Reseach on Score-Based Generative Models

– 深度学习导论 Final project –

中国科学技术大学

侯相龙

2023 年 7 月 7 日

姓名：侯相龙

邮箱：howdragon@mail.ustc.edu.cn

学号：PB20010429

所在学院：数学科学学院

摘要

扩散模型以其能生成高质量样本得到的研究者的广泛关注。在本报告中，我主要阅读了 Y Song 的工作，深入了解 score-based generative model 的发展。从应用 score function 的原因，到两个基本的离散加噪的 score-based generative model，最后详细解释了带随机微分方程（SDE）的 diffusion model，并指出许多常见的扩散模型都是它的特例.

关键词: 生成模型、得分函数、随机微分方程、概率流 ODE、Langevin 动力学、扩散过程

目录

1 Introduction

在课程学习了扩散模型以及 DDPM 后，我阅读了 DDPM 的论文（Ho, Jain, and Abbeel 2020），但其中并没有详细介绍 score function 和 Langevin dynamics，于是我调研了 Y Song 与此相关的工作（Song and Ermon 2019），知道了这个思路的背景；另外，基于此我更加详细地阅读了 Y Song 在 2021 年的工作Song et al. 2021：Score-Based Generative Model（with Stochastic Differential Equations），他将扩散模型这种离散的多步去噪过程统一成了一个连续的随机微分方程（SDE）的特殊形式. 更加有趣的是，DDPM 的祖先采样过程等价于 diffusion with SDE 的离散化形式，并且，score-based model 的优化目标和 DDPM 的优化目标是等价的.

本文旨在介绍 Score-Based Generative Models 的发展，并详细分析了 Y Song 在 Score-Based Generative Modeling through Stochastic Differential Equations 中提出的基于随机微分方程（SDE）的生成模型.

在本文中，第二部分 (2) 介绍基础理论：得分函数（对数似然的梯度）和基于 Langevin 动力学的采样方法. 第三部分 (3) 介绍了两个基础的 Score-Based Generative Models：Naive score-based generative model 和 Score-based generative modeling with multiple noise perturbations. 第四部分 (4) 介绍了构建 Score-Based Generative Model(with SDE) 的方法. 第五部分 (5) 介绍 Score-Based Generative Model(with SDE) 的具体细节和应用场景，例如 SDE 与概率流 ODE 的联系，可控制生成等. 第六部分 (6) 总结了 Score-Based Generative Model(with SDE) 与 DDPM 的联系、模型优势以及笔者的阅读收获.

2 Basic concepts: score function and Langevin dynamics

一般而言，生成模型可大致分为两类：

- **likelihood-based models：** 这类模型通过极大似然估计，直接学习分布的概率密度。例如 autoregressive models(Larochelle and Murray 2011)、variational auto-encoders (VAEs)(Kingma and Welling 2013)
- **implicit generative models：** 这类模型的概率分布通过隐式的采样过程表示。例如 GANs(Goodfellow et al. 2014)

但是，对于 likelihood-based models 而言，为了能直接表示模型的概率密度，对概率模型的结构要求很高需要（因为需要计算与为止参数有关的归一化常数），这就导致了它的表达能力有限；对于 implicit generative models 而言，又常常需要对抗训练，这是不稳定的（Salimans et al. 2016）

为了解决这样的问题，提出了 score function(Liu, Lee, and Jordan 2016)，即对数似然的梯度，以及基于 Langevin dynamics 的采样方法生成新样本。这样就不需要计算归一化常数，也可以直接通过 score-matching 来学习.

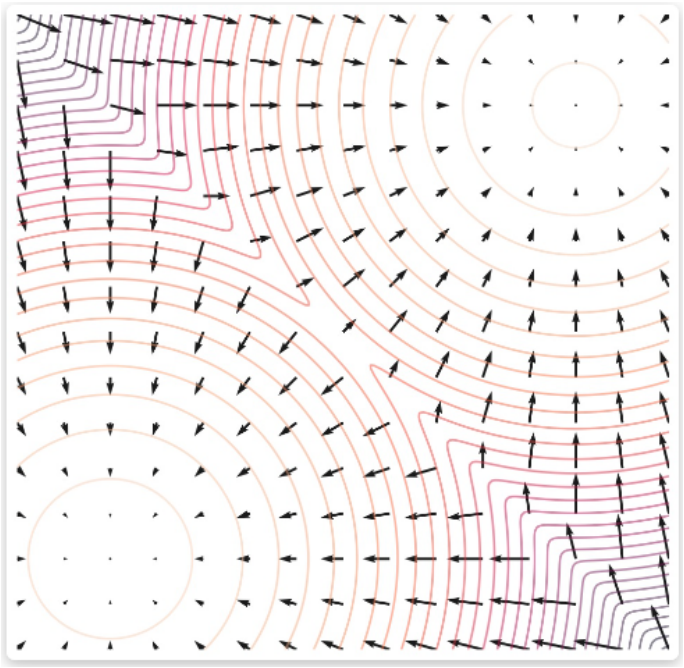

图 1: 2-Gauss 混合分布的 score function 示意图

2.1 score function

为了更好说明 score function 的优势，我们举一个例子：能量函数模型 (Song and Kingma 2021), 形如：

$$p_\theta(\mathbf{x}) = \frac{e^{-f_\theta(\mathbf{x})}}{Z_\theta} \tag{1}$$

其中，Z_θ 是与参数有关的归一化常量，使得 $\int p_\theta(\mathbf{x})\mathrm{d}\mathbf{x} = 1$。那么，我们在优化极大似然函数：

$$\max_\theta \sum_{i=1}^{N} \log p_\theta(\mathbf{x}_i) \tag{2}$$

不可避免地会涉及到对 Z_θ 求导操作或估计它的导数值，而这个运算代价较高。

而如果用 score function 处理这个问题的话 (Y Song：Song and Ermon 2019)，估计 $\mathbf{s}_\theta(\mathbf{x})$（可以用 NN），使得 $\mathbf{s}_\theta(\mathbf{x}) \approx \nabla_\mathbf{x} \log p(\mathbf{x})$。由于

$$\mathbf{s}_\theta(\mathbf{x}) \approx \nabla_\mathbf{x} \log p_\theta(\mathbf{x}) = -\nabla_\mathbf{x} f_\theta(\mathbf{x}) - \underbrace{\nabla_\mathbf{x} \log Z_\theta}_{=0} = -\nabla_\mathbf{x} f_\theta(\mathbf{x}) \tag{3}$$

并不会涉及对归一化常数的操作，这样我们就可以降低模型结构要求，增强模型表达能力。

另外，我们想直接最小化 Fisher divergence：

$$\mathbb{E}_{p(\mathbf{x})}[\|\nabla_\mathbf{x} \log p(\mathbf{x}) - \mathbf{s}_\theta(\mathbf{x})\|_2^2] \tag{4}$$

但是，我们仍然不知道数据的 score function: $\nabla_{\mathbf{x}} \log p(\mathbf{x})$，处理这个问题比较好的方法类为 score matching，例如可以用分部积分的方法可转化为优化 $\mathbb{E}_{p_{\text{data}}(\mathbf{x})}\left[\frac{1}{2}\|\mathbf{s}_\theta(\mathbf{x})\|_2^2 + \text{trace}(\underbrace{\nabla_{\mathbf{x}}\mathbf{s}_\theta(\mathbf{x})}_{\text{Jacobian of } \mathbf{s}_\theta(\mathbf{x})})\right]$ (Hyvärinen and Dayan 2005)，而这个期望可以简单的被样本上的平均估计. 所以我们也不需要进行对抗训练。

2.2 Langevin dynamics

在训练 $\mathbf{s}_\theta(\mathbf{x}) \approx \nabla_{\mathbf{x}} \log p(\mathbf{x})$ 后，我们可以用 Langevin dynamics 的采样方法生成新样本（即生成概率密度较大处的样本点）即：

$$\mathbf{x}_{i+1} \leftarrow \mathbf{x}_i + \epsilon \nabla_{\mathbf{x}} \log p(\mathbf{x}) + \sqrt{2\epsilon}\mathbf{z}_i, \quad i = 0, 1, \cdots, K, \tag{5}$$

其中，$\mathbf{z}_i \sim \mathcal{N}(0, I)$(为了避免坍缩到一个点，在（5）中加入了此随机项)。具体地，我们可以用估计值 $\mathbf{s}_\theta(\mathbf{x})$ 来代替 $\nabla_{\mathbf{x}} \log p(\mathbf{x})$ 带入（5）式进行计算。

3 Two basic score-based generative models

3.1 Naive score-based generative model

在 Y Song 之前的工作中 (Song and Ermon 2019), 基于前面提到的思想，有如下朴素的方法：

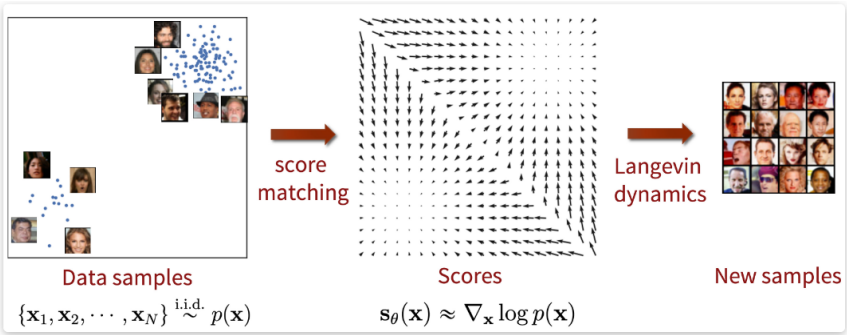

图 2: Naive score-based generative model

但是，这种朴素模型却又很大的局限性，注意到 Fisher divergence：

$$\mathbb{E}_{p(\mathbf{x})}[\|\nabla_{\mathbf{x}} \log p(\mathbf{x}) - \mathbf{s}_\theta(\mathbf{x})\|_2^2] = \int p(\mathbf{x})\|\nabla_{\mathbf{x}} \log p(\mathbf{x}) - \mathbf{s}_\theta(\mathbf{x})\|_2^2 \mathrm{d}\mathbf{x}. \tag{6}$$

那么当 $p(x)$ 很小时即低密度点，梯度的估计是不准确的，这样就会导致生成的结果的质量并不是很好，因为当数据维数较大时，我们并不能保证选取的初始迭代点为高密度点。

3.2 Score-based generative modeling with multiple noise perturbations

为了提高模型的性能，Y Song 提出了如下改进思路 (Song and Ermon 2020)：

对于朴素模型出现的问题，我们可以通过加 Gauss 噪声实现。当噪声足够大时，能够提高原来低密度点的密度，让低密度点的预测结果更加准确；但是我们又不能让噪声过大，以致于丢失原本分布的性质。为了实现这样的效果，我们可以采用 multiple noise perturbations，具体来说

首先，我们对原始数据加噪声：

$$p_{\sigma_i}(\mathbf{x}) = \int p(\mathbf{y})\mathcal{N}(\mathbf{x};\mathbf{y},\sigma_i^2 I)\mathrm{d}\mathbf{y}.$$

且满足 $\sigma_1 < \sigma_2 < \cdots < \sigma_L$。

接着对于每一个经过噪声扰动的数据，我们训练一个有噪声条件的 score-based model $\mathbf{s}_\theta(\mathbf{x}, i)$，即满足 $\mathbf{s}_\theta(\mathbf{x}, i) \approx \nabla_{\mathbf{x}} \log p_{\sigma_i}(\mathbf{x})$ for all $i = 1, 2, \cdots, L$.

最后，我们可按照 $i = L, L-1, \cdots, 1$ 的顺序，基于训练好的 $\mathbf{s}_\theta(\mathbf{x}, i)$，类似（5）式进行采样（annealed Langevin dynamics）. σ_i 单调递减保证了在初始时（低密度点）的样本移动与真实梯度方向相似，在接近高密度点后能够真实还原原始的真实分布.

4 Score-based generative model with SDEs

正如前面所提到的，多种噪声能使得模型效果取得不错的效果，那么一个自然的想法，当噪声规模趋于无穷的时候，是否会又更好的效果呢? 基于这个想法以及对扩散过程的思考，Y Song 提出了通过 SDEs 加噪声的方法 (Song et al. 2021)，实现了似然计算、可控制生成以及提供了提升 score-based model 的通用方法（如 DDPM）.

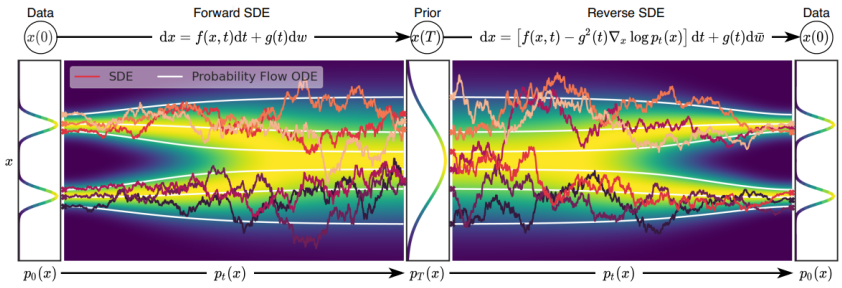

图 3: Overview of score-based generative modeling through SDEs

4.1 通过 SDE 对数据进行扰动

不同于前面进行离散的随机扰动，我们想让扰动形式为连续随机过程。而某些连续随机过程（尤其是扩散过程），通常是随机微分方程 (SDE) 的解。对于这个问题而言，SDE 有如下形式：

$$\mathrm{d}\mathbf{x} = \underbrace{\mathbf{f}(\mathbf{x}, t)}_{\text{drift coef}}\,\mathrm{d}t + \underbrace{g(t)}_{\text{diffusion coef}}\,\mathrm{d}\mathbf{w}, \tag{7}$$

其中，$f(\cdot, t) : \mathbb{R}^d \to \mathbb{R}^d$, $g(t) \in \mathbb{R}$(x free), $\mathbf{w}$ 为 Brown 运动. 这个 SDE 的解为随机变量列 $\{\mathbf{x}(t)\}_{t\in[0,T]}$.

令 $p_t(\mathbf{x})$ 为 $\mathbf{x}(t)$ 的概率分布。类别离散过程的加噪方法 (3.2)，我们知道：$p_0(\mathbf{x})$ 类似于 $p_{\sigma_0}(\mathbf{x})$，为数据的原始分布 $p(\mathbf{x})$；$p_T(\mathbf{x})$ 类似于 $p_{\sigma_L}(\mathbf{x})$，为充分加噪后的分布 $\pi(\mathbf{x})$, 可认为是先验分布。

特别地，我们考虑如下一维 SDE（Brwon 运动）：

$$\mathrm{d}\mathbf{x} = \sigma^2 \mathrm{d}\mathbf{w} \tag{8}$$

其对应的 Fokker–Planck 方程

$$\partial_t p - \frac{\sigma^2}{2}\partial_x^2 p = 0$$

有满足归一化的解：

$$p(t, x) = \frac{1}{\sigma\sqrt{2\pi t}} \exp\left(-\frac{x^2}{2\sigma^2 t}\right) \tag{9}$$

事实上，这个连续加噪过程与 (3.2) 中离散过程是非常相似的，噪声分布的方差随时间递增。

4.2 reverse SDE 生成样本

为了生成样本，我们需要考虑上面加噪的逆过程，即 SDE(7式) 对应的 reverse SDE：

$$\mathrm{d}\mathbf{x} = [\mathbf{f}(\mathbf{x}, t) - g^2(t)\nabla_{\mathbf{x}} \log p_t(\mathbf{x})]\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}. \tag{10}$$

其中 $\mathrm{d}t$ 是负的时间步长。从 (10式) 可以看出，我们只需要知道任意时间的 score-function：$\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$. 与前面（3）略有不同，我们需要训练依赖于时间的 score-function.

4.3 估计 score function

训练 time-dependent score function $\mathbf{s}_\theta(\mathbf{x}, t) \approx \nabla_{\mathbf{x}} \log p_t(\mathbf{x})$ 的目标函数为 Fisher divergence 的加权组合：

$$\mathbb{E}_{t\in\mathcal{U}(0,T)}\mathbb{E}_{p_t(\mathbf{x})}[\lambda(t)\|\nabla_{\mathbf{x}} \log p_t(\mathbf{x}) - \mathbf{s}_\theta(\mathbf{x}, t)\|_2^2], \tag{11}$$

同样的，利用 score matching 方法转化优化目标即可对此进行有效优化。

在训练完成 $\theta(\mathbf{x}, t)$，即可用类似 ODE 数值解的方法（Euler-Maruyama）求解逆向方程 (10)，具体来说令 $t = T$, 并选择足够小的步长 $\Delta t < 0$，迭代格式：

$$\begin{aligned}\Delta \mathbf{x} &\leftarrow [\mathbf{f}(\mathbf{x}, t) - g^2(t)\mathbf{s}_\theta(\mathbf{x}, t)]\Delta t + g(t)\sqrt{|\Delta t|}\mathbf{z}_t \\ \mathbf{x} &\leftarrow \mathbf{x} + \Delta \mathbf{x} \\ t &\leftarrow t + \Delta t,\end{aligned}$$

其中 z_t 为标准正态中采样得到的

5 Details about Score-based generative model with SDEs

5.1 概率流 ODE

为了能精确计算模型的精确对数似然，介绍了 SDE 对应的概率流 ODE。通过 Fokker–Planck 方程对原式变形，我们得到原 SDE 等价于一个具有确定过程的概率流 ODE：

$$\mathrm{d}\mathbf{x} = \left[\mathbf{f}(\mathbf{x}, t) - \frac{1}{2}g^2(t)\nabla_{\mathbf{x}} \log p_t(\mathbf{x})\right] \mathrm{d}t$$

它的分布同原 SDE 所确定的分布相同，且它的取样比 reverse SDE 更稳定，过程曲线更平滑。

事实上，概率流 ODE 是一种特殊的 neural ODE，我们可以通过 ODE 求解器直接求解。相比于前面的求解 reverse SDE 方法，求解概率流 ODE 的生成样本效率更高，并且可以得到精确的对数似然估计（即使我们只估计了 score function）

5.2 可控制生成

某些时候，我们需要生成的模型满足某些性质，比如产生特定标签的样本、图像补全等。假设控制信号为 y, 那么逆问题就变为求解在 $\mathbf{x} \mid \mathbf{y}$ 分布上采样。

注意到 Bayes 公式

$$\nabla_{\mathbf{x}} \log p(\mathbf{x} \mid \mathbf{y}) = \nabla_{\mathbf{x}} \log p(\mathbf{x}) + \nabla_{\mathbf{x}} \log p(\mathbf{y} \mid \mathbf{x})$$

其中，$\nabla \log p(\mathbf{x})$ 可以由前面提到的 $s_\theta(\mathbf{x})$ 近似，$p(\mathbf{y} \mid \mathbf{x})$ 可以通过训练新的神经网络或者先验知识确定。

5.3 结果

diffusion with SDE 在 CIFAR-10 上的效果，甚至优于 GAN 的最佳模型.

CIFAR-10 生成图片质量对比

Method	FID ↓	Inception score ↑
StyleGAN2 + ADA [38]	2.92	9.83
Ours [20]	**2.20**	**9.89**

CIFAR-10 上 log-likelihood 对比

Method	Negative log-likelihood (bits/dim) ↓
RealNVP	3.49
iResNet	3.45
Glow	3.35
FFJORD	3.40
Flow++	3.29
Ours	**2.99**

Probability flow ODE 生成的图片

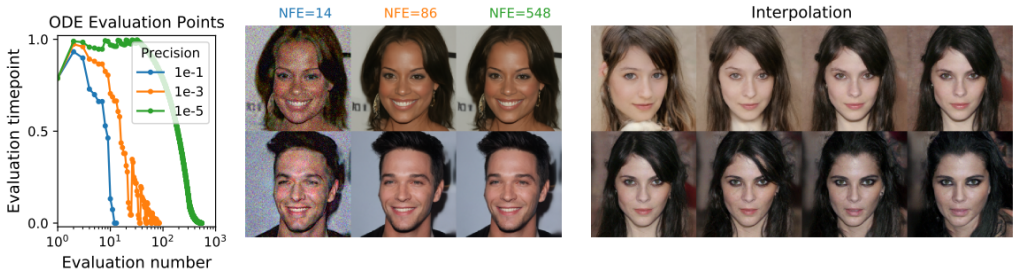

图 4: Probability flow ODE enables fast sampling with adaptive stepsizes as the numerical precision is varied (left), and reduces the number of score function evaluations (NFE) without harming quality (middle). The invertible mapping from latents to images allows for interpolations (right)

可控制生成

图 5: Left: Class-conditional samples on 32 ˆ 32 CIFAR-10. Top four rows are automobiles and bottom four rows are horses. Right: Inpainting (top two rows) and colorization (bottom two rows) results on 256 ˆ 256 LSUN. First column is the original image, second column is the masked/grayscale image, remaining columns are sampled image completions or colorizations

6 Conclusion

6.1 与 DDPM 的联系

课程中学习的 DDPM 实际上是 SDE 生成模型的特例. 下面具体说明:

扩散过程的 DDPM 满足:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I})$$

由 Taylor 展开，有:

$$\mathbf{x}_t \approx \mathbf{x}_{t-1} - \frac{\beta(t)\Delta t}{2}\mathbf{x}_{t-1} + \sqrt{\beta(t)\Delta t}\mathcal{N}(\mathbf{0}, \mathbf{I})$$

事实上，上式是离散化后的 SDE:

$$\mathrm{d}\mathbf{x}_t = -\frac{1}{2}\beta(t)\mathbf{x}_t\mathrm{d}t + \sqrt{\beta(t)}\mathrm{d}\boldsymbol{\omega}_t \tag{12}$$

(12) 对应的 reverse SDE 为:

$$\mathrm{d}\mathbf{x}_t = -\frac{1}{2}\beta(t)\left[\mathbf{x}_t + 2\nabla_{\mathbf{x}_t}\log q_t(\mathbf{x}_t)\right]\mathrm{d}t + \sqrt{\beta(t)}\mathrm{d}\bar{\boldsymbol{\omega}}_t \tag{13}$$

我们可以通过此式构建去噪过程.

(12) 对应的 Probabiliy Flow ODE 为

$$\mathrm{d}\mathbf{x}_t = -\frac{1}{2}\beta(t)\left[\mathbf{x}_t + \nabla_{\mathbf{x}_t}\log q_t(\mathbf{x}_t)\right]\mathrm{d}t \tag{14}$$

6.2 阅读收获

Y Song 关于 score-based model 做了一系列工作，最后用连续的 SDE 统一了扩散模型。我认为这个模型（理论）有如下价值：

- theoretical value：扩散过程本身便是一个连续的随机过程（Wiener process），而这个随机过程是可以用 SDE 描述的（Brown 运动方程）。这样去构建深度学习领域的模型有深刻的数学和物理基础。
- time-efficient：利用 fokker-planck 方程将 SDE 转化成了 ODE 之后，可以直接用 ODE 数值求解器求解，这避免了离散的扩散过程庞大的计算。
- high-quality sample: 扩散由离散向连续统一，将有限的加噪过程变为无限，提高了样本生成的质量

在读完 Y Song 有关 score-based model 的文章后，我更加深刻地体会到了扩散模型中的 score matching，Langevin dynamics, SDE and ODE 的广泛应用，学习到了科研中的基本研究思路和方法, 如何提升模型以及如何开展有价值的工作。

7 致谢

感谢在深度学习导论课程上王皓老师的精彩讲授，通过王皓老师的课程，我不仅学习到了传统的深度学习、神经网络的知识，也了解了比较前沿的工作和方法。另外，感谢两位助教在实验上的指导，我也在四次实验中提高了自己的代码实现能力。

References

Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. 2014. "Generative adversarial nets." *Advances in neural information processing systems* 27.

Ho, J., A. Jain, and P. Abbeel. 2020. "Denoising Diffusion Probabilistic Models." In *Advances in Neural Information Processing Systems,* edited by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, 33:6840–6851. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

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

Kingma, D. P., and M. Welling. 2013. "Auto-encoding variational bayes." *arXiv preprint arXiv:1312.6114*.

Larochelle, H., and I. Murray. 2011. "The neural autoregressive distribution estimator." In *Proceedings of the fourteenth international conference on artificial intelligence and statistics,* 29–37. JMLR Workshop and Conference Proceedings.

Liu, Q., J. D. Lee, and M. I. Jordan. 2016. *A Kernelized Stein Discrepancy for Goodness-of-fit Tests and Model Evaluation.* arXiv: 1602.03253 [stat.ML].

Salimans, T., I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. 2016. "Improved techniques for training gans." *Advances in neural information processing systems* 29.

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

———. 2020. "Improved techniques for training score-based generative models." *Advances in neural information processing systems* 33:12438–12448.

Song, Y., and D. P. Kingma. 2021. *How to Train Your Energy-Based Models.* arXiv: 2101.03288 [cs.LG].

Song, Y., J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. 2021. "Score-Based Generative Modeling through Stochastic Differential Equations." In *International Conference on Learning Representations.* https://openreview.net/forum?id=PxTIG12RRHS.