Compressed Sensing with Invertible Generative Models and Dependent Noise

Jay Whang
University of Texas at Austin
jaywhang@cs.utexas.edu

Qi Lei
Princeton University
qilei@princeton.edu

Alexandros G. Dimakis
University of Texas at Austin
dimakis@austin.utexas.edu

Abstract

We study image inverse problems with invertible generative priors, specifically normalizing flow models. Our formulation views the solution as the maximum a posteriori (MAP) estimate of the image given the measurements. Our general formulation allows for any differentiable noise model with long-range dependencies as well as non-linear differentiable forward operators. We establish theoretical recovery guarantees for denoising and compressed sensing under our framework. We also empirically validate our method on various inverse problems including 1-bit compressed sensing and denoising with highly structured noise patterns.

1 Introduction

Inverse problems seek to reconstruct an unknown signal from observations (or *measurements*), which are produced by some process that transforms the original signal. Because such processes are often lossy and noisy, inverse problems are typically formulated as reconstructing $\boldsymbol{x}$ from its measurements

$$\boldsymbol{y} = f(\boldsymbol{x}) + \boldsymbol{\delta}, \tag{1}$$

where f is a known deterministic *forward operator* and $\boldsymbol{\delta}$ is an additive noise which may have a complex structure itself. An impressively wide range of applications can be posed under this formulation with an appropriate choice of f and $\boldsymbol{\delta}$, such as compressed sensing [1, 2], computed tomography [3], magnetic resonance imaging (MRI) [4], and phase retrieval [5, 6].

In general for a non-invertible forward operator f, there can be potentially infinitely many signals that match given observations. Thus the recovery algorithm must critically rely on *a priori* knowledge about the original signal to find the most plausible solution among them. Sparsity has classically been a very influential structural prior for various inverse problems [1, 2, 7]. Alternatively, recent approaches introduced deep generative models as a powerful signal prior, showing significant gains in reconstruction quality compared to sparsity priors [8–10].

Our contributions

- We present a general formulation for obtaining maximum a posteriori (MAP) estimation reconstructions for dependent noise and general forward operators. Notably, our method can leverage generative models for both the original image and the noise.
- We empirically show that our method achieves excellent reconstruction in the presence of structured non-Gaussian noise. We also demonstrate the efficacy of our method on inverse problems with non-linear forward operators.
- We provide the initial theoretical characterization of likelihood-based priors for two common linear inverse problems: denoising and compressed sensing. Notably, unlike Bora et al. [8], we make no structural assumptions on the generator and provide bounds that depend on the likelihood of the prior.

NeurIPS 2020 Workshop on Deep Learning and Inverse Problems, virtual.

2 Background

Invertible Generative Models. Invertible generative models, also known as *normalizing flows*, are a class of likelihood-based generative models that approximate complex distributions by mapping a simple noise (e.g. standard Gaussian) through an invertible function G [11, 12]. Since G is invertible, change of variables formula allows us to efficiently compute the log-density of $\boldsymbol{x}$ as $\log p(\boldsymbol{x}) = \log p(\boldsymbol{z}) + \log |\det J_{G^{-1}}(\boldsymbol{x})|$, where $J_{G^{-1}}$ is the Jacobian of G^{-1}. Thus likelihood computation is straightforward as long as the log-determinant term can be efficiently evaluated. Importantly, the invertibility of G guarantees that it has an unrestricted range and can generate samples that are out-of-distribution, albeit at lower probability. This is a key distinction from a GAN, whose generator has a restricted range and cannot even *represent* out-of-distribution samples.

Related Work. While vast literature exists on various inverse problems, the idea of using a deep generative prior was introduced relatively recently by Bora et al. [8]. In that work, the generator from a GAN [13] was shown to be an effective prior for compressed sensing. Several recent studies have investigated different ways to extend this result. For example, Dhar et al. [14] considered allowing sparse deviations on the generator's output and Van Veen et al. [10] showed that the generator can be replaced by an *untrained* Deep Image Prior [15] for imaging tasks. More recently, Wu et al. [16] applied techniques from meta-learning to improve the reconstruction speed, and Ardizzone et al. [17] showed that one can *implicitly* learn the inverse process by modelling the forward process with a flow model, . Asim et al. [9] proposed to use a normalizing flow prior and reported excellent reconstruction performance, especially on out-of-distribution images.

3 Our Method

Notations and Setup. $\|\cdot\|$ denotes ℓ_2 norm, and $\odot$ denotes element-wise multiplication. We assume a differentiable forward operator f and a noise distribution p_Δ (which itself can be a generative model). Thus an observation is generated via $\boldsymbol{y} = f(\boldsymbol{x}) + \boldsymbol{\delta}$ where $\boldsymbol{x} \sim p$ and $\boldsymbol{\delta} \sim p_\Delta$, where $p(\boldsymbol{x})$ is assumed to be a flow model defined by an invertible mapping G.

3.1 MAP Formulation

For a given observation $\boldsymbol{y}$, our method tries to recover $\boldsymbol{x}$ as the MAP estimate of the conditional distribution $p(\boldsymbol{x}|\boldsymbol{y})$ where $p(\boldsymbol{x})$ is given by a flow model defined by an invertible mapping G:

$$\log p(\boldsymbol{x}|\boldsymbol{y}) = \log p(\boldsymbol{y}|\boldsymbol{x}) + \log p(\boldsymbol{x}) + C \triangleq -L_{\text{MAP}}(\boldsymbol{x};\boldsymbol{y}) + C,$$

where $L_{\text{MAP}}(\boldsymbol{x};\boldsymbol{y}) = -\log p_\Delta(\boldsymbol{y} - f(\boldsymbol{x})) - \log p(\boldsymbol{x})$. Recalling that the generative procedure for the flow model is $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, I), \boldsymbol{x} = G(\boldsymbol{z})$, we arrive at the following loss function:

$$L_{\text{G}}(\boldsymbol{z};\boldsymbol{y}) \triangleq -\log p_\Delta(\boldsymbol{y} - f(G(\boldsymbol{z}))) - \log p(G(\boldsymbol{z})) \tag{2}$$

Model Smoothing. Since our objective Equation (2) depends on the density $p(\boldsymbol{x})$ given by the flow model, our recovery of $\boldsymbol{x}$ depends heavily on the quality of density estimates from the model. Unfortunately likelihood-based models exhibit counter-intuitive properties, such as assigning higher density on out-of-distribution examples. [18–21]. Thus we introduce the *smoothing parameter* $\beta \geq 0$ that controls the degree to which we rely on $p(\boldsymbol{x})$:

$$L_{\text{G}}(\boldsymbol{z};\boldsymbol{y},\beta) = -\log p_\Delta(\boldsymbol{y} - f(G(\boldsymbol{z}))) - \beta \log p(G(\boldsymbol{z})) \tag{3}$$

3.2 Prior Work

This paper generalizes the work of Bora et al. [8] and Asim et al. [9], so we describe the methods proposed in those papers in detail. Importantly, we show that their approaches are special cases of our MAP formulation. Note that both papers considered linear inverse problems, i.e. $f(\boldsymbol{x}) = A\boldsymbol{x}$.

GAN Prior: [8] considers the loss $L_{\text{Bora}}(\boldsymbol{z};\boldsymbol{y}) = \|\boldsymbol{y} - AG(\boldsymbol{z})\|^2 + \lambda \|z\|^2$, which tries to project the input $\boldsymbol{y}$ onto the range of the generator G with ℓ_2 regularization on the latent variable. We note that G here is a GAN and is not invertible as it maps a noise to a higher-dimensional vector. Thus the image we wish to recover may not be in the range of G. This inability to represent arbitrary images leads to suboptimal reconstructions, as confirmed by Asim et al. [9] and our experiments.

Flow Prior: [9] consider the loss $L(\boldsymbol{z};\boldsymbol{y}) = \|\boldsymbol{y} - AG(\boldsymbol{z})\|^2 + \gamma \|\boldsymbol{z}\|$, which tries to simultaneously match the observation and regularize $\boldsymbol{z}$ to be close to zero. This is a special case of our MAP loss for isotropic Gaussian noise $\boldsymbol{\delta} \sim \mathcal{N}(\boldsymbol{0}, \gamma I)$ with a volume-preserving flow model p (i.e. the log-determinant term is constant). Continuing from eq. (2), $L_{\text{G}}(\boldsymbol{z};\boldsymbol{y}) = \frac{1}{2\sigma^2}\|\boldsymbol{y} - AG(\boldsymbol{z})\|^2 + \|\boldsymbol{z}\|^2 + C$,

where C absorbs the constant log-determinant term in $\log p(\boldsymbol{x}) = \log p(\boldsymbol{z}) + \log \|\det J_{G^{-1}}(\boldsymbol{x})\|$. This recovers the ℓ_2-regularized version of L_{Asim}.

4 Theoretical Analysis

Here we present a preliminary theoretical analysis for denoising and compressed sensing. Unlike most prior work, we take a probabilistic approach and avoid specific structural assumptions on $\boldsymbol{x}$, such as sparsity or being generated from a low-dimensional Gaussian prior. Detailed proofs can be found in Appendix A.

Recovery Guarantees for Denoising. Suppose we observe $\boldsymbol{y} = \boldsymbol{x}^* + \boldsymbol{\delta}$ with Gaussian noise $\boldsymbol{\delta} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$ with $\|\boldsymbol{\delta}\| = r$. We perform MAP inference by minimizing the following loss with gradient descent: $L_{\text{MAP}}(\boldsymbol{x}) = \frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{x}\|^2 + q(\boldsymbol{x})$, where we write $q(\boldsymbol{x}) \triangleq -\log p(\boldsymbol{x})$.

Theorem 4.1. *Let $\boldsymbol{x}^*$ be a local maximum of the model $p(\boldsymbol{x})$ and $\boldsymbol{y} = \boldsymbol{x}^* + \boldsymbol{\delta}$ be the noisy observation. Assume that q satisfies local strong convexity within the ball around $\boldsymbol{x}^*$ defined as $B_r^d(\boldsymbol{x}^*) \triangleq \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x} - \boldsymbol{x}^*\| \leq r\}$, i.e. the Hessian of q satisfies $H_q(\boldsymbol{x}) \succeq \mu I \ \forall \boldsymbol{x} \in B_r^d(\boldsymbol{x}^*)$ for some $\mu > 0$. Then gradient descent starting from $\boldsymbol{y}$ on the loss function $L_{MAP}(\boldsymbol{x})$ converges to $\bar{\boldsymbol{x}}$, a local minimizer of $L_{MAP}(\boldsymbol{x})$, that satisfies $\|\bar{\boldsymbol{x}} - \boldsymbol{x}^*\| \leq \frac{1}{\mu\sigma^2+1}\|\boldsymbol{\delta}\|$.*

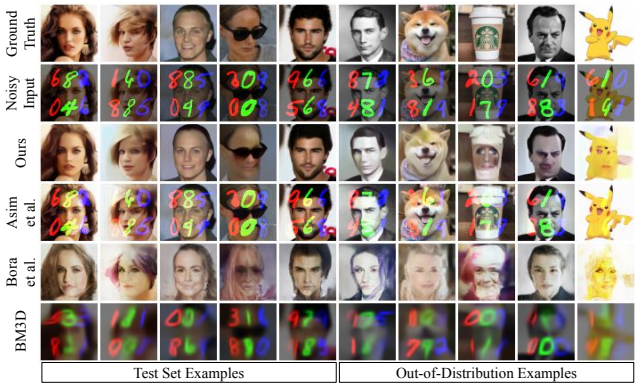
This theorem shows that a well-conditioned model with large μ leads to better denoising and confirms that our MAP formulation encourages reconstructions with high density.

Recovery Guarantees for Compressed Sensing. Here we show a worst-case bound for noisy compressed sensing where we observe $\boldsymbol{y} = A\boldsymbol{x}^* + \boldsymbol{\delta}$ with noise level $\epsilon = \|\boldsymbol{\delta}\|$ and $A \in \mathbb{R}^{m \times d}$ has entries drawn i.i.d. from $\mathcal{N}(0, 1/m)$. Note that the following constrained minimization problem is equivalent to the original unconstrained minimization of eq. (3) for some β via Lagrange multipliers.

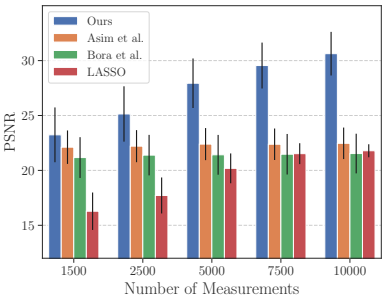
Theorem 4.2. *For a given $\boldsymbol{x}^* \in \mathbb{R}^d$ and $\rho = \log p(\boldsymbol{x}^*)$, define $S(\rho) = \{\boldsymbol{x} | \log p(\boldsymbol{x}) \geq \rho\}$ to be the set of images with density higher than the ground truth image $\boldsymbol{x}^*$. Recall that the observation for $\boldsymbol{x}^*$ is $\boldsymbol{y} = A\boldsymbol{x}^* + \boldsymbol{\delta}$. When we perform the MAP inference by solving the following optimization problem, $\bar{\boldsymbol{x}} = \underset{\boldsymbol{x} \in \mathbb{R}^d, \|A\boldsymbol{x} - \boldsymbol{y}\| \leq \epsilon}{\arg\min} \{-\log p(\boldsymbol{x})\}$, we have: $\mathbb{E}\|\bar{\boldsymbol{x}} - \boldsymbol{x}^*\| \leq \sqrt{8\pi}\left(\frac{w(S(\rho))}{\sqrt{m}} + \epsilon\right)$, where $w(\cdot)$ is the Gaussian mean width, and the expectation is over the randomness of A.*

We note that in high dimensions, Gaussian mean width scales with the set's normalized surface area. Since $S(\rho)$ becomes smaller as ρ increases, this suggests the above error bound is tighter for ground truth images $\boldsymbol{x}^*$ with higher density.

5 Experiments



(a) Result of removing MNIST noise from CelebA-HQ faces. Notice that without any understanding of the complex noise structure, baseline methods fail to produce good reconstructions.



(b) PSNR at different measurement counts (best viewed in color). The approaches by Asim et al. [9] and Bora et al. [8] show little improvements from having more measurements due to their inability to take the noise model into account.

Figure 1: Results from MNIST digit denoising (left) and noisy compressed sensing (right).

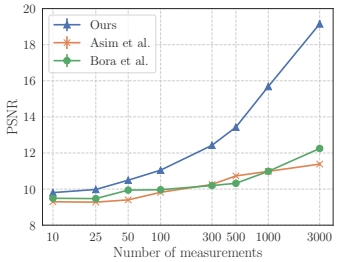
Our experiments are designed to evaluate our MAP formulation in two scenarios: (1) complex noise with dependencies and (2) non-linear forward operator. A detailed description of the datasets, models and hyperparameters are provided in the appendix.

Baseline Methods. We compare our approach to the methods of [8] and [9], as they are two recently proposed approaches that use deep generative prior on inverse problems. When applicable, we also compare against BM3D [22] and LASSO [23] with Discrete Cosine Transform basis as appropriate. We point out that the baselines methods are not designed to make use of the noise distribution, whereas our method does utilize it. Thus, the experiments are not meant to be taken as direct comparisons, but rather as an empirical evidence that the MAP formulation benefits from knowing the noise structure.

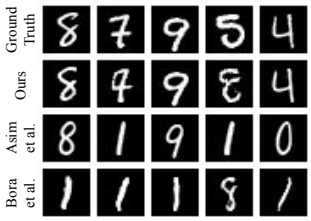
5.1 Results

Denoising MNIST Digits. The measurement process is $\boldsymbol{y} = 0.5 \cdot \boldsymbol{x} + \boldsymbol{\delta}_{\text{MNIST}}$, where $\boldsymbol{\delta}_{\text{MNIST}}$ represents MNIST digits added at different locations and color channels. Each digit itself comes from a flow model trained on the MNIST dataset. As shown in Figure 1a, our method successfully removes MNIST noise. While the method of Bora et al. [8] also removes MNIST digits because its outputs are forced to be in the range of the DCGAN used, the reconstructions are far from the ground truth – particularly on out-of-distribution samples that are not human faces.

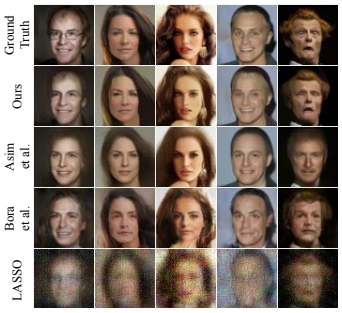
Noisy Compressed Sensing. Now we consider the measurement process $\boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{\delta}_{\text{sine}}$ where $A \in \mathbb{R}^{m \times d}$ is a random Gaussian measurement matrix and $\boldsymbol{\delta}$ has positive mean with variance that follows a sinusoidal pattern (details provided in Appendix B). Figure 1b shows that our method is able to make a better use of additional measurements. Interestingly in Figure 2c, all three methods with deep generative prior produced plausible human faces. However, the reconstructions from Asim et al. [9] and Bora et al. [8] significantly differ from the ground truth images. We posit that this is due to the implicit Gaussian noise assumption made by the two methods, again showing the benefits of explicitly incorporating the knowledge of noise distribution.



(a) Result of 1-bit noisy compressed sensing at different measurement counts. Our method achieves the same reconstruction performance using up to 2× fewer measurements compared to the best baseline method [9].



(b) Reconstructions for noisy 1-bit compressed sensing with 3000 binary measurements. Notice that our method fails more gracefully compared to other methods, i.e. even when the reconstructions differ from the ground truth, substantial parts of the reconstructions are still correct.



(c) Reconstructions for compressed sensing with 2500 random Gaussian measurements.

Figure 2: Experiment results for noisy compressed sensing on CelebA-HQ images.

1-bit Compressed Sensing. This task considers a combination of a non-linear forward operator as well as a non-Gaussian noise. The measurement process is the quantized version of noisy compressed sensing, i.e. $\boldsymbol{y} = \text{sign}(A\boldsymbol{x}) + \boldsymbol{\delta}_{\text{sine}}$. Note that this the most extreme form of quantized compressed sensing, since $\boldsymbol{y}$ only contains the noisy sign $\{+1, -1\}$ of the measurements. Because the gradient of sign function is zero everywhere, we use Straight-Through Estimator [24] for backpropagation. See Figure 2a and Figure 2b for a comparison of our method to the baselines at varying numbers of measurements.

6 Conclusion

We propose a novel method that generalized [9] to solve inverse problems for general differentiable forward operators and structured noise. The power of our approach stems from the flexiblity of invertible generative models which can be combined in a modular way to solve MAP inverse problems in very general settings, as we demonstrate. The central theoretical question that remains open is to analyze the optimization problem we formulated. In this paper we empirically minimize this loss using gradient descent, but some theoretical guarantees would be desirable, possibly under assumptions, e.g. random weights following the framework of [25].

Acknowledgments and Disclosure of Funding

This research has been supported by NSF Grants CCF 1763702,1934932, AF 1901292, 2008710, 2019844 research gifts by Western Digital, WNCG IAP, computing resources from TACC and the Archie Straiton Fellowship. QL is supported by the NSF under Grant #2030859 to the Computing Research Association for the CIFellows Project.

References

[1] Emmanuel J Candes, Justin K Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, 59(8): 1207–1223, 2006.

[2] David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4): 1289–1306, 2006.

[3] Guang-Hong Chen, Jie Tang, and Shuai Leng. Prior image constrained compressed sensing (piccs): a method to accurately reconstruct dynamic ct images from highly undersampled projection data sets. *Medical physics*, 35 2:660–3, 2008.

[4] Michael Lustig, David Donoho, and John M Pauly. Sparse mri: The application of compressed sensing for rapid mr imaging. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 58(6):1182–1195, 2007.

[5] Emmanuel J Candes, Yonina C Eldar, Thomas Strohmer, and Vladislav Voroninski. Phase retrieval via matrix completion. *SIAM review*, 57(2):225–251, 2015.

[6] Emmanuel J Candes, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

[7] Richard G Baraniuk, Volkan Cevher, Marco F Duarte, and Chinmay Hegde. Model-based compressive sensing. *IEEE Transactions on information theory*, 56(4):1982–2001, 2010.

[8] Ashish Bora, Ajil Jalal, Eric Price, and Alexandros G Dimakis. Compressed sensing using generative models. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 537–546. JMLR. org, 2017.

[9] Muhammad Asim, Ali Ahmed, and Paul Hand. Invertible generative models for inverse problems: mitigating representation error and dataset bias. *CoRR*, abs/1905.11672, 2019. URL `http://arxiv.org/abs/1905.11672`.

[10] Dave Van Veen, Ajil Jalal, Mahdi Soltanolkotabi, Eric Price, Sriram Vishwanath, and Alexandros G Dimakis. Compressed sensing with deep image prior and learned regularization. *arXiv preprint arXiv:1806.06438*, 2018.

[11] Esteban G Tabak and Cristina V Turner. A family of nonparametric density estimation algorithms. *Communications on Pure and Applied Mathematics*, 66(2):145–164, 2013.

[12] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *arXiv preprint arXiv:1912.02762*, 2019.

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[14] Manik Dhar, Aditya Grover, and Stefano Ermon. Modeling sparse deviations for compressed sensing using generative models, 2018.

[15] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9446–9454, 2018.

[16] Yan Wu, Mihaela Rosca, and Timothy P. Lillicrap. Deep compressed sensing. *CoRR*, abs/1905.06723, 2019. URL `http://arxiv.org/abs/1905.06723`.

[17] Lynton Ardizzone, Jakob Kruse, Sebastian J. Wirkert, Daniel Rahner, Eric W. Pellegrini, Ralf S. Klessen, Lena Maier-Hein, Carsten Rother, and Ullrich Köthe. Analyzing inverse problems with invertible neural networks. *CoRR*, abs/1808.04730, 2018. URL `http://arxiv.org/abs/1808.04730`.

[18] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? *arXiv preprint arXiv:1810.09136*, 2018.

[19] Hyunsun Choi, Eric Jang, and Alexander A Alemi. Waic, but why? generative ensembles for robust anomaly detection. *arXiv preprint arXiv:1810.01392*, 2018.

[20] Dan Hendrycks and Thomas G Dietterich. Benchmarking neural network robustness to common corruptions and surface variations. *arXiv preprint arXiv:1807.01697*, 2018.

[21] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, and Balaji Lakshminarayanan. Detecting out-of-distribution inputs to deep generative models using a test for typicality. *arXiv preprint arXiv:1906.02994*, 2019.

[22] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising with block-matching and 3d filtering. In *Image Processing: Algorithms and Systems, Neural Networks, and Machine Learning*, volume 6064, page 606414. International Society for Optics and Photonics, 2006.

[23] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[24] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

[25] Paul Hand and Vladislav Voroninski. Global guarantees for enforcing deep generative priors by empirical risk. *IEEE Transactions on Information Theory*, 2020.

[26] Roman Vershynin. Estimation in high dimensions: a geometric perspective. In *Sampling theory, a renaissance*, pages 3–66. Springer, 2015.

[27] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[28] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, pages 10215–10224, 2018.

[29] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

A Omitted Proofs

A.1 Proof for Denoising

Proof of Theorem 4.1. We first show that gradient descent with sufficiently small learning rate will converge to $\bar{\boldsymbol{x}}$, the locally-optimal minimizer of $L_{\text{MAP}}(\boldsymbol{x}) = \frac{1}{\sigma^2}\|\boldsymbol{x}-\boldsymbol{y}\|^2 + q(\boldsymbol{x})$, where we write $q(\boldsymbol{x}) \triangleq -\log p(\boldsymbol{x})$ (we subsume the constant coefficient $\frac{1}{2}$ into $\frac{1}{\sigma^2}$ without loss of generality). Notice in the ball $B_r^d(\boldsymbol{x}^*) := \left\{\boldsymbol{x} \in \mathbb{R}^d \mid \|\boldsymbol{x}-\boldsymbol{x}^*\| \le r\right\}$, L is $\left(\mu + \frac{1}{\sigma^2}\right)$ strongly-convex. We next show there is a stationary point $\bar{\boldsymbol{x}} \in B_r^d(\boldsymbol{x}^*)$ of $L(\boldsymbol{x})$.

$$\begin{aligned}
\nabla L(\bar{\boldsymbol{x}}) = 0 &\Longrightarrow \nabla q(\bar{\boldsymbol{x}}) + \frac{1}{\sigma^2}(\bar{\boldsymbol{x}} - \boldsymbol{y}) = 0 \\
&\Longrightarrow \nabla q(\bar{\boldsymbol{x}}) - \nabla q(\boldsymbol{x}^*) = \frac{1}{\sigma^2}(\boldsymbol{y} - \bar{\boldsymbol{x}}) \\
&\Longrightarrow \langle \nabla q(\bar{\boldsymbol{x}}) - \nabla q(\boldsymbol{x}^*), \bar{\boldsymbol{x}} - \boldsymbol{x}^* \rangle = \frac{1}{\sigma^2}\langle \boldsymbol{y} - \bar{\boldsymbol{x}}, \bar{\boldsymbol{x}} - \boldsymbol{x}^* \rangle
\end{aligned}$$

From strong convexity of q, $\langle \nabla q(\bar{\boldsymbol{x}}) - \nabla q(\boldsymbol{x}^*), \bar{\boldsymbol{x}} - \boldsymbol{x}^* \rangle \ge \mu\|\bar{\boldsymbol{x}} - \boldsymbol{x}^*\|^2$. Thus

$$\begin{aligned}
\frac{1}{\sigma^2}\langle \boldsymbol{y} - \boldsymbol{x}^*, \bar{\boldsymbol{x}} - \boldsymbol{x}^* \rangle &= \frac{1}{\sigma^2}\langle (\boldsymbol{y} - \bar{\boldsymbol{x}}) + (\bar{\boldsymbol{x}} - \boldsymbol{x}^*), \bar{\boldsymbol{x}} - \boldsymbol{x}^* \rangle \\
&= \frac{1}{\sigma^2}\langle \boldsymbol{y} - \bar{\boldsymbol{x}}, \bar{\boldsymbol{x}} - \boldsymbol{x}^* \rangle + \frac{1}{\sigma^2}\langle \bar{\boldsymbol{x}} - \boldsymbol{x}^*, \bar{\boldsymbol{x}} - \boldsymbol{x}^* \rangle \\
&= \langle \nabla q(\bar{\boldsymbol{x}}) - \nabla q(\boldsymbol{x}^*), \bar{\boldsymbol{x}} - \boldsymbol{x}^* \rangle + \frac{1}{\sigma^2}\|\bar{\boldsymbol{x}} - \boldsymbol{x}^*\|^2 \\
&\ge \mu\|\bar{\boldsymbol{x}} - \boldsymbol{x}^*\|^2 + \frac{1}{\sigma^2}\|\bar{\boldsymbol{x}} - \boldsymbol{x}^*\|^2 \\
&= \left(\mu + \frac{1}{\sigma^2}\right)\|\bar{\boldsymbol{x}} - \boldsymbol{x}^*\|^2
\end{aligned}$$

Finally, by Cauchy-Schwartz inequality $\langle \boldsymbol{y} - \boldsymbol{x}^*, \bar{\boldsymbol{x}} - \boldsymbol{x}^* \rangle \le \|\boldsymbol{y} - \boldsymbol{x}^*\| \cdot \|\bar{\boldsymbol{x}} - \boldsymbol{x}^*\|$. So we get $\|\bar{\boldsymbol{x}} - \boldsymbol{x}^*\| \le \frac{1}{1+\mu\sigma^2}\|\boldsymbol{y} - \boldsymbol{x}^*\| \le \|\boldsymbol{\delta}\| \le r$, i.e., $\bar{\boldsymbol{x}} \in B_r^d(\boldsymbol{x}^*)$. Notice L is $\left(\mu + \frac{1}{\sigma^2}\right)$ strongly-convex in $B_r^d(\boldsymbol{x}^*)$, which contains the stationary point $\bar{\boldsymbol{x}}$. Therefore $\bar{\boldsymbol{x}}$ is a local minimizer of $L(\boldsymbol{x})$. Also note that we implicitly require q to be twice differentiable, meaning in a compact set $B_r^d(\boldsymbol{x}^*)$ its smoothness is upper bounded by a constant M. Thus gradient descent starting from $\boldsymbol{y} \in B_r^d(\boldsymbol{x}^*)$ with learning rate smaller than $\frac{1}{M}$ will converge to $\bar{\boldsymbol{x}}$ without leaving the (convex) set $B_r^d(\boldsymbol{x}^*)$. □

A.2 Proof for Compressed sensing

Definition A.1 (Gaussian mean width)**.** *The Gaussian mean width of a set $K \subset \mathbb{R}^d$ is defined as:*

$$w(K) \triangleq \mathbb{E}_{\boldsymbol{g}\sim\mathcal{N}(0,I_d)}\left[\sup_{\boldsymbol{x}\in M(K)} |\langle \mathbf{g}, \boldsymbol{x}\rangle|\right],$$

where $M(K) = \{\boldsymbol{x} - \boldsymbol{y} : \boldsymbol{x}, \boldsymbol{y} \in K\}$ is the Minkowski sum of K and $-K$.

Intuitively, Gaussian mean width measures the complexity of the set K.

Theorem A.2 (Adapted from Theorem 6.1 by Vershynin [26])**.** *Let $K \subset \mathbb{R}^d$ be an arbitrary bounded set, and $A \in \mathbb{R}^{m\times d}$ be a random matrix with its entries sampled iid from Gaussian distribution $\mathcal{N}(0,1)$, observation $\boldsymbol{y} = A\boldsymbol{x}^* + \boldsymbol{\delta}$, where $\frac{1}{\sqrt{m}}\|\boldsymbol{\delta}\|_2 = \epsilon$, and $\boldsymbol{x}^*$ is unknown. Choose $\hat{\boldsymbol{x}}$ to be any vector satisfying $\hat{\boldsymbol{x}} \in K$ and $\frac{1}{\sqrt{m}}\|A\hat{\boldsymbol{x}} - \boldsymbol{y}\|_2 \le \epsilon$. Then*

$$\mathbb{E}\left[\sup_{\hat{\boldsymbol{x}}^*\in K} \|\hat{\boldsymbol{x}} - \boldsymbol{x}^*\|_2\right] \le \sqrt{8\pi}\left(\frac{w(K)}{\sqrt{m}} + \epsilon\right). \tag{4}$$

One consequence of this theorem is that $m = \Theta(w(K)^2)$ measurements are sufficient to guarantee small recovery error.

Remark: The statement of Theorem A.2 is only different from the original Theorem 6.1 in [26] where we replaced the ℓ_1 norm assumption by a stronger ℓ_2 norm assumption, and therefore is a slightly weaker version of the original theorem. As indicated in [26], this is still valid. To see this,

notice that for any $\hat{\boldsymbol{\delta}} \in \mathbb{R}^m$, we have $\left\|\hat{\boldsymbol{\delta}}\right\|_1 \leq \sqrt{m}\left\|\hat{\boldsymbol{\delta}}\right\|_2$. So the ℓ_2 bound $\frac{1}{\sqrt{m}}\left\|\hat{\boldsymbol{\delta}}\right\|_2 \leq \epsilon$ gives $\frac{1}{m}\left\|\hat{\boldsymbol{\delta}}\right\|_1 \leq \frac{1}{\sqrt{m}}\left\|\hat{\boldsymbol{\delta}}\right\|_2 \leq \epsilon$.

Theorem A.3 (Restatement of Theorem 4.2)**.** *For a given $\boldsymbol{x}^*$ and $q = \log p(\boldsymbol{x}^*)$, define $S(q) = \{\boldsymbol{x} | \log p(\boldsymbol{x}) \geq q\}$. Recall that the observation for $\boldsymbol{x}^*$ is $\boldsymbol{y} = A\boldsymbol{x}^* + \boldsymbol{\delta}$ where $A \in \mathbb{R}^{m\times d}$ has entries drawn i.i.d. from $\mathcal{N}(0, 1/m)$ and the noise level is $\epsilon = \|\boldsymbol{\delta}\|_2$. When we perform the MAP inference by solving the following problem,*

$$\bar{\boldsymbol{x}} \leftarrow \underset{\boldsymbol{x}\in\mathbb{R}^d, \|A\boldsymbol{x}-\boldsymbol{y}\|_2\leq\epsilon}{\arg\min} \{-\log p(\boldsymbol{x})\}, \tag{5}$$

we have:

$$\mathbb{E}\left\|\bar{\boldsymbol{x}} - \boldsymbol{x}^*\right\|_2 \leq \sqrt{8\pi}\left(\frac{w(S(q))}{\sqrt{m}} + \epsilon\right),$$

where the expectation is over the randomness of A.

Proof. Since $\mathbb{E}\left\|\bar{\boldsymbol{x}} - \boldsymbol{x}^*\right\|_2 \leq \mathbb{E}\left[\sup_{\boldsymbol{x}\in S(q)}\left\|\bar{\boldsymbol{x}} - \boldsymbol{x}\right\|_2\right]$, the proof directly follows from Theorem A.2 by choosing the set K to be $S(q)$. We only need to verify that $\bar{\boldsymbol{x}}$ is indeed inside $S(q)$. Notice that both $\bar{\boldsymbol{x}}$ and $\boldsymbol{x}^*$ are in the feasible set $\{\boldsymbol{x} : \|A\boldsymbol{x} - \boldsymbol{y}\|_2 \leq \epsilon\}$. Since $\bar{\boldsymbol{x}}$ maximizes $\log p(\boldsymbol{x})$ within this set, clearly its density should be at least that of $\boldsymbol{x}^*$, i.e. $\log p(\bar{\boldsymbol{x}}) \geq \log p(\boldsymbol{x}^*) = q$. Thus $\bar{\boldsymbol{x}} \in S(q)$.

Note that in our setting the variance of A's entries is set to be $1/m$ instead of 1 from Theorem A.2. Since the Gaussian mean width $w(S(q))$ is invariant to the scaling of A, we correct the scaling on the noise term and obtain $\mathbb{E}\left[\sup_{\boldsymbol{x}\in S(q)}\left\|\bar{\boldsymbol{x}} - \boldsymbol{x}^*\right\|_2\right] \leq \sqrt{8\pi}(\frac{w(S(q))}{\sqrt{m}} + \epsilon)$. □

Remark: Here we show the recovery guarantee when we optimize over the constrained version as in eq. (5). Note that for a suitable constant β, the constrained version $\min_{\boldsymbol{x}:\|A\boldsymbol{x}-\boldsymbol{y}\|_2\leq\epsilon}\{-\log p(\boldsymbol{x})\}$ is equivalent to $\min_{\boldsymbol{x}}\{-\log p(\boldsymbol{x}) + \beta\|A\boldsymbol{x} - \boldsymbol{y}\|_2^2\}$.

B Additional Experimental Results and Details

Here we include experimental results and details not included in the main text. Across all the experiments, we individually tuned the hyperparameters for each method.

B.1 Experimental Details

Dataset For MNIST, we used the default split of 60,000 training images and 10,000 test images of [27]. For CelebA-HQ, we used the split of 27,000 training images and 3,000 test images as provided by [28]. During evaluation, the following Python script was used to select 1000 MNIST images and 100 CelebA-HQ images from their respective test sets:

```
np.random.seed(0); indices_mnist  = np.random.choice(10000, 1000, False)
np.random.seed(0); indices_celeba = np.random.choice(3000,  100,  False)
```

Note that CelebA-HQ images were further resized to 64×64 resolution.

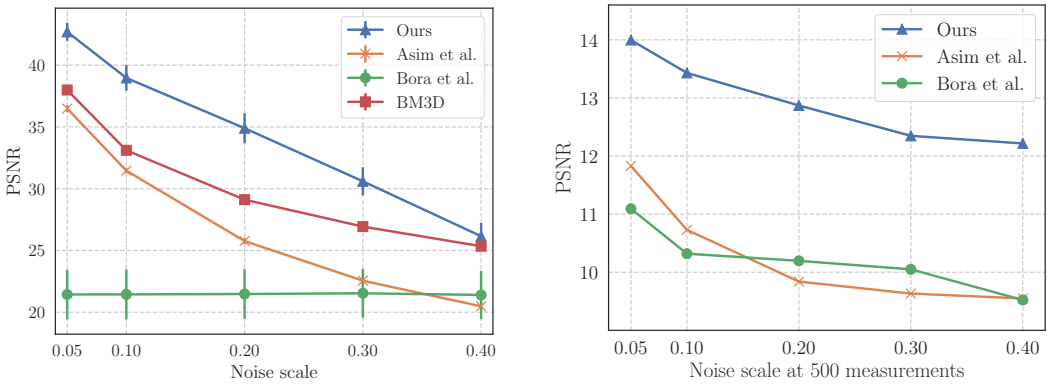
Noise Distribution For the sinusoidal noise used in the experiments, the standard deviation of the k-th pixel/row is calculated as: $\sigma_k = 0.1 \cdot \left(\exp\left(\sin(2\pi \cdot \frac{k}{16})\right) - 1\right)/(e-1)$, clamped to be in range $[0.001, 1]$. For Figure 4b, we used vary the coefficient 0.1 to values in $\{0.05, 0.1, 0.2, 0.3, 0.4\}$.

For the radial noise used in the additional experiment below, the standard deviation of each pixel with ℓ_2 distance is d from the center pixel (31, 31) is computed as: $\sigma_k = 0.1 \cdot \exp(-0.005 \cdot d^2)$, clamped to be in range $[0.001, 1000]$.

B.2 Additional Result: Removing RADIAL Noise

Consider the measurement process $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{\delta}_{\text{radial}}$, where each pixel follows a Gaussian distribution, but with variance that decays exponentially in distance to the center point. For a pixel whose ℓ_2 distance to the center pixel is d, the standard deviation is computed as $\sigma(d) = \exp\left(-0.005 \cdot d^2\right)$. See Figure 3 and Figure 4a for reconstructions as well as PSNR plot comparing the methods considered.

Figure 3: Result of denoising RADIAL noise on CelebA-HQ faces and out-of-distribution images.



(a) Result on denoising RADIAL noise at varying noise rates. Our method achieves the same reconstruction performance even when the noise has approximately $1.5\times$ higher noise scale compared to the best baseline method which is BM3D for this setting.

(b) Result of 1-bit compressed sensing at different noise scale. Our method obtains the best reconstructions, achieving similar PSNR as [9] when the noise scale is $8\times$ higher.

Figure 4: RADIAL denoising results (left) and 1-bit compressed sensing results at different noise levels (right).

B.3 Additional Result: 1-bit Compressed Sensing

Figure 4b shows the performance of each method at different noise scales for a fixed number of measurements. We observe that our method performs consistently better at all noise levels.

C Model Architecture and Hyperparameters

For the RealNVP models we trained, we used multiscale architecture as was done in [29], with residual networks and regularized weight normalization on convolutional layers. Following [28], we used 5-bit color depth for the CelebA-HQ model. Hyperparameters and samples from the models can be found in Table 1 and Figure 5.

Hyperparameter	**CelebA-HQ**	**MNIST**
Learning rate	$5e{-4}$	$1e{-3}$
Batch size	16	128
Image size	$64 \times 64 \times 3$	$28 \times 28 \times 1$
Pixel depth	5 bits	8 bits
Number of epochs	300	200
Number of scales	6	3
Residual blocks per scale	10	6
Learning rate halved every	60 epochs	40 epochs
Max gradient norm	500	100
Weightnorm regularization	$1e{-5}$	$5e{-5}$

Table 1: Hyperparameters used for RealNVP models.

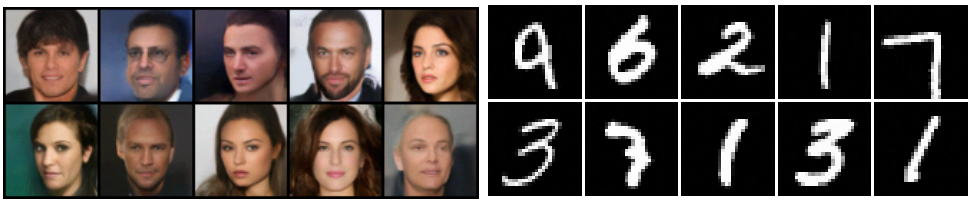


Figure 5: Samples from the RealNVP models used in our experiments.

Figure 6: Out-of-distribution images used in our experiments. We included different types of out-of-distribution instances including greyscale images and cartoons with flat image areas.