ANNEALED DENOISING SCORE MATCHING: LEARNING ENERGY BASED MODELS IN HIGH-DIMENSIONAL SPACES

Anonymous authors
Paper under double-blind review

ABSTRACT

Energy based models outputs unmormalized log-probability values given data samples. Such a estimation is essential in a variety of application problems such as sample generation, denoising, sample restoration, outlier detection, Bayesian reasoning, and many more. However, standard maximum likelihood training is computationally expensive due to the requirement of sampling the model distribution. Score matching potentially alleviates this problem, and denoising score matching (Vincent, 2011) is a particularly convenient version. However, previous attempts failed to produce models capable of high quality sample synthesis. We believe that it is because they only performed denoising score matching over a single noise scale. To overcome this limitation, here we instead learn an energy function using all noise scales. When sampled using Annealed Langevin dynamics and single step denoising jump, our model produced high-quality samples comparable to state-of-the-art techniques such as GANs, in addition to assigning likelihood to test data comparable to previous likelihood models. Our model set a new sample quality baseline in likelihood-based models. We further demonstrate that our model learns sample distribution and generalize well on an image inpainting task.

1 INTRODUCTION AND MOTIVATION

Treating data as stochastic samples from a probability distribution and developing models that can learn such distributions is at the core for solving a large variety of application problems, such as error correction/denoising Vincent et al. (2010), outlier/novelty detection (Zhai et al., 2016; Choi and Jang, 2018b), invariant pattern recognition, Bayesian reasoning (Welling and Teh, 2011) which relies on good data priors, and many others.

Energy based models (Ngiam et al., 2011) assign an energy $E(\boldsymbol{x})$ to each data point $\boldsymbol{x}$ which implicitly defines a probability by the Boltzmann distribution $p_m(\boldsymbol{x}) = e^{-E(\boldsymbol{x})}/Z$. Sampling from this distribution can be used as as generative process that yield plausible samples of $\boldsymbol{x}$. Compared to other generative models, like GANs (Goodfellow et al., 2014), flow-based models (Dinh et al., 2014; Kingma and Dhariwal, 2018), or auto-regressive models (Oord et al., 2016; Ostrovski et al., 2018), energy based models have significant advantages. First, they provide explicit (unnormalized) density information, compositionality (Hinton, 1999; Haarnoja et al., 2017), better mode coverage (Kumar et al., 2019) and flexibility (Du and Mordatch, 2019). Second, they do not require special model architecture (like auto-regressive and flow-based models). However, traditionally, energy based models are trained by maximum likelihood, which requires sampling of model distribution. As a result, learning high-dimensional benchmark datasets in energy based models was reported to be prohibitively slow (Nijkamp et al., 2019). Variants with a truncated sampling procedure have been proposed, such as contrastive divergence (Hinton, 2002). Such models learn much faster but have the draw back of not exploring the state space thoroughly (Tieleman, 2008).

Score matching (Hyvärinen, 2005), which can be viewed as a special case of minimum probability flow (Sohl-Dickstein et al., 2011), circumvents the requirement of sampling. In score matching, the score function is defined to be the gradient of log-density or the negative energy function. The expected squared distance of the model score function and the data score function are minimized. Surprisingly, this can be achieved without an estimation of the data score function in the first place

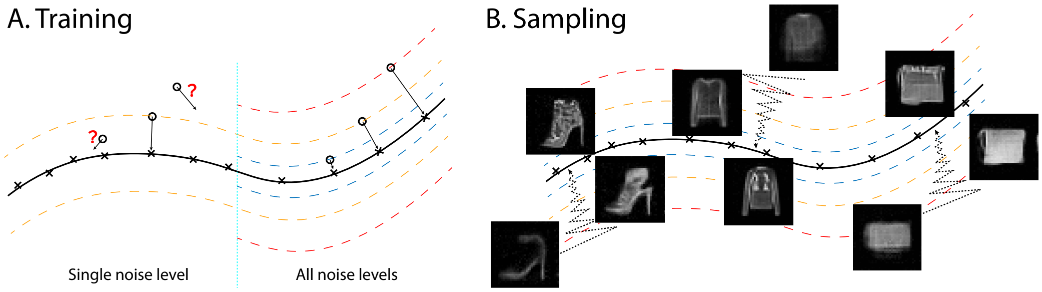

Figure 1: Illustration of anneal denoising score matching. A. During training, derivative of log-likelihood is forced to point toward data manifold, establishing energy difference between points within manifold and points outside. Note that energy is negative log-likelihood therefore energy is higher for point further away from data manifold. B. During annealed Langevin sampling, sample travel from outside data manifold to data manifold. Shown are singled step denoised sample during sampling of an energy function trained with ADSM on Fashion-MNIST (see text for details).

under mild condition. However, the original score matching algorithm is still computationally expensive (Song et al., 2019).

Our specific starting point here is *denoising score matching* (Vincent, 2011), a variation to score matching with a very efficient learning procedure. A simple noise structure with differentiable density kernel, such as Gaussian noise, is used to generate noisy surrogates from data points. At these noisy surrogates the scores can be matched during training. Previous models of denoising score matching were successfully demonstrated in denoising tasks, but sampling from these models was unable to generate good data samples, when initiated from random noise (Saremi and Hyvarinen, 2019). Wenliang et al. (2018) and Song and Ermon (2019) suspected this limitation to be due to a failure of the training to shape the energy function in low-density data regions. This could explain why sampling cannot reach high-density data regions when initialized at random (Saremi and Hyvarinen, 2019).

The weakness of learning in the previous models could be caused by the fact that training is performed at a single fixed noise level. One indication for this is that model performance strongly depended on the chosen noise level (Song et al., 2019). To further explore this explanation, we trained energy based models on different single noise scales in Appendix B. Indeed, these models capture some data structure but do not generate satisfactory data samples.

In order to construct an energy based model capable of high-quality sample synthesis while maintaining the speed advantage of score matching, we propose *annealed denoising score matching* (ADSM), a denoising score matching procedure using multiple noise levels. We expect this sampling-free learning method to shape an energy function that can drive samples anywhere in the state space towards high probability regions of the data. For sampling from the learned model, we propose *annealed Langevin dynamics* and additionally use single step denoise jump. We show that our model generates high-quality samples similar in quality to the-state-of-the-art image generation techniques, such as GANs (Goodfellow et al., 2014), as well as assigning competitive likelihood values to the test sets.

2 Energy based model with Annealed Denoising score matching

2.1 Score matching and denoising score matching

One convenient way of using score matching is learning the energy function corresponding to a Parzen window density estimator of the data: $p_{\sigma_0}(\tilde{\boldsymbol{x}}) = \int q_{\sigma_0}(\tilde{\boldsymbol{x}}|\boldsymbol{x})p(\boldsymbol{x})dx$. Though hard to evaluate, the data score is well defined: $s_d(\tilde{\boldsymbol{x}}) = \nabla_{\tilde{\boldsymbol{x}}} \log(p_{\sigma_0}(\tilde{\boldsymbol{x}}))$. And the corresponding objective is:

$$L_{SM}(\theta) = \mathbb{E}_{p_{\sigma 0}(\tilde{\boldsymbol{x}})} \parallel \nabla_{\tilde{\boldsymbol{x}}} \log(p_{\sigma_0}(\tilde{\boldsymbol{x}})) + \nabla_{\tilde{x}} E(\tilde{\boldsymbol{x}};\theta) \parallel^2 \tag{1}$$

A remarkable result proved in Vincent (2011) is that the following objective for denoising score matching (DSM) is equivalent to the objective above:

$$L_{DSM}(\theta) = \mathbb{E}_{p_{\sigma 0}(\tilde{\boldsymbol{x}},\boldsymbol{x})} \parallel \nabla_{\tilde{\boldsymbol{x}}} \log(q_{\sigma 0}(\tilde{\boldsymbol{x}}|\boldsymbol{x})) + \nabla_{\tilde{\boldsymbol{x}}} E(\tilde{\boldsymbol{x}};\theta) \parallel^2 \tag{2}$$

Note that the Parzen window density estimator score is replaced by the derivative of log density of the single noise kernel $\nabla_{\tilde{\boldsymbol{x}}} \log(q_\sigma(\tilde{\boldsymbol{x}}|\boldsymbol{x}))$, which is much easier to calculate. In the particular case of Gaussian noise $\log(q_{\sigma_0}(\tilde{\boldsymbol{x}}|\boldsymbol{x})) = -\frac{(\tilde{\boldsymbol{x}}-\boldsymbol{x}_i)^2}{2\sigma_0^2} + C$ and:

$$L_{DSM}(\theta) = \mathbb{E}_{p_{\sigma 0}(\tilde{\boldsymbol{x}},\boldsymbol{x})} \parallel \boldsymbol{x}_i - \tilde{\boldsymbol{x}}_i + {\sigma_0}^2 \nabla_{\tilde{\boldsymbol{x}}} E(\tilde{\boldsymbol{x}}_i;\theta) \parallel^2 \tag{3}$$

This objective (3) forces the energy gradient to align with the vector pointing from the noisy surrogate to the clean data sample, which explains the term “denoising” in the name. Note further, that in the particular case of Gaussian noise, the objective (3) coincides with the empirical Bayes objective proposed in Saremi and Hyvarinen (2019).

A tricky question is how to select σ_0, which is an important hyperparameter. It is reported in Song et al. (2019) that a grid search is required to give the optimal result, and heuristics do not provide enough guidance. Previous attempts of learning energy-based models with score matching over a single noise level also do not permit effective sampling to produce high-quality samples (Song et al., 2019; Saremi et al., 2018; Saremi and Hyvarinen, 2019). Given the drawbacks, we propose learning energy function using denoising score matching over a wide range of noise levels, effectively abolishing the noise scale σ_0 as a hyper-parameter.

2.2 A Geometric view

Before introducing our new model, we provide a geometrical picture for denoising score matching in high dimensional space. We adopt the common assumption that the data distribution to be learned is high-dimensional, but only has support around a relatively low-dimensional manifold (Tenenbaum et al., 2000; Roweis and Saul, 2000; Lawrence, 2005). We further assume that data is approximately uniformly distributed on the manifold $\mathcal{X}$. We recall two elementary properties of random Gaussian vectors in high-dimensional spaces: First, the length distribution of random vectors becomes concentrated at $\sqrt{d}\sigma$ (Vershynin, 2018), where σ^2 is the variance of a single component. Second, a random vector is almost always orthogonal to a fixed vector (Tao, 2012). With these premises one can visualize the configuration of noisy and noiseless data points used in the learning of denoising score matching: A data point $\boldsymbol{x}$ and its noisy version $\tilde{\boldsymbol{x}}$ always lie on a line which is almost perpendicular to the tangent space $T_{\boldsymbol{x}}\mathcal{X}$ of $\mathcal{X}$ at $\boldsymbol{x}$ and intersects $\mathcal{X}$ at $\boldsymbol{x}$. Further, the distance vectors between $(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ pairs all have similar length $\sqrt{d}\sigma$. As a consequence, the set of noisy data points concentrate on a set $\tilde{\mathcal{X}}_{\sqrt{d}\sigma,\epsilon}$ that has a distance with $(\sqrt{d}\sigma - \epsilon, \sqrt{d}\sigma + \epsilon)$ from the data manifold $\mathcal{X}$, where $\epsilon \ll 1$.

All told, performing denoising score matching learning with $(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ pairs generated with a fixed noise level σ, will enable the de-noising of noisy points lying in the described noisy set $\tilde{\mathcal{X}}_{\sqrt{d}\sigma,\epsilon}$. However, the learning provides little information about how to de-noise noisy points that lie outside the set, farther or closer to the data manifold. An illustration is presented in Figure 1A. Though $p_{\sigma_0}(\tilde{\boldsymbol{x}} \in \tilde{\mathcal{X}}^C_{\sqrt{d}\sigma,\epsilon})$ is very small in high-dimensional space, the score function in $\tilde{\mathcal{X}}^C_{\sqrt{d}\sigma,\epsilon}$ still plays a critical role in sampling from random noise. This phenomenon in high-dimensional space is not sufficiently appreciated by the classical score matching and denoising score matching objective.

Another property of denoising score matching was observed in the denoising auto-encoder literature (Vincent et al., 2010; Karklin and Simoncelli, 2011) that with higher noise level the learned representation tend to have larger spatial scale. An intuitive reason is that natural signals have multi-scale structure with $1/f$ power characteristic, meaning that features at low spatial frequencies have larger signal power than features at high spacial frequencies (Field, 1987). Thus, a higher noise levels will wash away features above certain spacial frequencies, but insufficiently corrupting features below certain spacial frequencies to drive learning. Thus, a model trained with only one noise level will only learn structures roughly of a single scale.

To illustrate this point, in Figure 1B we show some example of intermediate samples during annealed sampling process. Displayed are samples obtained by single step denoising from the noisy sample carrying different levels of noise. We provide more result from this and similar experiment in Figure 3 and Figure 7.

This geometric view provides a potential explanation why score matching has so far been unsuccessful in training generative models. Elementary properties of i.i.d. noise in high-dimensional spaces suggest that training at one fixed noise level is insufficient to gather the information required for denoising from different noise level in high-dimensional space. Further, the one-over-f structure found almost universally in interesting data sets with multi-scale structure causes models trained with one noise level to be myopic to one scale and ignorant of the others.

2.3 Annealed Denoising Score matching

Motivated by the geometric view, we strive to learn an energy function that captures the structure of data over different scales by introducing different levels of Gaussian noise corruptions.

The data distribution over different noise scales can be expressed as $p_{\sigma_0,T}(\tilde{\boldsymbol{x}}) = \int q_{\sigma_0,T}(\tilde{\boldsymbol{x}}|\boldsymbol{x})p(\boldsymbol{x})dx$, where $q_{\sigma_0,T}(\tilde{\boldsymbol{x}}|\boldsymbol{x}) = \mathcal{N}(\tilde{\boldsymbol{x}}|\boldsymbol{x}; T{\sigma_0}^2 I_d)$ and T is the temperature parameter that scales the size of Gaussian kernel (Mehrjou et al., 2017). Naively running traditional denoising score matching for different noise temperatures would lead to different energy functions $E_T(\tilde{\boldsymbol{x}};\theta)$. This is undesirable, the number of free model parameters will be multiplied by the number of different temperatures used during learning, and it is unclear how to to obtain density information from such a set of different energy functions. As an alternative, we propose a model in which the energy functions for different temperatures are collapsed into a single function $E(\tilde{\boldsymbol{x}};\theta)$. We borrow intuition from physics and simply set $E_T(\tilde{\boldsymbol{x}};\theta) = E(\tilde{\boldsymbol{x}};\theta)/T$ and use denoising score matching (2) to approximate $\log(p_{\sigma_0,T}(\tilde{\boldsymbol{x}}))$ by $E(\tilde{\boldsymbol{x}};\theta)/T$:

$$L_{DSM}(\theta;T) = \mathbb{E}_{p_{\sigma 0,T}(\tilde{\boldsymbol{x}},\boldsymbol{x})} \parallel \nabla_{\tilde{\boldsymbol{x}}} \log(q_{\sigma 0,T}(\tilde{\boldsymbol{x}}|\boldsymbol{x})) + \nabla_{\tilde{\boldsymbol{x}}} E(\tilde{\boldsymbol{x}};\theta)/T \parallel^2 \tag{4}$$

In the case of Gaussian noise (3) and by combining objective (4) over a temperature range the objective of *Annealing Denoising Score Matching* is:

$$L_{ADSM}(\theta) = \sum_T l(T)\, \mathbb{E}_{p_{\sigma_0,T}(\tilde{\boldsymbol{x}},\boldsymbol{x})} \parallel \boldsymbol{x}_i - \tilde{\boldsymbol{x}}_{i,T} + \sigma_0^2 \nabla_{\tilde{\boldsymbol{x}}} E(\tilde{\boldsymbol{x}}_{i,T};\theta) \parallel^2 \tag{5}$$

where $l(T)$ is a weighting function that normalizes terms from different temperatures. Here $\boldsymbol{x}_i$ and $\tilde{\boldsymbol{x}}_{i,T}$ denote a clean data point and the same data point corrupted with a noise level corresponding to temperature T. Since The difference $\boldsymbol{x}_i - \tilde{\boldsymbol{x}}_{i,T}$ scales linearly with the noise level, after successful training, we expect $\nabla_{\tilde{\boldsymbol{x}}} E(\tilde{\boldsymbol{x}}_{i,T};\theta)$ to scale similarly, thus the energy $E(\tilde{\boldsymbol{x}}_{i,T};\theta)$ should be roughly quadratic along the noise direction. We present a plausible physics illustration of the learned energy function in Appendix A.

It has to be emphasized that modeling $\log(p_{\sigma_0,T}(\tilde{\boldsymbol{x}}))$ by $E(\tilde{\boldsymbol{x}};\theta)/T$ is an approximation. Specifically, one can think of $p_{\sigma_0,T}(\tilde{\boldsymbol{x}}))$ as a mixture of Gaussians. However, there is no analytic expression for an energy function whose corresponding Boltzmann distribution would equal to this distribution. Therefore it is hard to write down the exact functional form of $E(\tilde{\boldsymbol{x}};\theta)$. The ADSM objective could be seen as a means to obtain an estimate of the true data distribution $p(\boldsymbol{x})$ via the Boltzmann distribution defined by the energy function $E(\boldsymbol{x};\theta)$. Despite the theoretical difficulty, we argue the proposed learning method is desirable by noting the following:

1. Learning is tractable because, like the original denoising score matching, training does not require sampling from the model.
2. As the gradient of energy function is required to point away from the noisy data point in the opposite direction of the clean sample, the noisy sample will have higher energy, and thus lower probability than the clean data point.
3. As temperature increases, the equilibrium distribution will have fewer and fewer modes, as expected from physical intuition. This statement is true since $E(\boldsymbol{x};\theta)/T$ is trained to match $\log(p_{\sigma_0,T}(\tilde{\boldsymbol{x}}))$ and the latter was shown to posses a monotonically decreasing number of modes as T increases (Lindeberg, 2011).
4. Because training is performed over a range of temperatures, the resulting energy function will naturally permit annealing sampling which helps with mode exploration.

It seems that the temperature range and σ_0 in (5) are both important hyper-parameters. However, all that matters is the range of $T\sigma_0^2$, the range of noise magnitude we apply to create the surrogates. Ideally we want our model to work across all noise levels, but applying denoising score matching to very large or very small noise is rather pointless. For very large noise the original sample's information is completely lost. Conversely, in the limit of small noise, the noisy sample is virtually indistinguishable from real data. In neither case can one expect to learn an informative gradient. Thus, in practice the noise range needs only be chosen broad enough to encourage learning of data features over all scales.

2.4 Sampling by Annealed Langevin Dynamics and Jump

Simulated annealing (Kirkpatrick et al., 1983; Neal, 2001) has been successfully applied to challenging computational problems, such as combinatorial optimization. In essence, the mode exploration in an objective function can be improved by sampling first at high temperature and then cooling down gradually. Langevin dynamics (Welling and Teh, 2011) has been applied to sampling from neural network energy functions (Du and Mordatch, 2019; Nijkamp et al., 2019). However, Du and Mordatch (2019) reported that mode exploration on the CIFAR dataset was rather limited. Further, Nijkamp et al. (2019) described that a staggering amount of Langvin steps, about 20k, were required to fully sample the energy function. To improve mode exploration even with small numbers of sampling steps, we propose Annealed Langevin dynamics. In this sampling process the temperature parameter in the Langevin dynamics is adjusted from high to low according to an annealing schedule.

In the model of a physical particle undergoing Brownian motion, temperature in its Langevin equation enters as a $\sqrt{T}$ factor in front of the noise term. Adopting this yields the following sampling process:

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \frac{\epsilon^2}{2}\nabla_{\boldsymbol{x}} E(\boldsymbol{x}_t;\theta) + \epsilon\sqrt{T_t}\mathcal{N}(0, I_d) \tag{6}$$

where T_t follows some annealing schedule, and ϵ denotes step length. During sampling, samples behave very much like physical particles under Brownian motion in a potential field. Because the particles have average energies close to the their current thermic energy, they explore the state space at different distances from data manifold depending on temperature. Eventually, they settle somewhere on the data manifold. A typical annealing process is depicted in Appendix Figure 11B. If the obtained sample is still noisy, we can make a single step gradient denoise jump to improve sample quality:

$$\boldsymbol{x}_{clean} = \boldsymbol{x}_{noisy} - \sigma_0^2\nabla_{\boldsymbol{x}} E(\boldsymbol{x}_{noisy};\theta) \tag{7}$$

This denoising procedure can be applied to noisy sample with any level of Gaussian noise because in our model the gradient automatically has the right magnitude to denoise the sample. Previously, this process has been investigated experimentally in a neural network setting (Saremi et al., 2018; Saremi and Hyvarinen, 2019). In addition, earlier theoretical work has proven that denoising by the gradient of the correct energy function yields the least square estimator for the sample, given the noisy sample (Raphan and Simoncelli, 2011).

3 Image modeling Results

Training and Sampling Details. The proposed energy based model is trained on standard image datasets, namely MNIST, Fashion MNIST, CelebA (Liu et al., 2015) and CIFAR-10 (Krizhevsky et al., 2009). During training we set $\sigma_0 = 0.1$ and train over a noise range of $\sqrt{T}\sigma_0 \in [0.05, 1.2]$, with the noise uniformly spaced on the batch dimension, for MNIST and Fashion MNIST we used geometrically distributed noise in the range $[0.1, 3]$. $l(T)$ is set to $1/T$ at all times to make each terms in the objective from different noise scale roughly the same size. We fix the batch size to 128 and use the Adam optimizer with a learning rate of 5×10^{-5}. For MNIST and Fashion MNIST, we use a 12-Layer ResNet He et al. (2016) with 64 filters, for the CelebA and CIFAT-10 data sets we used a 18-Layer ResNet with 128 filters. No normalization layer is used in any of the networks. We designed the output layer of all networks to take a generalized quadratic form Fan et al. (2018), because the energy function is anticipated to be approximately quadratic with respect to the noise level, we found this modification to boost the performance significantly. For more detail on training

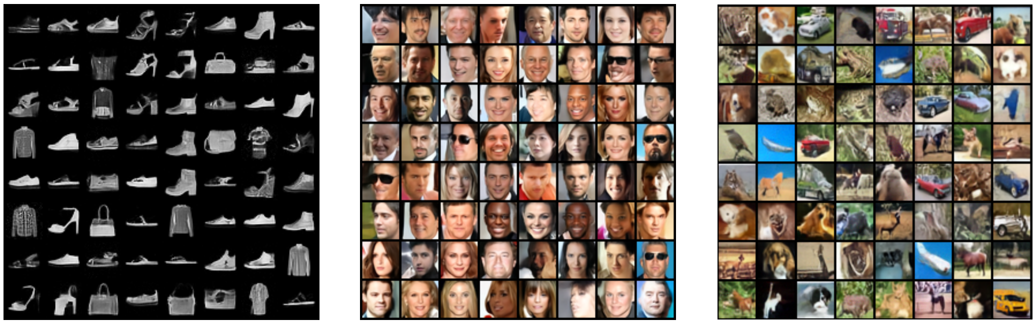

Figure 2: Samples from our model trained on Fashion MNIST, CelebA and CIFAR-10. See Figure 9 and Figure 10 in Appendix for more samples and comparison with training data.

and model architecture, see Appendix. One notable result is that since our training method does not involve sampling, we achieved a speed up of roughly an order of magnitude compared to the common maximum-likelihood training [1]. Our method thus enables the training of energy based models, even when limited computational resources prohibit maximum likelihood methods.

We found that the choice of the maximum noise level has little effect on learning as long as it is large enough to encourage learning of the longest range features in the data. However, as expected, learning with too small or too large noise levels is not beneficial and can even destabilize the training process. Further, our method appeared to be relatively insensitive to how the noise levels are distributed over a chosen range. Geometrically spaced noise as in (Song and Ermon, 2019) and linearly spaced noise both work, although in our case learning with linearly spaced noise was somewhat more robust.

For sampling the learned energy function we used annealed Langevin dynamics with an empirically optimized annealing schedule,see Figure 11 B for a the particular shape of annealing schedule we used. In contrast, annealing schedules with theoretical guaranteed convergence property takes extremely long (Geman and Geman, 1984). For step length ϵ we generally used 0.02, although any value within the range $[0.015, 0.05]$ seemed to work fine. After annealing process we perform a single step denoising jump to further enhance sample quality.

Table 1: Unconditional Inception score, FID scores and Likelihoods for CIFAR-10

Model	IS	FID	Likelihood	NNL (bits/dim)
iResNet (Behrmann et al., 2018)	-	65.01	Yes	3.45
PixelCNN (Oord et al., 2016)	4.60	65.93	Yes	**3.14**
PixelIQN (Ostrovski et al., 2018)	5.29	49.46	Yes	-
Residual Flow (Chen et al., 2019)	-	46.37	Yes	3.28
GLOW (Kingma and Dhariwal, 2018)	-	46.90	Yes	3.35
EBM (ensemble) (Du and Mordatch, 2019)	6.78	38.2	Yes	-[2]
SNGAN (Miyato et al., 2018)	8.22	**21.7**	No	-
ADSM(Ours)	8.31	31.7	Yes	3.76[3]
NCSN (Song and Ermon, 2019)	**8.91**	25.32	No	-

Unconditional Image Generation. We demonstrate the ability of our model to generalize by displaying samples obtained by annealed Langevin and jump sampling in the trained model. We evaluated 50k sampled images after training on CIFAR-10 with two performance scores, Inception (Salimans et al., 2016) and FID (Heusel et al., 2017). We achieved Inception Score of 8.31 and FID of 31.7. Scores for CelebA dataset are not reported here as they are not commonly reported and

[1] For example, on a single GPU, training MNIST with 12-layer Resnet takes 0.3s per batch with our method, while maximum likelihood training with a modest 30 Langevin step per weight update takes 3s per batch.

[2] Author reported difficulties evaluating Likelihood

[3] Upper Bound obtained by reverse AIS

may depend on the specific pre-processing used. More samples and training images are provided in Appendix for visual inspection. We believe that visual assessment is still essential because of issues with the Inception score (Barratt and Sharma, 2018). Indeed, we also found that the visually impressive network was not the one achieving the highest Inception Score.

Although overfitting is not a common concern for generative models, we still tested our model for overfitting. We found no indication for overfitting by comparing model samples with their nearest neighbors in the data set, see Figure 6 in Appendix.

Mode Coverage. We repeated with our model the 3 channel MNIST mode coverage experiment (Kumar et al., 2019). An energy based model was trained on 3-channel data where each channel is a random MNIST digit. Then 8000 samples were taken from the model and each channel is classified using a small MNIST classifier network. We obtained results of the 966 modes, comparable to GAN approaches. Training was successful and our model assigned low energy to all the learned modes, but some modes were not accessed during sampling, likely due to the Langevin Dynamics failing to explore these modes. A better sampling technique such as Neal et al. (2011) or a Maximum Entropy Generator (Kumar et al., 2019) could improve this result.

Image Inpainting. Image impainting can be achieved with our model by clamping part of the image to ground truth and performing the same annealed Langevin and Jump sampling procedure on the missing part of the image. Noise appropriate to the sampling temperature need to be added to the clamped inputs. The quality of inpainting results of our model trained on CelebA and CIFAR-10 can be assessed in Figure 3. For CIFAR inpainting we used test set,

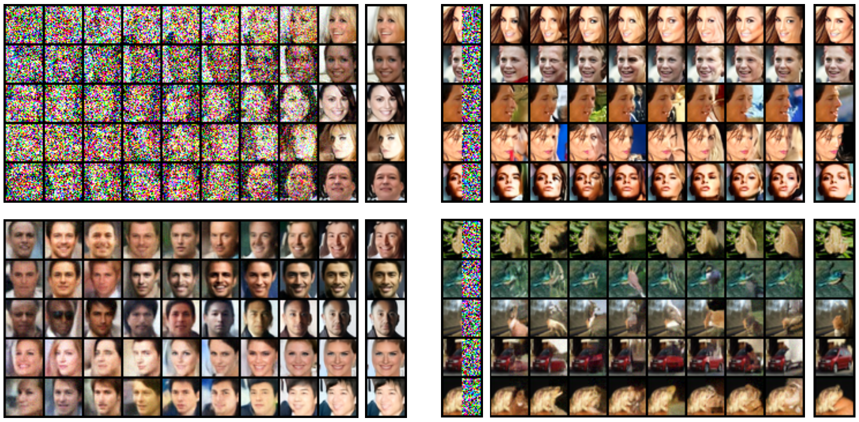

Figure 3: Demonstration of the sampling process (left), and image inpainting (right). The sampling process is shown with Gaussian noise (top left), and denoised by single step gradient jump (lower left). The column next to sampling process shows samples after the last denoising jump at the end of sampling. Inpainting results are shown next to initial image (left column) and the ground truth image (right column).

Log likelihood estimation. For energy based models the log likelihood can be obtained by estimating the partition function with Annealed Importance Sampling (AIS) (Salakhutdinov and Murray, 2008) or Reverse AIS (Burda et al., 2015). AIS is known to underestimate the partition function (Burda et al., 2015), and this was confirmed in our experiment. Therefore, we used only Reverse AIS to estimate the lower bound of the log probability of test samples. We optimized the sampling parameters to obtain the best likelihood results at $T = 1$ with the variance of the estimation still being low.

Table 1 shows the upper bound for the negative log likelihood (lower bound on probability) on CIFAR-10 from our model, in comparison to the log likelihood results from previous probabilistic generative models. More details on this experiment is provided in the Appendix. Further, our likelihood results on MNIST and Fashion MNIST were 1.89 (bits/dim) and 3.01 (bits/dim), respectively.

Outlier Detection. Choi and Jang (2018a) and Nalisnick et al. (2018) have reported intriguing behavior of high dimensional density models on out of distribution samples. Specifically, they showed that a lot of models assign higher likelihood to out of distribution samples than real data samples. We investigated whether our model behaves similarly.

Our energy function is only trained outside the data manifold where samples are noisy, so the energy value at clean data points is not always well behaved. Therefore, we added noise with magnitude σ_0 before measuring the energy value. We report that our network behaves similarly to previous likelihood models that it assigns lower energy, thus higher likelihood, to some OOD samples. We show one example of this phenomenon in Appendix Figure 11 A.

4 Related works

Previous efforts of learning energy based models with score matching Kingma and Cun (2010); Song et al. (2019) were either computationally intensive or unable to produce high-quality samples comparable to those obtained by other generative models such as GANs Goodfellow et al. (2014). Recently, Saremi et al. (2018) and Saremi and Hyvarinen (2019) proposed to train energy based models with denoising score matching, using a single noise scale for training. These models were capable of denoising data with noise levels similar to the one used during training, but the ability to generate data sample from random initialization is limited (See Appendix Figure 5). Other recent efforts on training energy based models used maximum likelihood training (Du and Mordatch, 2019)(Nijkamp et al., 2019). These models have high computational demands, even when combined with acceleration methods such as persistent initialization Tieleman (2008). Also, without applying annealing, the sampling process in these models suffers from poor mixing.

Compared to the described previous studies, our model has various desirable properties, it requires fewer computational resources, is able to denoise inputs with various levels of noise, and it generates high quality samples. The idea of combining denoising score matching and annealing was mentioned already in (Geras and Sutton, 2014; Chandra and Sharma, 2014; Zhang and Zhang, 2018), but in the context of pre-training neural networks for classification applications. Our approach is perhaps most closely related to the NCSN model (Song and Ermon, 2019), a network that is also trained with noisy samples corrupted with Gaussian noise at different levels. In essence, this model learns $p_{\sigma_0,T}(\tilde{\boldsymbol{x}})$ for each T as a separate model, in addition to the input image it receives a signal that represents the noise level explicitly. Our model differs in that it learns one energy model for all temperatures. Further it does not rely on explicit receive noise magnitude information but rather infers the noise magnitude from the given image, which enables applications in single step denoising. Moreover, our model provides density information, whereas it is not obvious how to transform the score output of the NCSN model into a density. Performance scores of the different models can be compared in Table 1. Although not the top performer in any single category, our model assigns competitive likelihood values to test samples, and the same time produces high quality samples comparable to none-likelihood based state-of-the-art models, such as GANs.

5 Conclusion

Our results confirms that the limitation of previous denoising score matching models is due to problems in the learning procedure. To remedy these problems we proposed a learning procedure using a wide range of different noise levels. To sample from the model we propose a novel annealed Langevin sampling and also used single step denoise jump at the end.

The resulting ADSM model is able to overcome several limitations of previous score matching methods. Specifically, we show empirically that the ADSM model is capable of denoising, producing high-quality samples, and performing image inpainting. Our model achieved those while assigning competitive likelihood to test samples, being roughly an order of magnitude faster over maximum likelihood method during training, and admitting more efficient sampling procedure using annealed Langevin sampling.

References

Shane Barratt and Rishi Sharma. A note on the inception score. *arXiv preprint arXiv:1801.01973*, 2018.

Jens Behrmann, David Duvenaud, and Jörn-Henrik Jacobsen. Invertible residual networks. *arXiv preprint arXiv:1811.00995*, 2018.

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Accurate and conservative estimates of mrf log-likelihood using reverse annealing. In *Artificial Intelligence and Statistics*, pages 102–110, 2015.

B Chandra and Rajesh Kumar Sharma. Adaptive noise schedule for denoising autoencoder. In *International conference on neural information processing*, pages 535–542. Springer, 2014.

Ricky TQ Chen, Jens Behrmann, David Duvenaud, and Jörn-Henrik Jacobsen. Residual flows for invertible generative modeling. *arXiv preprint arXiv:1906.02735*, 2019.

Hyunsun Choi and Eric Jang. Generative ensembles for robust anomaly detection. *arXiv preprint arXiv:1810.01392*, 2018a.

Hyunsun Choi and Eric Jang. Generative ensembles for robust anomaly detection. *arXiv preprint arXiv:1810.01392*, 2018b.

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

Yilun Du and Igor Mordatch. Implicit generation and generalization in energy-based models. *arXiv preprint arXiv:1903.08689*, 2019.

Fenglei Fan, Wenxiang Cong, and Ge Wang. A new type of neurons for machine learning. *International journal for numerical methods in biomedical engineering*, 34(2):e2920, 2018.

David J Field. Relations between the statistics of natural images and the response properties of cortical cells. *Josa a*, 4(12):2379–2394, 1987.

Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6): 721–741, 1984.

Krzysztof J Geras and Charles Sutton. Scheduled denoising autoencoders. *arXiv preprint arXiv:1406.3269*, 2014.

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

Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1352–1361. JMLR. org, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017.

Geoffrey E Hinton. Products of experts. 1999.

Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(Apr):695–709, 2005.

Yan Karklin and Eero P Simoncelli. Efficient coding of natural images with a population of noisy linear-nonlinear neurons. In *Advances in neural information processing systems*, pages 999–1007, 2011.

Durk P Kingma and Yann L Cun. Regularized estimation of image statistics by score matching. In *Advances in neural information processing systems*, pages 1126–1134, 2010.

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

Scott Kirkpatrick, C Daniel Gelatt, and Mario P Vecchi. Optimization by simulated annealing. *science*, 220(4598):671–680, 1983.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

Rithesh Kumar, Anirudh Goyal, Aaron Courville, and Yoshua Bengio. Maximum entropy generators for energy-based models. *arXiv preprint arXiv:1901.08508*, 2019.

Neil Lawrence. Probabilistic non-linear principal component analysis with gaussian process latent variable models. *Journal of machine learning research*, 6(Nov):1783–1816, 2005.

Tony Lindeberg. Generalized gaussian scale-space axiomatics comprising linear scale-space, affine scale-space and spatio-temporal scale-space. *Journal of Mathematical Imaging and Vision*, 40(1): 36–81, 2011.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

Arash Mehrjou, Bernhard Schölkopf, and Saeed Saremi. Annealed generative adversarial networks. *arXiv preprint arXiv:1705.07505*, 2017.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

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

Radford M Neal. Annealed importance sampling. *Statistics and computing*, 11(2):125–139, 2001.

Radford M Neal et al. Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2, 2011.

Jiquan Ngiam, Zhenghao Chen, Pang W Koh, and Andrew Y Ng. Learning deep energy models. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 1105–1112, 2011.

Erik Nijkamp, Mitch Hill, Tian Han, Song-Chun Zhu, and Ying Nian Wu. On the anatomy of mcmc-based maximum likelihood learning of energy-based models. *arXiv preprint arXiv:1903.12370*, 2019.

Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. *arXiv preprint arXiv:1601.06759*, 2016.

Georg Ostrovski, Will Dabney, and Rémi Munos. Autoregressive quantile networks for generative modeling. *arXiv preprint arXiv:1806.05575*, 2018.

Martin Raphan and Eero P Simoncelli. Least squares estimation without priors or supervision. *Neural computation*, 23(2):374–420, 2011.

Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.

Ruslan Salakhutdinov and Iain Murray. On the quantitative analysis of deep belief networks. In *Proceedings of the 25th international conference on Machine learning*, pages 872–879. ACM, 2008.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242, 2016.

Saeed Saremi and Aapo Hyvarinen. Neural empirical bayes. *arXiv preprint arXiv:1903.02334*, 2019.

Saeed Saremi, Arash Mehrjou, Bernhard Schölkopf, and Aapo Hyvärinen. Deep energy estimator networks. *arXiv preprint arXiv:1805.08306*, 2018.

Jascha Sohl-Dickstein, Peter B Battaglino, and Michael R DeWeese. New method for parameter estimation in probabilistic models: minimum probability flow. *Physical review letters*, 107(22): 220601, 2011.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *arXiv preprint arXiv:1907.05600*, 2019.

Yang Song, Sahaj Garg, Jiaxin Shi, and Stefano Ermon. Sliced score matching: A scalable approach to density and score estimation. *arXiv preprint arXiv:1905.07088*, 2019.

Terence Tao. *Topics in random matrix theory*, volume 132. American Mathematical Soc., 2012.

Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

Tijmen Tieleman. Training restricted boltzmann machines using approximations to the likelihood gradient. In *Proceedings of the 25th international conference on Machine learning*, pages 1064–1071. ACM, 2008.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(Dec):3371–3408, 2010.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688, 2011.

Li Wenliang, Dougal Sutherland, Heiko Strathmann, and Arthur Gretton. Learning deep kernels for exponential family densities. *arXiv preprint arXiv:1811.08357*, 2018.

Shuangfei Zhai, Yu Cheng, Weining Lu, and Zhongfei Zhang. Deep structured energy based models for anomaly detection. *arXiv preprint arXiv:1605.07717*, 2016.

Qianjun Zhang and Lei Zhang. Convolutional adaptive denoising autoencoders for hierarchical feature extraction. *Frontiers of Computer Science*, 12(6):1140–1148, 2018.

A A PHYSICS ANALOGY TO UNDERSTAND THE ADSM OBJECTIVE FUNCTION

Here we provide a concrete physics picture of the energy function learned by annealed score matching. Given a data manifold X, we define the energy function as:

$$E^*(d) = \frac{1}{2}kd^2 \tag{8}$$

where $d = \parallel \tilde{\boldsymbol{x}} - proj_X(\tilde{\boldsymbol{x}}) \parallel$ and $proj_X(\tilde{\boldsymbol{x}})$ is the projection operator to the data manifold X. One can easy verify that this is an energy function and describes a generalized Hookean spring system. At any point $\tilde{\boldsymbol{x}}$ away from the data manifold X, the force F is:

$$F = -\nabla_x E^*(\tilde{\boldsymbol{x}}) = k(\tilde{\boldsymbol{x}} - proj_X(\tilde{\boldsymbol{x}})) \tag{9}$$

as Figure 4 shows.

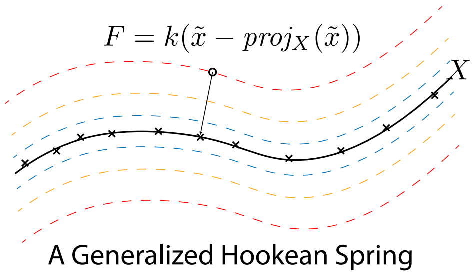
$F = k(\tilde{x} - proj_X(\tilde{x}))$
X
A Generalized Hookean Spring

Figure 4: Generalized Hookean spring system, where the force F at any noise-corrupted point $\tilde{\boldsymbol{x}}$ points towards its projection $proj_X(\tilde{\boldsymbol{x}})$ on the data manifold X and is proportional to the distance from $\tilde{\boldsymbol{x}}$ to X.

In high dimensional space, given a data point $\boldsymbol{x}$ and a Gaussian noise corrupted version $\tilde{\boldsymbol{x}}$, $\tilde{\boldsymbol{x}} - \boldsymbol{x}$ provides a good approximation to $\tilde{\boldsymbol{x}} - proj_X(\tilde{\boldsymbol{x}})$. Thus, by matching the score to $\tilde{\boldsymbol{x}} - \boldsymbol{x}$ at every noisy point, the ADSM objective is an approximation to the energy defined by Equation 8, trained at different distances. Empirically, we also verified that the $E(\tilde{\boldsymbol{x}}; \theta)$ increases in quadratic fashion $w.r.t.$ $\parallel \tilde{\boldsymbol{x}} - \boldsymbol{x} \parallel$.

B FAILURE OF SINGLE NOISE DENOISING SCORE MATCHING

To compare with previous method, we trained energy based model with denoising score matching using one noise level on MNIST, initialized the sampling with Gaussian noise of the same level, and sampled with Langevin dynamics at $T = 1$ for 1000 steps and perform one denoise jump to recover the model's best estimate of the clean sample, see Figure 5. We used the same 12-layer ResNet as other MNIST experiments. Models were trained for 100000 steps before sampling with 1000 Langevin steps.

C OVERFITTING TEST

We demonstrate that the model does not simply memorize training examples by comparing model samples with their nearest neighbors in the training set. We use Fashion MNIST for this demonstration because overfitting can occur there easier than on more complicated datasets, see Figure 6.

D DETAILS ON TRAINING AND SAMPLING

We used a custom designed ResNet architecture for all experiments. For MNIST and Fashion-MNIST we used a 12-layer ResNet with 64 filters on first layer, while for CelebA and CIFAR dataset we used a 18-layer ResNet with 128 filters on the first layer. All network used the ELU activation

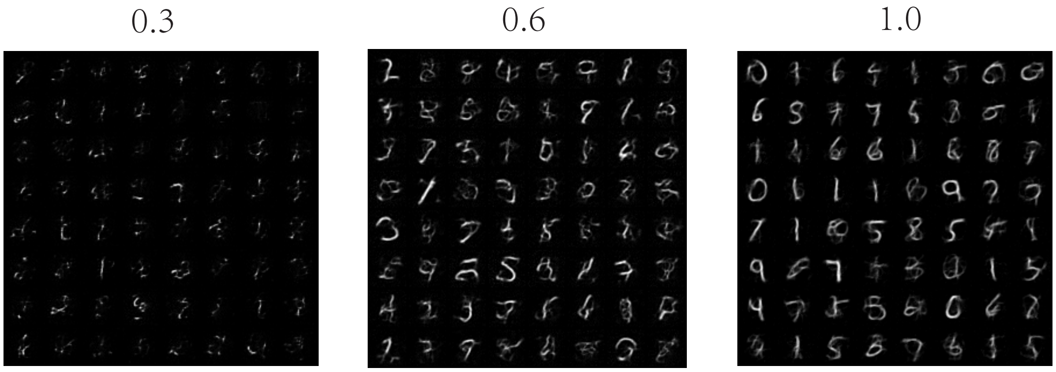

Figure 5: Denoised samples from energy based model trained with denoising score matching with single magnitude Gaussian noise on MNIST. Noise magnitude used in training is shown above samples.

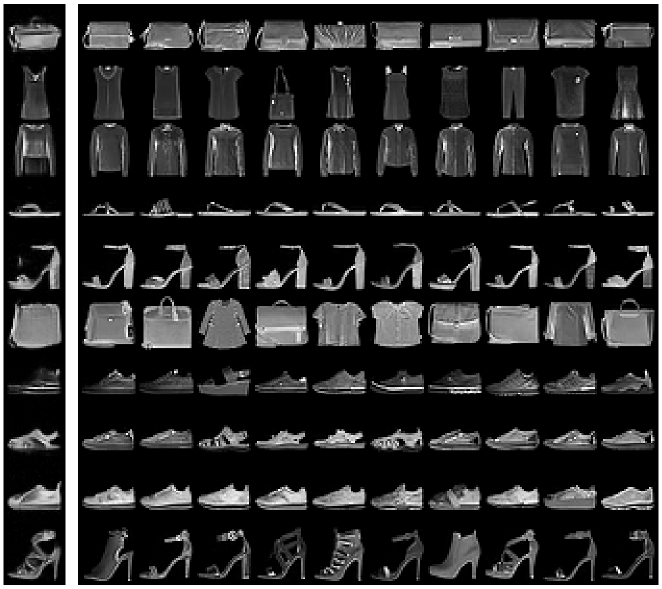

Figure 6: Samples from energy based model trained on Fashion MNIST (Left column) next to 10 (L2) nearest neighbors in the training set.

function. We did not use any normalization in the ResBlocks and the filer number is doubled at each downsampling block. Details about the structure of our networks used can be found in our code release.

Since the gradient of our energy model scales linearly with the noise, we expected our energy function to scale quadratically with noise magnitude. Therefore, we modified the standard energy based network output layer to take a flexible quadratic form (Fan et al., 2018):

$$E_{out} = (\sum_i a_i h_i + b_1)(\sum_i c_i h_i + b_2) + \sum_i d_i h_i^2 + b_3 \tag{10}$$

where a_i, c_i, c_i and b_1, b_2, b_3 are learnable parameters, and h_i is the (flattened) output of last residual block. We found this modification to significantly improve performance compared to using a simple linear last layer.

For CIFAR and CelebA results we trained for 300k weight updates, saving a checkpoint every 5000 updates. We then took 1000 samples from each saved networks and used the network with the lowest FID score. For MNIST and fashion MNIST we simply trained for 100k updates and used the last checkpoint. During training we pad MNIST and Fashion MNIST to 32*32 for convenience and randomly flipped CelebA images. No other modification was performed. Although we only constrained the gradient of the energy function, the energy value itself could in principle be unbounded. However, we observed that they naturally stabilized so we did not explicitly regularize them, unlike in Du and Mordatch (2019).

For the Log likelihood estimation we initialized reverse chain on test images plus $0.05 * \mathcal{N}(0, I_d)$ noise for stability, and then run 10000 Langevin updates towards an isotropic Gaussian reference distribution. The variance of estimation was generally less then 10%.

E Extended Samples and Inpainting results

We further demonstrate the mixing process by providing more denoised images throughout sampling process.

We provide more samples to enable the reader to visually judge the quality of our sample generation in Figure 8, 9 and 10. All samples are randomly selected.

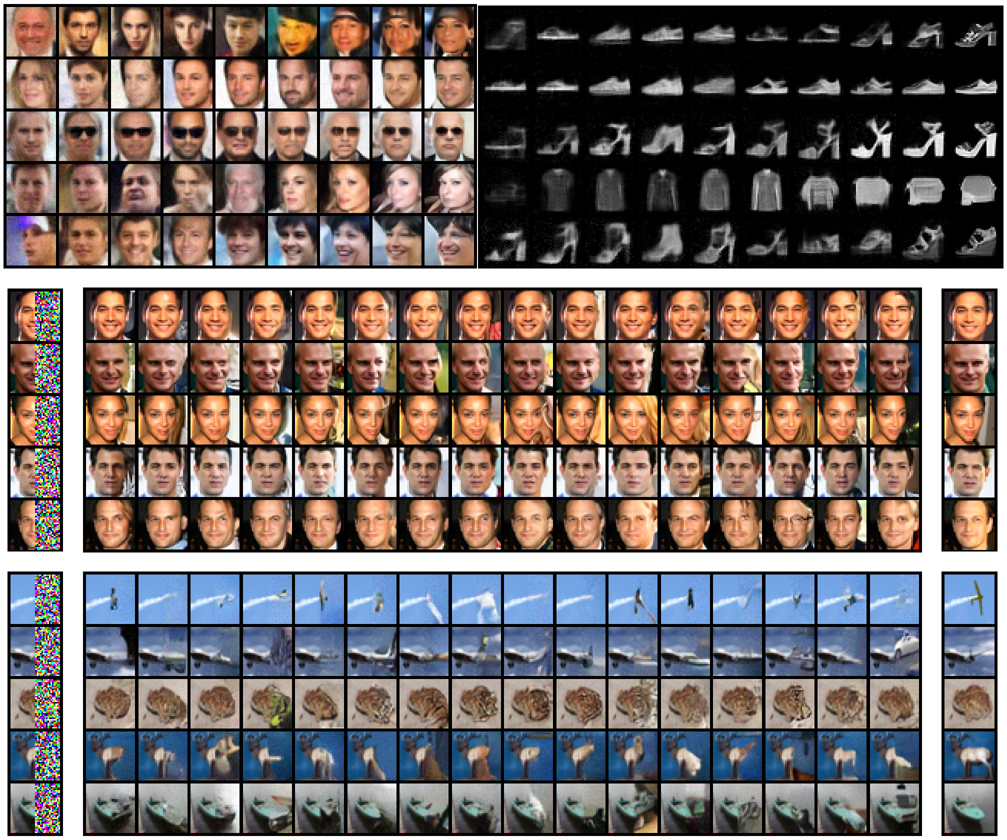

Figure 7: Denoised Sampling process and inpainting results. Sampling process is from left to right.

F Sampling process and Energy value comparisons

Here we show how the average energy of samples behaves vs the sampling temperature. We also show an example of our model making out of distribution error that is common in most other likelihood based models (Nalisnick et al., 2018) Figure 11.

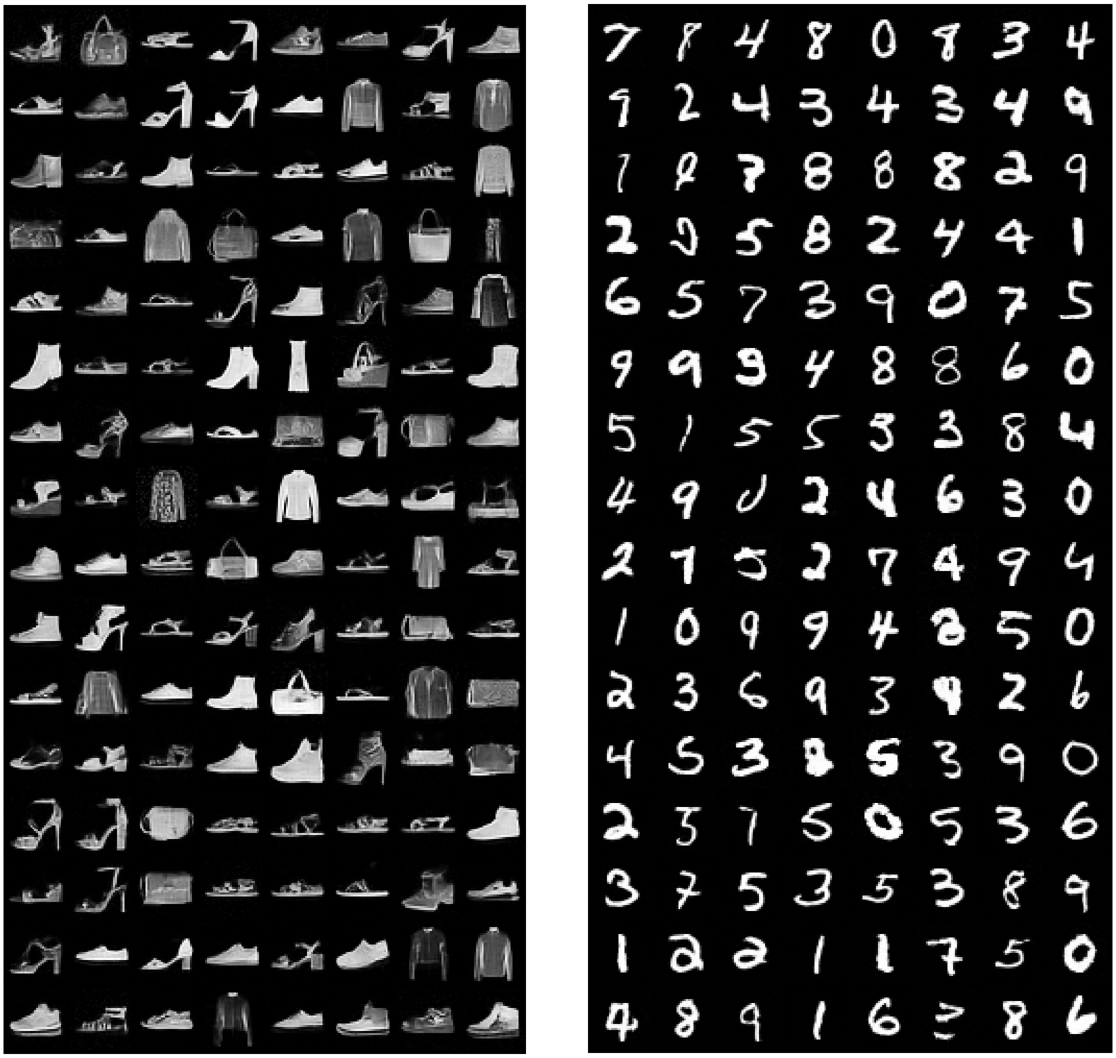

Figure 8: Extended Fashion MNIST and MNIST samples

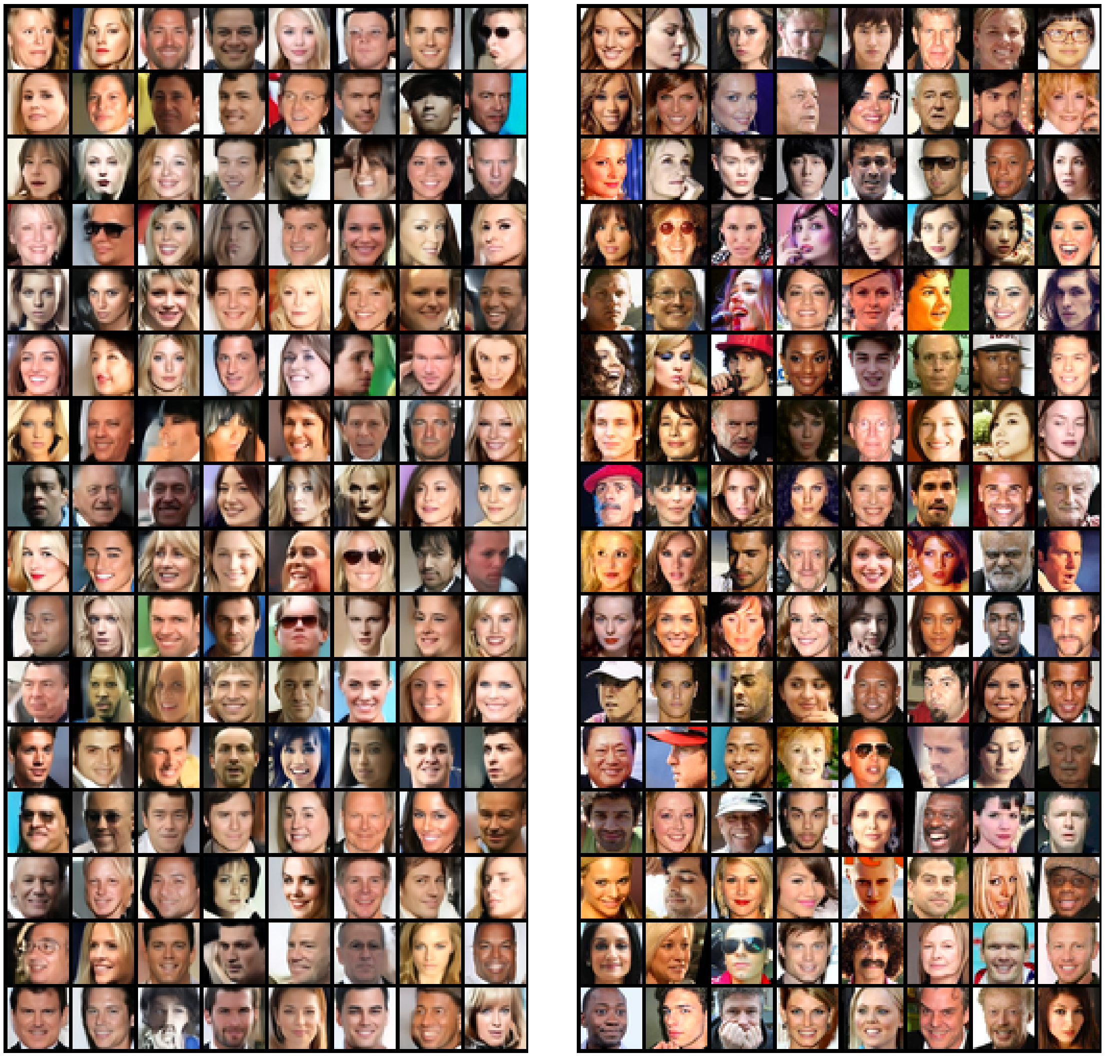

Figure 9: Samples (left panel) from network trained on CelebA, and training examples from the dataset (right panel).

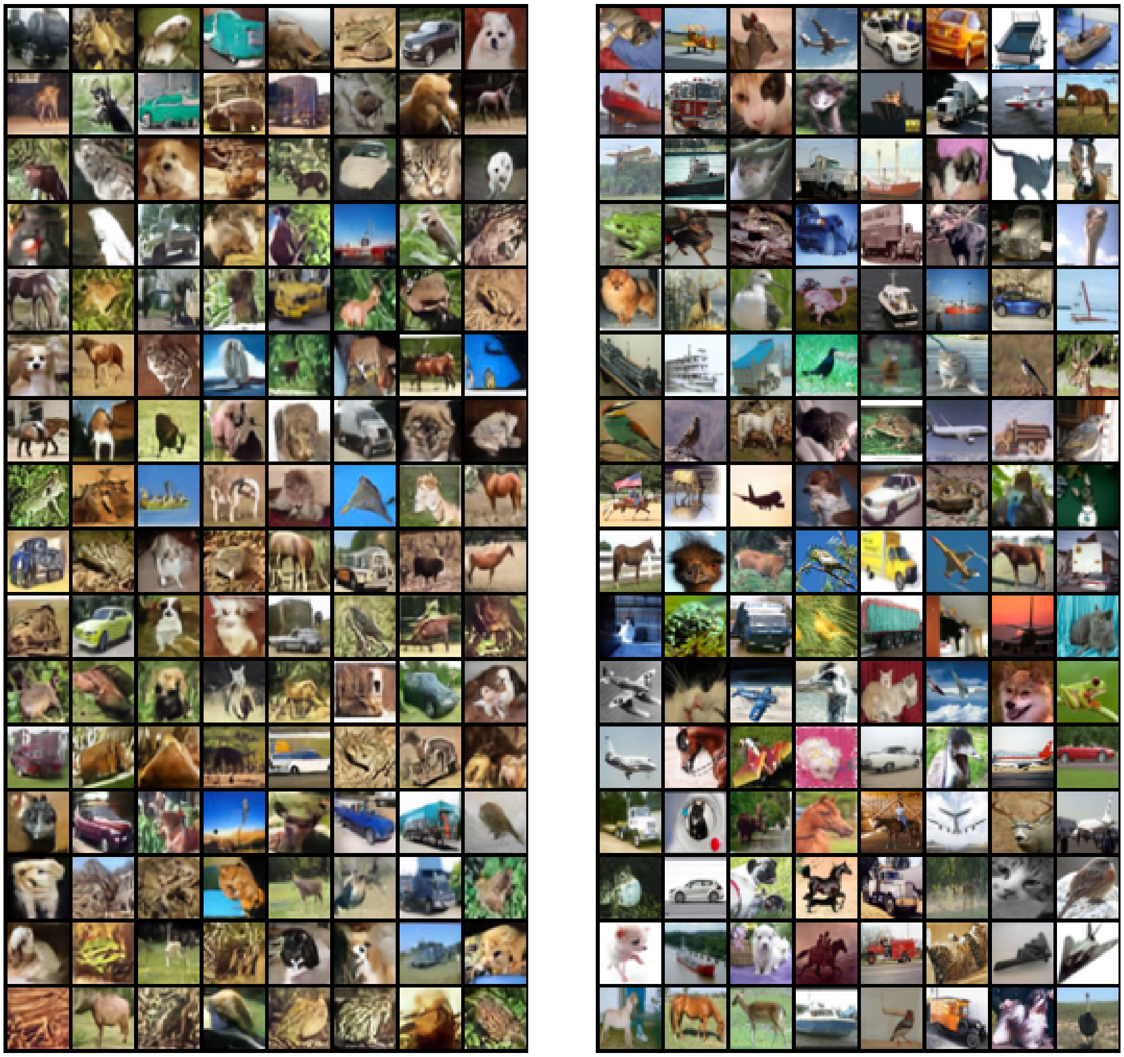

Figure 10: Samples (left panel) from energy based model trained on CIFAR-10 next to training examples (right panel).

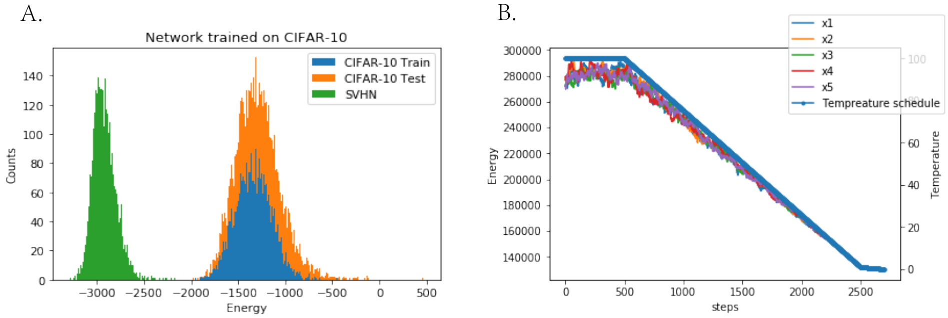

Figure 11: A. Energy values for CIFAR-10 train, CIFAR-10 test and SVHN datasets for a network trained on CIFAR-10 images. Note that the network does not over fit to the training set, but just like most deep likelihood model, it assigns lower energy to SVHN images than its own training data. B. Annealing schedule and a typical energy trace for a sample during Annealed Langevin Sampling. The energy of the sample is proportional to the temperature, indicating sampling is close to a quasi-static process.