A Coherence Parameter Characterizing Generative Compressed Sensing With Fourier Measurements

Aaron Berk†, Simone Brugiapaglia‡, Babhru Joshi§, Yaniv Plan¶, Matthew Scott∥, and Özgür Yilmaz, Member, IEEE

Abstract—In Bora et al. (2017), a mathematical framework was developed for compressed sensing guarantees in the setting where the measurement matrix is Gaussian and the signal structure is the range of a generative neural network (GNN). The problem of compressed sensing with GNNs has since been extensively analyzed when the measurement matrix and/or network weights follow a subgaussian distribution. We move beyond the subgaussian assumption, to measurement matrices that are derived by sampling uniformly at random rows of a unitary matrix (including subsampled Fourier measurements as a special case). Specifically, we prove the first known restricted isometry guarantee for generative compressed sensing (GCS) with subsampled isometries, and provide recovery bounds with nearly order-optimal sample complexity, addressing an open problem of (Scarlett et al., 2022, p. 10). Recovery efficacy is characterized by the coherence, a new parameter, which measures the interplay between the range of the network and the measurement matrix. Our approach relies on subspace counting arguments and ideas central to high-dimensional probability. Furthermore, we propose a regularization strategy for training GNNs to have favourable coherence with the measurement operator. We provide compelling numerical simulations that support this regularized training strategy: our strategy yields low coherence networks that require fewer measurements for signal recovery. This, together with our theoretical results, supports coherence as a natural quantity for characterizing GCS with subsampled isometries.

Index Terms—Generative neural network, subsampled isometry, compressed sensing, coherence, Fourier measurements.

I. INTRODUCTION

The solution of underdetermined linear inverse problems has many important applications including geophysics [3], [4] and medical imaging [5], [6]. In particular, compressed sensing permits accurate and stable recovery of signals that are well represented by one of a certain set of structural proxies (e.g., sparsity) [5], [7]. Moreover, this recovery is effected using an order-optimal number of random measurements [7]. In applications like medical imaging [5], the measurement matrices under consideration are derived from a bounded orthonormal system (a unitary matrix with bounded entries), which complicates the theoretical analysis. Furthermore, for such applications one desires a highly effective representation for encoding the images. Developing a theoretical analysis that properly accounts for realistic measurement paradigms and complexly designed image representations is nontrivial in general [1], [7], [8]. For example, there has been much work validating that generative neural networks (GNNs) are highly effective at representing natural signals [9], [10], [11]. In this vein, recent work has shown promising empirical results for compressed sensing with realistic measurement matrices when the structural proxy is a GNN. Other recent work has established recovery guarantees for compressed sensing when the structural proxy is a GNN and the measurement matrix is subgaussian [1]. (See Section I-A for a fuller depiction of related aspects of this problem.) However, an open problem is the following [2, p. 10]:

Open problem 1 (SubIso GCS): A theoretical analysis of compressed sensing when the measurement matrix is structured (e.g., a randomly subsampled unitary matrix) and the signal model proxy is a GNN.

Broadly, we approach a solution to subIso GCS as follows. For a matrix $A \in \mathbb{C}^{m \times n}$, a particular GNN architecture $G : \mathbb{R}^n \to \mathbb{R}^m$ and an unknown signal $x_0 \in \mathbb{R}(G)$, the range of $G$, we determine the conditions (on $A$, $G$, $x_0$, etc.) under which it is possible to approximately recover $x_0$ from noisy linear measurements $b = Ax_0 + \eta$ by (approximately) solving an optimization problem of the form

$$\min_{z \in \mathbb{R}^d} \| b - AG(z) \|_2.$$  \hspace{1cm} (1)

Above, $\eta \in \mathbb{C}^m$ is some unknown corruption. Specifically, we are interested in establishing sample complexity bounds (lower...
bounds on $m$) for realistic measurement matrices $A$ — where $A$ is an underdetermined matrix randomly subsampled from a unitary matrix. Namely, the rows of $A$ have been sampled uniformly at random without replacement from a unitary matrix $U \in \mathbb{C}^{n \times n}$. We next present a mathematical description of the subsampling of the rows, described similarly to [12, Sec. 4].

Definition 1 (Subsampled Isometry): Let $2 \leq m \leq n < \infty$ be integers and let $U \in \mathbb{C}^{n \times n}$ be a unitary matrix. Let $\theta := (\theta_i)_{i \in [n]}$ be an iid Bernoulli random vector; $\theta_i \sim \text{Ber}(m/n)$. Define the set $\mathcal{J} := \{j : \theta_j = 1\}$ and enumerate the elements of $\mathcal{J}$ as $j_1, \ldots, j_m$ where $|\mathcal{J}| = |\mathcal{J}|$ is a binomial random variable with $\mathbb{E} \mathcal{J} = m$. Let $A \in \mathbb{C}^{m \times n}$ be the matrix whose $i$th row is $\frac{\theta_i}{\sqrt{m}} U_{ji}, i \in [m]$. We call $A$ an $(m, U)$-subsampled isometry. When there is no risk of confusion we simply refer to $A$ as a subsampled isometry and implicitly acknowledge the existence of an $(m, \theta, U)$ giving rise to $A$.

With $A$ so defined, $A$ is isotropic: $\mathbb{E} A^* A = \sum_{i=1}^n U_i^* U_i = I_n$ where $I_n$ is the $n \times n$ identity matrix.

Remark 1: An important example of a matrix isometry is the discrete orthogonal system given by the (discrete) Fourier transform (DFT) matrix, and has important roles in signal processing and numerical computation [2], [5], [8]. Thus, all of our results apply, in particular, when the measurement matrix is a subsampled DFT.

Lastly, we introduce the kind of GNN to which we restrict our attention in this work. Namely, we study ReLU-activated expansive neural networks, where ReLU is the so-called rectified linear unit defined as $\sigma(x) := \max(x, 0)$, acting element-wise on the entries of $x$.

Definition 2 ((k, d, n)-Generative Network): Fix the integers $2 \leq k : k_0 \leq k_1, \ldots, k_d$ where $k_d := n < \infty$, and suppose for $i \in [d]$ that $W^{(i)}(i) \in \mathbb{R}^{k_i \times k_{i-1}}$. A $(k, d, n)$-generative network is a function $G : \mathbb{R}^k \rightarrow \mathbb{R}^n$ of the form

$$G(z) := W^{(d)} \sigma \left( \cdots W^{(2)} \sigma \left( W^{(1)} z \right) \right).$$

Remark 2: In practice, ReLU generative networks often use biases, which are learned parameters in addition to the weight matrices. Such networks have the form

$$G_{\text{bias}}(z) := W^{(d)} \sigma \left( \cdots \sigma \left( W^{(1)} z + b^{(1)} \right) \right) + b^{(d)},$$

where $b^{(i)} \in \mathbb{R}^{k_i}, i \in [d]$. Let $\mathcal{H} := \{z \in \mathbb{R}^{k+1} : \sum_{k+1} = 1\}$ and define the augmented matrices $\hat{W}^{(i)} := \begin{bmatrix} W^{(i)} & b^{(i)} \end{bmatrix}$. In the last layer, remove the last row of the augmented matrix. Let $\hat{G} : \mathbb{R}^{k+1} \rightarrow \mathbb{R}^n$ be the $(k + 1, d, n)$-generative network having weight matrices $\hat{W}^{(i)}$. Since $G_{\text{bias}} = \hat{G}|_{\mathcal{H}}$, we have $\mathcal{R}(G_{\text{bias}}) \subset \mathcal{R}(\hat{G})$ (i.e., any network with biases has range contained in a similar network without biases that has code dimension augmented by 1). Our theory applies directly to the network without biases, $\hat{G}$, and due to the containment, all results given for $\hat{G}$ extend to the biased network $G_{\text{bias}}$ (even so for Theorem 3, see Remark S2.4).

With these ingredients, we provide a suggestive “cartoon” of the main theoretical contribution of this work, which itself can be found in Theorem 1.

Theorem Sketch 1 (Cartoon): Let $G$ be a $(k, d, n)$-generative network, and $A$ a subsampled isometry. Suppose $\mathcal{R}(G) - \mathcal{R}(G)$ is “incoherent” with respect to the rows of $A$, quantified by a parameter $\alpha > 0$. If the number of measurements $m$ satisfies $m \geq \kappa d n^2$ (up to log factors), then, with high probability on $A$, it is possible to approximately recover an unknown signal $x_0 \in \mathcal{R}(G)$ from noisy underdetermined linear measurements $b = A x_0 + \eta$ with nearly order-optimal error.

The coherence parameter $\alpha$ is defined below in Definition 4 using the measurement norm introduced in Definition 3. The quantification of $\alpha$ is discussed thereafter, and fully elaborated in Section III. The notion of “incoherence” in the cartoon above is specified in Corollary 1. Coherence is related to the concept of incoherent bases [7, p. 373], while the measurement norm is closely related to the so-called $X$-norm in [13]. Effectively, coherence characterizes the alignment between the components comprising $\mathcal{R}(G)$ and the row vectors $A_i$ of the subsampled isometry $A$.

Definition 3 (Measurement Norm): Let $G \in \mathbb{C}^{n \times n}$ be a unitary matrix. Define the norm $\| \cdot \|_U : \mathbb{C}^n \rightarrow [0, \infty)$ by

$$\|x\|_U := \|Ux\|_{\infty} = \max_{i \in [n]} |\langle U_i, x \rangle|.$$

Definition 4 (Coherence): Let $T \subset \mathbb{R}^d$ be a set and $U \in \mathbb{C}^{n \times n}$ a unitary matrix. For $\alpha > 0$, say that $T$ is $\alpha$-coherent with respect to $\| \cdot \|_U$ if

$$\sup_{x \in T, x \in \mathbb{C}^{n-1}} \|x\|_U \leq \alpha.$$

We refer to the quantity on the left-hand side as the coherence.

The idea is that the structural proxy/prior under consideration should be incoherent with respect to the measurement process. Thus, we desire that chords in $\mathcal{R}(G)$ be not too closely aligned (in the sense controlled by $\alpha$) with the rows of $U$, with which the subsampled isometry $A$ is associated. This follows a paradigm from classic compressed sensing: democracy of the measurement process, i.e., no single measurement should be essential for signal recovery, rather the measurements should be used jointly. This is natural, since we are randomly sampling, and any potential measurement may not be sampled. The standard definition of coherence in compressed sensing, and its nonrandom origins, involves a bound on inner products between columns of the sensing matrix [7, Ch. 5]; it applies to deterministic measurement matrices, and is somewhat different than our definition. There is also a definition of incoherence in [7, Ch. 12] (see also [14]) or incoherence property [15]. The latter two are defined in the setup of random sampling and are somewhat analogous to the parameter we defined.

Though it is likely difficult to measure coherence precisely in practice, we propose a computationally efficient heuristic

---

1We will need to expand this set slightly via Definition 5.
that upper bounds the coherence. For perspective, we show that if $G$ has Gaussian weights, one may take $\alpha^2 \sim kd/n$ in Cartoon (see Theorem 1 and Theorem 3), thereby giving a sample complexity, $m$, proportional to $(kd)^2$ up to log factors. We leave improving the quadratic dependence as an open question, discussed further in Section VI.

We briefly itemize the main contributions of this paper:

- we introduce the coherence for characterizing recovery efficacy via the alignment of the network’s range with the measurement matrix (see Definition 4);
- we establish a restricted isometry property for $(k, d, n)$-generative networks with subsampled isometries (see Theorem 2 and Corollary 1);
- we prove sample complexity and recovery bounds in this setting (see Theorem 1);
- we propose a regularization strategy for training GNNs with low coherence (see Section IV-A) and demonstrate improved sample complexity for recovery (see Section IV-B);
- together with our theory, we provide compelling numerical simulations that support coherence as a natural quantity of interest linked to favourable deep generative recovery (see Section IV-B).

A. Related Work

Theoretically, [1] have analyzed compressed sensing problems in the so-called generative prior framework, focusing on Gaussian or subgaussian measurement matrices. This led to much follow-up work in the generative prior framework, albeit none in the subsampled Fourier setting to our knowledge. For example, [16] extends the analysis to the setting of demixing with subgaussian matrices, while [17] analyzes the semi-parametric single-index model with generative prior under Gaussian measurements. Finally, exact recovery of the underlying latent code for GNNs (i.e., seeking $z \in \mathbb{R}^k$ such that $x = G(u)$) has been analyzed; however, these analyses rely on the GNN having a suitable structure with weight matrices that possess a suitable randomness [18], [19], [20], [21]. For a review of these and related problems, see [2].

Promising empirical results of [8] suggest remarkable efficacy of generative compressed sensing (GCS) in realistic measurement paradigms. Furthermore, the authors provide a framework with theoretical guarantees for using Langevin dynamics to sample from a generative prior. Several recent works have developed sophisticated generative adversarial networks (GANs) (which are effectively a type of GNN) for compressed sensing in medical imaging [22], [23]. Other work has empirically explored multi-scale (non-Gaussian) sampling strategies for image compressed sensing using GANs [24]. Separately, see [25] for the use of GCS in uncertainty quantification of high-dimensional partial differential equations with random inputs. Recently popular is the use of untrained GNNs for signal recovery [26], [27]. For instance, [28] executed a promising empirical investigation of medical image compressed sensing using untrained GNNs.

Compressed sensing with subsampled isometries is well studied for sparse signal recovery. The original works developing such recovery guarantees are [29], [30], with improvements appearing in [13], [31]. See [7] for a thorough presentation of this material including relevant background. See [12, Sec. 4] for a clear presentation of this material via an extension of generic chaining. In this setting, the best-known number of log factors in the sample complexity bound sufficient to achieve the restricted isometry property is due to [32] with subsequent extensions and improvements in [33], [34], [35]. Reference [36] address compressed sensing with subsampled isometries when the structural proxy is a neural network with random weights.

Using a notion of coherence to analyze the solution of convex linear inverse problems was proposed in [29], [37]. Reference [38] relate this notion to the matrix norm $\| \cdot \|_{2 \rightarrow \infty}$ (defined in Section I-B) in order to analyze covariance estimation and singular subspace recovery. Additionally, see [14] or [7, p. 373] for a discussion of incoherent bases, and [13, p. 1034] for the analogue of our measurement norm in the sparsity case.

The present work relies on important ideas from high-dimensional probability, such as controlling the expected supremum of a random process on a geometric set. These ideas are well treated in [39], [40]; see [41] for a thorough treatment of high-dimensional probability. This work also relies on counting linear regions comprising the range of a ReLU-activated GNN. In this respect, we rely on a result that appears in [42]. Tighter but less analytically tractable bounds appear in [43], while a computational exploration of region counting has been performed in [44].

B. Notation

For an integer $n \geq 1$ denote $[n] := \{1, \ldots, n\}$. For $x \in \mathbb{C}^n$, denote the $\ell_p$ norm for $1 \leq p < \infty$ by $\|x\|_p := (\sum_{i=1}^n |x_i|^p)^{1/p}$ and for $p = \infty$ by $\|x\|_\infty := \max_{i \in [n]} |x_i|$. Here, if $x \in \mathbb{C}$ then $|x| = \sqrt{\text{Re}(x)^2 + \text{Im}(x)^2}$ and the conjugate is given by $\tilde{x} := \text{Re}(x) - i\text{Im}(x)$. If $X \in \mathbb{C}^{m \times n}$ is a matrix then the conjugate transpose is denoted $X^* = (\tilde{X})_{i \in [n], j \in [m]}$. The $\ell_p$ norm for real numbers, $1 \leq p \leq \infty$ is defined in the standard, analogous way. Denote the real and complex sphere each by $S^{n-1} := \{x : \|x\|_2 = 1\}$, disambiguating only where unclear from context. The operator norm of a matrix $X \in \mathbb{C}^{m \times n}$, induced by the Euclidean norm, is denoted $\|X\| := \sup_{\|x\|_2=1} \|Xx\|_2$. Unless otherwise noted, $X_i$ denotes the $i$th row of the matrix $X$, viewed as a column vector. The Frobenius norm of $X$ is denoted $\|X\|_F$ and satisfies $\|X\|^2_F = \sum_{i=1}^m \|X_i\|^2_2$. The matrix norm $\| \cdot \|_{p \rightarrow q}$ for $1 \leq p, q \leq \infty$ is $\|X\|_{p \rightarrow q} := \sup_{x \neq \theta} \frac{\|Xx\|_q}{\|x\|_p}$. We use $\Pi_{\mathcal{L}}$ to denote the standard $\ell_2$ projection operator onto the set $\mathcal{L}$, which selects a single point lexicographically, if necessary, to ensure uniqueness. $\text{Ber}(p)$ denotes the Bernoulli distribution with parameter $p$; $\text{Binom}(n, p)$ the binomial distribution for $n$ items with rate $p$.

Throughout this work, $C > 0$ represents an absolute constant having no dependence on any parameters, whose value may change from one appearance to the next. Constants with dependence on a parameter will be denoted with an appropriate subscript — e.g., $C_\delta$ is an absolute constant depending only on a parameter $\delta$. Likewise, for two quantities $a, b$, if
a ≤ b then a ≤ Cb; analogously for a ≥ b. Finally, given two sets A, B ⊆ \mathbb{R}^n, A ± B denotes the Minkowski sum/difference: A ± B := \{a ± b : a ∈ A, b ∈ B\}. Similarly, for a ∈ \mathbb{R}^n, a − B := \{a − b : b ∈ B\} and aB := \{ab : b ∈ B\}. The range of a function f : \mathbb{R}^n → \mathbb{R}^m is denoted \text{R}(f) := \{f(x) : x ∈ \mathbb{R}^n\} (e.g., if X is a matrix then \text{R}(X) denotes the column space of X). As above, \sigma(x) := \max(x, 0), which may act element-wise on a vector.

II. MAIN RESULTS

Proofs of results in this section are deferred to Section V-A.

Observe, if G is a (k, d, n)-generative network, then \text{R}(G) and \mathcal{G} := \text{R}(G) − \text{R}(G) are unions of polyhedral cones (see Lemma S2.2 and Remark S2.3). Note that polyhedral cones (Definition S2.2) are convex. We introduce the following definition to expand each cone into a full subspace.

Definition 5: Let \mathcal{C} ⊆ \mathbb{R}^n be the union of N convex cones: \mathcal{C} = \bigcup_{i=1}^N C_i. Define the piecewise linear expansion

\Delta(\mathcal{C}) := \bigcup_{i=1}^N \text{span}(C_i) = \bigcup_{i=1}^N (C_i − C_i).\n
The second equality follows from Proposition S3.1. See Remark S3.1 for a list of properties of \Delta, including uniqueness. Note each cone comprising \text{R}(G) has dimension at most k, hence \Delta(\text{R}(G)) is a union of linear subspaces each having dimension at most k.

We now present the main result of the paper, which establishes sample complexity and recovery bounds for generative compressed sensing with subsampled isometries. Below, \hat{x} := x_0 − \Pi_{\text{R}(G)} x_0.

Theorem 1 (Subsampled Isometry GCS): Let G : \mathbb{R}^k → \mathbb{R}^n be a (k, d, n)-generative network with layer widths k_i ≤ k_0 ≤ k_1, . . . , k_d where k_0 := n, \delta, \epsilon > 0, \mathcal{G} := \text{R}(G) − \text{R}(G) and A ∈ \mathbb{C}^{m×n} a subsampled isometry associated with a unitary matrix U ∈ \mathbb{C}^{n×n}. If \Delta(\mathcal{G}) is \alpha-3-\text{coherent} with respect to \|·\|_2, and

\begin{align*}
    m &\geq \alpha^2 n \left(2k \sum_{i=1}^{d-1} \log \left(\frac{2e k_i}{k_0}\right) + \log \frac{4k}{\epsilon}\right),
\end{align*}

then the following holds with probability at least 1 − \epsilon on the realization of A.

For any x_0 ∈ \mathbb{R}^n, let b := A x_0 + \eta where \eta ∈ \mathbb{C}^m. Let \hat{x} ∈ \mathbb{R}^n satisfy \|A\hat{x} − b\|_2 ≤ \min_{x ∈ \Delta(\text{R}(G))} \|Ax − b\|_2 + \delta. Then,

\begin{align*}
    \|\hat{x} − x_0\|_2 &\leq \|x\|_2 + 3\|A^{-1}\|_2 + 3\|\eta\|_2 + \frac{3\delta}{2}.
\end{align*}

Remark 3: Since \text{R}(G) is an union of polyhedral cones, \mathcal{G} ⊆ \Delta(\mathcal{G}) ⊆ \mathcal{G} − \mathcal{G}. Hence, it is sufficient to assume \mathcal{G} − \mathcal{G} be \alpha-3-\text{coherent} with respect to \|·\|_2. This containment may aid practitioners to control \alpha since one may sample from \mathcal{G} − \mathcal{G}.

Remark 4: The approximation error \|x\|_2 is controlled by the expressivity of G, satisfying (by definition)

\begin{align*}
    \|x\|_2 &\leq \min_{u ∈ \mathbb{R}^k} \|G(u) − x_0\|_2.
\end{align*}

The modelling error incurred via \|Ax\|_2 could be large compared to \|x\|_2 : \|Ax\|_2 ≤ \frac{\sqrt{n}}{\sqrt{m}}\|x\|_2 in general. However, if G admits a good representation of the modelled data distribution, then one might expect this term still to be small.

Certainly, if x_0 ∈ \mathbb{R}(G), the final expression in Theorem 1 reduces to

\begin{align*}
    \|\hat{x} − x_0\|_2 &\leq 3\|\eta\|_2 + \frac{3\delta}{2}.
\end{align*}

Otherwise, if x_1 is independent of A, \mathbb{E}\|A x_1\|_2 ≤ \|x_1\|_2 by Jensen’s inequality. Thus, by Markov’s inequality one has \mathbb{P}(\|A x_1\|_2 ≥ \kappa \|x_1\|_2) ≤ \kappa^{-1}. Finally, a strategy for more precisely controlling \|A x\|_2 is given in Section S4 (see especially Proposition S4.1).

Analogous to the restricted isometry property of compressed sensing or the set-restricted eigenvalue condition of [1], the proof of Theorem 1 relies on a restricted isometry condition. This condition guarantees that pairwise distances of points in \mathcal{G} are approximately preserved under the action of A. We first state a result controlling norms of points in \mathcal{G} under the action of A; control over pairwise distances then follows easily.

Theorem 2 (Gen-RIP): Let A ∈ \mathbb{C}^{m×n} be a subsampled isometry associated with a unitary matrix U ∈ \mathbb{C}^{n×n} and \delta, \epsilon > 0. Suppose that G : \mathbb{R}^k → \mathbb{R}^n is a (k, d, n)-generative network with layer widths k = k_0 ≤ k_1, . . . , k_d where k_d := n, G := \text{R}(G) − \text{R}(G) and A ∈ \mathbb{C}^{m×n} a subsampled isometry associated with a unitary matrix U ∈ \mathbb{C}^{n×n}. If \Delta(\mathcal{G}) is \alpha-3-\text{coherent} with respect to \|·\|_2, and

\begin{align*}
    m &\geq \alpha^2 n \left(2k \sum_{i=1}^{d-1} \log \left(\frac{2e k_i}{k_0}\right) + \log \frac{2k}{\epsilon}\right),
\end{align*}

then with probability at least 1 − \epsilon on the realization of A, it holds that

\begin{align*}
    \sup_{x ∈ \text{R}(G) \cap \mathbb{C}^{n×n}} \|Ax\|_2 − 1 &\leq \delta.
\end{align*}

Remark 5: In Section III we show that \alpha can have dependence on n proportional to n^{−1/2}, ignoring log factors (see Proposition 1 and Theorem 3). Therefore, the sample complexity can be independent of the ambient dimension n (again ignoring log factors).

Remark 6: Analogous to Remark 3, in Theorem 2 it is sufficient to assume \alpha-3-\text{coherence} of \mathcal{G} − \mathcal{G}, since \mathcal{G} ⊆ \Delta(\mathcal{G}) ⊆ \mathcal{G} − \mathcal{G}.

We now state the result that provides the notion of restricted isometry needed for Theorem 1. This result, which controls pairwise differences of elements in \mathcal{G}, is an immediate consequence of Theorem 2 using the observation in Remark S2.3.

Corollary 1 (Restricted Isometry on the Difference Set): Let G : \mathbb{R}^k → \mathbb{R}^n be a (k, d, n)-generative network with layer widths k = k_0 ≤ k_1, . . . , k_d where k_d := n, G := \text{R}(G) − \text{R}(G), \delta, \epsilon > 0, and suppose A ∈ \mathbb{C}^{m×n} is a subsampled isometry associated with a unitary matrix U ∈ \mathbb{C}^{n×n}. Assume that \Delta(\mathcal{G}) is \alpha-3-\text{coherent} with respect to \|·\|_2. If

\begin{align*}
    m &\geq \alpha^2 n \left(2k \sum_{i=1}^{d-1} \log \left(\frac{2e k_i}{k_0}\right) + \log \frac{4k}{\epsilon}\right),
\end{align*}

then with probability at least 1 − \epsilon on the realization of A, it holds that

\begin{align*}
    \sup_{x ∈ \text{R}(G) \cap \mathbb{C}^{n×n}} \|Ax\|_2 − 1 &\leq \delta.
\end{align*}
then with probability at least $1 - \varepsilon$ on the realization of $A$, it holds that
\[ \sup_{x \in \mathbb{G}^{d-1}} \|Ax\|_2 - 1 \leq \delta. \]

**Remark 7:** In fact, the proof of Theorem 2 yields the stronger restricted isometry bound
\[ \sup_{x \in \Delta(\mathcal{R}(G)) \cap \mathbb{G}^{d-1}} \|Ax\|_2 - 1 \leq \delta. \]
Consequently, the restricted isometry bound in Corollary 1 can be strengthened to
\[ \sup_{x \in \Delta(\mathcal{R}(G)) \cap \mathbb{G}^{d-1}} \|Ax\|_2 - 1 \leq \delta. \]

The proofs of Theorem 2 and Theorem 1 are deferred to Section V-A. The result from Lemma 1 and Lemma S2.2, the former of which is presented next. It characterizes restricted isometry of a subspace incoherent with $\| \cdot \|_U$. Its proof is deferred to Section V-A.

**Lemma 1 (RIP for Incoherent Subspace):** Let $A \in \mathbb{C}^{n \times n}$ be a subsampled isometry associated with a unitary matrix $U \in \mathbb{C}^{n \times n}$. Suppose that $\mathcal{L} \subseteq \mathbb{R}^n$ is a $k$-dimensional subspace that is $\alpha$-coherent with respect to $\| \cdot \|_U$. Then, for any $0 \leq \delta \leq 1$,
\[ P \left\{ \sup_{x \in \mathcal{L} \cap \mathbb{G}^{d-1}} \|Ax\|_2 - 1 \geq \delta \right\} \leq 2k \exp \left(- \frac{C\delta^2 m}{\alpha^2 n} \right). \]

**Remark 8:** Convincing empirical results of [44] suggest the number of linear regions for empirically observed neural networks may typically be linear in the number of nodes, rather than exponential in the width. Such a reduction would be a boon for the sample complexity obtained in Theorem 2, which depends on the number of linear regions comprising $\mathcal{R}(G)$ (using Lemma S2.2; see Section V-A).

### III. Typical Coherence

Proofs for results in this section are deferred to Section V-B. The first result of this section establishes a lower bound on the coherence parameter. Together with Corollary 1 this yields a quadratic “bottleneck” on the sample complexity in terms of the parameter $k$.

**Proposition 1:** For a unitary matrix $U \in \mathbb{C}^{n \times n}$, any $k$-dimensional subspace $T \subseteq \mathbb{R}^n$ has coherence with respect to $\| \cdot \|_U$ of at least $\sqrt{\frac{k}{n}}$. Furthermore, this lower bound is tight.

Under mild assumptions, when the generative network has random weights one may show that this is a typical coherence level between the network and the measurement operator.

**Theorem 3:** Let $U \in \mathbb{C}^{n \times n}$ be a unitary matrix and $G$ be a $(k,d,n)$-generative network with layer widths $k = k_0 \leq k_1, \ldots, k_d$ where $k_d := n$. Let the last weight matrix of $G$, $W^{(d)}$ be iid Gaussian: $W^{(d)}_{i,j} \sim \mathcal{N}(0, 1)$, $i \in [k_d], j \in [k_{d-1}]$. Let all other weights be arbitrary and fixed. Then, for any $\gamma \geq 0$, it holds with probability at least $1 - 2 \exp(-\gamma^2)$ that $\Delta(\mathcal{R}(G) - \mathcal{R}(G))$ is $\alpha$-coherent with respect to $\| \cdot \|_U$, where
\[ \alpha \leq \sqrt{\frac{k}{n}} + \sqrt{\frac{\log n}{n}} + \sqrt{\frac{k}{n} \sum_{i=1}^{d-1} \log \frac{2e k_i}{k}} + \sqrt{\frac{\gamma}{n}}. \]

**Remark 9:** We briefly comment on the behaviour of the third term, which, we argue, dominates for the principal case of interest. Assume the layers have approximately constant size: i.e., for two absolute constants $C_1, C_2 > 0$,
\[ \forall \ell \in [d], \quad C_1 \leq \log \frac{e k_i}{k} \leq C_2. \]

In this case, all terms in the sum in the third term will be of the same order, making this term have order $O\left(\sqrt{\frac{k}{n}}\right)$. If we further make the reasonable assumption that $dk > \log(n)$, then the third term dominates all others, hence
\[ \alpha = O\left(\sqrt{\frac{kd}{n}}\right). \]

**Remark 10:** Using Corollary 1 and Remark 9, one may take as the sample complexity for Theorem 1, in the case of a $(k,d,n)$-generative network with Gaussian weights,
\[ m \gtrsim 2k^2d \sum_{i=1}^{d-1} \log \left(\frac{2e k_i}{k}\right) + kd \frac{4k}{\varepsilon}. \]
We note in passing that an argument specialized to random weights is given in [36], with an improved sample complexity. Our goal in this section is not to find the optimal sample complexity for random weights, but to show the average case behaviour of the parameter $\alpha$.

### IV. Numerics

In this section we explore the connection between coherence and recovery error empirically, to suggest that coherence is indeed the salient quantity dictating recovery error. In addition, we propose a regularization strategy to train low coherence GNNs. This regularization strategy is new to our knowledge. The first experiment illustrates a phase portrait that empirically shows dependence on a coherence (proxy) and number of measurements for successful recovery. We also show, for a fixed number of measurements, that the probability of recovery failure increases with higher coherence (proxy). In the second experiment, we use the novel regularization approach to show that fewer measurements are required for signal recovery when a GNN is trained to have low coherence.

#### A. Experimental Methodology

1) **Coherence Heuristic and Regularization:** Ideally, in these experiments, one would calculate the coherence of the network exactly, via Definition 4. However, computing coherence is likely intractable in general. Instead, we use an upper bound on the coherence obtained as follows. Let $G$ be a $(k,d,n)$-generative network and let $W = W^{(d)}$ be its final weight matrix. Write the QR decomposition of $W$ as
\[ W = QR, \quad Q := \left[ Q_1 \ Q_2 \right], \quad R := \left[ R_1 \ 0 \right], \]
where $Q \in \mathbb{R}^{n \times n}$ is orthogonal, $R \in \mathbb{R}^{n \times k}$ has invertible submatrix $R_1 \in \mathbb{R}^{k \times k}$ and $Q_1 \in \mathbb{R}^{n \times k}$ is the submatrix multiplying with $R_1$. Let $G := \mathcal{R}(G) - \mathcal{R}(G)$, $W := \mathcal{R}(W) \cap S^{n-1}$ and let $D \in \mathbb{R}^{n \times n}$ be an orthogonal matrix. Using that
Δ(\mathcal{G}) \cap \mathbb{S}^{n-1} \subseteq \mathcal{W}, we bound the coherence with respect to \| \cdot \|_D as

\sup_{x \in \Delta(\mathcal{G}) \cap \mathbb{S}^{n-1}} \| Dx \|_\infty \leq \sup_{x \in \mathcal{W}} \| Dx \|_\infty

= \max_{i \in [n]} \max_{\gamma \in \mathbb{R}} \left\{ \| D_i^\top \gamma R_c \| : \| R_c \|_2 = 1 \right\}

= \max_{i \in [n]} \max_{\gamma \in \mathbb{R}} \left\{ \| D_i^\top Q_i v \| : \| v \|_2 = 1 \right\}

= \max_{i \in [n]} \| Q_i^\top D_i \|_2 = \| DQ_1 \|_{2 \to \infty}. \tag{2}

where the penultimate line uses \( z := R_i^{-1}v \). To re-phrase: \( \Delta(\mathcal{G}) \)

is always \( \| DQ_1 \|_{2 \to \infty} \)-coherent with respect to \( \| \cdot \|_D \). Our experiments and theory are consistent with the hypothesis that this is an effective heuristics for coherence.

Motivated by (2), we propose a strategy — novel, to our knowledge — to promote low coherence of the final layer \( W \) with respect to a fixed orthogonal matrix \( D \). This is achieved by applying the following regularization \( \rho \) to the final weight matrix of the GNN during training:

\[ \rho(W) = \| DW \|_{2 \to \infty} + \lambda \| W^T W - I \|_F, \tag{3} \]

Namely, the regularizer \( \rho \), with a fixed regularization parameter \( \lambda \geq 0 \), is added to the training loss function. Roughly, this regularization promotes low coherence because \( \| W^T W - I \|_F \) is smallest when \( W \) is orthonormal, making \( \| DW \|_{2 \to \infty} \) the coherence of \( \mathcal{R}(W) \) with respect to \( D \).

2) Network Architectures: In the experiments, we use three generative neural networks trained on the MNIST dataset [45], which consists of 60,000 28 × 28 images of handwritten digits. The GNNs are fully connected networks with three layers and parameters \( k = 20, k_1 = k_2 = 500, n = 784 \). Precisely, let the first one be \( G^{(1)} = s(W^{(1)} \sigma(W^{(1)} \sigma(W^{(1)} \sigma(x)))) \), where \( s(x) = (1 + \exp(-x))^{-1} \) is the sigmoid activation function. Let the remaining two GNNs be \( G^{(i)}(z) = W^{(i)} \sigma(W^{(i)} \sigma(W^{(i)} \sigma(z))) \), \( i = 2, 3 \). We use \( G^{(1)} \), which has a more realistic architecture for real applications, as a point of comparison with \( G^{(i)}, i = 2, 3 \). Variational autoencoders (VAEs) [9], with the decoder network as \( G^{(1)} \) and \( G^{(2)} \), were trained using the Adam optimizer [46] with a learning rate of 0.001 and a mini-batch size of 64 using Flux [47]. We trained another VAE with decoder network \( G^{(3)} \), using the same hyperparameters but using the regularization strategy described in Section IV-A1 to promote low coherence of the final layer \( \mathcal{W}^{(3,3)} \) with respect to a fixed orthogonal matrix \( D \). Specifically, the expression \( 10^4 \rho(W^{(3,3)}) \), with \( \lambda \) set to 1, was added to the VAE loss function. In all cases the VAE loss function was the usual one. See [48] for specific implementation details including the definition of the encoders, and refer to [9], [49] for further background on VAEs.

3) Measurement Matrix: Throughout the experiments, the matrix \( D \) was chosen to be the discrete cosine transform (DCT) matrix. For DCT implementation details, see for instance [50, fftpack.dct]. The matrix \( A \) is a slight variation of the subsampled isometry defined in Definition 1, modified to ensure that each realization of \( A \) has \( m \) rows. Namely, the random matrix \( A \) is subsampled from \( D \) by selecting the first \( m \) elements of a uniform random permutation of \([n]\). Note \( A \) is still re-normalized as in Definition 1.

4) First Experiment: For the first experiment, let \( G_\beta \) be a \((k, 2, n)\)-generative network with inner layers \( W^{(i)} = W^{(1)}, i = 1, 2 \) and last layer \( W^{(3)} = W_\beta \in \mathbb{R}^{n \times k} \) defined by

\[ W_\beta := \beta W^{(1,3)} + (1 - \beta) W^{(3,3)}. \]

Recall that \( W^{(1,3)} \) and \( W^{(3,3)} \) are the final layers of \( G^{(1)} \) and \( G^{(3)} \), respectively. Here, \( \beta \in [0, 1] \) is an interpolation parameter. The coherence, which was computed via (2), of \( \mathcal{R}(W^{(1,3)}) \) was 0.98, while the coherence of \( \mathcal{R}(W^{(3,3)}) \) was 0.82. As a result, for large \( \beta \), one should expect \( W_\beta \) to have large coherence with respect to \( \| \cdot \|_D \). We randomly sample \( z_0 \in \mathbb{R}^k \), fix the number of measurements \( m \in \{40, 60, \ldots, 440\} \), and set \( b = AG_\beta(z_0) \). For each measurement size \( m \) and coherence upper bound, we perform 20 independent trials. For each trial, we approximately solve (1) by running ADAM with a learning rate of 0.1 for 5000 iterations, or until the norm of the gradient is less than \( 10^{-7} \), and set \( \hat{x} \) to be the output. See [48] for specific implementation details. We say the target signal \( G_\beta(z_0) \) was successfully recovered if the relative reconstruction error (re) between \( G_\beta(z_0) \) and \( G_\beta(\hat{x}) \) is less than \( 10^{-5} \):

\[ \text{re}(x_0, \hat{x}) := \frac{\| x_0 - \hat{x} \|_2}{\| x_0 \|_2}. \]

5) Second Experiment: For the second experiment, we use each trained network \( G^{(i)}, i = 1, 2, 3 \). The coherence upper bounds of \( \Delta(\mathcal{R}(G^{(2)}) - \mathcal{R}(G^{(2)})) \) and \( \Delta(\mathcal{R}(G^{(3)}) - \mathcal{R}(G^{(3)})) \), computed using (2) are 0.96 and 0.81, respectively, which empirically shows that the regularization (3) promotes low coherence during training. For the networks \( G^{(i)} \), let \( E^{(i)} : \mathbb{R}^n \to \mathbb{R}^k \) be the corresponding encoder network from their shared VAE. We randomly sample an image \( x_0^i \) from the test set of the MNIST dataset and let \( x_0^i = G^{(i)}(E^{(i)}(x_0^2)) \). We use \( G^{(i)} \) which has a more realistic architecture for real applications, as a point of comparison with \( G^{(i)}, i = 2, 3 \). Variational autoencoders (VAEs) [9], with the decoder network as \( G^{(1)} \) and \( G^{(2)} \), were trained using the Adam optimizer [46] with a learning rate of 0.001 and a mini-batch size of 64 using Flux [47]. We trained another VAE with decoder network \( G^{(3)} \), using the same hyperparameters but using the regularization strategy described in Section IV-A1 to promote low coherence of the final layer \( \mathcal{W}^{(3,3)} \) with respect to a fixed orthogonal matrix \( D \). Specifically, the expression \( 10^4 \rho(W^{(3,3)}) \), with \( \lambda \) set to 1, was added to the VAE loss function. In all cases the VAE loss function was the usual one. See [48] for specific implementation details including the definition of the encoders, and refer to [9], [49] for further background on VAEs.

B. Numerical Results

1) Recovery Phase Transition: The results of the first experiment appear in Figure 1. Specifically, Figure 1(a) plots the fraction of successful recoveries from 20 independent trials as a function of the coherence heuristic (2) and number of measurements. White squares correspond to 100% successful recovery (all errors were below \( 10^{-5} \)), while black squares correspond to no successful recoveries (all errors were above \( 10^{-5} \)). In Figure 1(b), we show a slice of the phase plot for \( m = 100 \), plotting re as a function of the coherence heuristic (2). Each dot corresponds to one of 20 trials at each coherence level. The plot is shown on a log-\( y \) scale. The solid line plots the geometric mean of re as a function of coherence,
with an envelope representing 1 geometric standard deviation (see [51, Appendix A.1.3] for more information on this visualization strategy). Figure 1 indicates that coherence may be effectively controlled via the heuristic (2), and that coherence is a natural quantity associated with recovery performance. These findings corroborate our theoretical results.

2) Incoherent Networks Require Fewer Measurements: In the second experiment, we provide compelling numerical simulations that support our regularization strategy for lowering coherence of the trained network, resulting in stable recovery of the target signal with much fewer measurements. The results of the second experiment are shown in Figure 2 and Figure 3. In Figure 2, we show the recovered image for three images from the MNIST test set.

For each block of $3 \times 9$ images, the top row corresponds with the low coherence $G^{(3)} (\alpha = 0.82)$; the middle row, the high coherence $G^{(2)} (\alpha = 0.96)$; and the bottom row, $G^{(1)}$, which uses sigmoid activation. The left-most column is the target image belonging to $R(G^{(i)})$, labelled signal. All images were clamped to the range $[0, 1]$. The figure shows that a GNN with low coherence can effectively recover the target signal with much fewer measurements compared to a network with high coherence, even when that network uses a final sigmoid activation function (which is a realistic choice in practical settings). Remarkably, in some cases we observed that images could be recovered with fewer than $k$ measurements. This highlights the importance of regularizing for networks with low coherence during training. In Figure 3, we further provide empirical evidence of the benefit of low coherence for recovery. For each measurement, we show the results of 10 independent trials for $G^{(1)}$ (squares), $G^{(2)}$ (triangles) and $G^{(3)}$ (circles), respectively. The lines correspond to the empirical geometric mean ree as a function of $m$; shaded regions correspond to 1 geometric standard deviation. The solid line corresponds to $G^{(1)}$; the dashed line, the high coherence $G^{(2)} (\alpha = 0.96)$; and the dotted line, the low coherence $G^{(3)} (\alpha = 0.82)$. The leftmost column, signal, corresponds to the target image $x_0^0 \in R(G^{(i)})$. 

Fig. 1. Dependence of recovery on coherence and number of measurements $m$ for GNNs trained on MNIST.

![Figure 1](image1.png)

(a) Empirical recovery probability as a function of coherence and $m$. Each block corresponds to the average from 20 independent trials. White corresponds with 20 successful recoveries (ree $\leq 10^{-5}$); black with no successful recoveries.

![Figure 2](image2.png)

(b) Empirical ree as a function of coherence for $m = 100$. Each dot corresponds to one of 20 trials at each coherence level. The solid line shows the empirical geometric mean ree vs. coherence upper bound. The envelope shows 1 geometric standard deviation.

Fig. 2. Recovery comparison of MNIST images for various measurement sizes $m$ (denoted by column heading) for a low coherence network, high-coherence network and network with final sigmoid activation. In each block: the top row corresponds to $G^{(3)} (\alpha = 0.82)$; middle row $G^{(2)} (\alpha = 0.96)$; bottom row $G^{(1)}$ (labelled Sig). The leftmost column, signal, corresponds to the target image $x_0^0 \in R(G^{(i)})$.

![Figure 3](image3.png)

Fig. 3. Performance comparison for three GNNs trained on the MNIST dataset—one with low coherence, another with high coherence, and the last with sigmoid activation on the last layer. Plotted against number of measurements $m$ is ree. For each value of $m$, each dot corresponds to one of 10 trials. In each trial, a random image drawn from the MNIST test partition was used as the target signal. Lines depict the empirical geometric mean ree as a function of $m$; shaded regions correspond to 1 geometric standard deviation. The solid line corresponds to $G^{(1)}$; the dashed line, the high coherence $G^{(2)} (\alpha = 0.96)$; and the dotted line, the low coherence $G^{(3)} (\alpha = 0.82)$.
The latter quantity is bounded above by $\varepsilon A$. Proofs for Main Results

BERK et al.: COHERENCE PARAMETER CHARACTERIZING GCS

V. PROOFS

A. Proofs for Main Results

We proceed by proving Theorem 2, then Theorem 1. Note that Corollary 1, needed for Theorem 1, follows immediately from Theorem 2 using Remark S2.3.

Proof of Theorem 2: By construction, $\mathcal{R}(G) \subseteq \Delta(\mathcal{R}(G))$; the latter set is a union of linear subspaces (see Definition 5 and Remark S3.1). By Lemma S2.2, $\mathcal{R}(G)$ is a union of no more than $N$ polyhedral cones of dimension at most $k$, with $N$ satisfying

$$\log N \leq k \sum_{i=1}^{d-1} \log \left(\frac{2e k_i}{\delta} \right)$$

via Remark S2.2. In particular, $\Delta(\mathcal{R}(G))$ is a collection of at most $N$ subspaces. For any linear subspace $\mathcal{L} \in \Delta(\mathcal{R}(G))$, observe that $\mathcal{L}$ is $\alpha$-coherent with respect to $\| \cdot \|_U$ by assumption. Consequently, by a union bound and application of Lemma 1,

$$\mathbb{P}\left\{ \sup_{x \in \mathcal{R}(G) \cap S^{d-1}} \|Ax\|_2 - 1 \geq \delta \right\} \leq \sum_{\mathcal{L} \in \Delta(\mathcal{R}(G))} \mathbb{P}\left\{ \sup_{x \in \mathcal{L} \cap S^{d-1}} \|Ax\|_2 - 1 \geq \delta \right\} \leq 2 N \exp\left(-\frac{Cm\delta^2}{\alpha^2 n}\right).$$

The latter quantity is bounded above by $\varepsilon$ if

$$m \geq \frac{\alpha^2 n}{\delta^2} \left( \log N + \log \frac{2k}{\varepsilon} \right),$$

whence, by substituting the bound for $\log N$, it suffices to take

$$m \geq \frac{\alpha^2 n}{\delta^2} \left( \sum_{i=1}^{d-1} \log \left(\frac{2e k_i}{\delta} \right) + \log \frac{2k}{\varepsilon} \right).$$

Proof of Theorem 1: Recall $x^+ := x_0 - \Pi_{\mathcal{R}(G)} x_0$. By triangle inequality and the observation that $\Pi_{\mathcal{R}(G)} x_0 \in \mathcal{R}(G)$,

$$\|Ax - b\|_2 \leq \min_{x \in \mathcal{R}(G)} \|Ax - b\|_2 + \hat{\varepsilon} \leq \|A\Pi_{\mathcal{R}(G)} x_0 - b\|_2 + \hat{\varepsilon} \leq \|A x^+ + \eta\|_2^2 + \hat{\varepsilon} \leq \|A x^+\|_2^2 + \|\eta\|_2 + \hat{\varepsilon}.$$

Moreover, with probability at least $1 - \varepsilon$ on the realization of $A$, $A$ satisfies a restricted isometry condition on the difference set $\mathcal{R}(G) - \mathcal{R}(G)$ by Corollary 1. Therefore, since $\hat{x}, \Pi_{\mathcal{R}(G)} x_0 \in \mathcal{R}(G)$,

$$\|A\hat{x} - b\|_2 \leq \|A(x_0 - \Pi_{\mathcal{R}(G)} x_0) - \eta\|_2 \leq \|A\hat{x} - \Pi_{\mathcal{R}(G)} x_0\|_2 - \|A x^+\|_2 - \|\eta\|_2 \leq (1 - \delta)\|\hat{x} - \Pi_{\mathcal{R}(G)} x_0\|_2 - \|A x^+\|_2 - \|\eta\|_2.$$

Assembling the two inequalities gives

$$\|\hat{x} - \Pi_{\mathcal{R}(G)} x_0\|_2 \leq \frac{1}{1 - \delta} \left( 2\|A x^+\|_2 + 2\|\eta\|_2 + \hat{\varepsilon} \right).$$

Finally, apply triangle inequality and choose $\delta = \frac{1}{2}$ to get

$$\|\hat{x} - x_0\|_2 \leq \|x_0 - \Pi_{\mathcal{R}(G)} x_0\|_2 + \|\hat{x} - \Pi_{\mathcal{R}(G)} x_0\|_2 \leq \|x\|_2 + 3\|A x^+\|_2 + 3\|\eta\|_2 + \frac{3}{2} \hat{\varepsilon}.$$
Therefore, by Lemma S1.1 it follows that
\[
\mathbb{P}\left\{ \left\| \frac{1}{m} \sum_{i=1}^{n} U_i U_i^* (\theta_i - \frac{m}{n}) \right\| \geq \delta \right\} \\
\leq 2k \exp \left( \frac{-m\delta^2/2}{na^2 (1 - \frac{m}{n} + \frac{\delta}{3})} \right) \\
\leq 2k \exp \left( -C \cdot \min \left\{ \frac{m\delta^2}{na^2 (1 - \frac{m}{n})}, \frac{m\delta}{na^2} \right\} \right).
\]

To complete the proof, we adapt the argument from the proof of [41, Th. 3.1.1]. Indeed, for \( \delta, z \geq 0 \) note that \( |1 - z| > \delta \implies |z^2 - 1| > \max(\delta, \delta^2) \) yields the implication \( \max_i |1 - z_i| > \delta \implies \max_i |z_i^2 - 1| > \max(\delta, \delta^2) \).

Consequently,
\[
\mathbb{P}\left\{ \sup_{x \in \mathcal{L}^{m \times n}} \left\| Ax \right\|_2 - 1 \geq \delta \right\} \\
\leq \mathbb{P}\left\{ \sup_{x \in \mathcal{L}^{m \times n}} \left\| Ax \right\|_2 \geq \max(\delta, \delta^2) \right\} \\
\leq 2k \exp \left( -\frac{C \delta^2 m}{\alpha^2 n} \right).
\]

B. Proofs for Typical Coherence

The proof of Proposition 1 requires the following lemma.

**Lemma 2:** Let \( A \in \mathbb{C}^{n \times k} \) be a matrix with \( \ell_2 \)-normalized columns and let \( A_i \) denote the \( i \)-th row of \( A \). Then
\[
\max_{i \in [n]} \|A_i\|_2 \geq \sqrt{\frac{k}{n}},
\]

**Proof of Lemma 2:** Computing directly, using that each of the \( k \) columns has unit norm,
\[
\max_{i \in [n]} \|A_i\|_2 \geq \text{mean}_{i \in [n]} \|A_i\|_2^2 = \frac{1}{n} \|A\|_F^2 = \frac{k}{n}.
\]

Taking square roots completes the proof.

We now prove the proposition using the lemma.

**Proof of Proposition 1:** Take the set \( \mathcal{T} \) of all subspaces of dimension \( k \) in \( \mathbb{C}^n \). By rotational invariance of \( \mathcal{T} \), it suffices to show the result with respect to \( U = I \). Hence, let \( \{e_i\}_{i \in [n]} \) be the canonical basis. Any fixed \( T \in \mathcal{T} \) has coherence
\[
\alpha = \sup_{v \in T \cap \mathcal{S}^{d-1}} \|v\|_\infty.
\]

We will show a sharp lower bound on the coherence of all \( k \)-dimensional subspaces, namely
\[
\inf_{T \in \mathcal{T}} \sup_{v \in T \cap \mathcal{S}^{d-1}} \|v\|_\infty = \sqrt{\frac{k}{n}}.
\]

Apply Lemma 2 to lower bound the latter quantity. As Lemma 2 applies to any matrix in \( A \),
\[
\inf_{A \in \mathcal{A}} \sup_{v \in \mathcal{S}^{d-1}} \|A v\|_\infty = \inf_{A \in \mathcal{A}} \sup_{v \in \mathcal{S}^{d-1}} \|A v\|_\infty = \inf_{A \in \mathcal{A}} \max_{i \in [n]} \|A_i v\|_2.
\]

We next show that there exists a subspace such that equality holds. Take \( F \in \mathbb{C}^{n \times k} \) whose columns are the first \( k \) columns of the DFT matrix, as defined in Remark 1. The columns of \( F \) are orthonormal, so \( F \in \mathcal{A} \). Furthermore, each row of \( F \) has \( \ell_2 \) norm \( \sqrt{\frac{k}{n}} \). It follows that
\[
\inf_{A \in \mathcal{A}} \max_{i \in [n]} \|A_i\|_2 = \sqrt{\frac{k}{n}}.
\]

The proof of Theorem 3 uses Lemma S1.2 and the following lemma, which bounds the coherence of a random subspace sampled from the Grassmannian. The Grassmanian \( \Gamma_{n,k} \) consists of all \( k \)-dimensional subspaces of \( \mathbb{R}^n \) [41, Ch. 5.2.6].

**Lemma 3:** Let \( U \in \mathbb{C}^{n \times k} \) be a unitary matrix and denote by \( \mathcal{L} \in \Gamma_{n,k} \) a subspace distributed uniformly at random over \( \Gamma_{n,k} \). With probability at least \( 1 - 2 \exp(-\gamma^2) \), \( \mathcal{L} \) is \( \alpha \)-coherent with respect to \( \| \cdot \|_U \) with
\[
\alpha \leq \sqrt{\frac{k}{n}} + \sqrt{\frac{\log(n)}{n}} + \gamma \sqrt{n}.
\]

**Proof of Theorem 3:** Let \( \hat{G}(x) = \sigma(W^{(d-1)} \cdots \sigma(W^{(1)} x)) \) so that \( G(x) = W^{(d)} \hat{G}(x) \). Then,
\[
\Delta(\mathcal{R}(G) - \mathcal{R}(G)) = \Delta \left( W^{(d)} \left( \mathcal{R}(\hat{G}) - \mathcal{R}(\hat{G}) \right) \right) = W^{(d)} \Delta(\mathcal{R}(G) - \mathcal{R}(G)).
\]

By Lemma S2.2 and Remark S3.1, \( \Delta(\mathcal{R}(G) - \mathcal{R}(G)) \) is the union of \( M \) at-most \( 2k \)-dimensional linear subspaces with
\[
\log M \leq 2k \sum_{l=1}^{d-1} \log \left( \frac{2e k_l}{k} \right).
\]

Each subspace \( \mathcal{L} \) is uniformly distributed on \( \Gamma_{n,\dim \mathcal{L}} \), where \( \dim \mathcal{L} \leq 2k \), because the final weight matrix has id Gaussian entries independent of the other weight matrices (e.g., see [41, Ch. 3.3.2]). Enumerate the subspaces from 1 to \( M \) (independently of \( W^{(d)} \)). Then, applying Lemma 3, we see that the coherence of subspace \( i \) is a random variable \( \alpha_i \) such that
\[
\alpha_i \lesssim \sqrt{\frac{k}{n}} + \sqrt{\frac{\log(n)}{n}} + \frac{\gamma}{\sqrt{n}}
\]

with probability at least \( 1 - 2 \exp(-\gamma^2) \).

Let \( \alpha \) be the coherence of \( \Delta(\mathcal{R}(G) - \mathcal{R}(G)) \) and observe that \( \alpha \leq \max_{i \in [M]} \alpha_i \). Applying Lemma S1.2,
\[
\alpha \leq \max_{i \in [M]} \alpha_i \lesssim \sqrt{\frac{k}{n}} + \sqrt{\frac{\log(n)}{n}} + \sqrt{\frac{\log M}{n}} + \frac{\gamma}{\sqrt{n}}
\]

with probability at least \( 1 - 2 \exp(-\gamma^2) \).

We next prove Lemma 3.
Proof of Lemma 3: Let $A \in \mathbb{R}^{n \times k}$ with $A_{ij} \overset{\text{id}}{\sim} \mathcal{N}(0, 1)$. Then, $\mathcal{L} = \mathcal{R}(A)$ is a random subspace uniformly distributed over $\Gamma_{n,k}$. By rotation invariance of the Grassmannian, it suffices to show the result for $U = I$. Let $\{e_i\}_{i \in [n]}$ denote the canonical basis. Define

$$\alpha := \sup_{v \in \mathcal{L} \cap \mathcal{S}^{n-1}} \max_{i \in [n]} |\langle e_i, v \rangle|,$$

and note that $\mathcal{L}$ is $\alpha$-coherent with respect to $\|\cdot\|_2 = \|\cdot\|_{\infty}$. For each $i \in [n]$, we next analyze

$$\alpha_i := \sup_{v \in \mathcal{L} \cap \mathcal{S}^{n-1}} |\langle e_i, v \rangle| = \sup_{y \in \mathcal{S}^{k-1}} \frac{A_{iy}}{\|Ay\|_2} \tag{4}$$

We will show, with probability at least $1 - 4 \exp(-s^2)$,

$$\alpha_i \approx \sqrt{\frac{k}{n}} + \frac{s}{\sqrt{n}}.$$

To see why this result should hold, we focus our attention on the right hand side of (4). The denominator concentrates around $\sqrt{n}$ and the numerator is bounded by $\|A_i\|_2$, which concentrates around $\sqrt{k}$ with subgaussian tails.

We first obtain a lower bound on the smallest singular value of $A$ via [41, Th. 4.6.1], which guarantees with probability at least $1 - 2 \exp(-t^2)$ that

$$\inf_{y \in \mathcal{S}^{k-1}} \|Ay\|_2 \geq \sqrt{n} - C\sqrt{k} - Ct.$$  

By fixing $t = \frac{\sqrt{n}}{2}$ we define the event

$$B := \left\{ \inf_{y \in \mathcal{S}^{k-1}} \|Ay\|_2 \leq \frac{\sqrt{n}}{2} - C\sqrt{k} \right\}$$

satisfying $\mathbb{P}(B) \leq 2 \exp(-cn)$. We first limit $s$ so that $\sqrt{k} + s \leq \sqrt{n} - C\sqrt{k}$, which implies that $s < C\sqrt{n}$. Then

$$\mathbb{P} \left\{ \sup_{y \in \mathcal{S}^{k-1}} \frac{A_{iy}}{\|Ay\|_2} > \frac{\sqrt{k} + s}{\frac{\sqrt{n}}{2} - C\sqrt{k}} \right\} = \mathbb{P} \left\{ \sup_{y \in \mathcal{S}^{k-1}} \frac{A_{iy}}{\|Ay\|_2} > \frac{\sqrt{k} + s}{\frac{\sqrt{n}}{2} - C\sqrt{k}} \mid B \right\} \mathbb{P}(B)$$

$$+ \mathbb{P} \left\{ \sup_{y \in \mathcal{S}^{k-1}} \frac{A_{iy}}{\|Ay\|_2} > \frac{\sqrt{k} + s}{\frac{\sqrt{n}}{2} - C\sqrt{k}} \mid B^c \right\} \mathbb{P}(B^c)$$

$$\leq \mathbb{P}(B) + \mathbb{P} \left\{ \sup_{y \in \mathcal{S}^{k-1}} A_{iy} > \sqrt{k} + s \right\}$$

$$\leq 2 \exp(-cn) + \mathbb{P} \left\| Ay \right\|_2 > \sqrt{k} + s \right\}$$

$$\leq 2 \exp(-cn) + 2 \exp(-cs^2).$$

Above, we used the concentration of the norm of Gaussian vectors [41, Th. 3.1.1]. Since $s \leq C\sqrt{n}$, $\exp(-cn) \leq \exp(-cs^2)$. From this we find the desired subgaussian tail bound:

$$\mathbb{P} \left\{ \sup_{y \in \mathcal{S}^{k-1}} \frac{A_{iy}}{\|Ay\|_2} > \frac{\sqrt{k} + s}{\frac{\sqrt{n}}{2} - C\sqrt{k}} \right\} \leq 4 \exp(-cs^2).$$

The remaining values of $s$ satisfy $\sqrt{k} + s > \frac{\sqrt{n}}{2} - C\sqrt{k}$. Therefore, since $\sup_{y \in \mathcal{S}^{k-1}} \frac{A_{iy}}{\|Ay\|_2} \leq 1$,

$$\mathbb{P} \left\{ \sup_{y \in \mathcal{S}^{k-1}} \frac{A_{iy}}{\|Ay\|_2} > \frac{\sqrt{k} + s}{\frac{\sqrt{n}}{2} - C\sqrt{k}} \right\} = 0 \leq 4 \exp(-cs^2).$$

Therefore, the subgaussian bound applies for all values of $s$. We now scale $s$ by an absolute constant with $\gamma = cs$. Then

$$\alpha_i = \sup_{y \in \mathcal{S}^{k-1}} \frac{A_{iy}}{\|Ay\|_2} > \frac{\sqrt{k} + s}{\frac{\sqrt{n}}{2} - C\sqrt{k}} \geq \frac{\sqrt{k}}{n} + \frac{\gamma}{\sqrt{n}}$$

with probability less than $2 \exp(-\gamma^2)$. Changing the constant from 4 to 2 in the probability bound is achieved by suitable choice of $c$. Remembering that $\alpha$ is the coherence with the canonical basis, we apply Lemma S1.2 to find,

$$\alpha = \max_{i \in [n]} \alpha_i \lesssim \sqrt{\frac{k}{n}} + \frac{\log n}{n} + \frac{\gamma}{\sqrt{n}}$$

with probability at least $1 - 2 \exp(-\gamma^2)$. \hfill \blacksquare

VI. CONCLUSION

In this work, we have proved a restricted isometry property for a subsampled isometry with GNN structural proxy, Theorem 2. We used this to prove sample complexity and recovery bounds, Theorem 1. The recovery bound stated in Theorem 1 is uniform over ground truth signals, and permits a more finely tuned nonuniform control as discussed in Remark 4. To our knowledge, this provides the first theory for generative compressed sensing with subsampled isometries and non-random weights.

Our results rely on the notion of $\alpha$-coherence with respect to the measurement norm, introduced in Definition 4 and Definition 3, respectively. Closely related to the notion of incoherent bases [7, p. 373] and the $X$-norm of [13], we argue that $\alpha$-coherence is a natural quantity to measure the interplay between a GNN and the measurement operator. Indeed, in Section IV we propose a regularization strategy for promoting favourable coherence of GNNs during training, and connect this strategy with favourable recovery efficacy. Specifically, we show that our regularization strategy yields low coherence GNNs with improved sample complexity for recovery (Figure 3). Moreover, our numerics support that low coherence GNNs achieve better sample complexity than high coherence GNNs (Figure 1).

We suspect the $\Omega(k^2d^2)$ dependence in the sample complexity of our analysis is sub-optimal, and a consequence of our coherence-based approach. Ignoring logarithmic factors, it is an open question to prove recovery guarantees with $\Omega(kd)$ Fourier measurements and non-random weights, which would match the number of (sub-)gaussian measurements needed [1] and would also match the known worst-case lower bound [52]. In addition, it is an open problem to improve the regularization strategy for lowering coherence, possibly including middle layers. Finally, it is open to determine a notion of coherence for networks that have a final nonlinear activation function, and to characterize how this impacts recovery efficacy for such networks.
