A Primer on Score-based Diffusion Generative Modelling for Inverse Problems in Imaging

Riccardo Barbano & Alexander Denker

Computer Science Department, University College London
Mathematics Department, University of Bremen

A research topic at the intersections of Thermodynamics, Probabilistic Graphical Models, Bayesian Inference and Stochastic Differential Equations

Denoising Diffusion Probabilistic Models (DDPM) I

A diffusion probabilistic model is a parameterized Markov chain to produce samples matching the data after finite time [HJA20, SDWMG15]. This class of probabilistic models suddenly rose to prominence for the following reasons,

1. extreme flexibility in model construction
2. extreme flexibility in sampling and likelihood computation
3. unified framework (e.g., DDPM can be amalgamated into a wider framework of discretization of SDEs)
4. controllable and compositional generation [DDS+23]

Here we identify two processes: the forward process (or trajectory) and the reverse process.

Denoising Diffusion Probabilistic Models (DDPM) II

1. (forward process) seek to model a data distribution, labelled $q_{\mathrm{data}}(x)$, and for each training data-point, $x_0 \sim q_{\mathrm{data}}(x_0)$, we construct a series of latent variables (i.e., auxiliary variables of the same dimensionality of x_0), denoted as $\{x_t\}_{t=1}^{T}$ generated from a discrete Markov process, such that the approximate posterior is defined as a conditional joint distribution

$$q(x_1, \ldots, x_T | x_0) = \prod_{t=1}^{T} p(x_t | x_{t-1})$$

with each transition being defined as

$$p_{\beta_t}(x_t | x_{t-1}) = \mathcal{N}(x_t; (1 - \beta_t)^{1/2} x_{t-1}, \beta_t I_{d_x})$$

and with the unique property that all time marginals can be computed closed form

$$p_{\bar{\alpha}_t}(x_t | x_0) = \mathcal{N}(x_t; \bar{\alpha}_t^{1/2} x_0, (1 - \bar{\alpha}_t) I_{d_x})$$

Denoising Diffusion Probabilistic Models (DDPM) III

where $\alpha_t = 1 - \beta_t, \quad \bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. We refer to this as the forward diffusion process defined as a fixed discrete Markov chain that adds Gaussian noise according to a variance scheduler, i.e., $\beta_1, \ldots, \beta_T$, with $\forall \beta_t > 0$ taken to monotonically increase. The noise scales (or the diffusion rates) β_t are prescribed such that x_T (for large enough T) is approximately distributed according to $\mathcal{N}(x_T; 0, I_{d_x})$; thus, $0 < \beta_1, \beta_2, \ldots, \beta_T < 1$.

In sum, the diffusion process adds noise to the data until signal is destroyed.

2 (reverse process) takes the form of a Markov chain with learned transition, which are parametrised by θ

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \beta_t I_{d_x})$$

where

$$\mu_\theta(x_t, t) = (1 - \beta_t)^{-1/2}(x_t + \beta_t s_\theta(x_t, t))$$

but most importantly it reverses $q(x_t|x_{t-1})$ by learning a parametrised Markov chain in the reverse direction trained using variational inference. Diffusion models define a probabilistic generative process as the reverse of the noising process.

Denoising Diffusion Probabilistic Models (DDPM) IV

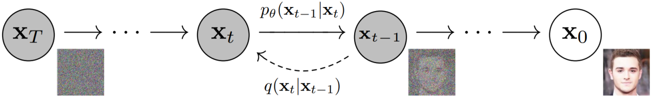

Figure: Ho et al. *Denoising Diffusion Probabilistic Models* (2020)[HJA20]

- $q(x_t|x_{t-1})$ is set such that $p(x_T) \approx \mathcal{N}(x_T; 0, I_{d_x})$
- $p_\theta(x_{t-1}|x_t)$ is learned to reverse the sampling

Denoising Diffusion Probabilistic Models (DDPM) V

Training is performed by optimising the diffusion objective (re-weighted ELBOs [KG23])

$$\mathcal{L}(\theta) = \mathbb{E}_{t\sim U}\mathbb{E}_{x_0\sim q}\mathbb{E}_{x_t|x_0\sim p_{\bar{\alpha}_t}} \|s_\theta(x_t, t) - \nabla_{x_t} \log p_{\bar{\alpha}_t}(x_t|x_0)\|_2^2$$

The objective described can be written as a re-weighted sum of denoising score matching objectives, which implies that the optimal model, s_{θ^*}, is the score of the perturbed data distribution [SSDK+20].

The training consists in the following steps,

1. $x_0 \sim q(x_0)$, $t \sim U(1, \ldots, T)$ and $\epsilon \sim \mathcal{N}(0, I_{d_x})$
2. $x_t = \bar{\alpha}_t^{1/2} x_0 + (1 - \bar{\alpha}_t)^{1/2}\epsilon$
3. Take gradient descent step, i.e., $\nabla_\theta \mathcal{L}(\theta)$
4. repeat until convergence

Denoising Diffusion Probabilistic Models (DDPM) VI

Once the model is learnt s_θ^*, samples can be generated by starting from x_T and [HJA20] follows the estimated reverse Markov chain using *ancestral sampling*

$$x_{t-1} = (1-\beta_t)^{-1/2}(x_t + \beta_t s_{\theta^*}(x_t, t)) + \beta_t^{1/2}\epsilon, \quad t = T, T-1, \cdots, 1, \epsilon \sim \mathcal{N}(0, I).$$

Denoising Score Matching with Langevin Dynamics (SMLD) I

Let's draw a connection between the two frameworks. Similarities are strikings!

1. [SSDK+20, SE19] construct a generative framework that uses a dataset to learn a model for generating new samples from $q_{\text{data}}(x)$. [SE19] trains a model parametrised by θ to learn the score of $q_{\text{data}}(x)$, i.e., $\nabla_x \log q_{\text{data}}(x)$.
2. [SE19] propose to train a Noise Conditional Score Network (NCSN), with a weighted sum of denoising score matching objectives [Vin11].
3. Similarly, [SE19] considers a sequences of noise scales $\sigma_{\min} = \sigma_1 < \sigma_2 < \cdots < \sigma_T = \sigma_{\max}$, such that $p_{\sigma_{\min}}(x) \approx q_{\text{data}}(x)$ and $p_{\sigma_{\max}}(x) \approx \mathcal{N}(x; 0, \sigma_{\max}^2 I_{d_x})$, and learn the optimal score-based model $s_{\theta^*}(x_t, \sigma_t)$, that matches $\nabla_x \log q_{\text{data}}(x)$ almost everywhere [SGSE20].
4. For sampling, [SSDK+20] runs annealed[1] Langevin MCMC.

[1] It goes from $\sigma_{\max}$ until $\sigma_{\min}$

A Unified Framework I

The goal is to construct a diffusion process $\{x_t\}_{t=0}^{T}$ indexed by a continuous time variable $t \in [0, T]$, such that $x_0 \sim p_0$ for which we have a dataset of i.i.d. samples, and $x_T \sim p_T$, for which we have a tractable form to generate samples efficiently,

- p_0 is the data distribution $q(x)$
- $p_T \approx \mathcal{N}(0, I_{d_x})$ is the prior distribution

1 This diffusion process can be modelled as the solution to an Ito SDE,

$$dx = f(x, t)dt + g(t)dw$$

where w[2] is the standard Wiener process (a.k.a., Brownian motion), $f(x, t)$ is a vector-valued function called the *drift coefficient* of $x(t)$, and $g(t)$[3] is a scalar function known as the *diffusion coefficient* of $x(t)$.

A Unified Framework II

2 [And82] states that the reverse of a diffusion process is also a diffusion process, running backwards in time and given by the reverse-time SDE,

$$dx = [f(x,t) - g(t)^2 \nabla_{x_t} \log p_t(x_t)]dt + g(t)d\bar{w}$$

where $\bar{w}$ is a standard Wiener process when time flows backwards from T to 0, and dt is an infinitesimal negative time-step.

3 [SSDK$^+$20] shows that the discrete Markov chain defined in DDPM as $T \to \infty$ converges to the following SDE,

$$dx = -\frac{1}{2}\beta(t)xdt + \beta(t)^{1/2}dw \tag{1}$$

In literature, the corresponding SDE model of DDPM goes under the name of Variance Preserving (VP) SDE.

[2] [Vin11] generalised to non-isotropic diffusion.

[3] [SSDK$^+$20] generalises to matrix-valued diffusion coefficients (ref. Appendix A).

Explore Diffusion Models on GoogleColab

References I

Brian DO Anderson, *Reverse-time diffusion equation models*, Stochastic Processes and their Applications **12** (1982), no. 3, 313–326.

Yilun Du, Conor Durkan, Robin Strudel, Joshua B Tenenbaum, Sander Dieleman, Rob Fergus, Jascha Sohl-Dickstein, Arnaud Doucet, and Will Grathwohl, *Reduce, reuse, recycle: Compositional generation with energy-based diffusion models and mcmc*, arXiv preprint arXiv:2302.11552 (2023).

Jonathan Ho, Ajay Jain, and Pieter Abbeel, *Denoising diffusion probabilistic models*, Advances in Neural Information Processing Systems **33** (2020), 6840–6851.

Diederik P Kingma and Ruiqi Gao, *Understanding the diffusion objective as a weighted integral of elbos*, arXiv preprint arXiv:2303.00848 (2023).

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli, *Deep unsupervised learning using nonequilibrium thermodynamics*, International Conference on Machine Learning, PMLR, 2015, pp. 2256–2265.

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

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

References II

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole, *Score-based generative modeling through stochastic differential equations*, arXiv preprint arXiv:2011.13456 (2020).

Pascal Vincent, *A connection between score matching and denoising autoencoders*, Neural computation **23** (2011), no. 7, 1661–1674.

UCL CENTRE FOR INVERSE PROBLEMS