AmbientFlow: Invertible generative models from incomplete, noisy measurements

Anonymous authors
Paper under double-blind review

Abstract

Generative models have gained popularity for their potential applications in imaging science, such as image reconstruction, posterior sampling and data sharing. Flow-based generative models are particularly attractive due to their ability to tractably provide exact density estimates along with fast, inexpensive and diverse samples. Training such models, however, requires a large, high quality dataset of objects. In applications such as computed imaging, it is often difficult to acquire such data due to requirements such as long acquisition time or high radiation dose, while acquiring noisy or partially observed measurements of these objects is more feasible. In this work, we propose AmbientFlow, a framework for learning flow-based generative models directly from noisy and incomplete data. Using variational Bayesian methods, a novel framework for establishing flow-based generative models from noisy, incomplete data is proposed. Extensive numerical studies demonstrate the effectiveness of AmbientFlow in correctly learning the object distribution. The utility of AmbientFlow in a downstream inference task of image reconstruction is demonstrated.

1 Introduction

Generative models have become a prominent focus of machine learning research in recent years. Modern generative models are neural network-based models of unknown data distributions learned from a large number of samples drawn from the distribution. They ideally provide an accurate representation of the distribution of interest and enable efficient, high-quality sampling and inference. Most modern generative models are implicit, i.e. they map a sample from a simple, tractable distribution such as the standard normal, to a sample from the distribution of interest. Popular generative models for image data distributions include variational autoencoders (VAEs), generative adversarial networks (GANs), normalizing flows and diffusion probabilistic models, among others (Kingma *et al.*, 2019; Goodfellow *et al.*, 2020; Kingma & Dhariwal, 2018; Ho *et al.*, 2020). Recently, many of these models have been successful in synthesizing high-quality perceptually realistic images from the underlying distribution.

Generative models have also been investigated for applications in imaging science. For example, computed imaging systems such as computed tomography (CT) or magnetic resonance imaging (MRI) rely on computational reconstruction to obtain an estimate of an object from noisy or incomplete imaging measurements. Generative models have been investigated for their use as priors in image reconstruction in order to mitigate the effects of data-incompleteness or noise in the measurements. While GANs have been explored for this purpose (Bora *et al.*, 2017; Menon *et al.*, 2020; Kelkar & Anastasio, 2021), they suffer from shortcomings such as inadequate mode coverage (Thanh-Tung & Tran, 2020), insufficient representation capacity (Karras *et al.*, 2020; Bora *et al.*, 2017) and misrepresentation of important domain-specific statistics (Kelkar *et al.*, 2023a). On the other hand, invertible generative models (IGMs), or normalizing flows offer exact density estimates, tractable log-likelihoods and useful representations of individual images (Dinh *et al.*, 2016; 2014; Kingma & Dhariwal, 2018; Kelkar *et al.*, 2021), making them more reliable for downstream inference tasks in imaging science (Asim *et al.*, 2020; Jalal *et al.*, 2021a). These models have shown potential for use in tasks such as image reconstruction, posterior sampling, uncertainty quantification and anomaly detection (Kelkar *et al.*, 2021; Asim *et al.*, 2020; Jalal *et al.*, 2021a; Zhao *et al.*).

Although IGMs are attractive for imaging applications, training them requires a large dataset of objects or high-quality image estimates as a proxy for the objects. Building such a dataset is challenging since acquiring a complete set of measurements to uniquely specify each object can be infeasible. Therefore, it is of interest to learn an IGM of objects directly from a dataset of noisy, incomplete imaging measurements. This problem was previously addressed in the context of GANs via the AmbientGAN framework, which augments a conventional GAN with the measurement operator (Bora *et al.*, 2018; Zhou *et al.*, 2022). It consists of generator and discriminator networks that are jointly optimized via an adversarial training strategy. Here, the generator attempts to synthesize fake objects that produce realistic measurements. The real and fake object distributions are then compared *indirectly* via a discriminator that distinguishes between real measurements and simulated measurements of the fake objects. This is fundamentally different from the training procedure of IGMs, which *directly* computes and maximizes the log-probability of the training data samples. Therefore, ideas from the AmbientGAN framework cannot be easily adapted to train an IGM of objects when only incomplete measurements is available.

The key contributions of this work are as follows. First a new framework named AmbientFlow is developed for training IGMs using noisy, incomplete measurements. Second, the accuracy of the object distribution recovered via AmbientFlow is theoretically analyzed under prescribed ideal conditions using compressed sensing. Next, numerical studies are presented to demonstrate the effectiveness of the proposed method on several different datasets and measurement operators. Finally, the utility of AmbientFlow for a downstream Bayesian inference task is illustrated using a case study of image reconstruction from simulated stylized MRI measurements.

The remainder of this manuscript is organized as follows. Section 2 describes the background of invertible generative models, computed imaging systems and image reconstruction from incomplete measurements. Section 3 describes the notation used and the proposed approach. Section 4 describes the setup for the numerical studies, with the results being presented in Section 5. Finally, the discussion and conclusion is presented in Section 6.

2 Background

Invertible generative models. Invertible generative models (IGMs) or flow-based generative models, are a class of generative models that employ an invertible neural network (INN) to learn an implicit mapping from a simple distribution such as an independent and identically distributed (iid) Gaussian distribution to the data distribution of interest. The INN is a bijective mapping $G_\theta : \mathbb{R}^n \to \mathbb{R}^n$ constructed using a composition of L simpler bijective functions $g_i : \mathbb{R}^n \to \mathbb{R}^n$, with

$$\mathbf{h}^{(i)} = g_i(\mathbf{h}^{(i-1)}) = (g_i \circ g_{i-1} \circ \cdots \circ g_1)(\mathbf{z}), \quad 0 < i \leq L, \quad \mathbf{z}, \mathbf{h}^{(i)} \in \mathbb{R}^n, \tag{1}$$

and $\mathbf{x} = G_\theta(\mathbf{z}) = \mathbf{h}^{(L)}$. As a consequence of bijectivity, the probability distribution function (PDF) p_θ of $\mathbf{x}$ represented by the IGM can be related to the PDF $q_\mathbf{z}$ of $\mathbf{z}$ as:

$$p_\theta(\mathbf{x}) \cdot |\det \nabla_\mathbf{z} G_\theta(\mathbf{z})| = q_\mathbf{z}(\mathbf{z}). \tag{2}$$

In order the establish the IGM, a dataset $\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^D$ is used, where $\mathbf{x}^{(i)}$ are assumed to be independent draws from the unknown true distribution $q_\mathbf{x}$. The INN is then trained by minimizing the following negative log-likelihood objective over the training dataset $\mathcal{D}$:

$$\mathcal{L}(\theta) = -\sum_{i=1}^D \log p_\theta(\mathbf{x}^{(i)}) = \sum_{i=1}^D \left[\log q_\mathbf{z}(\mathbf{z}^{(i)}) - \log|\det \nabla_\mathbf{z} G_\theta(\mathbf{z}^{(i)})|\right], \quad \mathbf{z}^{(i)} = G_\theta^{-1}(\mathbf{x}^{(i)}). \tag{3}$$

Minimizing the above loss function is equivalent to minimizing the Kullback–Leibler (KL) divergence $D_{\mathrm{KL}}(q_\mathbf{x} \| p_\theta) \coloneqq \mathbb{E}_{\mathbf{x}\sim q_\mathbf{x}}[\log q_\mathbf{x}(\mathbf{x}) - \log p_\theta(\mathbf{x})]$ between p_θ and the true data distribution $q_\mathbf{x}$. Most invertible generative models utilize a specialized architecture in order to guarantee invertibility of the constituent functions g_i and tractable computation of the Jacobian loss term $\log|\det \nabla_\mathbf{z} G_\theta(\mathbf{z})|$. Popular building blocks of such an architecture are affine coupling layers (Dinh *et al.*, 2016), random permutations (Dinh *et al.*, 2014) and invertible 1×1 convolutions (Kingma & Dhariwal, 2018).

Image reconstruction from incomplete imaging measurements. Many computed imaging systems can be modeled using an imaging equation described by the following linear system of equations:

$$\mathbf{g} = H\tilde{\mathbf{f}} + \mathbf{n}, \tag{4}$$

where $\tilde{\mathbf{f}} \in \mathbb{R}^n$ is approximates the object to-be-imaged, $H \in \mathbb{R}^{m\times n}$, known as the forward model, is a linear operator that models the physics of the imaging process, $\mathbf{n} \in \mathbb{R}^m, \mathbf{n} \sim q_{\mathbf{n}}$ models the measurement noise, and $\mathbf{g} \in \mathbb{R}^m$ are the measurements of the object. Often, H is ill-conditioned or rank deficient, in which case the measurements $\mathbf{g}$ are not sufficient to form a unique and stable estimate $\hat{\mathbf{f}}$ of the object $\tilde{\mathbf{f}}$, and prior knowledge about the nature of $\tilde{\mathbf{f}}$ is needed. Traditionally, one way to incorporate this prior information is to constrain the domain of H. For example, compressed sensing stipulates that if the true object 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 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) (Candes *et al.*, 2006):

Definition 2.1 (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).

3 Approach

In this section, an AmbientFlow method is proposed for obtaining an IGM of objects from a dataset of measurements. The following preliminary notation will be used in the remainder of the chapter.

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.

Conventionally, according to the discussion below Eq. (3), $D_{\mathrm{KL}}(q_{\mathbf{f}}\|p_\theta)$ would have to be minimized in order to train the IGM to estimate $q_{\mathbf{f}}$. However, in the present scenario, only samples from $q_{\mathbf{g}}$ are available. Therefore, we attempt to minimize the divergence $D_{\mathrm{KL}}(q_{\mathbf{g}}\|\psi_\theta)$, and show that for certain scenarios of interest, this is formally equivalent to approximately minimizing a distance between $q_{\mathbf{f}}$ and p_θ. However, computing $D_{\mathrm{KL}}(q_{\mathbf{g}}\|\psi_\theta)$ is non-trivial because a direct representation of $\psi_\theta(\mathbf{g})$ that could enable the computation of $\log \psi_\theta(\mathbf{g})$ is not available. Fortunately, a direct representation of p_θ is available via G_θ, which can be used to compute $\log p_\theta(\mathbf{f})$ for a given $\mathbf{f} \in \mathbb{R}^n$, using Eq. (2). Therefore, an additional INN, known as the posterior network $h_\phi(\cdot\,;\mathbf{g}) : \mathbb{R}^n \to \mathbb{R}^n$ is introduced that is represents the model posterior $p_\phi(\mathbf{f}\,|\,\mathbf{g})$ designed to approximate $p_\theta(\mathbf{f}\,|\,\mathbf{g})$ when jointly trained along with G_θ. The posterior network h_ϕ is designed to take two inputs – a new latent vector $\boldsymbol{\zeta} \sim q_{\boldsymbol{\zeta}} = \mathcal{N}(\mathbf{0}, I_n)$, and an auxiliary conditioning input $\mathbf{g} \sim q_{\mathbf{g}}$ from the training dataset, to produce $h_\phi(\boldsymbol{\zeta}\,;\mathbf{g}) \sim p_\phi(\mathbf{f}\,|\,\mathbf{g})$. The following theorem establishes a loss function that minimizes $D_{\mathrm{KL}}(q_{\mathbf{g}}\|\psi_\theta)$ using the posterior network, circumventing the need for direct access to $\psi_\theta(\mathbf{g})$, or samples of true objects from $q_{\mathbf{f}}$.

Theorem 3.1. *Let h_ϕ be such that $p_\phi(\mathbf{f}\,|\,\mathbf{g}) > 0$ over $\mathbb{R}^n$. Minimizing $D_{\mathrm{KL}}(q_{\mathbf{g}}\|\psi_\theta)$ is equivalent to maximizing the objective function $\mathcal{L}(\theta,\phi)$ over θ, ϕ, where $\mathcal{L}(\theta,\phi)$ is defined as*

$$\mathcal{L}(\theta,\phi) = \mathbb{E}_{\mathbf{g}\sim q_{\mathbf{g}}}\left[\log \mathbb{E}_{\boldsymbol{\zeta}\sim q_{\zeta}}\left\{\frac{p_\theta\big(h_\phi(\boldsymbol{\zeta};\mathbf{g})\big)\ q_{\mathbf{n}}\big(\mathbf{g}-Hh_\phi(\boldsymbol{\zeta};\mathbf{g})\big)}{p_\phi\big(h_\phi(\boldsymbol{\zeta};\mathbf{g})\,|\,\mathbf{g}\big)}\right\}\right] \tag{6}$$

The proof of Theorem 3.1 is provided in the supplementary document. A variational lower bound of $\mathcal{L}$ is employed, which promotes consistency between the modeled posterior $p_\phi(\cdot\,|\,\mathbf{g})$ and the posterior induced by the learned object distribution, $p_\theta(\cdot\,|\,\mathbf{g})$:

$$\mathcal{L}_M(\theta,\phi) = \mathbb{E}_{\mathbf{g},\boldsymbol{\zeta}_i}\underset{0<i\leq M}{\text{logavgexp}}\left[\log p_\theta\big(h_\phi(\boldsymbol{\zeta}_i;\mathbf{g})\big) + \log q_{\mathbf{n}}\big(\mathbf{g}-Hh_\phi(\boldsymbol{\zeta}_i;\mathbf{g})\big) - \log p_\phi\big(h_\phi(\boldsymbol{\zeta}_i;\mathbf{g})\,|\,\mathbf{g}\big)\right], \tag{7}$$

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

Intuitively, the three terms inside logavgexp in Eq. (7) can be interpreted as follows. The first term implies that G_θ is trained on samples produced by the posterior network h_ϕ. The second term is a data-fidelity term that makes sure that h_ϕ produces objects consistent with the measurement model. The third term penalizes degenerate h_ϕ for which $\nabla_{\boldsymbol{\zeta}} h_\phi(\boldsymbol{\zeta};\mathbf{g})$ is ill-conditioned, for example when h_ϕ produces no variation due to $\boldsymbol{\zeta}$ and only depends on $\mathbf{g}$. Note that the first and third terms are directly accessible via the INNs G_θ and h_ϕ. Also, for noise models commonly used in modeling computed imaging systems, such as the Gaussian noise model (Barrett & Myers, 2013), $q_{\mathbf{n}}$ can be explicitly computed.

For sufficiently expressive parametrizations for p_θ and h_ϕ, the maximum possible value of $\mathcal{L}_M$ is $\mathbb{E}_{\mathbf{g}\sim q_{\mathbf{g}}}\log q_{\mathbf{g}}(\mathbf{g})$, which corresponds to the scenario where the learned posteriors are consistent, i.e. $p_\phi(\mathbf{f}\,|\,\mathbf{g}) = p_\theta(\mathbf{f}\,|\,\mathbf{g})$, and the learned distribution of measurements matches the true measurement distribution, i.e. $\psi_\theta = q_{\mathbf{g}}$. It can be shown that for a class of forward operators, matching the measurement distribution is equivalent to matching the object distribution:

Lemma 3.1. *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 $\psi_\theta = q_{\mathbf{g}} \Rightarrow p_\theta = q_{\mathbf{f}}$.*

The proof of Lemma 3.1 is provided in the supplementary section. When the forward operator H has a null space, it is not possible to uniquely relate the learned object distribution p_θ to the learned measurement distribution ψ_θ without additional information about $q_{\mathbf{f}}$. Nevertheless, if the objects of interest are known to be compressible with respect to a sparsifying 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:

$$\hat{\theta},\hat{\phi} = \underset{\theta,\phi}{\arg\min} -\mathcal{L}_M(\theta,\phi) \quad \text{subject to} \quad \mathbb{E}_{\mathbf{g}\sim q_{\mathbf{g}}}\mathbb{E}_{\mathbf{f}\sim p_\phi(\cdot\,|\,\mathbf{g})}\|\Phi\mathbf{f} - \Phi\text{proj}_{\mathcal{S}_k}(\mathbf{f})\|_1 < \epsilon, \tag{8}$$

where $\text{proj}_{\mathcal{S}_k}(\mathbf{f})$ 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 the conditions of compressed sensing and the AmbientFlow is trained sufficiently well using Eq. (8), then the error between the true and recovered object distributions can be bounded. This is formalized as follows.

Theorem 3.2. *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. *$W_1(q_{\mathbf{f}}\ \|\ q_{\mathbf{f}}^{\mathcal{S}_k}) \leq \epsilon'$ (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 $p_\phi(\cdot\,|\,\mathbf{g}) = p_\theta(\cdot\,|\,\mathbf{g})$ is a feasible solution to Eq. (8) (G_θ and h_ϕ have sufficient capacity),*

then the distribution $p_{\hat{\theta}}$ recovered via Eq. (8) is close to the true object distribution, in terms of the Wasserstein distance i.e.

$$W_1(p_{\hat{\theta}} \parallel q_{\mathbf{f}}) \leq \left(1 + \frac{1}{\sqrt{1-\delta_k}}\|H\|_2\right)(\epsilon + \epsilon'). \tag{9}$$

The proof of Theorem 3.2 is deferred to the supplementary section.

In practice, Eq. (8) 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.

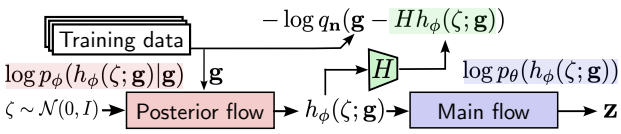

Figure 1: A schematic of the AmbientFlow framework

$$\tilde{\mathcal{L}}_M(\theta,\phi) = \mathbb{E}_{\mathbf{g}\sim q_{\mathbf{g}}}\mathbb{E}_{\boldsymbol{\zeta}_i \sim q_{\boldsymbol{\zeta}}}\Bigg[\underset{0<i\leq M}{\text{logavgexp}}\left\{\log p_\theta\big(h_\phi(\boldsymbol{\zeta}_i;\mathbf{g})\big) + \lambda \log q_{\mathbf{n}}\big(\mathbf{g} - Hh_\phi(\boldsymbol{\zeta}_i;\mathbf{g})\big) - \log p_\phi\big(h_\phi(\boldsymbol{\zeta}_i;\mathbf{g}) \,|\, \mathbf{g}\big)\right\}$$
$$- \mu\left\|\Phi h_\phi(\boldsymbol{\zeta}_i;\mathbf{g}) - \text{proj}_{\mathcal{S}_k}(\Phi h_\phi(\boldsymbol{\zeta}_i;\mathbf{g}))\right\|_1\Bigg] \tag{10}$$

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.

4 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, low-dimensional problems involving a distribution of handwritten digits from the MNIST dataset, problems involving face images from the CelebA-HQ dataset as well as the problem of recovering the object distribution from stylized magnetic resonance imaging measurements. A case study that demonstrates the utility of AmbientFlow in the downstream tasks of image reconstruction and posterior sampling is also described.

Datasets and imaging models.
1) Toy problems: First, a two-dimensional object distribution $q_{\mathbf{f}} : \mathbb{R}^2 \to \mathbb{R}$ was considered, which was created as a sum of eight Gaussian distributions $\mathcal{N}(\mathbf{c}_i, \sigma_{\mathbf{f}}^2 I_2)$, $1 \leq i \leq 8$, with centers $\mathbf{c}_i$ located at the vertices of a regular octagon centered at the origin, and standard deviation $\sigma_{\mathbf{f}} = 0.15$, as shown in Fig. 2a. The forward operator was the identity operator, and the noise $\mathbf{n}$ was distributed as a zero-mean Gaussian with covariance $\sigma_{\mathbf{n}}^2 I_2$, with $\sigma_{\mathbf{n}} = 0.45$. The distribution of the measurements $\mathbf{g} = \mathbf{f} + \mathbf{n}$ is shown in Fig. 2b. A training dataset of size $D = 5 \times 10^7$ was used.

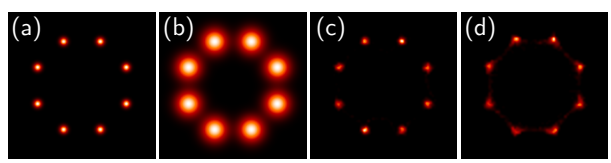

Figure 2: (a) True distribution $q_{\mathbf{f}}$, (b) distribution $q_{\mathbf{g}}$ of measurements, (c) distribution learned by a flow model trained on true objects, and (d) distribution learned by AmbientFlow trained on measurements.

Next, a problem of recovering the distribution of MNIST digits from noisy and/or blurred images of MNIST digits was considered (LeCun *et al.*, 1998). For this problem, three different forward models were considered, namely the identity operator, and two Gaussian blurring operators H_{blur1} and H_{blur2} with root-mean-squared (RMS) width values $\sigma_b = 1.5$ and 3.0 pixels. The measurement noise was distributed as $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma_{\mathbf{n}}^2 I_m)$, with $\sigma_{\mathbf{n}} = 0.3$.

2) Face image study: For the face image study, images of size $n = 64\times64\times3$ from the CelebA-HQ dataset were considered (Karras *et al.*, 2017). Two forward models were considered, namely the identity operator and the

blurring operator with RMS width $\sigma_b = 1.5$, and the measurement noise was distributed as $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma_{\mathbf{n}}^2 I_m)$, $\sigma_{\mathbf{n}} = 0.2$. A discrete gradient operator was used as the sparsifying transform Φ.

3) 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 $n = 128 \times 128$ from the FastMRI initiative database were considered (Zbontar *et al.*, 2018) as samples from the object distribution. Stylized MRI measurements with undersampling ratio $n/m = 1$ (fully sampled) and $n/m = 4$ were simulated using the fast Fourier transform (FFT). Complex valued iid Gaussian measurement noise with standard deviation 0.1 times the total range of ground truth gray values was considered. A discrete gradient operator was used as the sparsifying transform Φ.

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.

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 each forward model for the face image and stylized MRI study, a non-data-driven image restoration/reconstruction algorithm was used to generate a dataset of individual estimates of the object. For $H = I_m$ with $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma_{\mathbf{n}}^2 I_m)$, the block-matching and 3D filtering (BM3D) denoising algorithm was used to perform image restoration (Dabov *et al.*, 2007). For the blurring operator with Gaussian noise, Wienier deconvolution was used for this purpose (Blahut, 2004). For the stylized MRI study, a penalized least-squares with TV regularization (PLS-TV) algorithm was used for image reconstruction (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).

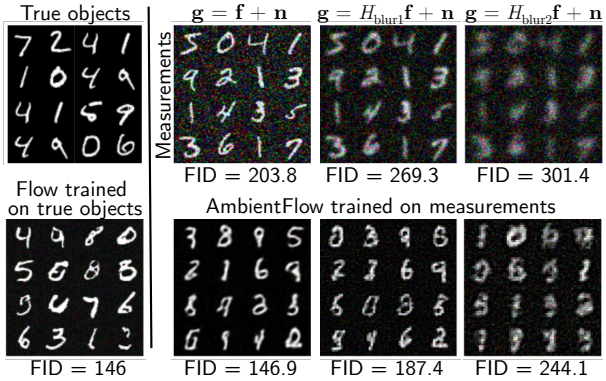

Figure 3: Samples from the flow model trained on true objects and AmbientFlow trained on measurements shown alongside samples of true objects and measurements for the MNIST dataset.

Case study. Next, the utility of the AmbientFlow in a downstream Bayesian inference task was examined. For this purpose, a case study of image reconstruction from incomplete measurements was considered, where the AmbientFlow was used as a prior. Preliminaries of generative priors for image reconstruction are discussed in (Dimakis *et al.*, 2022). In this study, the following two image reconstruction tasks are considered – (1) approximate maximum a posteriori (MAP) estimation, i.e. approximating the mode of the posterior $p_\theta(\cdot \,|\, \mathbf{g})$, and (2) approximate sampling from the posterior $p_\theta(\cdot \,|\, \mathbf{g})$.

Figure 4: True objects, noisy/blurred image measurements, reconstructed images and images synthesized by the flow model trained on the true objects, as well as the AmbientFlows trained on the measurements for the CelebA-HQ dataset.

For both the tasks, an AmbientFlow trained on the fully sampled, noisy simulated MRI measurements, as well as the flow trained on the true objects were considered. For the first task, the compressed sensing using generative models (CSGM) formalism was used to obtained approximate MAP estimates from measurements, for a held-out for a test dataset of size 45 (Asim *et al.*, 2020):

$$\hat{\mathbf{f}}_{\text{MAP}} = G_\theta(\hat{\mathbf{z}}_{\text{MAP}}), \quad \text{with } \hat{\mathbf{z}}_{\text{MAP}} = \arg\min_{\mathbf{z}} \|\mathbf{g} - HG_\theta(\mathbf{z})\|_2^2 + \lambda\|\mathbf{z}\|_2^2. \tag{11}$$

For the second task, approximate posterior sampling was performed with the flow models as priors using annealed Langevin dynamics (ALD) iterations proposed in Jalal, *et al.* (Jalal *et al.*, 2021a). For each true object, the minimum mean-squared error (MMSE) estimate $\hat{\mathbf{f}}_{\text{MMSE}}$ and the pixelwise standard deviation map $\hat{\sigma}$ were computed empirically using 40 samples obtained via ALD iterations. These image estimates were compared with reconstructed images obtained using the PLS-TV method. The regularization parameters for each image reconstruction method were tuned to achieve the best RMSE on a single validation image, and then kept constant for the entire test dataset.

5 Results

Figure 2 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. For the MNIST dataset, Fig. 3 shows samples from the flow model trained on true objects and AmbientFlow trained on measurements alongside samples of true objects and measurements, for different measurement models. It can be seen that when the degradation process is a simple noise addition, the AmbientFlow produces samples that visually appear denoised. When the degradation process consists of blurring and noise addition, the AmbientFlow produces images that do not contain blur or noise, although

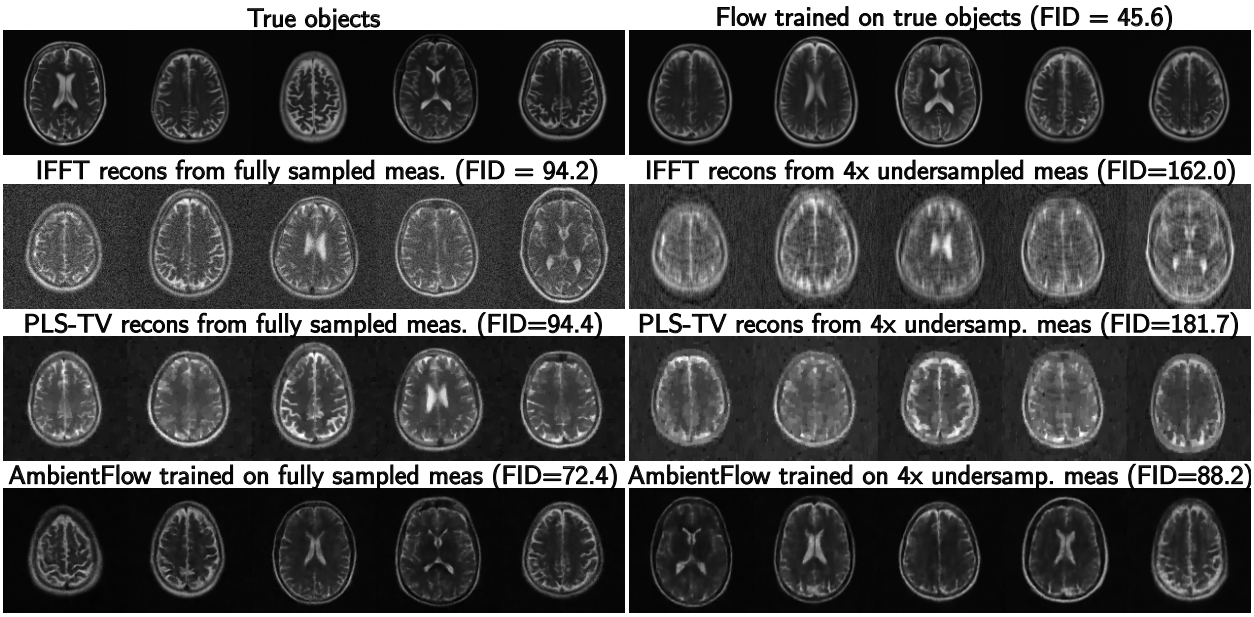

Figure 5: True objects, IFFT-based image estimates, PLS-TV based image estimates and images synthesized by the flow model trained on the true objects, as well as the AmbientFlows trained on the measurements for the stylized MRI study.

the visual image quality degrades as the blur increases. The visual findings are corroborated with quantitative results in terms of the FID score, shown in Fig. 3.

Figure 4 shows the results of the face image study for the two different measurement processes considered - (1) additive noise, and (2) Gaussian blur followed by additive noise. It can be seen that both visually and in terms of the FID score, the quality of images generated by the AmbientFlow models was second only to the flow trained directly on the true objects, for both the forward models considered. The images synthesized by the AmbientFlow models had better visual quality and better fidelity in distribution with the true objects than the ones produced by individually performing image restoration using BM3D and Wiener deconvolution for the two forward models respectively.

The results of the stylized MRI study are shown in Fig. 5. 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.

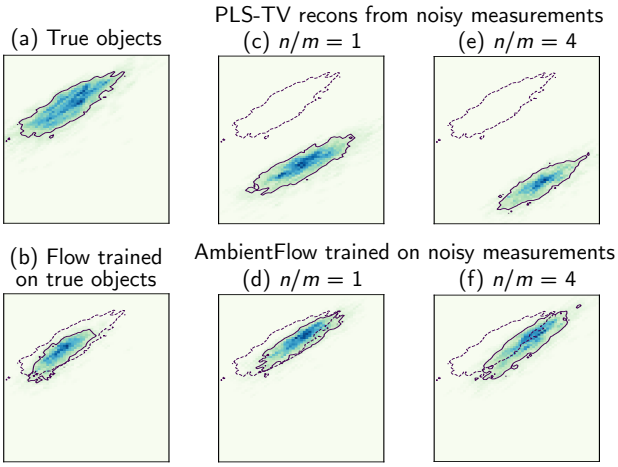

Figure 6: Empirical PDF over the first two principal components of the radiomic features extracted from MRI images. For each plot, the bold contour encloses the region containing 80% of the probability mass. For (b-f), the dotted contour encloses the region containing 80% of the probability mass of the true objects.

Table 1: The mean ± standard deviation of the RMSE and SSIM values computed over the following test image datasets – (1) the images reconstructed using the PLS-TV method, (2) the MAP and MMSE estimates using the flow prior trained on true objects, and (3) the MAP and MMSE estimates using the AmbientFlow prior trained on fully sampled noisy measurements.

Method		RMSE	SSIM
PLS-TV		0.038 ± 0.007	0.806 ± 0.019
Flow prior trained on true objects	MAP Estimate	0.025 ± 0.005	0.922 ± 0.013
	MMSE Estimate	**0.022 ± 0.003**	**0.940 ± 0.006**
AmbientFlow prior	MAP Estimate	0.025 ± 0.004	0.925 ± 0.012
	MMSE Estimate	**0.022 ± 0.004**	**0.936 ± 0.008**

Figure 6 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. 5, 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 relevant radiomic features, than the one defined by images individually reconstructed from the measurements using the PLS-TV method.

Next, the utility of the AmbientFlow for Bayesian inference is demonstrated with the help of an image reconstruction case study. Figure 7 shows a true object alongside images reconstructed from stylized 4-fold undersampled MR measurements simulated from the object, using the reconstruction methods described in Section 4. Recall that for both the flow-based priors shown, the MAP estimate was obtained using the CSGM framework (Bora *et al.*, 2017), and the MMSE estimate and the pixelwise standard deviation maps were computed empirically from samples from the posterior $p_\theta(\mathbf{f} \,|\, \mathbf{g})$ obtained using annealed Langevin dynamics (Jalal *et al.*, 2021a). Visually, it can be seen that the images reconstructed using the AmbientFlow prior were comparable to the images reconstructed using the flow prior trained on the true objects, and better than the image reconstructed using the PLS-TV method. Table 1 shows the root-mean-squared error (RMSE) and structural similarity (SSIM) (Wang *et al.*, 2004) of the reconstructed images with respect to the true object, averaged over a dataset of 45 test images. It can be seen that in terms of RMSE and SSIM, both the MAP and the MMSE image estimates obtained using the AmbientFlow prior are comparable to those obtained using the flow prior trained on true objects, despite the AmbientFlow being trained only using noisy stylized MR measurements.

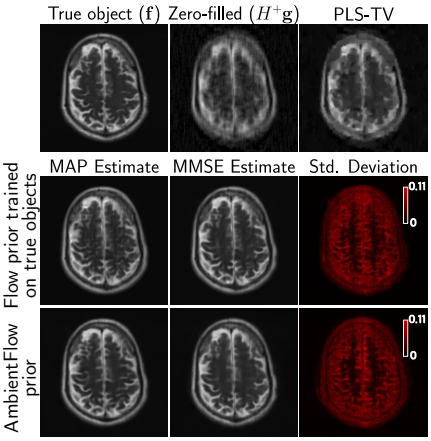

Figure 7: Image estimates and pixelwise standard deviation maps from the image reconstruction case study.

6 Discussion and conclusion

An estimate of the distribution of objects is known to be important for applications in imaging science. However, obtaining such an estimate in terms of a generative model that is useful for downstream tasks has remained challenging, especially when only noisy and incomplete measurements of the objects are available. In this work, a framework for learning flow-based generative models of objects from noisy/incomplete measurements was developed. The presented numerical studies show that the proposed AmbientFlow framework was able to mitigate the effects of data incompleteness and noise present in the measurements in order to build an accurate estimate of the object distribution and generate visually appealing images. In terms of perceptual

measures such as the FID score as well as domain specific radiomic features, the images synthesized by AmbientFlow maintained higher distributional fidelity with the true objects than the images individually reconstructed from the measurements. Furthermore, when the AmbientFlow trained on noisy measurements was employed as a prior in an image reconstruction task, the image estimates obtained were as accurate as those obtained when a flow model trained directly on the true objects was employed.

Some of the above observations also apply to the AmbientGAN framework developed by Bora, *et al.* (Bora *et al.*, 2018). Although the current state-of-the-art GANs may lead to images with better perceptual quality than IGMs, GANs trained on medical images have been shown to misrepresent medically relevant statistics despite producing visually appealing images (Kelkar *et al.*, 2023a). Also, an important drawback of GANs with in the context of imaging science is its unreliability in downstream Bayesian inference tasks. For example, GAN-constrained image reconstruction methods are known to be prone to hallucinations caused by dataset bias, distribution shifts and representation error, whereas IGMs used for the same purpose have been shown to be comparatively robust (Bhadra *et al.*, 2021; Asim *et al.*, 2020; Menon *et al.*, 2020; Kelkar *et al.*, 2021; Jalal *et al.*, 2021b;a; Zhao *et al.*, 2022). These drawbacks of GANs are in part due to the insufficient representation capacity and inability to access the log-probability, both of which are applicable for GANs learned in the ambient setting as well. IGMs, on the other hand, due to their ability to accurately represent images and compute fast, exact density estimates, are more suitable for some downstream inference tasks such as image reconstruction, posterior sampling and uncertainty quantification as compared to GANs.

The AmbientFlow framework bears some similarity with variational autoencoders (VAEs). In both cases, a variational lower bound of the evidence of the data is minimized. The object distribution in AmbientFlow is analogous to the latent variable distribution in VAEs. However, in VAEs, the latent distribution is typically simple and non-unique, and its desirable properties include disentanglement, tractable sampling, and accurate representation of the data via a trained decoder. However, in the AmbientFlow framework, the object distribution is a complex high dimensional distribution that is of primary interest and needs to be recovered as accurately as possible. It is related to the measured data via a physical measurement process, and may have known structure such as transform compressibility. Therefore, the aims and objectives of the two frameworks that guide their design are radically different.

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). However, such an extension is outside the scope of this manuscript, and will be addressed in a future work.

A key limitation of the proposed framework is that its performance depends heavily on the design of the posterior network. The posterior network architecture currently employed is inspired by Ardizzone, *et. al* (Ardizzone *et al.*, 2021). It displays favourable inductive biases for images due to several design choices such as wavelet-based downsampling layers, and a Laplacian pyramid feature extractor for the conditioning input. However, this architecture may not be able to properly model the posterior when H depends on a random parameter, such as a random view angle relevant for cryo-electron microscopy (Zhong *et al.*, 2021). Also, although this work involves preliminary assessments of AmbientFlow using the FID score and radiomic features, a proper evaluation of generative models for imaging applications is still an open problem (Sankaranarayanan *et al.*, 2023). Thorough evaluation of such models would involve assessing whether they can reproduce image statistics that are relevant to a wide variety of downstream tasks (Kelkar *et al.*, 2023a;b).

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.

Harrison H Barrett and Kyle J Myers. *Foundations of image science*. John Wiley & Sons, 2013.

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.

Sayantan Bhadra, Varun A Kelkar, Frank J Brooks, and Mark A Anastasio. On hallucinations in tomographic image reconstruction. *IEEE transactions on medical imaging*, 40(11):3249–3260, 2021.

Richard E Blahut. *Theory of remote image formation*. Cambridge University Press, 2004.

Ashish Bora, Ajil Jalal, Eric Price, and Alexandros G Dimakis. Compressed sensing using generative models. In *International Conference on Machine Learning*, pp. 537–546. PMLR, 2017.

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.

Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on image processing*, 16(8):2080–2095, 2007.

Alexandros G Dimakis, Ashish Bora, Dave Van Veen, Ajil Jalal, Sriram Vishwanath, and Eric Price. Deep generative models and inverse problems. *Mathematical Aspects of Deep Learning*, pp. 400, 2022.

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.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

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.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

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

Ajil Jalal, Sushrut Karmalkar, Jessica Hoffmann, Alex Dimakis, and Eric Price. Fairness for image generation with uncertain sensitive attributes. In *International Conference on Machine Learning*, pp. 4721–4732. PMLR, 2021b.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8110–8119, 2020.

Varun A Kelkar and Mark Anastasio. Prior image-constrained reconstruction using style-based generative models. In *International Conference on Machine Learning*, pp. 5367–5377. PMLR, 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, 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*, 2023a.

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, 2023b.

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.

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.

Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pp. 2437–2445, 2020.

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.

Swami Sankaranarayanan, Thomas Hartvigsen, Camille Bilodeau, Ryutaro Tanno, Cheng Zhang, Florian Tramer, and Phillip Isola. Challenges in deployable generative AI, ICML 2023 workshop. `https://deployinggenerativeai.github.io/index`, 2023.

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.

Hoang Thanh-Tung and Truyen Tran. Catastrophic forgetting and mode collapse in gans. In *2020 international joint conference on neural networks (ijcnn)*, pp. 1–10. IEEE, 2020.

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.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

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.

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*.

Ellen D Zhong, Tristan Bepler, Bonnie Berger, and Joseph H Davis. Cryodrgn: reconstruction of heterogeneous cryo-em structures using neural networks. *Nature methods*, 18(2):176–185, 2021.

Weimin Zhou, Sayantan Bhadra, Frank J Brooks, Hua Li, and Mark A Anastasio. Learning stochastic object models from medical imaging measurements by use of advanced ambient generative adversarial networks. *Journal of Medical Imaging*, 9(1):015503–015503, 2022.

A Theoretical analysis for Section 3

First, the notation defined in Section 3 is rehashed here 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.

A.1 Proof of Theorem 3.1

Theorem 3.1. *Let h_ϕ be such that $p_\phi(\mathbf{f} \,|\, \mathbf{g}) > 0$ over $\mathbb{R}^n$. Minimizing $D_{\mathrm{KL}}(q_{\mathbf{g}} \| \psi_\theta)$ is equivalent to maximizing the objective function $\mathcal{L}(\theta, \phi)$ over θ, ϕ, where $\mathcal{L}(\theta, \phi)$ is defined as*

$$\mathcal{L}(\theta, \phi) = \mathbb{E}_{\mathbf{g} \sim q_{\mathbf{g}}} \left[\log \mathbb{E}_{\boldsymbol{\zeta} \sim q_{\zeta}} \left\{ \frac{p_\theta\big(h_\phi(\boldsymbol{\zeta}; \mathbf{g})\big)\; q_{\mathbf{n}}\big(\mathbf{g} - H h_\phi(\boldsymbol{\zeta}; \mathbf{g})\big)}{p_\phi\big(h_\phi(\boldsymbol{\zeta}; \mathbf{g}) \,|\, \mathbf{g}\big)} \right\} \right] \tag{12}$$

Proof. From the definition of KL divergence, we have

$$\begin{aligned} D_{\mathrm{KL}}(q_{\mathbf{g}} \| \psi_\theta) &= \mathbb{E}_{\mathbf{g} \sim q_{\mathbf{g}}} \left[\log \frac{q_{\mathbf{g}}(\mathbf{g})}{\psi_\theta(\mathbf{g})} \right] && (13) \\ &= \mathbb{E}_{\mathbf{g} \sim q_{\mathbf{g}}} \log q_{\mathbf{g}}(\mathbf{g}) - \mathbb{E}_{\mathbf{g} \sim q_{\mathbf{g}}} \log \psi_\theta(\mathbf{g}). && (14) \end{aligned}$$

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

$$\begin{aligned} \psi_\theta(\mathbf{g}) &= \int q_{\mathbf{g}|\mathbf{f}}(\mathbf{g}|\mathbf{f}) p_\theta(\mathbf{f}) d\mathbf{f} && (15) \\ &= \int p_\phi(\mathbf{f}|\mathbf{g}) \frac{q_{\mathbf{n}}(\mathbf{g} - H\mathbf{f})\; p_\theta(\mathbf{f})}{p_\phi(\mathbf{f}|\mathbf{g})} d\mathbf{f} && (16) \\ &= \mathbb{E}_{\mathbf{f} \sim p_\phi(\cdot|\mathbf{g})} \left[\frac{q_{\mathbf{n}}(\mathbf{g} - H\mathbf{f})\; p_\theta(\mathbf{f})}{p_\phi(\mathbf{f}|\mathbf{g})} \right]. && (17) \end{aligned}$$

Therefore,

$$\begin{aligned} \mathbb{E}_{\mathbf{g} \sim q_{\mathbf{g}}} \log \psi_\theta(\mathbf{g}) &= \mathbb{E}_{\mathbf{g} \sim q_{\mathbf{g}}} \left[\log \mathbb{E}_{\mathbf{f} \sim p_\phi(\cdot|\mathbf{g})} \left\{ \frac{p_\theta(\mathbf{f})\; q_{\mathbf{n}}(\mathbf{g} - H\mathbf{f})}{p_\phi(\mathbf{f} \mid \mathbf{g})} \right\} \right] && (18) \\ &= \mathbb{E}_{\mathbf{g} \sim q_{\mathbf{g}}} \left[\log \mathbb{E}_{\boldsymbol{\zeta} \sim q_{\zeta}} \left\{ \frac{p_\theta\big(h_\phi(\boldsymbol{\zeta}; \mathbf{g})\big)\; q_{\mathbf{n}}\big(\mathbf{g} - H h_\phi(\boldsymbol{\zeta}; \mathbf{g})\big)}{p_\phi\big(h_\phi(\boldsymbol{\zeta}; \mathbf{g}) \mid \mathbf{g}\big)} \right\} \right] && (19) \\ &= \mathcal{L}(\theta, \phi). && (20) \end{aligned}$$

Therefore,

$$D_{\mathrm{KL}}(q_{\mathbf{g}} \| \psi_\theta) = \mathbb{E}_{\mathbf{g} \sim q_{\mathbf{g}}} \log q_{\mathbf{g}}(\mathbf{g}) - \mathcal{L}(\theta, \phi). \tag{21}$$

Since $\mathbb{E}_{\mathbf{g} \sim q_{\mathbf{g}}} \log q_{\mathbf{g}}(\mathbf{g})$ is a constant, minimizing $D_{\mathrm{KL}}(q_{\mathbf{g}} \| \psi_\theta)$ is equivalent to maximizing $\mathcal{L}(\theta, \phi)$. □

Remark. As described in Section 3, a variational lower bound of $\mathcal{L}$ is optimized in this work:

$$\mathcal{L}_M(\theta,\phi) = \mathbb{E}_{\mathbf{g},\boldsymbol{\zeta}_i} \underset{0<i\leq M}{\text{logavgexp}} \left[\log p_\theta\big(h_\phi(\boldsymbol{\zeta}_i;\mathbf{g})\big) + \log q_{\mathbf{n}}\big(\mathbf{g} - Hh_\phi(\boldsymbol{\zeta}_i;\mathbf{g})\big) - \log p_\phi\big(h_\phi(\boldsymbol{\zeta}_i;\mathbf{g}) \mid \mathbf{g}\big)\right], \tag{22}$$

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

For sufficiently expressive parametrizations for p_θ and h_ϕ, the maximum possible value of $\mathcal{L}_M$ is $\mathbb{E}_{\mathbf{g}\sim q_{\mathbf{g}}} \log q_{\mathbf{g}}(\mathbf{g})$, which corresponds to the scenario where the learned posteriors are consistent, i.e. $p_\phi(\mathbf{f}\,|\,\mathbf{g}) = p_\theta(\mathbf{f}\,|\,\mathbf{g})$, and the learned distribution of measurements matches the true measurement distribution, i.e. $\psi_\theta = q_{\mathbf{g}}$. This can be shown as follows.

For $M = 1$, $\mathcal{L}_M(\theta,\phi)$ reduces to the evidence lower bound (ELBO):

$$\begin{aligned}\mathcal{L}_{\text{ELBO}}(\theta,\phi) &= \mathbb{E}_{\mathbf{g}\sim q_{\mathbf{g}},\boldsymbol{\zeta}\sim q_{\boldsymbol{\zeta}}} \left[\log p_\theta\big(h_\phi(\boldsymbol{\zeta};\mathbf{g})\big) + \log q_{\mathbf{n}}\big(\mathbf{g} - Hh_\phi(\boldsymbol{\zeta};\mathbf{g})\big)\right. \\ &\qquad\qquad \left. - \log p_\phi\big(h_\phi(\boldsymbol{\zeta};\mathbf{g}) \mid \mathbf{g}\big)\right]. \qquad (23)\\ &= \mathbb{E}_{\mathbf{g}\sim q_{\mathbf{g}}} \log \psi_\theta(\mathbf{g}) - D_{\text{KL}}(p_\phi(\cdot|\mathbf{g})\|p_\theta(\cdot|\mathbf{g})). \qquad (24)\end{aligned}$$

Similar to importance-weighted autoencoders (IWAE), (Burda *et al.*, 2015), it can be shown that

$$\mathcal{L}_{\text{ELBO}}(\theta,\phi) \leq \mathcal{L}_M(\theta,\phi) \leq \mathbb{E}_{\mathbf{g}\sim q_{\mathbf{g}}} \log \psi_\theta(\mathbf{g}) \leq \mathbb{E}_{\mathbf{g}\sim q_{\mathbf{g}}} \log q_{\mathbf{g}}(\mathbf{g}), \tag{25}$$

with equality occurring when $\psi_\theta = q_{\mathbf{g}}$ and $p_\phi(\cdot|\mathbf{g}) = p_\theta(\cdot|\mathbf{g})$. Thus the maximum value that can be achieved by $\mathcal{L}_M(\theta,\phi)$ is $\mathbb{E}_{\mathbf{g}\sim q_{\mathbf{g}}} \log q_{\mathbf{g}}(\mathbf{g})$.

Lemma 3.1. *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,

$$\mathbf{g} = \mathbf{y} + \mathbf{n}, \tag{26}$$

$$\Rightarrow q_{\mathbf{g}} = q_{\mathbf{y}} * q_{\mathbf{n}}, \tag{27}$$

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

$$\chi_{\mathbf{g}}(\boldsymbol{\nu}) = \chi_{\mathbf{y}}(\boldsymbol{\nu})\chi_{\mathbf{n}}(\boldsymbol{\nu}), \quad \boldsymbol{\nu} \in \mathbb{R}^m. \tag{28}$$

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, $\psi_\theta = q_{\mathbf{g}} \Rightarrow p_\theta = q_{\mathbf{f}}$. □

A.2 Proof of Theorem 3.2

In order to prove Theorem 3.2, we first establish essential notation and intermediate results needed. Specifically, in Lemma A.1, 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 3.2.

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

$$\text{proj}_{\mathcal{S}}(\mathbf{f}) = \min_{\mathbf{f}'\in\mathcal{S}} \|\mathbf{f}' - \mathbf{f}\|_2 \tag{29}$$

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

$$W_1(q_1\|q_2) \coloneqq \inf_{\gamma\in\Gamma(q_1,q_2)} \mathbb{E}_{(\mathbf{x}_1,\mathbf{x}_2)\sim\gamma}\|\mathbf{x}_1 - \mathbf{x}_2\|_2 \tag{30}$$

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.

$$\int \gamma(\mathbf{x}_1, \mathbf{x}_2) d\mathbf{x}_2 = q_1(\mathbf{x}_1), \quad \int \gamma(\mathbf{x}_1, \mathbf{x}_2) d\mathbf{x}_1 = q_2(\mathbf{x}_2). \tag{31}$$

Lemma A.1. *Let $\mathbf{x} \in \mathbb{R}^n$ be a random vector with distribution q. Then, with the above notation,*

$$W_1(q \| q^{\mathcal{S}}) = \mathbb{E}_{\mathbf{x}\sim q} \left\| \mathbf{x} - \text{proj}_{\mathcal{S}}(\mathbf{x}) \right\|_2. \tag{32}$$

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

$$\gamma_0(\mathbf{x}, \mathbf{w}) = q(\mathbf{x})\delta\big(\mathbf{w} - \text{proj}_{\mathcal{S}}(\mathbf{x})\big), \quad \mathbf{x}, \mathbf{w} \in \mathbb{R}^n, \tag{33}$$

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

$$\begin{aligned} W_1(q \| q^{\mathcal{S}}) &\leq \mathbb{E}_{(\mathbf{x},\mathbf{w})\sim\gamma_0} \left\| \mathbf{x} - \mathbf{w} \right\|_2, && (34)\\ &= \int q(\mathbf{x})\delta\big(\mathbf{w} - \text{proj}_{\mathcal{S}}(\mathbf{x})\big) \left\| \mathbf{x} - \mathbf{w} \right\|_2 d\mathbf{x} d\mathbf{w}, && (35)\\ &= \int q(\mathbf{x}) \left\| \mathbf{x} - \text{proj}_{\mathcal{S}}(\mathbf{x}) \right\|_2 d\mathbf{x}, && (36)\\ &= \mathbb{E}_{\mathbf{x}\sim q} \left\| \mathbf{x} - \text{proj}_{\mathcal{S}}(\mathbf{x}) \right\|_2. && (37) \end{aligned}$$

On the other hand, by definition of orthogonal projection,

$$\left\| \mathbf{x} - \text{proj}_{\mathcal{S}}(\mathbf{x}) \right\|_2 \leq \left\| \mathbf{x} - \mathbf{w} \right\|_2, \quad \forall\ \mathbf{w} \in \text{supp}(q^{\mathcal{S}}). \tag{38}$$

Therefore,

$$\begin{aligned} \mathbb{E}_{\mathbf{x}\sim q} \left\| \mathbf{x} - \text{proj}_{\mathcal{S}}(\mathbf{x}) \right\|_2 &\leq \mathbb{E}_{(\mathbf{x},\mathbf{w})\sim\gamma} \left\| \mathbf{x} - \mathbf{w} \right\|_2, \quad \gamma \in \Gamma(q, q^{\mathcal{S}}). && (39)\\ \Rightarrow \mathbb{E}_{\mathbf{x}\sim q} \left\| \mathbf{x} - \text{proj}_{\mathcal{S}}(\mathbf{x}) \right\|_2 &\leq \inf_{\gamma\in\Gamma(q,q^{\mathcal{S}})} \mathbb{E}_{(\mathbf{x},\mathbf{w})\sim\gamma} \left\| \mathbf{x} - \mathbf{w} \right\|_2, && (40)\\ &= W_1(q \| q^{\mathcal{S}}). && (41) \end{aligned}$$

Equations (37) and (41) imply

$$W_1(q \| q^{\mathcal{S}}) = \mathbb{E}_{\mathbf{x}\sim q} \left\| x - \text{proj}_{\mathcal{S}}(\mathbf{x}) \right\|_2. \tag{42}$$

□

With all the tools in place, we now proceed to prove Theorem 3.2. Consider the optimization problem stated in Eq. (8):

$$\hat{\theta}, \hat{\phi} = \arg\min_{\theta,\phi} -\mathcal{L}_M(\theta, \phi) \quad \text{subject to} \quad \mathbb{E}_{\mathbf{g}\sim q_{\mathbf{g}}} \mathbb{E}_{\mathbf{f}\sim p_\phi(\cdot|\mathbf{g})} \|\Phi\mathbf{f} - \Phi\text{proj}_{\mathcal{S}_k}(\mathbf{f})\|_1 < \epsilon. \tag{8}$$

Theorem 3.2. *If the following hold:*

1. *$W_1(q_{\mathbf{f}} \parallel q_{\mathbf{f}}^{\mathcal{S}_k}) \leq \epsilon'$ (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 $\chi_{\mathbf{n}}(\mathbf{v})$ has full support over $\mathbb{C}^m$, and*
4. *(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. (8) (G_θ and h_ϕ have sufficient capacity),*

then the distrubution $p_{\hat{\theta}}$ recovered via Eq. (8) is close to the true object distribution, in terms of the Wasserstein distance i.e.

$$W_1(p_{\hat{\theta}} \parallel q_{\mathbf{f}}) \leq \left(1 + \frac{1}{\sqrt{1-\delta_k}} \|H\|_2\right)(\epsilon + \epsilon'). \tag{43}$$

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. (8), 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. (8), the maximum value of $\mathcal{L}_M$ under Eq. (8) is $\mathbb{E}_{\mathbf{g}\sim q_{\mathbf{g}}} \log q_{\mathbf{g}}(\mathbf{g})$. Therefore, according to Eq. (25), for the estimated $\hat{\theta}$ and $\hat{\phi}$, $\mathcal{L}(\hat{\theta}, \hat{\phi}) = \mathbb{E}_{\mathbf{g}\sim q_{\mathbf{g}}} \log q_{\mathbf{g}}(\mathbf{g})$,

$$\psi_{\hat{\theta}} = q_{\mathbf{g}} \text{ and } p_{\hat{\phi}}(\cdot|\mathbf{g}) = p_{\hat{\theta}}(\cdot|\mathbf{g}). \tag{44}$$

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

$$\begin{aligned} \|\mathbf{f}_1 - \mathbf{f}_2\|_2 &= \left\|\mathbf{f}_1 - \mathbf{f}_1^{\mathcal{S}} + \mathbf{f}_2^{\mathcal{S}} - \mathbf{f}_2 + \mathbf{f}_1^{\mathcal{S}} - \mathbf{f}_2^{\mathcal{S}}\right\|_2, & (45)\\ &\leq \left\|\mathbf{f}_1 - \mathbf{f}_1^{\mathcal{S}}\right\|_2 + \left\|\mathbf{f}_2^{\mathcal{S}} - \mathbf{f}_2\right\|_2 + \left\|\mathbf{f}_1^{\mathcal{S}} - \mathbf{f}_2^{\mathcal{S}}\right\|_2, & (46)\end{aligned}$$

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,

$$\|\mathbf{f}_1 - \mathbf{f}_2\|_2 \leq \left\|\mathbf{f}_1 - \mathbf{f}_1^{\mathcal{S}}\right\|_2 + \left\|\Phi^+\right\|_2 \left\|\mathbf{c}_2 - \mathbf{c}_2^{\mathcal{S}}\right\|_2 + \left\|\mathbf{f}_1^{\mathcal{S}} - \mathbf{f}_2^{\mathcal{S}}\right\|_2, \tag{47}$$

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,

$$\begin{aligned} \left\|\mathbf{y}_1^{\mathcal{S}} - \mathbf{y}_2^{\mathcal{S}}\right\|_2 &= \left\|\mathbf{y}_1^{\mathcal{S}} - \mathbf{y}_1 + \mathbf{y}_2 - \mathbf{y}_2^{\mathcal{S}} + \mathbf{y}_1 - \mathbf{y}_2\right\|_2 & (48)\\ &\leq \left\|\mathbf{y}_1 - \mathbf{y}_1^{\mathcal{S}}\right\|_2 + \left\|\mathbf{y}_2 - \mathbf{y}_2^{\mathcal{S}}\right\|_2 + \|\mathbf{y}_1 - \mathbf{y}_2\|_2, & (49)\\ &\leq \|H\|_2 \left\|\mathbf{f}_1 - \mathbf{f}_1^{\mathcal{S}}\right\|_2 + \left\|H\Phi^+\right\|_2 \left\|\mathbf{c}_2 - \mathbf{c}_2^{\mathcal{S}}\right\|_2 + \|\mathbf{y}_1 - \mathbf{y}_2\|_2. & (50)\end{aligned}$$

Now, the restricted isometry property (RIP) on H defined in Definition 2.1 implies

$$\left\|\mathbf{f}_1^{\mathcal{S}} - \mathbf{f}_2^{\mathcal{S}}\right\|_2 \leq \frac{1}{\sqrt{1-\delta_k}} \left\|\mathbf{y}_1^{\mathcal{S}} - \mathbf{y}_2^{\mathcal{S}}\right\|. \tag{51}$$

Therefore, Equations (47), (50) and (51) give

$$\begin{aligned} \|\mathbf{f}_1 - \mathbf{f}_2\| \leq{}& \left\|\mathbf{f}_1 - \mathbf{f}_1^{\mathcal{S}}\right\|_2 + \left\|\Phi^+\right\|_2 \left\|\mathbf{c}_2 - \mathbf{c}_2^{\mathcal{S}}\right\|_2 \\ &+ \frac{1}{\sqrt{1-\delta_k}}\Big[\|H\|_2 \left\|\mathbf{f}_1 - \mathbf{f}_1^{\mathcal{S}}\right\|_2 \\ &+ \left\|H\Phi^+\right\|_2 \left\|\mathbf{c}_2 - \mathbf{c}_2^{\mathcal{S}}\right\|_2 + \|\mathbf{y}_1 - \mathbf{y}_2\|_2\Big]. & (52)\\ \leq{}& \alpha\Big(\left\|\mathbf{f}_1 - \mathbf{f}_1^{\mathcal{S}}\right\|_2 + \left\|\Phi^+\right\|_2 \left\|\mathbf{c}_2 - \mathbf{c}_2^{\mathcal{S}}\right\|_2\Big) \\ &+ \frac{1}{\sqrt{1-\delta_k}} \|\mathbf{y}_1 - \mathbf{y}_2\|_2, & (53)\\ \text{where } \alpha ={}& 1 + \frac{1}{\sqrt{1-\delta_k}} \|H\|_2. & (54)\end{aligned}$$

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 $\beta \in B$,

$$\mathbb{E}_{(\mathbf{f}_1,\mathbf{f}_2)\sim\beta} \left\|\mathbf{f}_1 - \mathbf{f}_2\right\|_2 \leq \alpha\Big[\mathbb{E}_{\mathbf{f}_1\sim q_{\mathbf{f}}} \left\|\mathbf{f}_1 - \mathbf{f}_1^{\mathcal{S}}\right\|_2 + \|\Phi^+\|_2 \mathbb{E}_{\mathbf{f}_2\sim p_{\hat{\theta}}} \left\|\Phi\mathbf{f}_2 - \Phi\mathbf{f}_2^{\mathcal{S}}\right\|_2\Big] + \frac{1}{\sqrt{1-\delta_k}} \mathbb{E}_{(\mathbf{f}_1,\mathbf{f}_2)\sim\beta} \left\|H\mathbf{f}_1 - H\mathbf{f}_2\right\|_2 . \tag{55}$$

From Lemma A.1, we have

$$\mathbb{E}_{\mathbf{f}_1\sim q_{\mathbf{f}}} \left\|\mathbf{f}_1 - \mathbf{f}_1^{\mathcal{S}}\right\|_2 = W_1(q_{\mathbf{f}} \| q_{\mathbf{f}}^{\mathcal{S}_k}) \leq \epsilon'. \tag{56}$$

Also, from Eq. (6),

$$\begin{aligned}
\mathbb{E}_{\mathbf{f}_2\sim p_{\hat{\theta}}} \left\|\Phi\mathbf{f}_2 - \Phi\mathbf{f}_2^{\mathcal{S}}\right\|_2 &= \mathbb{E}_{\mathbf{g}\sim q_{\mathbf{g}}} \mathbb{E}_{\mathbf{f}_2\sim p_{\hat{\theta}}(\cdot|\mathbf{g})} \left\|\Phi\mathbf{f}_2 - \Phi\mathbf{f}_2^{\mathcal{S}}\right\|_2 , && (57)\\
&= \mathbb{E}_{\mathbf{g}\sim q_{\mathbf{g}}} \mathbb{E}_{\mathbf{f}\sim p_{\hat{\phi}}(\cdot|\mathbf{g})} \left\|\Phi\mathbf{f} - \Phi\mathrm{proj}_{\mathcal{S}_k}(\mathbf{f})\right\|_2 , && (58)\\
&\leq \mathbb{E}_{\mathbf{g}\sim q_{\mathbf{g}}} \mathbb{E}_{\mathbf{f}\sim p_{\hat{\phi}}(\cdot|\mathbf{g})} \left\|\Phi\mathbf{f} - \Phi\mathrm{proj}_{\mathcal{S}_k}(\mathbf{f})\right\|_1 , && (59)\\
&\leq \epsilon. && (60)
\end{aligned}$$

Taking the infimum of Eq. (55) over $\beta \in B$, we get

$$\inf_{\beta\in B} \mathbb{E}_{(\mathbf{f}_1,\mathbf{f}_2)\sim\beta} \left\|\mathbf{f}_1 - \mathbf{f}_2\right\|_2 \leq \alpha(\epsilon' + \|\Phi^+\|_2\epsilon) + \frac{1}{\sqrt{1-\delta_k}} \inf_{\beta\in B} \mathbb{E}_{(\mathbf{f}_1,\mathbf{f}_2)\sim\beta} \left\|H\mathbf{f}_1 - H\mathbf{f}_2\right\|_2 . \tag{61}$$

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. (61) is $W_1(q_{\mathbf{y}} \| \rho_{\hat{\theta}})$:

$$W_1(q_{\mathbf{y}} \| \rho_{\hat{\theta}}) = \inf_{\beta\in B} \mathbb{E}_{(\mathbf{f}_1,\mathbf{f}_2)\sim\beta} \left\|H\mathbf{f}_1 - H\mathbf{f}_2\right\|_2 . \tag{62}$$

From Eq. (44), since $q_{\mathbf{g}} = \psi_{\hat{\theta}}$, Lemma 3.1 implies $q_{\mathbf{y}} = \rho_{\hat{\theta}}$

$$\Rightarrow W_1(q_{\mathbf{y}} \| \rho_{\hat{\theta}}) = 0. \tag{63}$$

Combining with Eq. (61), and setting $\|\Phi^+\| \leq 1$ and α according to Eq. (54), we get

$$W_1(p_{\hat{\theta}} \| q_{\mathbf{f}}) \leq \left(1 + \frac{1}{\sqrt{1-\delta_k}} \|H\|_2\right)(\epsilon + \epsilon') \tag{64}$$

□