Michaelmas Term IIB Project Presentation

Variance Investigation of Learning Latent Space Energy-Based Prior Model

Prithvi Raj [pr478@cam.ac.uk]

November 20, 2023
IIB Project Code: D-mag92-1

Supervisors:
Prof. Mark Girolami
Mr. Justin Bunker

Context

Latent space modelling – low-dimensional representation of the data that encapsulates its fundamental characteristics.

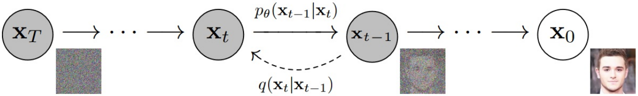


Figure: Example of latent-space learning with DDPM[1]

Some benefits:

- Dimensionality Reduction
- Interpretability
- Improved Generalisation

[1] Jonathan Ho, Ajay Jain, and Pieter Abbeel (2020). "Denoising Diffusion Probabilistic Models". In: *CoRR* abs/2006.11239. arXiv: 2006.11239. URL: https://arxiv.org/abs/2006.11239.

Latent Space Energy-Based Prior Model[2]

Energy-based learning to better capture the latent space prior distribution:

$$p_\alpha(z) = \frac{1}{Z_\alpha} \exp(f_\alpha(z)) \cdot \pi_0(z)$$

Top-down generator to map the acquired latent representations back to the original observable data:

$$x = g_\beta(z) + \epsilon$$

[2] Bo Pang et al. (2020). "Learning Latent Space Energy-Based Prior Model". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., pp. 21994–22008. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/fa3060edb66e6ff4507886f9912e1ab9-Paper.pdf.

Training

$$p_\alpha(z) = \frac{1}{Z_\alpha} \exp(f_\alpha(z)) \cdot \pi_0(z) \qquad x = g_\beta(z) + \epsilon$$

1. Sample: $z_{prior} \sim p_\alpha(z)$, $z_{posterior} \sim p_\theta(z|x) = p_\beta(x|z)p_\alpha(z)$
2. Predict: $x = g_\beta(z_{posterior}) + \epsilon$
3. Maximum Likelihood:

$$\nabla_\theta \log(p_\theta(x)) = \mathbb{E}_{p_\theta(z|x)}\left[\nabla_\theta \log(p(x,z))\right]$$

$$= \mathbb{E}_{p_\theta(z|x)}\left[\nabla_\theta \log(p_\alpha(z)) + \nabla_\theta \log(p_\beta(x|z))\right]$$

Closer Analysis of Maximum Likelihood Terms

$$\nabla_\theta \log(p_\alpha(z)) \propto \mathbb{E}_{p_\theta(z|x)}\left[\nabla_\alpha f_\alpha(z)\right] - \mathbb{E}_{p_\alpha(z)}\left[\nabla_\alpha f_\alpha(z)\right]$$

$$\nabla_\theta \log(p_\beta(x|z)) \propto \mathbb{E}_{p_\theta(z|x)}\left[\nabla_\beta \log(p_\beta(x|z))\right]$$

EBM-Learning involves closely "matching" $p_\alpha(z)$ to $p_\beta(z|x)$, (get a more informed prior). GEN-Learning involves maximising $p_\theta(x)$. essentially as an MSE loss between prediction and observation.

For generation, sample $z \sim p_\alpha(z)$, then pass through generator, $x = g_\beta(z)$. Assume $p_\alpha(z)$ has closely matched $p_\theta(z|x)$

Rough Around the Edges

Langevin sampling is employed to sample from complex/intractable probability distributions.

$$z_0 \sim p_0(z), \quad z_{k+1} = z_k + s\nabla_z \log \pi(z_k) + \sqrt{2s}\varepsilon_k, \quad k = 1, \ldots, K.$$

- No error correction, (all proposed steps were accepted)
- Short run MCMC, (20 iterations were used)

Despite the "statistical flexibility" highly-defined images can be generated...

Example

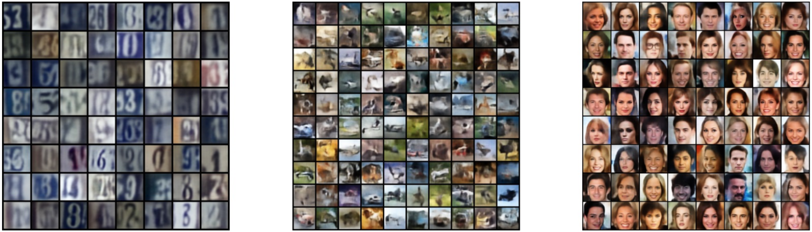


Figure: Model-generated images after training on SVHN, CIFAR-10, and Celeb-A respectively[3]

[3] Bo Pang et al. (2020). "Learning Latent Space Energy-Based Prior Model". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., pp. 21994–22008. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/fa3060edb66e6ff4507886f9912e1ab9-Paper.pdf.

My Project

Would enhancing the model's ability to produce samples with **lower variance** in their estimated likelihoods result in improved performance?

Taking a step further, we propose an investigation into the **gradient** of the log-marginal likelihood, given its existing availability as the learning gradient.

$$\nabla_\theta \log(p_\theta(\mathbf{x})) = \mathbb{E}_{p_\theta(\mathbf{z}|\mathbf{x})}\left[\nabla_\theta \log(p(\mathbf{x}, \mathbf{z}))\right]$$

By incorporating thermodynamic integration[4] into the assessment of the marginal likelihood, it is proposed that the variance in the MCMC estimate can be adjusted by leveraging power posteriors[5].

---

[4] Ben Calderhead and Mark Girolami (2009). "Estimating Bayes factors via thermodynamic integration and population MCMC". In: *Computational Statistics Data Analysis* 53.12, pp. 4028–4045. ISSN: 0167-9473. DOI: https://doi.org/10.1016/j.csda.2009.07.025. URL: https://www.sciencedirect.com/science/article/pii/S0167947309002722.

[5] N. Friel and A. N. Pettitt (2008). "Marginal Likelihood Estimation via Power Posteriors". In: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 70.3, pp. 589–607. ISSN: 13697412, 14679868. URL: http://www.jstor.org/stable/20203843 (visited on 10/26/2023).

Power Posterior Modification

$$p_\alpha(\mathbf{z}) = \frac{1}{Z_\alpha} \exp(f_\alpha(\mathbf{z}))\pi_0(\mathbf{z})$$

$$\mathbf{x} = g_\beta(\mathbf{z}) + \epsilon \quad \implies \quad p_\beta(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}; g_\beta(\mathbf{z}), \sigma_\epsilon^2 \mathbf{I})$$

$$p_\theta(\mathbf{z}|\mathbf{x}, t) = \frac{p_\beta(\mathbf{x}|\mathbf{z})^t p_\alpha(\mathbf{z})}{Z_{\alpha,\beta,t}}$$

Now the optimisation has some dependence on the temperature scheduling, t:

$$\theta^* = \arg\max_\theta \log(p_\theta(\mathbf{x})) \propto \int_0^1 \mathbb{E}_{p_\theta(\mathbf{z}|\mathbf{x},t)} \left[\log(p_\theta(\mathbf{x}|\mathbf{z}))\right] dt$$

Questions to Answer

Does thermodynamic integration reduce variance in $\nabla_\theta \log(p_\theta(\mathbf{x}))$ compared to the vanilla model?

Does reduced variance improve the model's performance?

Experimental Outline - Slide 1

- **Training:**
  - Train both models (vanilla and altered with thermodynamic integration) for an equal number of epochs.
- **Metrics Extraction:**
  - During training, periodically extract $\nabla_\theta \log(p(x|\theta))$ during training.
  - Calculate average variances and collect FID scores, similar to Faghri's approach[6].
- **Evaluation:**
  - After training, evaluate models on test datasets.
  - Collect variances, FID scores, and generate images.

[6] Fartash Faghri et al. (July 2020). "A study of gradient variance in deep learning". In: arXiv: 2007.04532 [cs.LG].

Experimental Outline - Slide 2

- **Iterative Comparison:**
  - If more data is required, modify temperature schedules or network architectures and repeat.
  - Obtain a comprehensive dataset.
- **Visualization:**
  - Plot variance against training iteration.
  - Plot FID score against training iteration.
  - Plot FID score against variance for each model.
  - Compare and contrast the results.

Progress

- Ramping up on literature and PyTorch. Read paper then implement.
- Implemented denoising diffusion model and denoising score matching model.
- Coded up my own Latent EBM/Generator model. Developed troubleshooting skills.

Target Dataset

Any examples are from training to generate make_moons dataset:

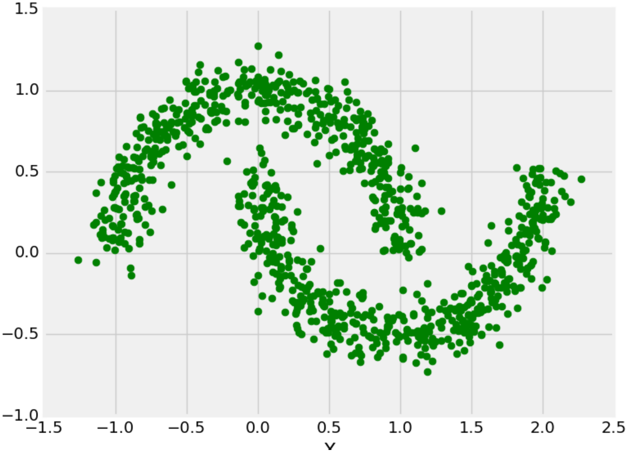


Figure: Sklearn's make_moons Dataset

Denoising Diffusion Model – Diffusion/Noise Addition

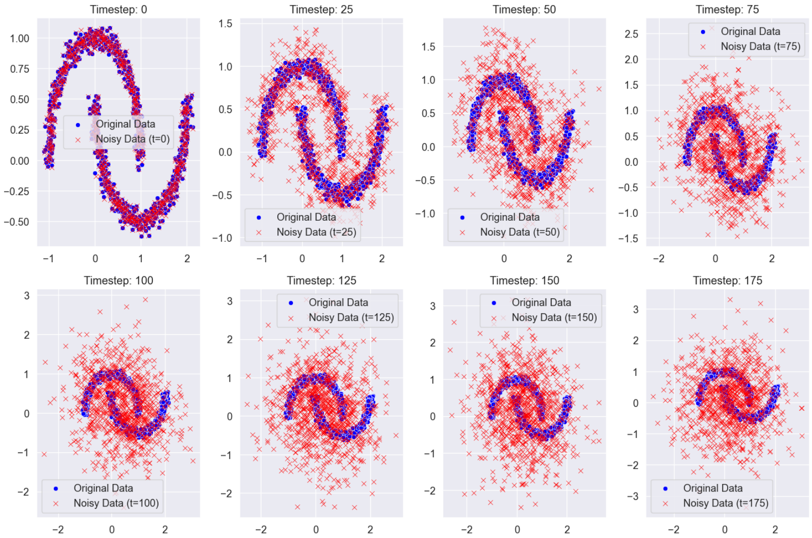


Figure: Adding noise to create a latent representation.

Denoising Diffusion Model – Sample Generation

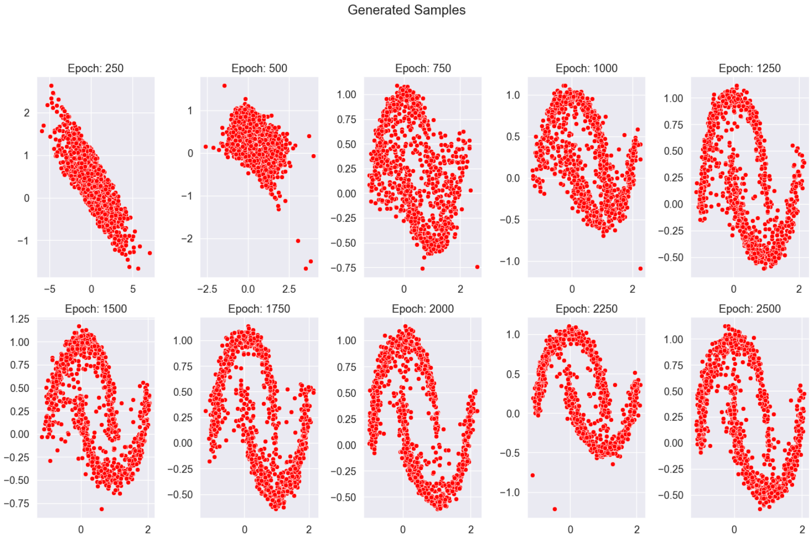


Figure: Samples generated by DDM during training.

Denoising Score Model – Noise Addition

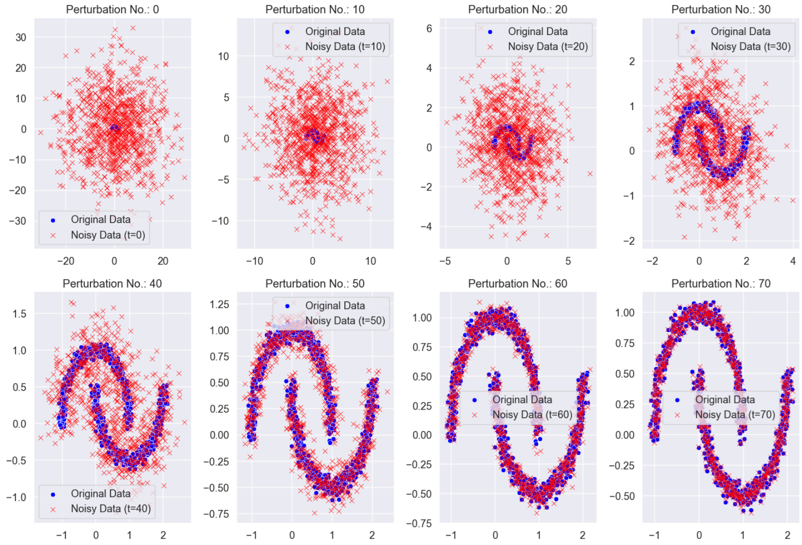


Figure: Noise addition in denoising score model.

Denoising Score Model – Sample Generation

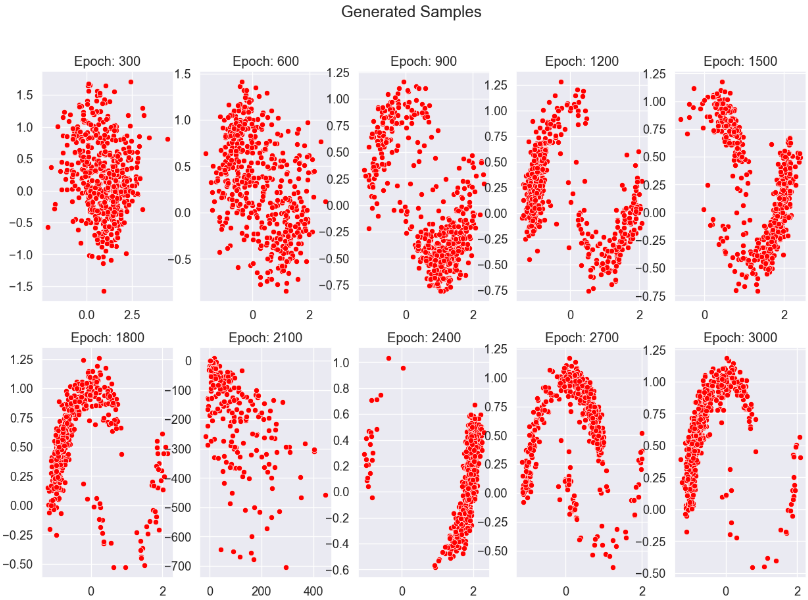


Figure: Samples generated by denoising score model during training.

Latent EBM/Generator Failed Attempt

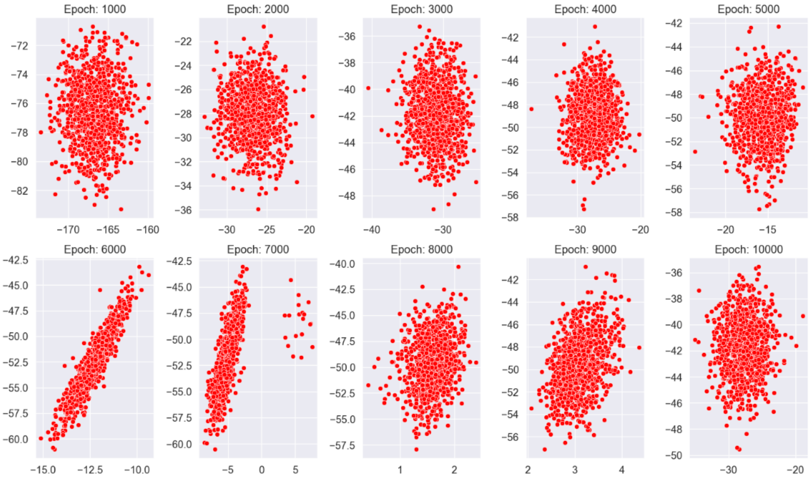
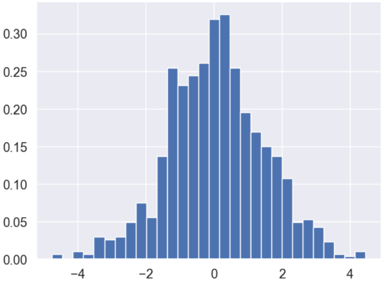
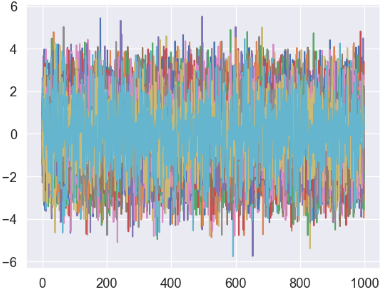


Figure: Model failed to generate samples!

Troubleshooting



(a) Histogram of Samples



(b) Evolution of Chains (too noisy?)

Figure: Langevin Sampler Troubleshooting

Troubleshooting

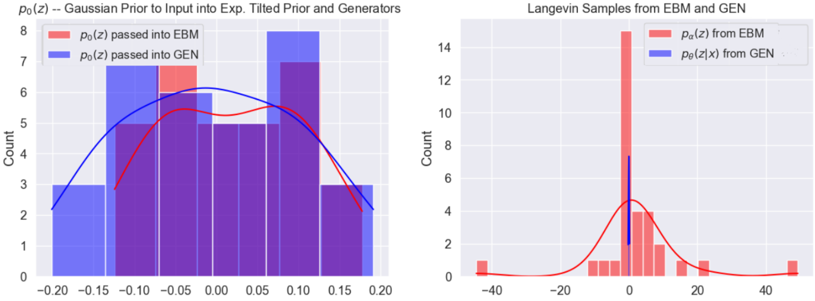
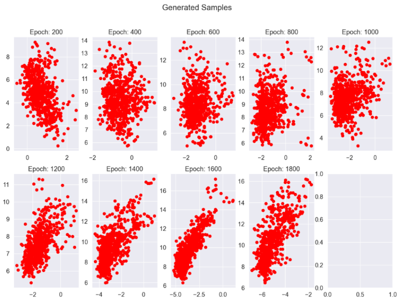
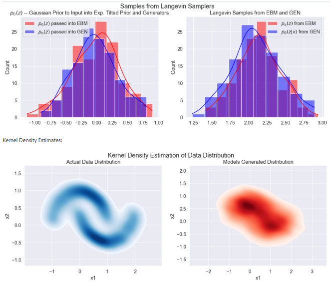


Figure: Histogram of samples from $\pi_0(z), p_(z), andp_\theta(z|x)$

Make Moons Retry



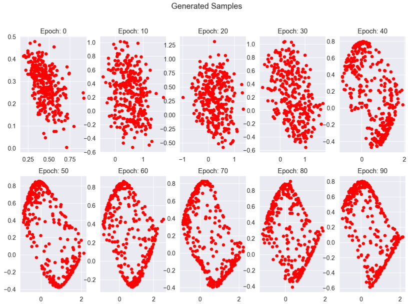
(a) Generated Samples



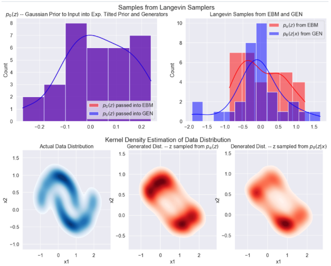
(b) Metrics

Figure: Learning the Make Moons Dataset

Make Moons Close-enough?



(a) Generated Samples



(b) Metrics

Figure: Successful EBM/Generator Make Moons Generation