# Generative Modeling by Estimating Grad. of the Data Dist.

## 组会汇报 | 论文阅读报告

**何瑞杰**

2025-09-10

中山大学 · 数学学院

# Outline

# Outline

# 1. 论文名称

- Title: Generative Modeling by Estimating Gradients of the Data Distribution
- Authors: Yang Song, Stefano Ermon
- Conference: NeurIPS 2019
- Link: https://arxiv.org/abs/1907.05600
- Keywords: Score Matching, Generative Models, Langevin Dynamics

# Outline

## 2. Score Matching

我们有从一个未知分布 $p_{\text{data}}(\boldsymbol{x})$ 中采样得到的数据集 $\{\boldsymbol{x}_i \in \mathbb{R}^D\}_{i=1}^n$。我们要尝试估计该分布。一个自然的假设是

$$p_{\text{data}}(\boldsymbol{x}) = \frac{\exp(-f_\theta(\boldsymbol{x}))}{Z(\theta)}$$

其中 $f_\theta : \mathbb{R}^D \to \mathbb{R}$ 是某个函数，$Z(\theta)$ 是归一化因子。若不加其他考虑，直接处理 $p_{\text{data}}(\boldsymbol{x})$ 将会不可避免地遇到计算 $Z(\theta)$ 的困难。因此 Score matching 的一个核心思路是转而去估计分布的 Score function，其“几何直观”的意义是指向概率密度增加的方向:

$$s_\theta(\boldsymbol{x}) := \nabla_{\boldsymbol{x}} \log p_\theta(\boldsymbol{x}) = -\nabla_{\boldsymbol{x}} f_\theta(\boldsymbol{x}) - \underbrace{\nabla_{\boldsymbol{x}} \log Z(\theta)}_{=0} = -\nabla_{\boldsymbol{x}} f_\theta(\boldsymbol{x}).$$

## 2. Score Matching

自然地，我们有 Score matching 的原始目标函数:

$$J(\theta) := \frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}[\|s_\theta(\boldsymbol{x}) - \nabla_{\boldsymbol{x}} \log p_{\text{data}}(\boldsymbol{x})\|_2^2].$$

但由于我们不知道原始数据的分布，因此我们无法求得 $\nabla_{\boldsymbol{x}} \log p_{\text{data}}(\boldsymbol{x})$，需要做下面转化

$$\begin{aligned}
\frac{1}{2}\mathbb{E}_{p_{\text{data}}}[\|s_\theta(\boldsymbol{x}) - \nabla_{\boldsymbol{x}} \log p_{\text{data}}(\boldsymbol{x})\|_2^2] &= \frac{1}{2}\mathbb{E}_{p_{\text{data}}}[\|s_\theta(\boldsymbol{x})\|_2^2] + \cancel{\mathbb{E}_{p_{\text{data}}}[\|\nabla_{\boldsymbol{x}} \log p_{\text{data}}(\boldsymbol{x})\|_2^2]} \\
&\quad -\mathbb{E}_{p_{\text{data}}}[\langle s_\theta(\boldsymbol{x}), \nabla_{\boldsymbol{x}} \log p_{\text{data}}(\boldsymbol{x})\rangle] \\
&= \frac{1}{2}\mathbb{E}_{p_{\text{data}}}[\|s_\theta(\boldsymbol{x})\|_2^2] + \int p(\boldsymbol{x})\ \text{div}(s_\theta(\boldsymbol{x}))\mathrm{d}x \\
&= \mathbb{E}_{p_{\text{data}}}\left[\frac{1}{2}\ \|s_\theta(\boldsymbol{x})\|_2^2 + \text{tr}(\nabla_{\boldsymbol{x}} s_\theta(\boldsymbol{x}))\right]
\end{aligned}$$

## 2. Score Matching

然而上式中的 $\mathrm{tr}(\nabla_{\boldsymbol{x}} s_\theta(\boldsymbol{x}))$ 计算成本还是太高。庆幸我们可以对原数据增加 Gaussian 噪声，从而将其经过一个条件分布 $q_\sigma(\bar{\boldsymbol{x}}|\boldsymbol{x}) \sim N(0, \sigma^2 I)$ 得到加噪声后的数据 $\bar{\boldsymbol{x}}$。

经计算后我们能得到更加实际的目标函数:

$$J(\theta) = \frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}},}[\|s_\theta(\bar{\boldsymbol{x}}) - \nabla_{\bar{\boldsymbol{x}}} \log q_\sigma(\bar{\boldsymbol{x}}|\boldsymbol{x})\|_2^2]$$

注意此时 $\nabla_{\bar{\boldsymbol{x}}} \log q_\sigma(\bar{\boldsymbol{x}}|\boldsymbol{x})$ 还可以写为 $-\frac{1}{\sigma}\cdot\varepsilon$，其中 $\varepsilon$ 是从 $N(0, I)$ 中采样得到的噪声，原目标函数就变成

$$J(\theta) = \frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}},}\left[\left\|s_\theta(\bar{\boldsymbol{x}}) + \frac{1}{\sigma}\cdot\varepsilon\right\|_2^2\right]$$

## 2. Score Matching

换句话说，score function 在这个意义下正在预测加入到训练数据中的噪声。在实际操作中，我们需要权衡 $\varepsilon$ 的取值，如果太小，该方法起不到明显效果；如果太大，加噪声后的分布和原分布的区别大，难以学习原分布的特征。

# Outline

## 3. Langevin Dynamics

假如我们已经训练好 $s_\theta(\boldsymbol{x})$，我们可以按照自然地想法做下面的采样:

$$\boldsymbol{x}_{\{t+1\}} \leftarrow \boldsymbol{x}_t - \varepsilon s_\theta(\boldsymbol{x}_t)$$

但这一方法总会使得若干步后的采样结果收敛于原始分布之密度函数的若干极大值点，与生成模型的多样性目标不符。因此为解决这一问题，我们可以用 Langevin Dynamics 来采样。其与原来方法的区别在于在每一步中加入噪声，最后迭代得到的结果将会服从原始分布 $p_{\text{data}}(\boldsymbol{x})$:

$$\boldsymbol{x}_{\{t+1\}} \leftarrow \boldsymbol{x}_t - \frac{\varepsilon}{2} s_\theta(\boldsymbol{x}_t) + \sqrt{\varepsilon} z_t$$

其中 $z_t \sim N(0, I)$。

关于 Langevin 采样为何结果会收敛到原分布我尚不清楚，需要后续继续搜集资料学习。

# Outline

# 4. Score-based 生成的问题

[流形假设造成的问题]

我们所处世界中的高维数据往往分布在一个低维流形上。但上文中提到的对数据分布密度函数在低维流形的环绕空间 $\mathbb{R}^D$ 中求梯度是没有意义的。

[低概率密度区域的估计问题]

如果原分布是一个混合分布，且两个 “峰” 中间存在一个低概率密度的区域，模型学习时将难以获取该区域的信息，最后训练的结果在该区域的表现将会很差。

如果我们使用 Gauss 分布对原分布做扰动，则得到的新分布的支持集将会是整个环绕空间，而不是流形。另一方面，恰当强度的扰动也会使得低概率密度区域的概率密度增加，从而更容易采样到该区域中的点，为模型训练提供更多的信息。

# Outline

## 5. 方法

上文中提到，对数据做扰动时，大强度的扰动会使得训练变得简单，但扰动后的分布与原分布相差很大；小强度的扰动使得扰动后分布近似原分布，但会有诸如低概率密度区域训练点不足的问题。

文中提出一个整合两者有点的方法，即不考虑单个扰动强度 $\sigma$，而是考虑一个序列 $\{\sigma_i\}_{i=1}^n$ 其中 $\sigma_n$ 是一个足够小的数 (例如 0.01 )，$\sigma_1$ 是一个足够大的数 (例如 25 )。我们训练一个条件 score function $s_\theta(\boldsymbol{x}, \sigma)$ 预测不同扰动强度下的噪声方向。此时目标函数变为

$$\ell(\theta, \sigma) := \frac{1}{2}\mathbb{E}_{p_{\text{data}}(\boldsymbol{x}), \bar{\boldsymbol{x}} \sim N(\boldsymbol{x}, \sigma^2 I)}\left[\left\| s_\theta(\bar{\boldsymbol{x}}) + \frac{\bar{\boldsymbol{x}} - \boldsymbol{x}}{\sigma^2} \right\|\right]$$

$$\ell(\theta, \{\sigma_i\}_{i=1}^n) := \frac{1}{n}\sum_{i=1}^n \lambda(\sigma_i)\ell(\theta, \sigma_i)$$

其中 $\lambda(\sigma_i)$ 是权重函数，常取为 $\lambda(\sigma_i) = \sigma_i^2$，以平衡不同扰动强度下 score function 的范数。

## 5. 方法

在采样时，我们就做类似模拟退火的采样动作。首先选取最高的扰动强度 $\sigma_1$ 然后再该噪声强度下迭代若干次，然后选取次高的扰动强度 $\sigma_2$ 再该噪声强度下迭代若干次，以此类推。

这样就可以综合利用大扰动强度和小扰动强度的有点，从而更好的采样。在实际实验中，作者提出的方法在 CIFAR-10 数据集上取得了不错的效果。

# 5. 方法

**Algorithm 1: Anneled Langevin Dynamics**

**procedure** ANNELED LANGEVIN DYNAMICS($\{\sigma_i\}_{i=1}^n$, $\varepsilon$, $T$)
  ▷ Initialize the search range
  $\bar{\boldsymbol{x}}_0 \leftarrow \boldsymbol{v}$
  **for** $t \leftarrow 1, ..., L$ **do**
    ▷ Set step size $\alpha_i$
    $\alpha_i \leftarrow \varepsilon \cdot \sigma_i^2 / \sigma_L^2$
    **for** $t \leftarrow 1, ..., T$ **do**
      Draw $\boldsymbol{z}_t \sim N(0, I)$
      $\bar{\boldsymbol{x}}_t \leftarrow \bar{\boldsymbol{x}}_{t-1} + \alpha_i/2 \cdot s_\theta(\bar{\boldsymbol{x}}_{t-1}, \sigma_i) + \sqrt{\alpha_i}\boldsymbol{z}_t$
    **end**
    $\bar{\boldsymbol{x}}_0 \leftarrow \bar{\boldsymbol{x}}_T$
  **end**
  **return** $\bar{\boldsymbol{x}}_T$
**end**