Model-adapted Fourier sampling for generative compressed sensing

Aaron Berk[*] **Simone Brugiapaglia**[†] **Yaniv Plan** **Matthew Scott**[‡] **Xia Sheng** [§] **Özgür Yılmaz**[¶]

Abstract

We study generative compressed sensing when the measurement matrix is randomly subsampled from a unitary matrix (with the DFT as an important special case). It was recently shown that $\mathcal{O}(kdn\|\boldsymbol{\alpha}\|_\infty^2)$ uniformly random Fourier measurements are sufficient to recover signals in the range of a neural network $G : \mathbb{R}^k \to \mathbb{R}^n$ of depth d, where each component of the so-called local coherence vector $\boldsymbol{\alpha}$ quantifies the alignment of a corresponding Fourier vector with the range of G. We construct a model-adapted sampling strategy with an improved sample complexity of $\mathcal{O}(kd\|\boldsymbol{\alpha}\|_2^2)$ measurements. This is enabled by: (1) new theoretical recovery guarantees that we develop for nonuniformly random sampling distributions and then (2) optimizing the sampling distribution to minimize the number of measurements needed for these guarantees. This development offers a sample complexity applicable to natural signal classes, which are often almost maximally coherent with low Fourier frequencies. Finally, we consider a surrogate sampling scheme, and validate its performance in recovery experiments using the CelebA dataset.

1 Introduction

Compressed sensing considers signals $\boldsymbol{x}_0 \in \mathbb{R}^n$ with high ambient dimension n that belong to (or can be well-approximated by elements of) a prior set $\mathcal{V} \subseteq \mathbb{R}^n$ with lower "complexity" than the ambient space. The aim is to recover (an approximation to) such signals with provable accuracy guarantees from the linear, typically noisy, measurements $\boldsymbol{b} = A\boldsymbol{x}_0 + \boldsymbol{\eta}$, where $\boldsymbol{\eta} \in \mathbb{C}^m$ denotes noise and $A \in \mathbb{C}^{m\times n}$ with $m \ll n$ is an appropriately chosen (possibly random) measurement matrix. The signal is to be recovered by means of a computationally feasible method that utilizes the structure of $\mathcal{V}$ and has access to only A and $\boldsymbol{b}$. In classical compressed sensing, $\mathcal{V}$ is the set of sparse vectors. In generative compressed sensing, the prior set $\mathcal{V}$ is chosen to be the range of a generative neural network $G : \mathbb{R}^k \to \mathbb{R}^n$, an idea that was first explored in [5]. Our results will hold for ReLU-activated neural networks as defined in [4, Definition I.2]. Here we denote the ReLU (Rectified Linear Unit) activation as $\sigma(z) := \max\{z, 0\}$, applied component-wise to a real vector $\boldsymbol{z}$.

Definition 1.1 ((k, d, n)-Generative Network)**.** With $k, d, n \in \mathbb{N}$, fix the integers $2 \le k := k_0 \le k_1, \dots, k_{d-1} \le k_d = n$, and for $i \in [d]$, let $W^{(i)} \in \mathbb{R}^{k_i \times k_{i-1}}$. A (k, d, n)-generative network is a function $G : \mathbb{R}^k \to \mathbb{R}^n$ of the form $G(\boldsymbol{z}) := W^{(d)}\sigma\left(\cdots W^{(2)}\sigma\left(W^{(1)}\boldsymbol{z}\right)\right)$.

In the same work [5], the authors provided a theoretical framework for generative compressed sensing with $A \in \mathbb{R}^{m\times n}$ having independent identically distributed (i.i.d.) Gaussian entries. However, the assumption of Gaussian measurements is unrealistic for applications like MRI, where measurements are spatial Fourier transforms. Limitations of the hardware in such an application restrict the set of possible measurements to the rows of a fixed unitary matrix (or approximately so, up to discretization of the measurements). Hence we consider the more realistic *subsampled unitary* measurement matrices, i.e., matrices with rows randomly subsampled from the rows of a unitary matrix $F \in \mathbb{C}^{n\times n}$.

[*]Dept. Mathematics and Statistics, McGill University, Montréal, Canada. Now affiliated with Deep Render, London UK

[†]Dept. Mathematics and Statistics, Concordia University, Montréal, Canada

[‡]Corresponding author: `matthewscott@math.ubc.ca`

[§]Authors listed in alphabetic order. MS was primarily responsible for developing the theory and writing the main body, XS was primarily responsible for numerical experiments and writing the numerical section.

[¶]YP, MS, XS and OY affiliated with Dept. Mathematics, University of British Columbia, Vancouver, Canada

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

A first result in the generative setting with subsampled unitary measurements [4] shows that for well-behaved networks, $m = \mathcal{O}(k^2 d^2)$ (up to log factors) measurements is sufficient for recovery with high probability. However, sampling uniformly is not efficient; sampling more informative measurements at a higher rate is known to improve performance in compressed sensing (e.g., low Fourier frequencies have strong correlation with natural images, they will therefore tend to be more informative) [3]. This idea is mirrored in the radial sampling strategy used in MRI scans, which takes a disproportionate number of low-frequency measurements [15]. We address this limitation by generalizing the theory to any measurement matrix of the form $A = SF$ where $F \in \mathbb{C}^{n\times n}$ is a unitary matrix and $S \in \mathbb{R}^{m\times n}$ is a *sampling matrix*, which we now define. We adopt the convention that $\boldsymbol{e}_j$ denotes a canonical basis vector and that Δ^{n-1} is the simplex in $\mathbb{R}^n$. See Section 1 for the definition of symbols throughout this paper.

Definition 1.2 (Sampling Matrix)**.** We define a *sampling matrix* to be any matrix $S \in \mathbb{R}^{m\times n}$ composed of i.i.d. row vectors $\boldsymbol{s}_1^*, \dots, \boldsymbol{s}_m^*$ such that $\mathbb{P}(\boldsymbol{s}_i = \boldsymbol{e}_j) = p_j$ for all $i \in [m], j \in [n]$, for some fixed probability vector $\boldsymbol{p} \in \Delta^{n-1}$.

We will further show, similarly to [18], that picking the sampling probabilities in a manner informed by the geometry of the problem yields improved recovery guarantees relative to the uniform case. Specifically, we provide a bound on the measurement complexity of $m = \mathcal{O}(k^2 d^2)$ even for models which are highly aligned with a small subset of the rows of F. To find good sampling probabilities, we must understand that measurements (i.e. the rows of the measurement matrix) are effective for a prior set $\mathcal{V} \subseteq \mathbb{R}^n$ if they help differentiate between signals in $\mathcal{V}$. Therefore, we consider the alignment of rows of F with the set $\mathcal{V} - \mathcal{V}$ (where the difference is in the sense of a Minkowski sum, see Section 1). We will sample rows of F that have a high degree of alignment with $\mathcal{V} - \mathcal{V}$ at a higher rate. For technical reasons, we consider alignments with a slightly larger set, given by the following set operator previously introduced in [4, Definition 2.1].

Definition 1.3 (Piecewise Linear Expansion)**.** Let $\mathcal{C} \subseteq \mathbb{R}^n$ be the union of N convex cones: $\mathcal{C} = \bigcup_{i=1}^N \mathcal{C}_i$. Define the piecewise linear expansion $\Delta(\mathcal{C}) := \bigcup_{i=1}^N \operatorname{span}(\mathcal{C}_i) = \bigcup_{i=1}^N (\mathcal{C}_i - \mathcal{C}_i)$.

The piecewise linear expansion is a well-defined set operator as shown in [4, Remark A.2]. Specifically, it is independent of the choice of convex cones $\mathcal{C}_i$.

We require the following quantity to quantify the alignment of the individual vectors with the prior set.

Definition 1.4 (local coherence)**.** The *local coherence* of a vector $\boldsymbol{\phi} \in \mathbb{C}^n$ with respect to a cone $\mathcal{T} \subseteq \mathbb{R}^n$ is defined as

$$\alpha_{\mathcal{T}}(\boldsymbol{\phi}) := \sup_{\boldsymbol{x} \in \Delta(\mathcal{T}) \cap \mathbb{S}^{n-1}} |\langle \boldsymbol{\phi}, \boldsymbol{x} \rangle|.$$

The *local coherences* of a unitary matrix $F \in \mathbb{C}^{n\times n}$ with respect to a cone $\mathcal{T} \subseteq \mathbb{R}^n$ are collected in the vector $\boldsymbol{\alpha}$ with entries $\alpha_j := \alpha_{\mathcal{T}}(\boldsymbol{f}_j)$, where $\boldsymbol{f}_j^*$ is the jth row of F.

By using the local coherences of F to inform sampling probabilities, we will show that recovery occurs from only $m = \mathcal{O}(kd\|\boldsymbol{\alpha}\|_2^2)$ (up to log factors) measurements with high probability.

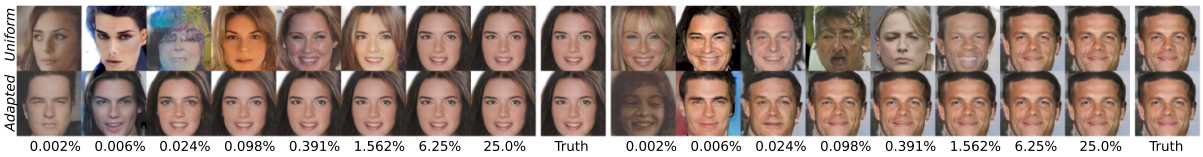


Figure 1: Samples recovered by different sampling schemes with different sampling rates

Prior Work Sampling with non-uniform probabilities was the subject of a line of research in classical compressed sensing. Seminal works involving the notion of local coherence in classical compressed sensing are [10, 11, 18, 19]. There have been quantities analogous to the local coherence appearing in the literature, such as Christoffel functions [16] and leverage scores [7, 14], which were recently introduced in machine learning in the context of, e.g., kernel-based methods [8] and deep learning [1]. While writing this manuscript, we became aware of the recent paper [2]. This work presents a framework for optimizing sampling in general scenarios based on the so-called *generalized Christoffel function*, a quantity that admits the (squared) local coherence considered here as a particular case. Furthermore, the method we use to numerically approximate the coherence in Section 3 can be seen as a special case of that proposed in [2]. However, the results in [2] assume that $\mathcal{V} - \mathcal{V}$ is a union of low-dimensional subspaces. Hence, they are not directly applicable to the case of generative compressed sensing with ReLU networks, for which $\mathcal{V} - \mathcal{V}$ is in general only *contained* in a union of low-dimensional subspaces. On the other hand, our theory explicitly covers the case of generative compressed sensing, and illustrates how sufficient conditions on m leading to successful signal recovery depend on the generative network's parameters (k, d, n). Second, we provide recovery guarantees that hold with high probability, as opposed to expectation.

The present work directly improves on results from [4] by improving the sample complexity from $n\|\boldsymbol{\alpha}\|_\infty^2$ to $\|\boldsymbol{\alpha}\|_2^2$ when the sampling probabilities are adapted to the generative model used. This is a sizable improvement in performance guarantees for a significant class of realistic generative models. It can be understood as extending the theory of generative compressed sensing with Fourier measurements to many realistic settings where we provide nearly order-optimal bounds on the sampling complexity. Indeed, in the context of the main result from [4], "favourable coherence" corresponds to $\|\boldsymbol{\alpha}\|_\infty \leq C\sqrt{kd/n}$ where C is an absolute constant. Despite the fact that the prior generated by a neural network with Gaussian weights will have such a coherence [4], we observe empirically in Figure 2d) that for a trained generative model, a small number of Fourier coefficients have values close to one. In such cases, the main result from [4] becomes vacuous while Theorem 2.1 remains meaningful.

Notation For any map f, we denote $\mathcal{R}(f)$ to be the range of f. We define the simplex $\Delta^{n-1} := \{\boldsymbol{p} \in \mathbb{R}^n : p_i \geq 0, \sum_{i=1}^n p_i = 1\}$, and the sphere $\mathbb{S}^{n-1} := \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\|_2 = 1\}$. For any matrix $A \in \mathbb{C}^{m\times n}$, we denote its pseudo-inverse by $A^+ \in \mathbb{C}^{n\times m}$. For a set $\mathcal{V} \in \mathbb{R}^n$ we denote its *self-difference* $\mathcal{V} - \mathcal{V} := \{\boldsymbol{v}_1 - \boldsymbol{v}_2 | \boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathcal{V}\}$. We define $\Pi_\mathcal{T} : \mathbb{R}^n \to \mathbb{R}^n$ to be the orthogonal projection on to a set $\mathcal{T} \subseteq \mathbb{R}^n$ in the sense that $\Pi_\mathcal{T}\,\boldsymbol{x}$ is a single element from the set $\arg\min_{\boldsymbol{t}\in\mathcal{T}} \|\boldsymbol{x} - \boldsymbol{t}\|_2$. We let $[l] = \{1, \ldots, l\}$. We denote $\langle\cdot,\cdot\rangle$ to be the the canonical inner product in $\mathbb{R}$ or $\mathbb{C}$ depending on the context.

2 Main Result

We now state the main result of this paper, where we give an upper bound on the sample complexity required for signal recovery. The accuracy of the recovery is dependent on the measurement noise, the modelling error (how far the signal is from the prior), and imperfect optimization. These are denoted, respectively, by $\boldsymbol{\eta}, \boldsymbol{x}^\perp$ and $\hat{\varepsilon}$ below.

Theorem 2.1. *Fix a (k, d, n)-generative network G, the cone $\mathcal{T} := \mathcal{R}(G) - \mathcal{R}(G) \subseteq \mathbb{R}^n$, the unitary matrix $F \in \mathbb{C}^{n\times n}$. Let $\boldsymbol{\alpha}$ be the vector of local coherences of F with respect to $\mathcal{T}$. Let $\boldsymbol{p} := (\alpha_j^2/\|\boldsymbol{\alpha}\|_2^2)_{j\in[n]} \in \Delta^{n-1}$ with $S \in \mathbb{R}^{m\times n}$ the corresponding random sampling matrix. Let $D \in \mathbb{R}^{n\times n}$ be the diagonal matrix with entries $D_{i,i} = 1/\sqrt{p_i}$. Let $\tilde{D} := SDS^+ \in \mathbb{R}^{m\times m}$. Let $\varepsilon > 0$. If*

$$m \geq C\|\boldsymbol{\alpha}\|_2^2 \left(kd \log\left(\frac{n}{k}\right) + \log\left(\frac{2}{\varepsilon}\right)\right)$$

for C an absolute constant, then with probability at least $1 - \varepsilon$ over the realization of S the following statement holds.

Statement 2.1. For any choice of $\boldsymbol{x}_0 \in \mathbb{R}^n$ and $\boldsymbol{\eta} \in \mathbb{C}^n$, let $\boldsymbol{b} := \frac{1}{\sqrt{m}} SF\boldsymbol{x} + \boldsymbol{\eta}$ and $\boldsymbol{x}^\perp := \boldsymbol{x}_0 - \Pi_\mathcal{T}\,\boldsymbol{x}_0$, and any $\hat{\boldsymbol{x}} \in \mathbb{R}^n, \hat{\varepsilon} > 0$ satisfying $\|\frac{1}{\sqrt{m}} SDF\hat{\boldsymbol{x}} - \tilde{D}\boldsymbol{b}\|_2 \leq \min_{\boldsymbol{x}\in\mathcal{T}} \|\frac{1}{\sqrt{m}} SDF\boldsymbol{x} - \tilde{D}\boldsymbol{b}\|_2 + \hat{\varepsilon}$. We have that

$$\|\hat{\boldsymbol{x}} - \boldsymbol{x}_0\|_2 \leq \|\boldsymbol{x}^\perp\|_2 + \frac{3}{\sqrt{m}}\|SDF\boldsymbol{x}^\perp\|_2 + 3\|\tilde{D}\boldsymbol{\eta}\|_2 + \frac{3}{2}\hat{\varepsilon}.$$

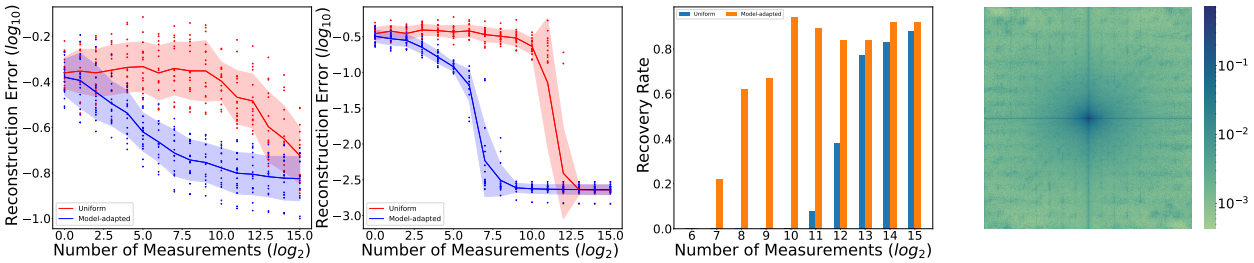
Remark 2.1. The matrix $\tilde{D}$ has a vector of diagonal entries $\mathrm{Diag}(\tilde{D})$ satisfying $\mathrm{Diag}(\tilde{D}) = S\mathrm{Diag}(D)$. ◇

Remark 2.2. We give a generalization of this result to arbitrary sampling probability vectors in Theorem A3.1. ◇

3 Numerics

In this section, we provide empirical evidence of the connection between coherence-based non-uniform sampling and recovery error. By presenting visual (Figure 1) and quantitative (Figure 2) evidence, we validate that model-adapted sampling using a coherence-informed probability vector can outperform a uniform sampling scheme — requiring fewer measurements for successful recovery.

Coherence heuristic Ideally, we would compute the local coherence $\boldsymbol{\alpha}$ using Definition 1.4, but to our knowledge computing local coherence is intractable for generative models relevant to practical settings [4]. Thus, we approximate the quantity by sampling points from the range of the generative model and computing the local coherence from the sampled points instead. Specifically, we sample *codes* from the latent space of the generative model to generate a batch of images with shape (B, C, H, W), where B, C, H, W stand for batch size, number of channels, image height and image width respectively. Then, we compute the set self-difference of the image batch, and normalize each difference vector. This gives a tensor of shape (B^2, C, H, W). We perform a channel-wise two-dimensional Discrete Fourier Transform (DFT) on the tensor, take the element-wise modulus, and then maximize over the batch dimension. This results in a coherence tensor a_0 with shape (C, H, W). To obtain the coherence-informed probability vector of each channel, we first square element-wise, then we normalize channel-wise. To estimate the local coherences of our generative model we use a batch size of 5000 and employ the DFT from PyTorch [17].



(a) Out-of-range Recovery (b) In-range Recovery (c) Recovery rate histogram (d) Fourier local coherences

Figure 2: In (a) we shows the reconstruction error in terms of number of measurements with out-of-range signals. In (b) we present the same quantity for in-range signals. In (c) we treat signal recoveries with relative error less than 3×10^{-3} as "successful", and display the proportion of successful recoveries out of 64 attempts on in-range signals. In (d) we plot the local coherences for one channel of one image. Observe that the local coherences take on values in many different orders of magnitude and peak sharply for a small number of local coherences.

In-Range vs Out-of-Range Signals We run signal recovery experiments for two kinds of images: when the signals are conditioned to be in the range of the generative model (in-range signals), and when they are directly picked from the validation set (out-of-range signals). The in-range signals are randomly generated (with Gaussian codes) by the same generative network that we use for recovery. This ensures that these signals lie within the prior set. In the out-of-range setting, a residual recovery error can be observed even with large numbers of measurements Figure 2. This error occurs because of the so-called model mismatch; there is some distance between the prior set and the signals.

Procedure for Signal Recovery The way we perform signal recovery goes as follows. For a given image $\boldsymbol{x}_0 \in \mathbb{R}^{C \times H \times W}$, we create a mask $M \in \{0,1\}^{C \times H \times W}$ by randomly sampling with replacement m times for each channel according to the probability vector. Let F be the channel-wise DFT operator and $G : \mathbb{R}^k \to \mathbb{R}^{C \times H \times W}$ be the generative neural network (where we omit the batch dimension for simplicity). We denote $\odot$ to be the element-wise tensor multiplication and $\|\cdot\|_F$ to be the Frobenius norm. We approximately solve the optimization program $\hat{\boldsymbol{z}} \in \arg\min_{\boldsymbol{z} \in \mathbb{R}^k} \|M \odot F\boldsymbol{x}_0 - M \odot FG(\boldsymbol{z})\|_F^2$ by running AdamW [13] with $lr = 0.003, \beta_1 = 0.9$ and $\beta_2 = 0.999$ for 20000 iterations on four different random initializations, and pick the code that achieves the lowest loss. The recovered signal is then $\hat{\boldsymbol{x}} := G(\hat{\boldsymbol{z}})$. We measure the quality of the signal recovery by using the relative recovery error (rre), $\text{rre}(\boldsymbol{x}_0, \hat{\boldsymbol{x}}) = \frac{\|\boldsymbol{x}_0 - \hat{\boldsymbol{x}}\|_2}{\|\boldsymbol{x}_0\|_2}$.

Observe that Figure 2b) demonstrates the efficiency of model-adapted sampling. Signal recovery with adapted sampling occurs with 16 times fewer measurements than when using uniform sampling. Similar performance gains can be observed visually in Figure 1 and Figure 3. Comparing the number of measurements to the ambient dimension, we see from Figure 2d) that signal recovery occurs with $\frac{m}{n} \approx 2^7/256^2 = 0.2\%$.

There are a few ways these numerical experiments do not directly match our theory. The sampling is done channel wise, which is technically block sampling [3]. Also, the signal recovery is performed without the preconditioning factor $\tilde{D}$ that appears in Theorem 2.1.

4 Conclusion

In this paper we bring together the ideas used to quantify the compatibility of generative models with subsampled unitary measurements, which were first explored in [4], with ideas of non-uniform sampling from classical compressed sensing. We present the first theoretical result applying coherence-based sampling (similar to leverage score sampling, or Christoffel function sampling) to the setting where the prior is a ReLU generative neural network. We find that adapting the sampling scheme to the geometry of the problem yields substantially improved sampling complexities for many realistic generative networks, and that this improvement is significant in empirical experiments.

Possible avenues for future research include extending the theory presented in [2] to ReLU nets by using methods introduced in the present work. This would yield the benefit of extending the theory from this paper to a number of realistic sampling schemes. A second research direction consists of investigating the optimality of the sample complexity bound that we present in this paper. The sample complexity that we guarantee includes a factor of k^2d^2 when the generative model is well-behaved. Whether this dependence can be reduced to kd, as is the case when the measurement matrix is Gaussian, is an interesting problem that remains open. Finally, the class of neural networks considered in this work could be expanded to include more realistic ones.

References

[1] Ben Adcock, Juan M Cardenas, and Nick Dexter. "CAS4DL: Christoffel adaptive sampling for function approximation via deep learning". In: *Sampling Theory, Signal Processing, and Data Analysis* 20.2 (2022), p. 21.

[2] Ben Adcock, Juan M. Cardenas, and Nick Dexter. *CS4ML: A General Framework for Active Learning with Arbitrary Data Based on Christoffel Functions*. June 2023. arXiv: 2306.00945 [cs, math]. (Visited on 09/12/2023).

[3] Ben Adcock and Anders C. Hansen. *Compressive Imaging: Structure, Sampling, Learning*. Cambridge: Cambridge University Press, 2021. ISBN: 978-1-108-42161-4. DOI: 10.1017/9781108377447. (Visited on 05/01/2023).

[4] Aaron Berk et al. "A Coherence Parameter Characterizing Generative Compressed Sensing with Fourier Measurements". In: *IEEE Journal on Selected Areas in Information Theory* (2022), pp. 1–1. ISSN: 2641-8770. DOI: 10.1109/JSAIT.2022.3220196.

[5] Ashish Bora et al. "Compressed Sensing Using Generative Models". In: *Proceedings of the 34th International Conference on Machine Learning*. PMLR, July 2017, pp. 537–546. (Visited on 01/07/2023).

[6] E.J. Candes, J. Romberg, and T. Tao. "Robust Uncertainty Principles: Exact Signal Reconstruction from Highly Incomplete Frequency Information". In: *IEEE Transactions on Information Theory* 52.2 (Feb. 2006), pp. 489–509. ISSN: 1557-9654. DOI: 10.1109/TIT.2005.862083.

[7] Samprit Chatterjee and Ali S Hadi. "Influential observations, high leverage points, and outliers in linear regression". In: *Statistical science* (1986), pp. 379–393.

[8] Tamás Erdélyi, Cameron Musco, and Christopher Musco. "Fourier sparse leverage scores and approximate kernel learning". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 109–122.

[9] Simon Foucart and Holger Rauhut. *A Mathematical Introduction to Compressive Sensing*. Springer New York, June 2013. ISBN: 978-1-4939-0063-3.

[10] Felix Krahmer, Holger Rauhut, and Rachel Ward. "Local Coherence Sampling in Compressed Sensing". In: *Proceedings of the 10th International Conference on Sampling Theory and Applications* ().

[11] Felix Krahmer and Rachel Ward. "Stable and Robust Sampling Strategies for Compressive Imaging". In: *IEEE Transactions on Image Processing* 23.2 (Feb. 2014), pp. 612–622. ISSN: 1057-7149, 1941-0042. DOI: 10.1109/TIP.2013.2288004. (Visited on 12/15/2022).

[12] Ziwei Liu et al. "Deep Learning Face Attributes in the Wild". In: *Proceedings of International Conference on Computer Vision (ICCV)*. Dec. 2015.

[13] Ilya Loshchilov and Frank Hutter. *Decoupled Weight Decay Regularization*. 2019. arXiv: 1711.05101 [cs.LG].

[14] Ping Ma, Michael Mahoney, and Bin Yu. "A statistical perspective on algorithmic leveraging". In: *International conference on machine learning*. PMLR. 2014, pp. 91–99.

[15] Maria Murad et al. "Radial Undersampling-Based Interpolation Scheme for Multislice CSMRI Reconstruction Techniques". In: *BioMed Research International* 2021 (2021), p. 6638588. ISSN: 2314-6141. DOI: 10.1155/2021/6638588.

[16] Paul Nevai. "Géza Freud, orthogonal polynomials and Christoffel functions. A case study". In: *Journal of Approximation Theory* 48.1 (1986), pp. 3–167. ISSN: 0021-9045. DOI: https://doi.org/10.1016/0021-9045(86)90016-X. URL: https://www.sciencedirect.com/science/article/pii/002190458690016X.

[17] Adam Paszke et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[18] Gilles Puy, Pierre Vandergheynst, and Yves Wiaux. "On Variable Density Compressive Sampling". In: *IEEE Signal Processing Letters* 18.10 (Oct. 2011), pp. 595–598. ISSN: 1070-9908, 1558-2361. DOI: 10.1109/LSP.2011.2163712. arXiv: 1109.6202 [cs, math]. (Visited on 12/16/2022).

[19] Holger Rauhut and Rachel Ward. "Sparse Legendre Expansions via 1-Minimization". In: *Journal of Approximation Theory* 164.5 (May 2012), pp. 517–533. ISSN: 0021-9045. DOI: 10.1016/j.jat.2012.01.008. (Visited on 04/23/2023).

[20] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, Sept. 2018. ISBN: 978-1-108-24454-1.

[21] Yuanbo Xiangli et al. "Real or Not Real, that is the Question". In: (2020).

This supplementary material contains acknowledgements, a generalization of Theorem 2.1, the proof of said generalization, properties of the piecewise linear expansion, and additional image recoveries both in-range and out-of-range.

A1 Acknowledgements

The authors acknowledge Ben Adcock for providing feedback on a preliminary version of this paper. S. Brugiapaglia acknowledges the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) through grant RGPIN-2020- 06766 and the Fonds de Recherche du Québec Nature et Technologies (FRQNT) through grant 313276. Y. Plan is partially supported by an NSERC Discovery Grant (GR009284), an NSERC Discovery Accelerator Supplement (GR007657), and a Tier II Canada Research Chair in Data Science (GR009243). O. Yilmaz was supported by an NSERC Discovery Grant (22R82411) O. Yilmaz also acknowledges support by the Pacific Institute for the Mathematical Sciences (PIMS).

A2 Additional Notation

In this work we make use of absolute constants which we always label C. Any constant labelled C is implicitly understood to be absolute and has a value that may differ from one appearance to the next. We write $a \lesssim b$ to mean $a \leq Cb$. We denote by $\|\cdot\|$ the operator norm, and by $\|\cdot\|_2$ the euclidean norm. We use capital letters for matrices and boldface lowercase letters for vectors. For some matrix $A \in \mathbb{C}^{m\times n}$, we denote by $\boldsymbol{a}_i^* \in \mathbb{C}^{1\times n}$ (lowercase of the letter symbolizing the matrix) the ith row vector of A, meaning that $A = \sum_{i=1}^m \boldsymbol{e}_i \boldsymbol{a}_i^*$. For a vector $\boldsymbol{u} \in \mathbb{R}^n$, we denote by u_i its ith entry. We let $\{\boldsymbol{e}_i\}_{i\in[n]}$ be the canonical basis of $\mathbb{R}^n$.

A3 Generalized Main Result

We present a generalization of Theorem 2.1 which provides recovery guarantees for arbitrary sampling probabilities. To quantify the quality of the interaction between the generative model G, the unitary matrix F, and the sampling probability vector $\boldsymbol{p}$, we introduce the following quantity.

Definition A3.1. Let $\mathcal{T} \subseteq \mathbb{R}^n$ be a cone. Let $F \in \mathbb{C}^{n\times n}$ be a unitary matrix and $\boldsymbol{p} \in \Delta^{n-1}$. Let $\boldsymbol{\alpha} \in \mathbb{R}^n$ the local coherences of F with respect to $\mathcal{T}$. Then define the quanity

$$\mu_{\mathcal{T}}(F, \boldsymbol{p}) := \max_{j\in[n]} \frac{\alpha_j}{\sqrt{p_j}}.$$

We can now state the following.

Theorem A3.1 (Generalized Main Result)**.** *Fix the (k, d, n)-generative network G, the cone $\mathcal{T} := \mathcal{R}(G) - \mathcal{R}(G) \subseteq \mathbb{R}^n$, the unitary matrix $F \in \mathbb{C}^{n\times n}$, the probability vector $\boldsymbol{p} \in \Delta^{n-1}$, and the corresponding random sampling matrix $S \in \mathbb{R}^{m\times n}$. Let $D \in \mathbb{R}^{n\times n}$ be a diagonal matrix with entries $D_{i,i} = \frac{1}{\sqrt{p_i}}$. Let $\tilde{D} := SDS^+ \in \mathbb{R}^{m\times m}$. Let $\boldsymbol{\alpha}$ be the vector of local coherences of F with respect to $\mathcal{T}$. Let $\varepsilon > 0$.*

Suppose that

$$m \geq C\mu_{\mathcal{T}}^2(F, \boldsymbol{p}) \left(kd \log\left(\frac{n}{k}\right) + \log\frac{2}{\varepsilon} \right).$$

Furthermore, if we pick the sampling probability vector

$$\boldsymbol{p}^* := \left(\frac{\alpha_j^2}{\|\boldsymbol{\alpha}\|_2^2} \right)_{j\in[n]},$$

we only require that

$$m \geq C\|\boldsymbol{\alpha}\|_2^2 \left(kd \log\left(\frac{n}{k}\right) + \log\frac{2}{\varepsilon} \right).$$

Then with probability at least $1 - \varepsilon$ over the realization of S, Statement 2.1 holds.

This theorem is strictly more general than Theorem 2.1. It is therefore sufficient to prove the generalized version, which we do in the next section.

Remark A3.1. The result [4, Theorem 2.1] is a corollary of Theorem 2.1; it follows from taking $\boldsymbol{p}$ as the uniform probability vector. ◇

A4 Proof of Theorem A3.1

Let us first introduce the so-called Restricted Isometry Property (RIP) [6, 9, 20].

Definition A4.1 (Restricted Isometry Property)**.** Let $\mathcal{T} \subseteq \mathbb{R}^n$ be a cone and $A \in \mathbb{C}^{m\times n}$ a matrix. We say that A satisfies the Restricted Isometry Property (RIP) when

$$\sup_{\boldsymbol{u}\in\mathcal{T}\cap\mathbb{S}^{n-1}} |\|A\boldsymbol{u}\|_2 - 1| \leq \frac{1}{3}.$$

Note that the constant $1/3$ is a specific choice made in order to simplify the presentation of this proof. It could be replaced by any generic absolute constant in $(0, 1)$.

The following lemma says that if, conditioning on S, $\frac{1}{\sqrt{m}}SDF$ has the RIP on $\mathcal{R}(G) - \mathcal{R}(G)$, then we have signal recovery.

Lemma A4.1 (RIP of a Subsampled and Preconditioned Matrix Yields Recovery)**.** *Let $\mathcal{V} \subseteq \mathbb{R}^n$ be a cone, $F \in \mathbb{C}^{n\times n}$ be a unitary matrix, $S \in \mathbb{R}^{m\times n}$ a matrix with all rows in the canonical basis of $\mathbb{R}^n$. Let $D \in \mathbb{R}^{n\times n}$ be a diagonal matrix. Let $\tilde{D} := SDS^+ \in \mathbb{R}^{m\times m}$.*

If $\frac{1}{\sqrt{m}}SDF$ has the RIP on the cone $\mathcal{T} := \mathcal{V} - \mathcal{V}$, then Statement 2.1 holds.

See the proof in Appendix Appendix A5.

We now proceed to prove a slightly stronger statement than what is required by Lemma A4.1; that the RIP holds on the piecewise linear expansion $\Delta(\mathcal{R}(G) - \mathcal{R}(G)) \supseteq \mathcal{R}(G) - \mathcal{R}(G)$.

To control the complexity of G, we count the number of affine pieces that it comprises. We do this with the result [4, Lemma A.6], which we re-write below for convenience.

Lemma A4.2 (Containing the Range of a ReLU Network in a Union of Subspaces)**.** *Let G be a (k, d, n)-generative network with layer widths $k = k_0 \leq k_1, \ldots, k_{d-1} \leq k_d$ where $k_d = n$ and $\bar{k} := \left(\prod_{\ell=1}^{d-1} k_\ell\right)^{1/(d-1)}$. Then $\mathcal{R}(G)$ is a union of no more than N at-most k-dimensional polyhedral cones where*

$$\log N \leq k(d-1)\log\left(\frac{2e\bar{k}}{k}\right) \lesssim kd\log\left(\frac{n}{k}\right).$$

From this result, we see that $\mathcal{T} := \mathcal{R}(G) - \mathcal{R}(G)$ is contained in a union of no more than N^2 affine pieces each of dimension no more than $2k$. Then from Remark A6.1 (for the proof, see [4, Remark A.2]), the cone $\Delta(\mathcal{T})$ is a union of no more than N^2 subspaces each of dimension at-most $2k$ (the factor of two will be absorbed into the absolute constant of the statement.) Fix $\mathcal{U} \subseteq \mathcal{T}$ to be any one of these subspaces. Then the following lemma implies that the matrix $\frac{1}{\sqrt{m}}SDF$ has the RIP on $\mathcal{U}$ with high probability.

Lemma A4.3 (Deviation of Subsampled Preconditioned Unitary Matrix on a Subspace)**.** *Let $F \in \mathbb{C}^{n\times n}$ be a unitary matrix, and $S \in \mathbb{R}^{m\times n}$ a random sampling matrix associated with the probability vector $\boldsymbol{p} \in \Delta^{n-1}$. Let $D \in \mathbb{R}^n$ be a diagonal pre-conditioning matrix with entries $D_{i,i} = \frac{1}{\sqrt{p_i}}$. Let $t > 0$. Let $\mathcal{U} \subseteq \mathbb{R}^n$ be a subspace of dimension k. Then*

$$\sup_{\boldsymbol{x}\in\mathcal{U}\cap\mathbb{S}^{n-1}} \left|\frac{1}{\sqrt{m}}\|SDF\boldsymbol{x}\|_2 - 1\right| \lesssim \frac{\mu_{\mathcal{U}}(F,\boldsymbol{p})}{\sqrt{m}}\sqrt{\log k} + \frac{\mu_{\mathcal{U}}(F,\boldsymbol{p})}{\sqrt{m}}t \tag{1}$$

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

See the proof in Appendix A5.

Since $\mathcal{U} \subseteq \Delta(\mathcal{T})$ we have that $\mu_{\mathcal{U}}(F,\boldsymbol{p}) \leq \mu_{\mathcal{T}}(F,\boldsymbol{p})$. Using this fact to upper-bound the r.h.s. of Equation 1, we find an identical concentration inequality that applies to each of the subspaces constituting $\Delta(\mathcal{T})$. By using Lemma A4.2 to bound the number of subspaces, we control the deviation of $\frac{1}{\sqrt{m}}SDF$ uniformly over all the subspaces constituting $\Delta(\mathcal{T})$ with a union bound. We find that, with probability at least $1 - 2\exp(-t^2)$,

$$\sup_{\boldsymbol{x}\in\Delta(\mathcal{T})\cap\mathbb{S}^{n-1}} \left|\frac{1}{\sqrt{m}}\|SDF\boldsymbol{x}\|_2 - 1\right|$$

$$\lesssim \frac{\mu_{\mathcal{T}}(F,\boldsymbol{p})}{\sqrt{m}}\sqrt{\log k} + \frac{\mu_{\mathcal{T}}(F,\boldsymbol{p})}{\sqrt{m}}\sqrt{kd\log\left(\frac{n}{k}\right)} + \frac{\mu_{\mathcal{T}}(F,\boldsymbol{p})}{\sqrt{m}}t.$$

For the method by which we applied the union bound, see [4, Lemma A.2]. In the r.h.s. of the equation above, the second term dominates the first, so the expression simplifies to

$$\sup_{\boldsymbol{x}\in\Delta(\mathcal{T})\cap\mathbb{S}^{n-1}}\left|\frac{1}{\sqrt{m}}\|SDF\boldsymbol{x}\|_2-1\right|\lesssim\frac{\mu_{\mathcal{T}}(F,\boldsymbol{p})}{\sqrt{m}}\sqrt{kd\log\left(\frac{n}{k}\right)}+\frac{\mu_{\mathcal{T}}(F,\boldsymbol{p})}{\sqrt{m}}t. \tag{2}$$

By fixing $t=\sqrt{\log\left(\frac{2}{\varepsilon}\right)}$ we find that the RIP holds with probability at least $1-\varepsilon$ on $\mathcal{T}$ when

$$m\geq C\mu_{\mathcal{T}}(F,\boldsymbol{p})^2\left(kd\log\left(\frac{n}{k}\right)+\log\frac{2}{\varepsilon}\right). \tag{3}$$

Then we find that the first part of Theorem A3.1 follows from Lemma A4.1.

The second sufficient condition on m follows from picking $\boldsymbol{p}$ so as to minimize the factor $\mu_{\mathcal{T}}(F,\boldsymbol{p})^2$ in Equation 3.

Lemma A4.4 (Adapting the Sampling Scheme to the Model)**.** *Let $F\in\mathbb{C}^{n\times n}$ be a unitary matrix, and $S\in\mathbb{R}^{m\times n}$ a random sampling matrix associated with the probability vector $\boldsymbol{p}\in\Delta^{n-1}$. Let $\boldsymbol{\alpha}$ be the local coherences of F with respect to a cone $\mathcal{T}\in\mathbb{R}^n$.*

$$\boldsymbol{p}^*_j:=\left(\frac{\alpha_j^2}{\|\boldsymbol{\alpha}\|_2^2}\right)_{j\in[n]}\in\operatorname*{arg\,min}_{\boldsymbol{p}\in\Delta^{n-1}}(\mu_{\mathcal{T}}(F,\boldsymbol{p})).$$

It achieves a value of

$$\mu_{\mathcal{T}}(F,\boldsymbol{p}^*)=\|\boldsymbol{\alpha}\|_2.$$

See the proof.

Applying Lemma A4.4 to Equation 3 concludes the proof of Theorem A3.1.

A5 Proof of the Lemmas

Proof of Lemma A4.1. Let $F\in\mathbb{C}^{n\times n}$ be a unitary matrix, and $S\in\mathbb{R}^{m\times n}$ a random sampling matrix associated with the probability vector $\boldsymbol{p}\in\Delta^{n-1}$. Let $D\in\mathbb{R}^{n\times n}$ be a diagonal matrix with entries $D_{i,i}=\frac{1}{\sqrt{p_i}}$. Let $\tilde{D}=SDS^+$.

We let $\tilde{\boldsymbol{b}}:=\tilde{D}\boldsymbol{b}$ and $\tilde{\boldsymbol{\eta}}:=\tilde{D}\boldsymbol{\eta}$. By left-multiplying the equation $\boldsymbol{b}=\frac{1}{\sqrt{m}}SF\boldsymbol{x}_0+\boldsymbol{\eta}$ by $\tilde{D}$ we get $\tilde{\boldsymbol{b}}=\frac{1}{\sqrt{m}}SDF\boldsymbol{x}_0+\tilde{\boldsymbol{\eta}}$. Notice that the linear operator $\frac{1}{\sqrt{m}}SDF$ has the RIP by assumption.

By triangle inequality and the observation that $\Pi_{\mathcal{T}}\,\boldsymbol{x}_0\in\mathcal{T}$,

$$\begin{aligned}\left\|\frac{1}{\sqrt{m}}SDF\hat{\boldsymbol{x}}-\tilde{\boldsymbol{b}}\right\|_2&\leq\min_{\boldsymbol{x}\in\mathcal{T}}\left\|\frac{1}{\sqrt{m}}SDF\boldsymbol{x}-\tilde{\boldsymbol{b}}\right\|_2+\hat{\varepsilon}\\&\leq\left\|\frac{1}{\sqrt{m}}SDF\,\Pi_{\mathcal{T}}\,\boldsymbol{x}_0-\tilde{\boldsymbol{b}}\right\|_2+\hat{\varepsilon}\\&=\left\|\frac{1}{\sqrt{m}}SDF\boldsymbol{x}^{\perp}+\tilde{\boldsymbol{\eta}}\right\|_2+\hat{\varepsilon}\\&\leq\frac{1}{\sqrt{m}}\|SDF\boldsymbol{x}^{\perp}\|_2+\|\tilde{\boldsymbol{\eta}}\|_2+\hat{\varepsilon}.\end{aligned}$$

Since $\hat{\boldsymbol{x}},\Pi_{\mathcal{T}}\,\boldsymbol{x}_0\in\mathcal{T}$, with the RIP property we find that

$$\begin{aligned}\left\|\frac{1}{\sqrt{m}}SDF\hat{\boldsymbol{x}}-\tilde{\boldsymbol{b}}\right\|_2&=\left\|\frac{1}{\sqrt{m}}SDF\left(\hat{\boldsymbol{x}}-\Pi_{\mathcal{T}}\,\boldsymbol{x}_0\right)-\frac{1}{\sqrt{m}}SDF\left(\boldsymbol{x}_0-\Pi_{\mathcal{T}}\,\boldsymbol{x}_0\right)-\tilde{\boldsymbol{\eta}}\right\|_2\\&\geq\frac{1}{\sqrt{m}}\|SDF\left(\hat{\boldsymbol{x}}-\Pi_{\mathcal{T}}\,\boldsymbol{x}_0\right)\|_2-\frac{1}{\sqrt{m}}\|SDF\boldsymbol{x}^{\perp}\|_2-\|\tilde{\boldsymbol{\eta}}\|_2\\&\geq\left(1-\frac{1}{3}\right)\|\hat{\boldsymbol{x}}-\Pi_{\mathcal{T}}\,\boldsymbol{x}_0\|_2-\frac{1}{\sqrt{m}}\|SDF\boldsymbol{x}^{\perp}\|_2-\|\tilde{\boldsymbol{\eta}}\|_2.\end{aligned}$$

Assembling the two inequalities gives

$$\|\hat{\boldsymbol{x}}-\Pi_{\mathcal{T}}\,\boldsymbol{x}_0\|_2\leq\frac{1}{1-\frac{1}{3}}\left[2\frac{1}{\sqrt{m}}\|SDF\boldsymbol{x}^{\perp}\|_2+2\|\tilde{\boldsymbol{\eta}}\|_2+\hat{\varepsilon}\right].$$

Finally, we apply triangle inequality to get

$$\begin{aligned}\|\hat{\boldsymbol{x}} - \boldsymbol{x}_0\|_2 &\leq \|\boldsymbol{x}_0 - \Pi_{\mathcal{T}}\, \boldsymbol{x}_0\|_2 + \|\hat{\boldsymbol{x}} - \Pi_{\mathcal{T}}\, \boldsymbol{x}_0\|_2 \\ &\leq \|\boldsymbol{x}^{\perp}\|_2 + \frac{3}{\sqrt{m}}\|SDF\boldsymbol{x}^{\perp}\|_2 + 3\|\tilde{\boldsymbol{\eta}}\|_2 + \frac{3}{2}\hat{\varepsilon}.\end{aligned}$$

□

Proof of Lemma A4.3. In what follows, we will use that $\forall \boldsymbol{x} \in \mathcal{U}$, $SDF\boldsymbol{x} = SDFP_{\mathcal{U}}^* P_{\mathcal{U}}\boldsymbol{x}$ where $P_{\mathcal{U}} \in \mathbb{R}^{k\times n}$ is the matrix with rows chosen to be any fixed orthonormal basis of $\mathcal{U}$. Indeed, notice that $P_{\mathcal{U}}^* P_{\mathcal{U}} = \Pi_{\mathcal{U}} \in \mathbb{R}^{n\times n}$, the orthonormal projection on to $\mathcal{U}$. Now consider

$$\begin{aligned}(\star) := \sup_{\boldsymbol{x}\in\mathcal{U}\cap\mathbb{S}^{n-1}} \left|\frac{1}{m}\,\|SDF\boldsymbol{x}\|_2^2 - 1\right| &= \sup_{\boldsymbol{x}\in\mathcal{U}\cap\mathbb{S}^{n-1}} \left|\frac{1}{m}\,\|SDFP_{\mathcal{U}}^* P_{\mathcal{U}}\boldsymbol{x}\|_2^2 - 1\right| \\ &= \sup_{\boldsymbol{u}\in\mathbb{R}^k\cap\mathbb{S}^{k-1}} \left|\frac{1}{m}\,\|SDFP_{\mathcal{U}}^*\boldsymbol{u}\|_2^2 - 1\right| \\ &= \frac{1}{m}\sup_{\boldsymbol{u}\in\mathbb{R}^k\cap\mathbb{S}^{k-1}} \left|\boldsymbol{u}^*\left[(SDFP_{\mathcal{U}}^*)^*(SDFP_{\mathcal{U}}^*) - mI\right]\boldsymbol{u}\right|.\end{aligned}$$

The second equality above follows from a change of variables $P_{\mathcal{U}}\boldsymbol{x} \to \boldsymbol{u} \in \mathbb{R}^k$. Since the matrix within the square bracket is symmetric, the last expression we find above corresponds to an operator norm.

$$(\star) = \frac{1}{m}\,\|P_{\mathcal{U}}F^*DS^*SDFP_{\mathcal{U}}^* - mI\| \tag{4}$$

$$= \frac{1}{m}\left\|\sum_{i=1}^{m}\left[P_{\mathcal{U}}F^*D\boldsymbol{s}_i\boldsymbol{s}_i^*DFP_{\mathcal{U}}^* - I\right]\right\| \tag{5}$$

This is a sum of independent random matrices because the sampling matrix matrix S is random and has independent rows. We now consider what will be the central ingredient of this proof: the Matrix Bernstein concentration bound [20, Theorem 5.4.1]. We will use it to bound Equation 5.

Lemma A5.1 (Matrix Bernstein)**.** *Let $X_1, ..., X_N$ be independent, mean zero, $n\times n$ symmetric random matrices, such that $||X_i|| \leq K$ almost surely for all i. Then, for every $t \geq 0$, we have*

$$\mathbb{P}\left\{\left\|\sum_{i=1}^{N} X_i\right\| \geq t\right\} \leq 2n\exp\left(-\frac{t^2/2}{\sigma^2 + Kt/3}\right),$$

where $\sigma^2 = \left\|\sum_{i=1}^{N}\mathbb{E}X_i^2\right\|$.

To compute σ^2 and K, we notice that we can write

$$\sum_{i=1}^{m}\left[P_{\mathcal{U}}F^*D\boldsymbol{s}_i\boldsymbol{s}_i^*DFP_{\mathcal{U}}^* - I\right] = \sum_{i\in[n]}\left[\boldsymbol{v}_i\boldsymbol{v}_i^* - I\right]$$

for the random vectors $\boldsymbol{v}_i := P_{\mathcal{U}}F^*D\boldsymbol{s}_i$. These vectors have two key properties. First, they are *isotropic*; this is the property that

$$\begin{aligned}\mathbb{E}[\boldsymbol{v}_i\boldsymbol{v}_i^*] &= \mathbb{E}[P_{\mathcal{U}}F^*D\boldsymbol{s}_i\boldsymbol{s}_i^*DFP_{\mathcal{U}}^*] \\ &= P_{\mathcal{U}}F^*D\mathbb{E}[\boldsymbol{s}_i\boldsymbol{s}_i^*]DFP_{\mathcal{U}}^* = I.\end{aligned}$$

The isotropic property gives us immediately that, as required, the matrices $\{\boldsymbol{v}_i\boldsymbol{v}_i^* - I\}_{i\in[m]}$ are mean-zero.

The second property of the vectors $\{\boldsymbol{v}_i\}_{i\in[m]}$ is that they have bounded magnitude almost surely.

$$\begin{aligned}\|\boldsymbol{v}_i\|_2 &= \|P_{\mathcal{U}}(F^*D\boldsymbol{s}_i)\|_2 \\ &= \frac{1}{\sqrt{p_i}}\|P_{\mathcal{U}}\boldsymbol{f}_i\|_2 \\ &= \frac{1}{\sqrt{p_i}}\sup_{\boldsymbol{x}\in\mathcal{U}\cap\mathbb{S}^{n-1}}\langle\boldsymbol{x}, \boldsymbol{f}_i\rangle \\ &\leq \mu_{\mathcal{U}}(F, \boldsymbol{p}).\end{aligned}$$

Let $\mu := \mu_{\mathcal{U}}(F, \boldsymbol{p})$ for conciseness. We proceed to compute a value for K. By triangle inequality and property of the operator norm of rank one matrices, we see that

$$\|\boldsymbol{v}_i\boldsymbol{v}_i^* - I\| \leq \|\boldsymbol{v}_i\|_2^2 + 1 \leq 2\mu^2.$$

The last inequality holds because of the lower bound $\mu^2 \geq 1$, which we now justify. Consider that from Lemma A4.4 we have that $\mu \geq \|\boldsymbol{\alpha}\|_2$, and furthermore that for any fixed one-dimensional subspace $\mathcal{U}_0 \subseteq \mathcal{U}$, we have that $\|\boldsymbol{\alpha}\|_2 = \|F\hat{\boldsymbol{u}}\|_2 = 1$ for a unit vector $\hat{\boldsymbol{u}} \in \mathcal{U}_0$. This gives us the desired lower bound by monotonicity of μ over set containment.

We now compute σ^2, similarly [3, Lemma 12.21]: Then

$$\begin{aligned}
\sigma^2 &= \left\| \sum_{i=1}^m \mathbb{E}\left[(\boldsymbol{v}_i\boldsymbol{v}_i^* - I)^2\right] \right\| \\
&= \sup_{\boldsymbol{u}\in\mathbb{R}^k\cap\mathbb{S}^{k-1}} \left\langle \boldsymbol{u}, \sum_{i=1}^m (\mathbb{E}[\boldsymbol{v}_i\boldsymbol{v}_i^*\boldsymbol{v}_i\boldsymbol{v}_i^*] - I)\boldsymbol{u} \right\rangle \\
&= \sup_{\boldsymbol{u}\in\mathbb{R}^k\cap\mathbb{S}^{k-1}} \left\langle \boldsymbol{u}, \sum_{i=1}^m \|\boldsymbol{v}_i\|_2^2 \mathbb{E}[\boldsymbol{v}_i\boldsymbol{v}_i^*]\boldsymbol{u} \right\rangle - \sum_{i=1}^m \|\boldsymbol{u}\|_2^2 \\
&\leq \sup_{\boldsymbol{u}\in\mathbb{R}^k\cap\mathbb{S}^{k-1}} \left\langle \boldsymbol{u}, \sum_{i=1}^m \mu^2 I\boldsymbol{u} \right\rangle - \sum_{i=1}^m \|\mathbb{E}[\boldsymbol{v}_i\boldsymbol{v}_i^*]\boldsymbol{u}\|_2^2 \\
&\leq \mu^2 m.
\end{aligned}$$

The second equality holds because the matrix is symmetric non-negative definite, and the last inequality is obtained by dropping the second negative term.

Then applying the Matrix Bernstein yields

$$\mathbb{P}\left\{ \left\| \sum_{i=1}^m [P_{\mathcal{U}}F^* D\boldsymbol{s}_i\boldsymbol{s}_i^* DFP_{\mathcal{U}}^* - I] \right\| \geq t \right\} \leq 2k\exp\left(-\frac{t^2/2}{\mu^2 m + 2\mu^2\frac{t}{3}}\right).$$

Substituting with Equation 5, we get

$$\mathbb{P}\left\{ \sup_{\boldsymbol{x}\in\mathcal{U}\cap\mathbb{S}^{n-1}} \left| \frac{1}{m}\|SDF\boldsymbol{x}\|_2^2 - 1 \right| \geq \frac{t}{m} \right\} \leq 2k\exp\left(-\frac{t^2/2}{\mu^2 m + 2\mu^2\frac{t}{3}}\right).$$

We would like to get our result in terms of the l_2 norm without the square. For this purpose we make use of the "square-root trick" that can be found in [20, Theorem 3.1.1]. We re-write the above as

$$\mathbb{P}\left\{ \sup_{\boldsymbol{x}\in\mathcal{U}\cap\mathbb{S}^{n-1}} \left| \frac{1}{m}\|SDF\boldsymbol{x}\|_2^2 - 1 \right| \geq \frac{t}{m} \right\} \leq 2k\exp\left(-C\min\left(\frac{t^2}{\mu^2 m}, \frac{t}{\mu^2}\right)\right).$$

We make the substitution $t \to m\max(\delta, \delta^2)$, which yields

$$\mathbb{P}\left\{ \sup_{\boldsymbol{x}\in\mathcal{U}\cap\mathbb{S}^{n-1}} \left| \frac{1}{m}\|SDF\boldsymbol{x}\|_2^2 - 1 \right| \geq \max(\delta, \delta^2) \right\} \leq 2k\exp\left(-C\frac{m\delta^2}{\mu^2}\right).$$

With the restricted inequality $\forall a, \delta > 0, |a-1| \geq \delta \implies |a^2 - 1| \geq \max(\delta, \delta^2)$, we infer that

$$\mathbb{P}\left\{ \sup_{\boldsymbol{x}\in\mathcal{U}\cap\mathbb{S}^{n-1}} \left| \frac{1}{\sqrt{m}}\|SDF\boldsymbol{x}\|_2 - 1 \right| \geq \delta \right\} \leq 2k\exp\left(-C\frac{m\delta^2}{\mu^2}\right).$$

Finally, with another substitution $\left(\frac{cm\delta^2}{\mu^2} - \log k\right) \to t^2$ we write that

$$\sup_{\boldsymbol{x}\in\mathcal{U}\cap\mathbb{S}^{n-1}} \left| \frac{1}{\sqrt{m}}\|SDF\boldsymbol{x}\|_2 - 1 \right| \lesssim \frac{\mu}{\sqrt{m}}\sqrt{\log k} + \frac{\mu}{\sqrt{m}}t$$

with probability at least $1 - 2\exp(-t^2)$. □

Proof of Lemma A4.4. It suffices to show that $\boldsymbol{p}^* := \left(\frac{\alpha_j^2}{\|\boldsymbol{\alpha}\|_2^2}\right)_{j\in[n]}$ satisfies

$$\boldsymbol{p}^* \in \arg\min_{\boldsymbol{p}\in\Delta^{n-1}} \mu_{\mathcal{T}}^2(F,\boldsymbol{p}) = \arg\min_{\boldsymbol{p}\in\Delta^{n-1}} \max_{j\in[n]} \frac{\alpha_j^2}{p_j}.$$

The vector $\boldsymbol{p}^*$ achieves a value of

$$\mu_{\mathcal{T}}^2(F,\boldsymbol{p}^*) = \|\boldsymbol{\alpha}\|_2^2$$

which is the minimum. Indeed, for any fixed vector $\boldsymbol{p} \in \Delta^{n-1}$,

$$\max_{j\in[n]} \frac{\alpha_j^2}{p_j} \geq \sum_{j=1}^{n} a_j \frac{\alpha_j^2}{p_j} \quad \forall \boldsymbol{a} \in \Delta^{n-1}.$$

The inequality above holds because the r.h.s. is a convex combination of the terms $\left\{\frac{\alpha_j^2}{p_j}\right\}_{j\in[n]}$, and is therefore upper-bounded by the maximum element of the combination. By letting $\boldsymbol{a} = \boldsymbol{p}$, we get

$$\max_{j\in[n]} \frac{\alpha_j^2}{p_j} \geq \|\boldsymbol{\alpha}\|_2^2.$$

Therefore, $\boldsymbol{p}^* \in \arg\min_{\boldsymbol{p}\in\Delta^{n-1}} \mu_{\mathcal{T}}^2(F,\boldsymbol{p})$. □

A6 Properties of the Piecewise Linear Expansion

The following is a subset of the elements in remark [4, Remark A.2], to which we refer the reader for the proof.

Remark A6.1 (Properties of the Piecewise Linear Expansion). Below we list several properties about Δ. Let $\mathcal{C} = \bigcup_{i=1}^{N} \mathcal{C}_i$ be the union of $N \in \mathbb{N}$ convex cones $\mathcal{C}_i$.

1. The set $\Delta(\mathcal{C})$ is uniquely defined. In particular, it is independent of the (finite) decomposition of $\mathcal{C}$ into convex cones.
2. If $\max_{i\in[N]} \dim \mathcal{C}_i \leq k$, then $\Delta(\mathcal{C})$ is a union of no more than N at-most k-dimensional linear subspaces.
3. The set $\Delta(\mathcal{C})$ satisfies $\mathcal{C} \subseteq \Delta(\mathcal{C}) \subseteq \mathcal{C} - \mathcal{C}$.
4. There are choices of $\mathcal{C}$ for which $\mathcal{C} \subsetneq \Delta(\mathcal{C})$ (for instance, refer to the example at the end of this section).

◇

A7 Experimental Specifications

CelebA with RealnessGAN CelebFaces Attributes Dataset (CelebA) is a dataset with over 200,000 celebrity face images [12]. We train a model on most images of the CelebA dataset, leaving out 2000 images to comprise a validation set. We crop the colour images to 256 by 256, leading to 256 × 256 × 3 = 196608 pixels per image. On this dataset, we train a RealnessGAN with the same training setup as described in [21], substituting the last Tanh layer with HardTanh, a linearized version of Tanh, to fit in our theoretical framework. See [21] for more training and architecture details.

A8 Additional Image Recoveries

Figure 3: In-range signal recovery of images using uniform sampling and model-adapted sampling. The sampling rate on the bottom is computed using the ratio between number of measurements and number of pixels.

Figure 4: Out-of-range signal recovery of images using uniform sampling and model-adapted sampling.