

---

# Uncertainty Modeling in Generative Compressed Sensing

---

Yilang Zhang <sup>\*1</sup> Mengchu Xu <sup>\*1</sup> Xiaojun Mao <sup>2</sup> Jian Wang <sup>1</sup>

## Abstract

Compressed sensing (CS) aims to recover a high-dimensional signal with structural priors from its low-dimensional linear measurements. Inspired by the huge success of deep neural networks in modeling the priors of natural signals, generative neural networks have been recently used to replace the hand-crafted structural priors in CS. However, the reconstruction capability of the generative model is fundamentally limited by the range of its generator, typically a small subset of the signal space of interest. To break this bottleneck and thus reconstruct those out-of-range signals, this paper presents a novel method called CS-BGM that can effectively expands the range of generator. Specifically, CS-BGM introduces uncertainties to the latent variable and parameters of the generator, while adopting the variational inference (VI) and maximum a posteriori (MAP) to infer them. Theoretical analysis demonstrates that expanding the range of generators is necessary for reducing the reconstruction error in generative CS. Extensive experiments show a consistent improvement of CS-BGM over the baselines.

## 1. Introduction

Compressed sensing (CS) is a paradigm that recovers an unknown signal  $\mathbf{x} \in \mathbb{R}^n$  from a small number of linear measurements

$$\mathbf{y} = \mathbf{Ax} + \mathbf{n}, \quad (1)$$

where  $\mathbf{A} \in \mathbb{R}^{m \times n}$  ( $m \ll n$ ) is the measurement or sensing matrix, and  $\mathbf{n} \in \mathbb{R}^m$  is the additive white Gaussian noise (AWGN). Since the data acquisition of CS allows for a sampling rate far below the Nyquist rate, it provides

<sup>\*</sup>Equal contribution <sup>1</sup>School of Data Science, Fudan University, Shanghai, China <sup>2</sup>School of Mathematical Sciences, Shanghai Jiao Tong University, Shanghai, China. Correspondence to: Xiaojun Mao and Jian Wang <maoxj@sjtu.edu.cn; jian.wang@fudan.edu.cn>.

an efficient approach for data encoding and compression. In particular, a broad range of real-world applications fall into the field of CS. Examples include magnetic resonance imaging (MRI) (Lustig et al., 2007), computed tomography (CT) (Chen et al., 2008), wireless channel estimation (Haupt et al., 2010), super-resolution (Fang et al., 2016), and single-pixel camera (Duarte et al., 2008).

Generally speaking, to reconstruct  $\mathbf{x}$  from the ill-posed linear system (1), additional information is needed to ensure a unique solution. A widely adopted approach is to make use of the sparsity prior of natural signals in some transform domains. For example, natural sounds and images are often observed to be sparse under Fourier and wavelet transformation, respectively. Under the sparsity assumption, traditional CS seeks to find out the sparsest signal  $\mathbf{x}$  that fits the measurements  $\mathbf{y}$ :

$$\min \|\mathbf{x}\|_0, \quad \text{s.t. } \mathbf{y} = \mathbf{Ax} + \mathbf{n}, \quad (2)$$

where  $\|\mathbf{x}\|_0 := |\{x_i | x_i \neq 0\}|$  is the  $\ell_0$ -norm (number of non-zero elements) of  $\mathbf{x}$ . It has been shown that if the sensing matrix  $\mathbf{A}$  satisfies some conditions such as the restricted isometry property (RIP) or restricted eigenvalue condition (REC), CS guarantees robust recovery of  $k$ -sparse signals (Candès & Tao, 2005). However, searching the sparsest solution to a linear system is known to be an NP-hard problem (Natarajan, 1995), for which no polynomial-time algorithm has been developed yet. In order to get a solution within feasible time, therefore, existing CS algorithms often employ greedy search principles, or make some relaxations to problem (2). Representative examples include matching pursuit (MP)-type algorithms (Tropp & Gilbert, 2007; Chen et al., 1989; Needell & Tropp, 2009) and optimization-based algorithms (Chen et al., 2001; Candès & Tao, 2005).

Although traditional CS algorithms have achieved great success in many applications, due to the fact that most natural signals are not strictly sparse in given transform domains, relying on sparsity as the sole prior for reconstruction could lead to inaccurate results. Fortunately, natural signals often possess a variety of features besides sparsity (Kim et al., 2020). Hence, an alternate for CS reconstruction is to combine sparsity with additional refined priors of signals, such as low-rank assumption (Fazel et al., 2008), total variation (Candes et al., 2006), and dictionary models (Shen et al., 2015). Despite their progress in both practical appli-

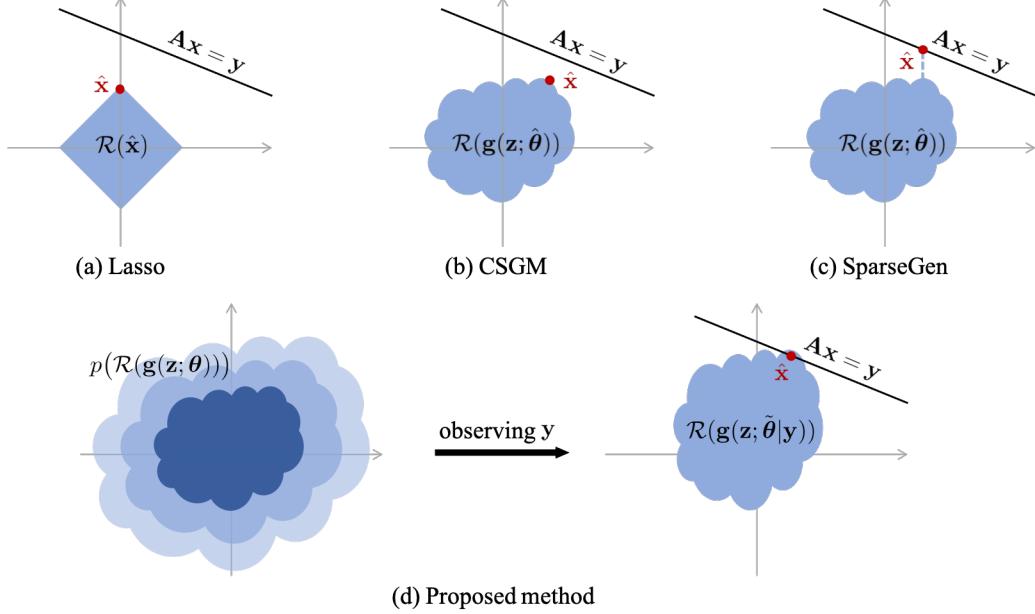


Figure 1. Comparative illustration between Lasso, CSGM, Sparse-Gen and the proposed CS-BGM method.

cations and theoretical guarantees, these hand-crafted priors can only cover a limited range of natural signals, which are hard to generalize to signals from various sources.

To address this generalization issue, Bora et al. (2017) proposed to perform CS using generative models (CSGM). In a nutshell, CSGM uses a generator, which is a pre-trained unsupervised generative model (e.g., variational autoencoder (VAE) (Kingma & Welling, 2014) or generative adversarial network (GAN) (Goodfellow et al., 2014)), as the distributional prior of the training data. Parameterized by  $\theta$ , the generator  $g$  that maps a low-dimensional latent variable  $\mathbf{z} \in \mathbb{R}^k$  to a high-dimensional signal  $g(\mathbf{z}; \theta) \in \mathbb{R}^n$  ( $k \ll n$ ) is trained to output signals resembling those in the training set. The pre-trained generator models the conditional distribution  $q(\mathbf{x}|\mathbf{z}, \theta) = \delta_D[\mathbf{x} - g(\mathbf{z}; \theta)]$ , where  $\delta_D[\cdot]$  is Dirac delta function, and  $\mathbf{z}$  is typically set as standard Gaussian  $\mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I})$ . Hence,  $q(\mathbf{x}) := \int \int q(\mathbf{x}|\mathbf{z}, \theta)p(\theta)p(\mathbf{z})d\theta d\mathbf{z}$  is called a generative prior.

While testing, CSGM fixes  $\theta$  to be the point estimation  $\hat{\theta}$  obtained from training, and optimizes the latent variable  $\mathbf{z}$  to produce an estimation  $\hat{\mathbf{x}} = g(\hat{\mathbf{z}}; \hat{\theta})$  that has the minimum fitting error to the test measurements, i.e.,

$$\hat{\mathbf{z}} = \arg \min_{\mathbf{z}} \|\mathbf{A}g(\mathbf{z}; \hat{\theta}) - \mathbf{y}\|_2^2. \quad (3)$$

This optimization can be performed via gradient descent (GD) or other search methods. Empirically, using generative models leads to a substantial gain in the reconstruction accuracy over conventional CS algorithms like Lasso (Candès & Tao, 2005), especially when the number of measurements is

small (Whang et al., 2021). Moreover, theoretical analyses have been established for CSGM; see (Hand & Voroninski, 2018; Kamath et al., 2020; Liu & Scarlett, 2020).

Following CSGM, plenty of studies have recently been developed to leverage the power of generative models in CS. On the one hand, much effort has been made to accelerate the optimization process in the test phase. Shah & Hegde (2018) proposed to use projected gradient descent (PGD) in the signal space to speed up convergence, for which provable guarantees were derived. Raj et al. (2019) showed that a network-based PGD (NPGD) is empirically faster. Wu et al. (2019) utilized meta-learning to learn the initialization of  $\theta$  as well as the measurement matrix  $\mathbf{A}$ , so that the test optimization can be completed within fewer GD iterations. Asim et al. (2020) used a flow-based invertible model where  $\mathbf{z}$  can be computed inversely from  $\mathbf{x}$ .

On the other hand, it is reported in (Bora et al., 2017) that CSGM often presents inferior performance when the test signals lie outside the range  $\mathcal{R}(g(\mathbf{z}; \theta)) := \{g(\mathbf{z}; \theta) | \mathbf{z} \in \mathbb{R}^k\}$  of the generator. Research has been done to ameliorate the reconstruction quality for these out-of-range signals. In (Dhar et al., 2018), an extra sparse deviation vector was introduced to expand the range of the generator (which is called Sparse-Gen). This approach has been extended by Yang et al. (2021) using non-convex  $\ell_q$ -norm. Furthermore, it has been shown in (Kabkab et al., 2018; Kim et al., 2020) that better reconstruction performances can be achieved via training generative models in supervised fashions (i.e., with access to  $\mathbf{y}$ ). Due to the use of  $\mathbf{y}$ , how-

ever, the supervised modifications only work with a fixed measurement matrix  $\mathbf{A}$ . As a result, retraining is required when  $\mathbf{A}$  changes. Besides, it was shown in (Daras et al., 2021) that optimizing the high-dimensional intermediate layer (in addition to the low-dimensional latent variable) helps improve the expressiveness of the generator. This framework of CSGM has also been applied to MRI, where posterior sampling via Langevin dynamics was employed to generate higher-quality reconstructions (Jalal et al., 2021).

In this paper, in order to effectively reconstruct those out-of-range signals, we propose to expand the range of the generator in CSGM by modeling uncertainties in the network parameter  $\theta$ . Indeed, when the generative models are trained using empirical loss minimization, the uncertainties emerge naturally because the finite training data are imperfect estimations of the signal space of interest. To perform reconstruction (i.e., inference of  $\mathbf{x}$ ), we use variational inference and maximum a posteriori (MAP) for the low-dimensional latent variable  $\mathbf{z}$  and high-dimensional parameter  $\theta$ , respectively. By doing so, the range of the generator is adjusted “adaptively” to embrace those out-of-range test signals. We thus refer to our method as CS with Bayesian generative model (CS-BGM). A comparative illustration between Lasso, CSGM, Sparse-Gen and the proposed method is given in Fig. 1. Our work will justify the necessity of adjusting the range of the generative model for fundamentally improving the reconstruction quality. Also, numerical experiments will show the superiority of the proposed method over the baselines.

Similar to (Bora et al., 2017), we only consider the case where the noise  $\mathbf{n}$  is assumed Gaussian. Whereas, we also mention that the proposed method can be readily generalized to non-Gaussian cases by applying techniques in e.g., (Jalal et al., 2020; Whang et al., 2020).

## 2. Preliminaries

Before proceeding to introduce the proposed CS-BGM, we review several important notions and definitions. The following two properties constrain the eigenvalues of the measurement matrix, which are commonly adopted to establish recovery guarantees for the sparsity-based CS algorithms.

**Definition 2.1** (REC). A matrix  $\mathbf{A} \in \mathbb{R}^{m \times n}$  is said to satisfy the restricted eigenvalue condition (REC) of order  $K$ , provided there exists some constant  $\gamma > 0$  such that

$$\|\mathbf{Ax}\|_2^2 \geq \gamma \|\mathbf{x}\|_2^2 \quad (4)$$

holds for all  $K$ -sparse vector  $\mathbf{x} \in \mathbb{R}^n$ .

**Definition 2.2** (RIP). A matrix  $\mathbf{A} \in \mathbb{R}^{m \times n}$  is said to satisfy the restricted isometry property (RIP) of order  $K$ , provided there exists some constant  $\delta \in (0, 1)$  such that

$$(1 - \delta) \|\mathbf{x}\|_2^2 \leq \|\mathbf{Ax}\|_2^2 \leq (1 + \delta) \|\mathbf{x}\|_2^2 \quad (5)$$

holds for all  $K$ -sparse vector  $\mathbf{x} \in \mathbb{R}^n$ . The minimum  $\delta$  satisfying (5) is called restricted isometry constant (RIC), which is denoted as  $\delta_K(\mathbf{A})$ , or  $\delta_K$  for notational simplicity.

While REC requires that all the eigenvalues of  $\mathbf{A}^* \mathbf{A}$  should be greater than  $\gamma$ , RIP further restricts them being inside the interval  $[1 - \delta, 1 + \delta]$ . From another perspective,  $\gamma_K$  measures the distance between the space of all  $K$ -sparse vectors and the null space of  $\mathbf{A}$ , and  $\delta_K$  is an indicator of how Euclidean norms of the  $K$ -sparse vectors can be preserved under the transform  $\mathbf{A}$  (Dhar et al., 2018). It is well-known that many random matrices satisfy the RIP and REC with high probability when the number  $m$  of measurements scales almost linearly with  $K$ . For example, it has been shown in (Candès & Tao, 2005; Baraniuk et al., 2008) that a random Gaussian matrix  $\mathbf{A}$  whose entries are drawn i.i.d. from  $\mathcal{N}(\mathbf{A}_{ij}; 0, \frac{1}{m})$  obeys the RIP with order- $K$  RIC  $\delta_K$  with overwhelming probability if  $m = \mathcal{O}(\frac{K}{\delta_K^2} \log \frac{n}{K})$ .

To generalize the recovery guarantees from the sparsity-based prior to the generative prior, Bora et al. (2017) introduced the set-restricted eigenvalue condition (S-REC).

**Definition 2.3** (Bora et al. 2017, S-REC). Let  $S \subseteq \mathbb{R}^n$  be a set. For some parameters  $\gamma > 0$  and  $\epsilon > 0$ , a matrix  $\mathbf{A} \in \mathbb{R}^{m \times n}$  is said to satisfy the set-restricted eigenvalue condition S-REC( $S, \gamma, \epsilon$ ), provided that

$$\|\mathbf{A}(\mathbf{x}_1 - \mathbf{x}_2)\|_2 \geq \gamma \|\mathbf{x}_1 - \mathbf{x}_2\|_2 - \epsilon, \quad \forall \mathbf{x}_1, \mathbf{x}_2 \in S. \quad (6)$$

Compared to REC that is defined with  $K$ -sparse vectors, S-REC can be applied to non-sparse vectors by choosing the set  $S$  properly. Thus, it can be used in the generative models by setting  $S$  as the range of the generator. Besides, there is an extra slack variable  $\epsilon$  for achieving the generalization. Similar to the REC, S-REC can also be satisfied by common random matrices with overwhelming probability when  $m$  scales almost linearly with  $k$ , leading to the reconstruction guarantees for the generative model; see (Bora et al., 2017).

## 3. The CS-BGM Method

In this section, we will explain how CS-BGM expands the range of the generator of CSGM to reconstruct the out-of-range signals. The key idea is to consider the uncertainties of the model parameters. We will also theoretically justify the necessity of expanding the range of generator.

### 3.1. A Bayesian View of Generative CS

We start with a Bayesian view of the generative CS, which makes it easier to increase the flexibility of the generative model. The training stage of CS-BGM is similar to that of CSGM. Specifically, the generator is trained such that the generative prior  $q(\mathbf{x})$  resembles the data distribution  $p(\mathbf{x})$  of the training signals. For example, it has been shown in

GAN (Goodfellow et al., 2014) that

$$\hat{\theta} = \arg \min_{\theta} \text{JSD}(q(\mathbf{x}|\theta; \mathbf{X}_{\text{tr}}) \| p(\mathbf{x}; \mathbf{X}_{\text{tr}})), \quad (7)$$

where JSD is the Jensen–Shannon divergence, and  $\mathbf{X}_{\text{tr}}$  is the matrix collecting all the training signals. In fact,  $\hat{\theta}$  is a point estimation  $p(\theta; \mathbf{X}_{\text{tr}}) \approx p(\theta; \hat{\theta}) = \delta_D[\theta - \hat{\theta}]$  from the Bayesian view. As a result, the generative prior  $q(\mathbf{x}; \mathbf{X}_{\text{tr}}) = \iint q(\mathbf{x}|\mathbf{z}, \theta)p(\theta; \mathbf{X}_{\text{tr}})p(\mathbf{z})d\theta d\mathbf{z} \approx q(\mathbf{x}; \hat{\theta})$  serves as a good approximation of the training signal distribution  $p(\mathbf{x}; \mathbf{X}_{\text{tr}})$ .

In testing, after observing the test measurements  $\mathbf{y}$ , we infer the conditional expectation of the signal through

$$\begin{aligned} \mathbb{E}[\mathbf{x}|\mathbf{y}; \mathbf{X}_{\text{tr}}] &\approx \int \mathbf{x} q(\mathbf{x}|\mathbf{y}; \mathbf{X}_{\text{tr}}) \\ &= \iint q(\mathbf{x}|\mathbf{z}, \theta)p(\mathbf{z}, \theta|\mathbf{y}; \mathbf{X}_{\text{tr}})d\mathbf{z}d\theta d\mathbf{x} \\ &= \mathbb{E}_{p(\mathbf{z}, \theta|\mathbf{y}; \mathbf{X}_{\text{tr}})}[\mathbf{g}(\mathbf{z}; \theta)]. \end{aligned} \quad (8)$$

where the last equality follows from  $q(\mathbf{x}|\mathbf{z}, \theta) = \delta_D[\mathbf{x} - \mathbf{g}(\mathbf{z}; \theta)]$ . Thus, what remains is to infer the posterior

$$p(\mathbf{z}, \theta|\mathbf{y}; \mathbf{X}_{\text{tr}}) \propto p(\mathbf{y}|\mathbf{z}, \theta)p(\theta; \mathbf{X}_{\text{tr}})p(\mathbf{z}), \quad (9)$$

where we have used the Bayes' rule.

For CSGM, the test prior (i.e. training posterior) for  $\theta$  is taken to be the point estimation  $p(\theta; \mathbf{X}_{\text{tr}}) \approx \delta_D[\theta - \hat{\theta}]$ , and  $\mathbf{z}$  is taken to be its MAP estimation:

$$\begin{aligned} \hat{\mathbf{z}} &= \arg \max_{\mathbf{z}} p(\mathbf{z}, \theta|\mathbf{y}; \mathbf{X}_{\text{tr}}) \\ &\approx \arg \max_{\mathbf{z}} p(\mathbf{z}, \hat{\theta}|\mathbf{y}; \mathbf{X}_{\text{tr}}) \\ &= \arg \min_{\mathbf{z}} -\log p(\mathbf{y}|\mathbf{z}; \hat{\theta}) - \log p(\mathbf{z}) \\ &= \arg \min_{\mathbf{z}} \|\mathbf{A}\mathbf{g}(\mathbf{z}; \hat{\theta}) - \mathbf{y}\|_2^2 + \lambda_{\mathbf{z}}\|\mathbf{z}\|_2^2, \end{aligned} \quad (10)$$

where  $\lambda_{\mathbf{z}}$  is the relative weights for the prior of  $\mathbf{z}$ . As a result, the conditional expectation (8) in this case is given by  $\mathbb{E}[\mathbf{x}|\mathbf{y}; \mathbf{X}_{\text{tr}}] \approx \mathbb{E}_{p(\mathbf{z}, \theta|\mathbf{y}; \mathbf{X}_{\text{tr}})}[\mathbf{g}(\mathbf{z}; \theta)] \approx \mathbf{g}(\hat{\mathbf{z}}; \hat{\theta})$ . Notice that there is an extra term  $\lambda_{\mathbf{z}}\|\mathbf{z}\|_2^2$  compared with (3). In fact, precisely due to this extra term, an improvement on the performance of CSGM has been reported in (Bora et al., 2017), which well matches with our analysis. However, since the training set contains only finite samples from the signal space, directly using the point estimations  $\hat{\theta}$  and  $\hat{\mathbf{z}}$  will ignore the uncertainties in the test prior  $p(\theta; \mathbf{X}_{\text{tr}})$  and posterior  $p(\mathbf{z}, \theta|\mathbf{y}; \mathbf{X}_{\text{tr}})$ . In particular, since the parameter  $\theta$  is deterministic after training, it will result in a fixed range  $\mathcal{R}(\mathbf{g}(\mathbf{z}; \hat{\theta}))$  of the generator, leading to inferior reconstruction quality of test signals outside the range.

Different from CSGM, therefore, we will take these uncertainties into consideration in the following. However, due

### Algorithm 1 Alternate optimization for CS-BGM

---

**Input:** Sensing matrix  $\mathbf{A} \in \mathbb{R}^{m \times n}$ , alternation number  $L$ , iteration numbers  $R_{\mathbf{z}}, R_{\theta}$ , learning rates  $\eta_{\mathbf{z}}, \eta_{\theta}$ , pre-trained model parameters  $\hat{\theta}$ , measurements  $\mathbf{y} \in \mathbb{R}^n$ .

**Initialization:** Initialize  $\mathbf{v}$  randomly,  $\tilde{\theta} \leftarrow \hat{\theta}$ .

**repeat**

- $\mathbf{v}^{(0)} \leftarrow \mathbf{v}$ ;
- for**  $r$  from 1 to  $R_{\mathbf{z}}$  **do**

  - $\mathbf{v}^{(r)} \leftarrow \mathbf{v}^{(r-1)} + \eta_{\mathbf{z}} \frac{\partial \text{ELBO}}{\partial \mathbf{v}^{(r-1)}}$  using MC sampling;

- end for**
- $\mathbf{v} \leftarrow \mathbf{v}^{(R_{\mathbf{z}})}$ ;
- $\theta^{(0)} \leftarrow \tilde{\theta}$ ;
- for**  $r$  from 1 to  $R_{\theta}$  **do**

  - $\theta^{(r)} \leftarrow \theta^{(r-1)} - \eta_{\theta} \frac{\partial \mathcal{L}(\theta^{(r-1)}, \mathbf{y})}{\partial \theta^{(r-1)}}$  using MC sam-
  - pling;

- end for**
- $\tilde{\theta} \leftarrow \theta^{(R_{\theta})}$ ;

**until** do  $L$  times

**Output:** Image estimation  $\mathbb{E}_{q(\mathbf{z}|\mathbf{v})}[\mathbf{g}(\mathbf{z}; \tilde{\theta})]$ .

---

to the high-dimensionality of  $\theta$  and the non-linearity of the neural network, it is difficult to directly infer, or even approximately infer  $p(\theta; \mathbf{X}_{\text{tr}})$  during training. Nevertheless, considering that the training and test signals are sampled from the same signal space, their distribution should be empirically close to each other. Hence, with a mild modification to  $\hat{\theta}$ , we can hopefully find a new generator whose range can well cover the test signals. Specifically, we equip  $\theta$  with a multivariate Gaussian prior  $p(\theta; \mathbf{X}_{\text{tr}}) = \mathcal{N}(\theta; \hat{\theta}, \lambda \mathbf{I})$ , where  $\lambda$  is a hyperparameter determined by the discrepancy between training and testing.

To infer the joint posterior of  $\mathbf{z}$  and  $\theta$  which are dependent of each other (cf. (9)), we propose to solve for them with *alternating optimization*. Moreover, we choose to infer  $\mathbf{z}$  and  $\theta$  using variational inference (VI) and MAP estimation, respectively. To be more specific, we perform the following two processes iteratively:

- i) fixing  $p(\theta|\mathbf{y}; \mathbf{X}_{\text{tr}}) \approx \delta[\theta - \hat{\theta}]$  to be the MAP estimation, find a surrogate variational posterior  $q(\mathbf{z}|\mathbf{v})$  that minimizes the Kullback-Leibler (KL) divergence  $\text{KL}(q(\mathbf{z}|\mathbf{v}) \| p(\mathbf{z}|\mathbf{y}; \hat{\theta}))$ , where  $\mathbf{v}$  is the variational parameter to be optimized. It is well-known (see e.g., (Blei et al., 2017)) that minimizing this KL divergence above is equivalent to maximizing the evidence lower bound

$$\text{ELBO} := \mathbb{E}_{q(\mathbf{z}|\mathbf{v})} [\log p(\mathbf{y}|\mathbf{z}; \hat{\theta})] - \text{KL}(q(\mathbf{z}|\mathbf{v}) \| p(\mathbf{z})); \quad (11)$$

- ii) fixing  $q(\mathbf{z}|\mathbf{v})$ , find a MAP estimator  $\tilde{\theta}$  that maximizes the posterior  $p(\theta|\mathbf{z}, \mathbf{y}; \mathbf{X}_{\text{tr}}) \propto p(\mathbf{y}|\mathbf{z}, \theta)p(\theta; \mathbf{X}_{\text{tr}})$ . Or

equivalently, minimize the loss

$$\mathcal{L}(\boldsymbol{\theta}, \mathbf{y}) := \mathbb{E}_{q(\mathbf{z}|\mathbf{v})} [\|\mathbf{A}\mathbf{g}(\mathbf{z}; \boldsymbol{\theta}) - \mathbf{y}\|_2^2] + \lambda_{\boldsymbol{\theta}} \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|_2^2, \quad (12)$$

where  $\lambda_{\boldsymbol{\theta}}$  is the relative weight trading-off the data fidelity and the prior of  $\boldsymbol{\theta}$  during testing.

It should be noted that the dimension of  $\boldsymbol{\theta}$  is much larger than  $\mathbf{z}$ , so it would require a considerable number of Monte Carlo (MC) samples to perform VI on it. For practical consideration, we therefore choose MAP rather than VI. Indeed, since the reconstruction results are more sensitive to the generator parameters  $\boldsymbol{\theta}$  compared to the latent variable  $\mathbf{z}$ , the posterior of  $\boldsymbol{\theta}$  will have much smaller variance, and thus can be well approximated by a MAP estimation. Finally, the signal estimated by CS-BGM is given by  $\mathbb{E}[\mathbf{x}|\mathbf{y}; \mathbf{X}_{\text{tr}}] \approx \mathbb{E}_{p(\mathbf{z}, \boldsymbol{\theta}|\mathbf{y}; \mathbf{X}_{\text{tr}})}[\mathbf{g}(\mathbf{z}; \boldsymbol{\theta})] \approx \mathbb{E}_{q(\mathbf{z}|\mathbf{v})}[\mathbf{g}(\mathbf{z}; \tilde{\boldsymbol{\theta}})]$ . The alternating optimization strategy is summarized in Alg. 1.

Alternatively, one can use the MAP estimation (instead of VI) of  $\mathbf{z}$ , i.e., (10), in the alternate optimization. Interestingly, this modification leads to a variant of the proposed CS-BGM without MC sampling, thus shortening the reconstruction time. Experiments in Sec. 4 suggest that this variant also achieves a satisfactory performance.

So far, we have explained how CS-BGM expands the range of generator via uncertainty modeling to include the out-of-range signals. We mention that adjusting the parameters of generative models has also been suggested in CS with untrained neural networks (Ulyanov et al., 2018). However, the proposed method mainly differs in two aspects. Firstly, this work focuses on CS using *random measurements*, while untrained neural networks were used in inverse problems such as denoising, super-resolution, and inpainting removal, for which a rough estimation of the original signal can be obtained directly from the *downsampled or perturbed measurements*. Secondly, the training process of the generative model provides a *strong prior* for  $\boldsymbol{\theta}$  in the proposed method. With this prior term, we are only allowed to modify  $\boldsymbol{\theta}$  slightly; cf. (12). Whereas in CS with untrained neural networks, there is *no prior* so that  $\boldsymbol{\theta}$  is randomly initialized and can change significantly. In the next section, we will justify that expanding the range of the generative model is necessary for fundamentally reducing the reconstruction error.

### 3.2. Why Expanding the Range of Generator?

Although theoretical guarantees have been well developed for CSGM, empirical tests suggest that a large measurement error per pixel (i.e.,  $\frac{1}{n} \|\mathbf{A}\mathbf{g}(\hat{\mathbf{z}}; \boldsymbol{\theta}) - \mathbf{y}\|_2^2$ ) can still be observed on the test set even after optimizing over  $\mathbf{z}$  (Bora et al., 2017); see also Fig. 2 for such an example. In the context of CS, to reduce the reconstruction error  $\|\mathbf{g}(\hat{\mathbf{z}}; \hat{\boldsymbol{\theta}}) - \mathbf{x}\|_2^2$ , there are two major approaches in general.

- i) One is by adding additional priors of  $\mathbf{x}$  to refine the reconstruction results. As  $\hat{\mathbf{z}}$  is already the minimizer of the measurement error in CSGM, these priors help to find a  $\mathbf{z}'$  such that  $\|\mathbf{A}\mathbf{g}(\mathbf{z}'; \hat{\boldsymbol{\theta}}) - \mathbf{y}\|_2^2 \geq \|\mathbf{A}\mathbf{g}(\hat{\mathbf{z}}; \hat{\boldsymbol{\theta}}) - \mathbf{y}\|_2^2$  but  $\|\mathbf{g}(\mathbf{z}'; \hat{\boldsymbol{\theta}}) - \mathbf{x}\|_2^2 \leq \|\mathbf{g}(\hat{\mathbf{z}}; \hat{\boldsymbol{\theta}}) - \mathbf{x}\|_2^2$ .
- ii) The other is by expanding the range of reconstruction model to cover more signals. Ablation tests by Bora et al. (2017) showed that the representative capability of the generative model is mainly limited by its range, although no theoretical evidence has been available.

Our analysis is motivated from these two approaches. We will show theoretically that as long as the measurement error is large, the reconstruction error can never be significantly reduced, no matter what additional prior is imposed. This is because the reconstruction error is essentially lower bounded by its measurement error. Therefore, to fundamentally reduce the reconstruction error, breaking the range limit of the generative model and thus enhancing its representative capability is necessary.

To formalize our analysis, we first give a useful framework called the set-restricted isometry property (S-RIP).

**Definition 3.1** (S-RIP). Let  $S \subseteq \mathbb{R}^n$  be a set. For some parameters  $\delta \in (0, 1)$  and  $\epsilon > 0$ , a matrix  $\mathbf{A} \in \mathbb{R}^{m \times n}$  is said to satisfy the S-RIP( $S, \delta, \epsilon$ ), provided that

$$\begin{aligned} \sqrt{1-\delta} \|\mathbf{x}_1 - \mathbf{x}_2\|_2 - \epsilon &\leq \|\mathbf{A}(\mathbf{x}_1 - \mathbf{x}_2)\|_2 \\ &\leq \sqrt{1+\delta} \|\mathbf{x}_1 - \mathbf{x}_2\|_2 + \epsilon, \quad \forall \mathbf{x}_1, \mathbf{x}_2 \in S. \end{aligned}$$

While the definition of S-RIP resembles that of S-REC, it further requires the measurement distance  $\|\mathbf{A}(\mathbf{x}_1 - \mathbf{x}_2)\|_2$  to be upper bounded by the signal distance  $\|\mathbf{x}_1 - \mathbf{x}_2\|_2$ . Precisely, the upper bound is of vital importance to our analysis.

Next, we show that with high probability, the S-RIP can be satisfied by random Gaussian matrices.

**Lemma 3.2.** Let  $\mathbf{g} : \mathbb{R}^k \mapsto \mathbb{R}^n$  be  $L$ -Lipschitz. Let  $B^k(r) = \{\mathbf{z} | \mathbf{z} \in \mathbb{R}^k, \|\mathbf{z}\|_2 \leq r\}$  be an  $\ell_2$ -norm ball in  $\mathbb{R}^k$ . And denote by  $\mathbf{g}(B^k(r)) := \{\mathbf{g}(\mathbf{z}; \boldsymbol{\theta}) | \mathbf{z} \in B^k(r)\}$  the range of the generator restricted to inputs from the  $\ell_2$ -norm ball. For  $\delta \in (0, 1)$ , if

$$m = \Omega\left(\frac{k}{\delta^2} \log \frac{Lr}{\epsilon}\right), \quad (13)$$

then a random matrix  $\mathbf{A} \in \mathbb{R}^{m \times n}$  with i.i.d. entries  $\mathcal{N}(\mathbf{A}_{ij}; 0, \frac{1}{m})$  satisfies the S-RIP( $\mathbf{g}(B^k(r)), \delta, \epsilon$ ) with probability exceeding  $1 - e^{-\Omega(\delta^2 m)}$ .

The proofs are left to the appendices. This lemma suggests that, with  $m$  increasing almost linearly in  $k$ , the S-RIP can be satisfied with overwhelming probability in  $\mathbf{g}(B^k(r))$ . It

Original	CSGM	CS-BGM
CSGM	CS-BGM	
Measurement error $\frac{1}{n} \ \mathbf{A}\mathbf{g}(\hat{\mathbf{z}}; \theta) - \mathbf{y}\ _2^2:$	1.073	0.047
Reconstruction error $\frac{1}{n} \ \mathbf{g}(\hat{\mathbf{z}}; \theta) - \mathbf{x}\ _2^2:$	0.0137	0.0044

Figure 2. An example of reconstructing an RGB face image of  $64 \times 64 \times 3$ , where CS-BGM can effectively reduce the measurement and reconstruction errors compared to CSGM.

should be noted that although the Lipschitz constant  $L$  of the generator varies for different network architecture, its effect is much weaker than that of  $k$  due to the logarithm term. We also remark that in practice,  $\mathbf{z}$  could hardly take values corresponding to tiny prior probabilities, otherwise the regularization terms in Eqs. (10) and (11) would become too large. Empirically, when those  $\mathbf{z}$  are given as inputs, the decoder outputs can be interpreted as “extrapolation”, which are of inferior qualities compared to “interpolation” ones. Therefore,  $\mathbf{z}$  is generally required to have a bounded norm  $r$ ; see the rationale in (Bora et al., 2017).

Then, with the aid of Lemma 3.2, we have the following theorem, which applies to the entire range of the generator.

**Theorem 3.3.** Let  $\mathbf{g} : \mathbb{R}^k \mapsto \mathbb{R}^n$  be a  $d$ -layer neural network, where each layer is composed by a linear transform and a point-wise non-linearity. Also, denote by  $\mathcal{R}(\mathbf{g}(\mathbf{z}; \theta)) := \{\mathbf{g}(\mathbf{z}; \theta) | \mathbf{z} \in \mathbb{R}^k\}$  the range of the generator. Suppose that there are at most  $c$  nodes per layer, and all the non-linearities are piecewise linear with at most two pieces. Then, for  $\delta \in (0, 1)$ , if

$$m = \Omega\left(\frac{kd}{\delta^2} \log c\right), \quad (14)$$

a random matrix  $\mathbf{A} \in \mathbb{R}^{m \times n}$  with i.i.d. entries  $\mathcal{N}(\mathbf{A}_{ij}; 0, \frac{1}{m})$  satisfies the S-RIP( $\mathcal{R}(\mathbf{g}(\mathbf{z}; \theta)), \delta, \epsilon$ ) with probability exceeding  $1 - e^{-\Omega(\delta^2 m)}$ .

Finally, we show that a necessary condition for a generator to cover the test signal  $\mathbf{x}$  is that a small measurement error can be achieved by optimizing over  $\mathbf{z}$ . Consider an ideal generator  $\mathbf{g}(\mathbf{z}; \theta^*)$  such that  $\min_{\mathbf{z}} \|\mathbf{g}(\mathbf{z}; \theta^*) - \mathbf{x}\| = 0$  with minimizer denoted by  $\mathbf{z}^*$ . Then, we know from Theorem 3.3 that if  $m = \Omega\left(\frac{kd}{\delta^2} \log c\right)$ , the Gaussian measurement matrix  $\mathbf{A}$  with i.i.d. entries  $\mathcal{N}(\mathbf{A}_{ij}; 0, \frac{1}{m})$  satisfies the

S-RIP( $\mathcal{R}(\mathbf{g}(\mathbf{z}; \theta^*)), \delta, \epsilon$ ) with probability of  $1 - e^{-\Omega(\delta^2 m)}$ . Also, let  $\hat{\mathbf{z}} = \arg \min_{\mathbf{z}} \|\mathbf{A}\mathbf{g}(\mathbf{z}; \theta^*) - \mathbf{y}\|_2^2$ , then we have

$$\begin{aligned} \|\mathbf{A}\mathbf{g}(\hat{\mathbf{z}}; \theta^*) - \mathbf{y}\|_2 &\leq \|\mathbf{A}\mathbf{g}(\mathbf{z}^*; \theta^*) - \mathbf{y}\|_2 \\ &= \|\mathbf{A}\mathbf{g}(\mathbf{z}^*; \theta^*) - (\mathbf{A}\mathbf{x} + \mathbf{n})\|_2 \\ &\leq \|\mathbf{A}(\mathbf{g}(\mathbf{z}^*; \theta^*) - \mathbf{x})\|_2 + \|\mathbf{n}\|_2 \\ &\leq \sqrt{1 + \delta} \|\mathbf{g}(\mathbf{z}^*; \theta^*) - \mathbf{x}\|_2 + \epsilon + \|\mathbf{n}\|_2 \\ &= \epsilon + \|\mathbf{n}\|_2 \end{aligned} \quad (15)$$

with overwhelming probability.<sup>1</sup> In other words, the reconstruction error is essentially lower bounded by its measurement error, which completes the justification.

We mention that our analysis also explains Fig. 2. In CSGM, the minimum measurement error is still large, suggesting that the generator cannot cover the test signal, hence the reconstruction error cannot be small. Whereas for CS-BGM, the measurement error is adequately small so that a small reconstruction error can be possibly achieved.

## 4. Experiments

In this section, we will state the setup of the numerical experiments and evaluate the performance of our method. Comparisons among CS-BGM, LASSO, CSGM, PGD-GAN, and Sparse-Gen will be presented to empirically appraise the recovery capabilities. All experiments are run using Tensorflow (Abadi et al., 2015) on one Intel(R) Xeon(R) Silver 4116 CPU and four GeForce RTX 2080 Ti GPUs. For convenient reproducibility, our codes are available at [https://github.com/347325285/CS\\_BGM](https://github.com/347325285/CS_BGM).

### 4.1. Datasets

We consider two datasets: i) the MNIST handwritten digit dataset (LeCun & Cortes, 2010) and ii) the CelebFaces Attributes (CelebA) dataset (Liu et al., 2015). For the MNIST dataset, each digit is a grayscale image of size  $28 \times 28$ , where the value of each pixel is 0 or 1. For the CelebA dataset, the dimension of each RGB image is cropped to  $64 \times 64 \times 3$  for consistent use. In this case, the input of MNIST is 784-dimensional and sparse. Whereas for CelebA, the input is as high as 12288-dimensional and dense.

### 4.2. Models and Hyperparameters

To ensure that the measurement matrix satisfies the S-RIP with high probability, we consider the random Gaussian measurement matrix with each entry i.i.d. drawn from  $\mathcal{N}(\mathbf{A}_{ij}; 0, \frac{1}{m})$ . We use pre-trained VAE models as in (Bora et al., 2017) on the MNIST dataset. The VAE model is a fully connected neural network with the architecture 20-500-

<sup>1</sup>This doesn't hold with the S-REC, because the latter fails to lower bound the recovery error with the measurement error.

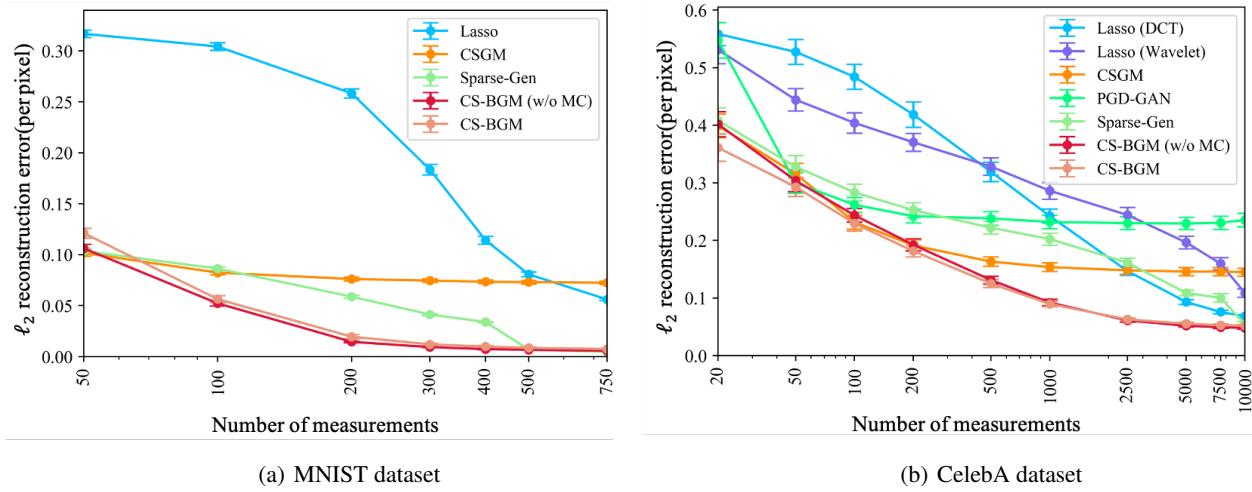


Figure 3. Comparative results among Lasso, CSGM, Sparse-Gen and CS-BGM. The vertical bars are 95% confidence intervals.

500-784, i.e., we generate the digit of size  $28 \times 28$  from a standard normal vector  $\mathbf{z} \in \mathbb{R}^{20}$ . For the experiments on CelebA dataset, we use the pre-trained generative model as in (Bora et al., 2017). The latent representation space for  $\mathbf{z}$  is of dimension 100. For both models, we use Adam optimizer (Kingma & Ba, 2015) with learning rate 0.001 for  $\mathbf{v}$  and 0.002 for  $\theta$ , respectively. In the fast implementation named CS-BGM (w/o MC), which infers  $\mathbf{z}$  with MAP estimation like  $\theta$ , the learning rate for  $\mathbf{z}$  is set to be 0.0011. The regular coefficient  $\lambda_\theta$  is 0.1. To compare the performance of different methods, we choose 1024 digits from the MNIST test dataset and 64 images from CelebA, with a batch size of 64. For the implementations of other comparative algorithms, we use hyperparameters suggested by their authors.

### 4.3. Optimization Strategy

Here, we detail the optimization strategy of the proposed CS-BGM algorithm. Following the common choice, we use a multivariate Gaussian distribution with diagonal covariance matrix as the variational posterior  $q(\mathbf{z}|\mathbf{v})$ , and adopt the reparameterization trick (Kingma & Welling, 2014) when performing VI. The number of MC samples for inference is set to 20 on the MNIST dataset and 10 on the celebA dataset. In our experiments, we perform the optimization of  $\mathbf{z}$  for 2000 iterations and then  $\theta$  for 500 iterations with only 1 alternations. In particular, we set  $J = 1$ ,  $K = 2000$ , and  $L = 500$  to Alg. 1. In the fast implementation, we set  $J = 1$ ,  $K = 500$ , and  $L = 200$  to Alg. 1. We do not perform more alternations or iterations since the loss has already converged under such setting. We stress that optimization with more alternations for different  $m$  could produce better performance under an elaborate parameter selection. For the sake of brevity, however, we do not include such fine-tuning

in our experiments.

#### 4.4. Results

We present our experimental results on the MNIST and CelebA datasets, respectively. For comparison, we choose the per-pixel  $\ell_2$  reconstruction loss ( $\frac{1}{n}\|\mathbf{x} - \hat{\mathbf{x}}\|_2$ ) as our performance metric.

#### 4.4.1. MNIST

According to the previous hyperparameter settings, we construct the corresponding measurement matrix  $\mathbf{A}$  of size  $m \times 784$  (with  $m$  varying from 50 to 750) and compare the reconstructed pixel error between different methods in Fig. 3(a) and Fig. 4. Overall, it can be observed that our proposed method performs the best (i.e., with the lowest  $\ell_2$  loss) when the number  $m$  of measurements is between 100 and 400. When  $m$  is 500 or larger, the performance of Sparse-Gen improves and becomes comparable to that of CS-BGM. This is perhaps because Sparse-Gen optimizes the following problem:

$$\min_{\mathbf{z}, \boldsymbol{\nu}} \|\mathbf{D}\boldsymbol{\nu}\|_1 + \|\mathbf{A}(\mathbf{g}(\mathbf{z}; \boldsymbol{\theta}) + \boldsymbol{\nu}) - \mathbf{y}\|_2^2, \quad (16)$$

where  $D$  is some transform basis, which is set as an identity matrix in MNIST. Thus, when  $m$  becomes larger, the assumption made by Sparse-Gen (i.e.,  $\nu$  being sparse) is more likely to holds.

#### 4.4.2. CELEBA

We test all methods for  $\mathbf{A} \in \mathbb{R}^{m \times 12288}$  with  $m$  varying from 20 to  $10^4$ . As shown in Fig. 3(b), the CS-BGM method has the minimum reconstruction error for all tested region of  $m$ . The gap between CS-BGM and CSGM even increases with

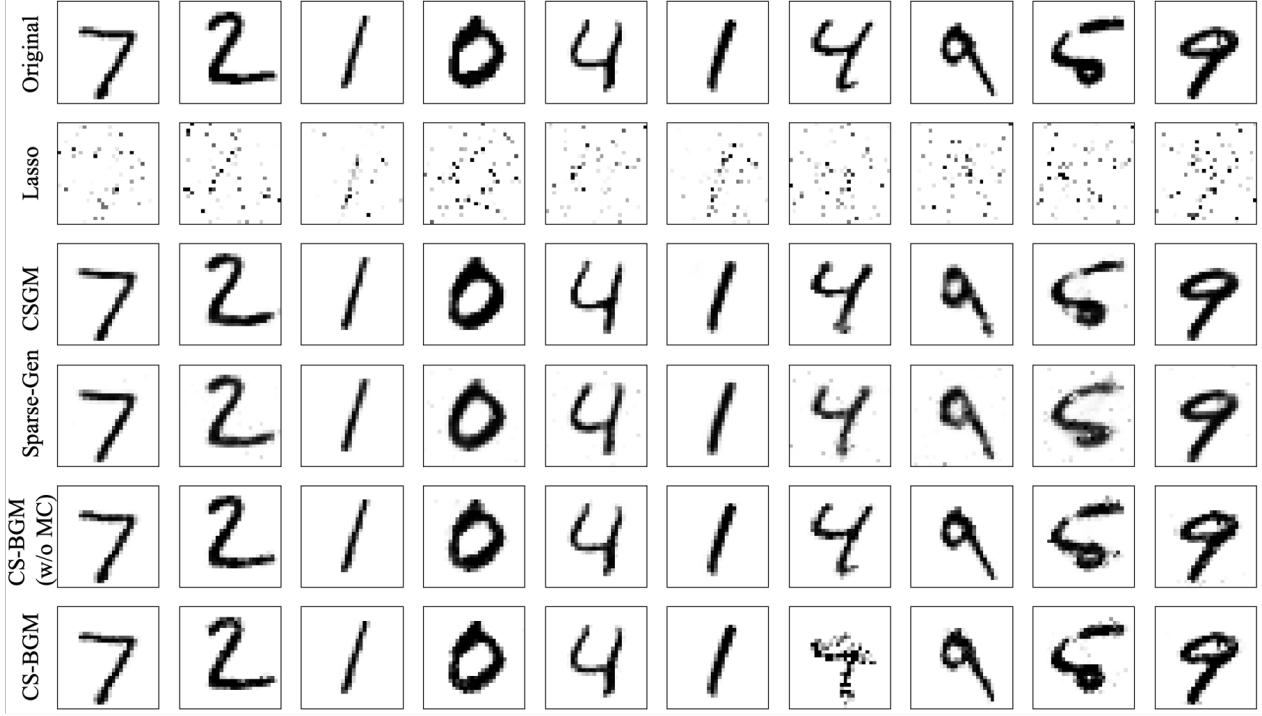


Figure 4. Reconstruction results on MNIST dataset when the measurement number  $m = 100$ .

$m$ , which clearly demonstrates the superiority of expanding the range of generator.

Fig. 5 shows the comparative reconstruction results of face images among different methods on CelebA dataset. Since it has been reported by Dhar et al. (2018) that Sparse-Gen performs better with discrete wavelet transform (DWT) than discrete cosine transform (DCT), we run Sparse-Gen with DWT in our experiments. The number  $m$  of measurements is set to 1000. One can observe that Lasso works poorly in most cases, since the face images (after DCT or DWT) are not sparse enough for accurate reconstruction.

One can also observe that the faces recovered by other methods (CSGM, PGD-GAN, and Sparse-Gen) are more or less deformed or distorted. The main reason is that the distribution of the training images are inconsistent with that of the test ones, so that the latter cannot be reconstructed with the learned prior. Indeed, most of these distorted faces are combinations of the elements from the training images, while they are not similar in detail to the test ones. This observation confirms the ‘‘inconsistency’’ of the distributions. In addition, the residual (i.e.  $\mathbf{x} - \hat{\mathbf{x}}$ ) is not strictly sparse under the hand-crafted transform domain, and hence could not be well modeled using a sparse  $\nu$ . In comparison, CS-BGM can accurately restore the details of the test images, as it can adaptively adjust the range of the generator to promote the reconstruction quality.

## 5. Conclusion and Future Work

In this paper, we have proposed a method called CS-BGM for fundamentally reducing the reconstruction error of signals in CS with generative models. Our analysis has shown that the quality of reconstructed images is limited by the measurement error, regardless of what additional prior is imposed. Nevertheless, by introducing uncertainties, CS-BGM can effectively break through the range limit of representative capability for the generator, thus offering big potential for quality improvement in generative image reconstruction. Empirically experiments on common datasets have demonstrated that the performance gains brought by expanding the range of the generator are indeed nontrivial.

Our future work will focus on several feasible directions. First, we will try different priors (e.g., sparsity) on  $\theta$  in CS-BGM, or learn a prior for  $\theta$  from the training data. Second, as mentioned, the upper bound of the reconstruction error is dominated by the slack constant  $\epsilon$  when the measurement error is adequately small. To further refine the reconstruction results, therefore, attention will be cast to the reduction of  $\epsilon$ , with possible use of additional priors on the signal  $\mathbf{x}$ . Third, we will try to generalize the proposed method to larger and more sophisticated models such as StyleGAN (Karras et al., 2019). Since the complexity of CS-BGM is dominated by the alternating optimization between  $\mathbf{z}$  and  $\theta$ , a potential solution is to train a neural network (e.g., LSTM) to infer their joint posterior given  $\mathbf{y}$ . Thus, the burdensome optimization

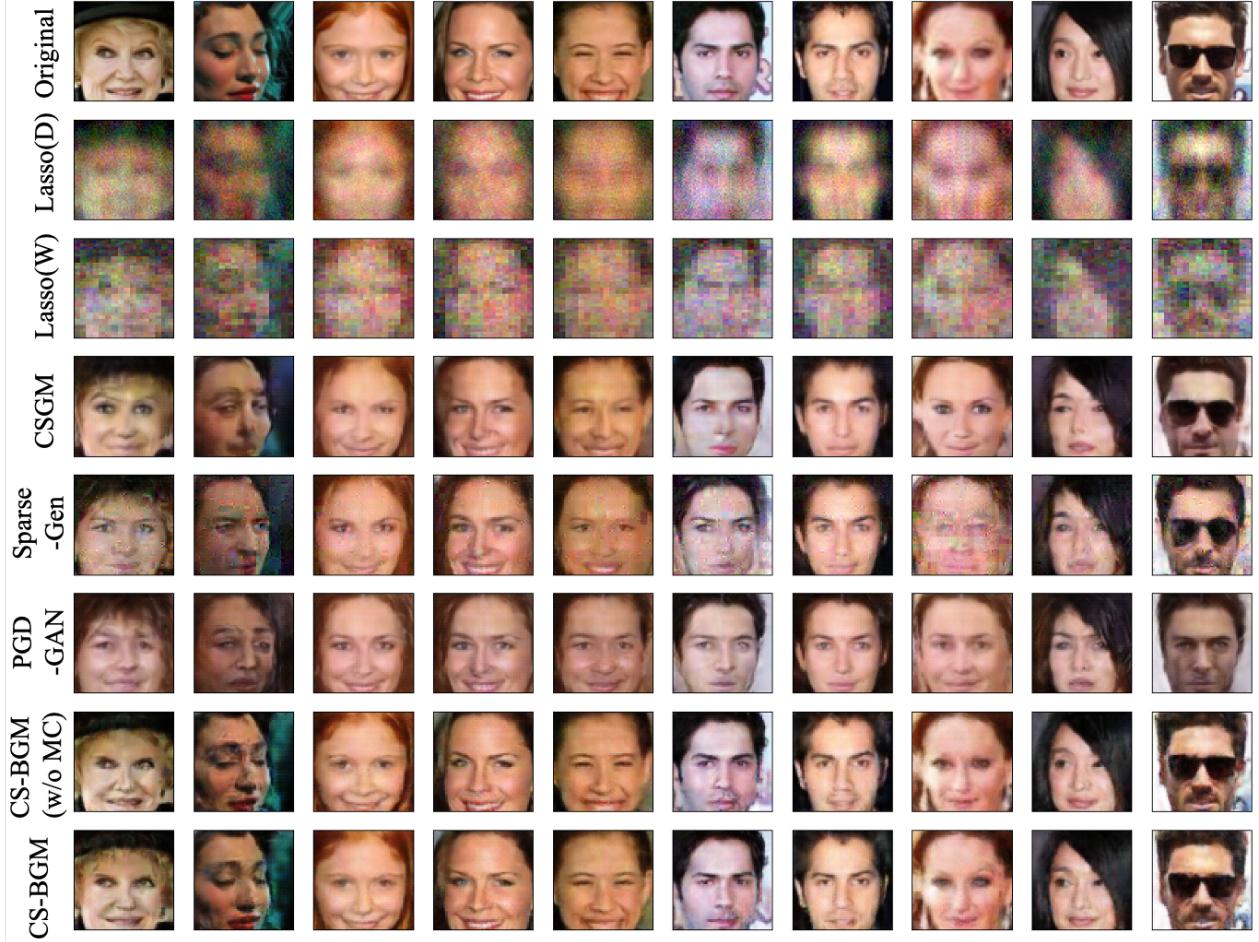


Figure 5. Reconstruction results on CelebA dataset when the measurement number  $m = 1000$ . Lasso (C) and Lasso (W) mean that we perform Lasso on the DCT and DWT, respectively. Sparse-Gen is performed on the wavelet basis as well.

process can be substituted by a simple forward mapping.

## Acknowledgements

Shanghai Municipal Science and Technology Major Project (2018SHZDZX01); National Natural Science Foundation of China (12001109, 92046021, 61971146); Science and Technology Commission of Shanghai Municipality grant 20dz1200600; ZJ Lab and Shanghai Center for Brain Science and Brain-Inspired Technology; Innovation Cross and Cooperation Team Project of Chinese Academy of Sciences (JCTD-2020-15).

## References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Lev-

enberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

Asim, M., Daniels, M., Leong, O., Ahmed, A., and Hand, P. Invertible generative models for inverse problems: mitigating representation error and dataset bias. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 399–409. PMLR, 13–18 Jul 2020.

Baraniuk, R., Davenport, M., DeVore, R., and Wakin, M. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28(3): 253–263, 2008. Communicated by Emmanuel J. Candès.

- Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.
- Bora, A., Jalal, A., Price, E., and Dimakis, A. G. Compressed sensing using generative models. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 537–546. PMLR, 06–11 Aug 2017.
- Candès, E. and Tao, T. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.
- Candes, E., Romberg, J., and Tao, T. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.
- Chen, G.-H., Tang, J., and Leng, S. Prior image constrained compressed sensing (piccs): A method to accurately reconstruct dynamic ct images from highly undersampled projection data sets. *Medical Physics*, 35(2):660–663, 2008.
- Chen, S., Billings, S. A., and Luo, W. Orthogonal least squares methods and their application to non-linear system identification. *International Journal of Control*, 50 (5):1873–1896, 1989.
- Chen, S. S., Donoho, D. L., and Saunders, M. A. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1): 129–159, 2001.
- Daras, G., Dean, J., Jalal, A., and Dimakis, A. Intermediate layer optimization for inverse problems using deep generative models. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2421–2432. PMLR, 18–24 Jul 2021.
- Dhar, M., Grover, A., and Ermon, S. Modeling sparse deviations for compressed sensing using generative models. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1214–1223. PMLR, 10–15 Jul 2018.
- Duarte, M. F., Davenport, M. A., Takhar, D., Laska, J. N., Sun, T., Kelly, K. F., and Baraniuk, R. G. Single-pixel imaging via compressive sampling. *IEEE Signal Processing Magazine*, 25(2):83–91, 2008.
- Fang, J., Wang, F., Shen, Y., Li, H., and Blum, R. S. Super-resolution compressed sensing for line spectral estimation: An iterative reweighted approach. *IEEE Transactions on Signal Processing*, 64(18):4649–4662, 2016.
- Fazel, M., Candes, E., Recht, B., and Parrilo, P. Compressed sensing and robust recovery of low rank matrices. In *2008 42nd Asilomar Conference on Signals, Systems and Computers*, pp. 1043–1047, 2008.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.
- Hand, P. and Voroninski, V. Global guarantees for enforcing deep generative priors by empirical risk. In Bubeck, S., Perchet, V., and Rigollet, P. (eds.), *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 970–978. PMLR, 06–09 Jul 2018.
- Haupt, J., Bajwa, W. U., Raz, G., and Nowak, R. Toeplitz compressed sensing matrices with applications to sparse channel estimation. *IEEE Transactions on Information Theory*, 56(11):5862–5875, 2010.
- Jalal, A., Liu, L., Dimakis, A. G., and Caramanis, C. Robust compressed sensing using generative models. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 713–727. Curran Associates, Inc., 2020.
- Jalal, A., Arvinte, M., Daras, G., Price, E., Dimakis, A. G., and Tamir, J. Robust compressed sensing mri with deep generative priors. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 14938–14954. Curran Associates, Inc., 2021. URL <https://proceedings.neurips.cc/paper/2021/file/7d6044e95a16761171b130dcba476a43e-Paper.pdf>.
- Kabkab, M., Samangouei, P., and Chellappa, R. Task-aware compressed sensing with generative adversarial networks. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- Kamath, A., Price, E., and Karmalkar, S. On the power of compressed sensing with generative models. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5101–5109. PMLR, 13–18 Jul 2020.

- Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- Kim, K.-S., Lee, J. H., and Yang, E. Compressed sensing via measurement-conditional generative models. *arXiv preprint arXiv:2007.00873*, 2020.
- Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.
- Kingma, D. P. and Welling, M. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR*, 2014.
- Laurent, B. and Massart, P. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302 – 1338, 2000.
- LeCun, Y. and Cortes, C. MNIST handwritten digit database. 2010.
- Liu, Z. and Scarlett, J. Information-theoretic lower bounds for compressive sensing with generative models. *IEEE Journal on Selected Areas in Information Theory*, 1(1):292–303, 2020.
- Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.
- Lustig, M., Donoho, D., and Pauly, J. M. Sparse mri: The application of compressed sensing for rapid mr imaging. *Magnetic Resonance in Medicine*, 58(6):1182–1195, 2007.
- Natarajan, B. K. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24(2):227–234, 1995.
- Needell, D. and Tropp, J. Cosamp: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, 26(3):301–321, 2009.
- Raj, A., Li, Y., and Bresler, Y. Gan-based projector for faster recovery with convergence guarantees in linear inverse problems. In *Proceedings of International Conference on Computer Vision (ICCV)*, October 2019.
- Shah, V. and Hegde, C. Solving linear inverse problems using gan priors: An algorithm with provable guarantees. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4609–4613, 2018.
- Shen, Y., Li, J., Zhu, Z., Cao, W., and Song, Y. Image reconstruction algorithm from compressed sensing measurements by dictionary learning. *Neurocomputing*, 151: 1153–1162, 2015.
- Tropp, J. A. and Gilbert, A. C. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory*, 53(12):4655–4666, 2007.
- Ulyanov, D., Vedaldi, A., and Lempitsky, V. Deep image prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- Whang, J., Lei, Q., and Dimakis, A. Compressed sensing with invertible generative models and dependent noise. In *NeurIPS 2020 Workshop on Deep Learning and Inverse Problems*, 2020.
- Whang, J., Lei, Q., and Dimakis, A. Solving inverse problems with a flow-based noise model. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11146–11157. PMLR, 18–24 Jul 2021.
- Wu, Y., Rosca, M., and Lillicrap, T. Deep compressed sensing. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6850–6860. PMLR, 09–15 Jun 2019.
- Yang, Y., Wang, H., Qiu, H., Wang, J., and Wang, Y. Non-convex sparse deviation modeling via generative models. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2345–2349, 2021.

## A. Proof of Lemma 3.2

**Lemma A.1** (Restated). *Let  $\mathbf{g} : \mathbb{R}^k \mapsto \mathbb{R}^n$  be  $L$ -Lipschitz. Let  $B^k(r) = \{\mathbf{z} | \mathbf{z} \in \mathbb{R}^k, \|\mathbf{z}\|_2 \leq r\}$  be an  $\ell_2$ -norm ball in  $\mathbb{R}^k$ . And denote by  $\mathbf{g}(B^k(r)) := \{\mathbf{g}(\mathbf{z}; \theta) | \mathbf{z} \in B^k(r)\}$  the range of the generator restricted to inputs from the  $\ell_2$ -norm ball. For  $\delta \in (0, 1)$ , if*

$$m = \Omega\left(\frac{k}{\delta^2} \log \frac{Lr}{\epsilon}\right), \quad (17)$$

*then a random matrix  $\mathbf{A} \in \mathbb{R}^{m \times n}$  with i.i.d. entries  $\mathcal{N}(\mathbf{A}_{ij}; 0, \frac{1}{m})$  satisfies the S-RIP( $\mathbf{g}(B^k(r)), \delta, \epsilon$ ) with probability exceeding  $1 - e^{-\Omega(\delta^2 m)}$ .*

*Proof.* The proof extends that of (Bora et al., 2017, Lemma 4.1) using Laurent-Massart bounds and Johnson–Lindenstrauss lemma.

First, Let  $\mathbf{b}$  be a vector such that

$$b_i = \frac{\sqrt{m}}{\|\mathbf{x}\|_2} (\mathbf{Ax})_i, \quad i = 1, \dots, m.$$

Since the entries of  $\mathbf{A}$  are i.i.d. Gaussian  $\mathcal{N}(\mathbf{A}_{ij}; 0, \frac{1}{m})$ , it is straightforward to verify that

$$b_i \sim \mathcal{N}(0, 1).$$

Thus we have

$$\|\mathbf{b}\|_2^2 \sim \chi^2(m).$$

Using the concentration of measure for chi-squared distribution (Laurent & Massart, 2000), we obtain

$$\mathbb{P}[\|\mathbf{b}\|_2^2 - m \geq 2\sqrt{mt} + 2t] \leq e^{-t} \quad (18)$$

and

$$\mathbb{P}[\|\mathbf{b}\|_2^2 - m \leq -2\sqrt{mt}] \leq e^{-t}. \quad (19)$$

Substituting  $2\sqrt{mt} + 2t$  with  $\delta m$  in (18), and  $2\sqrt{mt}$  with  $\delta m$  in (19), respectively, we get

$$\mathbb{P}[\|\mathbf{b}\|_2^2 \geq (1 + \delta)m] \leq e^{-\frac{m}{2}(\delta+1+\sqrt{2\delta+1})}$$

and

$$\mathbb{P}[\|\mathbf{b}\|_2^2 \leq (1 - \delta)m] \leq e^{-\frac{\delta^2 m}{4}}.$$

Notice that

$$\|\mathbf{Ax}\|_2^2 = \frac{\|\mathbf{x}\|_2^2}{m} \|\mathbf{b}\|_2^2,$$

by the union bound we have for a fixed  $\mathbf{x} \in \mathbb{R}^n$  that

$$\begin{aligned} \mathbb{P}[(1 - \delta)\|\mathbf{x}\|_2^2 \leq \|\mathbf{Ax}\|_2^2 \leq (1 + \delta)\|\mathbf{x}\|_2^2] &\geq 1 - e^{-\frac{m}{2}(\delta+1+\sqrt{2\delta+1})} - e^{-\frac{\delta^2 m}{4}} \\ &= 1 - e^{-\Omega(\delta m)} - e^{-\Omega(\delta^2 m)} \\ &= 1 - e^{-\Omega(\delta^2 m)}, \end{aligned} \quad (20)$$

where the last equation is because  $\delta \in (0, 1)$ .

Then, we construct an  $\frac{\epsilon}{L}$ -net  $N$  on  $B^k(r)$  such that

$$\log |N| \leq k \log \left( \frac{4Lr}{\epsilon} \right).$$

Since  $\mathbf{g}$  is  $L$ -Lipschitz, we know that  $\mathbf{g}(N)$  is an  $\epsilon$ -net of  $\mathbf{g}(B^k(r))$ .

Let  $T := \{\mathbf{g}(\mathbf{z}_1) - \mathbf{g}(\mathbf{z}_2) | \mathbf{z}_1, \mathbf{z}_2 \in N\}$  be the set of pairwise differences for vectors in  $\mathbf{g}(N)$ . Then we have

$$\log |T| \leq \log |N|^2 \leq 2k \log \left( \frac{4Lr}{\epsilon} \right). \quad (21)$$

For any  $\mathbf{z}, \mathbf{z}' \in B^k(r)$ , there exists  $\mathbf{z}_1, \mathbf{z}_2 \in N$  such that  $\mathbf{g}(\mathbf{z}_1), \mathbf{g}(\mathbf{z}_2)$  are  $\epsilon$ -close to  $\mathbf{g}(\mathbf{z}), \mathbf{g}(\mathbf{z}')$ , respectively. Therefore we obtain

$$\begin{aligned} \|\mathbf{g}(\mathbf{z}) - \mathbf{g}(\mathbf{z}')\|_2 &\leq \|\mathbf{g}(\mathbf{z}_1) - \mathbf{g}(\mathbf{z}_2)\|_2 + \|\mathbf{g}(\mathbf{z}) - \mathbf{g}(\mathbf{z}_1)\|_2 + \|\mathbf{g}(\mathbf{z}_2) - \mathbf{g}(\mathbf{z}')\|_2 \\ &\leq \|\mathbf{g}(\mathbf{z}_1) - \mathbf{g}(\mathbf{z}_2)\|_2 + 2\epsilon \end{aligned} \quad (22)$$

and

$$\begin{aligned} \|\mathbf{g}(\mathbf{z}) - \mathbf{g}(\mathbf{z}')\|_2 &\geq \|\mathbf{g}(\mathbf{z}_1) - \mathbf{g}(\mathbf{z}_2)\|_2 - \|\mathbf{g}(\mathbf{z}) - \mathbf{g}(\mathbf{z}_1)\|_2 - \|\mathbf{g}(\mathbf{z}_2) - \mathbf{g}(\mathbf{z}')\|_2 \\ &\geq \|\mathbf{g}(\mathbf{z}_1) - \mathbf{g}(\mathbf{z}_2)\|_2 - 2\epsilon. \end{aligned} \quad (23)$$

Employing (Bora et al., 2017, Lemma 8.2), with probability of  $1 - e^{-\Omega(m)}$ , we get

$$\|\mathbf{A}\mathbf{g}(\mathbf{z}_1; \boldsymbol{\theta}) - \mathbf{A}\mathbf{g}(\mathbf{z}; \boldsymbol{\theta})\|_2 = \mathcal{O}(\epsilon)$$

and

$$\|\mathbf{A}\mathbf{g}(\mathbf{z}_2; \boldsymbol{\theta}) - \mathbf{A}\mathbf{g}(\mathbf{z}'; \boldsymbol{\theta})\|_2 = \mathcal{O}(\epsilon).$$

Therefore, with probability of  $1 - e^{-\Omega(m)}$ , we have

$$\begin{aligned} \|\mathbf{A}\mathbf{g}(\mathbf{z}) - \mathbf{A}\mathbf{g}(\mathbf{z}')\|_2 &\leq \|\mathbf{A}\mathbf{g}(\mathbf{z}_1) - \mathbf{A}\mathbf{g}(\mathbf{z}_2)\|_2 + \|\mathbf{A}\mathbf{g}(\mathbf{z}) - \mathbf{A}\mathbf{g}(\mathbf{z}_1)\|_2 + \|\mathbf{A}\mathbf{g}(\mathbf{z}_2) - \mathbf{A}\mathbf{g}(\mathbf{z}')\|_2 \\ &\leq \|\mathbf{A}\mathbf{g}(\mathbf{z}_1) - \mathbf{A}\mathbf{g}(\mathbf{z}_2)\|_2 + \mathcal{O}(\epsilon) \end{aligned} \quad (24)$$

and

$$\|\mathbf{A}\mathbf{g}(\mathbf{z}) - \mathbf{A}\mathbf{g}(\mathbf{z}')\|_2 \geq \|\mathbf{A}\mathbf{g}(\mathbf{z}_1) - \mathbf{A}\mathbf{g}(\mathbf{z}_2)\|_2 - \mathcal{O}(\epsilon). \quad (25)$$

Substituting  $\mathbf{x}$  in (20) with  $\mathbf{g}(\mathbf{z}_1) - \mathbf{g}(\mathbf{z}_2)$ , it then holds with probability of  $1 - e^{-\Omega(\delta^2 m)}$  that

$$(1 - \delta)\|\mathbf{g}(\mathbf{z}_1) - \mathbf{g}(\mathbf{z}_2)\|_2^2 \leq \|\mathbf{A}\mathbf{g}(\mathbf{z}_1) - \mathbf{A}\mathbf{g}(\mathbf{z}_2)\|_2^2 \leq (1 + \delta)\|\mathbf{g}(\mathbf{z}_1) - \mathbf{g}(\mathbf{z}_2)\|_2^2 \quad (26)$$

for a fixed  $\mathbf{g}(\mathbf{z}_1) - \mathbf{g}(\mathbf{z}_2)$ .

By the union bound over all the vectors in  $T$ , we have (26) holds for  $\forall \mathbf{z}_1, \mathbf{z}_2 \in N$  with probability of

$$1 - 2|T|e^{-\Omega(\delta^2 m)} = 1 - 2e^{-\Omega(\delta^2 m)}$$

where the equality follows from (13) and (21).

Since (24), (25) and (26) are independent with each other, then with probability of

$$(1 - e^{-\Omega(m)})(1 - e^{-\Omega(\delta^2 m)}) = 1 - e^{-\Omega(\delta^2 m)},$$

it holds that

$$\begin{aligned} \|\mathbf{A}\mathbf{g}(\mathbf{z}) - \mathbf{A}\mathbf{g}(\mathbf{z}')\|_2 &\leq \|\mathbf{A}\mathbf{g}(\mathbf{z}_1) - \mathbf{A}\mathbf{g}(\mathbf{z}_2)\|_2 + \mathcal{O}(\epsilon) \\ &\leq \sqrt{1 + \delta}\|\mathbf{g}(\mathbf{z}_1) - \mathbf{g}(\mathbf{z}_2)\|_2 + \mathcal{O}(\epsilon) \end{aligned}$$

and that

$$\begin{aligned} \|\mathbf{A}\mathbf{g}(\mathbf{z}) - \mathbf{A}\mathbf{g}(\mathbf{z}')\|_2 &\geq \|\mathbf{A}\mathbf{g}(\mathbf{z}_1) - \mathbf{A}\mathbf{g}(\mathbf{z}_2)\|_2 - \mathcal{O}(\epsilon) \\ &\geq \sqrt{1 + \delta}\|\mathbf{g}(\mathbf{z}_1) - \mathbf{g}(\mathbf{z}_2)\|_2 - \mathcal{O}(\epsilon), \end{aligned}$$

which is the desired S-RIP( $\mathbf{g}(B^k(r)), \delta, \epsilon$ ).  $\square$

## B. Proof of Theorem 3.3

**Theorem B.1** (Restated). *Let  $\mathbf{g} : \mathbb{R}^k \mapsto \mathbb{R}^n$  be a  $d$ -layer neural network, where each layer is composed by a linear transform and a point-wise non-linearity. And denote by  $\mathcal{R}(\mathbf{g}(\mathbf{z}; \boldsymbol{\theta})) := \{\mathbf{g}(\mathbf{z}; \boldsymbol{\theta}) | \mathbf{z} \in \mathbb{R}^k\}$  the range of the generator. Suppose there are at most  $c$  nodes per layer, and all the non-linearities are piecewise linear with at most two pieces. For  $\delta \in (0, 1)$ , if*

$$m = \Omega\left(\frac{kd}{\delta^2} \log c\right), \quad (27)$$

*then a random matrix  $\mathbf{A} \in \mathbb{R}^{m \times n}$  with i.i.d. entries  $\mathcal{N}(\mathbf{A}_{ij}; 0, \frac{1}{m})$  satisfies the S-RIP( $\mathcal{R}(\mathbf{g}(\mathbf{z}; \boldsymbol{\theta}))$ ,  $\delta, \epsilon$ ) with probability of  $1 - e^{-\Omega(\delta^2 m)}$ .*

*Proof.* From (Bora et al., 2017, Section 8.3) we know that  $\mathcal{R}(\mathbf{g}(\mathbf{z}; \boldsymbol{\theta}))$  is a union of  $\mathcal{O}(c^{kd})$  different  $k$ -dimensional faces ( $k$ -faces) in  $\mathbb{R}^n$ .

For each  $k$ -face  $F \subseteq \mathcal{R}(\mathbf{g}(\mathbf{z}; \boldsymbol{\theta}))$ , applying Lemma 3.2 (which still holds for any compact set in addition to  $\ell_2$ -norm balls), we get that  $\mathbf{A}$  satisfies S-RIP( $F, \delta, \epsilon$ ) with probability of  $1 - e^{-\Omega(\delta^2 m)}$  if

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

Then, for the range  $\mathcal{R}(\mathbf{g}(\mathbf{z}; \boldsymbol{\theta}))$  which is a union of  $\mathcal{O}(c^{kd})$  different  $k$ -faces, the union bound gives that  $\mathbf{A}$  satisfies S-RIP( $F, \delta, \epsilon$ ) with probability exceeding  $1 - \mathcal{O}(c^{kd})e^{-\Omega(\delta^2 m)}$  if

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

Therefore, let

$$m = \Omega\left(\frac{kd}{\delta^2} \log c\right),$$

then  $\mathbf{A}$  can satisfy S-RIP( $F, \delta, \epsilon$ ) with probability of  $1 - e^{-\Omega(\delta^2 m)}$ . The proof is thus completed.  $\square$