Beyond Independent Measurements: General Compressed Sensing with GNN Application

Alireza Naderi
Department of Mathematics
University of British Columbia
Vancouver, BC V6T 1Z2
alireza@math.ubc.ca

Yaniv Plan
Department of Mathematics
University of British Columbia
Vancouver, BC V6T 1Z2
yaniv@math.ubc.ca

Abstract

We consider the problem of recovering a structured signal $\mathbf{x} \in \mathbb{R}^n$ from noisy linear observations $\mathbf{y} = \mathbf{M}\mathbf{x} + \mathbf{w}$. The measurement matrix is modeled as $\mathbf{M} = \mathbf{B}\mathbf{A}$, where $\mathbf{B} \in \mathbb{R}^{l \times m}$ is arbitrary and $\mathbf{A} \in \mathbb{R}^{m \times n}$ has independent sub-gaussian rows. By varying $\mathbf{B}$, and the sub-gaussian distribution of $\mathbf{A}$, this gives a family of measurement matrices which may have heavy tails, dependent rows and columns, and singular values with a large dynamic range. When the structure is given as a possibly non-convex cone $T \subset \mathbb{R}^n$, an approximate empirical risk minimizer is proven to be a robust estimator if the effective number of measurements is sufficient, even in the presence of a model mismatch. In classical compressed sensing with independent (sub-)gaussian measurements, one asks *how many measurements are needed to recover* $\mathbf{x}$*?* In our setting, however, the effective number of measurements depends on the properties of $\mathbf{B}$. We show that the *effective rank* of $\mathbf{B}$ may be used as a surrogate for the number of measurements, and if this exceeds the squared *Gaussian mean width* of $(T - T) \cap \mathbb{S}^{n-1}$, then accurate recovery is guaranteed. Furthermore, we examine the special case of generative priors in detail, that is when $\mathbf{x}$ lies close to $T = \mathrm{ran}(G)$ and $G : \mathbb{R}^k \to \mathbb{R}^n$ is a Generative Neural Network (GNN) with ReLU activation functions. Our work relies on a recent result in random matrix theory by Jeong, Li, Plan, and Yılmaz [10].

1 Introduction

In compressed sensing [7], the goal is to reconstruct a high-dimensional signal $\mathbf{x} \in \mathbb{R}^n$ from a noisy low-dimensional linear transformation of it, $\mathbf{y} = \mathbf{M}\mathbf{x} + \mathbf{w} \in \mathbb{R}^l$, $l < n$. Even in the absence of noise, the reconstruction would not be possible without a further assumption of *signal structure*, i.e., some restriction of the possible values of $\mathbf{x}$. In early works in compressed sensing, the signal structure was encoded via sparsity, or sparsity with respect to a dictionary. A sufficient condition on the *measurement matrix* $\mathbf{M}$ that enables robust recovery is the celebrated Restricted Isometry Property (RIP) [4]. Intriguingly, all known algorithms for certifying RIP either are computationally intractable or only function in the parameter regimes which are highly suboptimal [6], but a sub-gaussian measurement matrix with independent rows satisfies the RIP with high probability in near optimal parameter regimes [2].

More recently, compressed sensing ideas have been generalized to allow the signal structure to be almost arbitrary, encoded as a subset $T \subset \mathbb{R}^n$, provided that the measurements are appropriately random. Gaussian measurement matrices are rotationally invariant [23], which allows them to be universally effective for general compressed sensing. This was proven using [5, 18] or generalizing [19, 15, 20] Gordon's theorem [8], and also through a foundational result in conic geometry [1]. Sub-gaussian matrices, as defined later, are essentially the largest class of matrices that approximately

NeurIPS 2021 Workshop on Deep Learning and Inverse Problems, virtual.

satisfy rotation invariance, and also enjoy universal guarantees, although with different proof techniques relying on chaining arguments [14, 11, 10]. We also note a few results outside of the Gaussian or sub-gaussian framework: [21, 16, 13, 12].

The recent results of [10] extend previous theory by allowing the measurement matrix to take the form $\mathbf{M} = \mathbf{BA}$ with arbitrary $\mathbf{B}$ and sub-gaussian $\mathbf{A}$ with independent rows. They prove a general restricted isometry property, thereby giving basic signal recovery guarantees for measurement matrices with dependent rows. This random model is compelling because real-data measurement matrices often have highly variable singular values, but sub-gaussian matrices with independent rows typically have nearly uniform singular values. In fact, under mild assumptions, the singular values of $\mathbf{BA}$ concentrate around the singular values of $\mathbf{B}$ (after rescaling), and so, one may target any singular value vector by adjusting the singular values of $\mathbf{B}$. We build upon the work of [10] to show accurate signal recovery by the generalized Lasso, even in the presence of inexact optimization and/or a mismatch in the signal structure. We believe this provides the first general compressed sensing guarantees allowing any two of the following three items simultaneously (we allow all three): (i) the measurement matrix need not have independent rows, (ii) the signal may be estimated by generalized Lasso, but without exact optimization, and (iii) the signal is allowed to be only *approximately* structured, i.e. close to the structure set and not belonging to it. The improved dependence on the sub-gaussian parameter of the measurement matrix in [10] is reflected in our work as well.

Items (ii) and (iii) above make our framework well-suited to the setting in which the signal structure is the range of a neural network. Indeed, in recent years, it has been shown empirically that learning the appropriate signal structure by fitting it to the range of a neural network can be much more effective than using a predetermined structure such as sparsity [3, 22, 9]. In this case, one assumes that T is the range of a generative model $G : \mathbb{R}^k \to \mathbb{R}^n$ which is already trained. While training often approximates the signal structure well, any new signal outside of the training set will typically be close, but not in T; thereby necessitating item (iii). The standard method of estimating $\mathbf{x}$ is to find the restricted least squares fit, i.e., to let the estimate $\hat{\mathbf{x}}$ be the solution to the program, sometimes called generalized Lasso:

$$\text{minimize } ||\mathbf{y} - \mathbf{M}\mathbf{x}'||_2^2 \quad \text{s.t. } \mathbf{x}' \in T. \tag{P}$$

Since $T = \text{ran}(G)$ is non-convex, and no convex relaxation of such a program is known in general, typical gradient-descent-based algorithms would not be guaranteed to approach a global minimum; thereby necessitating item (ii). Our theory requires an upper bound of the Gaussian mean width of $(T - T) \cap \mathbb{S}^{n-1}$, which we give in Proposition 3.2 below; we believe this is novel.

2 Main Results

2.1 Problem Setup

Recall that a random variable Z is called *sub-gaussian*, if its tail probability is dominated by that of a Gaussian random variable. An equivalent rigorous definition requires the *sub-gaussian norm* to be finite: $||Z||_{\psi_2} := \inf\left\{t > 0 \,\middle|\, \mathbb{E}(Z^2/t^2) \leq 2\right\} < \infty$. A random vector $\mathbf{v} \in \mathbb{R}^n$ is *sub-gaussian*, if all of its one-dimensional marginals are sub-gaussian random variables. Mathematically, if $||\langle\theta, \mathbf{v}\rangle||_{\psi_2}$ is finite for all $\theta \in \mathbb{S}^{n-1}$, we define $||\mathbf{v}||_{\psi_2} := \sup_{\theta\in\mathbb{S}^{n-1}} ||\langle\theta, \mathbf{v}\rangle||_{\psi_2}$.

We will consider a random matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ whose rows $\{\mathbf{a}_1^\top, \cdots, \mathbf{a}_m^\top\}$ are statistically independent, mean-zero ($\mathbb{E}\mathbf{a}_i = \mathbf{0}$), isotropic ($\mathbb{E}\mathbf{a}_i\mathbf{a}_i^\top = \mathbf{I}_n$), and sub-gaussian with parameter K ($||\mathbf{a}_i||_{\psi_2} \leq K$). We let the measurement matrix be the product $\mathbf{BA}$, where $\mathbf{B} \in \mathbb{R}^{l\times m}$ is arbitrary. In other words, we let every row of our measurement matrix be an arbitrary linear combination of $\mathbf{a}_1^\top, \cdots, \mathbf{a}_m^\top$. Our goal is to determine the criteria that $\mathbf{B}$, $\mathbf{A}$, and the structure set T must satisfy for accurate recovery to be possible. Informally, one requires $\mathbf{B}$ to be far from low-rank, otherwise, the number of independent effective measurements would not be sufficient. An appropriate quantity is the *stable rank* defined as

$$\text{sr}(\mathbf{B}) := \frac{||\mathbf{B}||_F^2}{||\mathbf{B}||^2} = \frac{\sum_{i=1}^{\text{rank}(\mathbf{B})} \sigma_i^2}{\max_i \sigma_i^2} \leq \text{rank}(\mathbf{B}), \tag{1}$$

where σ_i's are the singular values of $\mathbf{B}$. The main advantage of this definition over rank itself is that it is robust to the small non-zero singular values that increase the rank but do not contribute to an

effective measurement, hence the name stable rank. Furthermore, one requires T to be *small* in some sense, ideally not to intersect the null space of $\mathbf{M} = \mathbf{BA}$. A useful notion of size is the *Gaussian mean width* defined as

$$w(T) := \mathbb{E} \sup_{\mathbf{v} \in T} \langle \mathbf{v}, \mathbf{g} \rangle, \tag{2}$$

where the expectation is calculated with respect to $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$. The reader may consult [17] for some basic properties of Gaussian mean width. The following theorem gives the desired estimation bound.

2.2 Main Theorem

Theorem 2.1. *Let $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{B} \in \mathbb{R}^{l \times m}$ be an arbitrary fixed matrix, and $\mathbf{A} \in \mathbb{R}^{m \times n}$ be a matrix whose rows are independent, mean-zero, isotropic, and sub-gaussian vectors with sub-gaussian parameter K. Let $T \subset \mathbb{R}^n$ be a closed cone and define $T' := (T - T) \cap \mathbb{S}^{n-1}$. Let $\mathbf{y} = \mathbf{BAx} + \mathbf{w}$ for some fixed unknown $\mathbf{w} \in \mathbb{R}^m$. Let $\hat{\mathbf{x}} \in T$ satisfy $||\mathbf{y} - \mathbf{BA}\hat{\mathbf{x}}||_2^2 \leq \min_{\mathbf{x}' \in T} ||\mathbf{y} - \mathbf{BAx}'||_2^2 + \epsilon^2$. If $\mathrm{sr}(B) \gg K^2 \log K \cdot w^2(T')$, then with probability larger than $1 - 11e^{-w^2(T')}$,*

$$||\mathbf{x} - \hat{\mathbf{x}}||_2 \lesssim \frac{Kw(T')}{||\mathbf{B}||_F \sqrt{\mathrm{sr}(\mathbf{B})}} ||\mathbf{w}||_2 + \frac{\epsilon}{||\mathbf{B}||_F} + \frac{K\sqrt{l}}{\sqrt{\mathrm{sr}(\mathbf{B})}} \mathrm{dist}(\mathbf{x}, T). \tag{3}$$

There are three sources of error present here: the additive noise $\mathbf{w}$, the inaccuracy in optimization ϵ, and the model mismatch, i.e., $\mathbf{x}$ not exactly belonging to T. Note that $\mathrm{sr}(\mathbf{B}) \leq \mathrm{rank}(\mathbf{B}) \leq \min(l, m)$, so basically the effective number of measurements is bounded by the number of underlying independent measurements m. When $\mathbf{B}$ is a multiple of identity, the equality case $\mathrm{sr}(\mathbf{B}) = m$ occurs.

What happens if we have more effective measurements than needed? The energy of the noise that appears in the total error in (3) is decreased by the *oversampling factor* $\mathrm{sr}(\mathbf{B})/w^2(T')$. In other words, if we have some control over the measurement matrix, then we are able to denoise the original signal by increasing $\mathrm{sr}(\mathbf{B})$. In case of $\mathbf{B} = \mathbf{I}_m$, this simply means increasing the number of measurements m. This denoising effect is well known for stochastic noise, but we believe that it had previously been shown only for non-random noise under the assumption of a Gaussian measurement matrix, as in [20]. We note, it is important here that the noise is *fixed* i.e., not chosen adversarially depending on the realization of $\mathbf{A}$.

3 Application on Generative Neural Networks

Due to the success of deep generative neural networks to learn complex structures, recent works on compressed sensing have considered generative priors instead of the traditional sparsity [3]. While some works require the trained network $G : \mathbb{R}^k \to \mathbb{R}^n$ to be L-Lipschitz [3] and prove that $\mathcal{O}(k \log L)$ number of measurements would suffice for a recovery guarantee, a drawback of such analysis is that the Lipschitz constant of the network cannot be calculated solely based on its architecture. Here we present a different approach: we give an upper bound of the Gaussian mean width of $T' = \big(\mathrm{ran}(G) - \mathrm{ran}(G)\big) \cap \mathbb{S}^{n-1}$, which only depends on the hyperparameters of the model. Then we apply Theorem 2.1 to achieve a recovery guarantee for compressed sensing with generative priors.

3.1 GNN and Guassian Mean Width

A d-layer GNN with ReLU activation function is a function $G : \mathbb{R}^k \to \mathbb{R}^n$ of the form

$$G(\mathbf{z}) = \sigma(\mathbf{A}_d \sigma(\mathbf{A}_{d-1} \sigma(\dots \mathbf{A}_2 \sigma(\mathbf{A}_1 \mathbf{z}) \dots))), \tag{4}$$

where $\sigma(\cdot) = \max(\cdot, 0)$ is applied entrywise and $\mathbf{A}_i \in \mathbb{R}^{p_i \times p_{i-1}}$, $p_0 = k$, and $p_d = n$. The weight matrices $\mathbf{A}_i$ are sometimes assumed to have iid Gaussian entries, however, our analysis puts no requirement on $\mathbf{A}_i$. To bound the Gaussian mean width of $\mathrm{ran}(G)$, we first show that the set is contained in a union of subspaces with minimal count. Note that the choice of ReLU activation function is not essential to what follows, and any piecewise linear function with only two pieces (e.g. leaky ReLU) could be substituted. Similar counting arguments have appeared in various theoretical works on expansive neural nets.

Lemma 3.1. *Let G be as in (4) and $T = \mathrm{ran}(G)$. We have $T \subset \bigcup_{i=1}^{N} E_i$, where each E_i is a subspace of dimension at most k and*

$$N \leq \Big[(\frac{2e}{k})^d \Big(\prod_{i=1}^{d} p_i \Big) \Big]^k. \tag{5}$$

Consequently,

$$T - T \subset \bigcup_{i=1}^{\binom{N}{2}} F_i,$$

where each F_i is a subspace of dimension at most $2k$.

Proposition 3.2. *Let G be as in (4) and $T = \mathrm{ran}(G) \subset \mathbb{R}^n$. Then the following holds:*

$$w\big(T \cap \mathbb{S}^{n-1}\big) \leq w\big((T-T) \cap \mathbb{S}^{n-1}\big) \lesssim \sqrt{kd \log\Big(\frac{p'}{k}\Big)}, \tag{6}$$

where $p' = \Big(\prod_{j=1}^{d} p_j \Big)^{1/d}$ is the geometric mean of $p_1, \cdots, p_d$.

3.2 Compressed Sensing with Generative Priors

The effective number of measurements required for any recovery algorithm to be successful is $\mathcal{O}\big(w^2(T')\big) = \mathcal{O}\big(kd \log(p'/k)\big)$. Typically, a GNN is assumed to be expansive, that is $k = p_0 \leq p_1 \leq \cdots \leq p_d = n$. Thus, $k \leq p' \leq n$, which slightly improves $\mathcal{O}(kd \log n)$ in [3]. Moreover, by combining Theorem 2.1 and Proposition 3.2, we present a recovery guarantee for the compressed sensing with generative priors that not only allows for a much broader range of measurement matrices, but also demonstrates the dependence on their parameters.

Corollary 3.2.1. *Consider the settings of Theorem 2.1 for $T = \mathrm{ran}(G)$ as in* (4)*. If $\mathrm{sr}(B) \gg K^2 \log K \cdot kd \log(p'/k)$, then*

$$||\mathbf{x} - \hat{\mathbf{x}}||_2 \lesssim \frac{K\sqrt{kd\log(p'/k)}}{||\mathbf{B}||_F \sqrt{\mathrm{sr}(\mathbf{B})}} ||\mathbf{w}||_2 + \frac{\epsilon}{||\mathbf{B}||_F} + \frac{K\sqrt{l}}{\sqrt{\mathrm{sr}(\mathbf{B})}} \mathrm{dist}\big(\mathbf{x}, \mathrm{ran}(G)\big). \tag{7}$$

Our analysis improves upon the best known results in three ways: Firstly, our result suggests a *denoising* behaviour as the number of measurements (or the stable rank of $\mathbf{B}$) increases, even though the noise is not assumed to be random. Intuitively, one expects a smaller error bound associated with less compression, and our theory highlights this dependence. In contrast, the bound in [3] shows constant dependence on the noise level, the optimization margin, and the model mismatch, regardless of the number of measurements. Secondly, we improve the logarithmic factor in the compression bound, i.e. we need $\mathcal{O}(kd\log(p'/k)) \leq \mathcal{O}(kd\log(n/k))$ effective measurements, rather than $\mathcal{O}(kd \log n)$. This was also available to the authors of [3], had they used a tighter inequality in proof of Lemma 8.3. Finally, we require milder conditions for the measurement matrix, allowing for a fixed *mixing* matrix to create dependence among the rows. This improvement is based on the geometry-preserving properties discussed in [10].

4 Summary

We achieve an estimation bound for the reconstruction of structured signals using noisy and dependent random measurements. The generality of our model enables its application on the non-traditional structure sets such as the range of a ReLU generative neural network. We believe that this is the first general compressed sensing result that specializes well to the generative structure. Whether a similar recovery guarantee is obtainable for GNNs with other activation functions (e.g. sigmoid or tanh), remains an open question.

References

[1] Dennis Amelunxen et al. "Living on the edge: Phase transitions in convex programs with random data". In: *Information and Inference: A Journal of the IMA* 3.3 (2014), pp. 224–294.

[2] Richard Baraniuk et al. "A simple proof of the restricted isometry property for random matrices". In: *Constructive Approximation* 28.3 (2008), pp. 253–263.

[3] Ashish Bora et al. "Compressed sensing using generative models". In: *International Conference on Machine Learning*. PMLR. 2017, pp. 537–546.

[4] Emmanuel J Candes. "The restricted isometry property and its implications for compressed sensing". In: *Comptes rendus mathematique* 346.9-10 (2008), pp. 589–592.

[5] Venkat Chandrasekaran et al. "The convex geometry of linear inverse problems". In: *Foundations of Computational mathematics* 12.6 (2012), pp. 805–849.

[6] Yunzi Ding et al. "The average-case time complexity of certifying the restricted isometry property". In: *arXiv preprint arXiv:2005.11270* (2020).

[7] Simon Foucart and Holger Rauhut. *A Mathematical Introduction to Compressive Sensing*. Springer Science & Business Media, 2013.

[8] Yehoram Gordon. "On Milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$". In: *Geometric aspects of functional analysis*. Springer, 1988, pp. 84–106.

[9] Ajil Jalal et al. "Robust compressed sensing using generative models". In: *Advances in Neural Information Processing Systems* (2020).

[10] Halyun Jeong et al. "Sub-Gaussian matrices on sets: Optimal tail dependence and applications". In: *arXiv preprint arXiv:2001.10631* (2020).

[11] Christopher Liaw et al. "A simple tool for bounding the deviation of random matrices on geometric sets". In: *Geometric aspects of functional analysis*. Springer, 2017, pp. 277–299.

[12] Shahar Mendelson. "Empirical processes with a bounded ψ_1 diameter". In: *Geometric and Functional Analysis* 20.4 (2010), pp. 988–1027.

[13] Shahar Mendelson. "Learning without concentration". In: *Conference on Learning Theory*. PMLR. 2014, pp. 25–39.

[14] Shahar Mendelson, Alain Pajor, and Nicole Tomczak-Jaegermann. "Reconstruction and subgaussian operators in asymptotic geometric analysis". In: *Geometric and Functional Analysis* 17.4 (2007), pp. 1248–1282.

[15] Samet Oymak, Christos Thrampoulidis, and Babak Hassibi. "The squared-error of generalized lasso: A precise analysis". In: *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE. 2013, pp. 1002–1009.

[16] Samet Oymak and Joel A Tropp. "Universality laws for randomized dimension reduction, with applications". In: *Information and Inference: A Journal of the IMA* 7.3 (2018), pp. 337–446.

[17] Yaniv Plan and Roman Vershynin. "Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach". In: *IEEE Transactions on Information Theory* 59.1 (2012), pp. 482–494.

[18] Mark Rudelson and Roman Vershynin. "On sparse reconstruction from Fourier and Gaussian measurements". In: *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences* 61.8 (2008), pp. 1025–1045.

[19] Mihailo Stojnic. "A framework to characterize performance of lasso algorithms". In: *arXiv preprint arXiv:1303.7291* (2013).

[20] Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. "Simple error bounds for regularized noisy linear inverse problems". In: *2014 IEEE International Symposium on Information Theory*. IEEE. 2014, pp. 3007–3011.

[21] Joel A Tropp. "Convex recovery of a structured signal from independent random linear measurements". In: *Sampling Theory, a Renaissance*. Springer, 2015, pp. 67–101.

[22] Dave Van Veen et al. "Compressed sensing with deep image prior and learned regularization". In: *arXiv preprint arXiv:1806.06438* (2018).

[23] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. Vol. 47. Cambridge university press, 2018.

A Appendix

Proof of Theorem 2.1. Since T is closed, there exists a (possibly not unique) point $\mathbf{x}_0 = \operatorname{argmin}_{\mathbf{x}' \in T} ||\mathbf{x} - \mathbf{x}'||_2$. Let $\mathbf{r} := \mathbf{x} - \mathbf{x}_0$ and $\mathbf{h} := \hat{\mathbf{x}} - \mathbf{x}_0 \in T - T$. We have

$$
\begin{aligned}
||\mathbf{BAh} - (\mathbf{BAr} + \mathbf{w})||_2^2 &= ||\mathbf{BA}\hat{\mathbf{x}} - \mathbf{BAx}_0 - \mathbf{BAr} - \mathbf{w}||_2^2 && (8)\\
&= ||\mathbf{BA}\hat{\mathbf{x}} - (\mathbf{BAx} + \mathbf{w})||_2^2 && (9)\\
&= ||\mathbf{y} - \mathbf{BA}\hat{\mathbf{x}}||_2^2 && (10)\\
&\leq ||\mathbf{y} - \mathbf{BAx}_0||_2^2 + \epsilon^2 = ||\mathbf{BAr} + \mathbf{w}||_2^2 + \epsilon^2. && (11)
\end{aligned}
$$

Expanding the LHS and rearranging yields

$$||\mathbf{BAh}||_2^2 \leq 2\langle \mathbf{BAh}, \mathbf{w}\rangle + 2\langle \mathbf{BAh}, \mathbf{BAr}\rangle + \epsilon^2. \tag{12}$$

The LHS is concentrated about $||\mathbf{B}||_F^2 \cdot ||\mathbf{h}||_2^2$. In fact, by Theorem 1.1 of [10] we can write

$$||\mathbf{BAh}||_2^2 \geq ||\mathbf{h}||_2^2 \Big[||\mathbf{B}||_F - CK\sqrt{\log K}||\mathbf{B}||\big(w(T') + \alpha \cdot \mathrm{rad}(T')\big)\Big]^2 \tag{13}$$

with probability at least $1 - 3e^{-\alpha^2}$. Choosing $\alpha = w(T')$ and using the fact that $\mathrm{sr}(\mathbf{B}) \gg K^2 \log K \cdot w^2(T')$, we get

$$||\mathbf{BAh}||_2^2 \gtrsim ||\mathbf{B}||_F^2 \cdot ||\mathbf{h}||_2^2, \tag{14}$$

with probability at least $1 - 3e^{-w^2(T')}$.

To bound the first term on the RHS, define the random process $X_{\mathbf{t}} := \langle \mathbf{BAt}, \mathbf{w}\rangle$, for $\mathbf{t} \in T'$. Recall that $||\cdot||_{\psi_2} = \sup_{||\mathbf{u}||_2 = 1} ||\langle \mathbf{u}, \cdot\rangle||_{\psi_2}$. So, $||X_{\mathbf{t}} - X_{\mathbf{s}}||_{\psi_2} = ||\langle \mathbf{t} - \mathbf{s}, \mathbf{A}^\top \mathbf{B}^\top \mathbf{w}\rangle||_{\psi_2} \leq ||\mathbf{t} - \mathbf{s}||_2 ||\mathbf{A}^\top \mathbf{B}^\top \mathbf{w}||_{\psi_2} \lesssim K||\mathbf{B}^\top \mathbf{w}||_2 ||\mathbf{t} - \mathbf{s}||_2 \leq K||\mathbf{B}|| \cdot ||\mathbf{w}||_2 \cdot ||\mathbf{t} - \mathbf{s}||_2$. Therefore, by Talagrand's comparison inequality (Exercise 8.6.5 of [23]),

$$\sup_{\mathbf{t} \in T'} |X_{\mathbf{t}}| = \sup_{\mathbf{t} \in T'} \langle \mathbf{BAt}, \mathbf{w}\rangle \lesssim K||\mathbf{B}|| \cdot ||\mathbf{w}||_2 \big(w(T') + \beta \cdot \mathrm{rad}(T')\big) \tag{15}$$

with probability at least $1 - 2e^{-\beta^2}$. Again, by setting $\beta = w(T')$ we get

$$\langle \mathbf{BAh}, \mathbf{w}\rangle \leq ||\mathbf{h}||_2 \cdot \sup_{\mathbf{t} \in T'} \langle \mathbf{BAt}, \mathbf{w}\rangle \lesssim ||\mathbf{h}||_2 \cdot K||\mathbf{B}|| \cdot ||\mathbf{w}||_2 w(T') \tag{16}$$

with probability at least $1 - 2e^{-w^2(T')}$.

Now let us bound the second term on the RHS. By Cauchy-Schwarz inequality, $\langle \mathbf{BAh}, \mathbf{BAr}\rangle \leq ||\mathbf{BAh}||_2 ||\mathbf{BAr}||_2$. Theorem 1.1 of [10] implies that $||\mathbf{BAh}||_2 \lesssim ||\mathbf{B}||_F ||\mathbf{h}||_2$ with probability at least $1 - 3e^{-w^2(T')}$, considering that $\mathrm{sr}(\mathbf{B}) \gg K^2 \log K \cdot w^2(T')$. The same theorem can be used on singleton $\{\mathbf{r}\}$ to bound $||\mathbf{BAr}||_2$. Since $w(\{\mathbf{r}\}) = 0$ and $\mathrm{rad}(\{\mathbf{r}\}) = ||\mathbf{r}||$, we get

$$
\begin{aligned}
||\mathbf{BAr}||_2 &\leq ||\mathbf{B}||_F ||\mathbf{r}||_2 + CuK\sqrt{logK}||\mathbf{B}||||\mathbf{r}||_2 && (17)\\
&= ||\mathbf{B}||_F ||\mathbf{r}||_2 \Big(1 + \frac{CuK\sqrt{logK}}{\sqrt{\mathrm{sr}(\mathbf{B})}}\Big), && (18)
\end{aligned}
$$

with probability at least $1 - 3e^{-u^2}$. Choosing $u = w(T')$ and using $\mathrm{sr}(\mathbf{B}) \gg K^2 \log K \cdot w^2(T')$ yields $||\mathbf{BAr}||_2 \lesssim ||\mathbf{B}||_F ||\mathbf{r}||_2$, with probability at least $1 - 3e^{-w^2(T')}$. Thus,

$$\langle \mathbf{BAh}, \mathbf{BAr}\rangle \lesssim ||\mathbf{B}||_F^2 \cdot ||\mathbf{h}||_2 ||\mathbf{r}||_2, \tag{19}$$

with probability at least $1 - 3e^{-w^2(T')} - 3e^{-w^2(T')} \geq 1 - 6e^{-w^2(T')}$.

Combining equations (12)-(19) gives us

$$||\mathbf{B}||_F^2 ||\mathbf{h}||_2^2 - C||\mathbf{h}||_2 \big(K||\mathbf{B}||w(T')||\mathbf{w}||_2 + ||\mathbf{B}||_F^2 ||\mathbf{r}||_2\big) - \epsilon^2 \leq 0, \tag{20}$$

which implies

$$||\mathbf{h}||_2 \leq \frac{C(K||\mathbf{B}||w(T')||\mathbf{w}||_2 + ||\mathbf{B}||_F^2||\mathbf{r}||_2)}{2||\mathbf{B}||_F^2} \tag{21}$$

$$+ \sqrt{\Big(\frac{C(K||\mathbf{B}||w(T')||\mathbf{w}||_2 + ||\mathbf{B}||_F^2||\mathbf{r}||_2)}{2||\mathbf{B}||_F^2}\Big)^2 + \frac{\epsilon^2}{||\mathbf{B}||_F}} \tag{22}$$

$$\leq \frac{C(K||\mathbf{B}||w(T')||\mathbf{w}||_2 + ||\mathbf{B}||_F^2||\mathbf{r}||_2)}{||\mathbf{B}||_F^2} + \frac{\epsilon}{||\mathbf{B}||_F} \tag{23}$$

$$\lesssim \frac{Kw(T')}{||\mathbf{B}||_F\sqrt{\mathrm{sr}(\mathbf{B})}}||\mathbf{w}||_2 + \frac{\epsilon}{||\mathbf{B}||_F} + ||\mathbf{r}||_2, \tag{24}$$

with probability at least $1 - 3e^{-w^2(T')} - 2e^{-w^2(T')} - 6e^{-w^2(T')} = 1 - 11e^{-w^2(T')}$. Finally,

$$||\hat{\mathbf{x}} - \mathbf{x}||_2 \leq ||\hat{\mathbf{x}} - \mathbf{x}_0||_2 + ||\mathbf{x} - \mathbf{x}_0||_2 = ||\mathbf{h}||_2 + ||\mathbf{r}||_2 \tag{25}$$

$$\lesssim \frac{Kw(T')}{||\mathbf{B}||_F\sqrt{\mathrm{sr}(\mathbf{B})}}||\mathbf{w}||_2 + \frac{\epsilon}{||\mathbf{B}||_F} + ||\mathbf{r}||_2, \tag{26}$$

with the aforesaid probability. □

Lemma A.1. *A k-dimensional subspace in $\mathbb{R}^n$ intersects at most $2^k\binom{n}{k}$ different orthants.*

Lemma A.2. *Let $D \subset \mathbb{R}^n$ be a k-dimensional subspace, and Q be an orthant with q number of positive (and $n - q$ negative) coordinates. Then $\sigma(D \cap Q)$ is contained in a subspace of dimension at most $\min(k, q)$.*

Proof of Lemma 3.1. By the rank-nullity theorem, the subspace $\mathrm{im}(\mathbf{A}_1) \subset \mathbb{R}^{p_1}$ is at most k-dimensional. By Lemma A.1, this particular subspace hits at most $2^k\binom{p_1}{k}$ different orthants. Therefore, by Lemma A.2, $\sigma(\mathrm{im}(\mathbf{A}_1))$ is contained in a union of at most $2^k\binom{p_1}{k}$ subspaces of dimension at most k. Each of these subspaces, when multiplied by $\mathbf{A}_2$, is mapped to another subspace of dimension at most k in $\mathbb{R}^{p_2}$. Thus, the linear transformations do not increase the number nor the dimension of the subspaces. Hence, at each layer i, every subspace breaks into $2^k\binom{p_i}{k}$ subspaces, at most. All in all, after d layers, we end up having a union of N subspaces of dimension at most k, where

$$N \leq \prod_{i=1}^{d} 2^k\binom{p_i}{k} \leq \prod_{i=1}^{d}\Big(\frac{2ep_i}{k}\Big)^k = \Big[(\frac{2e}{k})^d\Big(\prod_{i=1}^{d} p_i\Big)\Big]^k. \tag{27}$$

Thus,

$$T = \mathrm{ran}(G) \subset \bigcup_{i=1}^{N} E_i, \tag{28}$$

and

$$T - T \subset \bigcup_{i=1}^{N}\bigcup_{j=1}^{N}(E_i - E_j) = \bigcup_{1\leq i<j\leq N}(E_i + E_j), \tag{29}$$

where $\dim(E_i) \leq k$ and $\dim(E_i + E_j) \leq 2k$. □

Lemma A.3. *Let $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I}_n)$ and $f : \mathbb{R}^n \to \mathbb{R}$ be a Lipschitz function. Then*

$$||f(\mathbf{g}) - \mathbb{E}f(\mathbf{g})||_{\psi_2} \lesssim ||f||_{\mathrm{Lip}}. \tag{30}$$

Lemma A.4. *Let $X_1, X_2, \cdots, X_N$ be sub-gaussian random variables with $K = \max_i ||X_i||_{\psi_2}$. Then*

$$\mathbb{E}\max_i |X_i| \lesssim K\sqrt{\log N}. \tag{31}$$

Lemma A.5. *Let $T_1, \cdots, T_N \subset \mathbb{S}^{n-1} \subset \mathbb{R}^n$. Then*

$$w\big(\bigcup_{i=1}^{N} T_i\big) \lesssim \max_i w(T_i) + \sqrt{\log N}. \tag{32}$$

Proof of Lemma A.5. Let $\mathbf{g}$ be a standard Gaussian vector and define $z_i := \sup_{\mathbf{x}\in T_i} \langle \mathbf{x}, \mathbf{g}\rangle - w(T_i)$. Equivalently, let $f_i(\cdot) := \sup_{\mathbf{x}\in T_i} \langle \mathbf{x}, \cdot\rangle$ and check that $z_i = f_i(\mathbf{g}) - \mathbb{E} f_i(\mathbf{g})$. Lemma A.3 implies $||z_i||_{\psi_2} = ||f_i(\mathbf{g}) - \mathbb{E} f_i(\mathbf{g})||_{\psi_2} \lesssim ||f_i||_{\mathrm{Lip}} = 1$. Then, by Lemma A.4, we have $\mathbb{E}\max_i z_i \lesssim \sqrt{\log N}\max_i ||z_i||_{\psi_2} \lesssim \sqrt{\log N}$. Hence,

$$\begin{aligned}
w\Big(\bigcup_{i=1}^{N} T_i\Big) &= \mathbb{E} \sup_{\mathbf{x}\in \cup T_i} \langle \mathbf{x}, \mathbf{g}\rangle && (33)\\
&= \mathbb{E} \max_i \sup_{\mathbf{x}\in T_i} \langle \mathbf{x}, \mathbf{g}\rangle && (34)\\
&= \mathbb{E} \max_i \big(z_i + w(T_i)\big) && (35)\\
&= \max_i w(T_i) + \mathbb{E}\max_i z_i && (36)\\
&\lesssim \max_i w(T_i) + \sqrt{\log N}. && (37)
\end{aligned}$$

□

Proof of Proposition 3.2. By Lemma 3.1, $T - T \subset \bigcup_{i=1}^{\binom{N}{2}} F_i$, where $\dim(F_i) \le 2k$. Thus,

$$w\big((T-T)\cap \mathbb{S}^{n-1}\big) \le w\big(\bigcup_{i=1}^{\binom{N}{2}} (F_i \cap \mathbb{S}^{n-1})\big). \tag{38}$$

Using Lemma A.5 and Lemma 3.1 respectively, then we have

$$\begin{aligned}
w\big(\bigcup_{i=1}^{\binom{N}{2}} (F_i \cap \mathbb{S}^{n-1})\big) &\lesssim \max_i w(F_i \cap \mathbb{S}^{n-1}) + \sqrt{\log \binom{N}{2}} && (39)\\
&\le \max_i \sqrt{\dim(F_i)} + \sqrt{2\log N} && (40)\\
&\le \sqrt{2k} + \sqrt{2kd\log(\frac{2ep'}{k})} && (41)\\
&\lesssim \sqrt{kd\log(\frac{p'}{k})}. && (42)
\end{aligned}$$

□