AmbientFlow: Invertible generative models from incomplete, noisy imaging measurements

Varun A. Kelkar
Analog Devices, Inc.,
Boston, MA 02110
varun.kelkar@analog.com

Rucha Deshpande
Washington University in St. Louis
St. Louis, MO 63130
r.deshpande@wustl.edu

Arindam Banerjee
University of Illinois at Urbana-Champaign
Urbana, IL 61801
arindamb@illinois.edu

Mark A. Anastasio
University of Illinois at Urbana-Champaign
Urbana, IL 61801
maa@illinois.edu

Abstract

Generative models, including normalizing flows, are gaining popularity in imaging science for tasks such as image reconstruction, posterior sampling, and data sharing. However, training them requires a high-quality dataset of objects, which can be challenging to obtain in fields such as tomographic imaging. This work proposes AmbientFlow, a framework for training flow-based generative models directly from noisy and incomplete data using variational Bayesian methods. The effectiveness of AmbientFlow in learning invertible generative models of objects from noisy, incomplete stylized imaging measurements is demonstrated via numerical studies.

1 Introduction

Invertible generative models (IGMs) are a class of machine learning models that approximate an unknown data distribution by learning an implicit bijective mapping from a simple, tractable distribution such as the standard normal distribution, to the data distribution of interest. Due to their bijectivity, they enable both efficient, high-quality sampling as well as tractable density estimation. Due to these properties, invertible generative models have been investigated for diverse applications, such as data generation and editing (Kingma & Dhariwal, 2018), variational inference (Rezende & Mohamed, 2015) and probabilistic modeling for physical systems (Papamakarios *et al.*, 2021).

IGMs are also promising for potential applications in imaging science. For instance, IGMs have been investigated for their use as priors for regularizing inverse problems in tomographic imaging, where a computational procedure is required to estimate an object from noisy or incomplete imaging measurements (Zhao *et al.*, 2022; Zhang & Curtis, 2021; Kelkar *et al.*, 2021). IGMs offer exact density estimates, tractable log-likelihoods and useful representations of individual images (Dinh *et al.*, 2016, 2014; Kingma & Dhariwal, 2018). This makes them more reliable for downstream inference tasks in imaging science as compared to certain other types of generative models, such as generative adversarial networks (GANs), that exhibit insufficient representation capacity and misrepresentation of domain-specific statistics (Asim *et al.*, 2020; Jalal *et al.*, 2021; Kelkar *et al.*, 2023b). IGMs have shown potential for use in tasks such as image reconstruction, posterior sampling, uncertainty quantification and anomaly detection (Kelkar *et al.*, 2021; Jalal *et al.*, 2021; Zhao *et al.*).

Despite their promise, training IGMs requires a large dataset of objects or high-quality image estimates. This poses challenges in scenarios such as tomographic imaging, where acquiring complete measurements of objects to build such a dataset can be impractical. In the context of GANs, the AmbientGAN framework was proposed to address this problem, where a conventional GAN was augmented with the measurement operator. It employs an adversarial training strategy, where the real and fake object distributions are *indirectly* compared by a discriminator acting on the real measurements, and measurements simulated from the fake objects. This approach is fundamentally

NeurIPS 2023 Workshop on Deep Learning and Inverse Problems .

distinct from the direct log-probability maximization used in training IGMs, which isn't easily adaptable when only incomplete measurements are available.

In this work, a new framework named AmbientFlow is developed for training IGMs using noisy, incomplete measurements. The accuracy of the object distribution recovered via AmbientFlow is theoretically analyzed under prescribed ideal conditions using compressed sensing. Finally, numerical studies are presented to demonstrate the effectiveness of the proposed method on a two-dimensional toy problem, as well as a stylized magnetic resonance imaging (MRI) numerical study. Additional experiments can be found in our extended paper (Kelkar *et al.*, 2023a).

2 Approach

Conventionally, invertible generative models (IGMs) are trained to bijectively map a latent variable $\mathbf{z}$ with a simple distribution, such as the standard normal distribution, to an object $\mathbf{f} \in \mathbb{R}^n$ that follows the distribution of interest, using an invertible neural network (INN) $G_\theta : \mathbb{R}^n \to \mathbb{R}^n$. The IGM is trained by minimizing the KL divergence $D_{\mathrm{KL}}(q_{\mathbf{f}} \| p_\theta)$ between the learned distribution p_θ and the true distribution $q_{\mathbf{f}}$, or equivalently, by maximizing the log-likelihood objective over a dataset of samples drawn from $q_{\mathbf{f}}$:

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathbf{f}\sim q_{\mathbf{f}}} \log p_\theta(\mathbf{f}) = \mathbb{E}_{\mathbf{f}\sim q_{\mathbf{f}}}[\log q_{\mathbf{z}}(\mathbf{z}) - \log|\det \nabla_{\mathbf{z}} G_\theta(\mathbf{z})|], \;\; \mathbf{z} = G_\theta^{-1}(\mathbf{f}). \tag{1}$$

In the present scenario, instead of samples from $q_{\mathbf{f}}$, a training dataset of noisy and potentially incomplete imaging measurements $\mathbf{g} = H\mathbf{f} + \mathbf{n}$ from objects $\mathbf{f} \sim q_{\mathbf{f}}$ is available. Here, $H \in \mathbb{R}^{m\times n}$, known as the forward model, is a linear operator that models the physics of the imaging process, and $\mathbf{n} \in \mathbb{R}^m, \mathbf{n} \sim q_{\mathbf{n}}$ models the measurement noise.

Let $q_{\mathbf{g}}$ be the true distribution of the measurements, and ψ_θ be the distribution of fake measurements, i.e. for $\mathbf{f} \sim p_\theta$, $H\mathbf{f} + \mathbf{n} \sim \psi_\theta$. Since samples from $q_{\mathbf{f}}$ are not available, an additional INN, known as the posterior network $h_\phi(\cdot\,; \mathbf{g}) : \mathbb{R}^n \to \mathbb{R}^n$ is introduced that takes two inputs – a latent vector $\boldsymbol{\zeta} \sim q_{\boldsymbol{\zeta}} = \mathcal{N}(\mathbf{0}, I_n)$, and a conditioning input $\mathbf{g} \sim q_{\mathbf{g}}$ from the training dataset, to produce samples from the model posterior $p_\phi(\mathbf{f} \,|\, \mathbf{g})$. We propose to minimize $D_{\mathrm{KL}}(q_{\mathbf{g}} \| \psi_\theta)$ in lieu of $D_{\mathrm{KL}}(q_{\mathbf{f}} \| p_\theta)$ and establish conditions under which matching ψ_θ to $q_{\mathbf{g}}$ upper-bounds a distance between p_θ and $q_{\mathbf{f}}$. As formalized in Theorem A.1 in the appendix, for sufficiently expressive parametrizations of p_θ and p_ϕ, minimizing $D_{\mathrm{KL}}(q_{\mathbf{g}} \| \psi_\theta)$ is equivalent to maximizing the following objective:

$$\mathcal{L}_M(\theta, \phi) = \mathbb{E}_{\mathbf{g}, \boldsymbol{\zeta}_i} \underset{0<i\leq M}{\mathrm{logavgexp}} \left[\log p_\theta\big(h_\phi(\boldsymbol{\zeta}_i; \mathbf{g})\big) + \log q_{\mathbf{n}}\big(\mathbf{g} - H h_\phi(\boldsymbol{\zeta}_i; \mathbf{g})\big) - \log p_\phi\big(h_\phi(\boldsymbol{\zeta}_i; \mathbf{g}) \,|\, \mathbf{g}\big)\right],$$

$$\text{where } \boldsymbol{\zeta}_i \sim q_{\boldsymbol{\zeta}},\ 0 < i \leq M, \text{ and } \underset{0<i\leq M}{\mathrm{logavgexp}}(x_i) \coloneqq \log\left[\frac{1}{M}\sum_{i=1}^{M} \exp(x_i)\right]. \tag{2}$$

The objective $\mathcal{L}_M(\theta, \phi)$ uses the posterior network to circumvent the need for direct access to $\psi_\theta(\mathbf{g})$, or samples of true objects from $q_{\mathbf{f}}$.

It can be shown that for a system with an injective forward operator H and measurement noise $\mathbf{n}$ having a non-zero characteristic function, $\psi_\theta = q_{\mathbf{g}} \Rightarrow p_\theta = q_{\mathbf{f}}$ (Bora *et al.*, 2018). However, when the forward operator H has a null space, it is not possible to uniquely relate p_θ to the measurement distribution without additional information about $q_{\mathbf{f}}$. Nevertheless, if the objects of interest are known to be close to vectors that are sparse with respect to a full-rank transform Φ, p_θ can be constrained to the set of distributions concentrated on these compressible objects. In order to recover a distribution p_θ concentrated on objects that are compressible with respect to Φ, the following optimization problem is proposed:

$$\hat{\theta}, \hat{\phi} = \underset{\theta,\phi}{\arg\min} -\mathcal{L}_M(\theta, \phi) \quad \text{subject to } \mathbb{E}_{\mathbf{g}\sim q_{\mathbf{g}}} \mathbb{E}_{\mathbf{f}\sim p_\phi(\cdot\,|\,\mathbf{g})} \|\Phi\mathbf{f} - \Phi \mathrm{proj}_{\mathcal{S}_k}(\mathbf{f})\|_1 < \epsilon, \tag{3}$$

where $\mathrm{proj}_{\mathcal{S}_k}(\mathbf{f})$ denotes the orthogonal projection of $\mathbf{f} \in \mathbb{R}^n$ onto the set $\mathcal{S}_k$ of objects for which $\Phi\mathbf{f}$ is $k-$sparse. It can be shown that if H and $\mathbf{f} \sim q_{\mathbf{f}}$ satisfy specific conditions of compressed sensing (Candes *et al.*, 2006) and the AmbientFlow is trained sufficiently well using Eq. (3), then the error between the true and recovered object distributions can be bounded. This is formally established via Theorem A.2 in the appendix.

In practice, Eq. (3) is reformulated in its Lagrangian form, and a regularization parameter μ is used to control the strength of the sparsity-promoting constraint. Also, inspired by the β-VAE framework (Higgins *et al.*, 2017), an additional regularization parameter λ was used to control the strength of the likelihood term $\log q_{\mathbf{n}}(\mathbf{g} - Hh_\phi(\boldsymbol{\zeta}_i; \mathbf{g}))$. This modifies the problem to maximizing the following objective function, which was optimized using gradient-based methods.

$$\tilde{\mathcal{L}}_M(\theta, \phi) = \mathbb{E}_{\mathbf{g}\sim q_{\mathbf{g}}} \mathbb{E}_{\boldsymbol{\zeta}_i \sim q_{\boldsymbol{\zeta}}} \Bigg[\underset{0<i\leq M}{\text{logavgexp}} \left\{ \log p_\theta\big(h_\phi(\boldsymbol{\zeta}_i; \mathbf{g})\big) + \lambda \log q_{\mathbf{n}}\big(\mathbf{g} - Hh_\phi(\boldsymbol{\zeta}_i; \mathbf{g})\big) \right.$$
$$\left. - \log p_\phi\big(h_\phi(\boldsymbol{\zeta}_i; \mathbf{g}) \,|\, \mathbf{g}\big) \right\} - \mu \big\| \Phi h_\phi(\boldsymbol{\zeta}_i; \mathbf{g}) - \text{proj}_{\mathcal{S}_k}(\Phi h_\phi(\boldsymbol{\zeta}_i; \mathbf{g})) \big\|_1 \Bigg] \quad (4)$$

Empirically, we observe that the proposed ℓ_1 penalty also promotes sparse deviation of the output of h_ϕ from $\mathcal{S}_k$, which improves the quality of the images generated by AmbientFlow.

3 Numerical Studies

This section describes the numerical studies used to demonstrate the utility of AmbientFlow for learning object distributions from noisy and incomplete imaging measurements. The studies include toy problems in two dimensions, as well as the problem of recovering the object distribution from stylized magnetic resonance imaging measurements.

1) Toy problem: First, a two-dimensional object distribution was considered, which was created as a sum of eight Gaussian distributions with centers forming a regular octagon, as shown in Fig. 1a. The forward operator was the identity operator, and the noise $\mathbf{n}$ was distributed as a zero-mean Gaussian. The distribution of the measurements $\mathbf{g} = \mathbf{f} + \mathbf{n}$ is shown in Fig. 1b.

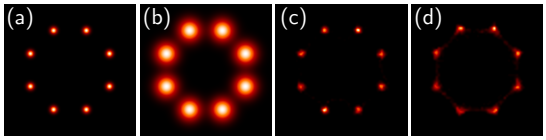

Figure 1: (a) True distribution $q_{\mathbf{f}}$, (b) distribution $q_{\mathbf{g}}$ of measurements, (c) distribution learned by a flow model trained on true objects, and (d) distribution learned by AmbientFlow trained on measurements.

2) Stylized MRI study: In this study, the problem of recovering the distribution of objects from simulated, stylized MRI measurements was considered. T2-weighted brain images of size $n = 128\times128$ from the FastMRI initiative database were considered (Zbontar *et al.*, 2018) as samples from the object distribution. Two separate AmbientFlow models were trained using simulated, stylized MRI measurements with undersampling ratio $n/m = 1$ and $n/m = 4$ containing complex valued iid Gaussian measurement noise. A discrete gradient operator was used as the sparsifying transform Φ. The images generated by AmbientFlow were compared alongside the images individually reconstructed from the measurements using penalized least-squares with TV regularization (PLS-TV) (Beck & Teboulle, 2009), as well as the inverse fast Fourier transform (IFFT) based estimates. The Frechet Inception distance (FID) score was used for comparisons. Additionally, radiomic features commonly computed in medical imaging applications were employed (Van Griethuysen *et al.*, 2017). Details of AmbientFlow training and evaluation are provided in the appendix.

4 Results

Figure 1 shows the true object distribution, the distribution learned by a flow model trained on objects, the measurement distribution, and the object distribution recovered by AmbientFlow trained using the measurements. It can be seen that AmbientFlow is successful in generating nearly noiseless samples that belong to one of the eight Gaussian blobs, although a small number of generated samples lie in the connecting region between the blobs.

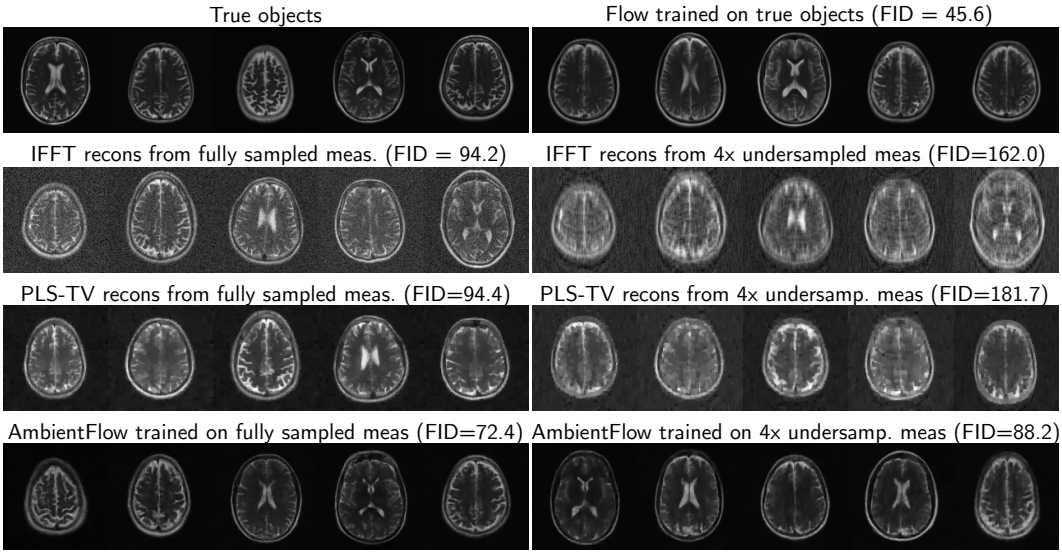

Figure 3: True objects, IFFT-based image estimates, PLS-TV based image estimates and images synthesized by the flow model trained on the true objects, as well as the AmbientFlows trained on the measurements for the stylized MRI study.

The results of the stylized MRI study are shown in Fig. 3. The visual and FID-based quality of images synthesized by the AmbientFlow models was inferior only to the images synthesized by the flow trained directly on objects, and was superior to the images reconstructed individually from the measurements using the PLS-TV method. Since the underlying Inception network used to compute the FID score is not directly related to medical images, additional evaluation was performed in terms of radiomic features relevant to medical image assessments.

Figure 2 plots the empirical PDF over the first two principal components of the radiomic features extracted from each of the MR image sets shown in Fig. 3, except the IFFT image estimates. It can be seen that there is a significant disparity between the principal radiomic feature PDFs of the true objects and the images reconstructed individually using PLS-TV. On the other hand, the AmbientFlow-generated images have a radiomic feature distribution closer to the true objects for both the fully sampled and 4-fold undersampled cases. This implies that, training an AmbientFlow on the noisy/incomplete measurements yielded an estimate of the object distribution that was more accurate in terms of the computed radiomic features, than the one defined by images individually reconstructed from the measurements using the PLS-TV method.

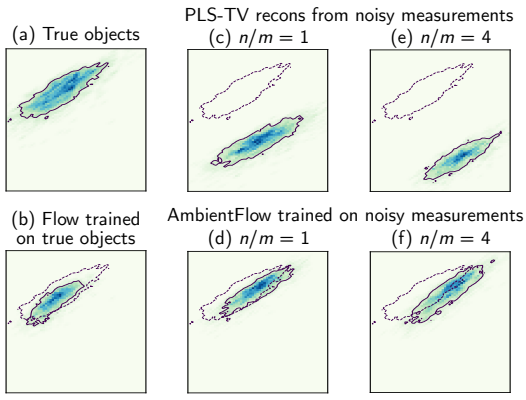

Figure 2: Empirical PDF over the first two principal components of the radiomic features extracted from MRI images. For each plot, the bold contour encloses the region containing 80% of the probability mass. For (b-f), the dotted contour encloses the region containing 80% of the probability mass of the true objects.

5 Discussion and conclusion

In imaging science, obtaining an estimate of the object distribution that is useful for downstream tasks is challenging when only noisy and incomplete measurements of the objects are available. In this work, a framework for learning flow-based generative models of objects directly from noisy/incomplete measurements was developed. The presented numerical studies show that AmbientFlow mitigated the effects of data incompleteness and measurement noise, and accurately approximated the object

distribution in terms of perceptual measures such as the FID score, as well as domain specific radiomic features.

The AmbientFlow framework bears some similarity with variational autoencoders (VAEs), with the latent distribution in VAEs being analogous to the object distribution in AmbientFlow. However, the latent distribution in VAEs is typically simple and non-unique, whereas the object distribution in AmbientFlow is a complex high dimensional distribution that is of primary interest and needs to be recovered as accurately as possible. Hence, the two frameworks have distinct goals that guide their design. The presented framework can be adopted to other generative models that utilize a log-likelihood-based training objective, such as denoising diffusion probabilistic models (DDPMs) which enable high-quality Bayesian inference in imaging (Song *et al.*, 2021; Daras *et al.*, 2023).

A limitation of the proposed framework is that its performance is limited by the capacity of the posterior network h_ϕ to accurately model the posterior. Also, although this work involves preliminary assessments of AmbientFlow using the FID score and radiomic features, a thorough evaluation of generative models for imaging applications would involve assessing whether they can reproduce image statistics that are relevant to a wide variety of downstream tasks (Kelkar *et al.*, 2023b,c).

References

Lynton Ardizzone, Jakob Kruse, Carsten Lüth, Niels Bracher, Carsten Rother, and Ullrich Köthe. Conditional invertible neural networks for diverse image-to-image translation. In *Pattern Recognition: 42nd DAGM German Conference, DAGM GCPR 2020, Tübingen, Germany, September 28–October 1, 2020, Proceedings 42*, pp. 373–387. Springer, 2021.

Muhammad Asim, Max Daniels, Oscar Leong, Ali Ahmed, and Paul Hand. Invertible generative models for inverse problems: mitigating representation error and dataset bias. In *International Conference on Machine Learning*, pp. 399–409. PMLR, 2020.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

Ashish Bora, Eric Price, and Alexandros G Dimakis. Ambientgan: Generative models from lossy measurements. In *International conference on learning representations*, 2018.

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.

Emmanuel J Candes, Justin K Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 59(8):1207–1223, 2006.

Giannis Daras, Kulin Shah, Yuval Dagan, Aravind Gollakota, Alexandros G Dimakis, and Adam Klivans. Ambient diffusion: Learning clean distributions from corrupted data. *arXiv preprint arXiv:2305.19256*, 2023.

Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.

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

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2017.

Ajil Jalal, Sushrut Karmalkar, Alexandros G Dimakis, and Eric Price. Instance-optimal compressed sensing via posterior sampling. *arXiv preprint arXiv:2106.11438*, 2021.

Varun A Kelkar, Sayantan Bhadra, and Mark A Anastasio. Compressible latent-space invertible networks for generative model-constrained image reconstruction. *IEEE transactions on computational imaging*, 7:209–223, 2021.

Varun A Kelkar, Rucha Deshpande, Arindam Banerjee, and Mark A Anastasio. Ambientflow: Invertible generative models from incomplete, noisy measurements. *arXiv preprint arXiv:2309.04856*, 2023a.

Varun A Kelkar, Dimitrios S Gotsis, Frank J Brooks, KC Prabhat, Kyle J Myers, Rongping Zeng, and Mark A Anastasio. Assessing the ability of generative adversarial networks to learn canonical medical image statistics. *IEEE transactions on medical imaging*, 2023b.

Varun A Kelkar, Dimitrios S Gotsis, Rucha Deshpande, Frank J Brooks, KC Prabhat, Kyle J Myers, Rongping Zeng, and Mark A Anastasio. Evaluating generative stochastic image models using task-based image quality measures. In *Medical Imaging 2023: Image Perception, Observer Performance, and Technology Assessment*, volume 12467, pp. 304–310. SPIE, 2023c.

Diederik P Kingma, Max Welling, *et al.* An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019.

Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018.

Bhagwandas Pannalal Lathi and Roger A Green. *Linear systems and signals*, volume 2. Oxford University Press New York, 2005.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *The Journal of Machine Learning Research*, 22 (1):2617–2680, 2021.

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11410–11420, 2022.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pp. 1530–1538. PMLR, 2015.

Kim Seonghyeon. PyTorch Implementation of Glow: Generative Flow with Invertible 1x1 Convolutions. URL `https://github.com/rosinality/glow-pytorch`.

Yang Song, Liyue Shen, Lei Xing, and Stefano Ermon. Solving inverse problems in medical imaging with score-based generative models. *arXiv preprint arXiv:2111.08005*, 2021.

Joost JM Van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina GH Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo JWL Aerts. Computational radiomics system to decode the radiographic phenotype. *Cancer research*, 77(21):e104–e107, 2017.

Jure Zbontar, Florian Knoll, Anuroop Sriram, Tullie Murrell, Zhengnan Huang, Matthew J. Muckley, Aaron Defazio, Ruben Stern, Patricia Johnson, Mary Bruno, Marc Parente, Krzysztof J. Geras, Joe Katsnelson, Hersh Chandarana, Zizhao Zhang, Michal Drozdzal, Adriana Romero, Michael Rabbat, Pascal Vincent, Nafissa Yakubova, James Pinkerton, Duo Wang, Erich Owens, C. Lawrence Zitnick, Michael P. Recht, Daniel K. Sodickson, and Yvonne W. Lui. fastMRI: An open dataset and benchmarks for accelerated MRI, 2018.

Xin Zhang and Andrew Curtis. Bayesian geophysical inversion using invertible neural networks. *Journal of Geophysical Research: Solid Earth*, 126(7):e2021JB022320, 2021.

Xuebin Zhao, Andrew Curtis, and Xin Zhang. Bayesian seismic tomography using normalizing flows. *Geophysical Journal International*, 228(1):213–239, 2022.

Yuzhong Zhao, Qiaoqiao Ding, and Xiaoqun Zhang. Ae-flow: Autoencoders with normalizing flows for medical images anomaly detection. In *The Eleventh International Conference on Learning Representations*.

A Theoretical analysis

First, the notation used in the manuscript is defined again for convenience.

Notation. Let $q_{\mathbf{f}}$, $q_{\mathbf{g}}$ and $q_{\mathbf{n}}$ denote the unknown true object distribution to-be-recovered, the true measurement distribution and the known measurement noise distribution, respectively. Let $\mathcal{D} = \{\mathbf{g}^{(i)}\}_{i=1}^{D}$ be a dataset of independent and identically distributed (iid) measurements drawn from $q_{\mathbf{g}}$. Let $G_\theta : \mathbb{R}^n \to \mathbb{R}^n$ be an INN. Let p_θ be the distribution represented by G_θ, i.e. given a latent distribution $q_{\mathbf{z}} = \mathcal{N}(\mathbf{0}, I_n)$, $G_\theta(\mathbf{z}) \sim p_\theta$ for $\mathbf{z} \sim q_{\mathbf{z}}$. Also, let ψ_θ be the distribution of fake measurements, i.e. for $\mathbf{f} \sim p_\theta$, $H\mathbf{f} + \mathbf{n} \sim \psi_\theta$. Let $p_\theta(\mathbf{f} \,|\, \mathbf{g}) \propto q_{\mathbf{n}}(\mathbf{g} - H\mathbf{f})\, p_\theta(\mathbf{f})$ denote the posterior induced by the learned object distribution represented by G_θ. Let $\Phi \in \mathbb{R}^{l \times n}$, $l \geq n$ be a full-rank linear transformation (henceforth referred to as a sparsifying transform). Also, let $\mathcal{S}_k = \{\mathbf{v} \in \mathbb{R}^n \text{ s.t. } \|\Phi\mathbf{v}\|_0 \leq k\}$ be the set of vectors k-sparse with respect to Φ. Since Φ is full-rank, throughout this chapter we assume without the loss of generality, that $\|\Phi^+\|_2 \leq 1$, where Φ^+ is the Moore-Penrose pseudoinverse of Φ. Throughout the manuscript, we also assume that $q_{\mathbf{f}}$ is absolutely continuous with respect to p_θ, and $q_{\mathbf{g}}$ is absolutely continuous with respect to ψ_θ.

Next, we define the restricted isometry property which characterizes matrices that are well-conditioned when operating on sparse vectors.

Definition A.1 (Restricted isometry property)**.** For $s \in \mathbb{N}$, define the restricted isometry constant (RIC) δ_s as the smallest constant that satisfies

$$(1 - \delta_s)\|\mathbf{v}\|_2^2 \leq \|H\mathbf{v}\|_2^2 \leq (1 + \delta_s)\|\mathbf{v}\|_2^2, \tag{5}$$

for all $\mathbf{v}$ such that $\|\Phi\mathbf{v}\|_0 \leq s$. H is said to satisfy the restricted isometry property for all $\mathbf{v}$ such that $\|\Phi\mathbf{v}\|_0 \leq k$, if $\delta_k + \delta_{2k} + \delta_{3k} < 1$ (Candes *et al.*, 2006).

Compressed sensing stipulates that if an object $\mathbf{f}$ is k-sparse after a full-rank linear transformation $\Phi \in \mathbb{R}^{l \times n}$, $l \geq n$, then the object can be stably estimated from noisy, incomplete measurements $H\mathbf{f} + \mathbf{n}$ if for all vectors $\mathbf{v} \in \mathbb{R}^n$ that are k-sparse in the transform domain Φ, H satisfies the restricted isometry property (RIP) defined above (Candes *et al.*, 2006).

The following theorem formulates an objective that is equivalent to minimizing $D_{\mathrm{KL}}(q_{\mathbf{g}} \| \psi_\theta)$.

Theorem A.1. *Let h_ϕ be such that $p_\phi(\mathbf{f} \,|\, \mathbf{g}) > 0$ over $\mathbb{R}^n$. Also, assume that h_ϕ and G_θ have sufficient capacity, i.e. $\exists\, \theta^*, \phi^*$ such that $\psi_{\theta^*} = q_{\mathbf{g}}$, and $p_{\phi^*}(\cdot|\mathbf{g}) = p_{\theta^*}(\cdot|\mathbf{g})$. Then, minimizing $D_{\mathrm{KL}}(q_{\mathbf{g}} \| \psi_\theta)$ is equivalent to maximizing the following objective function over θ, ϕ:*

$$\mathcal{L}_M(\theta, \phi) = \mathbb{E}_{\mathbf{g}, \boldsymbol{\zeta}_i} \underset{0<i\leq M}{\operatorname{logavgexp}} \big[\log p_\theta\big(h_\phi(\boldsymbol{\zeta}_i; \mathbf{g})\big) + \log q_{\mathbf{n}}\big(\mathbf{g} - H h_\phi(\boldsymbol{\zeta}_i; \mathbf{g})\big) - \log p_\phi\big(h_\phi(\boldsymbol{\zeta}_i; \mathbf{g}) \mid \mathbf{g}\big)\big], \tag{6}$$

where $\boldsymbol{\zeta}_i \sim q_{\boldsymbol{\zeta}}$, $0 < i \leq M, M \in \mathbb{N}$, and $\operatorname{logavgexp}_{0<i\leq M}(x_i) \coloneqq \log\left[\frac{1}{M}\sum_{i=1}^{M} \exp(x_i)\right]$.

Proof. From the definition of KL divergence, we have

$$D_{\mathrm{KL}}(q_{\mathbf{g}} \| \psi_\theta) = \mathbb{E}_{\mathbf{g} \sim q_{\mathbf{g}}}\left[\log \frac{q_{\mathbf{g}}(\mathbf{g})}{\psi_\theta(\mathbf{g})}\right] \tag{7}$$

$$= \mathbb{E}_{\mathbf{g} \sim q_{\mathbf{g}}} \log q_{\mathbf{g}}(\mathbf{g}) - \mathbb{E}_{\mathbf{g} \sim q_{\mathbf{g}}} \log \psi_\theta(\mathbf{g}). \tag{8}$$

Now, $\psi_\theta(\mathbf{g})$ can be written as

$$\psi_\theta(\mathbf{g}) = \int q_{\mathbf{g}|\mathbf{f}}(\mathbf{g}|\mathbf{f}) p_\theta(\mathbf{f}) d\mathbf{f} \tag{9}$$

$$= \int p_\phi(\mathbf{f}|\mathbf{g}) \frac{q_{\mathbf{n}}(\mathbf{g} - H\mathbf{f})\, p_\theta(\mathbf{f})}{p_\phi(\mathbf{f}|\mathbf{g})} d\mathbf{f} \tag{10}$$

$$= \mathbb{E}_{\mathbf{f} \sim p_\phi(\cdot|\mathbf{g})}\left[\frac{q_{\mathbf{n}}(\mathbf{g} - H\mathbf{f})\, p_\theta(\mathbf{f})}{p_\phi(\mathbf{f}|\mathbf{g})}\right]. \tag{11}$$

Therefore,

$$\mathbb{E}_{\mathbf{g}\sim q_{\mathbf{g}}} \log \psi_\theta(\mathbf{g}) = \mathbb{E}_{\mathbf{g}\sim q_{\mathbf{g}}} \left[\log \mathbb{E}_{\mathbf{f}\sim p_\phi(\cdot|\mathbf{g})} \left\{ \frac{p_\theta(\mathbf{f})\, q_{\mathbf{n}}(\mathbf{g} - H\mathbf{f})}{p_\phi(\mathbf{f} \mid \mathbf{g})} \right\} \right] \tag{12}$$

$$= \mathbb{E}_{\mathbf{g}\sim q_{\mathbf{g}}} \left[\log \mathbb{E}_{\boldsymbol{\zeta}\sim q_{\boldsymbol{\zeta}}} \left\{ \frac{p_\theta\big(h_\phi(\boldsymbol{\zeta};\mathbf{g})\big)\, q_{\mathbf{n}}\big(\mathbf{g} - H h_\phi(\boldsymbol{\zeta};\mathbf{g})\big)}{p_\phi\big(h_\phi(\boldsymbol{\zeta};\mathbf{g}) \mid \mathbf{g}\big)} \right\} \right]. \tag{13}$$

$$\geq \mathbb{E}_{\mathbf{g}\sim q_{\mathbf{g}}, \boldsymbol{\zeta}_i\sim q_{\boldsymbol{\zeta}}} \left[\log \frac{1}{M} \sum_{i=1}^{M} \left\{ \frac{p_\theta\big(h_\phi(\boldsymbol{\zeta}_i;\mathbf{g})\big)\, q_{\mathbf{n}}\big(\mathbf{g} - H h_\phi(\boldsymbol{\zeta}_i;\mathbf{g})\big)}{p_\phi\big(h_\phi(\boldsymbol{\zeta}_i;\mathbf{g}) \mid \mathbf{g}\big)} \right\} \right] \tag{14}$$

$$= \mathcal{L}_M(\theta, \phi). \qquad \text{(by Jensen's inequality)}$$

Furthermore, following the discussion provided in Burda, *et al.* (Burda *et al.*, 2015),

$$\mathcal{L}_M(\theta,\phi) \geq \mathbb{E}_{\mathbf{g}\sim q_{\mathbf{g}}} \log \psi_\theta(\mathbf{g}) - D_{\mathrm{KL}}(p_\phi(\cdot|\mathbf{g}) \| p_\theta(\cdot|\mathbf{g})), \tag{15}$$

$$= \mathcal{L}_{\mathrm{ELBO}}(\theta,\phi), \tag{16}$$

where $\mathcal{L}_{\mathrm{ELBO}}(\theta,\phi)$ denotes the evidence lower bound (Kingma *et al.*, 2019).

Therefore,

$$\mathcal{L}_{\mathrm{ELBO}}(\theta,\phi) \leq \mathcal{L}_M(\theta,\phi) \leq \mathbb{E}_{\mathbf{g}\sim q_{\mathbf{g}}} \log \psi_\theta(\mathbf{g}) \leq \mathbb{E}_{\mathbf{g}\sim q_{\mathbf{g}}} \log q_{\mathbf{g}}(\mathbf{g}). \tag{17}$$

Now, since h_ϕ and G_θ have sufficient capacity, i.e. $\exists\ \theta^*, \phi^*$ such that $\psi_{\theta^*} = q_{\mathbf{g}}$, and $p_{\phi^*}(\cdot|\mathbf{g}) = p_{\theta^*}(\cdot|\mathbf{g})$,

$$\max_{\theta,\phi} \mathcal{L}_{\mathrm{ELBO}}(\theta,\phi) = \max_{\theta} \mathbb{E}_{\mathbf{g}\sim q_{\mathbf{g}}} \log \psi_\theta(\mathbf{g}) = \mathbb{E}_{\mathbf{g}\sim q_{\mathbf{g}}} \log q_{\mathbf{g}}(\mathbf{g}). \tag{18}$$

Therefore,

$$\max_{\theta,\phi} \mathcal{L}_M(\theta,\phi) = \mathbb{E}_{\mathbf{g}\sim q_{\mathbf{g}}} \log q_{\mathbf{g}}(\mathbf{g}). \tag{19}$$

Now, $D_{\mathrm{KL}}(q_{\mathbf{g}} \| \psi_\theta) = \mathbb{E}_{\mathbf{g}\sim q_{\mathbf{g}}} \log q_{\mathbf{g}}(\mathbf{g}) - \mathbb{E}_{\mathbf{g}\sim q_{\mathbf{g}}} \log \psi_\theta(\mathbf{g})$. Therefore, since $\mathbb{E}_{\mathbf{g}\sim q_{\mathbf{g}}} \log q_{\mathbf{g}}(\mathbf{g})$ is a constant,

$$\min_{\theta} D_{\mathrm{KL}}(q_{\mathbf{g}} \| \psi_\theta) = \mathbb{E}_{\mathbf{g}\sim q_{\mathbf{g}}} \log q_{\mathbf{g}}(\mathbf{g}) - \max_{\theta} \mathbb{E}_{\mathbf{g}\sim q_{\mathbf{g}}} \log \psi_\theta(\mathbf{g}) \tag{20}$$

$$= \mathbb{E}_{\mathbf{g}\sim q_{\mathbf{g}}} \log q_{\mathbf{g}}(\mathbf{g}) - \max_{\theta,\phi} \mathcal{L}_M(\theta,\phi) \tag{21}$$

Also, $\mathcal{L}_M(\theta,\phi)$ attains its maximum at θ^*, ϕ^* defined in the statement of Theorem A.1, whereas $D_{\mathrm{KL}}(q_{\mathbf{g}} \| \psi_\theta)$ attains its minimum at θ^*. Therefore, minimizing $D_{\mathrm{KL}}(q_{\mathbf{g}} \| \psi_\theta)$ with respect to θ is equivalent to maximizing $\mathcal{L}_M(\theta,\phi)$ with respect to θ and ϕ under the assumptions of Theorem A.1. □

The following lemma, adapted from (Bora *et al.*, 2018), establishes a condition for unique recovery of the object distribution for certain types of forward operators.

Lemma A.1. *If H is a square matrix ($n = m$) with full-rank, if the noise $\mathbf{n}$ is independent of the object, and if the characteristic function of the noise $\chi_{\mathbf{n}}(\boldsymbol{\nu}) = \mathbb{E}_{\mathbf{n}\sim q_{\mathbf{n}}} \exp(\iota \boldsymbol{\nu}^\top \mathbf{n})$ has full support over $\mathbb{R}^m$ (ι is the square-root of -1), then $\psi_\theta = q_{\mathbf{g}} \Rightarrow p_\theta = q_{\mathbf{f}}$.*

Proof. This proof as been adapted from the AmbientGAN work (Bora *et al.*, 2018). Let $\mathbf{y} = H\mathbf{f}$ represent the noiseless measurements. Therefore,

$$\mathbf{g} = \mathbf{y} + \mathbf{n}, \tag{22}$$

$$\Rightarrow q_{\mathbf{g}} = q_{\mathbf{y}} * q_{\mathbf{n}}, \tag{23}$$

where $*$ represents a convolution (in the sense of linear systems theory) (Lathi & Green, 2005). Therefore,

$$\chi_{\mathbf{g}}(\boldsymbol{\nu}) = \chi_{\mathbf{y}}(\boldsymbol{\nu}) \chi_{\mathbf{n}}(\boldsymbol{\nu}), \quad \boldsymbol{\nu} \in \mathbb{R}^m. \tag{24}$$

Since $\chi_{\mathbf{n}}$ has full support over $\mathbb{R}^m$, $\chi_{\mathbf{g}}$ uniquely determines $\chi_{\mathbf{y}}$. Therefore, $q_{\mathbf{g}}$ uniquely determines $q_{\mathbf{y}}$.

Also, since H is bijective, $q_{\mathbf{y}}$ uniquely determines $q_{\mathbf{f}}$. Therefore, $\psi_\theta = q_{\mathbf{g}} \Rightarrow p_\theta = q_{\mathbf{f}}$. □

Now, if H is rank-deficient, it is in general not possible to uniquely relate p_θ to the measurement distribution. However, for object distributions that are concentrated on transform-sparse vectors, if H satisfies the conditions of compressed sensing, we show that it is possible to accurately recover the object distribution by solving the following optimization problem:

$$\hat{\theta}, \hat{\phi} = \arg\min_{\theta,\phi} -\mathcal{L}_M(\theta, \phi) \quad \text{subject to } \mathbb{E}_{\mathbf{g}\sim q_{\mathbf{g}}} \mathbb{E}_{\mathbf{f}\sim p_\phi(\cdot|\mathbf{g})} \|\Phi\mathbf{f} - \Phi \text{proj}_{\mathcal{S}_k}(\mathbf{f})\|_1 < \epsilon. \tag{3}$$

This is formalized as follows.

Theorem A.2. *For a PDF $q : \mathbb{R}^n \to \mathbb{R}$, let $q^{\mathcal{S}_k}$ denote the distribution of $\text{proj}_{S_k}(\mathbf{x})$, for $\mathbf{x} \sim q$. Also, for distributions q_1, q_2, let $W_1(q_1\|q_2) \coloneqq \inf_{q\in\Gamma} \mathbb{E}_{(\mathbf{x}_1,\mathbf{x}_2)\sim q}\|\mathbf{x}_1 - \mathbf{x}_2\|_2$, denote the Wasserstein 1-distance, with Γ being the set of all joint distributions $q : \mathbb{R}^{n\times n} \to \mathbb{R}$ with marginals q_1, q_2, i.e. $\int q(\mathbf{x}_1, \mathbf{x}_2)d\mathbf{x}_2 = q_1(\mathbf{x}_1),\ \int q(\mathbf{x}_1, \mathbf{x}_2)d\mathbf{x}_1 = q_2(\mathbf{x}_2)$.*

If the following hold:

1. *$W_1(q_\mathbf{f} \parallel q_\mathbf{f}^{\mathcal{S}_k}) \le \epsilon'$ (the true object distribution is concentrated on k-sparse objects under Φ),*
2. *H satisfies the RIP for objects k-sparse w.r.t. Φ, with isometry constant δ_k,*
3. *the characteristic function of noise $\chi_\mathbf{n}(\boldsymbol{\nu})$ has full support over $\mathbb{C}^m$, and*
4. *(θ, ϕ) satisfying $p_\theta = q_\mathbf{f}$ and $p_\phi(\cdot \mid \mathbf{g}) = p_\theta(\cdot \mid \mathbf{g})$ is a feasible solution to Eq. (3) (G_θ and h_ϕ have sufficient capacity),*

then the distribution $p_{\hat{\theta}}$ recovered via Eq. (3) is close to the true object distribution, in terms of the Wasserstein distance i.e.

$$W_1(p_{\hat{\theta}} \parallel q_\mathbf{f}) \le \left(1 + \frac{1}{\sqrt{1-\delta_k}}\|H\|_2\right)(\epsilon + \epsilon'). \tag{25}$$

In order to prove Theorem A.2, we first establish essential notation and intermediate results needed. Specifically, in Lemma A.2, we derive an expression for the wasserstein distance between a distribution of a random variable, and the distribution of its projection onto a set. We then proceed to prove Theorem A.2.

Notation. For a closed set $\mathcal{S} \subset \mathbb{R}^n$, let $\text{proj}_\mathcal{S}(\mathbf{f})$ denote the orthogonal projection of $\mathbf{f}$ onto $\mathcal{S}$, defined as

$$\text{proj}_\mathcal{S}(\mathbf{f}) = \min_{\mathbf{f}'\in\mathcal{S}} \|\mathbf{f}' - \mathbf{f}\|_2 \tag{26}$$

For a PDF $q : \mathbb{R}^n \to \mathbb{R}$, let $q^\mathcal{S}$ denote the distribution of $\text{proj}_S(\mathbf{x})$, for $\mathbf{x} \sim q$. Also, for distributions q_1, q_2, let

$$W_1(q_1\|q_2) \coloneqq \inf_{\gamma\in\Gamma(q_1,q_2)} \mathbb{E}_{(\mathbf{x}_1,\mathbf{x}_2)\sim\gamma}\|\mathbf{x}_1 - \mathbf{x}_2\|_2 \tag{27}$$

denote the Wasserstein 1-distance, with $\Gamma(q_1, q_2)$ being the set of all joint distributions $\gamma : \mathbb{R}^{n\times n} \to \mathbb{R}$ with marginals q_1, q_2, i.e.

$$\int \gamma(\mathbf{x}_1, \mathbf{x}_2)d\mathbf{x}_2 = q_1(\mathbf{x}_1),\ \int \gamma(\mathbf{x}_1, \mathbf{x}_2)d\mathbf{x}_1 = q_2(\mathbf{x}_2). \tag{28}$$

Lemma A.2. *Let $\mathbf{x} \in \mathbb{R}^n$ be a random vector with distribution q. Then, with the above notation,*

$$W_1(q\|q^\mathcal{S}) = \mathbb{E}_{\mathbf{x}\sim q} \|\mathbf{x} - \text{proj}_\mathcal{S}(\mathbf{x})\|_2 . \tag{29}$$

Proof. Let $\gamma_0 : \mathbb{R}^{n\times n} \to \mathbb{R}$ be a degenerate joint distribution given by

$$\gamma_0(\mathbf{x}, \mathbf{w}) = q(\mathbf{x})\delta\big(\mathbf{w} - \text{proj}_\mathcal{S}(\mathbf{x})\big), \quad \mathbf{x}, \mathbf{w} \in \mathbb{R}^n, \tag{30}$$

where, $\delta(\mathbf{w})$ denotes the Dirac delta. Therefore, by definition of the Wasserstein distance,

$$\begin{aligned} W_1(q\|q^\mathcal{S}) &\le \mathbb{E}_{(\mathbf{x},\mathbf{w})\sim\gamma_0} \|\mathbf{x} - \mathbf{w}\|_2 , && (31)\\ &= \int q(\mathbf{x})\delta\big(\mathbf{w} - \text{proj}_\mathcal{S}(\mathbf{x})\big) \|\mathbf{x} - \mathbf{w}\|_2 \, d\mathbf{x}d\mathbf{w}, && (32)\\ &= \int q(\mathbf{x}) \|\mathbf{x} - \text{proj}_\mathcal{S}(\mathbf{x})\|_2 \, d\mathbf{x}, && (33)\\ &= \mathbb{E}_{\mathbf{x}\sim q} \|\mathbf{x} - \text{proj}_\mathcal{S}(\mathbf{x})\|_2 . && (34) \end{aligned}$$

On the other hand, by definition of orthogonal projection,

$$\|\mathbf{x} - \text{proj}_{\mathcal{S}}(\mathbf{x})\|_2 \leq \|\mathbf{x} - \mathbf{w}\|_2, \quad \forall\, \mathbf{w} \in \text{supp}(q^{\mathcal{S}}). \tag{35}$$

Therefore,

$$\mathbb{E}_{\mathbf{x}\sim q} \|\mathbf{x} - \text{proj}_{\mathcal{S}}(\mathbf{x})\|_2 \leq \mathbb{E}_{(\mathbf{x},\mathbf{w})\sim\gamma} \|\mathbf{x} - \mathbf{w}\|_2, \quad \gamma \in \Gamma(q, q^{\mathcal{S}}). \tag{36}$$

$$\Rightarrow \mathbb{E}_{\mathbf{x}\sim q} \|\mathbf{x} - \text{proj}_{\mathcal{S}}(\mathbf{x})\|_2 \leq \inf_{\gamma\in\Gamma(q,q^{\mathcal{S}})} \mathbb{E}_{(\mathbf{x},\mathbf{w})\sim\gamma} \|\mathbf{x} - \mathbf{w}\|_2, \tag{37}$$

$$= W_1(q\|q^{\mathcal{S}}). \tag{38}$$

Equations (34) and (38) imply

$$W_1(q\|q^{\mathcal{S}}) = \mathbb{E}_{\mathbf{x}\sim q} \|x - \text{proj}_{\mathcal{S}}(\mathbf{x})\|_2. \tag{39}$$

□

With all the tools in place, we now proceed to prove Theorem A.2.

The intuitive idea behind the proof of this theorem is as follows. Compressed sensing stipulates that under precribed conditions, the forward operator is injective on a set of sparse vectors. Thus, if an object distribution is sparse, then the distribution of its measurements should be uniquely linked to it. If the object distribution $q_{\mathbf{f}}$ is compressible, and if it is ensured that a compressible distribution $p_{\hat{\theta}}$ is recovered via Eq. (3), then both $q_{\mathbf{f}}$ and $p_{\hat{\theta}}$ will be concentrated on the sparse vectors, and will associated with the same measurement distribution $q_{\mathbf{g}}$. Since the sparse vectors are uniquely determined by the measurements, $p_{\hat{\theta}}$ and $q_{\mathbf{f}}$ must themselves be close.

Proof. Since (θ, ϕ) satisfying $p_\theta = q_{\mathbf{f}}$, and $p_\phi(\cdot|\mathbf{g}) = p_\theta(\cdot|\mathbf{g})$ is a feasible solution to Eq. (3), the maximum value of $\mathcal{L}_M$ under Eq. (3) is $\mathbb{E}_{\mathbf{g}\sim q_{\mathbf{g}}} \log q_{\mathbf{g}}(\mathbf{g})$. Therefore, according to Eq. (17), for the estimated $\hat{\theta}$ and $\hat{\phi}$, $\mathcal{L}(\hat{\theta}, \hat{\phi}) = \mathbb{E}_{\mathbf{g}\sim q_{\mathbf{g}}} \log q_{\mathbf{g}}(\mathbf{g})$,

$$\psi_{\hat{\theta}} = q_{\mathbf{g}} \text{ and } p_{\hat{\phi}}(\cdot|\mathbf{g}) = p_{\hat{\theta}}(\cdot|\mathbf{g}). \tag{40}$$

Let $\mathbf{f}_1, \mathbf{f}_2 \in \mathbb{R}^n$. Therefore, by triangle inequality,

$$\|\mathbf{f}_1 - \mathbf{f}_2\|_2 = \left\|\mathbf{f}_1 - \mathbf{f}_1^{\mathcal{S}} + \mathbf{f}_2^{\mathcal{S}} - \mathbf{f}_2 + \mathbf{f}_1^{\mathcal{S}} - \mathbf{f}_2^{\mathcal{S}}\right\|_2, \tag{41}$$

$$\leq \left\|\mathbf{f}_1 - \mathbf{f}_1^{\mathcal{S}}\right\|_2 + \left\|\mathbf{f}_2^{\mathcal{S}} - \mathbf{f}_2\right\|_2 + \left\|\mathbf{f}_1^{\mathcal{S}} - \mathbf{f}_2^{\mathcal{S}}\right\|_2, \tag{42}$$

where $\mathbf{f}^{\mathcal{S}}$ is a shorthand for $\text{proj}_{\mathcal{S}_k}(\mathbf{f})$ for $\mathbf{f} \in \mathbb{R}^n$.

$\mathbf{f}_1, \mathbf{f}_2$ can be represented in terms of the spasifying transform Φ. Let $\mathbf{c}_i = \Phi\mathbf{f}_i$ and $\mathbf{c}_i^{\mathcal{S}} = \Phi\mathbf{f}_i^{\mathcal{S}}$, for $i = 1, 2$. Therefore,

$$\|\mathbf{f}_1 - \mathbf{f}_2\|_2 \leq \left\|\mathbf{f}_1 - \mathbf{f}_1^{\mathcal{S}}\right\|_2 + \left\|\Phi^+\right\|_2 \left\|\mathbf{c}_2 - \mathbf{c}_2^{\mathcal{S}}\right\|_2 + \left\|\mathbf{f}_1^{\mathcal{S}} - \mathbf{f}_2^{\mathcal{S}}\right\|_2, \tag{43}$$

where Φ^+ is the Moore-Penrose pseudoinverse of Φ. Also, by definition, recall that $\mathbf{c}_1^{\mathcal{S}}$ and $\mathbf{c}_2^{\mathcal{S}}$ have at most k non-zero values.

Now, let $\mathbf{y}_i = H\mathbf{f}_i$ for $i = 1, 2$. Therefore,

$$\left\|\mathbf{y}_1^{\mathcal{S}} - \mathbf{y}_2^{\mathcal{S}}\right\|_2 = \left\|\mathbf{y}_1^{\mathcal{S}} - \mathbf{y}_1 + \mathbf{y}_2 - \mathbf{y}_2^{\mathcal{S}} + \mathbf{y}_1 - \mathbf{y}_2\right\|_2 \tag{44}$$

$$\leq \left\|\mathbf{y}_1 - \mathbf{y}_1^{\mathcal{S}}\right\|_2 + \left\|\mathbf{y}_2 - \mathbf{y}_2^{\mathcal{S}}\right\|_2 + \|\mathbf{y}_1 - \mathbf{y}_2\|_2, \tag{45}$$

$$\leq \|H\|_2 \left\|\mathbf{f}_1 - \mathbf{f}_1^{\mathcal{S}}\right\|_2 + \left\|H\Phi^+\right\|_2 \left\|\mathbf{c}_2 - \mathbf{c}_2^{\mathcal{S}}\right\|_2 + \|\mathbf{y}_1 - \mathbf{y}_2\|_2. \tag{46}$$

Now, the restricted isometry property (RIP) on H defined in Definition A.1 implies

$$\left\|\mathbf{f}_1^{\mathcal{S}} - \mathbf{f}_2^{\mathcal{S}}\right\|_2 \leq \frac{1}{\sqrt{1-\delta_k}} \left\|\mathbf{y}_1^{\mathcal{S}} - \mathbf{y}_2^{\mathcal{S}}\right\|. \tag{47}$$

Therefore, Equations (43), (46) and (47) give

$$\begin{aligned}\|\mathbf{f}_1 - \mathbf{f}_2\| &\leq \left\|\mathbf{f}_1 - \mathbf{f}_1^{\mathcal{S}}\right\|_2 + \left\|\Phi^+\right\|_2 \left\|\mathbf{c}_2 - \mathbf{c}_2^{\mathcal{S}}\right\|_2 \\ &\quad + \frac{1}{\sqrt{1-\delta_k}}\Big[\, \|H\|_2 \left\|\mathbf{f}_1 - \mathbf{f}_1^{\mathcal{S}}\right\|_2 \\ &\quad + \left\|H\Phi^+\right\|_2 \left\|\mathbf{c}_2 - \mathbf{c}_2^{\mathcal{S}}\right\|_2 + \|\mathbf{y}_1 - \mathbf{y}_2\|_2 \Big]. \qquad (48) \\ &\leq \alpha\Big(\left\|\mathbf{f}_1 - \mathbf{f}_1^{\mathcal{S}}\right\|_2 + \left\|\Phi^+\right\|_2 \left\|\mathbf{c}_2 - \mathbf{c}_2^{\mathcal{S}}\right\|_2 \Big) \\ &\quad + \frac{1}{\sqrt{1-\delta_k}} \|\mathbf{y}_1 - \mathbf{y}_2\|_2 , \qquad (49) \\ \text{where } \alpha &= 1 + \frac{1}{\sqrt{1-\delta_k}} \|H\|_2 . \qquad (50)\end{aligned}$$

Now, let $B = \Gamma(q_{\mathbf{f}}, p_{\hat{\theta}})$, i.e. the set of all joint distributions $\beta : \mathbb{R}^{n\times n} \to \mathbb{R}$ that have marginals $q_{\mathbf{f}}$ and $p_{\hat{\theta}}$. Also, let $\rho_{\hat{\theta}}$ be the distribution of $\mathbf{y} = H\mathbf{f}$ for $\mathbf{f} \sim p_{\hat{\theta}}$, i.e. the noiseless version of $\psi_{\hat{\theta}}$. Therefore, for $\beta \in B$,

$$\begin{aligned}\mathbb{E}_{(\mathbf{f}_1,\mathbf{f}_2)\sim\beta} \|\mathbf{f}_1 - \mathbf{f}_2\|_2 &\leq \alpha\Big[\mathbb{E}_{\mathbf{f}_1\sim q_{\mathbf{f}}} \left\|\mathbf{f}_1 - \mathbf{f}_1^{\mathcal{S}}\right\|_2 + \|\Phi^+\|_2 \mathbb{E}_{\mathbf{f}_2\sim p_{\hat{\theta}}} \left\|\Phi\mathbf{f}_2 - \Phi\mathbf{f}_2^{\mathcal{S}}\right\|_2 \Big] \\ &\quad + \frac{1}{\sqrt{1-\delta_k}} \mathbb{E}_{(\mathbf{f}_1,\mathbf{f}_2)\sim\beta} \|H\mathbf{f}_1 - H\mathbf{f}_2\|_2 . \qquad (51)\end{aligned}$$

From Lemma A.2, we have

$$\mathbb{E}_{\mathbf{f}_1\sim q_{\mathbf{f}}} \left\|\mathbf{f}_1 - \mathbf{f}_1^{\mathcal{S}}\right\|_2 = W_1(q_{\mathbf{f}} \| q_{\mathbf{f}}^{\mathcal{S}_k}) \leq \epsilon'. \qquad (52)$$

Also, from Eq. (3),

$$\begin{aligned}\mathbb{E}_{\mathbf{f}_2\sim p_{\hat{\theta}}} \left\|\Phi\mathbf{f}_2 - \Phi\mathbf{f}_2^{\mathcal{S}}\right\|_2 &= \mathbb{E}_{\mathbf{g}\sim q_{\mathbf{g}}} \mathbb{E}_{\mathbf{f}_2\sim p_{\hat{\theta}}(\cdot|\mathbf{g})} \left\|\Phi\mathbf{f}_2 - \Phi\mathbf{f}_2^{\mathcal{S}}\right\|_2 , \qquad (53) \\ &= \mathbb{E}_{\mathbf{g}\sim q_{\mathbf{g}}} \mathbb{E}_{\mathbf{f}\sim p_{\hat{\phi}}(\cdot|\mathbf{g})} \left\|\Phi\mathbf{f} - \Phi\text{proj}_{\mathcal{S}_k}(\mathbf{f})\right\|_2 , \qquad (54) \\ &\leq \mathbb{E}_{\mathbf{g}\sim q_{\mathbf{g}}} \mathbb{E}_{\mathbf{f}\sim p_{\hat{\phi}}(\cdot|\mathbf{g})} \left\|\Phi\mathbf{f} - \Phi\text{proj}_{\mathcal{S}_k}(\mathbf{f})\right\|_1 , \qquad (55) \\ &\leq \epsilon. \qquad (56)\end{aligned}$$

Taking the infimum of Eq. (51) over $\beta \in B$, we get

$$\inf_{\beta\in B} \mathbb{E}_{(\mathbf{f}_1,\mathbf{f}_2)\sim\beta} \|\mathbf{f}_1 - \mathbf{f}_2\|_2 \leq \alpha(\epsilon' + \|\Phi^+\|_2\epsilon) + \frac{1}{\sqrt{1-\delta_k}} \inf_{\beta\in B} \mathbb{E}_{(\mathbf{f}_1,\mathbf{f}_2)\sim\beta} \|H\mathbf{f}_1 - H\mathbf{f}_2\|_2 . \qquad (57)$$

Note that the left-hand side of the above equation is precisely $W_1(p_{\hat{\theta}} \| q_{\mathbf{f}})$. Also, note that the rightmost term in Eq. (57) is $W_1(q_{\mathbf{y}} \| \rho_{\hat{\theta}})$:

$$W_1(q_{\mathbf{y}} \| \rho_{\hat{\theta}}) = \inf_{\beta\in B} \mathbb{E}_{(\mathbf{f}_1,\mathbf{f}_2)\sim\beta} \|H\mathbf{f}_1 - H\mathbf{f}_2\|_2 . \qquad (58)$$

From Eq. (40), since $q_{\mathbf{g}} = \psi_{\hat{\theta}}$, Lemma A.1 implies $q_{\mathbf{y}} = \rho_{\hat{\theta}}$

$$\Rightarrow W_1(q_{\mathbf{y}} \| \rho_{\hat{\theta}}) = 0. \qquad (59)$$

Combining with Eq. (57), and setting $\|\Phi^+\| \leq 1$ and α according to Eq. (50), we get

$$W_1(p_{\hat{\theta}} \| q_{\mathbf{f}}) \leq \left(1 + \frac{1}{\sqrt{1-\delta_k}} \|H\|_2\right)(\epsilon + \epsilon') \qquad (60)$$

□

B Addional details regarding the setup of the numerical studies

Network architecture and training. The architecture of the main flow model G_θ was adapted from the Glow architecture (Kingma & Dhariwal, 2018). The posterior network was adapted from the conditional INN architecture proposed by Ardizzone, *et. al* (Ardizzone *et al.*, 2021). AmbientFlow was trained using PyTorch using an NVIDIA A100 GPU. All hyperparameters for the main INN were fixed based on a PyTorch implementation of the Glow architecture (Seonghyeon), except the number of blocks, which was set to scale logarithmically by the image dimension. The size of the training dataset was 10^7 for the toy study, and 21,000 for the stylized MRI study.

Baselines and evaluation metrics. For each dataset, an INN was trained on the ground-truth objects. The architecture and hyperparameters used for this INN for a particular dataset were identical to the ones used for main flow G_θ within the AmbientFlow framework trained on that dataset. Besides, for the stylized MRI study, a dataset of individual estimates of the object was generated using the penalized least-squares with TV regularization (PLS-TV) algorithm (Beck & Teboulle, 2009). The regularization parameter for the image reconstruction method was tuned to give the lowest RMS error (RMSE) for every individual reconstructed image. Although this method of tuning the parameters is not feasible in real systems, it gives the best set of reconstructed images in terms of the RMSE, thus providing a strong baseline.

The Frechet Inception distance (FID) score, computed using the Clean-FID package (Parmar *et al.*, 2022), was used to compare a dataset of 5,000 true objects with an equivalent number of images synthesized using (1) the INN trained on the true objects and (2) the AmbientFlow trained on the measurements, and (3) images individually reconstructed from their corresponding measurements. Additionally, for the stylized MRI study, radiomic features meaningful to medical imaging were computed on the true objects, generated objects, and reconstructed images (Van Griethuysen *et al.*, 2017).