Uncertainty Autoencoders: Learning Compressed Representations via Variational Information Maximization

Aditya Grover
Stanford University

Stefano Ermon
Stanford University

Abstract

Compressed sensing techniques enable efficient acquisition and recovery of sparse, high-dimensional data signals via low-dimensional projections. In this work, we propose *Uncertainty Autoencoders*, a learning framework for unsupervised representation learning inspired by compressed sensing. We treat the low-dimensional projections as noisy latent representations of an autoencoder and directly learn both the acquisition (*i.e.*, encoding) and amortized recovery (*i.e.*, decoding) procedures. Our learning objective optimizes for a tractable variational lower bound to the mutual information between the datapoints and the latent representations. We show how our framework provides a unified treatment to several lines of research in dimensionality reduction, compressed sensing, and generative modeling. Empirically, we demonstrate a 32% improvement on average over competing approaches for the task of statistical compressed sensing of high-dimensional datasets.

1 INTRODUCTION

The goal of unsupervised representation learning is to learn transformations of the input data which succinctly capture the statistics of an underlying data distribution [1]. In this work, we propose a learning framework for unsupervised representation learning inspired by compressed sensing. Compressed sensing is a class of techniques used to efficiently acquire and recover high-dimensional data using compressed measurements much fewer than the data dimensionality. The celebrated results in compressed sensing posit that *sparse*, high-dimensional datapoints can be acquired using much fewer measurements (roughly logarithmic) than the data dimensionality [2, 3, 4]. The acquisition is done using certain classes of random matrices and the recovery procedure is based on LASSO [5, 6].

The assumptions of sparsity are fairly general and can be applied "out-of-the-box" for many data modalities, *e.g.*, images and audio are typically sparse in the wavelet and Fourier basis respectively. However, such assumptions ignore the statistical nature of many real-world problems. For representation learning in particular, we have access to a training dataset from an underlying domain. In this work, we use this data to *learn* the acquisition and recovery procedures, thereby sidestepping generic sparsity assumptions. In particular, we view the compressed measurements as the latent representations of an *uncertainty autoencoder*.

An uncertainty autoencoder (UAE) parameterizes both the acquisition and recovery procedures for compressed sensing. The learning objective for a UAE is based on the InfoMax principle, which seeks to learn encodings that maximize the mutual information between the observed datapoints and noisy representations [7]. Since the mutual information is typically intractable in high-dimensions, we instead maximize tractable variational lower bounds [8, 9]. In doing so, we introduce a parameteric decoder that is trained to recover the original datapoint via its noisy representation. Unlike LASSO-based recovery, a parametric decoder *amortizes* the recovery process, which requires only a forward pass through the decoder at test time and thus enables scalability to large datasets [10, 11].

Notably, the framework of uncertainty autoencoders unifies and extends several lines of prior research in unsupervised representation learning. First, we show theoretically under suitable assumptions that an uncertainty autoencoder is an *implicit* generative model of the underlying data distribution [12], *i.e.*, a UAE permits sampling from the learned data distribution even

A preliminary version titled "Variational Compressive Sensing using Uncertainty Autoencoders" appeared at the Uncertainty in Deep Learning Workshop at UAI, 2018.

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

though it does not specify an explicit likelihood function. Hence, it directly contrasts with variational autoencoders (VAE) which specify a likelihood function (which is intractable and approximated by a tractable evidence lower bound) [13]. Unlike a VAE, a UAE does not require specifying a prior over the latent representations and hence offsets pathologically observed scenarios that cause the latent representations to be uninformative when used with expressive decoders [14].

Next, we show that an uncertainty autoencoder, under suitable assumptions, is a generalization of principal component analysis (PCA). While earlier results connecting standard autoencoders with PCA assume linear encodings and decodings [15, 16, 17], our result surprisingly holds even for non-linear decodings. In practice, *linear* encodings learned jointly with non-linear decodings based on the UAE objective vastly outperform the linear encodings obtained via PCA. For dimensionality reduction on the MNIST dataset, we observed an average improvement of 5.33% over PCA when the low-dimensional representations are used for classification under a wide range of settings.

We evaluate UAEs for statistical compressed sensing of high-dimensional datasets. On the MNIST, Omniglot, and CelebA datasets, we observe average improvements of 38%, 31%, and 28% in recovery over the closest benchmark across all measurements considered. Finally, we demonstrate that uncertainty autoencoders demonstrate good generalization performance across domains in experiments where the encoder/decoder trained on a source dataset are transferred over for compressed sensing of another target dataset.

2 PRELIMINARIES

We use upper case to denote probability distributions and assume they admit absolutely continuous densities on a suitable reference measure, denoted by lower case notation. We also use upper and lower case for random variables and their realizations respectively.

Compressed sensing (CS). Let the datapoint and measurements be denoted with multivariate random variables $X \in \mathbb{R}^n$ and $Y \in \mathbb{R}^m$ respectively. The goal is to recover X given the measurements Y. For the purpose of compressed sensing, we assume $m < n$ and relate these variables through a measurement matrix $W \in \mathbb{R}^{m \times l}$ and a parameterized acquisition function $f_\psi : \mathbb{R}^n \to \mathbb{R}^l$ (for any integer $l > 0$) such that:

$$y = W f_\psi(x) + \epsilon \tag{1}$$

where ϵ is the measurement noise. If we let $f_\psi(\cdot)$ be the identity function (*i.e.*, $f_\psi(x) = x$ for all x), then we recover a standard system of underdetermined linear equations where measurements are linear combinations of the datapoint corrupted by noise. In all other cases, the acquisition function transforms x such that $f_\psi(x)$ is potentially more amenable for compressed sensing. For instance, $f_\psi(\cdot)$ could specify a change of basis that encourages sparsity, *e.g.*, a Fourier basis for audio. Note that we allow the codomain of the mapping $f_\psi(\cdot)$ to be defined on a higher or lower dimensional space (*i.e.*, $l \neq n$ in general).

Sparse CS. To obtain nontrivial solutions to an underdetermined system, X is assumed to be sparse in some basis B. We are not given any additional information about X. The measurement matrix W is a random Gaussian matrix and the recovery is done via LASSO [2, 3, 4]. LASSO solves for a convex ℓ_1-minimization problem such that the reconstruction $\widehat{x}$ for any datapoint x is given as: $\hat{x} = \arg\min_x \|Bx\|_1 + \lambda \|y - Wx\|_2^2$ where $\lambda > 0$ is a tunable hyperparameter.

Statistical CS. In *statistical* compressed sensing [18], we are additionally given access to a set of signals $\mathcal{D}$, such that each $x \in \mathcal{D}$ is assumed to be sampled i.i.d. from a data distribution Q_{data}. Using this dataset, we learn the the measurement matrix W and the acquisition function $f_\psi(\cdot)$ in Eq. (1).

At test time, we directly observe the measurements y_{test} that are assumed to satisfy Eq. (1) for a target datapoint $x_{\text{test}} \sim Q_{\text{data}}(X)$ and the task is to provide an accurate reconstruction $\hat{x}_{\text{test}}$. Evaluation is based on the reconstruction error between x_{test} and $\hat{x}_{\text{test}}$. Particularly relevant to this work, we can optionally *learn* a recovery function $g_\theta : \mathbb{R}^m \to \mathbb{R}^n$ to reconstruct X given the measurements Y.

This amortized approach [11] is in contrast to standard LASSO-based decoding which solves an optimization problem *for every new datapoint* at test time. If we learned the recovery function $g_\theta(\cdot)$ during training, then $\hat{x}_{\text{test}} = g_\theta(y_{\text{test}})$ and the ℓ_2 error is given by $\|x_{\text{test}} - g_\theta(y_{\text{test}})\|_2$. Such a recovery process requires only a function evaluation at test time and permits scaling to large datasets [10, 11].

Autoencoders. An autoencoder is a pair of parameterized functions (e, d) designed to encode and decode datapoints. For a standard autoencoder, let $e : \mathbb{R}^n \to \mathbb{R}^m$ and $d : \mathbb{R}^m \to \mathbb{R}^n$ denote the encoding and decoding functions respectively for an n-dimensional datapoint and an m-dimensional latent space. The learning objective minimizes the l_2 reconstruction error over a dataset $\mathcal{D}$:

$$\min_{e,d} \sum_{x \in \mathcal{D}} \|x - d(e(x))\|_2^2 \tag{2}$$

where the encoding and decoding functions are typically parameterized using neural networks.

3 UNCERTAINTY AUTOENCODER

Consider a joint distribution between the signals X and the measurements Y, which factorizes as $Q_\phi(X,Y) = Q_{\text{data}}(X)Q_\phi(Y|X)$. Here, $Q_{\text{data}}(X)$ is a fixed data distribution and $Q_\phi(Y|X)$ is a parameterized observation model that depends on the measurement noise ϵ, as given by Eq. (1). In particular, ϕ corresponds to collectively the set of measurement matrix parameters W and the acquisition function parameters ψ. For instance, for isotropic Gaussian noise ϵ with a fixed variance σ^2, we have $Q_\phi(Y|X) = \mathcal{N}(Wf_\psi(X), \sigma^2 I_m)$.

In an uncertainty autoencoder, we wish to learn the parameters ϕ that permit efficient and accurate recovery of a signal X using the measurements Y. In order to do so, we propose to maximize the mutual information between X and Y:

$$\begin{aligned}\max_{\phi} \; I_\phi(X,Y) &= \int q_\phi(x,y) \log \frac{q_\phi(x,y)}{q_{\text{data}}(x)q_\phi(y)} \mathrm{d}x\mathrm{d}y \\ &= H(X) - H_\phi(X|Y) \end{aligned} \tag{3}$$

where H denotes differential entropy. The intuition is simple: if the measurements preserve maximum information about the signal, we can hope that recovery will have low reconstruction error. We formalize this intuition by noting that this objective is equivalent to maximizing the average log-posterior probability of X given Y. In fact, in Eq. (3), we can omit the term corresponding to the data entropy (since it is independent of ϕ) to get the following equivalent objective:

$$\max_{\phi} -H_\phi(X|Y) = \mathbb{E}_{Q_\phi(X,Y)}[\log q_\phi(x|y)]. \tag{4}$$

Even though the mutual information is maximized and equals the data entropy when $Y = X$, the dimensionality constraints on $m \ll n$, the parametric assumptions on $f_\psi(\cdot)$, and the noise model prohibit learning an identity mapping. Note that the properties of noise ϵ such as the distributional family and sufficient statistics are externally specified. For example, these could be specified based on properties of the measurement device for compressed sensing. More generally for unsupervised representation learning, we treat these properties as hyperparameters tuned based on the reconstruction loss on a held-out set, or any other form of available supervision. It is not suggested to optimize for these statistics during learning as the UAE would tend to shrink this noise to zero to maximize mutual information, thus ignoring measurement uncertainty in the context of compressed sensing and preventing generalization to out-of-distribution examples for representation learning. The theoretical results in Section 4 analyze the effect of noise more formally.

Estimating mutual information between arbitrary high dimensional random variables can be challenging. However, we can lower bound the mutual information by introducing a variational approximation to the model posterior $Q_\phi(X|Y)$ [8]. Denoting this approximation as $P_\theta(X|Y)$, we get the following lower bound:

$$I_\phi(X,Y) \geq H(X) + \mathbb{E}_{Q_\phi(X,Y)}\left[\log p_\theta(x|y)\right]. \tag{5}$$

Comparing Eqs. (3, 4, 5), we can see that the second term in Eq. (5) approximates the intractable negative conditional entropy, $-H_\phi(X|Y)$ with a variational lower bound. Optimizing this bound leads to a decoding distribution given by $P_\theta(X|Y)$ with variational parameters θ. The bound is tight when there is no distortion during recovery, or equivalently when the decoding distribution $P_\theta(X|Y)$ matches the true posterior $Q_\phi(X|Y)$ (*i.e.*, the Bayes optimal decoder).

Stochastic optimization. Formally, the uncertainty autoencoder (UAE) objective is given by:

$$\max_{\theta,\phi} \mathbb{E}_{Q_\phi(X,Y)}\left[\log p_\theta(x|y)\right]. \tag{6}$$

In practice, the data distribution $Q_{\text{data}}(X)$ is unknown and accessible only via a finite dataset $\mathcal{D}$. Hence, expectations with respect to $Q_{\text{data}}(X)$ and its gradients can be estimated using Monte Carlo methods. This allows us to express the UAE objective as:

$$\max_{\theta,\phi} \sum_{x \in \mathcal{D}} \mathbb{E}_{Q_\phi(Y|x)}\left[\log p_\theta(x|y)\right] := \mathcal{L}(\phi, \theta; \mathcal{D}). \tag{7}$$

Tractable evaluation of the above objective is closely tied to the distributional assumptions on the noise model. This could be specified externally based on, *e.g.*, properties of the sensing device in compressed sensing. For the typical case of an isotropic Gaussian noise model, we know that $Q_\phi(Y|X) = \mathcal{N}(Wf_\psi(X), \sigma^2 I_m)$, which is easy-to-sample.

While Monte Carlo gradient estimates with respect to θ can be efficiently obtained via linearity of expectation, gradient estimation with respect to ϕ is challenging since these parameters specify the sampling distribution $Q_\phi(Y|X)$. One solution is to evaluate *score function* gradient estimates along with control variates [19, 20, 21]. Alternatively, many continuous distributions (*e.g.*, the isotropic Gaussian and Laplace distributions) can be *reparameterized* such that it is possible to obtain samples by applying a deterministic transformation to samples from a fixed distribution and typically leads to low-variance gradient estimates [13, 22, 23, 24].

4 THEORETICAL ANALYSIS

In this section, we derive connections of uncertainty autoencoders with generative modeling and Principal Component Analysis (PCA). The proofs of all theoretical results in this section are in Appendix A.

4.1 Implicit generative modeling

Starting from an arbitrary point $x^{(0)} \in \mathbb{R}^n$, define a Markov chain over X, Y with the following transitions:

$$y^{(t)} \sim Q_\phi(Y|x^{(t)}) \tag{8}$$

$$x^{(t+1)} \sim P_\theta(X|y^{(t)}) \tag{9}$$

Theorem 1. *Let θ^*, ϕ^* denote an optimal solution to the UAE objective in Eq. (6). If there exists a ϕ such that $q_\phi(x|y) = p_{\theta^*}(x|y)$ and the Markov chain defined in Eqs. (8, 9) is ergodic, then the stationary distribution of the chain for the parameters ϕ^* and θ^* is given by $Q_{\phi^*}(X, Y)$.*

The above theorem suggests an interesting insight into the behavior of UAEs. Under idealized conditions, the learned model specifies an implicit generative model for $Q_{\phi^*}(X, Y)$. Further, ergodicity can be shown to hold for the isotropic Gaussian noise model.

Corollary 1. *Let θ^*, ϕ^* denote an optimal solution to the UAE objective in Eq. (6). If there exists a ϕ such that $q_\phi(x|y) = p_{\theta^*}(x|y)$ and the noise model is Gaussian, then the stationary distribution of the chain for the parameters ϕ^* and θ^* is given by $Q_{\phi^*}(X, Y)$.*

The marginal of the joint distribution $Q_\phi(X, Y)$ with respect to X corresponds to the data distribution. A UAE hence seeks to learn an *implicit* generative model of the data distribution [25, 12], *i.e.*, even though we do not have a tractable estimate for the likelihood of the model, we can generate samples using the Markov chain transitions defined in Eqs. (8, 9).

4.2 Optimal encodings

A UAE can also be viewed as a dimensionality reduction technique for the dataset $\mathcal{D}$. While in general the encoding performing this reduction can be nonlinear, the case of a linear encoding is one where the projection vectors are given as the rows of the measurement matrix W. The result below characterizes the optimal encoding of the dataset $\mathcal{D}$ with respect to the UAE objective for an isotropic Gaussian noise model.

Theorem 2. *Assume a uniform data distribution over a finite dataset $\mathcal{D}$. Further, we assume that expectations in the UAE objective exist, and the signals and measurement matrices are bounded in ℓ_2/Frobenius norms, i.e., $\|x\|_2 \leq k_1$ for all $x \in \mathcal{D}$, $\|W\|_F \leq k_2$ for some positive constants $k_1, k_2 \in \mathbb{R}^+$. For a linear encoder and isotropic Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$, the optimal measurement matrix W^* that maximizes the mutual information for an optimal decoder in the limit $\sigma \to \infty$ is given as:*

$$W^* = \text{eig}_m \left(\sum_{x_i, x_j \in \mathcal{D}} \left[(x_i - x_j)(x_i - x_j)^T \right] \right)$$

where $\text{eig}_m(M)$ denotes the top-m eigenvectors of the matrix M with the largest eigenvalues (specified up to a positive scaling constant).

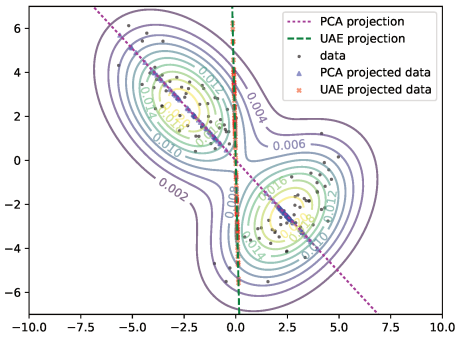
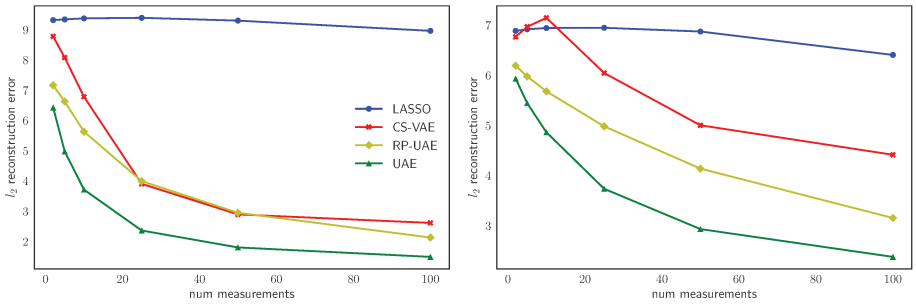


Figure 1: Dimensionality reduction using PCA vs. UAE. Projections of the data (**black points**) on the UAE direction (**green line**) maximize the likelihood of decoding unlike the PCA projection axis (**magenta line**) which collapses many points in a narrow region.

Under the stated assumptions, the above result suggests an interesting connection between UAE and PCA. PCA seeks to find the directions that explain the most variance in the data. Theorem 2 suggests that when the noise in the projected signal is very high, the optimal projection directions (*i.e.*, the rows of W^*) correspond to the principal components of the data signals. We note that this observation comes with a caveat; when the noise variance is high, it will dominate the contribution to the measurements Y in Eq. (1) as one would expect. Hence, the measurements and the signal will have low mutual information even under the optimal measurement matrix W^*.

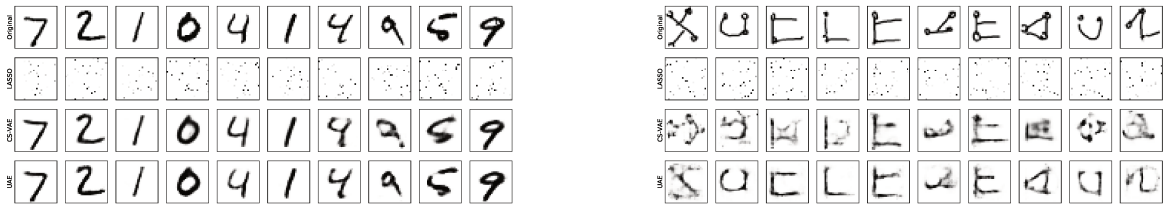
Our assumptions are notably different from prior results in autoencoding drawing connections with PCA. Prior results show that linear encoding *and* decoding in a standard autoencoder recovers the principal components of the data (Eq. (3) in [15], Eq. (1) in [16]). In contrast, Theorem 2 is derived from variational principles and *does not assume linear decoding.*

In general, the behaviors of UAE and PCA can be vastly different. As noted in prior work [8, 26], the principal components may not be the the most informative low-dimensional projections for recovering the



(a) MNIST (b) Omniglot

Figure 2: Test ℓ_2 reconstruction error (per image) for compressed sensing.



(a) MNIST (b) Omniglot

Figure 3: Reconstructions for $m = 25$. **Top:** Original. **Second:** LASSO. **Third:** CS-VAE. **Last:** UAE. 25 projections of the data are sufficient for UAE to reconstruct the original image with high accuracy.

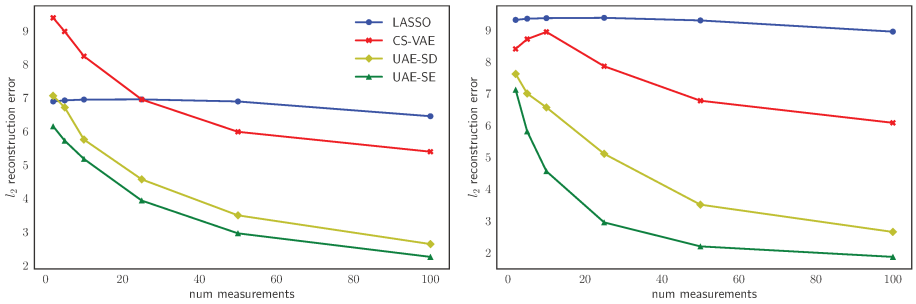
original high-dimensional data back from its projections. A UAE, on the other hand, is explicitly designed to preserve as much information as possible (see Eq. (4)). We illustrate the differences in a synthetic experiment in Figure 1. The true data distribution is an equiweighted mixture of two Gaussians stretched along orthogonal directions. We sample 100 points (black) from this mixture and consider two dimensionality reductions. In the first case, we project the data on the first principal component (blue points on magenta line). This axis captures a large fraction of the variance in the data but collapses data sampled from the bottom right Gaussian in a narrow region. The projections of the data on the UAE axis (red points on green line) are more spread out. This suggest that recovery is easier, even if doing so increases the total variance in the projected space compared to PCA.

5 EXPERIMENTS

5.1 Statistical compressed sensing

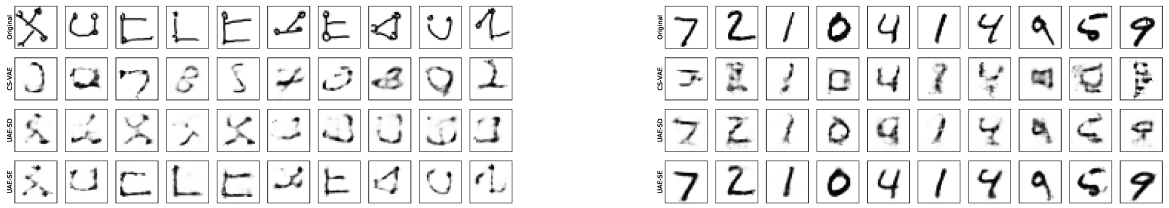
We perform compressed sensing on three datasets: MNIST [27], Omniglot [28], and CelebA dataset [29], with extremely low number of measurements $m \in \{2, 5, 10, 25, 50, 100\}$. We discuss the MNIST and Omniglot datasets here since they have a similar setup. To save space, results on the CelebA dataset are deferred to Appendix B.3. Every image in MNIST and Omniglot has a dimensionality of 28×28. In all our experiments, we assume a Gaussian noise model with $\sigma = 0.1$. We evaluated UAE against:

- **LASSO** decoding with random Gaussian matrices. The MNIST and Omniglot datasets are reasonably sparse in the canonical pixel basis, and hence, we did not observe any gains after applying Discrete Cosine Transform and Daubechies-1 Wavelet Transform.
- **CS-VAE**. This approach to compressed sensing was proposed by [30] and learns a latent variable generative model over the observed variables X and the latent variables Z. Such a model defines a mapping $G : \mathbb{R}^k \to \mathbb{R}^n$ from Z to X, which is given by either the mean function of the observation model for a VAE or the forward deterministic mapping to generate samples for a GAN. We use VAEs in our experiments. Thereafter, using a classic acquisition matrix satisfying a generalized Restricted Eigenvalue Condition (say W) (*e.g.*, random Gaussian matrices), the reconstruction $\hat{x}$ for any datapoint is given as: $\hat{x} = G(\arg\min_z \|y - WG(z)\|_2)$. Intuitively, this procedure seeks the latent vector z such that the corresponding point on the range of G can best approximate the measurements y under the mapping W. We used the default parameter settings and architectures proposed in [30].



(a) Source: MNIST, Target: Omniglot (b) Source: Omniglot, Target: MNIST

Figure 4: Test ℓ_2 reconstruction error (per image) for transfer compressed sensing.



(a) Source: MNIST, Target: Omniglot (b) Source: Omniglot, Target: MNIST

Figure 5: Reconstructions for $m = 25$. **Top:** Target. **Second:** CS-VAE. **Third:** UAE-SD. **Last:** UAE-SE.

- **RP-UAE.** To independently evaluate the effect of variational decoding, this ablation baseline encodes the data using Gaussian random projections (RP) and trains the decoder based on the UAE objective. Since LASSO and CS-VAE both use an RP encoding, the differences in performance would arise only due to the decoding procedures.

The UAE decoder and the CS-VAE encoder/decoder are multi-layer perceptrons consisting of two hidden layers with 500 units each. For a fair comparison with random Gaussian matrices, the UAE encoder is linear. Further, we perform ℓ_2 regularization on the norm of W. This helps in generalization to test signals outside the train set and is equivalent to solving the Lagrangian of a constrained UAE objective:

$$\max_{\theta,\phi} \mathbb{E}_{Q_\phi(X,Y)} \left[\log p_\theta(x|y)\right] \text{ subject to } \|W\|_F \leq k.$$

The Lagrangian parameter is chosen by line search on the above objective. The constraint ensures that UAE does not learn encodings W that trivially scale the measurement matrix to overcome noise. For each m, we choose k to be the expected norm of a random Gaussian matrix of dimensions $n \times m$ for fair comparisons with other baselines. In practice, the norm of the learned W for a UAE is much smaller than those of random Gaussian matrices suggesting that the observed performance improvements are non-trivial.

Results. The ℓ_2 reconstruction errors on the standard test sets are shown in Figure 2. For both datasets, we observe that UAE drastically outperforms both LASSO and CS-VAE for all values of m considered. LASSO (blue curves) is unable to reconstruct with such few measurements. The CS-VAE (red) error decays much more slowly compared to UAE as m grows. Even the RP-UAE baseline (yellow), which trains the decoder keeping the encoding fixed to a random projection, outperforms CS-VAE. Jointly training the encoder and the decoder using the UAE objective (green) exhibits the best performance. These results are also reflected qualitatively for the reconstructed test signals shown in Figure 3 for $m = 25$ measurements.

5.2 Transfer compressed sensing

To test the generalization of the learned models to similar, unseen datasets, we consider the task of *transfer* compressed sensing task introduced in [31].

Experimental setup. We train the models on a source domain that is related to a target domain. Since the dimensions of MNIST and Omniglot images match, transferring from one domain to another requires no additional processing. For UAE, we consider two variants. In UAE-SE, we used the encodings from the source domain and retrain the decoder on the target domain. For UAE-SD, we use source decoder and retrain the encoder on the target domain.

Table 1: PCA vs. UAE. Average test classification accuracy for the MNIST dataset.

Dimensions	Method	kNN	DT	RF	MLP	AdaB	NB	QDA	SVM
2	PCA	0.4078	0.4283	0.4484	0.4695	0.4002	0.4455	0.4576	0.4503
	UAE	**0.4644**	**0.5085**	**0.5341**	**0.5437**	**0.4248**	**0.5226**	**0.5316**	**0.5256**
5	PCA	0.7291	0.5640	0.6257	0.7475	0.5570	0.6587	0.7321	0.7102
	UAE	**0.8115**	**0.6331**	**0.7094**	**0.8262**	**0.6164**	**0.7286**	**0.7961**	**0.7873**
10	PCA	0.9257	**0.6354**	0.6956	0.9006	0.7025	0.7789	0.8918	0.8440
	UAE	**0.9323**	0.5583	**0.7362**	**0.9258**	**0.7165**	**0.7895**	**0.9098**	**0.8753**
25	PCA	**0.9734**	**0.6382**	0.6889	0.9521	0.7234	**0.8635**	0.9572	0.9194
	UAE	0.9730	0.5407	**0.7022**	**0.9614**	**0.7398**	0.8306	**0.9580**	**0.9218**
50	PCA	0.9751	**0.6381**	0.6059	0.9580	**0.7390**	**0.8786**	0.9632	0.9376
	UAE	**0.9754**	0.5424	**0.6765**	**0.9597**	0.7330	0.8579	**0.9638**	**0.9384**
100	PCA	**0.9734**	0.6380	0.4040	0.9584	0.7136	0.8763	0.9570	0.9428
	UAE	0.9731	**0.6446**	**0.6241**	**0.9597**	**0.7170**	**0.8809**	**0.9595**	**0.9431**

Results. The ℓ_2 reconstruction errors are shown in Figure 4. LASSO (blue curves) does not involve any learning, and hence its performance is same as Figure 2. The CS-VAE (red) performance degrades significantly in comparison, even performing worse than LASSO in some cases. The UAE based methods outperform these approaches and UAE-SE (green) fares better than UAE-SD (yellow). Qualitative differences are highlighted in Figure 5 for $m = 25$ measurements.

5.3 Dimensionality reduction

Dimensionality reduction is a common preprocessing technique for specifying features for classification. We compare PCA and UAE on this task. While Theorem 2 posits that the two techniques are equivalent in the regime of high noise given optimal UAE decodings, we set the noise as a hyperparameter based on a validation set to enable out-of-sample generalization.

Setup. We learn the principal components and UAE projections on the MNIST training set for varying number of dimensions. We then learn classifiers based on the these projections. Again, we use a linear encoder for the UAE for a fair evaluation. Since the inductive biases vary across different classifiers, we considered 8 commonly used classifiers: k-Nearest Neighbors (kNN), Decision Trees (DT), Random Forests (RF), Multilayer Perceptron (MLP), AdaBoost (AdaB), Gaussian Naive Bayes (NB), Quadratic Discriminant Analysis (QDA), and Support Vector Machines (SVM) with a linear kernel.

Results. The performance of the PCA and UAE feature representations for different number of dimensions is shown in Table 1. We find that UAE outperforms PCA in a majority of the cases. Further, this trend is largely consistent across classifiers. The improvements are especially high when the number of dimensions is low, suggesting the benefits of UAE as a dimensionality reduction technique for classification.

6 RELATED WORK

In this section, we contrast uncertainty autoencoders with related works in autoencoding, compressed sensing, and mutual information maximization.

Autoencoders. To contrast uncertainty autoencoders with other commonly used autoencoding schemes, consider a UAE with a Gaussian observation model with fixed isotropic covariance for the decoder of all the autoencoding objectives we discuss subsequently. The UAE objective can be simplified as:

$$\min_{\theta,\phi} \mathbb{E}_{x,y\sim Q_\phi(X,Y)} \left[\|x - g_\theta(y)\|_2^2\right]$$

Standard Autoencoder. If we assume no measurement noise (*i.e.*, $\epsilon = 0$) and assume the observation model $P_\theta(X|Y)$ to be a Gaussian with mean $g_\theta(Y)$ and a fixed isotropic Σ, then the UAE objective reduces to minimizing the mean squared error between the true and recovered datapoint:

$$\min_{\theta,W,\psi} \mathbb{E}_{x\sim Q_{\text{data}}(X)} \left[\|x - g_\theta(W f_\psi(x))\|_2^2\right]$$

This special case of a UAE corresponds to a standard autoencoder [32] where the measurements Y signify a hidden representation for X. However, this case lacks the interpretation of an implicit generative model since the assumptions of Theorem 1 do not hold.

Denoising Autoencoders. A DAE [33] adds noise at the level of the input datapoint X to learn robust representations. For a UAE, the noise model is defined at the level of the compressed measurements. Again, with the assumptions of a Gaussian decoder, the DAE objective can be expressed as:

$$\min_{\theta,W,\psi} \mathbb{E}_{x\sim Q_{\text{data}}(X),\tilde{x}\sim C(\tilde{X}|x)} \left[\|x - g(W f_\psi(\tilde{x}))\|_2^2\right]$$

where $C(\cdot|X)$ is some predefined noise corruption model. Similar to Theorem 1, a DAE also learns an implicit model of the data distribution [34, 35].

Variational Autoencoders. A VAE [13, 22] explicitly learns a latent variable model $P_\theta(X, Y)$ for the dataset. The learning objective is a variational lower bound to the marginal log-likelihood assigned by the model to the data $\mathcal{X}$, which notationally corresponds to $\mathbb{E}_{Q_{\text{data}}(X)}[\log P_\theta(x)]$. The variational objective that maximizes this quantity can be simplified as:

$$\min_{\theta,\phi} \quad \mathbb{E}_{x,y\sim Q_\phi(X,Y)}\left[\|x - g_\theta(y)\|_2^2\right] + \mathbb{E}_{x\sim Q_{\text{data}}}\left[KL(Q_\phi(Y|x), P(Y))\right]$$

The learning objective includes a reconstruction error term, akin to the UAE objective. Crucially, it also includes a regularization term to minimize the KL divergence of the variational posterior over Y with a prior distribution over Y. A key difference is that a UAE does not explicitly need to model the prior distribution over Y. On the downside, a VAE can perform efficient ancestral sampling while a UAE requires running relatively expensive Markov Chains to obtain samples.

Recent works have attempted to unify the variants of variational autoencoders through the lens of mutual information [36, 37, 14]. These works also highlight scenarios where the VAE can learn to ignore the latent code in the presence of a strong decoder thereby affecting the reconstructions to attain a lower KL loss. One particular variant, the β-VAE, weighs the additional KL regularization term with a positive factor β and can effectively learn disentangled representations [38, 39]. Although [38] does not consider this case, the UAE can be seen as a β-VAE with $\beta = 0$.

To summarize, our uncertainty autoencoding formulation provides a combination of unique desirable properties for representation learning that are absent in prior autoencoders. As discussed, a UAE defines an implicit generative model without specifying a prior (Theorem 1) even under realistic conditions (Corollary 1; unlike DAEs) and has rich connections with PCA even for non-linear decoders (Theorem 2; unlike any kind of existing autoencoder).

Generative modeling and compressed sensing. The closely related works of [30, 31] also use generative models for compressed sensing. As highlighted in Section 5, their approach is radically different from UAE. Similar to [30], a UAE learns a data distribution. However, in doing so, it additionally learns an acquisition/encoding function and a recovery/decoding function, unlike [30, 31] which rely on generic random matrices and ℓ_2 decoding. The cost of implicit learning in a UAE is that some of its inference capabilities, such as likelihood evaluation and sampling, are intractable or require running Markov chains. However, these inference queries are orthogonal to compressed sensing. Finally, our decoding is amortized and scales to large datasets, unlike [30, 31] which solve an independent optimization problem for each test datapoint.

Mutual information maximization. The principle of mutual information maximization, often referred to as InfoMax in prior work, was first proposed for learning encodings for communication over a noisy channel [7]. The InfoMax objective has also been applied for statistical compressed sensing for learning both linear and non-linear encodings [26, 40, 41]. Our work differs from these existing frameworks in two fundamental ways. First, we optimize for a tractable variational lower bound to the MI that which allows our method to scale to high-dimensional data. Second, we learn an amortized [10, 11] decoder in addition to the encoder that sidesteps expensive, per-example optimization for the test datapoints.

Further, we improve upon the *IM algorithm* proposed originally for variational information maximization [8]. While the IM algorithm proposes to optimize the lower bound on the mutual information in alternating "wake-sleep" phases for optimizing the encoder ("wake") and decoder ("sleep") analogous to the expectation-maximization procedure used in [26], we optimize the encoder and decoder jointly using a single consistent objective leveraging recent advancements in gradient based variational stochastic optimization.

7 CONCLUSION

In this work, we presented uncertainty autoencoders (UAE), a framework for unsupervised representation learning via variational maximization of mutual information between an input signal and its latent representation. We presented connections of our framework with many related threads of research, in particular with respect to implicit generative modeling and principal component analysis. Empirically, we showed that UAEs are a natural candidate for statistical compressed sensing, wherein we can learn the acquisition and recovery functions jointly.

In the future, it would be interesting to incorporate advancements in compressed sensing based on complex neural network architectures [42, 43, 44, 45, 46] within the UAE framework for real world applications, *e.g.*, medical imaging. Unlike the rich theory surrounding the compressed sensing of sparse signals, a similar theory surrounding generative model-based priors on the signal distribution is lacking. Recent works have made promising progress in developing a theory of SGD based recovery methods for nonconvex inverse problems, which continues to be an exciting direction for future work [30, 47, 31, 48].

Acknowledgements

This research was supported by NSF (#1651565, #1522054, #1733686), ONR (N00014-19-1-2145), AFOSR (FA9550-19-1-0024), and FLI. AG is supported by a Microsoft Research Ph.D. fellowship and a Stanford Data Science scholarship. We are thankful to Daniel Levy and Yang Song for helpful discussions on a proof and Kristy Choi, Manik Dhar, Neal Jean, and Ben Poole for helpful comments.

References

[1] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[2] Emmanuel J Candès and Terence Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

[3] David L Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[4] Emmanuel J Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.

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

[6] Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and Dantzig selector. *The Annals of Statistics*, pages 1705–1732, 2009.

[7] Ralph Linsker. How to generate ordered maps by maximizing the mutual information between input and output signals. *Neural computation*, 1(3):402–411, 1989.

[8] David Barber and Felix Agakov. The IM algorithm: A variational approach to information maximization. In *Advances in Neural Information Processing Systems*, 2003.

[9] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017.

[10] Samuel Gershman and Noah Goodman. Amortized inference in probabilistic reasoning. In *Proceedings of the annual meeting of the cognitive science society*, volume 36, 2014.

[11] Rui Shu, Hung H Bui, Shengjia Zhao, Mykel J Kochenderfer, and Stefano Ermon. Amortized inference regularization. In *Advances in Neural Information Processing Systems*, 2018.

[12] Shakir Mohamed and Balaji Lakshminarayanan. Learning in implicit generative models. *arXiv preprint arXiv:1610.03483*, 2016.

[13] Diederik Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

[14] Xi Chen, Diederik P Kingma, Tim Salimans, Yan Duan, Prafulla Dhariwal, John Schulman, Ilya Sutskever, and Pieter Abbeel. Variational lossy autoencoder. In *International Conference on Learning Representations*, 2017.

[15] Hervé Bourlard and Yves Kamp. Auto-association by multilayer perceptrons and singular value decomposition. *Biological cybernetics*, 59(4-5):291–294, 1988.

[16] Pierre Baldi and Kurt Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks*, 2(1):53–58, 1989.

[17] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[18] Guoshen Yu and Guillermo Sapiro. Statistical compressed sensing of Gaussian mixture models. *IEEE Transactions on Signal Processing*, 59(12):5842–5858, 2011.

[19] Michael C Fu. Gradient estimation. *Handbooks in operations research and management science*, 13:575–616, 2006.

[20] Peter W Glynn. Likelihood ratio gradient estimation for stochastic systems. *Communications of the ACM*, 33(10):75–84, 1990.

[21] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

[22] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, 2014.

[23] Paul Glasserman. *Monte Carlo methods in financial engineering*, volume 53. Springer Science & Business Media, 2013.

[24] John Schulman, Nicolas Heess, Theophane Weber, and Pieter Abbeel. Gradient estimation using stochastic computation graphs. In *Advances in Neural Information Processing Systems*, 2015.

[25] Peter J Diggle and Richard J Gratton. Monte carlo methods of inference for implicit statistical models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 193–227, 1984.

[26] Yair Weiss, Hyun Sung Chang, and William T Freeman. Learning compressed sensing. In *Snowbird Learning Workshop*, 2007.

[27] Yann LeCun, Corinna Cortes, and Christopher JC Burges. MNIST handwritten digit database. *http://yann. lecun. com/exdb/mnist*, 2010.

[28] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

[29] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *International Conference on Computer Vision*, 2015.

[30] Ashish Bora, Ajil Jalal, Eric Price, and Alexandros G Dimakis. Compressed sensing using generative models. In *International Conference on Machine Learning*, 2017.

[31] Manik Dhar, Aditya Grover, and Stefano Ermon. Modeling sparse deviations for compressed sensing using generative models. In *International Conference on Machine Learning*, 2018.

[32] Yoshua Bengio et al. Learning deep architectures for ai. *Foundations and trends® in Machine Learning*, 2(1):1–127, 2009.

[33] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *International Conference on Machine Learning*, 2008.

[34] Yoshua Bengio, Li Yao, Guillaume Alain, and Pascal Vincent. Generalized denoising auto-encoders as generative models. In *Advances in Neural Information Processing Systems*, 2013.

[35] Guillaume Alain and Yoshua Bengio. What regularized auto-encoders learn from the data-generating distribution. *Journal of Machine Learning Research*, 15(1):3563–3593, 2014.

[36] Alexander A Alemi, Ben Poole, Ian Fischer, Joshua V Dillon, Rif A Saurous, and Kevin Murphy. Fixing a broken ELBO. In *International Conference on Machine Learning*, 2018.

[37] Shengjia Zhao, Jiaming Song, and Stefano Ermon. The information autoencoding family: A lagrangian perspective on latent variable generative models. In *Conference on Uncertainty in Artificial Intelligence*, 2018.

[38] 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*, 2016.

[39] Shengjia Zhao, Jiaming Song, and Stefano Ermon. Infovae: Information maximizing variational autoencoders. In *AAAI Conference on Artificial Intelligence*, 2019.

[40] William R Carson, Minhua Chen, Miguel RD Rodrigues, Robert Calderbank, and Lawrence Carin. Communications-inspired projection design with application to compressive sensing. *SIAM Journal on Imaging Sciences*, 5(4):1185–1212, 2012.

[41] Liming Wang, Abolfazl Razi, Miguel Rodrigues, Robert Calderbank, and Lawrence Carin. Nonlinear information-theoretic compressive measurement design. In *International Conference on Machine Learning*, 2014.

[42] Ali Mousavi, Ankit B Patel, and Richard G Baraniuk. A deep learning approach to structured signal recovery. In *Annual Allerton Conference on Communication, Control, and Computing*, 2015.

[43] Kuldeep Kulkarni, Suhas Lohit, Pavan Turaga, Ronan Kerviche, and Amit Ashok. Reconnet: Non-iterative reconstruction of images from compressively sensed measurements. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[44] JH Rick Chang, Chun-Liang Li, Barnabas Poczos, BVK Vijaya Kumar, and Aswin C Sankaranarayanan. One network to solve them all—solving linear inverse problems using deep projection models. *arXiv preprint*, 2017.

[45] Xiaotong Lu, Weisheng Dong, Peiyao Wang, Guangming Shi, and Xuemei Xie. Convcsnet: A convolutional compressive sensing framework based on deep learning. *arXiv preprint arXiv:1801.10342*, 2018.

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

[47] Paul Hand and Vladislav Voroninski. Global guarantees for enforcing deep generative priors by empirical risk. In *Conference on Learning Theory*, 2018.

[48] Risheng Liu, Shichao Cheng, Yi He, Xin Fan, Zhouchen Lin, and Zhongxuan Luo. On the convergence of learning-based iterative methods for nonconvex inverse problems. *arXiv preprint arXiv:1808.05331*, 2018.

[49] Gareth O Roberts and Jeffrey S Rosenthal. Harris recurrence of metropolis-within-gibbs and trans-dimensional markov chains. *The Annals of Applied Probability*, pages 2123–2139, 2006.

[50] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations*, 2016.

[51] Guangliang Chen and Deanna Needell. Compressed sensing and dictionary learning. *Preprint*, 106, 2015.

[52] Shuyang Gao, Greg Ver Steeg, and Aram Galstyan. Variational information maximization for feature selection. In *Advances in Neural Information Processing Systems*, 2016.

[53] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2016.

[54] Yunzhu Li, Jiaming Song, and Stefano Ermon. Infogail: Interpretable imitation learning from visual demonstrations. In *Advances in Neural Information Processing Systems*, 2017.

[55] Shakir Mohamed and Danilo Jimenez Rezende. Variational information maximisation for intrinsically motivated reinforcement learning. In *Advances in Neural Information Processing Systems*, 2015.

A Proofs of Theoretical Results

A.1 Proof of Theorem 1

Proof. We can rewrite the UAE objective in Eq. (6) as:

$$\mathbb{E}_{Q_\phi(X,Y)}\left[\log p_\theta(x|y)\right] = \mathbb{E}_{Q_\phi(Y)}\left[\int q_\phi(x|y)\log p_\theta(x|y)\mathrm{d}x\right] \tag{10}$$

$$= -H_\phi(X|Y) - \mathbb{E}_{Q_\phi(Y)}\left[\mathrm{KL}(Q_\phi(X|y)\|P_\theta(X|y))\right]. \tag{11}$$

The KL-divergence is non-negative and minimized when its argument distributions are identical. Hence, for a fixed optimal value of $\theta = \theta^*$, if there exists a ϕ in the space of encoders being optimized that satisfies:

$$p_{\theta^*}(X|Y) = q_\phi(X|Y) \tag{12}$$

for all X, Y with $p^*_\theta(Y) \neq 0$, then it corresponds to the optimal encoder, *i.e.*,

$$\phi = \phi^*. \tag{13}$$

For any value of ϕ, we know the following Gibbs chain converges to $Q_\phi(X,Y)$ if the chain is ergodic:

$$y^{(t)} \sim Q_\phi(Y|x^{(t)}) \tag{14}$$

$$x^{(t+1)} \sim Q_\phi(X|y^{(t)}). \tag{15}$$

Substituting the results from Eqs. (12-15) in the Markov chain transitions in Eqs. (8, 9) finishes the proof. □

A.2 Proof of Corollary 1

Proof. By using earlier results (Proposition 2 in [49]), we need to show that the Markov chain defined in Eqs. (8)-(9) is Φ-irreducible with a Gaussian noise model.[1] That is, there exists a measure such that there is a non-zero probability of transitioning from every set of non-zero measure to every other such set defined on the same measure using this Markov chain.

Consider the Lebesgue measure. Formally, given any (x, y) and (x', y') such that the density $q(x, y) > 0$ and $q(x', y') > 0$ for the Lebesgue measure, we need to show that the probability density of transitioning $q(x', y'|x, y) > 0$.

(1) Since $q(y|x) > 0$ for all $x \in \mathbb{R}^n$, $y \in \mathbb{R}^m$ (by Gaussian noise model assumption), we can use Eq. (8) to transition from (x, y) to (x, y') with non-zero probability.

(2) Next, we claim that the transition probability $q(x'|y)$ is non-negative for all x', y. By Bayes rule, we have:

$$q(x'|y) = \frac{q(y|x')q(x')}{q(y)}.$$

Since $q(x, y) > 0$ and $q(x', y') > 0$, the marginals $q(y)$ and $q(x')$ are positive. Again, $q(y|x') > 0$ for all $x' \in \mathbb{R}^n$, $y \in \mathbb{R}^m$ by the Gaussian noise model assumption. Hence, $q(x'|y)$ is positive. Finally, using the optimality assumption for the posteriors $p(x'|y)$ matching $q(x'|y)$ for all x', y', we can use Eq. (9) to transition from (x, y') to (x', y') with non-zero probability.

From (1) and (2), we see that there is a non-zero probability of transitioning from (x, y) to (x', y'). Hence, under the assumptions of the corollary the Markov chain in Eqs. (8, 9) is ergodic.

□

[1] Note that the symbol Φ here is different from the parameters denoted by little ϕ used in the rest of the paper.

A.3 Proof of Theorem 2

Proof. Under an optimal decoder, the model posterior $P_\theta(X|Y)$ matches the true posterior $Q_\phi(X|Y)$ and hence, the UAE objective can be simplified as:

$$\begin{aligned}\mathbb{E}_{Q_\phi(X,Y)}[\log q_\phi(x|y)] &= \mathbb{E}_{Q_\phi(X,Y)}[\log q_\phi(x,y) - \log q_\phi(y)] \\ &= -H(X) - \mathbb{E}_{Q_{\text{data}}(X)}[H(Y|x)] - \mathbb{E}_{Q_\phi(X,Y)}[\log q_\phi(y)]. \end{aligned} \tag{16}$$

The first term corresponds to the negative of the data entropy, is independent of ϕ and σ, and hence it can be removed. For the second term, note that $Y|x$ is a normal distributed random variable and hence its entropy is given by a constant $\frac{1}{2}\log 2\pi e\sigma^2$. Only the third term depends on ϕ.

Removing the data entropy term since it is a constant independent of both ϕ and σ, we can define a modified objective $M(W, \mathcal{D}, \sigma)$ as:

$$M(W, \mathcal{D}, \sigma) := \mathbb{E}_{Q_\phi(X,Y)}[\log q_\phi(y)] + \frac{1}{2}\log 2\pi e\sigma^2. \tag{17}$$

As $\sigma \to \infty$, the optimal encodings maximizing the mutual information can be specified as:

$$W^* = \lim_{\sigma\to\infty} \arg\max_W -M(W, \mathcal{D}, \sigma). \tag{18}$$

We can lower-bound $M(W, \mathcal{D}, \sigma)$ using Jensen's inequality:

$$\begin{aligned} M(W, \mathcal{D}, \sigma) &= \mathbb{E}_{Q_\phi(X,Y)}[\log \mathbb{E}_{x_j\sim Q_{\text{data}}(X)}\left[q_\phi(y|x_j)\right]] + \frac{1}{2}\log 2\pi e\sigma^2 \\ &= \frac{1}{|\mathcal{D}|}\sum_{x_i\in\mathcal{D}} \mathbb{E}_{Q_\phi(Y|X)}\left[\log \frac{1}{|\mathcal{D}|}\sum_{x_j\in\mathcal{D}} q_\phi(y|x_j)\right] + \frac{1}{2}\log 2\pi e\sigma^2 \\ &\geq \frac{1}{|\mathcal{D}|}\sum_{x_i\in\mathcal{D}} \mathbb{E}_{Q_\phi(Y|X)}\left[\sum_{x_j\in\mathcal{D}} \frac{1}{|\mathcal{D}|}\log q_\phi(y|x_j)\right] + \frac{1}{2}\log 2\pi e\sigma^2 \\ &:= C(W, \mathcal{D}, \sigma) \end{aligned} \tag{19}$$

where we have used the fact that the data distribution is uniform over the entire dataset (by assumption).

Finally, we denote the non-negative slack term for the above inequality as $S(W, \mathcal{D}, \sigma)$ such that:

$$M(W, \mathcal{D}, \sigma) = C(W, \mathcal{D}, \sigma) + S(W, \mathcal{D}, \sigma). \tag{20}$$

Overview of proof strategy: We will first simplify expressions for the lower bound $C(W, \mathcal{D}, \sigma)$ and slack term $S(W, \mathcal{D}, \sigma)$. Then, we will show that as $\sigma \to \infty$, the ratio of the slack term and the lower bound converges pointwise to 0 and hence, the lower bound is arbitrarily close to $M(W, \mathcal{D}, \sigma)$ in this regime for a fixed W. Further, we will show that the convergence is uniform in W. Finally, we will note that the optimal encodings W^* for the lower bound correspond to the stated expressions for W in the proof statement.

As a first step, we consider simplifications of the lower bound and the slack term.

Lower bound: $C(W, \mathcal{D}, \sigma)$

$$
\begin{aligned}
C(W, \mathcal{D}, \sigma) &= \frac{1}{|\mathcal{D}|} \sum_{x_i \in \mathcal{D}} \mathbb{E}_{\epsilon \sim \mathcal{N}(0, \sigma^2 I)} \left[\frac{1}{|\mathcal{D}|} \sum_{x_j \in \mathcal{D}} [\log q_\phi(W x_i + \epsilon | x_j)] \right] + \frac{1}{2} \log 2\pi e \sigma^2 \\
&= \frac{1}{|\mathcal{D}|^2} \sum_{x_i, x_j \in \mathcal{D}} \mathbb{E}_{\epsilon \sim \mathcal{N}(0, \sigma^2 I)} [\log q_\phi(W x_i + \epsilon | x_j)] + \frac{1}{2} \log 2\pi e \sigma^2 \\
&= -\frac{1}{|\mathcal{D}|^2} \sum_{x_i, x_j \in \mathcal{D}} \mathbb{E}_{\epsilon \sim \mathcal{N}(0, \sigma^2 I)} \left[\frac{(W x_i + \epsilon - W x_j)^T (W x_i + \epsilon - W x_j)}{2\sigma^2} + \frac{1}{2} \log 2\pi \sigma^2 \right] + \frac{1}{2} \log 2\pi e \sigma^2 \\
&= -\frac{1}{|\mathcal{D}|^2} \sum_{x_i, x_j \in \mathcal{D}} \mathbb{E}_{\epsilon \sim \mathcal{N}(0, \sigma^2 I)} \left[\frac{(W x_i - W x_j)^T (W x_i - W x_j) + 2\epsilon^T (W x_i - W x_j) + \epsilon^T \epsilon}{2\sigma^2} \right] + \frac{1}{2} \\
&= -\frac{1}{|\mathcal{D}|^2} \sum_{x_i, x_j \in \mathcal{D}} \left(\frac{(W x_i - W x_j)^T (W x_i - W x_j)}{2\sigma^2} + \frac{1}{2} \right) + \frac{1}{2} \\
&= -\frac{1}{|\mathcal{D}|^2} \sum_{x_i, x_j \in \mathcal{D}} \left(\frac{(W x_i - W x_j)^T (W x_i - W x_j)}{2\sigma^2} \right).
\end{aligned} \tag{21}
$$

Slack: $S(W, \mathcal{D}, \sigma)$

$$
\begin{aligned}
S(W, \mathcal{D}, \sigma) &= -C(W, \mathcal{D}, \sigma) + M(W, \mathcal{D}, \sigma) \\
&= -\frac{1}{|\mathcal{D}|} \sum_{x_i \in \mathcal{D}} \mathbb{E}_{Q_\phi(Y|X)} \left[\frac{1}{|\mathcal{D}|} \sum_{x_j \in \mathcal{D}} \log q_\phi(y | x_j) \right] + \frac{1}{|\mathcal{D}|} \sum_{x_i \in \mathcal{D}} \mathbb{E}_{Q_\phi(Y|X)} \left[\log \frac{1}{|\mathcal{D}|} \sum_{x_j \in \mathcal{D}} q_\phi(y | x_j) \right] \\
&= -\frac{1}{|\mathcal{D}|} \sum_{x_i \in \mathcal{D}} \mathbb{E}_{Q_\phi(Y|X)} \left[\mathbb{E}_{Q_{\text{data}}(X)} \left[\log \frac{q_\phi(y, x_j)}{q_{\text{data}}(x_j)} \right] \right] + \frac{1}{|\mathcal{D}|} \sum_{x_i \in \mathcal{D}} \mathbb{E}_{Q_\phi(Y|X)} [\log q_\phi(y)] \\
&= \frac{1}{|\mathcal{D}|} \sum_{x_i \in \mathcal{D}} \mathbb{E}_{Q_\phi(Y|X)} \left[\mathbb{E}_{Q_{\text{data}}(X)} \left[\log \frac{q_{\text{data}}(x_j)}{q_\phi(x_j | y) q_\phi(y)} \right] \right] + \frac{1}{|\mathcal{D}|} \sum_{x_i \in \mathcal{D}} \mathbb{E}_{Q_\phi(Y|X)} [\log q_\phi(y)] \\
&= \frac{1}{|\mathcal{D}|} \sum_{x_i \in \mathcal{D}} \mathbb{E}_{Q_\phi(Y|X)} \left[KL \left(Q_{\text{data}}(X), Q_\phi(X | y) \right) \right] \\
&= -\frac{1}{|\mathcal{D}|^2} \sum_{x_i, x_j \in \mathcal{D}} \mathbb{E}_{Q_\phi(Y|X)} [\log q_\phi(x_j | y) + \log |\mathcal{D}|] \\
&= -\log |\mathcal{D}| - \frac{1}{|\mathcal{D}|^2} \sum_{x_i, x_j \in \mathcal{D}} \mathbb{E}_{\epsilon \sim \mathcal{N}(0, \sigma^2 I)} [\log q_\phi(x_j | W x_i + \epsilon)]
\end{aligned} \tag{22}
$$

We can simplify the posteriors $Q_\phi(x_j | W x_i + \epsilon)$ as:

$$
\begin{aligned}
Q_\phi(x_j | W x_i + \epsilon) &= \frac{Q_\phi(x_j, W x_i + \epsilon)}{Q_\phi(W x_i + \epsilon)} \\
&= \frac{Q_\phi(x_j) Q_\phi(W x_i + \epsilon | x_j)}{\sum_{x_k \in \mathcal{D}} Q_\phi(W x_i + \epsilon | x_k) Q_\phi(x_k)} = \frac{\exp\left(-{(W(x_i - x_j) + \epsilon)^T (W(x_i - x_j) + \epsilon)}/{2\sigma^2} \right)}{\sum_{x_k \in \mathcal{D}} \exp\left(-{(W(x_i - x_k) + \epsilon)^T (W(x_i - x_k) + \epsilon)}/{2\sigma^2} \right)}
\end{aligned} \tag{23}
$$

where we have used the fact that the data distribution is uniform and the decoder is isotropic Gaussian.

Substituting the above expression for the slack term:

$$\begin{aligned}
S(W, \mathcal{D}, \sigma) &= -\log|\mathcal{D}| - \frac{1}{|\mathcal{D}|^2} \sum_{x_i, x_j \in \mathcal{D}} \mathbb{E}_{\epsilon \sim \mathcal{N}(0, \sigma^2 I)} \Bigg[\left(-\frac{(W(x_i - x_j) + \epsilon)^T (W(x_i - x_j) + \epsilon)}{2\sigma^2} \right) \\
&\quad - \log \sum_{x_k \in \mathcal{D}} \exp\left(-\frac{(W(x_i - x_k) + \epsilon)^T (W(x_i - x_k) + \epsilon)}{2\sigma^2} \right) \Bigg] \\
&= \frac{1}{|\mathcal{D}|^2} \sum_{x_i, x_j \in \mathcal{D}} \Bigg[\frac{(Wx_i - Wx_j)^T (Wx_i - Wx_j)}{2\sigma^2} + \frac{1}{2} \\
&\quad + \underbrace{\mathbb{E}_{\epsilon \sim \mathcal{N}(0, \sigma^2 I)} \left(\log \sum_{x_k \in \mathcal{D}} \exp\left(-\frac{(W(x_i - x_k) + \epsilon)^T (W(x_i - x_k) + \epsilon)}{2\sigma^2} \right) - \log|\mathcal{D}| \right)}_{\int \gamma(\sigma, \epsilon, W, \mathcal{D}, x_i) \mathrm{d}\epsilon} \Bigg]. \qquad (24)
\end{aligned}$$

For any fixed x_i, W, the integrand $\gamma(\sigma, \epsilon, W, \mathcal{D}, x_i)$ in Eq. (24) can be seen as a sequence of functions indexed by σ. We next make the claim that dominated convergence in ϵ holds for this sequence for all x_i, W. We show so by first observing that $\gamma(\sigma, \epsilon, W, \mathcal{D}, x_i)$ converges pointwise to a constant (=0) as $\sigma \to \infty$ and thereafter deriving integrable upper and lower bounds for $\gamma(\sigma, \epsilon, W, \mathcal{D}, x_i)$ below that are independent of σ.

For the upper bound, we note that:

$$\begin{aligned}
\log \sum_{x_j \in \mathcal{D}} \exp\left(-\frac{(Wx_i + \epsilon - Wx_j)^T (Wx_i + \epsilon - Wx_j)}{2\sigma^2} \right) &\leq \max_{x_j \in \mathcal{D}} -\frac{(Wx_i + \epsilon - Wx_j)^T (Wx_i + \epsilon - Wx_j)}{2\sigma^2} + \log|\mathcal{D}| \\
&\leq \log|\mathcal{D}|. \qquad (25)
\end{aligned}$$

This gives an upper bound on the integrand $\gamma(\sigma, \epsilon, W, \mathcal{D}, x_i)$:

$$\begin{aligned}
\gamma(\sigma, \epsilon, W, \mathcal{D}, x_i) &\leq -\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\epsilon^T \epsilon}{2\sigma^2} \right) [\log|\mathcal{D}| - \log|\mathcal{D}|] \\
&= 0. \qquad (26)
\end{aligned}$$

For the lower bound, we note that:

$$\begin{aligned}
\log \sum_{x_j \in \mathcal{D}} \exp\left(-\frac{(Wx_i + \epsilon - Wx_j)^T (Wx_i + \epsilon - Wx_j)}{2\sigma^2} \right) &\geq \max_{x_j \in \mathcal{D}} \left(-\frac{(Wx_i + \epsilon - Wx_j)^T (Wx_i + \epsilon - Wx_j)}{2\sigma^2} \right) \\
&= -\min_{x_j \in \mathcal{D}} \left(\frac{(Wx_i + \epsilon - Wx_j)^T (Wx_i + \epsilon - Wx_j)}{2\sigma^2} \right). \qquad (27)
\end{aligned}$$

.

Hence, we have the following lower bound:

$$\begin{aligned}
\gamma(\sigma, \epsilon, W, \mathcal{D}, x_i) &\geq -\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\epsilon^T \epsilon}{2\sigma^2} \right) \left(\min_{x_j \in \mathcal{D}} \frac{(Wx_i + \epsilon - Wx_j)^T (Wx_i + \epsilon - Wx_j)}{2\sigma^2} - \log|\mathcal{D}| \right) \\
&\geq -\frac{1}{\sqrt{\pi \epsilon^T \epsilon}} \left(\frac{1}{\epsilon^T \epsilon} \left[\min_{x_j \in \mathcal{D}} (Wx_i + \epsilon - Wx_j)^T (Wx_i + \epsilon - Wx_j) \right] - 2\log|\mathcal{D}| \right) \\
&= -\frac{1}{\sqrt{\pi \epsilon^T \epsilon}} \left(\frac{1}{\epsilon^T \epsilon} \left[\min_{x_j \in \mathcal{D}} (Wx_i - Wx_j)^T (Wx_i - Wx_j) + 2(Wx_i - Wx_j)^T \epsilon + \epsilon^T \epsilon \right] - 2\log|\mathcal{D}| \right) \\
&\geq -\frac{1}{\sqrt{\pi \epsilon^T \epsilon}} \left(\frac{4k_1^2 k_2^2}{\epsilon^T \epsilon} + \frac{4k_1 k_2}{\sqrt{\epsilon^T \epsilon}} + 1 - 2\log|\mathcal{D}| \right) \qquad (28)
\end{aligned}$$

where we used the inequalities $\exp(-1/z) \leq z^{3/2}, \exp(-1/z) \leq z^{1/2}$ for any $z > 0$ in the second step (with $z = {}^{2\sigma^2}/_{\epsilon^T \epsilon}$) and Cauchy-Schwarz for the last step (since $\|x\|_2 \leq k_1$ for all $x \in \mathcal{D}$, $\|W\|_F \leq k_2$ for some positive constants $k_1, k_2 \in \mathbb{R}^+$ by assumption).

Since both the upper and lower bounds for the integrand are independent of σ, dominated convergence holds for the third term in Eq. (24).

Consequently, we can evaluate limits to obtain a limiting ratio between the slack term and the lower bound:

$$\lim_{\sigma \to \infty} \frac{S(W, \mathcal{D}, \sigma)}{C(W, \mathcal{D}, \sigma)} = 0 \tag{29}$$

using the expressions derived in Eq. (21) and Eq. (24), dominated convergence for interchanging limits and expectations, along with L'Hôpital's rule.

We can now rewrite Eq. (20) as:

$$M(W, \mathcal{D}, \sigma) = C(W, \mathcal{D}, \sigma) \left(1 + \frac{S(W, \mathcal{D}, \sigma)}{C(W, \mathcal{D}, \sigma)}\right). \tag{30}$$

By the (ϵ, δ) definition of limit, we know that for any fixed W that satisfies $\|W\|_F \leq k_2$ and $\forall \epsilon > 0$, there exists a $\delta > 0$ such that $\forall \sigma > \delta$, we have:

$$|M(W, \mathcal{D}, \sigma) - C(W, \mathcal{D}, \sigma)| < \epsilon. \tag{31}$$

Next, we note that the slack term $S(W, \mathcal{D}, \sigma)$ is monotonic in σ and converges pointwise for any fixed W that satisfies $\|W\|_F \leq k_2$.

$$\lim_{\sigma \to \infty} S(W, \mathcal{D}, \sigma) = \lim_{\sigma \to \infty} M(W, \mathcal{D}, \sigma) - C(W, \mathcal{D}, \sigma) = 0. \tag{32}$$

Using Dini's Theorem, this implies the convergence of the slack term is uniform in W as $\sigma \to \infty$. Hence, for all W that satisfy $\|W\|_F \leq k_2$ and $\forall \epsilon > 0$, there exists a $\delta > 0$ such that $\forall \sigma > \delta$, we have:

$$|M(W, \mathcal{D}, \sigma) - C(W, \mathcal{D}, \sigma)| < \epsilon. \tag{33}$$

Since the arg max operator preserves continuity (via Berge's maximum theorem) and is assumed to be identifiable, we conclude that $\forall W$ satisfying $\|W\|_F \leq k_2$ and $\forall \epsilon > 0$, there exists a $\delta > 0$ such that $\forall \sigma > \delta$, we have:

$$|W^* - \arg\max_W \sum_{x_i, x_j \in \mathcal{D}} \left(Wx_i - Wx_j\right)^T (Wx_i - Wx_j)\big) \,| < \epsilon \tag{34}$$

which finishes the proof. □

Table 2: Frobenius norms of the UAE encodings and random Gaussian projections for MNIST and Omniglot datasets.

m	Random Gaussian Matrices	MNIST-UAE	Omniglot-UAE
2	39.57	6.42	2.17
5	63.15	5.98	2.66
10	88.98	7.24	3.50
25	139.56	8.53	4.71
50	198.28	9.44	5.45
100	280.25	10.62	6.02

B Experimental details

For MNIST, we use the train/valid/test split of $50,000/10,000/10,000$ images. For Omniglot, we use train/valid/test split of $23,845/500/8,070$ images. For CelebA, we used the splits as provided by [29] on the dataset website. All images were scaled such that pixel values are between 0 and 1. We used the Adam optimizer with a learning rate of 0.001 for all the learned models. For MNIST and Omniglot, we used a batch size of 100. For CelebA, we used a batch size of 64. Further, we implemented early stopping based on the best validation bounds after 200 epochs for MNIST, 500 epochs for Omniglot, and 200 epochs for CelebA.

B.1 Hyperparameters for compressed sensing on MNIST and Omniglot

For both datasets, the UAE decoder used 2 hidden layers of 500 units each with ReLU activations. The encoder was a single linear layer with only weight parameters and no bias parameters. The encoder and decoder architectures for the *VAE baseline* are symmetrical with 2 hidden layers of 500 units each and 20 latent units. We used the *LASSO baseline* implementation from sklearn and tuned the Lagrange parameter on the validation sets. For the baselines, we do 10 random restarts with $1,000$ steps per restart and pick the reconstruction with best measurement error as prescribed in [30]. Refer to [30] for further details of the baseline implementations.

Table 2 shows the average norms for the random Gaussian matrices used in the baselines and the learned UAE encodings. The lower norms for the UAE encodings suggest that the UAE baseline is not trivially overcoming noise by increasing the norm of W.

B.2 Hyperparameters for dimensionality reduction

For PCA and each of the classifiers, we used the standard implementations in sklearn with default parameters and the following exceptions:

- KNN: n_neighbors = 3
- DT: max_depth = 5
- RF: max_depth = 5, n_estimators = 10, max_features = 1
- MLP: alpha=1
- SVC: kernel=linear, C=0.025

B.3 Statistical compressed sensing on CelebA dataset

For the CelebA dataset, the dimensions of the images are $64 \times 64 \times 3$ and $\sigma = 0.01$. The naive pixel basis does not augur well for compressed sensing on such high-dimensional RGB datasets. Following [30], we experimented with the Discrete Cosine Transform (DCT) and Wavelet basis for the LASSO baseline. Further, we used the DCGAN architecture [50] as in [30] as our main baseline. For the UAE approach, we used additional convolutional layers in the encoder to learn a 256 dimensional feature space for the image before projecting it down to m dimensions.

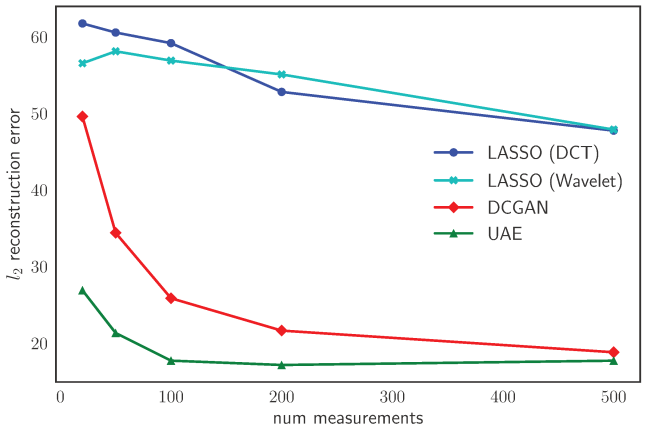


Figure 6: Test ℓ_2 reconstruction error (per image) for compressed sensing on CelebA.

Encoder architecture:

Signal → Conv[Kernel: 4x4, Stride: 2, Filters: 32, Padding: Same, Activation: Relu]
→ Conv[Kernel: 4x4, Stride: 2, Filters: 32, Padding: Same, Activation: Relu]
→ Conv[Kernel: 4x4, Stride: 2, Filters: 64, Padding: Same, Activation: Relu]
→ Conv[Kernel: 4x4, Stride: 2, Filters: 64, Padding: Same, Activation: Relu]
→ Conv[Kernel: 4x4, Stride: 1, Filters: 256, Padding: Valid, Activation: Relu]
→ Fully_Connected[Units: m, Activation: None]

Decoder architecture:

Measurements → Fully_Connected[Units: 256, Activation: Relu]
→ Conv_transpose[Kernel: 4x4, Stride: 1, Filters: 256, Padding: Valid, Activation: Relu]
→ Conv_transpose[Kernel: 4x4, Stride: 2, Filters: 64, Padding: Same, Activation: Relu]
→ Conv_transpose[Kernel: 4x4, Stride: 2, Filters: 64, Padding: Same, Activation: Relu]
→ Conv_transpose[Kernel: 4x4, Stride: 2, Filters: 32, Padding: Same, Activation: Relu]
→ Conv_transpose[Kernel: 4x4, Stride: 2, Filters: 3, Padding: Valid, Activation: Sigmoid]

We consider $m = \{20, 50, 100, 200, 500\}$ measurements. The results are shown in Figure 6. While the performance of DCGAN is comparable with that of UAE for $m = 500$, UAE outperforms DCGAN significantly when m is low. The LASSO baselines do not perform well, consistent with the observations made for the experiments on the MNIST and Omniglot datasets. Qualitative evaluations are shown in Figure 7 for $m = 50$ measurements.

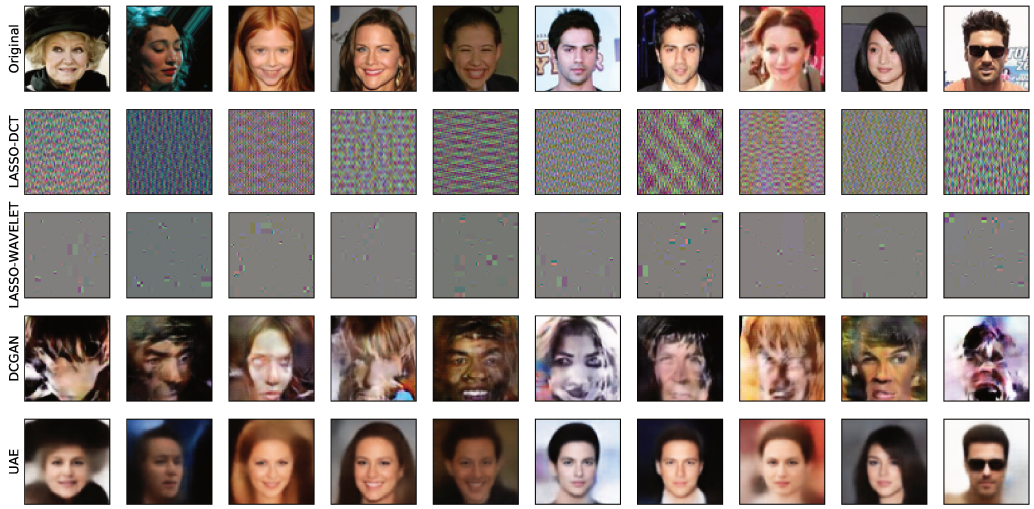


Figure 7: Reconstructions for $m = 50$ on the CelebA dataset. **Top:** Target. **Second:** LASSO-DCT. **Third:** LASSO-Wavelet. **Fourth:** DCGAN. **Last:** UAE.

C Additional related work

Dictionary learning. An uncertainty autoencoder can be also seen as a more flexible, generalized form of *undercomplete* dictionary learning with non-linear encoding and decoding. To see the connection, consider the simplified noise-free setting where the decoding distribution is a Gaussian with fixed variance and the mean of the decoding function is linear in a linear function of the measurements. That is, we are considering a standard *linear* autoencoder with $Y = WX$ and $P_\theta(X|Y) = \mathcal{N}(\widehat{W}Y, \Sigma)$, where $\widehat{W}$ is some decoding matrix. Under these assumptions, the UAE objective simplifies to:

$$\min_{W,\widehat{W}} \mathbb{E}_{x\sim Q_{\text{data}}} \left[\|x - \widehat{W}Wx\|_2^2 \right].$$

If we think of the decoding $\widehat{W}$ as a dictionary and the encoding WX as the representation then we arrive at an undercomplete dictionary learning. A large body of prior research has focussed on *overcomplete* dictionary learning for compressed sensing. Here, the goal is to learn an *encoding dictionary* and an overcomplete basis in which the original signal is sparse. This basis allows us to leverage algorithms for compressed sensing that are designed based on sparsity assumptions over the signals (see [51] and references therein). An uncertainty autoencoder makes no sparsity assumptions, and it crucially learns a *decoding dictionary* and an encoding basis. By adding more (non-linear) layers to the encoder, one could also learn an (over/under) complete basis for the dataset with desired properties such as sparsity.

Further applications of variational information maximization. The variational information maximization principle underlies many recent algorithms and tasks, such as feature selection [52], interpretable representation learning in generative adversarial networks [53, 54], and intrinsic motivation in reinforcement learning [55].