

DDPM, DDIM and Tweedie's Formula.

References

1. Sohl-Dickstein et. al., Deep unsupervised learning using non-equilibrium thermodynamics, arXiv: 1503.03585, 2015.
2. Song & Ermon, Generative modeling by estimating gradients of the data dist., arXiv: 1907.05600, 2019.
3. Ho, Jain & Abbeel, Denoising diffusion probabilistic models arXiv: 2006.11239, 2020.
4. Song et. al., Score-based generative modeling through stochastic differential equations, arXiv 2011.13456, 2020.
5. Song et. al., Denoising diffusion implicit models, arXiv: 2010.02902, 2020.
6. B. Anderson, Reverse-time diffusion equation models, Stochastic Processes and their Applications, 1982.
7. Tweedie's Formula ~ 1950 ; B. Efron, Tweedie's formula and Selection Bias, 2011
8. Special thanks to Litu Rout for extensive discussions.

Motivation: Images can be represented as high-dimensional vector of pixels; a 512×512 image in RGB $\approx$ vector in 800,000 dimensions.

Tasks:

- * Generate a new image 'x'
- * Generate an image consistent with text 'z'
- * Denoise / increase resolution / deblur an image given measurement vector 'y'
-

Observation: These generation tasks can be interpreted as sampling from some (high dimensional) distribution or conditional sampling:

$$x \sim p_*(x), \quad x \in \mathbb{R}^d$$

$\hookrightarrow$ distribution over images

$$x \sim p_*(x|y), \quad y \in \mathbb{R}^m$$

$\hookrightarrow$ measurement

$$x \sim p_*(x|z), \quad z \in \mathbb{R}^l$$

$\hookrightarrow$ "text embedding"

Approach: Construct a Markov process whose stationary distribution is $P_*(\cdot)$. Initialize the MP with a random (Gaussian) initialization, and evolve the MP. Draw a sample after time T large enough, and this should be "close" to an image. (This is approach (a) discussed at the beginning of Note 8).

The DDPM (Denoising Diffusion Probabilistic Model) approach: Reverse a "forward" process that converts images to noise.

Forward Process: SDE that converts images into noise

Reverse Process: SDE that converts noise into images.

The reverse process is our sampler. The forward process is only used to construct the reverse process.

The forward process

The forward SDE is the standard OU process:

$$d\vec{X}_t = -\vec{X}_t dt + \sqrt{2} dB_t, \quad t \geq 0. \quad X_t \in \mathbb{R}^d$$

We initial $\vec{X}_0 \sim p_*(x)$ (data dist.)

Note that for this (or any other initial dist.), the density evolves to a standard Gaussian. Concretely, the above is a special case of the Langevin Dynamics, with steady-state dist. $\pi_*(x) = N(0, I_d)$

Furthermore, from LSI notes, we know that the density at time t , denoted by $\vec{p}_t(x)$ converges exponentially fast to $\pi_*(x)$, i.e.,

$$\vec{p}_0(x) = \underset{\text{data dist}}{p_*(x)} \dots \dots \dots \Rightarrow \vec{p}_\infty(x) = \pi_*(x) = N(0, I_d)$$

Note: We use the notation $\vec{(\cdot)}$ to denote random variables and densities associated with the forward process that evolves "forward" in time, i.e., t goes from 0 to ∞ .

This also differentiates from the reverse process that goes "backward" in time from some pre-selected $T > 0$ to $t = 0$. Variables associated with the reverse process are denoted by $\overleftarrow{(\cdot)}$.

The reverse process.

We now time-reverse the forward OU process. We fix $T > 0$, and consider the forward process:

$$d\vec{X}_t = -\vec{X}_t dt + \sqrt{2} dB_t, \quad X_0 \sim \vec{P}_0(\cdot), \quad t \in [0, T].$$

Recall from above that $\vec{X}_t \sim \vec{P}_t(\cdot)$, and

$$\vec{X}_t \approx N(0, I_d)$$

(See Notes 4, pp. 15 for a more careful characterization)

Our goal is to construct a stochastic process that runs backwards in time, i.e., $t \in [T, 0]$, s.t. $\overleftarrow{X}_T \sim \vec{P}_T$, i.e., $\overleftarrow{X}_T \stackrel{d}{=} \vec{X}_T$, and also $\overleftarrow{X}_0 \stackrel{d}{=} \vec{X}_0$.

In other words, this process is initialized with a sample from $\vec{P}_T(\cdot)$ (the terminal state of the forward OU process, which is roughly a standard Gaussian, for T large enough).

After running this new "reverse process", $\overleftarrow{X}_t, t \in [T, 0]$, at time '0', we will get a sample from $\vec{P}_0(\cdot)$, i.e., the distribution over images.

Let us define:

$$d\overleftarrow{X}_t = \left(-\overleftarrow{X}_t - 2\nabla_{\vec{x}} \ln \vec{p}_t(\vec{x}_t) \right) dt + \sqrt{2} d\tilde{B}_t,$$

$$t \in [T, 0], \quad \overleftarrow{X}_T \sim \vec{P}_T(\cdot).$$

Denote the pdf of this reverse process by $\overleftarrow{P}_t(\cdot)$, $t \in [\tau, 0]$. By construction, we have $\overleftarrow{P}_\tau(\cdot) = \overrightarrow{P}_\tau(\cdot)$.

Claim: $\forall t \in [0, \tau]$, $\overleftarrow{P}_t(\cdot) = \overrightarrow{P}_t(\cdot)$.

The above claim was derived in [3] (Anderson 1982). The proof of the claim below follows the discussion in [2]. (Anderson derives the reverse SDE for any general forward process $d\overrightarrow{X}_t = \mu(X_t)dt + \sigma(X_t)d\tilde{B}_t$, see [3] for details.)

Proof: Consider the reverse SDE:

$$d\overleftarrow{X}_t = \left(-\overleftarrow{X}_t - 2 \nabla \ln \overrightarrow{P}_t(\overleftarrow{X}_t) \right) dt + \sqrt{2} d\tilde{B}_t$$

$t \in [\tau, 0]$, $\overleftarrow{X}_\tau \sim \overrightarrow{P}_\tau(\cdot)$.

$t \rightarrow \tau - t$: (time direction reverse)

$$d\overleftarrow{X}_{\tau-t} = -\overleftarrow{X}_{\tau-t} - 2 \nabla \ln \overrightarrow{P}_{\tau-t}(\overleftarrow{X}_{\tau-t}) d(\tau-t) + \sqrt{2} d\tilde{B}_{\tau-t}.$$

Now $d(\tau-t) = -dt$ $\therefore (\tilde{B}_\tau - \tilde{B}_{\tau-t}) \stackrel{d}{=} (\hat{B}_t - \hat{B}_0)$

$$\therefore d\overleftarrow{X}_{\tau-t} = \left(\overleftarrow{X}_{\tau-t} + 2\nabla \ln \vec{P}_{\tau-t}(\overleftarrow{X}_{\tau-t}) \right) dt + \sqrt{2} d\hat{B}_t.$$

Let $Y_t = \overleftarrow{X}_{\tau-t}$

Then, $dY_t = (Y_t + 2\nabla \ln \vec{P}_{\tau-t}(Y_t))dt + \sqrt{2} d\hat{B}_t$,
 $t \in [0, \tau]$, $Y_0 \sim \vec{P}_\tau(\cdot)$.

Let $q_t(\cdot)$ be the pdf of Y_t , with $q_0(\cdot) = \vec{P}_\tau(\cdot)$.

The FP Equation for Y_t : $\left(\nabla_1(\cdot) = \frac{\partial}{\partial x}(\cdot) \right)$

$$\frac{\partial}{\partial t} q_t(y) = -\frac{\partial}{\partial y} \left((y + 2\frac{\partial \ln \vec{P}_{\tau-t}(y)}{\partial y}) q_t(y) \right) + \frac{\partial^2}{\partial y^2} q_t(y).$$

Let $\overleftarrow{p}_t(\cdot)$ be the pdf of $\overleftarrow{X}_t$. Then, by

construction, $\overleftarrow{p}_t(\cdot) = q_{T-t}(\cdot)$.

$\therefore$ The FP Equation for $\overleftarrow{p}_t(\cdot)$ is given by:

$$\begin{aligned} \frac{\partial}{\partial t} \overleftarrow{p}_{T-t}(x) = & -\frac{\partial}{\partial x} \left(\left(x + 2 \frac{\partial \ln \overrightarrow{p}_{T-t}(x)}{\partial x} \right) \cdot \overleftarrow{p}_{T-t}(x) \right) \\ & + \frac{\partial^2}{\partial x^2} \overleftarrow{p}_{T-t}(x). \end{aligned}$$

$(T-t) \rightarrow s :$

$$-\frac{\partial}{\partial s} \overleftarrow{p}_s(x) = -\frac{\partial}{\partial x} \left(\left(x + 2 \frac{\partial \ln \overrightarrow{p}_s(x)}{\partial x} \right) \overleftarrow{p}_s(x) \right) + \frac{\partial^2}{\partial x^2} \overleftarrow{p}_s(x)$$

i.e.,

$$\begin{aligned} \frac{\partial}{\partial s} \overleftarrow{p}_s(x) = & \frac{\partial}{\partial x} (x \overleftarrow{p}_s(x)) + \frac{\partial}{\partial x} \left(\left(2 \cdot \frac{1}{\overrightarrow{p}_s(x)} \frac{\partial \overrightarrow{p}_s(x)}{\partial x} \right) \overleftarrow{p}_s(x) \right) \\ & - \frac{\partial^2}{\partial x^2} \overleftarrow{p}_s(x) \end{aligned} \quad \longrightarrow \textcircled{3}$$

We now claim that $\overleftarrow{p}_s(x) = \overrightarrow{p}_s(x)$. To show this, we substitute $\overrightarrow{p}_s(x) = \overleftarrow{p}_s(x)$ in $\textcircled{3}$ above:

$$\frac{\partial}{\partial s} \overleftarrow{p}_s(x) = \frac{\partial}{\partial x} (x \overleftarrow{p}_s(x)) + 2 \cdot \frac{\partial^2}{\partial x^2} \overleftarrow{p}_s(x) - \frac{\partial^2}{\partial x^2} \overleftarrow{p}_s(x)$$

i.e.,

$$\frac{\partial}{\partial s} \overleftarrow{p}_s(x) = \frac{\partial}{\partial x} (x \overleftarrow{p}_s(x)) + \frac{\partial^2}{\partial x^2} \overleftarrow{p}_s(x) \rightarrow \textcircled{4}$$

Recall the FP Equation for the standard OU Process $\vec{X}(t)$:

$$\frac{\partial}{\partial t} \vec{p}_t(x) = \frac{\partial}{\partial x} (x \vec{p}_t(x)) + \frac{\partial^2}{\partial x^2} \vec{p}_t(x), \rightarrow \textcircled{5}$$

Observe that $\textcircled{4} = \textcircled{5}$. From uniqueness of solution of the FP Equation, we conclude that our claim that $\overleftarrow{p}_s(x) = \vec{p}_s(x)$ is correct. $\square$

Note: $\nabla \ln \vec{p}_t(x)$ is called the **score function** in the generative modeling literature. Also, see (Song & Ermon, 2019) for applying this approach for image generation.

Tweedie's Formula to compute $\nabla \ln p_t(x)$

The reverse diffusion starts with a Gaussian sample at $t=T$, and approximates a sample from the image distribution $p_0(\cdot)$ at $t=0$. This evolution, however, requires access to the score function $\nabla \ln p_t(\cdot)$.

There are several approaches to learning this function. One line of work (Hyvarinen 2005) develops an approach to learn this function by matching to minimize a Fisher information metric. This requires estimates of higher order derivatives, and in sliced score matching (Song et. al. 2019) develop an approach using projections to get around higher order derivatives.

The most popular approach today is based on denoising score matching, proposed by (Vincent 2010). This technique goes back to work by (Tweedie 1950) as discussed by (Efron 2011).

Outline of this approach:

- ① Suppose we have access to a large collection of samples of images. Denote the r.v. corresponding to an image by $X_0 \sim p_0(\cdot)$.

We "noise" the image by adding '0' mean Gaussian noise, with appropriate variance σ^2 s.t.

$$X_\sigma = X_0 + \sigma Z, \quad Z \sim N(0, I),$$

Z indep. of X_0 .

and X_σ has the marginal pdf given by the forward OU process

$$\min_f E[(X_0 - f(X_\sigma))^2]$$

- ② We now "denoise" the image using an MMSE estimator:

$$\hat{X}_0 = E[X_0 | X_\sigma = x_\sigma]$$

$$= g_\theta(x_\sigma)$$

parameters
of a neural network

(learned
function).

This estimator is learned using squared-loss regression:

$$\min_{g_\theta(\cdot)} E[(x_0 - g_\theta(x_\sigma))^2]$$

If $g_\theta(\cdot)$ is a large enough class of functions, we know that

$$g_\theta(x_\sigma) = E[x_0 | x_\sigma = x_\sigma].$$

The optimization above is solved using a large number of (image, noised image) pairs, i.e., (x_0, x_σ) pairs.

③ **Tweedie's Formula:** This goes back to 1950s, and states: Suppose $x_\sigma = x_0 + \sigma Z$ as above, then, $\xrightarrow{\sim p_0}$ $\xrightarrow{\sim N(0,1)}$

$$E[x_0 | x_\sigma = x_\sigma] = x_\sigma + \sigma^2 \nabla \ln p_\sigma(x_\sigma).$$

Thus, by learning the LHS, we effectively learn a score function $\nabla \ln p_t(\cdot)$.

(The actual implementations differ from above pipeline, but is based on the same idea).

Proof of Tweedie's formula (for scalar case; $\sigma=1$)

Suppose $Y = X + Z$, $X \sim p(\cdot)$, $X \perp\!\!\!\perp Z$
 $Z \sim N(0,1)$ $\downarrow$
indep

$$Y \sim q(\cdot) = p(x) * \left(\frac{1}{\sqrt{2\pi}} e^{-z^2/2} \right)$$

$$E[X|Y=y] = \int_{-\infty}^{\infty} f_{X|Y}(x|y) \cdot x \, dx$$

$$= \int_{-\infty}^{\infty} \frac{f_{X,Y}(x,y) \cdot x \, dx}{f_Y(y)}.$$

$$= \frac{1}{f_Y(y)} \int_{-\infty}^{\infty} x f_{Y|X}(y|x) \cdot f_X(x) dx$$

Now, $f_Y(y) = q(y)$, $f_X(x) = p(x)$

$$f_{Y|X}(y|x) = \frac{1}{\sqrt{2\pi}} e^{-(y-x)^2/2}$$

$$\therefore E[X|Y=y] = \frac{1}{q(y)} \int_{-\infty}^{\infty} x \cdot p(x) \cdot \frac{e^{-(y-x)^2/2}}{\sqrt{2\pi}} dx$$

→ (1)

Next, we have:

$$q(y) = p(y) \cdot \frac{1}{\sqrt{2\pi}} e^{-y^2/2}$$

$$= \int_{-\infty}^{\infty} p(x) \cdot \frac{1}{\sqrt{2\pi}} e^{-(y-x)^2/2} dx$$

$$\therefore \frac{\partial}{\partial y} q(y) = \int_{-\infty}^{\infty} p(x) \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-x)^2}{2}} \cdot \left(-\frac{2(y-x)}{2}\right) dx$$

$$= -y \int_{-\infty}^{\infty} \underbrace{p(x)}_{f_x(x)} \cdot \underbrace{\frac{1}{\sqrt{2\pi}} e^{-\frac{(y-x)^2}{2}}}_{f_{y|x}(y|x)} dx$$

$$+ \int_{-\infty}^{\infty} x p(x) \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-x)^2}{2}} dx$$

$$q(y) \cdot E[x|y=y]$$

from (1')

$$= -y \int_{-\infty}^{\infty} f_{x,y}(x,y) dx + q(y) E[x|y=y]$$

$$= -y q(y) + q(y) E[x|y=y]$$

$$\therefore \frac{\partial}{\partial y} q(y) = -y q(y) + q(y) E[x|y=y]$$

→ (2')

$$\text{i.e., } E[X|Y=y] = y + \frac{1}{q(y)} \cdot \frac{\partial}{\partial y} q(y) \\ = y + \frac{\partial}{\partial y} \ln q(y). \quad \square$$

Generative Flow Models. (DDIM)

Refs: (a) Song et. al., Score-based generative modeling through stochastic differential equations, arXiv 2011.13456, 2020.

(b) Song et. al., Denoising diffusion implicit models, arXiv: 2010.02502, 2020.

High level idea so far: Construct an SDE running backward in time to sample from the distribution over images.

This section: Construct an ODE, initialized with a RANDOM initial condition, so that the terminal state is an image. We do this through a 2-step process:

① Construct a time-dependent vector field $\vec{V}(x, t)$, $x \in \mathbb{R}^d$, such that:

$$d\vec{Z}_t = \vec{V}(\vec{Z}_t, t) dt, \quad \vec{Z}_0 \sim p_0(\cdot).$$

i.e., sample an image $\vec{Z}_0$ from $p_0(\cdot)$, and evolve the deterministic dynamic forward in time.

The goal is to construct $\vec{V}(\cdot, \cdot)$ such that $\vec{Z}_\infty \rightarrow N(0, I)$. (i.e., the ODE flows to a sample of a Gaussian r.v.; note that because the IC is a random variable, $\vec{Z}_t$ is a r.v. $\forall t \geq 0$).

② Reverse the Field: Construct a new reverse vector field $\overleftarrow{V}(z, t)$ such that it is initialized as: $\overleftarrow{Z}_0 = \vec{Z}_T$. This ODE also runs forwards in time, and the field $\overleftarrow{V}(\cdot, \cdot)$ is constructed such that $\overleftarrow{Z}_T = \vec{Z}_0$, i.e., you sample an image.

Note: This approach is much faster to implement by using fast ODE solvers.

Key Idea: Interpreting an SDE density evolution as an ODE density evolution:

Consider: $dx_t = \mu(x_t, t)dt + \sigma(x_t, t)dB_t$,
 $t \in [0, T]$, $x_t \in \mathbb{R}$ (scalar case for simple notation)

The FP Equation is: ($P_t(x)$ is the density)

$$\frac{\partial}{\partial t} P_t(x) = -\frac{\partial}{\partial x} (\mu(x, t) P_t(x)) + \frac{1}{2} \frac{\partial^2}{\partial x^2} (\sigma^2(x, t) P_t(x)).$$

Observe:

$$\frac{1}{2} \frac{\partial^2}{\partial x^2} (\sigma^2 P_t) = \frac{1}{2} \frac{\partial}{\partial x} \left(\frac{\partial}{\partial x} \sigma^2 P_t \right)$$

$$= \frac{1}{2} \frac{\partial}{\partial x} \left(\left(\frac{\partial}{\partial x} \sigma^2 \right) P_t + \sigma^2 \cdot \frac{\partial}{\partial x} P_t \right)$$

$$\left(\because \frac{\partial}{\partial x} \ln P_t = \frac{1}{P_t} \frac{\partial}{\partial x} P_t \right) \\ = \frac{1}{2} \frac{\partial}{\partial x} \left(\left(\frac{\partial}{\partial x} \sigma^2 \right) P_t + \sigma^2 \cdot \left(\frac{\partial}{\partial x} \ln P_t \right) P_t \right)$$

We can now absorb the second term in the FP Equation into the first.

$$\frac{\partial}{\partial t} P_t(x) = - \frac{\partial}{\partial x} \left(\underbrace{\left(\mu(x,t) - \frac{1}{2} \frac{\partial}{\partial x} \sigma^2(x,t) - \frac{\sigma^2(x,t)}{2} \frac{\partial}{\partial x} \ln P_t(x) \right)}_{\vec{V}(x,t)} P_t(x) \right)$$

$$= - \frac{\partial}{\partial x} \left(\vec{V}(x,t) P_t(x) \right).$$

This is the density flow equation for an ODE in a vector field $\vec{V}(x,t)$, with initial condition $P_0(x)$.

(Why: In the FP equation, set $\sigma(\cdot) = 0$, and we get the continuity equation, which describes the density flow for an ODE; see page 8).

① Using reasoning as above, define

$$\vec{V}(x, t) = -(x + \nabla \ln p_t(x)) , \quad x \in \mathbb{R}^d.$$

where $p_t(\cdot)$ is the pdf of the OU process:

$$d\vec{X}_t = -\vec{X}_t dt + \sqrt{2} dB_t , \quad X_0 \sim p_0(\cdot).$$

Then, consider the ODE

$$d\vec{Z}_t = \vec{V}(\vec{Z}_t, t) dt , \quad Z_0 \sim p_0(\cdot).$$

Then, $\vec{Z}_t \stackrel{d}{=} \vec{X}_t$.

i.e., the ODE has the same marginal as the SDE.

② Reversing the Field

Define $\overleftarrow{V}(x, t) = -\vec{V}(x, T-t)$, $t \in [0, T]$
as the reverse vector field, and let

$$d\overleftarrow{Z}_t = \overleftarrow{V}(\overleftarrow{Z}_t, t) dt, \quad t \in [0, \tau]$$

be the reverse field ODE.

Claim: Suppose $\overleftarrow{Z}_0 = \overrightarrow{Z}_\tau$ (initialization of reverse field)

Then, $\overleftarrow{Z}_\tau = \overrightarrow{Z}_0$. In words, this ODE flow samples from the image distribution.

Pf: $\frac{d\overleftarrow{Z}_t}{dt} = \overleftarrow{V}(\overleftarrow{Z}_t, t) = -\overrightarrow{V}(\overleftarrow{Z}_t, \tau - t).$

Change: $\tau - t \rightarrow s$ $\frac{d\overleftarrow{Z}_{\tau-s}}{ds} = \overrightarrow{V}(\overleftarrow{Z}_{\tau-s}, s)$

Also, $\frac{d\overrightarrow{Z}_s}{ds} = \overrightarrow{V}(\overrightarrow{Z}_s, s)$

i.e., $\overleftarrow{Z}_{\tau-s} = \overrightarrow{Z}_s \quad \forall s. \quad \square$