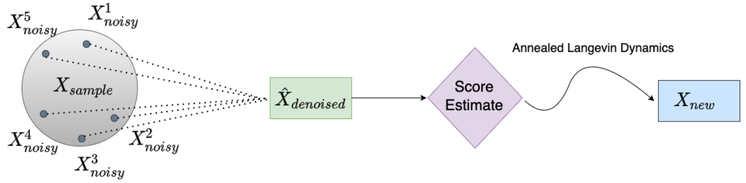
Score-Based Generative Modeling with Multi-Sample Denoiser

Anuj Nagpal

ICME, Stanford University

Objective

- **Score-based generative models** are gaining a lot of traction due to their GAN-level image sampling quality without adversarial training.
- **Denoising autoencoders** are equivalent to score-matching models, with the relation mathematically captured by **Tweedie's formula**.
- Conventionally, a single data point is perturbed and used for **denoising score matching** objective which incurs some approximation error.
- We extend the denoising score matching to **multiple perturbed samples** and show that it can decrease denoising error and give better score estimates.



Mathematical Idea

Tweedie formula states that if we have a denoised least squares estimate $\hat{x}$ from a perturbed sample y, we can estimate the score function as:

$$\nabla_x \log p(x) = \frac{\hat{x} - y}{\sigma^2}$$

where σ is the noise scale of perturbation.

For k **independent noisy observations** $y_1, y_2, \ldots, y_k$ of a random variable x where $P(y_i|x) = \mathcal{N}(y_i; x, \sigma^2)\ \forall\ i$, it can be proved that:

$$\frac{\partial \log P(y_1, y_2, ..., y_k)}{\partial y_i} = \frac{\hat{x}(y_1, y_2, ..., y_k) - y_i}{\sigma^2}$$

Thus given a denoiser based on k samples, we can get a score estimate and run **chained langevin dynamics** to sample from $P(y_1, y_2, ..., y_k)$ as follows:

$$\begin{bmatrix} y_{1,t+1} \\ y_{2,t+1} \\ \cdots \\ y_{k,t+1} \end{bmatrix} \leftarrow \begin{bmatrix} y_{1,t} \\ y_{2,t} \\ \cdots \\ y_{k,t} \end{bmatrix} + \epsilon \begin{bmatrix} \nabla_{y_1} \log P(y_1, ..., y_k) \\ \nabla_{y_2} \log P(y_1, ..., y_k) \\ \cdots \\ \nabla_{y_k} \log P(y_1, ..., y_k) \end{bmatrix} + \sqrt{2\epsilon} \begin{bmatrix} z_1 \\ z_2 \\ \cdots \\ z_k \end{bmatrix}$$

where $z_1, z_2, ..., z_k \sim \mathcal{N}(0, I)$

Technical Approach

- We use a **diffusion process** to slowly add noise to the input data, using the stochastic differential equation $dx = \sigma^t dw$.
- A **multi-sample denoiser** is then trained using **MSE loss** which can extract a denoised estimate of the original data point given multiple independent perturbations of it.
- A score estimate can be obtained from the denoised output using the **multi-sample tweedie formula**.
- This score estimate can be used to generate new samples from model distribution by running chains of **annealed langevin dynamics**.

Why this is better?

We claim that using multiple noisy samples instead of one gives our generative model an increased capacity of modeling the true distribution and incurs lesser approximation error.

Model Design

- We implemented a **U-net denoiser** which takes in multiple noisy images concatenated along the channel layer so that instead of 3 RGB channels, we will now have **3 x #samples number of channels**.
- **Multiple independent noisy perturbations** are sampled using a single noise scale, i.e, $\tilde{x}_i = x + \sigma z_i$ where $z_i \sim N(0, I)$.
- Noise scale embedding is also fed into the denoiser via **Gaussian Random Feature** encoding of the time step t of the SDE.

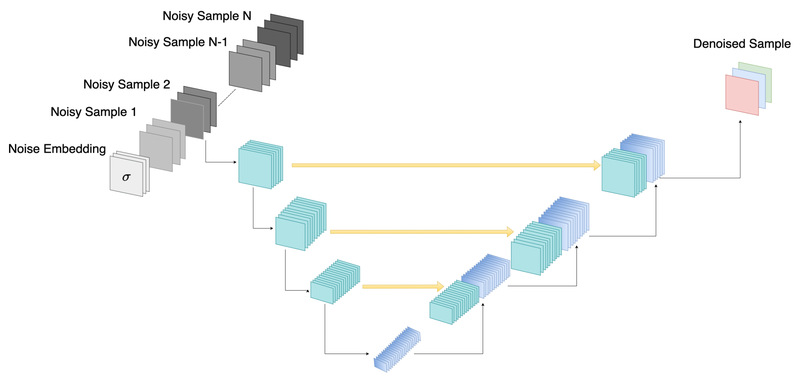
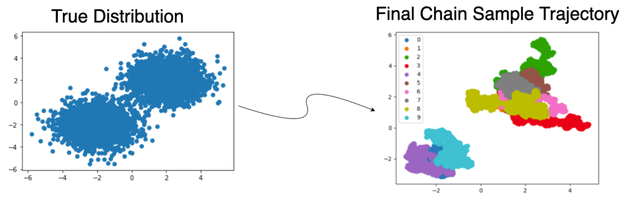


Figure 1. Multi-Sample Denoiser

Experimental Results

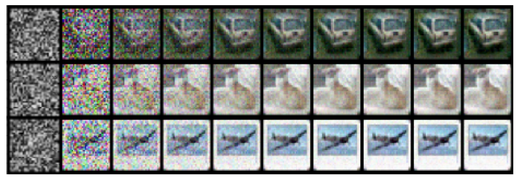
We trained our denoising model on the following datasets in the order of increasing complexity and were able to generate **realistic samples** from complete noise:

Gaussian Mixture Model



MNIST Images

CIFAR-10 Images



We observed **increased denoising capacity** with an increase in number of samples, thus capable of giving a **better score estimate for sampling**:

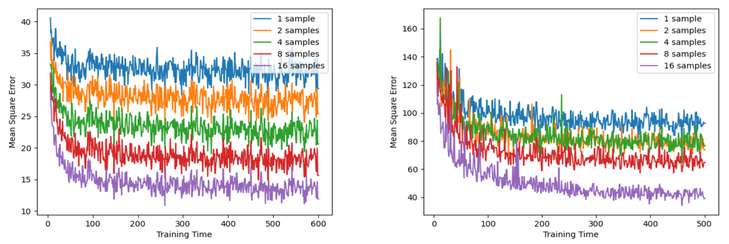


Figure 2. Denoising Error vs Training Time for MNIST (left) and CIFAR-10 (right)

- Special thanks to Chris Cundy from SAIL's Ermon Group