Score Matching for Generative Modeling

Halvard Bariller [1] Franki Nguimatsia Tiofack [2] Junior Cedric Tonga [1]

[1]ENS Paris-Saclay [2]ENSAE

Summary

- Score-Based approach for Generative Modeling: distribution's score estimation followed by Langevin Dynamics
- Addresses issues of manifold hypothesis and low-density areas of the distribution's support
- Train a single network on different levels of noise and sample new data by tuning down the amount of added noise
- Allows to generate high-quality samples from complex distributions

Context: Generative Models

Generative Models: **Data generation** (ever more complex models require more data for training), **data quality enhancement** (e.g., image inpainting), **robustness improvement** (e.g., outlier detection) ...

General objective: Learn unknown and complex probability distributions (e.g., images), and potentially sample data from it.

Classical Statistical Approach:

- $\mathcal{D}_N = \{\mathbf{x}_n \in \mathbb{R}^d\}_{1\leq n\leq N}$ i.i.d. samples from unknown probability distribution $\mathbf{x}_n \sim p_{data}(\mathbf{x})$
- Family of parameterized models $\mathcal{P} = (p_\theta)_{\theta\in\mathbb{R}^d}$
- Estimate θ s.t. p_θ such that $p_\theta(\mathbf{x}) \approx p_{data}(\mathbf{x})$ (unknown)

$$p_\theta(\mathbf{x}) = \frac{e^{q_\theta(\mathbf{x})}}{Z(\theta)} \quad \text{where} \quad Z(\theta) = \int_{\mathbf{x}} e^{q_\theta(\mathbf{x})} d\mathbf{x}$$

.

Limitation: Normalizing term $Z(\theta)$ (usually) intractable

Score

Definition: *(Score function)* $\mathbf{s}_\theta(\mathbf{x}) \triangleq \nabla_{\mathbf{x}} \log p_\theta(\mathbf{x})$

Circumvent normalizing term computation: $\mathbf{s}_\theta(\mathbf{x}) = \nabla_{\mathbf{x}} q_\theta(\mathbf{x}) - \underbrace{\nabla_{\mathbf{x}} \log Z(\theta)}_{=0}$

Score-Based Approach: $\mathbf{s}_\theta(\mathbf{x}) \approx \nabla_{\mathbf{x}} \log p_{data}(\mathbf{x})$

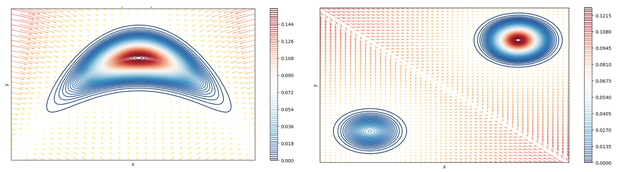

Figure 1. Density functions and associated Score functions (vector fields)

Langevin Monte-Carlo Sampling

Problem: How to sample data knowing only the score (and not the density) of a distribution?

Idea: Use $\nabla_{\mathbf{x}} \log p_\theta(\mathbf{x})$ to head towards high-density areas of the distribution's support

Definition: *(Discrete-time Langevin Dynamics)*

$$\mathbf{x}_t = \mathbf{x}_{t-1} + \frac{\epsilon}{2}\nabla_{\mathbf{x}} \log p_\theta(\mathbf{x}_{t-1}) + \sqrt{\epsilon}\mathbf{z}_t \quad \text{where} \quad \begin{cases} \mathbf{z}_t \sim \mathcal{N}(0, I_d) & \text{(Gaussian noise)} \\ \epsilon > 0 & \text{(Step size)} \end{cases}$$

Score Matching

Context: $\{\mathbf{x}_n\}_{1\leq n\leq N}$ i.i.d. drawn from unknown data distribution $\mathbf{x} \sim p_{data}$.
Goal: Estimate the score function $\mathbf{s}_\theta : \mathbb{R}^d \to \mathbb{R}^d$ s.t. $\mathbf{s}_\theta(\mathbf{x}) \approx \nabla_{\mathbf{x}} \log p_{data}(\mathbf{x})$

Objective function: Fisher Divergence (Explicit Score Matching)

$$J_{ESM}(\theta) \triangleq \frac{1}{2}\mathbb{E}_{p_{data}(\mathbf{x})}\left[||\mathbf{s}_\theta(\mathbf{x}) - \nabla_{\mathbf{x}} \log p_{data}(\mathbf{x})||_2^2\right] \tag{1}$$

Implicit Score Matching [Hyvärinen and Dayan, 2005]

$$J_{ISM}(\theta) \triangleq \mathbb{E}_{p_{data}(\mathbf{x})}\left[\text{tr}\left(\nabla_{\mathbf{x}}\mathbf{s}_\theta(\mathbf{x})\right) + \frac{1}{2}||\mathbf{s}_\theta(\mathbf{x})||_2^2\right] + c \tag{2}$$

Denoising Score Matching [Vincent, 2011]

$$J_{DSM_\sigma}(\theta) = \frac{1}{2}\mathbb{E}_{q_\sigma(\tilde{\mathbf{x}},\mathbf{x})}\left[||\mathbf{s}_\theta(\tilde{\mathbf{x}}) - \nabla_{\tilde{\mathbf{x}}} \log q_\sigma(\tilde{\mathbf{x}}|\mathbf{x})||_2^2\right] \tag{3}$$

where $\tilde{\mathbf{x}} \sim \mathcal{N}(\mathbf{x}, \sigma^2 I_d)$

Sliced Score Matching [Song et al., 2020]

$$J_{SSM}(\theta) \triangleq \mathbb{E}_{p_\mathbf{v}}\mathbb{E}_{p_{data}(\mathbf{x})}\left[||\mathbf{v}^T\nabla_{\mathbf{x}}\mathbf{s}_\theta(\mathbf{x})\mathbf{v} + \frac{1}{2}||\mathbf{s}_\theta(\mathbf{x})||_2^2\right] \tag{4}$$

where $p_\mathbf{v}$ is a distribution over random vectors (e.g., $\mathbf{v} \sim \mathcal{N}(0, I_d)$)

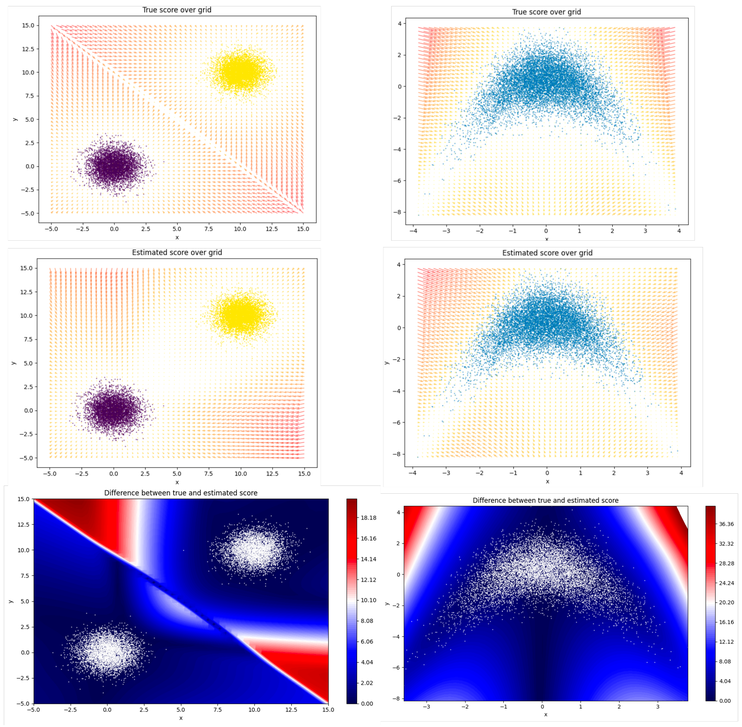

Figure 2. Score Estimation with Sliced Score Matching
Upper: Exact Score; **Middle:** Estimated Score; **Bottom:** Difference between vector fields.

Limitations

- Estimation through empirical mean does not penalize enough errors in low-density regions of the support (c.f. Figure 2)
- Data tend to be contained in low-dimensional manifolds

Noise Perturbation

Proposal: Corrupt the data with a Gaussian white noise

Trade-off on the level of noise to add:

More noise $\implies$ $\begin{cases} \text{data more scattered over the support: better score matching} \\ \text{data more altered: } p_{data}(\tilde{\mathbf{x}}) \text{ further from } p_{data}(\mathbf{x}) \end{cases}$

Solution: Construct a decreasing sequence of different noise levels

$$\{\sigma_i \in \mathbb{R}_+^*\}_{1\leq i\leq L} \text{ s.t. } \frac{\sigma_1}{\sigma_2} = \cdots = \frac{\sigma_{L-1}}{\sigma_L} > 1$$

L perturbed distributions: $\forall i \in \{1, \ldots, L\},\ p_{\sigma_i}(\tilde{\mathbf{x}}) = \int_{\mathbf{x}} p_{data}(\mathbf{x})\mathcal{N}(\tilde{\mathbf{x}}; \mathbf{x}, \sigma_i)d\mathbf{x}$

Unified Objective and Sampling Process

Noise-Conditional Score Network [Song and Ermon, 2019]:

Train a unique Score Network conditioned on the noise level

$$J_{NCSN}(\theta) \triangleq \frac{1}{L}\sum_{i=1}^{L}\lambda(\sigma_i)\mathbb{E}_{q_{\sigma_i}(\tilde{\mathbf{x}},\mathbf{x})}\left[||\mathbf{s}_\theta(\tilde{\mathbf{x}}, \sigma_i) + \frac{(\tilde{\mathbf{x}} - \mathbf{x})}{\sigma_i^2}||_2^2\right] \tag{5}$$

where $\lambda(\sigma_i) = \sigma_i^2$

Annealed Langevin Dynamics [Song and Ermon, 2019]:

- Sequential sampling of L different chains
- Use final states obtained for $p_{\sigma_i}(\tilde{\mathbf{x}})$ as initial states of the Langevin process for $p_{\sigma_{i+1}}(\tilde{\mathbf{x}})$
- Allows to better reconcile cluster weights in multi-modal distributions

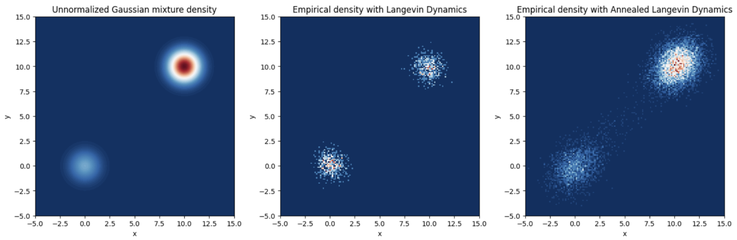

Figure 3. **Left**: Original Density; **Middle**: Estimated Density with Metropolis-Adjusted Langevin Algorithm using the exact score functions; **Right**: Estimated Density with Annealed Langevin algorithm using a NCSN.

References

Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

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

Yang Song, Sahaj Garg, Jiaxin Shi, and Stefano Ermon. Sliced score matching: A scalable approach to density and score estimation. In *Uncertainty in Artificial Intelligence*, pages 574–584. PMLR, 2020.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.