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-**??**) 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, W, the final term 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 term for all x, W. We show so by deriving upper and lower bounds on the integrand $\gamma(\sigma,\epsilon,W,\mathcal{D},x_i)$ that are independent of σ.

For the upper bound, we note that:

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

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

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

For the lower bound, we note that:

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

.

Hence, we have the following lower bound:

$$\begin{aligned}
\gamma(\sigma,\epsilon,W,\mathcal{D},x_i) &\ge -\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{\epsilon^2}{2\sigma^2}\right)\left(\min_{x_j\in\mathcal{D}}\frac{(Wx_i+\epsilon-Wx_j)^T(Wx_i+\epsilon-Wx_j)}{2\sigma^2}\right)-\log|\mathcal{D}| \\
&\ge -\frac{1}{\sqrt{\pi}\epsilon^3}\left(\min_{x_j\in\mathcal{D}}(Wx_i+\epsilon-Wx_j)^T(Wx_i+\epsilon-Wx_j)\right)-\log|\mathcal{D}| \\
&= -\frac{1}{\sqrt{\pi}\epsilon^3}\left(\min_{x_j\in\mathcal{D}}(Wx_i-Wx_j)^T(Wx_i-Wx_j)+2\epsilon(Wx_i-Wx_j)+\epsilon^T\epsilon\right)-\log|\mathcal{D}| \\
&\ge -\frac{1}{\sqrt{\pi}\epsilon^3}\left(4k_1^2k_2^2+4k_1k_2\epsilon^T\epsilon\right)-\log|\mathcal{D}| \qquad (28)
\end{aligned}$$

where we used the inequalities $\exp(-1/z)\le z^{3/2}$ for any $z>0$ in the second step (with $z=2\sigma^2/\epsilon^2$) and Cauchy-Schwarz for the last step (since $\|x\|_2\le k_1$ for all $x\in\mathcal{D}$, $\|W\|_F\le k_2$ 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.

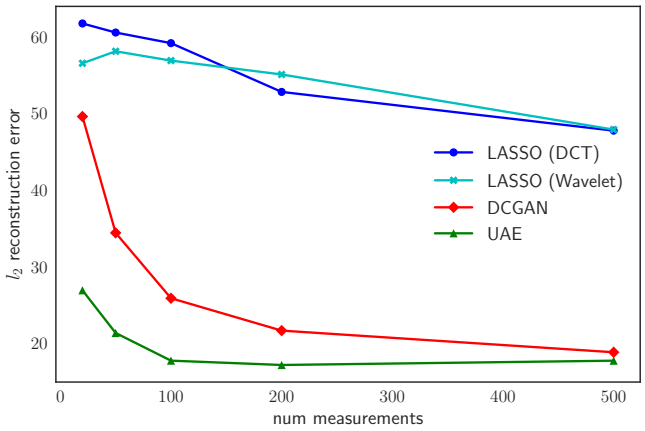

Figure 6: Test ℓ_2 reconstruction error (per image) for compressed sensing on CelebA.

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.

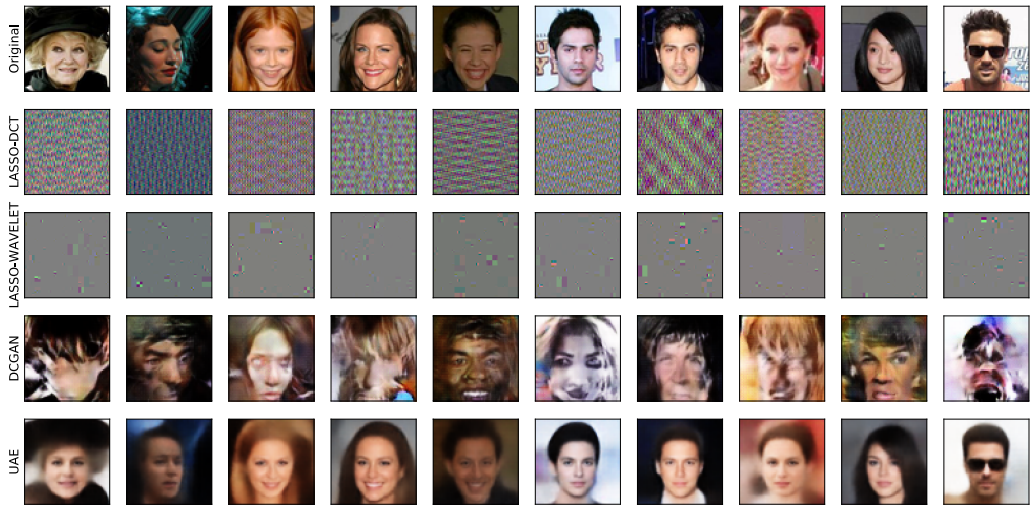

Figure 7: Reconstructions for $m = 50$ on the CelebA dataset. **Top:** Target. **Second:** LASSO-DCT. **Third:** LASSO-Wavelet. **Fourth:** DCGAN. **Last:** UAE.

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].