldsymbol{\Pi}^\perp}\boldsymbol{\nu}_{k_0}\|_\infty &= \|\boldsymbol{\Psi}_{S^c}\sum_{k=1}^{k_0}\mathrm{P}_{\boldsymbol{\Pi}^\perp}R_{\Omega_k}\mathbf{w}_{k-1}\|_\infty \\ (\mathbf{w}_k \in \boldsymbol{\Pi}) &= \|\boldsymbol{\Psi}_{S^c}\sum_{k=1}^{k_0}\mathrm{P}_{\boldsymbol{\Pi}^\perp}(R_{\Omega_k} - I)\mathbf{w}_{k-1}\|_\infty \\ \text{(triangle inequality)} &\le \sum_{k=1}^{k_0}\|\boldsymbol{\Psi}_{S^c}\mathrm{P}_{\boldsymbol{\Pi}^\perp}(R_{\Omega_k} - I)\mathbf{w}_{k-1}\|_\infty \\ (\boldsymbol{\Psi}_{S^c}\mathrm{P}_{\boldsymbol{\Pi}} = 0) &= \sum_{k=1}^{k_0}\|\boldsymbol{\Psi}_{S^c}(R_{\Omega_k} - I)\mathbf{w}_{k-1}\|_\infty \\ \text{(Lemma 5)} &\le \sum_{k=1}^{k_0}\frac{1}{\sqrt{C_0}}\|\mathbf{w}_{k-1}\|_{\boldsymbol{\Phi},\infty} \\ \text{(Lemma 6)} &\le \sum_{k=1}^{k_0}\frac{1}{\sqrt{C_0}}\frac{1}{\log s}\frac{1}{2^{k-1}}\|\mathbf{w}_0\|_{\boldsymbol{\Phi},\infty} \\ (\mathbf{w}_0 = -\boldsymbol{\gamma}_S) &\le \frac{\|\boldsymbol{\gamma}_S\|_{\boldsymbol{\Phi},\infty}}{\sqrt{C_0}\log s} \\ \text{(Lemma 7)} &\le \frac{1}{4}. \end{aligned}$$

To this end we have shown that the constructed ν_{k_0} satisfies all conditions, and the succeed probability over all the proof is dominated by $1 - s^{-\sqrt{C_0}/4}$.

APPENDIX B
PROOF OF THEOREM 2

Let's prove the first inequality. For the two terms involved in $\boldsymbol{\pi}$ coherence, we have

$$\begin{aligned} \|\mathrm{P}_{\boldsymbol{\Pi}}\boldsymbol{\phi}_j\|_2^2 &= \langle \boldsymbol{\Psi}\mathrm{P}_{\boldsymbol{\Pi}}\boldsymbol{\phi}_j, \boldsymbol{\Psi}\mathrm{P}_{\boldsymbol{\Pi}}\boldsymbol{\phi}_j\rangle \\ \text{(dual norm)} &\le \|\boldsymbol{\Psi}\mathrm{P}_{\boldsymbol{\Pi}}\boldsymbol{\phi}_j\|_1\|\boldsymbol{\Psi}\mathrm{P}_{\boldsymbol{\Pi}}\boldsymbol{\phi}_j\|_\infty \\ &\le \|\boldsymbol{\Psi}\mathrm{P}_{\boldsymbol{\Pi}}\boldsymbol{\phi}_j\|_1\|\boldsymbol{\Psi}\boldsymbol{\phi}_j\|_\infty \\ &= \|\boldsymbol{\Psi}\mathrm{P}_{\boldsymbol{\Pi}}\boldsymbol{\phi}_j\|_1\sqrt{\mu(\boldsymbol{\Psi}, \boldsymbol{\phi}_j)}, \end{aligned}$$

and

$$\begin{aligned} \|\mathrm{P}_{\boldsymbol{\Pi}}\boldsymbol{\phi}_j\|_2\|\boldsymbol{\Psi}\mathrm{P}_{\boldsymbol{\Pi}^\perp}\boldsymbol{\phi}_j\|_\infty &\le \|\boldsymbol{\Psi}\mathrm{P}_{\boldsymbol{\Pi}}\boldsymbol{\phi}_j\|_2\|\boldsymbol{\Psi}\boldsymbol{\phi}_j\|_\infty \\ (l_2 \le l_1) &\le \|\boldsymbol{\Psi}\mathrm{P}_{\boldsymbol{\Pi}}\boldsymbol{\phi}_j\|_1\|\boldsymbol{\Psi}\boldsymbol{\phi}_j\|_\infty \\ &= \|\boldsymbol{\Psi}\mathrm{P}_{\boldsymbol{\Pi}}\boldsymbol{\phi}_j\|_1\sqrt{\mu(\boldsymbol{\Psi}, \boldsymbol{\phi}_j)}. \end{aligned}$$

Hence, the first inequality holds. For the second inequality, we begin with

$$\begin{aligned} \|\boldsymbol{\Psi}\mathrm{P}_{\boldsymbol{\Pi}}\boldsymbol{\phi}_j\|_1\sqrt{\mu(\boldsymbol{\Psi}, \boldsymbol{\phi}_j)} &= \|\boldsymbol{\Psi}\mathrm{P}_{\boldsymbol{\Pi}}\boldsymbol{\phi}_j\|_1\|\boldsymbol{\Psi}\boldsymbol{\phi}_j\|_\infty \\ (\dim(\boldsymbol{\Pi}) = s) &\le s\|\boldsymbol{\Psi}\mathrm{P}_{\boldsymbol{\Pi}}\boldsymbol{\phi}_j\|_\infty\|\boldsymbol{\Psi}\boldsymbol{\phi}_j\|_\infty \\ &\le s\|\boldsymbol{\Psi}\boldsymbol{\phi}_j\|_\infty^2 \\ &= s\mu(\boldsymbol{\Psi}, \boldsymbol{\phi}_j). \end{aligned}$$

Now the third inequality holds trivially.

The sampling scheme in [10] has sample complexity of the order $\mathcal{O}(\sum_j \mu(\boldsymbol{\Psi}, \boldsymbol{\phi}_j)s\log^3 s\log n)$, while ours is of the order $\mathcal{O}(\sum_j \boldsymbol{\pi}(\boldsymbol{\Pi}, \boldsymbol{\phi}_j)\log s\log n)$. Thus, by the second inequality our sample complexity is lower. The sample complexity in [11] is of the order $\mathcal{O}(\max\{\Theta, \Gamma\}\log s\log n)$, where Θ, Γ are

two complicated notions of coherence that both depend on the sampling probability. There's no general closed form for the optimal sampling scheme that minimizes the two coherences. However, we can consider the lower bound $\mathcal{O}(\Theta \log s \log n)$, which can be minimized by the sampling scheme

$$p_j \propto \frac{\|\mathbf{\Psi} \mathrm{P}_{\mathbf{\Pi}} \boldsymbol{\phi}_j\|_1 \sqrt{\mu(\mathbf{\Psi}, \boldsymbol{\phi}_j)}}{\sum_j \|\mathbf{\Psi} \mathrm{P}_{\mathbf{\Pi}} \boldsymbol{\phi}_j\|_1 \sqrt{\mu(\mathbf{\Psi}, \boldsymbol{\phi}_j)}} \tag{34}$$

with complexity $\mathcal{O}(\sum_j \|\mathbf{\Psi} \mathrm{P}_{\mathbf{\Pi}} \boldsymbol{\phi}_j\|_1 \sqrt{\mu(\mathbf{\Psi}, \boldsymbol{\phi}_j)} \log s \log n)$. Hence by our first inequality, our sample complexity is lower.

Appendix C
Proof of Lemma 1

Adopt the definition of local coherence, we define

$$\|\mathbf{\Psi} \mathrm{P}_{\mathbf{\Pi}_l} \boldsymbol{\phi}_j\|_\infty = \sqrt{\mu(\mathbf{\Pi}_l, \boldsymbol{\phi}_j)} \tag{35}$$

$$\|\mathbf{\Psi} \mathrm{P}_{\mathbf{\Pi}^\perp} \boldsymbol{\phi}_j\|_\infty = \sqrt{\mu(\mathbf{\Pi}^\perp, \boldsymbol{\phi}_j)}. \tag{36}$$

Since $\dim(\mathbf{\Pi}_l) = s_l$, we obtain by the relationship between l_2 and l_∞ that

$$\|\mathrm{P}_{\mathbf{\Pi}_l} \boldsymbol{\phi}_j\|_2^2 \le s_l \|\mathbf{\Psi} \mathrm{P}_{\mathbf{\Pi}_l} \boldsymbol{\phi}_j\|_\infty^2 \overset{\text{eq. (35)}}{=} s_l \mu(\mathbf{\Pi}_l, \boldsymbol{\phi}_j). \tag{37}$$

Thus, we obtain the upper bound

$$\|\mathrm{P}_{\mathbf{\Pi}} \boldsymbol{\phi}_j\|_2 \|\mathbf{\Psi} \mathrm{P}_{\mathbf{\Pi}^\perp} \boldsymbol{\phi}_j\|_\infty \tag{38}$$

$$\text{(eq. (3), eq. (36))} = \sqrt{\sum_{l=1}^{p} \|\mathrm{P}_{\mathbf{\Pi}_l} \boldsymbol{\phi}_j\|_2^2} \sqrt{\mu(\mathbf{\Pi}^\perp, \boldsymbol{\phi}_j)} \tag{39}$$

$$\text{(eq. (37))} \le \sqrt{\sum_{l=1}^{p} s_l \mu(\mathbf{\Pi}_l, \boldsymbol{\phi}_j) \mu(\mathbf{\Pi}^\perp, \boldsymbol{\phi}_j)}. \tag{40}$$

Next we decompose $\mathbf{\Pi}$ w.r.t. different dyadic levels. Specifically, we denote $\mathbf{\Pi}_l : \mathbf{\Pi} \cap \mathrm{span}\{\boldsymbol{\psi}_j \mid j \in L_l\}$, i.e. the part of $\mathbf{\Pi}$ in the subspace spanned by the wavelet basis in level l, and it follows that $\mathbf{\Pi} = \bigoplus_{l=0}^{p} \mathbf{\Pi}_l$. Now we have

$$\begin{aligned}
&\sum_{j=1}^{n} \|\mathrm{P}_{\mathbf{\Pi}} \boldsymbol{\phi}_j\|_2 \|\mathbf{\Psi} \mathrm{P}_{\mathbf{\Pi}^\perp} \boldsymbol{\phi}_j\|_\infty \\
\text{(eq. (40))} &\le \sum_{j=1}^{n} \sqrt{\sum_{l=0}^{p} s_l \mu(\mathbf{\Pi}_l, \boldsymbol{\phi}_j) \mu(\mathbf{\Pi}^\perp, \boldsymbol{\phi}_j)} \\
&\le \sum_{j=1}^{n} \sqrt{\sum_{l=0}^{p} s_l \mu(\mathbf{\Pi}_l, \boldsymbol{\phi}_j) \mu(\mathbf{\Psi}, \boldsymbol{\phi}_j)} \\
&\le \sum_{j=1}^{n} \sqrt{\sum_{l=0}^{p} s_l \max_{i \in L_l} \mu(\boldsymbol{\psi}_i, \boldsymbol{\phi}_j) \max_i \mu(\boldsymbol{\psi}_i, \boldsymbol{\phi}_j)} \\
\text{(eq. (11))} &\le \sum_{j=1}^{n} \sqrt{\sum_{l=0}^{p} s_l 2^{-k(j)} 2^{-|k(j)-l|} 2^{-k(j)}} \\
&\le \sum_{j=1}^{n} \sum_{l=0}^{p} \sqrt{s_l} 2^{-k(j)} 2^{-|k(j)-l|/2} \\
(|L_{k(j)}| \le 2^{k(j)}) &\le \sum_{l=0}^{p} \sqrt{s_l} \sum_{k(j)=0}^{p} 2^{-|k(j)-l|/2} \\
&\le \frac{2\sqrt{2}}{\sqrt{2}-1} \sum_{l=0}^{p} \sqrt{s_l} \\
&\le \frac{2\sqrt{2}}{\sqrt{2}-1} s.
\end{aligned}$$

Since we have shown that $\sum_{j=1}^{n} \|\mathrm{P}_{\mathbf{\Pi}} \boldsymbol{\phi}_j\|_2^2 = s$, the overall sample complexity of eq. (6) is upper bounded by $\mathcal{O}(s \log s \log n)$.

Appendix D
Proof of Theorem 3

Let us introduce the Bernoulli random variables $\delta_j \sim \mathrm{Bernoulli}(p_j), j = 1, \cdots, n$, and the number of sampled measurements can be represented by $|\Omega| = \sum_{j=1}^{n} \delta_j$. Then the converse scheme in eq. (14) implies that

$$\mathbb{E}|\Omega| = \sum_{j=1}^{n} p_j \le \sum_{j=1}^{n} \|\mathrm{P}_{\mathbf{\Pi}} \boldsymbol{\phi}_j\|_2^2 \le s. \tag{41}$$

Hence, a simple Hoeffding's bound gives that $|\Omega| < s-2$ with high probability.

Now let us identify a set of candidates $\hat{\mathbf{x}}$ that achieves the same objective value as $\mathbf{x}_0$. Recall that $\theta := \mathbf{\Psi} \mathbf{x}_0$ with support set S, and our constructed $\hat{\mathbf{x}}$ is as follows

$$\hat{\mathbf{x}} := \mathbf{x}_0 - \mathbf{\Psi}^T (\boldsymbol{\lambda} \cdot \mathrm{sgn}(\boldsymbol{\theta})), \quad \text{where} \quad |\boldsymbol{\lambda}| \preccurlyeq |\boldsymbol{\theta}|. \tag{42}$$

In the above expression, $\boldsymbol{\lambda}$ is a vector, '$\cdot$' denotes the element-wise product, and '$\preccurlyeq$' means $\le$ holds element-wise. Consequently, the objective value of problem **(P)** for $\hat{\mathbf{x}}$ becomes

$$\begin{aligned}
\|\mathbf{\Psi} \hat{\mathbf{x}}\|_1 &= \|\boldsymbol{\theta} - \boldsymbol{\lambda} \cdot \mathrm{sgn}(\boldsymbol{\theta})\|_1 \\
&= \sum_{i \in S} |\boldsymbol{\theta}_i - \boldsymbol{\lambda}_i \mathrm{sgn}(\boldsymbol{\theta}_i)| \\
&\overset{\text{(i)}}{=} \sum_{i \in S} |\boldsymbol{\theta}_i| - \boldsymbol{\lambda}_i \\
&= \|\mathbf{\Psi} \mathbf{x}_0\|_1 - \langle \boldsymbol{\lambda}, \mathbf{1}_S \rangle,
\end{aligned}$$

where (i) follows from $|\boldsymbol{\lambda}| \preccurlyeq |\boldsymbol{\theta}|$, and $\mathbf{1}_S$ denotes a vector that has 1s on the support set S. Thus, to achieve the same objective value we only need to require

$$\langle \boldsymbol{\lambda}, \mathbf{1}_S \rangle = 0. \tag{43}$$

Moreover, the linear constraint of problem **(P)** requires $\mathbf{\Phi}_\Omega \hat{\mathbf{x}} = \mathbf{\Phi}_\Omega \mathbf{x}_0$. Then with the form of $\hat{\mathbf{x}}$, this constraint becomes

$$\mathbf{\Phi}_\Omega \mathbf{\Psi}^T (\boldsymbol{\lambda} \cdot \mathrm{sgn}(\boldsymbol{\theta})) = 0. \tag{44}$$

Notice from eqs. (43) and (44) that we only need to identify the entries of $\boldsymbol{\lambda}$ that are supported on S, and the total number of linear constrains are $1 + |\Omega| < s - 1$. Thus, we can identify infinite many $\boldsymbol{\lambda}$s that satisfy eqs. (43) and (44). Finally, to ensure that $|\boldsymbol{\lambda}| \preccurlyeq |\boldsymbol{\theta}|$ we simply apply the normalization

$$\forall i \in S, \quad \boldsymbol{\lambda}_i \leftarrow \boldsymbol{\lambda}_i \frac{\min_i |\boldsymbol{\theta}_i|}{\|\boldsymbol{\lambda}\|_\infty}. \tag{45}$$

APPENDIX E
PROOF OF LEMMA 3

The operator can be equivalently expressed as:

$$(\mathrm{P}_{\mathbf{\Pi}}\mathrm{R}_{\Omega}\mathrm{P}_{\mathbf{\Pi}}-\mathrm{P}_{\mathbf{\Pi}})(\cdot)=\sum_{j=1}^{n}\left(\frac{\delta_j}{p_j}-1\right)\langle\boldsymbol{\phi}_j,\mathrm{P}_{\mathbf{\Pi}}(\cdot)\rangle\mathrm{P}_{\mathbf{\Pi}}\boldsymbol{\phi}_j \overset{\text{def}}{=}\sum_{j=1}^{n}\mathcal{X}_j(\cdot).$$

Then for each operator $\mathcal{X}_j(\cdot)$, its operator norm can be bounded as

$$\|\mathcal{X}_j(\cdot)\|=\max_{\mathbf{z}}\frac{\|\mathcal{X}_j(\mathbf{z})\|_2}{\|\mathbf{z}\|_2}\leq\max_{\mathbf{z}}\frac{\|\mathrm{P}_{\mathbf{\Pi}}\boldsymbol{\phi}_j\|_2^2\|\mathbf{z}\|_2}{p_j\|\mathbf{z}\|_2}\overset{\text{eq. (6)}}{\leq}\frac{1}{C_0\log n}.$$

Moreover, the operator norm of its variance $\|\sum_{j=1}^{n}\mathbb{E}\mathcal{X}_j^2(\cdot)\|$ can be bounded as

$$\begin{aligned}
&\left\|\sum_{j=1}^{n}\mathbb{E}\mathcal{X}_j^2(\cdot)\right\|=\max_{\mathbf{z}}\frac{\|\sum_{j=1}^{n}\mathbb{E}\mathcal{X}_j^2(\mathbf{z})\|_2}{\|\mathbf{z}\|_2}\\
&=\max_{\mathbf{z}}\left\|\sum_{j=1}^{n}\mathbb{E}\left(\frac{\delta_j}{p_j}-1\right)^2\langle\boldsymbol{\phi}_j,\mathrm{P}_{\mathbf{\Pi}}\mathbf{z}\rangle\langle\boldsymbol{\phi}_j,\mathrm{P}_{\mathbf{\Pi}}\boldsymbol{\phi}_j\rangle\mathrm{P}_{\mathbf{\Pi}}\boldsymbol{\phi}_j\right\|_2/\|\mathbf{z}\|_2\\
&=\max_{\mathbf{z}}\left\|\sum_{j=1}^{n}\frac{(1-p_j)\|\mathrm{P}_{\mathbf{\Pi}}\boldsymbol{\phi}_j\|_2^2}{p_j}\langle\boldsymbol{\phi}_j,\mathrm{P}_{\mathbf{\Pi}}\mathbf{z}\rangle\mathrm{P}_{\mathbf{\Pi}}\boldsymbol{\phi}_j\right\|_2/\|\mathbf{z}\|_2\\
&\leq\max_{\mathbf{z}}\left\{\max_j\frac{(1-p_j)\|\mathrm{P}_{\mathbf{\Pi}}\boldsymbol{\phi}_j\|_2^2}{p_j}\right\}\left\|\sum_{j=1}^{n}\langle\boldsymbol{\phi}_j,\mathrm{P}_{\mathbf{\Pi}}\mathbf{z}\rangle\boldsymbol{\phi}_j\right\|_2/\|\mathbf{z}\|_2\\
&\leq\max_{\mathbf{z}}\left\{\max_j\frac{\|\mathrm{P}_{\mathbf{\Pi}}\boldsymbol{\phi}_j\|_2^2}{p_j}\right\}\frac{\|\mathrm{P}_{\mathbf{\Pi}}\mathbf{z}\|_2}{\|\mathbf{z}\|_2}\\
&\overset{\text{(i)}}{\leq}\frac{1}{C_0\log n},
\end{aligned}$$

where (i) follows from eq. (6). Then apply the matrix Bernstein's inequality in Theorem 6 gives the result.

APPENDIX F
PROOF OF LEMMA 4

For all $k=1,\cdots,k_0$, we expand $(\mathrm{P}_{\mathbf{\Pi}}\mathrm{R}_{\Omega_k}\mathrm{P}_{\mathbf{\Pi}}-\mathrm{P}_{\mathbf{\Pi}})\mathbf{w}$ as

$$(\mathrm{P}_{\mathbf{\Pi}}\mathrm{R}_{\Omega_k}\mathrm{P}_{\mathbf{\Pi}}-\mathrm{P}_{\mathbf{\Pi}})\mathbf{w}=\sum_{j=1}^{n}\left(\frac{\delta_j^k}{q_j^k}-1\right)\langle\boldsymbol{\phi}_j,\mathrm{P}_{\mathbf{\Pi}}\mathbf{w}\rangle\mathrm{P}_{\mathbf{\Pi}}\boldsymbol{\phi}_j\overset{\text{def}}{=}\sum_{j=1}^{n}\mathcal{X}_j^k.$$

Then for all j we have

$$\begin{aligned}
\|\mathcal{X}_j^k\|_2&=\left\|\left(\frac{\delta_j^k}{q_j^k}-1\right)\langle\mathrm{P}_{\mathbf{\Pi}}\boldsymbol{\phi}_j,\mathbf{w}\rangle\mathrm{P}_{\mathbf{\Pi}}\boldsymbol{\phi}_j\right\|_2\\
&\leq\frac{\|\mathrm{P}_{\mathbf{\Pi}}\boldsymbol{\phi}_j\|_2^2}{q_j^k}\|\mathbf{w}\|_2\\
&\overset{\text{(i)}}{\leq}\frac{\|\mathbf{w}\|_2}{C_0\log n},k=1,2...k_0
\end{aligned}$$

where (i) is by the specification of q_j^k of the golfing scheme. Moreover, for the variance part we have

$$\begin{aligned}
\sum_{j=1}^{n}\mathbb{E}\|\mathcal{X}_j^k\|_2^2&=\sum_{j=1}^{n}\mathbb{E}\left(\frac{\delta_j^k}{q_j^k}-1\right)^2|\langle\boldsymbol{\phi}_j,\mathrm{P}_{\mathbf{\Pi}}\mathbf{w}\rangle|^2\|\mathrm{P}_{\mathbf{\Pi}}\boldsymbol{\phi}_j\|_2^2\\
&\leq\sum_{j=1}^{n}\frac{\|\mathrm{P}_{\mathbf{\Pi}}\boldsymbol{\phi}_j\|_2^2}{q_j^k}|\langle\boldsymbol{\phi}_j,\mathrm{P}_{\mathbf{\Pi}}\mathbf{w}\rangle|^2\\
&\leq\max_j\left(\frac{\|\mathrm{P}_{\mathbf{\Pi}}\boldsymbol{\phi}_j\|_2^2}{q_j^k}\right)\|\mathrm{P}_{\mathbf{\Pi}}\mathbf{w}\|_2^2\\
&\overset{\text{(i)}}{\leq}\frac{\|\mathbf{w}\|_2^2}{C_0\log n},k=1,2...k_0
\end{aligned}$$

where (i) also follows from the specification of q_j^k of the golfing scheme. The result follows from the vector Bernstein's inequality in Theorem 5.

APPENDIX G
PROOF OF LEMMA 5

Since $\mathbf{\Psi}_{S^c}$ is restricted on the rows $\boldsymbol{\psi}_i^T, i\in S^c$, $\mathbf{\Psi}_{S^c}(R_{\Omega_k}-I)\mathbf{w}$ is thus supported on S^c. The i-th entry in S^c of vector $\mathbf{\Psi}_{S^c}(R_{\Omega_k}-I)\mathbf{w}$ can be alternatively expressed as

$$[\mathbf{\Psi}_{S^c}(R_{\Omega_k}-I)\mathbf{w}]_i=\sum_{j=1}^{n}\left(\frac{\delta_j^k}{q_j^k}-1\right)\langle\boldsymbol{\phi}_j,\mathbf{w}\rangle\langle\boldsymbol{\phi}_j,\boldsymbol{\psi}_i\rangle\overset{\text{def}}{=}\sum_{j=1}^{n}\mathcal{X}_j^k,\quad i\in S^c.$$

By the specifications of the golfing scheme, we have for all k that $q_j^k\geq C_0\|\mathrm{P}_{\mathbf{\Pi}}\boldsymbol{\phi}_j\|_2\|\mathbf{\Psi}\mathrm{P}_{\mathbf{\Pi}^\perp}\boldsymbol{\phi}_j\|_\infty\log n$. We then obtain

$$\begin{aligned}
|\mathcal{X}_j^k|&=\left|\left(\frac{\delta_j^k}{q_j^k}-1\right)\langle\boldsymbol{\phi}_j,\mathbf{w}\rangle\langle\boldsymbol{\phi}_j,\boldsymbol{\psi}_i\rangle\right|\\
&\leq\left|\frac{\langle\boldsymbol{\phi}_j,\mathbf{w}\rangle\langle\boldsymbol{\phi}_j,\boldsymbol{\psi}_i\rangle}{C_0\|\mathrm{P}_{\mathbf{\Pi}}\boldsymbol{\phi}_j\|_2\|\mathbf{\Psi}\mathrm{P}_{\mathbf{\Pi}^\perp}\boldsymbol{\phi}_j\|_\infty\log n}\right|\\
&\overset{\text{(i)}}{\leq}\frac{\|\mathbf{w}\|_{\mathbf{\Phi},\infty}}{C_0\log n},
\end{aligned}$$

where (i) follows from the fact that $|\langle\boldsymbol{\phi}_j,\boldsymbol{\psi}_i\rangle|\leq\|\mathbf{\Psi}\mathrm{P}_{\mathbf{\Pi}^\perp}\boldsymbol{\phi}_j\|_\infty$ for all $i\in S^c$. Moreover, we also have $q_j^k\geq C_0\|\mathrm{P}_{\mathbf{\Pi}}\boldsymbol{\phi}_j\|_2^2\log n$. Then we obtain

$$\begin{aligned}
\sum_{j=1}^{n}\mathbb{E}|\mathcal{X}_j^k|^2&=\sum_{j=1}^{n}\mathbb{E}\left(\frac{\delta_j^k}{q_j^k}-1\right)^2|\langle\boldsymbol{\phi}_j,\mathbf{w}\rangle\langle\boldsymbol{\phi}_j,\boldsymbol{\psi}_i\rangle|^2\\
&\leq\sum_{j=1}^{n}\frac{|\langle\boldsymbol{\phi}_j,\mathbf{w}\rangle\langle\boldsymbol{\phi}_j,\boldsymbol{\psi}_i\rangle|^2}{q_j^k}\\
&\leq\max_j\frac{|\langle\boldsymbol{\phi}_j,\mathbf{w}\rangle|^2}{q_j^k}\sum_{j=1}^{n}|\langle\boldsymbol{\phi}_j,\boldsymbol{\psi}_i\rangle|^2\\
&\leq\max_j\frac{|\langle\boldsymbol{\phi}_j,\mathbf{w}\rangle|^2}{C_0\|\mathrm{P}_{\mathbf{\Pi}}\boldsymbol{\phi}_j\|_2^2\log n}\\
&\leq\frac{\|\mathbf{w}\|_{\mathbf{\Phi},\infty}^2}{C_0\log n}.
\end{aligned}$$

The result follows from scalar Bernstein's inequality in Theorem 4 and a union bound.

APPENDIX H
PROOF OF LEMMA 6

By the definition of the weighted norm, we have

$$\|(\mathrm{P}_{\mathbf{\Pi}}\mathrm{R}_{\Omega_k}\mathrm{P}_{\mathbf{\Pi}}-\mathrm{P}_{\mathbf{\Pi}})\mathbf{w}\|_{\mathbf{\Phi},\infty}=\max_j\frac{|\langle(\mathrm{P}_{\mathbf{\Pi}}\mathrm{R}_{\Omega_k}\mathrm{P}_{\mathbf{\Pi}}-\mathrm{P}_{\mathbf{\Pi}})\mathbf{w},\boldsymbol{\phi}_j\rangle|}{\|\mathrm{P}_{\mathbf{\Pi}}\boldsymbol{\phi}_j\|_2}$$
$$=\max_j\left|\frac{\sum_{i=1}^n\left(\frac{\delta_i^k}{q_i^k}-1\right)\langle\boldsymbol{\phi}_i,\mathrm{P}_{\mathbf{\Pi}}\mathbf{w}\rangle\langle\mathrm{P}_{\mathbf{\Pi}}\boldsymbol{\phi}_i,\boldsymbol{\phi}_j\rangle}{\|\mathrm{P}_{\mathbf{\Pi}}\boldsymbol{\phi}_j\|_2}\right|.$$

Now consider the j-th term, i.e.,

$$\left|\sum_{i=1}^n\frac{\left(\frac{\delta_i^k}{q_i^k}-1\right)\langle\boldsymbol{\phi}_i,\mathrm{P}_{\mathbf{\Pi}}\mathbf{w}\rangle\langle\mathrm{P}_{\mathbf{\Pi}}\boldsymbol{\phi}_i,\boldsymbol{\phi}_j\rangle}{\|\mathrm{P}_{\mathbf{\Pi}}\boldsymbol{\phi}_j\|_2}\right|\overset{\text{def}}{=}|\sum_{i=1}^n\mathcal{X}_i^k|$$

we have that

$$|\mathcal{X}_i^k|=\left|\frac{\left(\frac{\delta_i^k}{q_i^k}-1\right)\langle\boldsymbol{\phi}_i,\mathrm{P}_{\mathbf{\Pi}}\mathbf{w}\rangle\langle\mathrm{P}_{\mathbf{\Pi}}\boldsymbol{\phi}_i,\boldsymbol{\phi}_j\rangle}{\|\mathrm{P}_{\mathbf{\Pi}}\boldsymbol{\phi}_j\|_2}\right|$$
$$\overset{\text{(i)}}{\leq}\left|\frac{\langle\boldsymbol{\phi}_i,\mathbf{w}\rangle\|\mathrm{P}_{\mathbf{\Pi}}\boldsymbol{\phi}_i\|_2}{q_i^k}\right|$$
$$\leq\begin{cases}[C_0\log n\log s]^{-1}\|\mathbf{w}\|_{\mathbf{\Phi},\infty},\ k\in\{1,2\}\\ [C_0\log n]^{-1}\|\mathbf{w}\|_{\mathbf{\Phi},\infty},\ k=3,4...k_0\end{cases}$$

where (i) follows from Cauchy-Swartz and the fact that $\mathbf{w}\in\mathbf{\Pi}$. Moreover, we have

$$\sum_{i=1}^n\mathbb{E}|\mathcal{X}_i^k|^2$$
$$=\sum_{i=1}^n\mathbb{E}\left(\frac{\delta_i^k}{q_i^k}-1\right)^2\left|\frac{|\langle\boldsymbol{\phi}_i,\mathrm{P}_{\mathbf{\Pi}}\mathbf{w}\rangle|^2|\langle\mathrm{P}_{\mathbf{\Pi}}\boldsymbol{\phi}_i,\boldsymbol{\phi}_j\rangle|^2}{\|\mathrm{P}_{\mathbf{\Pi}}\boldsymbol{\phi}_j\|_2^2}\right|$$
$$\leq\frac{1}{\|\mathrm{P}_{\mathbf{\Pi}}\boldsymbol{\phi}_j\|_2^2}\sum_{i=1}^n\left|\frac{|\langle\boldsymbol{\phi}_i,\mathrm{P}_{\mathbf{\Pi}}\mathbf{w}\rangle|^2|\langle\boldsymbol{\phi}_i,\mathrm{P}_{\mathbf{\Pi}}\boldsymbol{\phi}_j\rangle|^2}{q_i^k}\right|$$
$$\overset{\text{(i)}}{\leq}\max_i\frac{|\langle\boldsymbol{\phi}_i,\mathbf{w}\rangle|^2}{q_i^k}\frac{1}{\|\mathrm{P}_{\mathbf{\Pi}}\boldsymbol{\phi}_j\|_2^2}\sum_{i=1}^n|\langle\boldsymbol{\phi}_i,\mathrm{P}_{\mathbf{\Pi}}\boldsymbol{\phi}_j\rangle|^2$$
$$=\max_i\frac{|\langle\boldsymbol{\phi}_i,\mathbf{w}\rangle|^2}{q_i^k}$$
$$\overset{\text{(ii)}}{\leq}\begin{cases}[C_0\log n\log s]^{-1}\|\mathbf{w}\|_{\mathbf{\Phi},\infty}^2,\ k\in\{1,2\}\\ [C_0\log n]^{-1}\|\mathbf{w}\|_{\mathbf{\Phi},\infty}^2,\ k=3,4...k_0\end{cases}$$

where (i) follows from the fact that $\mathbf{w}\in\mathbf{\Pi}$, and (ii) utilizes the specifications of q_i^k of the golfing scheme. The result follows from the scalar Bernstein's inequality in Theorem 4 and a union bound.

APPENDIX I
PROOF OF LEMMA 7

Recall that $\boldsymbol{\gamma}_S=\sum_{i\in S}\mathrm{sgn}(\boldsymbol{\theta})_i\boldsymbol{\psi}_i$, and by definition

$$\|\boldsymbol{\gamma}_S\|_{\mathbf{\Phi},\infty}=\max_j\frac{|\langle\boldsymbol{\phi}_j,\boldsymbol{\gamma}_S\rangle|}{\|\mathrm{P}_{\mathbf{\Pi}}\boldsymbol{\phi}_j\|_2}$$
$$=\max_j\frac{|\sum_{i\in S}\mathrm{sgn}(\boldsymbol{\theta})_i\langle\boldsymbol{\phi}_j,\boldsymbol{\psi}_i\rangle|}{\|\mathrm{P}_{\mathbf{\Pi}}\boldsymbol{\phi}_j\|_2}$$
$$\overset{\text{def}}{=}\max_j\frac{|\sum_{i\in S}\mathcal{X}_i|}{\|\mathrm{P}_{\mathbf{\Pi}}\boldsymbol{\phi}_j\|_2}.$$

Now for for each j we bound the term $|\sum_{i\in S}\mathcal{X}_i|$. We have

$$|\mathcal{X}_i|=|\mathrm{sgn}(\boldsymbol{\theta})_i\langle\boldsymbol{\phi}_j,\boldsymbol{\psi}_i\rangle|\leq\|\mathrm{P}_{\mathbf{\Pi}}\boldsymbol{\phi}_j\|_2$$
$$=\frac{\sqrt{C_0}\|\mathrm{P}_{\mathbf{\Pi}}\boldsymbol{\phi}_j\|_2\log s/4}{\sqrt{C_0}\log s/4},$$

where we have used the fact that $\boldsymbol{\psi}_i\in\mathbf{\Pi}$ for $i\in S$. Moreover, we have for $C_0>1$

$$\sum_{i\in S}\mathbb{E}|\mathcal{X}_i|^2\leq\sum_{i\in S}|\langle\boldsymbol{\phi}_j,\boldsymbol{\psi}_i\rangle|^2=\|\mathrm{P}_{\mathbf{\Pi}}\boldsymbol{\phi}_j\|_2^2$$
$$\leq\frac{C_0\|\mathrm{P}_{\mathbf{\Pi}}\boldsymbol{\phi}_j\|_2^2\log^2 s/16}{\sqrt{C_0}\log s/4}.$$

Then scalar Bernstein's inequality in Theorem 4 and a union bound gives the result.

APPENDIX J
BERNSTEIN'S INEQUALITIES

Theorem 4 (Scalar Bernstein's inequality)**.** *Let $x_1,\cdots,x_m\in\mathbb{R}$ be independent, zero mean random variables. If for all $i=1,\cdots,m$, it holds that $|x_i|\leq B$ and $\sum_{i=1}^m\mathbb{E}|x_i|^2\leq\sigma^2$, then*

$$\mathbf{P}\left(\left|\sum_{i=1}^m x_i\right|>t\right)\leq 2\mathrm{exp}\left(-\frac{t^2/2}{\sigma^2+Bt/3}\right).\qquad(46)$$

Theorem 5 (Vector Bernstein's inequality)**.** *Let $\mathbf{x}_1,\cdots,\mathbf{x}_m\in\mathbb{R}^n$ be independent, zero mean random vectors. If for all $i=1,\cdots,m$, it holds that $\|\mathbf{x}_i\|_2\leq B$ and $\sum_{i=1}^m\mathbb{E}\|\mathbf{x}_i\|_2^2\leq\sigma^2$, then for all $0<t<\sigma^2/B$*

$$\mathbf{P}\left(\left\|\sum_{i=1}^m\mathbf{x}_i\right\|_2>t\right)\leq\exp\left(-\frac{t^2}{8\sigma^2}+\frac{1}{4}\right).\qquad(47)$$

Theorem 6 (Matrix Bernstein's inequality)**.** *Let $\mathbf{X}_1,\cdots,\mathbf{X}_m\in\mathbb{R}^{n\times n}$ be independent, zero mean random matrices. If for all $i=1,\cdots,m$, it holds that $\|\mathbf{X}_i\|\leq B$ and $\max\{\|\sum_{i=1}^m\mathbb{E}\mathbf{X}_i^T\mathbf{X}_i\|,\|\sum_{i=1}^m\mathbb{E}\mathbf{X}_i\mathbf{X}_i^T\|\}\leq\sigma^2$, then for all $c>0$*

$$\left\|\sum_{i=1}^m\mathbf{X}_i\right\|\leq 2\sqrt{c\sigma^2\mathrm{log}2n}+cB\mathrm{log}2n\qquad(48)$$

with probability at least $1-2n^{-(c-1)}$.