A Roadmap

In Appendix C we establish the main consequences of the R2WDC that are then used to prove Theorem 4.4. Then in Appendix D, we prove Lemma 5.2 showing that a Gaussian generative network satisfies the R2WDC with high probability. In Appendix E we analyze the perturbation of the gradient and objective function due to the noise term η, and provide the proof of Lemma 5.3. Extension of the recovery guarantees for Phase Retrieval, Denosing, and Spiked Matrix Recovery are discussed in Appendices F.1, F.2, F.3 respectively. Finally, in Appendix G we give an example of a network with contractive layers, satisfying the assumptions of our main theorems, and in Appendix H we verify the prediction of our theory with synthetic experiments.

B Notation

For any vector x we denote with $\|x\|$ its Euclidean norm and for any matrix A we denote with $\|A\|$ its spectral norm and with $\|A\|_F$ its Frobenius norm. The euclidean inner product between two vectors a and b is $\langle a, b\rangle$. For a set S we will write $|S|$ for its cardinality and S^c for its complement. Let $\mathcal{B}(x, r)$ be the Euclidean ball of radius r centered at x, and $\mathcal{S}^{k-1}$ be the unit sphere in $\mathbb{R}^k$. We will use $a = b + O_1(\delta)$ when $\|a - b\| \leq \delta$, where the norm is understood to be the absolute value for scalars, the Euclidean norm for vectors and the spectral norm for matrices.

C Consequences of the R2WDC

Following [18], we define the function $g : [0, \pi] \to \mathbb{R}$ which describes how the operator $x \mapsto W_{+,x}$ distorts angles:

$$g(\theta) := \cos^{-1}\Big(\frac{(\pi - \theta)\cos\theta + \sin\theta}{\pi}\Big). \tag{18}$$

For two nonzero vectors x, y we let $\bar{\theta}_0 = \angle(x, y)$ and define inductively $\bar{\theta}_i := g(\bar{\theta}_{i-1})$. Then we set

$$\tilde{h}_{x,y} := \frac{1}{2^d}\left[\Big(\prod_{i=0}^{d-1}\frac{\pi - \bar{\theta}_i}{\pi}\Big)y + \sum_{i=1}^{d-1}\frac{\sin\bar{\theta}_i}{\pi}\Big(\prod_{j=i+1}^{d-1}\frac{\pi - \bar{\theta}_j}{\pi}\Big)\|y\|\hat{x}\right]. \tag{19}$$

Proposition C.1. *Fix $\epsilon > 0$ such that $\max(2d\epsilon, 10\epsilon) < 1$. Let G be a generative network as in (2) satisfying the R2WDC with constant ϵ. Then for any $x \in \mathbb{R}^k$ and $j \in [d]$*

$$\|x\|^2\Big(\frac{1}{2} - \epsilon\Big)^j \leq \|G_j(x)\|^2, \leq \Big(\frac{1}{2} + \epsilon\Big)^j\|x\|^2 \tag{20a}$$

$$\|G(x)\|^2 \leq \frac{1 + 4\epsilon d}{2^d}\|x\|^2. \tag{20b}$$

Moreover, for any $x \neq 0$, $y \neq 0$, $j \in [d]$, the angle $\theta_j = \angle(G_j(x), G_j(y))$ is well-defined and

$$|\theta_j - g(\theta_{j-1})| \leq 4\sqrt{\epsilon} \tag{21a}$$

$$\langle G(x), G(y)\rangle \geq \frac{1}{4\pi}\frac{1}{2^d}\|x\|\|y\| \tag{21b}$$

$$|\langle G(x), G(y)\rangle - \langle x, \tilde{h}_{x,y}\rangle| \leq 24\frac{d^3\sqrt{\epsilon}}{2^d}\|x\|\|y\| \tag{21c}$$

where g is given in (18) and $\tilde{h}$ in (19).

The next result is used to prove concentration of the gradient of the objective function around its expectation.

Proposition C.2. *Fix $0 < \epsilon < d^{-4}/(16\pi)^2$ and $d \geq 2$. Suppose that G as in (2) satisfies the R2WDC with constant ϵ. Let $x \in \mathbb{R}^k$ be a point where $G(x)$ is differentiable, and $y \in \mathbb{R}^k \setminus \{0\}$, then*

$$\|\Lambda_{d,x}\|^2 \leq \frac{1 + 4\epsilon d}{2^d} \leq \frac{13}{12}\frac{1}{2^d} \tag{22}$$

$$\|\Lambda_{d,x}^t\Lambda_{d,x} - \frac{1}{2^d}I_k\| \leq \frac{4\epsilon d}{2^d} \tag{23}$$

$$\|\Lambda_{d,x}^t\Lambda_{d,y}y - \tilde{h}_{x,y}\| \leq 24\frac{d^3\sqrt{\epsilon}}{2^d}\|y\| \tag{24}$$

The next proposition uses the R2WDC to bound the local Lipschitz constant of the ReLU-networks $\{G_j\}_{j\in[d]}$.

Proposition C.3. *Suppose that $x \in \mathcal{B}(x, d\sqrt{\epsilon}\|y\|)$ and G satisfies the R2WDC with $\epsilon < 1/(200)^4/d^6$. Then for every $i \in [d]$, it holds that*

$$\|G_i(x) - G_i(y)\| \leq \frac{1.2}{2^{i/2}}\|x - y\| \tag{25}$$

The next proposition is used to show that when x is close to y, the gradient of the objective function points in a direction that decreases the distance between of x and y.

Proposition C.4. *Suppose $x \in \mathcal{B}(y, d\sqrt{\epsilon}\|y\|)$ is a differentiable point for G, and the R2WDC holds with $\epsilon < 1/(200)^4/d^6$. Then it holds that*

$$\Lambda_x^T(\Lambda_x x - \Lambda_y y) = \frac{1}{2^d}(x - y) + \frac{1}{2^d}\frac{1}{16}\|x - y\|O_1(1) \tag{26}$$

We can now prove Theorem 4.4.

Proof of Theorem 4.4. The proof of Theorem 3.1 in [21] only uses the inequality (20a)-(26), which are proved for a network satisfying the WDC. The previous propositions have shown that such inequalities hold under the weaker R2WDC. Therefore from the proof of Theorem 3.1 in [21] combined with the Propositions C.1-C.4, we obtain automatically the proof of Theorem 4.4. □

C.1 Supplemental Results for Section C

Proof of Proposition C.1

Proof. For $x, y \in \mathbb{R}^k$ and $j \in [d]$, below we write $x_j := G_j(x)$ and $y_j := G_j(y)$.

- *Proof of* (20a)
Notice that by (9) for $x \in \mathbb{R}^k$

$$\Big(\frac{1}{2} - \epsilon\Big)\|x_{j-1}\|^2 \leq \|x_j\|^2 \leq \Big(\frac{1}{2} + \epsilon\Big)\|x_{j-1}\|^2,$$

which proceeding by induction gives (20a).

- *Proof of* (20b)
Next observe that since $\log(1 + z) \leq z$, $e^z \leq 1 + 2z$ for $z < 1$ and $2d\epsilon \leq 1$, from (20a) we have

$$\|G_d(x)\|^2 \leq \frac{(1 + 2\epsilon)^d}{2^d}\|x\|^2 \leq \frac{1}{2^d}e^{d\log(1+2\epsilon)}\|x\| \leq \frac{1 + 4\epsilon d}{2^d}\|x\|,$$

which corresponds to (20b).

- *Proof of* (21a)
Assume that $x, y \in \mathbb{R}^k \setminus \{0\}$. Then, the assumption $2d\epsilon \leq 1$ and the lower bound in (20a) imply that θ_j are well-defined for all $j \in [d]$. To prove then (21a) notice that it is sufficient to prove that for any $j \in [d]$ it holds that

$$\Big|\cos\theta_j - \frac{(\pi - \theta_{j-1})\cos\theta_{j-1} + \sin\theta_{j-1}}{\pi}\Big| \leq 5\epsilon$$

By homogeneity of the ReLU activation function, we can assume without loss of generality that $\|x_{j-1}\| = \|y_{j-1}\| = 1$. Let

$$\begin{aligned}
\delta_1 &:= \langle x_{j-1}, \big(W_{j,+,x}^T W_{j,+,y} - Q_{x_{j-1},y_{j-1}}\big)y_{j-1}\rangle \\
\delta_2 &:= \langle x_{j-1}, \big(W_{j,+,x}^T W_{j,+,x} - I_k/2\big)y_{j-1}\rangle \\
\delta_3 &:= \langle y_{j-1}, \big(W_{j,+,y}^T W_{j,+,y} - I_k/2\big)y_{j-1}\rangle
\end{aligned}$$

and notice that by the R2WDC we have $\max(|\delta_1|, |\delta_2|, |\delta_3|) \leq \epsilon$. Thus,

$$\begin{aligned}
\cos\theta_j &= \frac{\langle x_j, y_j\rangle}{\|x_j\|\|y_j\|} \\
&= \frac{\langle x_{j-1}, W_{j,+,x}^T W_{j,+,y} y_{j-1}\rangle}{\sqrt{\langle x_{j-1}, W_{j,+,x}^T W_{j,+,x} x_{j-1}\rangle \langle y_{j-1}, W_{j,+,y}^T W_{j,+,y} y_{j-1}\rangle}} \\
&= 2\frac{\langle x_{j-1}, Q_{x_{j-1},y_{j-1}} y_{j-1}\rangle + \delta_1}{\sqrt{(1+2\delta_2)(1+2\delta_3)}}.
\end{aligned}$$

Finally, notice that $2\langle x_{j-1}, Q_{x_{j-1},y_{j-1}} y_{j-1}\rangle = [(\pi - \theta_{j-1})\cos\theta_{j-1} + \sin\theta_{j-1}]/\pi$ so

$$\begin{aligned}
|\cos\theta_j - 2\langle x_{j-1}, Q_{x_{j-1},y_{j-1}} y_{j-1}\rangle| &\leq 2|\langle x_{j-1}, Q_{x_{j-1},y_{j-1}} y_{j-1}\rangle| \left|1 - \frac{1}{\sqrt{(1+2\delta_2)(1+2\delta_3)}}\right| \\
&\quad + \frac{2|\delta_1|}{\sqrt{(1+2\delta_2)(1+2\delta_3)}} \\
&\leq \left|1 - \frac{1}{(1-2\epsilon)}\right| + \frac{2\epsilon}{(1-2\epsilon)} \\
&\leq 5\epsilon
\end{aligned}$$

where the second inequality follows from $|2\langle x_{j-1}, Q_{x_{j-1},y_{j-1}} y_{j-1}\rangle| \leq 1$ and $\max(|\delta_1|, |\delta_2|, |\delta_3|) \leq \epsilon$, and the third inequality from $10\epsilon < 1$.

- *Proof of* (21b)

By (20a) and $\epsilon \leq 1/2$, it follows that $\|x_d\|\|y_d\| \geq \frac{(1-2\epsilon)^d}{2^d}\|x\|\|y\|\frac{1-2d\epsilon}{2^d}$. Moreover, let $\delta := 4\sqrt{\epsilon}$, then by (21a) we have that $\theta_j = g(\theta_{j-1}) + O_1(\delta)$. Thus, $\theta_d = g(g(\cdots g(g(\theta_0) + O_1(\delta)) + O_1(\delta)\cdots) + O_1(\delta)))) + O_1(\delta)$ and for $\bar\theta = g^{\circ d}(\theta_0)$, so that, using $g'(\theta) \leq 1$ for all θ, we have

$$|\theta_d - \bar\theta_d| \leq d\delta. \tag{27}$$

Then by (27), $\bar\theta_d \leq \cos^{-1}(1/\pi)$ for $d \geq 2$, and $16\pi d\sqrt{\epsilon} < 1$, follows that $\cos\theta_d \geq 3/(4\pi)$.

Finally, if $2d\epsilon \leq 2/3$, we can then conclude that

$$\langle G(x), G(y)\rangle \geq \cos(\theta_d)\|x_d\|\|y_d\| \geq \frac{1}{4\pi}\frac{1}{2^d}\|x\|\|y\|.$$

- *Proof of* (21c)

The following result on a recurrence relation will be used in the subsequent analysis

$$\Gamma_d = s_d\Gamma_{d-1} + r_d, \quad \Gamma_0 = y \implies \Gamma_d = \Big(\prod_{i=1}^d s_i\Big)y + \sum_{i=1}^d \Big(r_i \prod_{j=i+1}^d s_j\Big) \tag{28}$$

Define $\Gamma_d := \langle x_d, y_d\rangle$, then

$$\begin{aligned}
\Gamma_d &= \langle x_{d-1}, W_{d-1,+,x}^T W_{d-1,+,x} y_{d-1}\rangle \\
&= \langle x_{d-1}, Q_{x_{d-1},y_{d-1}} y_{d-1}\rangle + O_1(\epsilon)\|y_{d-1}\|\|x_{d-1}\|, \\
&= \frac{\pi - \theta_{d-1}}{2\pi}\Gamma_{d-1} + \frac{\sin\theta_{d-1}}{2\pi}\|x_{d-1}\|\|y_{d-1}\| + O_1(\epsilon)\|y_{d-1}\|\|x_{d-1}\|, \\
&= \frac{\pi - \theta_{d-1}}{2\pi}\Gamma_{d-1} + \frac{\sin\theta_{d-1}}{2\pi}\frac{\|x\|\|y\|}{2^{d-1}} + \frac{\epsilon}{2^d}\Big(\frac{4\epsilon d}{\pi} + 2(1+4\epsilon d)\Big)\|y\|\|x\|O_1(1), \\
&= \frac{\pi - \theta_{d-1}}{2\pi}\Gamma_{d-1} + \frac{\sin\theta_{d-1}}{2\pi}\frac{\|x\|\|y\|}{2^{d-1}} + 11d\epsilon\frac{\|y\|\|x\|}{2^d}O_1(1),
\end{aligned}$$

Where the second equality follows from the R2WDC , the third from the definition of $Q_{p,q}$. The rest of the proof proceeds as in the proof of Lemma 8 in [18]. □

Proof of Proposition C.2

Proof.

- *Proof of* (23).
Let $x \in \mathbb{R}^k$ be a point where G is differentiable, and notice that for small enough z, by local linearity of G, we have $G(x+z) = \Lambda_x z$. Then the R2WDC gives for $j \in [d]$

$$\left|\langle (W_{j,+,x}^T W_{j,+,y} - I_k/2)\Lambda_{j-1,x} z, \Lambda_{j-1,x} z\rangle\right| \leq \epsilon \|\Lambda_{j-1,x}\|^2 \|z\|^2$$

for all z, which in turn implies

$$\|\Lambda_{j,x}^T \Lambda_{j,x} - \frac{1}{2}\Lambda_{j-1,x}^T \Lambda_{j-1,x}\| \leq \epsilon \|\Lambda_{j-1,x}^T \Lambda_{j-1,x}\|. \tag{30}$$

Let now $M_d := \Lambda_{d,x}^T \Lambda_{d,x}$ with $M_0 = I_k$, then

$$M_d = \frac{1}{2} M_{d-1} + \|M_{d-1}\| O_1(\epsilon). \tag{31}$$

We then obtain

$$\|M_d\| \leq \Big(\frac{1}{2} + \epsilon\Big)\|M_{d-1}\| \leq \frac{(1+2\epsilon)^d}{2^d}\|M_0\| \leq \frac{1+4\epsilon d}{2^d},$$

where the second inequality, and the third inequality uses $2d\epsilon \leq 1$ and the same reasoning as in the proof of (20b). From (22) and (31) we obtain the following recurrence relation

$$M_d = \frac{1}{2} M_{d-1} + O_1\Big(\epsilon \frac{1+4\epsilon(d-1)}{2^{d-1}}\Big),$$

which, using (28) and $4\epsilon d \leq 1$, gives

$$\begin{aligned} M_d &= \frac{1}{2} I_k + \sum_{i=1}^{d} O_1\Big(\epsilon \frac{1+4\epsilon(i-1)}{2^{i-1}}\Big)\frac{1}{2^{d-i}} \\ &= \frac{1}{2} I_k + \frac{4\epsilon d}{2^d} O_1(1) \end{aligned}$$

- *Proof of* (24).
Notice again that if $x \in \mathbb{R}^k$ is a differentiable point for G, the R2WDC gives for any $j \in [d]$

$$\|\Lambda_{j,x}^T \Lambda_{j,y} y - \Lambda_{j-1,x}^T Q_{x_{j-1},y_{j-1}} \Lambda_{j-1,y} y\| \leq \epsilon \|\Lambda_{j-1,x}\| \|G_{j-1}(y)\|. \tag{32}$$

We then let $\Gamma_d := \Lambda_{d,x}^T \Lambda_{d,y} y$ and observe that

$$\begin{aligned} \Gamma_d &= \Lambda_{d-1,x}^T Q_{x_{d-1},y_{d-1}} \Lambda_{d-1,y} y + \|\Lambda_{d-1,x}\| \|G_{d-1}(y)\| O_1(\epsilon) \\ &= \frac{\pi - \theta_{d-1}}{2\pi}\Gamma_{d-1} + \frac{\sin\theta_{d-1}}{2\pi}\frac{\|y_{d-1}\|}{\|x_{d-1}\|}\Lambda_{d-1,x}^T \Lambda_{d-1,x} x + \epsilon\Big(\frac{1+4\epsilon d}{2^{d-1}}\Big)\|y\| \end{aligned}$$

where the first equality is from (32), and the second uses the definition of $Q_{x,y}$, (20b) and (22). The rest of the proof follows as in the proof of Equation (7) in Lemma 8 in [18]. □

Proof of Proposition C.3

Lemma C.5. *Suppose G satisfies the R2WDC with constant ϵ. Then for any $x, y \in \mathbb{R}^k \setminus \{0\}$ and $i \in [d]$, it holds that*

$$\|G_i(x) - G_i(y)\| \leq \left(\sqrt{\frac{1}{2} + \epsilon} + \sqrt{2(2\epsilon + \theta_{i-1})}\right)\|G_{i-1}(x) - G_{j-1}(y)\|$$

where $\theta_{i-1} = \angle(G_i(x), G_i(y))$.

Proof of Lemma C.5. We have

$$\|G_j(x) - G_j(y)\| \leq \|(W_j)_{+,x_{j-1}}(x_{j-1} - y_{j-1})\| + \|\big(W_{j,+,x} - W_{j,+,y}\big)y_{j-1}\|. \tag{33}$$

We begin analyzing the first term, noticing that by the R2WDC

$$\begin{aligned}\|W_{j,+,x}(x_{j-1} - y_{j-1})\|^2 &= (x_{j-1} - y_{j-1})^T(W_{j,+,x}^T W_{j,+,x} - \frac{1}{2}I_{n_1})(x_{j-1} - y_{j-1}) + \frac{1}{2}\|x_{j-1} - y_{j-1}\|^2 \\ &\leq \Big(\frac{1}{2} + \epsilon\Big)\|x_{j-1} - y_{j-1}\|^2 \end{aligned} \tag{34}$$

We next analyze the second term. Let $W_{j,i} \in \mathbb{R}^{1\times n_{j-1}}$ be the i-th row of W_j then

$$\begin{aligned}\|\big(W_{j,+,x} - W_{j,+,y}\big)y_{j-1}\|^2 &= \sum_{i=1}^{n}\big(\mathbb{1}_{W_{j,i}x_{j-1}>0} - \mathbb{1}_{W_{j,i}y_{j-1}>0}\big)^2(W_{j,i}\,y_{j-1})^2 \\ &\leq \sum_{i=1}^{n}\big(\mathbb{1}_{W_{j,i}x>0} - \mathbb{1}_{W_{j,i}y>0}\big)(W_{j,i}(x_{j-1} - y_{j-1}) \\ &= \sum_{i=1}^{n}\mathbb{1}_{W_{j,i}x>0}\mathbb{1}_{W_{j,i}y\leq 0}W_{j,i}(x_{j-1} - y_{j-1}) \\ &\quad + \sum_{i=1}^{n}\mathbb{1}_{W_{j,i}x\leq 0}\mathbb{1}_{W_{j,i}y>0}W_{j,i}(x_{j-1} - y_{j-1}) \\ &= (x_{j-1} - y_{j-1})^T(W_j)_{+,x_{j-1}}^T\Big((W_j)_{+,x_{j-1}} - (W_j)_{+,y_{j-1}}\Big)(x_{j-1} - y_{j-1}) \\ &\quad + (x_{j-1} - y_{j-1})^T(W_j)_{+,y_{j-1}}^T\Big((W_j)_{+,y_{j-1}} - (W_j)_{+,x_{j-1}}\Big)(x_{j-1} - y_{j-1}).\end{aligned} \tag{35}$$

Observe now that by the R2WDC we have

$$\begin{aligned}&|(x_{j-1} - y_{j-1})^T(W_j)_{+,y_{j-1}}^T\Big((W_j)_{+,y_{j-1}} - (W_j)_{+,x_{j-1}}\Big)(x_{j-1} - y_{j-1})| \\ &\quad\leq |(x_{j-1} - y_{j-1})^T\Big((W_j)_{+,y_{j-1}}^T(W_j)_{+,y_{j-1}} - I_k/2\Big)(x_{j-1} - y_{j-1})| \\ &\quad + |(x_{j-1} - y_{j-1})^T\Big((W_j)_{+,y_{j-1}}^T(W_j)_{+,x_{j-1}} - Q_{x_{j-1},y_{j-1}}\Big)(x_{j-1} - y_{j-1})| \\ &\quad + |(x_{j-1} - y_{j-1})^T\Big(I_{n_{i-1}}/2 - Q_{x_{j-1},y_{j-1}}\Big)(x_{j-1} - y_{j-1})| \\ &\quad\leq (2\epsilon + \theta_{j-1})\|x_{j-1} - y_{j-1}\|^2,\end{aligned}$$

which together with (35) gives

$$\|\big(W_{j,+,x} - W_{j,+,y}\big)y_{j-1}\|^2 \leq 2(2\epsilon + \theta_{j-1})\|x_{j-1} - y_{j-1}\|^2. \tag{36}$$

We conclude using (34) and (36) in (33). □

Proof of Proposition C.4

We next prove the convexity-like property in Proposition C.4.

Proof of Proposition C.4. We begin observing that by (27) we have $|\theta_i - \bar{\theta}_i| \leq 4i\sqrt{\epsilon} \leq 4d\sqrt{\epsilon}$. Furthermore, since $x \in \mathcal{B}(y, d\sqrt{\epsilon}\|y\|)$ it follows that

$$\bar{\theta}_i \leq \bar{\theta}_0 \leq 2d\sqrt{\epsilon}.$$

Thus by the assumption on ϵ, we have

$$\sqrt{2}\sqrt{\theta_i + 2\epsilon} \leq \sqrt{2}\sqrt{\bar{\theta}_i + 4d\sqrt{\epsilon} + 2\epsilon} \leq \sqrt{2}\sqrt{2d\sqrt{\epsilon} + 4d\sqrt{\epsilon} + 2\epsilon} \leq \frac{1}{30\sqrt{2}d} \tag{37}$$

Let now $\Gamma_d := \Lambda_{d,x}^T(\Lambda_{d,x}x - \Lambda_{d,y}y)$. Then notice that

$$\begin{aligned}
\Gamma_d &= \Lambda_{d-1,x}^T W_{d,+,x}^T(W_{d,+,x}\Lambda_{d-1,x}x - W_{d,+,y}\Lambda_{d-1,y}y) \\
&= \Lambda_{d-1,x}^T W_{d,+,x}^T W_{d,+,x}(\Lambda_{d-1,x}x - \Lambda_{d-1,y}y) + \Lambda_{d,x}^T(W_{d,+,x} - W_{d,+,y})\Lambda_{d-1,y}y \\
&= \frac{1}{2}\Gamma_{d-1} + \epsilon\|\Lambda_{d-1,x}\|\|\Lambda_{d-1,x}x - \Lambda_{d-1,y}y\|O_1(1) + \|\Lambda_{d,x}\|\|(W_{d,+,x} - W_{d,+,y})\Lambda_{d-1,y}y\|O_1(1) \\
&= \frac{1}{2}\Gamma_{d-1} + \Big(\epsilon + \sqrt{\frac{1}{2} + \epsilon}\sqrt{2(2\epsilon + \theta_{d-1})}\Big)\|\Lambda_{d-1}\|\|\Lambda_{d-1,x}x - \Lambda_{d-1,y}y\|O_1(1) \\
&= \frac{1}{2}\Gamma_{d-1} + \Big(\epsilon + \sqrt{\frac{1}{2} + \epsilon}\sqrt{2(2\epsilon + \theta_{d-1})}\Big)\frac{1.2\sqrt{1+4\epsilon d}}{2^{d-1}}\|x-y\|O_1(1) \\
&= \frac{1}{2}\Gamma_{d-1} + 2\Big(\frac{1}{200^4 d^6} + \frac{1}{30\sqrt{2}d}\Big)\frac{\|x-y\|}{2^{d-1}}O_1(1) \qquad (38)
\end{aligned}$$

where the third equality follows from the R2WDC, the fourth the R2WDC and (36), the fifth from (22) and Proposition C.3, and sixth from (37) and the assumption on ϵ. Finally, from (38) and (28) we obtain

$$\Gamma_d = \frac{1}{2^d}\|x-y\| + \frac{1}{16}\frac{\|x-y\|}{2^d}O_1(1)$$

□

D Proof of Lemma 5.2

In this section, we prove that a generative network G with random weights satisfies the R2WDC with high-probability (Lemma 5.2). Our proof is inspired by the proof of Proposition 3 in [17].

Notice that because of the piecewise-linear nature of the ReLU activation function, the output of a ReLU network is a subset of a union of affine subspaces. The following lemma from [22] provides an upper bound on the number of such subspaces.

Lemma D.1 (Lemma 7 in [22]). *Consider a generative network G as in (2) and assume that $n_i \geq k$ for $i \in [d]$. Then for $i \in [d]$, range(G_i) is contained in a union of affine subspaces. Precisely,*

$$\textit{range}(G_i) \subseteq \cup_{j\in[\Psi_i]} S_{i,j} \quad \textit{where} \quad \Psi_i \leq \prod_{j=1}^{i}\Big(\frac{en_j}{k}\Big)^k.$$

Here each $S_{i,j}$ is some k-dimensional affine subspace (which depends on $\{W_\ell\}_{\ell\in[i]}$) in $\mathbb{R}^{n_i}$.

We next give the main result upon which the proof of Proposition 5.2 rests.

Proposition D.2. *Fix $0 < \epsilon < 1$ and $\ell < n$. Let $W \in \mathbb{R}^{m\times n}$ have i.i.d. $\mathcal{N}(0, 1/m)$. Let R, S be ℓ-dimensional subspaces of $\mathbb{R}^n$, and T be an ℓ'-dimensional subspaces of $\mathbb{R}^n$ with $l' \geq l$. Then if $m \geq C_\epsilon \ell'$, we have that*

$$|\langle W_{+,r}^T W_{+,s}u, v\rangle - \langle Q_{r,s}u, v\rangle| \leq \epsilon\|u\|_2\|v\|_2 \quad \forall\, u, v \in T, \forall r \in R, \forall s \in S, \qquad (39)$$

with probability exceeding

$$1 - \gamma\Big(\frac{e\,m}{\ell}\Big)^{2\ell}\exp(-c_\epsilon m)$$

Furthermore, let $U = \bigcup_{i=1}^{N_1} U_i$, $V = \bigcup_{j=1}^{N_2} V_j$ $V = \bigcup_{j=1}^{N_2} V_j$, $R = \bigcup_{p=1}^{N_3} R_p$, and $S = \bigcup_{q=1}^{N_4} S_q$ be union of subspaces of $\mathbb{R}^n$ of dimension at most ℓ. Then if $m \geq 2C_\epsilon\ell'$

$$|\langle W_{+,r}^T W_{+,s}u, v\rangle - \langle Q_{r,s}u, v\rangle| \leq \epsilon\|u\|_2\|v\|_2 \quad \forall\, u \in U, v \in V, \forall r \in R, \forall s \in S, \qquad (40)$$

with probability exceeding

$$1 - \gamma N_1 N_2 N_3 N_4\Big(\frac{e\,m}{\ell}\Big)^{2\ell}\exp(-c_\epsilon m).$$

Here c_ϵ depends polynomially on ϵ, C_ϵ depends polynomially on ϵ^{-1}, and γ is a positive universal constant.

With the above two results, we are in a position to prove Lemma 5.2.

Proof of Lemma 5.2. We begin establishing the proposition in the $d = 2$ case.

If $n_1 \geq 2C_\epsilon k$ by the second part of Proposition D.2 with $U, V, R, S = \mathbb{R}^k$, W_1 satisfies (9) with probability at least

$$1 - \gamma\Big(\frac{en_1}{k}\Big)^{2k} \exp(-c_\epsilon n_1).$$

We next consider the bound (9) for $j = 2$. Fix W_1 and observe that, by Lemma D.1, range(G_1) is contained in the union of at most Ψ_1 number of k-dimensional affine subspaces of $\mathbb{R}^{n_1}$ and $\{G_1(x_1) - G_1(x_2) : x_1, x_2 \in \mathbb{R}^k\}$ is contained in the union of at most Ψ_1^2 number of $2k$-dimensional affine subspaces of $\mathbb{R}^{n_1}$. Since then an ℓ-dimensional affine subspace is also contained in an $\ell + 1$ subspace. We have that range$(G_1) \subset \mathcal{R}_1$ where $\mathcal{R}_1$ is the union of at most Ψ_1 number of $k+1$-dimensional subspaces and $\{G_1(x_1) - G_1(x_2) : x_1, x_2 \in \mathbb{R}^k\} \subset \mathcal{U}_1$ where $\mathcal{U}_1$ is the union of at most Ψ_1^2 number of $2k+1$-dimensional subspaces.

By applying the second part of Proposition D.2 to the sets $\mathcal{U}_1, \mathcal{U}_1, \mathcal{R}_1$ and $\mathcal{R}_1$, we have that for fixed W_1,

$$\begin{aligned}\Big|\langle\Big((W_2)^T_{+,G_1(x)}(W_2)_{+,G_1(y)} - Q_{G_1(x),G_1(y)}\Big)\big(G_1(x_1) - G_1(x_2)\big), G_1(x_3) - G_1(x_4)\rangle\Big| \\ \leq \epsilon\|G_1(x_1) - G_1(x_2)\|_2\|G_1(x_3) - G_1(x_4)\|_2 \quad (41)\end{aligned}$$

with probability at least

$$1 - \gamma\Psi_1^6\Big(\frac{e\,n_2}{k+1}\Big)^{2k+2} e^{-c_\epsilon n_2} \geq 1 - \gamma\Big(\frac{e\,n_2}{k+1}\Big)^{4k} e^{-c_\epsilon n_2/2}$$

provided that $n_2 \geq 12c_\epsilon^{-1} \log \Psi_1$ and $n_2 \geq 2C_\epsilon(2k+1)$. In particular the above holds provided that $n_2 \geq \widetilde{C}_\epsilon k \log(en_1/k)$ where $\widetilde{C}_\epsilon$ depends polynomially on ϵ^{-1}.

Integrating over the probability space of W_1, independence of W_2 and W_1 implies that (41) holds for random W_1 with the same probability bound. This allows us to conclude that a two-layer random generative network G satisfies the R2WDC with probability at least

$$1 - \gamma\Big(\frac{en_1}{k}\Big)^{2k} e^{-c_\epsilon n_1} - \gamma\Big(\frac{e\,n_2}{k+1}\Big)^{4k} e^{-c_\epsilon n_2/2}.$$

The proof of the $d \geq 2$ case follows similarly. In particular, to establish (9) for W_i notice that range(G_{i-1}) is contained in the union of at most Ψ_{i-1} number $k+1$ subspaces, and $\{G_{i-1}(x_1) - G_{i-1}(x_2) : x_1, x_2 \in \mathbb{R}^k\}$ in the union of at most Ψ_{i-1}^2 number of $2k+1$-dimensional subspaces. Applying Proposition D.2 to these subspaces we have that for fixed $\{W_j\}_{j\in[i-1]}$

$$\begin{aligned}\Big|\langle\Big((W_i)^T_{+,G_{i-1}(x)}(W_i)_{+,G_{i-1}(y)} - Q_{G_{i-1}(x),G_{i-1}(y)}\Big)\big(G_{i-1}(x_1) - G_{i-1}(x_2)\big), G_{i-1}(x_3) - G_{i-1}(x_4)\rangle\Big| \\ \leq \epsilon\|G_{i-1}(x_1) - G_{i-1}(x_2)\|_2\|G_{i-1}(x_3) - G_{i-1}(x_4)\|_2 \quad (42)\end{aligned}$$

with probability at least

$$1 - \gamma\Big(\frac{e\,n_i}{k+1}\Big)^{4k} e^{-c_\epsilon n_i/2}$$

provided that

$$n_i \geq \widetilde{C}_\epsilon \cdot k \cdot \prod_{j=1}^{i-1} \frac{e\,n_j}{k}.$$

Integrating over the probability space of $\{W_j\}_{j\in[i-1]}$) indpendence of W_i and $(W_1, \ldots, W_{i-1})$ gives that (42) holds with the same probability bound. □

We will devote the following section to the proof of Proposition D.2.

D.1 Proof of Proposition D.2

We begin by proving a weaker form of Proposition D.2, that characterizes the concentration of $W_{+,r}^T W_{+,s}$ around its mean for fixed r, s and when acting on ℓ-dimensional subspaces.

Lemma D.3. *Fix $0 < \epsilon < 1$ and $k < n$. Let $W \in \mathbb{R}^{m\times n}$ have i.i.d. $\mathcal{N}(0, 1/m)$ entries and fix $r, s \in \mathbb{R}^n$. Let T be a ℓ-dimensional subspace of $\mathbb{R}^n$. Then if $m \geq \tilde{C}_1 \ell$, we have that with probability exceeding $1 - 2\exp(-\tilde{c}_1\, m)$,*

$$|\langle W_{+,r}^T W_{+,s} u, u\rangle - \langle Q_{r,s} u, u\rangle| \leq \epsilon \|u\|_2^2 \quad \forall\, u \in T \tag{43}$$

and

$$|\langle W_{+,r}^T W_{+,s} u, v\rangle - \langle Q_{r,s} u, v\rangle| \leq 3\epsilon \|u\|_2 \|v\|_2 \quad \forall\, u, v \in T, \tag{44}$$

Furthermore, let $U = \bigcup_{i=1}^{N_1} U_i$ and $V = \bigcup_{j=1}^{N_2} V_j$ where U_i and V_j are subspaces of $\mathbb{R}^n$ of dimension at most ℓ for all $i \in [N_1]$ and $j \in [N_2]$. Then if $m \geq 2\tilde{C}_1 \ell$

$$|\langle W_{+,r}^T W_{+,s} u, v\rangle - \langle Q_{r,s} u, v\rangle| \leq 3\epsilon \|u\|_2 \|v\|_2 \quad \forall\, u \in U, \forall\, v \in V, \tag{45}$$

with probability exceeding $1 - 2N_1 N_2 \exp(-\tilde{c}_1 m)$. Here $\tilde{c}_1$ depends polynomially on ϵ and $\tilde{C}_1 = \Omega(\epsilon^{-1} \log \epsilon^{-1})$.

Proof. The proof follows the one in Proposition 4 of [17] with minor variations. Set $\Sigma_{r,s} := W_{+,r}^T W_{+,s} - Q_{r,s}$, and notice that for fixed $u \in \mathbb{R}^{n-1}$, $\langle \Sigma_{r,s} u, u\rangle = \sum_{i=1}^m Y_i$ where $Y_i = X_i - \mathbb{E}[X_i]$, $X_i = \mathbb{1}_{\langle w_i, r\rangle > 0} \mathbb{1}_{\langle w_i, s\rangle > 0} \langle w_i, u\rangle^2$ and each $w_i \sim \mathcal{N}(0, I_n/m)$. We then notice that the Y_i are sub-exponential random variables and by standard ϵ-net argument we can show that (43) holds with high-probability. Proposition 5 in [17] can then be adapted to this case as well and used to derive (44) from (43). Finally (45) follows by a union bound over all subspaces of the form $\text{span}(U_i, V_j)$. □

We next observe that the rows of a sufficiently tall random matrix W tessellate the unit sphere in regions of small diameter.

Lemma D.4. *Fix $0 < \epsilon < 1$. Let $W \in \mathbb{R}^{m\times n}$ have i.i.d. $\mathcal{N}(0, 1/m)$ entries with rows $\{w_\ell\}_{\ell=1}^m$. Let Z be a ℓ-dimensional subspace of $\mathbb{R}^n$. Define $E_{Z,W}$ to be the event that there exists a set $Z_0 \subset Z$ with the following properties:*

i) each $z_0 \in Z_0$ satisfies $\langle w_\ell, z_0\rangle \neq 0$ for all $\ell \in [m]$,

ii) $|Z_0| \leq (\frac{e\, m}{\ell})^\ell$, and

iii) for all $z \in Z$ such that $\|z\|_2 = 1$, there exists $z_0 \in Z_0$ such that $\|z - z_0\|_2 \leq \epsilon$.

If $m \geq \tilde{C}_2 \ell$, then $\mathbb{P}(E_{Z,W}) \geq 1 - C_2 \exp(-c_2 \epsilon m)$. Here C_2 and c_2 are positive absolute constants and $\tilde{C}_2$ depends polynomially on ϵ^{-1}.

Proof. The proof of this lemma follows the one in Lemma 24 in [17]. The upper bound $|Z_0| \leq (\frac{e\, m}{\ell})^\ell$ is due to Lemma D.6 in Appendix D.2. □

We are now ready to present the proof of Proposition D.2.

Proof of Proposition D.2. Let $E_{R,W}$ be the event defined in Lemma D.4 corresponding to the matrix W and subspace R. On the event $E_{R,W}$ there exists a finite set $R_0 \subset R$ satisfying properties i) - iii) of Lemma D.4. Similarly, we can define the event $E_{S,W}$ for the matrix W and subspace S, and the finite set $S_0 \subset S$ satisfying properties i) - iii).

We can then define the event $E_{R,S} := E_{R,W} \cap E_{S,W}$ so that if $m \geq \tilde{C}_2 \ell'$ by Lemma D.4 we have

$$\mathbb{P}(E_{R,S}) \geq 1 - 2C_2 \exp(-c_2 \epsilon m).$$

For fixed $r_0 \in R_0$ and $s_0 \in S_0$, Lemma D.3 gives that if $m \geq 2\tilde{C}_1 \ell$ with probability at least $1 - 2\exp(-\tilde{c}_1 m)$

$$|\langle W_{+,r_0}^T W_{+,s_0} u, v\rangle - \langle Q_{r_0,v_0} u, v\rangle| \leq 3\epsilon \|u\|_2 \|v\|_2 \quad \forall u, v \in T.$$

Next, let E_0 be the event that

$$|\langle W_{+,r_0}^T W_{+,s_0} u, v\rangle - \langle Q_{r_0,v_0} u, v\rangle| \le 3\epsilon \|u\|_2 \|v\|_2 \quad \forall\, u, v \in T,\ r_0 \in R_0,\ s_0 \in S_0.$$

Then, on $E_{R,S}$, a union bound gives

$$\mathbb{P}(E_0) \ge 1 - 2|R_0||S_0| \exp(-\tilde{c}_1 m/2) \ge 1 - 2\Big(\frac{e\,m}{\ell}\Big)^{2\ell} \exp(-\tilde{c}_1 m/2).$$

We will next work on the event $E_0 \cap E_{R,S}$. Fix nonzero $r \in R$ and $s \in S$, and define the set of indices

$$\Omega_{r,s} := \{j \in [m] : \langle w_j, r\rangle = 0 \text{ or } \langle w_j, s\rangle = 0\}$$

Observe then that by the definition of $W_{+,r}$ and $\Omega_{r,s}$ the following holds

$$\begin{aligned} W_{+,r}^T W_{+,s} &= \sum_{j=1}^m \mathbb{1}_{\langle w_j, r\rangle > 0} \mathbb{1}_{\langle w_j, s\rangle > 0} w_j w_j^T \\ &= \sum_{j \in \Omega_{r,s}} \mathbb{1}_{\langle w_j, r\rangle > 0} \mathbb{1}_{\langle w_j, s\rangle > 0} w_j w_j^T + \sum_{j \in \Omega_{r,s}^c} \mathbb{1}_{\langle w_j, r\rangle > 0} \mathbb{1}_{\langle w_j, s\rangle > 0} w_j w_j^T \\ &= \sum_{j \in \Omega_{r,s}^c} \mathbb{1}_{\langle w_j, r\rangle > 0} \mathbb{1}_{\langle w_j, s\rangle > 0} w_j w_j^T \end{aligned}$$

On the event $E_{R,S}$, there exist therefore $r_0 \in R_0$ and $s_0 \in S_0$ such that for all $j \in \Omega_{r,s}^c$ it holds that

$$\operatorname{sgn}(\langle w_j, r\rangle) = \operatorname{sgn}(\langle w_j, r_0\rangle) \quad \text{and} \quad \operatorname{sgn}(\langle w_j, s\rangle) = \operatorname{sgn}(\langle w_j, s_0\rangle).$$

In particular, we can write

$$\begin{aligned} W_{+,r}^T W_{+,s} &= \sum_{j \in \Omega_{r,s}^c} \mathbb{1}_{\langle w_j, r\rangle > 0} \mathbb{1}_{\langle w_j, s\rangle > 0} w_j w_j^T \\ &= W_{+,r_0}^T W_{+,s_0} - \sum_{j \in \Omega_{r,s}} \mathbb{1}_{\langle w_j, r_0\rangle > 0} \mathbb{1}_{\langle w_j, s_0\rangle > 0} w_j w_j^T \\ &=: W_{+,r_0}^T W_{+,s_0} - \widetilde{W}_{+,r_0}^T \widetilde{W}_{+,s_0} \end{aligned}$$

The next lemma shows that the residual $\widetilde{W}_{+,r_0}^T \widetilde{W}_{+,s_0}$ has small norm when acting on T.

Lemma D.5. *Fix $0 < \epsilon < 1$ and $\ell < m$. Suppose that $W \in \mathbb{R}^{m \times n}$ has i.i.d. $\mathcal{N}(0, 1/m)$ entries. Let $T \subset \mathbb{R}^n$ be an ℓ-dimensional subspace and R_0 and S_0 be subsets of $\mathbb{R}^n$. Let E_1 be the event the following inequality holds for all set of indexes $\Omega \subset [m]$ with cardinality $|\Omega| \le 2\ell$:*

$$|\langle \widetilde{W}_{+,r_0}^T \widetilde{W}_{+,s_0} u, v\rangle| \le \epsilon \|u\|_2 \|v\|_2 \quad \forall u, v \in T,\ r_0 \in R_0,\ s_0 \in S_0$$

where

$$\widetilde{W}_{+,r_0}^T \widetilde{W}_{+,s_0} := \sum_{j \in \Omega} \mathbb{1}_{\langle w_j, r_0\rangle > 0} \mathbb{1}_{\langle w_j, s_0\rangle > 0} w_j w_j^T.$$

There exists a $\delta_\epsilon > 0$ such that if $m \ge 9\epsilon^{-1}\ell$ an $2\ell \le \delta_\epsilon m$, then $\mathbb{P}(E_1) \ge 1 - 2m \exp(-\epsilon m/36)$.

We now consider the event $E := E_1 \cap E_0 \cap E_{R,S}$ where E_1 is the event defined in the previous lemma. On E for all $r \in R$, $s \in S$ and $u, v \in T$,

$$\begin{aligned} |\langle W_{+,r}^T W_{+,s} u, v\rangle - \langle Q_{r,s} u, v\rangle| &= \Big|\langle W_{+,r_0}^T W_{+,s_0} u, v\rangle - \langle \widetilde{W}_{+,r_0}^T \widetilde{W}_{+,s_0} u, v\rangle - \langle Q_{r,s} u, v\rangle\Big| \\ &\le \Big|\langle W_{+,r_0}^T W_{+,s_0} u, v\rangle - \langle Q_{r_0,s_0} u, v\rangle\Big| \\ &\quad + \Big|\langle Q_{r_0,s_0} u, v\rangle - \langle Q_{r_0,s_0} u, v\rangle\Big| \\ &\quad + \Big|\langle \widetilde{W}_{+,r_0}^T \widetilde{W}_{+,s_0} u, v\rangle\Big| \\ &\le 3\epsilon \|u\|_2 \|v\|_2 + \frac{60}{\pi} \epsilon \|u\|_2 \|v\|_2 + \epsilon \|u\|_2 \|v\|_2 \\ &\le 24\epsilon \|u\|_2 \|v\|_2, \end{aligned} \tag{46}$$

where the first equality used the event $E_{R,S}$ and the definition of $\widetilde{W}^T_{+,r_0}\widetilde{W}_{+,s_0}$. The second inequality used instead the event $E_1 \cap E_0$ and the Lipschitz continuity of $Q_{r,s}$ (Lemma D.7).

In conclusion, there exist C_ϵ and c_ϵ such that if $m \geq C_\epsilon \ell'$ then

$$\begin{aligned}\mathbb{P}(E_1 \cap E_0 \cap E_{R,S}) &\geq 1 - 2m\exp(-\epsilon m/36) - 2\Big(\frac{e\,m}{\ell}\Big)^{2\ell}\exp(-\tilde{c}_1 m/2) - 2C_2\exp(-c_2\epsilon m) \\ &\geq 1 - \gamma\Big(\frac{e\,m}{\ell}\Big)^{2\ell}\exp(-c_\epsilon m)\end{aligned}$$

Here C_ϵ depends polynomially on ϵ^{-1} and c_ϵ depends polynomially on ϵ, and γ is positive absolute constant.

Notice that (46) gives a bound in terms of $24\epsilon\|u\|_2\|v\|_2$. To obtain a bound as in (39) simply rescale ϵ by $1/24$ in the discussion above, and modify c_ϵ and C_ϵ accordingly.

To extend (39) to the union of subspaces, we consider the subspace $T_{i,j} = \text{span}(U_i, V_j)$ with dimension at most $2\ell'$. Then use (39) with subspaces $T_{i,j}$, R_p and S_q, and take a union bound. □

D.2 Supplemental Results for Section D

We begin this section by providing an upper bound on the number of activation patterns of a ReLU layer. This result is used in the proof of Lemma D.4.

Lemma D.6. *Let S be an ℓ-dimensional subspace of $\mathbb{R}^n$ and $m \geq \ell$. Let $W \in \mathbb{R}^{m\times n}$ have i.i.d $\mathcal{N}(0, 1/m)$ entries. Then with probability 1,*

$$|\{\mathit{diag}(Ws > 0)W \ |s \in S\}| \leq \Big(\frac{em}{\ell}\Big)^\ell$$

Proof. Observe that by rotational invariance of the Gaussian distribution we may take, without loss of generality, S to be the span of the first ℓ standard basis vector, i.e. $S = \text{span}(e_1, \ldots, e_\ell)$. We can then also take $W \in \mathbb{R}^{m\times\ell}$ and $S = \mathbb{R}^\ell$.

Let $\{w_j\}_{j=1}^m$ be the rows of the matrix W. Notice that for fixed W, $|\{\text{diag}(Ws > 0)W \ |s \in S\}|$ equals the number of binary vectors $(\mathbb{1}_{\langle w_j, v\rangle > 0})_{j\in[n]}$ for $v \in \mathcal{S}^{\ell-1}$. Each $(\mathbb{1}_{\langle w_j, v\rangle>0})_{j\in[n]}$ uniquely identifies a region of the partitioning of $\mathbb{R}^l$ induced by the set of hyperplanes $\mathcal{H} := \{x : \langle w_j, x\rangle = 0\}$. From the theory of hyperplane arrangements [29] we know that $m \geq \ell$ hyperplanes in $\mathbb{R}^\ell$ partition the space in at most $\sum_{j=0}^\ell \binom{m}{j}$. Thus, with probability 1 we have

$$\begin{aligned}|\{\text{diag}(Ws > 0)W \ |s \in S\}| &\leq \sum_{j=0}^{\ell}\binom{m}{j} \\ &\leq \sum_{j=0}^{\ell}\frac{m^j}{j!} \leq \sum_{j=0}^{\ell}\frac{\ell^j}{j!}\Big(\frac{m}{\ell}\Big)^j \leq \Big(\frac{m}{\ell}\Big)^\ell\sum_{j=0}^{\infty}\frac{\ell^j}{j!} = \Big(\frac{em}{\ell}\Big)^\ell\end{aligned}$$

□

Next we prove Lemma D.5, providing an upper bound for the random matrix $\widetilde{W}^T\widetilde{W}$ when acting on low-dimensional subspaces.

Proof of Lemma D.5. Notice that for any $\Omega \subset [m]$, $u, v \in T$, $r_0 \in R_0$ and $s_0 \in S_0$, it holds that

$$\begin{aligned}|\langle\widetilde{W}^T_{+,r_0}\widetilde{W}_{+,s_0}u, v\rangle| &= |\langle\text{diag}(W_\Omega r_0 > 0)\odot\text{diag}(W_\Omega s_0 > 0)W_\Omega u, W_\Omega v\rangle| \\ &\leq \|\text{diag}(W_\Omega r_0 > 0)\odot\text{diag}(W_\Omega s_0 > 0)\|\|W_\Omega v\|\|W_\Omega u\| \\ &\leq \|W_\Omega v\|\|W_\Omega u\|.\end{aligned}$$

Therefore, it is sufficient to show that

$$\|W_\Omega u\| \leq \sqrt{\epsilon}\|u\| \qquad \forall u \in T \ \ \forall\Omega \subset [m] \ \text{ satisfying } \ |\Omega| \leq 2\ell \leq \delta_\epsilon m.$$

The rest of the proof follows, *mutatis mutandis*, as in Lemma 26 of [17]. □

We will next show that $Q_{x,y}$ is a Lipschitz function of its arguments.

Lemma D.7. *Fix $0 < \epsilon < 1$ and $x, \tilde{x}, y, \tilde{y} \in \mathcal{S}^{n-1}$. If $\|\tilde{x} - x\| \leq \epsilon$ and $\|\tilde{y} - y\| \leq \epsilon$, then*

$$\|Q_{\tilde{x},\tilde{y}} - Q_{\tilde{x},\tilde{y}}\| \leq \Big(\frac{2}{\pi} + 2\sqrt{79}\Big)\epsilon$$

Proof. Recall the following facts:

$$\|x - y\| \geq 2\sin(\angle(x,y)/2), \qquad \forall x, y \in \mathcal{S}^{n-1} \tag{47}$$
$$|\angle(x_1, x_2)| \geq |\angle(x_1, y) - \angle(x_2, y)|, \qquad \forall x_1, x_2, y \in \mathcal{S}^{n-1} \tag{48}$$
$$\sin(\theta/2) \geq \theta/4, \qquad \forall \theta \in [0, \pi] \tag{49}$$

Let $\theta_{\tilde{x},x} = \angle(\tilde{x}, x)$ and $\theta_{\tilde{y},y} = \angle(\tilde{y}, y)$, then

$$\|Q_{x,y} - Q_{\tilde{x},\tilde{y}}\| \leq \frac{|\theta_{x,y} - \theta_{\tilde{x},\tilde{y}}|}{2\pi} + \Big\|\frac{\sin\theta_{x,y}}{2\pi} M_{x\leftrightarrow y} - \frac{\sin\theta_{\tilde{x},\tilde{y}}}{2\pi} M_{\tilde{x}\leftrightarrow\tilde{y}}\Big\|.$$

By (48) it holds that

$$|\theta_{x,y} - \theta_{\tilde{x},\tilde{y}}| \leq |\theta_{x,y} - \theta_{\tilde{x},y}| + |\theta_{\tilde{x},y} - \theta_{\tilde{x},\tilde{y}}| \leq |\theta_{\tilde{x},x}| + |\theta_{\tilde{y},y}|,$$

while from (47) and (49) it follows that

$$|\theta_{\tilde{x},x}| \leq 4\sin(\theta_{\tilde{x},x}/2) \leq 2\epsilon,$$
$$|\theta_{\tilde{y},y}| \leq 4\sin(\theta_{\tilde{y},y}/2) \leq 2\epsilon.$$

Thus $|\theta_{x,y} - \theta_{\tilde{x},\tilde{y}}| \leq 4\epsilon$. Lemma B.3 in [10] then proves that

$$\Big\|\frac{\sin\theta_{x,y}}{2\pi} M_{x\leftrightarrow y} - \frac{\sin\theta_{\tilde{x},\tilde{y}}}{2\pi} M_{\tilde{x}\leftrightarrow\tilde{y}}\Big\| \leq 2\sqrt{79}\epsilon,$$

which concludes the proof. □

E Proof of Lemma 5.3

In this section we prove Lemma 5.3 which is used to bound the perturbation of the objective function f_{cs} and its gradient due to the presence of the noise term η.

Proof of Lemma 5.3. Fix $x, z \in \mathcal{S}^{k-1}$ and notice that by the properties of the Gaussian distribution, for $t \geq 0$ it holds that

$$\mathbb{P}_A\left[\langle z, \Lambda_x^T A^T \eta\rangle \geq \frac{\|\Lambda_x z\|}{\sqrt{m}}\|\eta\| t\right] = \mathbb{P}_{y\sim\mathcal{N}(0,1)}\left[\frac{\|\Lambda_x z\|}{\sqrt{m}}\|\eta\| y \geq \frac{\|\Lambda_x z\|}{\sqrt{m}}\|\eta\| t\right] \leq e^{-\frac{t^2}{2}}.$$

If $z = x$ use (20b), while if $z \neq x$ and G differentiable at x use (22), to obtain that

$$\mathbb{P}_A\left[\langle z, \Lambda_x^T A^T \eta\rangle \geq \sqrt{\frac{13}{12}}\frac{\|\eta\|}{2^{d/2}}\frac{t}{\sqrt{m}}\right] \leq e^{-\frac{t^2}{2}}$$

Let $\mathcal{N}_{1/2}$ be a $\frac{1}{2}$-net over $\mathcal{S}^{k-1}$ such that $|\mathcal{N}_{1/2}| \leq 5^k$ (see for example [38]). Recall that by Lemma D.1 the number of different matrices Λ_x is bounded by Ψ_d. Thus, a union bound gives

$$\mathbb{P}\left[\langle z, \Lambda_x^T A^T \eta\rangle \geq \sqrt{\frac{13}{12}}\frac{\|\eta\|}{2^{d/2}}\frac{t}{\sqrt{m}},\ \ \forall x, z \in \mathcal{S}^{k-1}\right] \leq |\mathcal{N}_{\frac{1}{2}}|\,\Psi_d\,\mathbb{P}\left[\langle z, \Lambda_x^T A^T \eta\rangle \geq \sqrt{\frac{13}{12}}\frac{\|\eta\|}{2^{d/2}}\frac{t}{\sqrt{m}}\right]$$
$$\leq \exp(-\frac{t}{2}^2 + \log 5 + \log\Psi_d)$$

Choosing $t = 2\sqrt{k\log(5\prod_{i=1}^d \frac{e\,n_i}{k})}$ we obtain the theses.

□

F Extensions

F.1 Compressive Phase Retrieval with a Generative Prior

Consider a generative network $G : \mathbb{R}^k \to \mathbb{R}^n$ as in (2). The compressive phase retrieval problem with a generative network prior can be formulated as follows.

COMPRESSIVE PHASE RETRIEVAL WITH A DEEP GENERATIVE PRIOR

Let: $G : \mathbb{R}^k \to \mathbb{R}^n$ generative network, $A \in \mathbb{R}^{m \times n}$ measurement matrix.
Let: $y_\star = G(x_\star)$ for some unknown $x_\star \in \mathbb{R}^k$.

Given: G and A.
Given: Measurements $b = |Ay_\star| + \eta \in \mathbb{R}^m$ with $m \ll n$ and $\eta \in \mathbb{R}^m$ noise.

Estimate: $y_\star$.

To estimate $y_\star$, [16] proposes to find the latent code $\hat{x}$ that minimizes the reconstruction error

$$\tilde{x} = \arg\min_{x \in \mathbb{R}^x} f_{\text{pr}}(x) := \frac{1}{2}\|b - |AG(x)|\|_2^2, \qquad (50)$$
$$y_\star \approx G(\tilde{x}).$$

In [17] it is shown that Algorithm 1 with inputs f_{pr}, small enough step size and arbitrary initial condition estimates $y_\star$ up to the noise level in polynomial time, provided that the number of phaseless measurements is up-to log-factors $m \geq k \cdot \text{poly}(d)$ and the generative network is logarithmically expansive. The proof uses the WDC and an isometry condition akin to the RRIC. As before, the RWDC can be replaced by the R2WDC and obtain the same convergence guarantees. Moreover, as in the case of compressed sensing, the logarithmic factor in the number of measurements can be improved using Lemma D.1.

F.2 Denoising with a Generative Prior

Consider a generative network $G : \mathbb{R}^k \to \mathbb{R}^n$ as in (2). The denoising problem with a generative network prior can be formulated as follows.

DENOISING WITH A DEEP GENERATIVE PRIOR

Let: $G : \mathbb{R}^k \to \mathbb{R}^n$ generative network.
Let: $y_\star = G(x_\star)$ for some unknown $x_\star \in \mathbb{R}^k$.

Given: G.
Given: Noisy signal $b = y_\star + \eta \in \mathbb{R}^m$ with $\eta \sim \mathcal{N}(0, \sigma^2 \, \mathrm{I}_n)$ noise.

Estimate: $y_\star$.

To estimate $y_\star$, [20] proposes to find the latent code $\hat{x}$ that minimizes the reconstruction error

$$\tilde{x} = \arg\min_{x \in \mathbb{R}^x} f_{\text{den}}(x) := \frac{1}{2}\|b - G(x)\|_2^2, \qquad (51)$$
$$y_\star \approx G(\tilde{x}).$$

In [20] recovery guarantees based on this minimization problem are given for an expansive generative network G. Specifically, it is shown that Algorithm 1 with input f_{den}, small enough step size α and arbitrary initial point x_0, reconstructs the signal $y_\star$ up to an $O(k/n)$ error. The random network G is assumed to be logarithmically expansive in order to satisfy the WDC with high-probability, but inspecting the proof it can be seen that the R2WDC is enough. Using Lemma 5.2 we can extend the result of [20] to the case of a generative network satisfying Assumptions B.

F.3 Spiked Matrix Recovery with a Generative Prior

Consider a generative network $G : \mathbb{R}^k \to \mathbb{R}^n$ as in (2). The spiked Wishart matrix recovery with a generative prior is formulated as follows.

SPIKED WISHART MATRIX RECOVERY WITH A DEEP GENERATIVE PRIOR

Let: $G : \mathbb{R}^k \to \mathbb{R}^n$ generative network.
Let: $y_\star = G(x_\star)$ for some unknown $x_\star \in \mathbb{R}^k$.

Given: G.
Given: Noisy matrix $B = u\, y_\star{}^T + \sigma \mathcal{Z} \in \mathbb{R}^{N \times n}$, with $u \sim \mathcal{N}(0, I_N)$ and $\mathcal{Z}$ with i.i.d. $\mathcal{N}(0,1)$ entries.

Estimate: $y_\star$.

Similarly, the spiked Wigner matrix recovery with a generative prior is formulated as follows.

SPIKED WIGNER MATRIX RECOVERY WITH A DEEP GENERATIVE PRIOR

Let: $G : \mathbb{R}^k \to \mathbb{R}^n$ generative network.
Let: $y_\star = G(x_\star)$ for some unknown $x_\star \in \mathbb{R}^k$.

Given: G.
Given: Noisy matrix $B = y_\star\, y_\star{}^T + \sigma \mathcal{H} \in \mathbb{R}^{n \times n}$, with $\mathcal{H}$ from a Gaussian Orthogonal Ensemble.

Estimate: $y_\star$.

To estimate $y_\star$, [8] proposes to find the latent code $\hat{x}$ that minimizes the reconstruction error

$$\tilde{x} = \arg\min_{x \in \mathbb{R}^x} f_{\text{spiked}}(x) := \frac{1}{2}\|M - G(x)G(x)^T\|_F^2,$$
$$y_\star \approx G(\tilde{x}),$$

where

- in the spiked Wishart model $M = B^T B/N - \sigma^2 I_n$;
- in the spiked Wigner model $M = B$.

As shown in [9] Algorithm 1 with inputs f_{spiked}, appropriate α and arbitrary initial point x_0, estimates in polynomial time the signal $y_\star$ with rate-optimal dependence on the noise level or sample complexity. In particular, this shows that the absence of a computational-statistical gap in spiked matrix recovery with an expansive (random) generative network prior. The proof uses the fact that for G satisfying the WDC the bounds in Proposition C.1, C.2, C.3 and C.4 hold. Since these bounds hold under the weaker R2WDC we can directly extend the results in [9] to non-expansive generative networks G satisfying Assumptions B.

G An example of a contractive generative network

In this section we give an example of a generative network as in (2) satisfying the conditions (12) and (17), and with contractive layers.

Let $d \geq 2$ and $\bar{C}_\epsilon := \max(\widetilde{C}_\epsilon, 16 c_\epsilon^{-1}/\log(2))$. Then consider a d-layer generative network G such that for $i \in [d]$

$$n_i := \bar{C}_\epsilon \cdot k \cdot d(2d - i) \cdot \alpha,$$

where $\alpha \cdot \bar{C}_\epsilon \in \mathbb{N}$ and

$$\alpha \geq \max\left\{\frac{2\log\left(\bar{C}_\epsilon \cdot k\right)}{d^2}, \log\left(e^2\bar{C}_\epsilon\right)\right\}. \tag{52}$$

We now demonstrate that n_i satisfies (12). Notice that

$$\begin{aligned}
\log\Big(\prod_{j=1}^{i-1}\frac{en_j}{k}\Big) &= \sum_{j=1}^{i-1}\log\left(\alpha\cdot\bar{C}_\epsilon\cdot d(2d-j)\cdot e\right)\\
&\leq (d-1)\log\left(\alpha\cdot\bar{C}_\epsilon\cdot 2d^2\cdot e\right)\\
&= (d-1)\big[\log(e\bar{C}_\epsilon)+2\log(d)\big]+(d-1)\log(2\alpha)\\
&\leq (d-1)d\big[\log(e\bar{C}_\epsilon)+1\big]+(d-1)\alpha\\
&\leq (d-1)d\alpha+(d-1)\alpha\\
&= (d^2-1)\alpha
\end{aligned}$$

where in the second inequality we have used $2\log(x)\leq x$ and $\log(2x)\leq x$ for $x>0$ and in the third (52). Next since $d(2d-i)\geq(d^2-1)$ for every $i\in[d]$, n_i satisfies (12) for every $i\in[d]$.

We now show that n_i satisfies (17). We have

$$\begin{aligned}
\log(n_i) &= \log(\bar{C}_\epsilon\cdot k\cdot d(2d-i)\cdot\alpha)\\
&= \log(d(2d-i)\alpha)+\log(\bar{C}_\epsilon k)\\
&\leq \frac{d(2d-i)\alpha}{2}+\log(\bar{C}_\epsilon k)\\
&\leq \frac{d(2d-i)\alpha}{2}+\frac{d^2\alpha}{2}\\
&\leq d(2d-i)\alpha,
\end{aligned}$$

where in the first inequality we have used $2\log(x)\leq x$ for $x>0$, in the second inequality (52) and in the third $2d^2\leq d(2d-i)$ for every $i\in[d]$. We therefore have

$$\log(n_i)\cdot\frac{16\cdot k\cdot c_\epsilon^{-1}}{\log(2)}\leq d(2d-i)\cdot\alpha\cdot\frac{16\cdot k\cdot c_\epsilon^{-1}}{\log(2)}\leq d(2d-i)\cdot\alpha\cdot\bar{C}_\epsilon\cdot k=n_i.$$

H Experiments

In this section, we verify the prediction of our theory (Theorem 5.4) on synthetic data. We consider the compressed sensing problem with generative prior formulated in Se ction 3 and show that a practical implementation of Algorithm 1 can estimates the sought signal in the case of generative networks with contractive layers.

At each step Algorithm 1 requires the computation of the subgradient of f_{cs} at a point $\tilde{x}$. This could be achieved for example by computing $\nabla f_{\text{cs}}(\tilde{x}+\delta)$ where δ is sufficiently small and f_{cs} is differentiable at $f_{\text{cs}}(\tilde{x}+\delta)$. As already shown in [21], in practice, one can replace Algorithm 1 with Algorithm 2 below. In the Algorithm 2, $v'_{\tilde{x}_t}$ corresponds to the gradient of f_{cs} at a point of differentiability. At a point of non-differentiability $v'_{\tilde{x}_t}$ corresponds to the output of autograd applied to f_{cs} as implemented by the current deep learning libraries. Notice that the set of non-differentiability of f_{cs} has measure zero and it is extremely unlikely that the iterates of the subgradient descent will hit this set. Hence, in practice Algorithm 1 and Algorithm 2 are equivalent.

We then the performance of Algorithm 2 when solving compressed sensing with generative networks G with two and three layers. The entries of the weights W_i of G are drawn from $\mathcal{N}(0, I/n_i)$, while the entries of the measurements matrix A are drawn from $\mathcal{N}(0, I/m)$. In all the experiments we run Algorithm 2 for 10,000 iterations, or until the norm of the $v'_{\tilde{x}_t}$ is below 10^{-10}, whichever comes first. We fix the learning rate ν to 0.7 and randomly initialize x_0. The target $x_\star$ is chosen so that $y_\star = G(x_\star)$ has unit norm.

In the first experiment, we consider the noiseless case with $\eta = 0$ and a fixed number of measurements $m = 300$. The 2-layers generative networks have fixed layer widths of dimensions $n_1 = 700$ and

Algorithm 2: PRACTICAL SUBGRADIENT DESCENT [21]

Input: Objective function f, initial point $x_0 \in \mathbb{R}^k \setminus \{0\}$ and step size α
Output: An estimate of the target signal $y_\star = G(x_\star)$ and latent vector $x_\star$

for $t = 0, 1, \ldots$ **do**
  if $f(-x_t) < f(x_t)$ **then** $\tilde{x}_t \leftarrow -x_t$
  else $\tilde{x}_t \leftarrow x_t$
  $v'_{\tilde{x}_t} = \Lambda^T_{d,\tilde{x}_t} A^T (A \Lambda_{d,\tilde{x}_t} \tilde{x}_t - b)$
  $x_{t+1} \leftarrow \tilde{x}_t - \alpha v'_{\tilde{x}_t}$
end
return $x_t, G(x_t)$

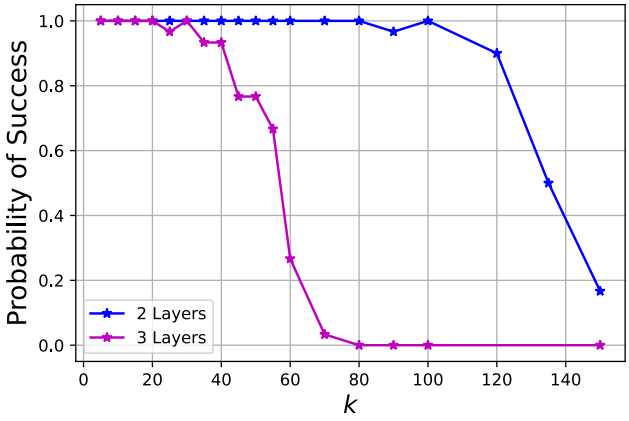

Figure 1: Empirical probability of successful recovery the latent vector from compressed noiseless random measurements versus the latent dimension. The 2-layers generative networks have fixed layer widths of dimensions $n_1 = 700$ and $n = n_2 = 600$ and the 3-layers generative networks have fixed layer widths of dimensions $n_1 = 150$, $n_2 = 700$ and $n = n_3 = 600$.

$n = n_2 = 650$ with varying latent dimension $n_0 = k$. The 3-layers generative networks have fixed layer widths of dimensions $n_1 = 150$, $n_2 = 700$ and $n = n_3 = 600$ with varying latent dimension $n_0 = k$.

Figure 1 reports the empirical probability of successful recovery from 30 random independent trials (over A and G) for of noiseless problems, where a run Algorithm 2 is called successful if the relative reconstruction error $\|x_\star - x_T\|/\|x_\star\|$ is below 10^{-3}. These experiments show that recovery of $x_\star$ from undersampled measurements can be achieved even in the case of non-expansive generative networks and adding more layers makes the problem harder. Nonetheless, the algorithm succeeds in wider range of parameters as predicted by our theory, in particular with a milder dependence on the depth d.

In the second experiment we study the noisy compressive sensing problem with a fixed number of measurements $m = 300$. We consider 2-layers generative networks with fixed layer widths of dimensions $n_1 = 700$ and $n = n_2 = 600$, and varying latent dimension $n_0 = k$ in $\{5, 10, 15, 20, 25, 30, 35, 40, 45\}$. The measurements are taken to be $b = Ay_\star + \eta$ where $\eta = \tau\, e$ where e is taken uniformly at random over the sphere $\mathcal{S}^{m-1}$ and τ is chosen so that the signal to noise ratio (SNR) varies in $\{20, 40, 80\}$. The SNR is defined as $10 \log_{10} \left(\|AG(x_\star)\|/\|\eta\| \right)$.

In Figure 2 we plot the average reconstruction error $\|G(x_T) - y_\star\|/\|y_\star\|$ over 30 independent random draws of A and G. As predicted by our theory this quantity scales linearly with the ratio k/m. Furthermore, the error is proportional to the magnitude of the noise $\|\eta\|$.

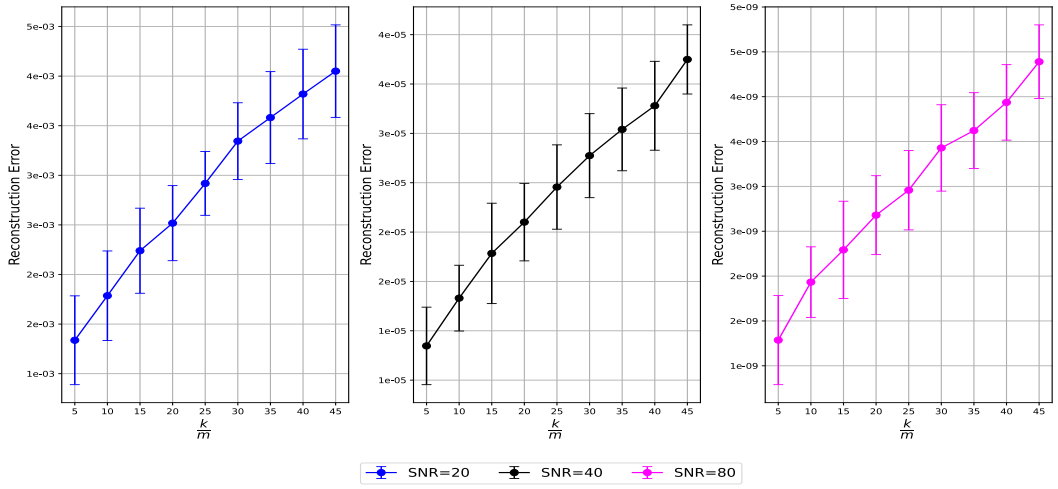

Figure 2: Reconstruction error $\|y_\star - G(x)\|/\|y_\star\|$ from noisy compressed linear measurements at varying level of SNR. The number of measurements m is fixed at 300 and latent dimension k is varied. The generative networks have outputs of dimension $n = n_2 = 600$ and hidden layers of dimension $n_1 = 700$.