A. Proofs of theoretical results

A.1. Lemma 1

To account for measurement noise in our analysis, we define the ϵ-tube set T of a matrix A as,

$$\mathrm{T}_A(\epsilon) = \{w : \|Aw\|_2 \leq \epsilon\}.$$

Note that in the absence of noise, $\mathrm{T}_A(0)$ corresponds to the nullspace of A. Next, we define a difference function, $G' : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}^n$ such that $G'(z_1, z_2) = G(z_1) - G(z_2)$. Consequently, we obtain a difference set $S_{l,G'}$ as the Minkowski sum of $S_l(0)$ (the space of l sparse vectors) and range of G',

$$S_{l,G'} = \cup_{z_1,z_2} S_l(G'(z_1, z_2)).$$

This allows us to define $\sigma_{l,G'}(x)$ as,

$$\sigma_{S_{l,G'}}(x) = \inf_{\hat{x} \in S_{l,G'}} \|x - \hat{x}\|_1.$$

Now, in order to prove Lemma 1, we state and derive a couple of lemmas. The proofs of the next two Lemmas (3 and 4) are modeled along the theory developed in Cohen et al. (2009) for the sensing of l-sparse vectors. We extend it to the case of $S_{l,G}$. Lemma 3 encodes the idea that for sensing to be successful any two points in $S_{l,G}$ should not be very close when acted upon by the measurement map A. This can be equivalently stated as requiring that any point in the nullspace of A should not be approximated very well by points in $S_{2l,G'}$. Because we are working with bounded noise we need these results on the tube $T_A(2\epsilon)$, instead of just the nullspace. A point of interest is that informally the next lemma provides a sufficient condition for a good decoder to exist and also provides a different set of similar necessary conditions for good decoding.

Lemma 3. *Given a measurement matrix $A \in \mathbb{R}^{m \times n}$, measurement noise ϵ such that $\|\epsilon\|_2 \leq \epsilon_{\max}$, and a generative model function $G : \mathbb{R}^k \to \mathbb{R}^n$ we want a decoder $\Delta : \mathbb{R}^m \to \mathbb{R}^n$ which provides the following (ℓ_2, ℓ_1)-mixed norm approximation guarantee on the set of l-sparse vectors S_l,*

$$\|x - \Delta(Ax + \epsilon)\|_2 \leq C_0 l^{-t} \sigma_{l,G}(x) + C_1 \epsilon_{\max} + \delta$$

for some constants $C_0, \delta, t \geq 0$.

The sufficient condition for such a decoder to exist is given by,

$$\|\eta\|_2 \leq \frac{C_0}{2} l^{-t} \sigma_{2l,G'}(\eta) + C_1 \epsilon_{\max} + \delta, \forall \eta \in T_A(2\epsilon_{\max}).$$

We call this the (ℓ_2, ℓ_1)-mixed norm null space property.

A necessary condition for the same follows,

$$\|\eta\|_2 \leq C_0 l^{-t} \sigma_{2l,G'}(\eta) + 2C_1 \epsilon_{\max} + 2\delta, \forall \eta \in T_A(\epsilon_{\max}).$$

Proof. To prove the sufficiency of the null space condition, we define a decoder $\Delta : \mathbb{R}^m \to \mathbb{R}^n$ as follows,

$$\Delta(y) = \underset{x : \|Ax - y\|_2 \leq \epsilon_{\max}}{\arg\min} \sigma_{l,G}(x).$$

We will prove this decoder satisfies the mixed norm guarantee given the (ℓ_2, ℓ_1)-mixed norm null space property. Using the definition of Δ, we have,

$$\|A(x - \Delta(Ax + \epsilon))\|_2 \leq \|Ax + \epsilon - A\Delta(Ax + \epsilon))\|_2 + \epsilon_{\max} \leq 2\epsilon_{\max}.$$

This implies $x - \Delta(Ax + \epsilon) \in T_A(2\epsilon_{\max})$. Combining with the mixed norm guarantee, we have,

$$\begin{aligned} \|x - \Delta(Ax + \epsilon)\|_2 - \delta - C_1 \epsilon_{\max} &\leq \frac{C_0}{2} l^{-t} \sigma_{2l,G'}(x - \Delta(Ax + \epsilon)) \\ &\leq \frac{C_0}{2} l^{-t} (\sigma_{l,G}(x) + \sigma_{l,G}(\Delta(Ax + \epsilon))) \\ &\leq C_0 l^{-t} \sigma_{l,G}(x). \end{aligned}$$

The second last step follows from the triangle inequality whereby $\sigma_{2l,G'}(x+y) \leq \sigma_{l,G}(x) + \sigma_{l,G}(y)$ and the last step uses the fact that the decoder is the minimizer of $\sigma_{l,G}(x)$.

For the necessary condition, consider any decoder Δ which provides the needed guarantee. Consider $\eta \in T_A(\epsilon_{\max})$ and now pick $z_0, z_1 \in \mathbb{R}^k, \eta_0 \in S_{2l}$ such that the following inequality is satisfied,

$$\|\eta - (G(z_0) - G(z_1) + \eta_0)\|_1 - \varepsilon' \leq \sigma_{2l,G'}(\eta), \tag{13}$$

where $\varepsilon' > 0$. We can find a z_0, z_1, η_0 for any arbitrarily small and positive ε'. This is the case because we have,

$$\sigma_{2l,G'}(\eta) = \inf_{\hat{\eta} \in S_{2l,G'}} \|\eta - \hat{\eta}\|_1 = \inf_{\hat{z}_0, \hat{z}_1 \in \mathbb{R}^k, \hat{\eta}_0 \in S_{2l}} \|\eta - (G(\hat{z}_0) - G(\hat{z}_1) + \hat{\eta}_0)\|_1,$$

which we obtain by parameterizing $\hat{\eta} \in S_{2l,G'}$ as $\hat{\eta} = G(\hat{z}_0) - G(\hat{z}_1) + \hat{\eta}_0$ for $\hat{z}_0, \hat{z}_1 \in \mathbb{R}^k, \hat{\eta}_0 \in S_{2l}$. We cannot necessarily find z_0, z_1, η_0 such that $\varepsilon' = 0$ because $S_{2l,G'}$ may not be a closed set. For convenience, we let $G_0 = G(z_0)$ and $G_1 = G(z_1)$ which means $G'(z_0, z_1) = G(z_0) - G(z_1) = G_0 - G_1$. We can split η_0 as $\eta_0 = \eta_1 + \eta_2$ for some $\eta_1, \eta_2 \in S_l$, and for convenience define $\eta_3 = \eta - \eta_0 - G_0 + G_1$. Note, we can now rewrite (13) as,

$$\|\eta_3\|_1 \leq \sigma_{2l,G'}(\eta) + \varepsilon' \tag{14}$$

Since $G_0 + \eta_1 \in S_{l,G}$, we have $\sigma_{l,G}(G_0 + \eta_1) = 0$. This simplifies the (ℓ_2, ℓ_1)-mixed norm guarantee of our decoder when applied to $G_0 + \eta_1$,

$$\|G_0 + \eta_1 - \Delta(A(G_0 + \eta_1))\|_2 \leq \delta + C_1 \epsilon_{\max}. \tag{15}$$

Plugging in all the above, we have:

$$\begin{aligned}
\|\eta\|_2 &= \|\eta_2 + \eta_1 + \eta_3 + G_0 - G_1\|_2 \\
&\leq \|\eta_1 + G_0 - \Delta(A(\eta_1 + G_0))\|_2 + \|\eta_3 + \eta_2 - G_1 + \Delta(A(\eta_1 + G_0))\|_2 \\
&\leq \delta + C_1\epsilon_{\max} + \| - \eta_3 - \eta_2 + G_1 - \Delta(A\eta + A(G_1 - \eta_2 - \eta_3))\|_2 && \text{(from Eq. (15))} \\
&\leq 2\delta + 2C_1\epsilon_{\max} + C_0 l^{-t}\sigma_{l,G}(G_1 - \eta_2 - \eta_3) && \text{(since } \eta \in T_A(\epsilon_{\max}) \text{ and the } (\ell_2, \ell_1)\text{-guarantee)} \\
&\leq 2\delta + 2C_1\epsilon_{\max} + C_0 l^{-t}\|\eta_3\|_1 \\
&= C_0 l^{-t}\sigma_{2l,G'}(\eta) + 2\delta + 2C_1\epsilon_{\max} + C_0 l^{-t}\varepsilon'. && \text{(from Eq. (14))}
\end{aligned}$$

As we can make ε' arbitrarily small we can make it tend to 0 providing us with the required result. □

The next lemma basically shows that if A satisfies the S-REC and RIP conditions then we operate in the constraint regime required by the previous lemma.

Lemma 4. *If the measurement matrix $A \in \mathbb{R}^{m \times n}$ satisfies S-REC($S_{(a+b)l/2,G'}, 1-\alpha, \delta$) and RIP($bl, \alpha$) for integers $a, b, l > 0$ and function $G : \mathbb{R}^k \to \mathbb{R}^m$, then we have for any vector $\eta \in T_A(\epsilon)$,*

$$\|\eta\|_2 \leq (bl)^{-1/2}(C_0 + 1)\sigma_{al,G'}(\eta) + C_1\epsilon + \delta'$$

where $C_0 = (1-\alpha)^{-1}(1+\alpha), C_1 = (1-\alpha)^{-1}$, $\delta' = \delta(1-\alpha)^{-1}$.

Proof. For any choice of $\eta \in T_A(\epsilon)$ and $G(z_1), G(z_2)$, let $\nu \in S_{al}$ be the minimizer of $\|\eta - G(z_1) + G(z_2) - \nu\|_1$. We can find this ν because S_{al} is closed, concretely we can construct this ν by taking a n dimensional vector which has everything but the top al magnitude components in $\eta - G(z_1) + G(z_2)$ zeroed out. As the choice of $G(z_1)$ and $G(z_2)$ is arbitrary, it suffices to prove the statement for $\|\eta - G(z_1) + G(z_2) - \nu\|_1$ (instead of $\sigma_{al,G'}(\eta)$).

Given a set of indices $\mathcal{I}$ for a n dimensional vector we use $\mathcal{I}^c$ to denote the set of indices not in $\mathcal{I}$. Now note that ν corresponds to the al largest coordinates of $\eta' = \eta - G(z_1) + G(z_2)$. Let the indices corresponding to those coordinates be $\mathcal{T}_0$. We take $\mathcal{T}_1$ to be the indices of the next bl (and not al) largest coordinates. Similarly, define $\mathcal{T}_2, \ldots, \mathcal{T}_s$ to be subsequent indices for the next bl largest coordinates. The final set $\mathcal{T}_s$ can contain indices of less than bl coordinates. Let $\mathcal{T}_0 \cup \mathcal{T}_1 = \mathcal{T}$. We will use $x_{\mathcal{I}}$ to denote the vector obtained by zeroing out values in x for all indices in the set $\mathcal{I}^c$. We can write $\eta_{\mathcal{T}} + (G(z_1) - G(z_2))_{\mathcal{T}^c}$ as $\eta_{\mathcal{T}} - (G(z_1) - G(z_2))_{\mathcal{T}} + (G(z_1) - G(z_2))$ where $\eta_{\mathcal{T}}, (G(z_1) - G(z_2))_{\mathcal{T}} \in S_{(a+b)l}$.

We can write $\eta_{\mathcal{T}} + (G(z_1) - G(z_2))_{\mathcal{T}}$ as $s_1 - s_2$ where $s_1, s_2 \in S_{(a+b)l/2}$. This allows us to write $\eta_{\mathcal{T}} + (G(z_1) - G(z_2))_{\mathcal{T}^c}$ as $G(z_1) + s_1 - (G(z_2) + s_2)$ where $G(z_1) + s_1, G(z_2) + s_2 \in S_{(a+b)l/2,G'}$. Now we use the fact that A satisfies S-REC($S_{(a+b)l/2,G'}$) to get,

$$\begin{aligned}
\|\eta_{\mathcal{T}} + (G(z_1) - G(z_2))_{\mathcal{T}^c}\|_2 &= \|G(z_1) + s_1 - (G(z_2) + s_2)\|_2 \\
&\leq (1-\alpha)^{-1}\|A(G(z_1) + s_1 - (G(z_2) + s_2))\|_2 + (1-\alpha)^{-1}\delta \qquad \text{(using S-REC)} \\
&\leq (1-\alpha)^{-1}\|A(\eta_{\mathcal{T}} + (G(z_1) - G(z_2))_{\mathcal{T}^c})\|_2 + (1-\alpha)^{-1}\delta. \qquad (16)
\end{aligned}$$

We can write $\eta = \eta_{\mathcal{T}} + \eta_{\mathcal{T}_2} + ... + \eta_{\mathcal{T}_s}$. As $\eta \in T_A(\epsilon)$ we can write $A\eta_{\mathcal{T}} = -A(\eta_{\mathcal{T}_2} + ... + \eta_{\mathcal{T}_s}) + \gamma$ where $\|\gamma\|_2 \leq \epsilon$. Hence,

$$\begin{aligned}
\|A(\eta_{\mathcal{T}} + (G(z_1) - G(z_2))_{\mathcal{T}^c})\|_2 &= \|A((\eta - G(z_1) + G(z_2))_{\mathcal{T}_2} + .. + (\eta - G(z_1) + G(z_2))_{\mathcal{T}_s}) - \gamma\|_2 \\
&= \|A\eta'_{\mathcal{T}_2} + ... + A\eta'_{\mathcal{T}_s} - \gamma\|_2 \\
&\leq \sum_{j=2}^{s} \|A\eta'_{\mathcal{T}_j}\|_2 + \|\gamma\|_2 \\
&\leq (1+\alpha)\sum_{j=2}^{s} \|\eta'_{\mathcal{T}_j}\|_2 + \epsilon. \qquad \text{(using RIP)} \\
& \qquad (17)
\end{aligned}$$

From Eq. (16) and Eq. (17), we get,

$$\|\eta_{\mathcal{T}} + (G(z_1) - G(z_2))_{T^c}\|_2 - \delta' \leq (1-\alpha)^{-1}(1+\alpha)\sum_{j=2}^{s} \|\eta'_{T_j}\|_2 + (1-\alpha)^{-1}\epsilon.$$

Adding $\|\eta'_{\mathcal{T}^c}\|_2$ on both sides and applying the triangle inequality, we get,

$$\begin{aligned}
\|\eta\|_2 &\leq \|\eta_{\mathcal{T}} + (G(z_1) - G(z_2))_{T^c}\|_2 + \|\eta'_{\mathcal{T}^c}\|_2 \\
&\leq ((1-\alpha)^{-1}(1+\alpha) + 1)\sum_{j=2}^{s} \|\eta'_{\mathcal{T}_j}\|_2 + \delta' + C_1\epsilon. \qquad (18)
\end{aligned}$$

For any $i \geq 1$, $j_1 \in \mathcal{T}_{i+1}$,and $j_2 \in \mathcal{T}_i$ we have $|\eta'_{j_1}| \leq |\eta'_{j_2}|$ which in turn implies that $|\eta'_{j_1}| \leq (bl)^{-1}\|\eta'_{\mathcal{T}_i}\|_1$. Squaring and adding the inequalities for all such indices in T_i and T_{i+1}, we get,

$$\|\eta'_{i+1}\|_2 \leq (bl)^{-1/2}\|\eta'_i\|_1.$$

Substituting the result we obtained above in Eq. (18), we get,

$$\|\eta\|_2 - \delta' - C_1\epsilon \leq (bl)^{-1/2}((1-\alpha)^{-1}(1+\alpha) + 1)\sum_{j=1}^{s} \|\eta'_{\mathcal{T}_j}\|_1 = (bl)^{-1/2}(C_0 + 1)\|\eta'_{\mathcal{T}_0^c}\|_1$$

finishing the proof. □

Lemma 1 follows directly from Lemma 3 and Lemma 4 after substituting $a = 1$ and $b = 2$.

A.2. Lemma 2

Recall that random Gaussian matrices satisfy RIP and S-REC properties with high probability (Candès & Tao, 2005; Bora et al., 2017). For completeness and notation, we restate these facts before proving Lemma 2.

Fact 1. *Let $A \in \mathbb{R}^{m\times n}$ be a random Gaussian matrix with each entry sampled i.i.d. from $\mathcal{N}(0, 1/m)$. $\alpha \in (0,1)$. For*

$$m = \Omega\left(\frac{l}{\alpha^2}\log(n/l)\right),$$

A satisfies RIP(l, α) with probability at least $1 - e^{-\Omega(\alpha^2 m)}$.

Fact 2. *Let $A \in \mathbb{R}^{m\times n}$ be a random Gaussian matrix with each entry sampled i.i.d. from $\mathcal{N}(0, 1/m)$. Let $G : \mathbb{R}^k \to \mathbb{R}^n$ be an L-Lipschitz function and define $B^k(r) = \{z : \|z\|_2 \leq r\}$ to be the l_2 norm ball. For*

$$m = \Omega\left(\frac{k}{\alpha^2}\log\left(\frac{Lr}{\delta}\right)\right),$$

A satisfies S-REC($G(B^k(r)), 1-\alpha, \delta$) with probability at least $1 - e^{-\Omega(\alpha^2 m)}$.

Note the proofs of the next two results basically involve small modifications in the proofs presented in Bora et al. (2017) at a few key places to extend them from the setting of the range of the generative model G to the set $S_{l,G}$.

Proof. We will use the mathematical constructs of ϵ-nets for proving the lemma. Let M be a δ/L-net for $B^k(r)$. Then there exists a net such that,

$$\log(|M|) \leq k\log\left(\frac{Lr}{\delta}\right).$$

As this net is δ/L-cover for $B^k(r)$, we will have that $G(M)$ is a δ-cover of $G(B^k(r))$.
For any two points $z_1, z_2 \in B^k(r)$ we can find points $z_1', z_2' \in M$ such that distance in l_2 norm between $G(z_1)$ and $G(z_1')$ is less than δ (similarly for $G(z_2)$ and $G(z_2')$). Now consider some set of indices I of size l and ν be an l-sparse vector with support I (that is all elements outside the indices in I are zero). Using the triangle inequality, we get,

$$\begin{aligned}\|G(z_1) - G(z_2) + \nu\|_2 &\leq \|G(z_1) - G(z_1')\|_2 + \|G(z_1') - G(z_2') + \nu\|_2 + \|G(z_2') - G(z_2)\|_2 \\ &\leq \|G(z_1') - G(z_2') + \nu\|_2 + 2\delta.\end{aligned}$$

Again using the triangle inequality, we have,

$$\|AG(z_1') - AG(z_2') + A\nu\|_2 \leq \|AG(z_1') - AG(z_1)\|_2 + \|AG(z_1) - AG(z_2) + A\nu\|_2 + \|AG(z_2) - AG(z_2')\|_2.$$

From Lemma 8.3 in Bora et al. (2017), we have $\|AG(z_1') - AG(z_2)\|_2 = \mathcal{O}(\delta)$, and $\|AG(z_2) - AG(z_2')\|_2 = \mathcal{O}(\delta)$ with probability $1 - e^{-\Omega(m)}$. Applying this to the previous inequality gives us,

$$\|AG(z_1') - AG(z_2') + A\nu\|_2 \leq \|AG(z_1) - AG(z_2) + A\nu\|_2 + \mathcal{O}(\delta).$$

We note for fixed z_1', z_2' and ν varying over points with support I, $G(z_1') - G(z_2') + \nu$ lie in a subspace of size at most $l + 1$ (*i.e.,* the subspace generated by $G(z_1') - G(z_2')$ and the basis for the subspace with support I). Using the machinery of oblivious subspace embeddings, we get,

$$(1-\alpha)\|G(z_1') - G(z_2') + \nu\|_2 \leq \|AG(z_1') - AG(z_2') + A\nu\|_2$$

will hold with probability $1 - e^{-\Omega(\alpha^2 m)}$ when $m = \Omega(l/\alpha^2)$. We take a union bound over all choices of z_1', z_2' and choices of I (choosing l indices from n). Let the number of choices be N. Using the simple bound $\binom{n}{l} \leq \left(\frac{ne}{l}\right)^l$ we have:

$$\log(N) \leq 2\log(|M|) + l\log\left(\frac{en}{l}\right) \leq 2k\log\left(\frac{Lr}{\delta}\right) + l\log\left(\frac{en}{l}\right).$$

Now we conclude when,

$$m = \Omega\Big(\frac{1}{\alpha^2}\Big(k\log\left(\frac{Lr}{\delta}\right) + l\log\left(\frac{n}{l}\right)\Big)\Big),$$

the following holds with probability $1 - e^{-\Omega(\alpha^2 m)}$ for all $z_1, z_2 \in B^k(r)$ and $\nu \in S_l$ (the set of l-sparse vectors),

$$(1-\alpha)\|G(z_1) - G(z_2) + \nu\|_2 \leq \|A(G(z_1) - G(z_2) + \nu)\|_2 + \mathcal{O}(\delta).$$

The $\mathcal{O}(\delta)$ can be scaled so that we just have δ there and that would not affect the bound on m in the form it is stated. □

Finally, we note that Theorem 1 follows directly from the statements of Lemma 1 and Lemma 2.

A.3. Theorem 2

We first restate the full statement of Theorem 2 for completeness:

Theorem 2. ***(restated)*** *Let $G : \mathbb{R}^k \to \mathbb{R}^n$ be a neural network of depth d. For any $\alpha \in (0, 1)$, $l > 0$, let $A \in \mathbb{R}^{m\times n}$ be a random Gaussian matrix with*

$$m = \Omega\Big(\frac{1}{\alpha^2}\Big((k + l)d \log c + (k + l) \log(n/l)\Big)\Big).$$

rows of i.i.d. entries scaled such that $A_{i,j} \sim N(0, 1/m)$. Let Δ be the decoder satisfying Lemma 1. Then, we have with $1 - e^{-\Omega(\alpha^2 m)}$ probability,

$$\|x - \Delta(Ax + \epsilon)\|_2 \leq (2l)^{-1/2} C_0 \sigma_{l,G}(x) + C_1 \epsilon_{\max} + \delta'$$

for all $x \in \mathbb{R}^n, \|\epsilon\|_2 \leq \epsilon_{\max}$, where $C_0 = 2((1 + \alpha)(1 - \alpha)^{-1} + 1), C_1 = 2(1 - \alpha)^{-1}$, and $\delta' = \delta(1 - \alpha)^{-1}$.

The proof technique for Corollary 2 is closely related to Matouek (2002) and Bora et al. (2017). We provide a geometrical proof sketch and refer the reader to the above works for further details.

Proof. Each individual layer of a neural network function G consists of at most c hyperplanes and the ReLU unit gets activated whenever the input of the previous layer crosses these hyperplanes. This implies that the partitions made by the hyperplanes on the input space of the previous layer describe regions where the function is defined by single matrix. From Lemma 8.3 of (Bora et al., 2017), the number of such partitions is at most $\mathcal{O}(c^k)$. Hence, the total number of partitions from the output space to the input space across d-layers will be $\mathcal{O}(c^{kd})$. Consequently, the range of G will be a union of $\mathcal{O}(c^{kd})$ possibly truncated faces of dimension k in $\mathbb{R}^n$.

Now if we consider the Minkowski sum $S_{l,G}$, then we observe that this set will be a union of $\mathcal{O}(c^{kd}(n/l)^l)$ possibly truncated face of dimension $k + l$. Consider any two faces in $S_{l,G}$. The space defined by the difference of vectors (one from each face) will be part of a subspace of size $2k + 2l + 1$. This is because each face can be parameterized as $v_0 + \sum t_i b_i$ where v_0 is fixed, b_i is a basis for this face, and t_i are the parameters. Hence the difference of two faces will have the same parametrization with at most $2k + 2l$ basis vectors and a fixed point. Adding the fixed point to the basis gives us the required subspace.

Finally, we use oblivious subspace embeddings to note that a random Gaussian matrix A with each entry sampled from $\mathcal{N}(0, 1/m)$ leads to a subspace emebedding with distortion α with a probability of $1 - e^{-\Omega(\alpha^2 m)}$ if $m = \Omega((k + l)/\alpha^2)$. Since there are $\mathcal{O}(c^{kd}(n/l)^l)$ such faces we take a union bound over all pairs of them to see that A satisfies S-REC($S_{l,G}, (1 - \alpha)^{-1}, 0$) with probability $1 - c^{2kd}(n/l)^{2l} e^{-\Omega(\alpha^2 m)}$. This implies that if we have,

$$m = \Omega\left(\frac{1}{\alpha^2}\Big((k + l)(d \log c + \log(n/l))\Big)\right)$$

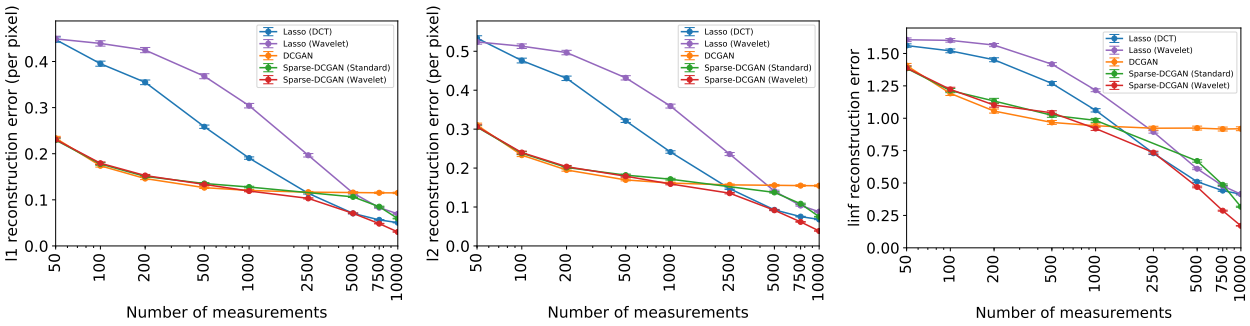
then A satisfies S-REC($S_{l,G}, (1 - \alpha)^{-1}, 0$) with probability $1 - e^{-\Omega(\alpha^2 m)}$ finishing the proof.

□

B. Architectures and hyperparameter details

For both the MNIST and Omniglot dataset, the network architecture was fixed to $784 - 500 - 500 - 20$ for both the generative network and the inference network (except for the final layer of the inference network which has 40 units since both the mean and the variance of the Gaussian variational posterior are learned). Learning is done using Adam (Kingma & Ba, 2015) with a learning rate of 0.001.

For LASSO-based recovery the signal recovery algorithms/libraries were from CVXOPT (Andersen et al., 2013). The Adam (Kingma & Ba, 2015) implementations in Tensorflow (Abadi et al., 2016) was used for generative model-based recovery and Sparse-Gen recovery. The step sizes were selected by evaluating different step sizes via grid search over a held-out validation set (distinct from the held-out test set for which the scores are reported). The recovery procedure was run 10 times each for the generative model based method and the Sparse-gen method, the value with the smallest measurement error is then returned.



(a) CelebA - l_1 (b) CelebA - l_2 (c) CelebA - l_∞

Figure 5. Reconstruction error in terms of l_1 (**left**), l_2 (**center**), and l_∞ (**right**) norms for the CelebA datasets. The performance of Sparse-DCGAN is competitive with DCAN for low measurements and it matches Lasso at high measurements as expected.

C. Additional results for CelebA dataset

For CelebA we train models with the DCGAN (Radford et al., 2015) architecture using adversarial training. The step-sizes were chosen again using grid search with a held-out validation set. As natural images are not sparse in the standard basis we use basis vectors obtained from wavelets and discrete cosine transform for LASSO and Sparse-Gen (called Sparse-DCGAN here). The graphs show that the trends are similar to the MNIST and Omniglot experiments. Sparse-DCGAN shows comparable performance to DCGAN for low measurements and does better than LASSO and DCGAN as the number of measurements increase. The wavelet basis works better than the DCT basis for Sparse-DCGAN.