BUILDING DIFFUSION MODEL'S THEORY FROM GROUND UP

Ayan Das *(University of Surrey, UK)*

I. INTRODUCTION

In this blogpost, we present a 'ground up' view of Diffusion Models, starting from its core element (i.e. the 'score'), building up to the modern form. While other models use a direct parametric function (with parameters θ) to transform noise into data distribution $q_{\text{data}}(x)$, i.e. $x = f_\theta(z)$, where $z \sim \mathcal{N}(0, I)$, diffusion model do so iteratively that involves a parametric "per-iteration" function (estimate of the true 'score')

$$x = g_1(g_2(g_3(...z..., s_\theta), s_\theta), s_\theta), \text{ where } z \sim \mathcal{N}(0, I). \quad (1)$$

The 'Score' of a distribution:

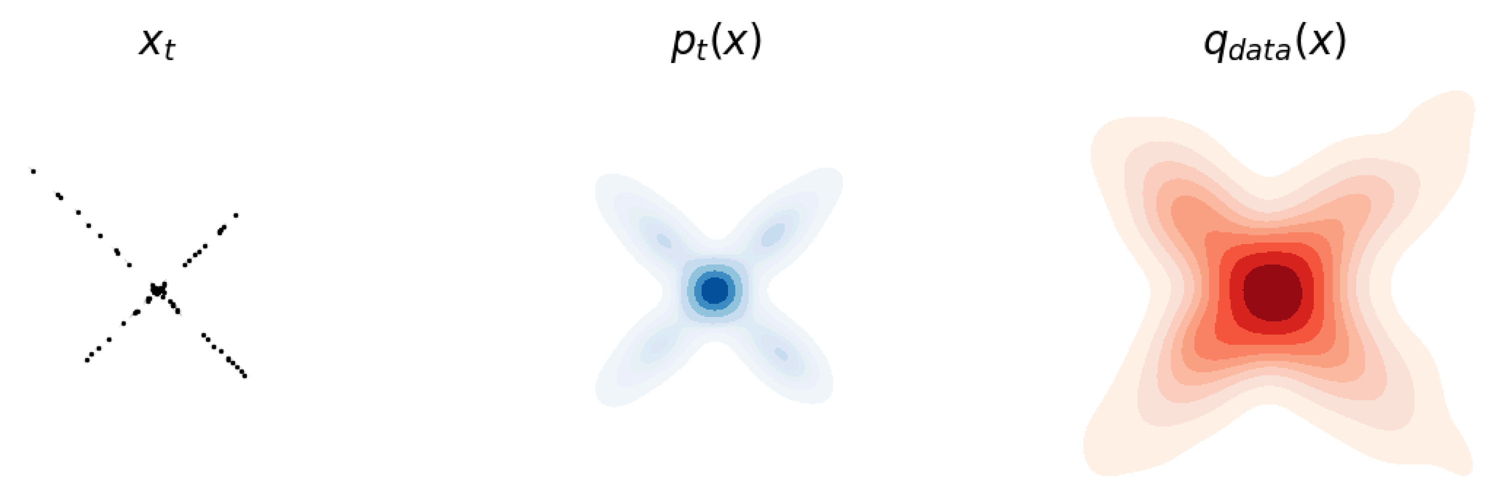
The 'score' of a distribution $q_{\text{data}}(x)$ is simply the gradient of the log-density, i.e. $\nabla_x \log q_{\text{data}}(x)$. This term was originally coined [1] long back in 1935 by Ronald Fisher in a slightly different context. But in machine learning, it is interpreted as a *guide* to go uphill in the log-density surface. An infinitesimal step in the direction of the score can get us to a state of higher likelihood

$$x_t' = x_t + \delta \cdot \nabla_x \log q_{\text{data}}(x) \quad (2)$$

II. GENERATIVE MODELLING WITH SCORE FUNCTION

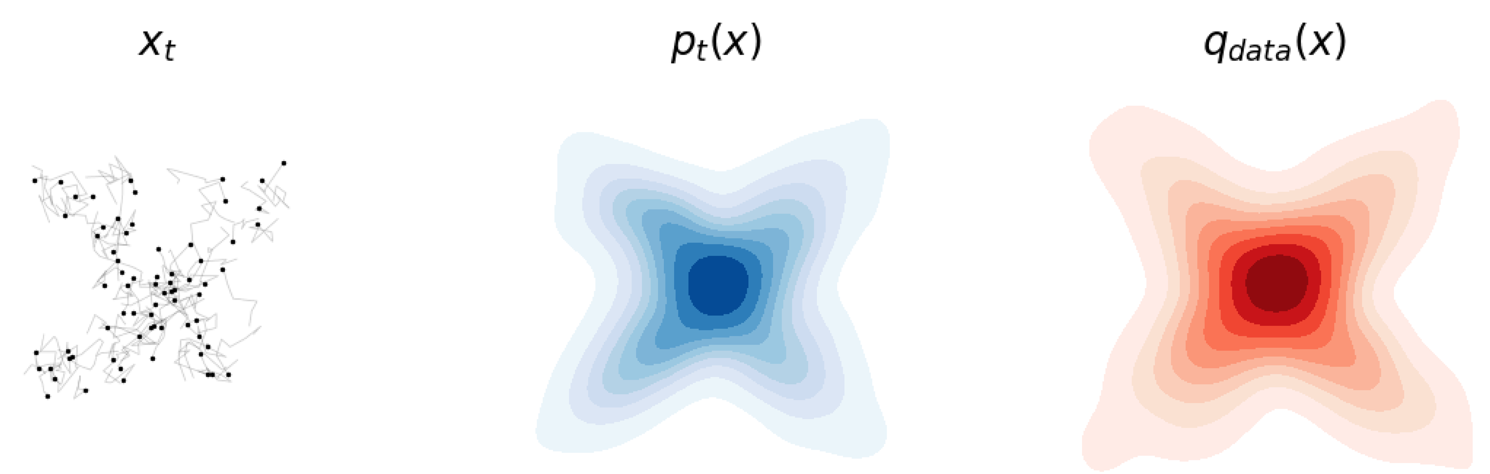
One can craft an iterative "sampling rule" solely based on the intutive interpretation of score provided in Equation 2

$$dx = \nabla_x \log q_{\text{data}}(x) \cdot dt \quad (3)$$



However, this process is NOT guarenteed to converge to the true distribution $q_{\text{data}}(x)$. Turns out that this problem has been studied [2] in particle physics long ago by Paul Langevin, in order to explain movement of particles suspended in fluid. Accoding to their theory, adding a little noise term fixes it.

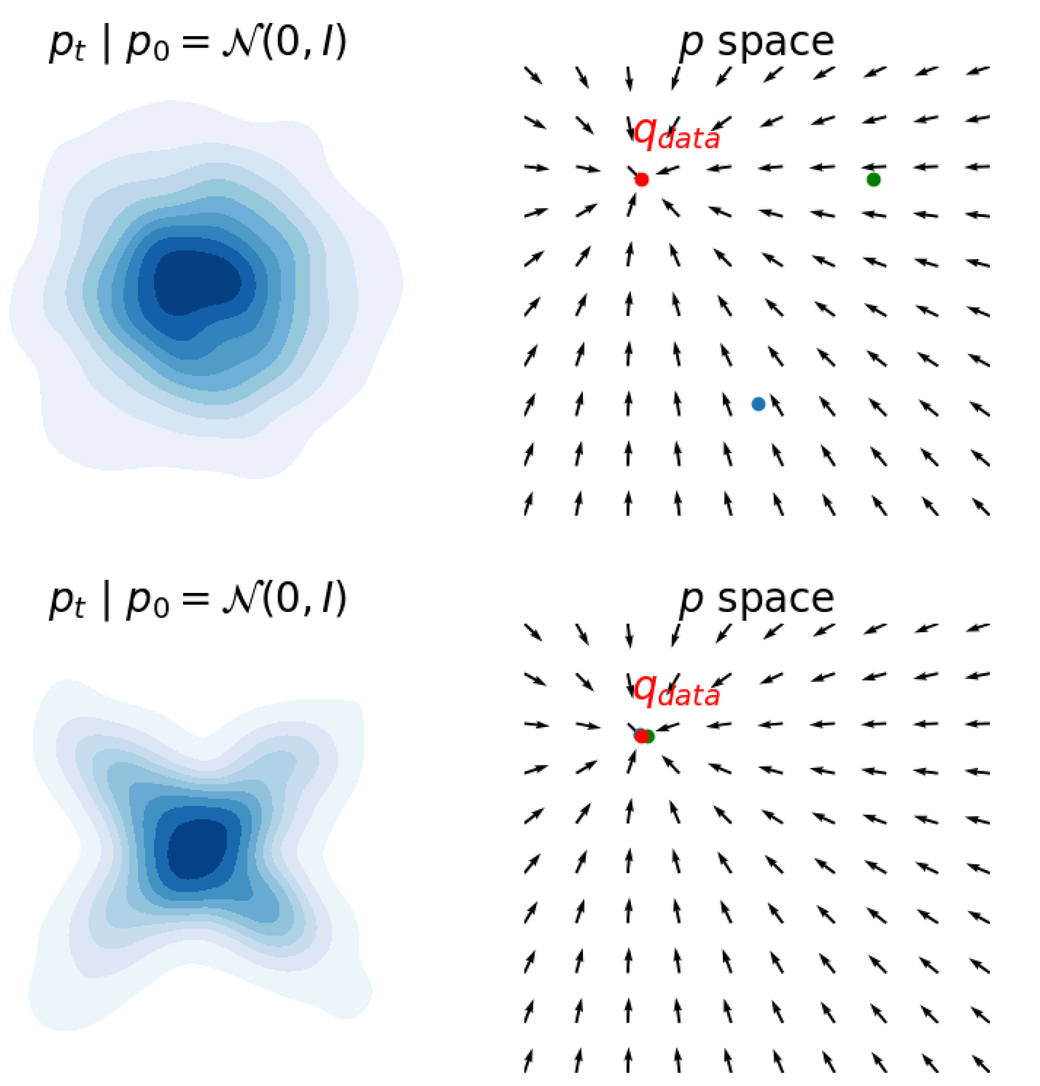
$$dx = \nabla_x \log q_{\text{data}}(x) \cdot dt + \sqrt{2} \cdot dB_t, \text{ where } dB_t = \mathcal{N}(0, dt) \quad (4)$$



Fokker-Planck Equation & a probability path:

The process in Equation 4 has a guarantee of convergence as $t \to \infty$. This can be validated by first noting $\mu_t(x) = \nabla_x \log q_{\text{data}}(x), \sigma_t(x) = \sqrt{2}$ and using the Fokker-Planck equation where we set $p_\infty(x) := q_{\text{data}}(x)$

$$\frac{\partial}{\partial t} p_\infty(x) = -\frac{\partial}{\partial x} p_\infty(x)\mu_t(x) + \frac{1}{2}\frac{\partial^2}{\partial x^2} p_\infty(x)\sigma_t^2(x) \quad (5)$$



The forward process & 'schedule':

It was argued [3] that learning an estimate of the true score everywhere on the data space x is extremely hard. The most popular solution to this problem is to learn a score *specialized* for t. To do so, one need samples from $p_t(x)$. The "forward process" is thus designed as an *ahead-of-time* description of the "path" taken by $p_t(x)$ on the probability space (refer to the above figure). One can revert the path by using the same Langevin Equation 4 but with end target being $\mathcal{N}(0, I)$ and starting from $q_{\text{data}}(x)$

$$dx = \nabla_x \log \mathcal{N}(0, I) \cdot dt + \sqrt{2} \cdot dB_t = -x \cdot dt + \sqrt{2} \cdot dB_t \quad (6)$$

To see its similarity to the 'modern' form of forward process, one must redefine the interpretation of time to contain it within a finite interval (e.g.$[0, 1]$). This is required due to the fact that langevin equation only guarantees convergence with $t \to \infty$. One can do so by first discretizing the above equation and plugging a new time mapping $t' = \mathcal{T}(t) = 1 - \exp(-t)$

$$x_{t'+dt'} = (1 - e^t dt')x_t + \sqrt{2e^t} dB_t \quad (7)$$

This resembles DDPM's [4] forward process where $e^t dt'$ is analogous to β_t in DDPM (small and increating in time t). We can then sample x_t for any t by simulating Equation 7.

$$x_t \sim q_t(x) \quad (8)$$

Now the score of a t-specialized density, i.e $\nabla_x \log q_t(x)$ must be learned.

III. ESTIMATING THE SCORE FUNCTION

The previous section deals with the generative modelling problem – given that we have access to the true score $\nabla_x \log q_{\text{data}}(x)$, which in reality, we don't. The very first credible solution to the score learning was proposed by [5], which turns the original Score Matching objective (not computable)

$$J(\theta) = \frac{1}{2}\mathbb{E}_{x \sim q_{\text{data}}(x)}[\| s_\theta(x) - \nabla_x \log q_{\text{data}}(x) \|^2] \quad (9)$$

.. to a practically computable version named "Implicit Score Matching (ISM)" which does not require the true score (unavailable)

$$J_I(\theta) = \mathbb{E}_{x \sim q_{\text{data}}(x)}\left[\frac{1}{2} \| s_\theta(x) \|^2 + \texttt{Tr}(\nabla_x s_\theta(x))\right]. \quad (10)$$

This objective was further upgraded by Vincent Pascal [6] as the "Denoising Score Matching", the variant still used in modern Diffusion Models

$$J_D(\theta) = \mathbb{E}_{x \sim q_{\text{data}}(x), \epsilon \sim \mathcal{N}(0,I)}\left[\frac{1}{2} \| s_\theta\left(\underbrace{x + \sigma\epsilon}_{\tilde{x}}\right) - \left(-\frac{\epsilon}{\sigma}\right) \|^2\right]. \quad (11)$$

A slight modification (reparameterization) of the denoising score matching loss above leads to the widely used variant called "noise estimation" where instead of score, we learn the noise direction from a noisy sample

$$J_\epsilon(\theta) = \mathbb{E}_{x \sim q_{\text{data}}(x), \epsilon \sim \mathcal{N}(0,I)}\left[\frac{1}{2\sigma^2} \| \epsilon_\theta(\tilde{x}) - \epsilon \|^2\right] \quad (12)$$

Yet another variant, named as "end point estimation" can be derived easily from the above equation, which predicts the clean sample from noisy one

$$J_x(\theta) = \mathbb{E}_{x \sim q_{\text{data}}(x), \epsilon \sim \mathcal{N}(0,I)}\left[\frac{1}{2\sigma^4} \| x_\theta(\tilde{x}) - x \|^2\right] \quad (13)$$

Connection to Tweedie's formula:

Equation 13 above has an interesting interpretation – as long as the noise is gaussian, it can be seen as learning posterior mean of clean quantity from noisy samples. In bayesian "inverse problem" literature, this is known as *Tweedie's formula* [7]

$$\mathbb{E}_{x \sim q(x \mid \tilde{x})}[x] = \tilde{x} + \sigma^2 \nabla_{\tilde{x}} \log p(\tilde{x}). \quad (14)$$

The similarity is obvious when we observe the fact that $x_\theta(\tilde{x}) = \tilde{x} + \sigma^2 s_\theta(\tilde{x})$.

BIBLIOGRAPHY

[1] R. A. Fisher, "The detection of linkage with "dominant" abnormalities," *Annals of Eugenics*, vol. 6, 1935.

[2] D. S. Lemons and A. Gythiel, "Paul langevin's 1908 paper "on the theory of brownian motion," *American Journal of Physics*, vol. 65, 1997.

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

[4] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," in *NeurIPS*, 2020.

[5] A. Hyvärinen, "Estimation of Non-Normalized Statistical Models by Score Matching," *Journal of Machine Learning Research*, 2005.

[6] P. Vincent, "A connection between score matching and denoising autoencoders," *Neural computation*, 2011.

[7] H. E. Robbins, "An empirical Bayes approach to statistics," *Breakthroughs in Statistics: Foundations and basic theory*. Springer, pp. 388–394, 1992.

 ayan05das@gmail.com | ayandas.me