TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2023

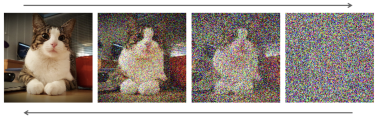
The Mathematics of Diffusion Models

McAllester, arXiv 2023

Denoising Diffusion Probabilistic Models (DDPM)

Ho, Jain and Abbeel, June 2020

The Diffusion SDE



Consider a discrete time process $z(0), z(\Delta t), z(2\Delta t), z(3\Delta t), \ldots$

$$z(0) \;=\; y, \quad y \sim \mathrm{Pop}(y)$$

$$z(t+\Delta t) \;=\; z(t) + \sqrt{\Delta t}\,\epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$$

A sum of two Gaussians is a Gaussian whose **variance** is the sum of the two variances.

$$z(t+n\Delta t) = z(t) + \sqrt{n\Delta t}\,\epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$$

Here $\sqrt{n\Delta t}$ is the **standard deviation** of the added noise.

The Diffusion SDE

The stochastic differential equation is the limit as the discrete step size Δt goes to zero.

$$z(0) = y, \quad y \sim \text{Pop}(y)$$

$$z(t + \Delta t) = z(t) + \sqrt{\Delta t}\, \epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \tag{1}$$

In this limit equation (1) holds for all t and Δt.

The Markovian ELBO

For a Markovian VAE we have

$$
\begin{aligned}
-\ln \mathrm{Pop}(y) &= -\ln \frac{P_{\mathrm{enc}}(z_N) P_{\mathrm{enc}}(z_{N-1}|z_N) \cdots P_{\mathrm{enc}}(z_1|z_2) P_{\mathrm{enc}}(y|z_1)}{P_{\mathrm{enc}}(z_N|y) P_{\mathrm{enc}}(z_{N-1}|z_N, y) \cdots P_{\mathrm{enc}}(z_1|z_2, y)} \\
&= \begin{cases} KL(P_{\mathrm{enc}}(z_N|y),\ P_{\mathrm{enc}}(z_N)) \\ +\sum_{i=2}^{N}\ E_{\mathrm{enc}}\ KL(P_{\mathrm{enc}}(z_{i-1}|z_i, y),\ P_{\mathrm{enc}}(z_{i-1}|z_i)) \\ -E_{\mathrm{enc}}\ [\ln P_{\mathrm{enc}}(y|z_1)] \end{cases} \qquad (2) \\
H(y) &\leq \begin{cases} E_{\mathrm{enc}}\ KL(P_{\mathrm{enc}}(z_N|y),\ p_{\mathrm{gen}}(z_N)) \\ +\sum_{i=2}^{N}\ E_{\mathrm{enc}}\ KL(P_{\mathrm{enc}}(z_{i-1}|z_i, y),\ p_{\mathrm{gen}}(z_{i-1}|z_i)) \\ -\ln p_{\mathrm{gen}}(y|z_1) \end{cases} \qquad (3)
\end{aligned}
$$

Here we focus on (2) rather than (3). We will first derive exact expressions rather than inequalities.

Exact Equalities

$$-\ln \mathrm{Pop}(y) \;=\; \begin{cases} KL(P_{\mathrm{enc}}(z_N|y),\; P_{\mathrm{enc}}(z_N)) \\ +\sum_{i=2}^{N}\; E_{\mathrm{enc}}\; KL(P_{\mathrm{enc}}(z_{i-1}|z_i, y),\; P_{\mathrm{enc}}(z_{i-1}|z_i)) \\ -E_{\mathrm{enc}}\; [\ln P_{\mathrm{enc}}(y|z_1)] \end{cases}$$

Reverse-Time Probabilities

In the limit of small Δt it is possible to derive the following.

$$p(z(t-\Delta t)|z(t),y) = \mathcal{N}\left(z(t) + \frac{\Delta t(y - z(t))}{t},\ \ \Delta t I \right)$$

$$p(z(t-\Delta t)|z(t)) = \mathcal{N}\left(z(t) + \frac{\Delta t(E[y|t,z(t)] - z(t))}{t},\ \ \Delta t I \right)$$

The Reverse-Diffusion SDE

$$p(z(t-\Delta t)|z(t)) = \mathcal{N}\left(z(t) + \frac{\Delta t(E[y|t,z(t)]-z(t))}{t}, \Delta t I \right)$$

This equation defines a reverse-diffusion SDE which we can write as

$$z(t-\Delta t) = z(t) + \left(\frac{E[y|t,z(t)] - z(t)}{t} \right) \Delta t + \sqrt{\Delta t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$$

KL-Divergence

$$-\ln \mathrm{Pop}(y) \;=\; \begin{cases} KL(P_{\mathrm{enc}}(z_N|y),\; P_{\mathrm{enc}}(z_N)) \\ \\ +\sum_{i=2}^{N}\; E_{\mathrm{enc}}\; KL(P_{\mathrm{enc}}(z_{i-1}|z_i, y),\; P_{\mathrm{enc}}(z_{i-1}|z_i)) \\ \\ -E_{\mathrm{enc}}\; [\ln P_{\mathrm{enc}}(y|z_1)] \end{cases}$$

For two Gaussian distributions with the same isotropic covariance we have

$$KL\left(\begin{array}{c}\mathcal{N}(\mu_1, \sigma^2 I),\\ \mathcal{N}(\mu_2, \sigma^2 I)\end{array}\right) = \frac{||u_1 - \mu_2||^2}{2\sigma^2}$$

KL-Divergence

$$-\ln \mathrm{Pop}(y) = \begin{cases} KL(P_{\mathrm{enc}}(z_N|y),\ P_{\mathrm{enc}}(z_N)) \\ +\sum_{i=2}^{N}\ E_{\mathrm{enc}}\ KL(P_{\mathrm{enc}}(z_{i-1}|z_i, y),\ P_{\mathrm{enc}}(z_{i-1}|z_i)) \\ -E_{\mathrm{enc}}\ [\ln P_{\mathrm{enc}}(y|z_1)] \end{cases}$$

$$p(z(t-\Delta t)|z(t), y) = \mathcal{N}\left(\ z(t) + \frac{\Delta t(y - z(t))}{t},\ \ \Delta t I\ \right)$$

$$p(z(t-\Delta t)|z(t)) = \mathcal{N}\left(\ z(t) + \frac{\Delta t(E[y|t, z(t)] - z(t))}{t},\ \ \Delta t I\ \right)$$

KL-Divergences

$$p(z(t-\Delta t)|z(t),y) \;=\; \mathcal{N}\left(\; z(t) + \tfrac{\Delta t(y-z(t))}{t},\;\; \Delta t I \;\right)$$

$$p(z(t-\Delta t)|z(t)) \;=\; \mathcal{N}\left(\; z(t) + \tfrac{\Delta t(E[y|t,z(t)]-z(t))}{t},\;\; \Delta t I \;\right)$$

$$KL\left(\begin{array}{l} p(z(t-\Delta t)|z(t),y), \\ p(z(t-\Delta t)|z(t)) \end{array}\right) \;=\; \left(\frac{||y-E[y|t,z(t)]||^2 \Delta t^2}{2t^2 \Delta t}\right)$$

$$= \left(\frac{||y-E[y|t,z(t)]||^2}{2t^2}\right)\Delta t$$

KL-Divergences

$$
-\ln \mathrm{Pop}(y) = \begin{cases} KL(P_{\mathrm{enc}}(z_N|y),\ P_{\mathrm{enc}}(z_N)) \\ +\sum_{i=2}^{N}\ E_{\mathrm{enc}}\ KL(P_{\mathrm{enc}}(z_{i-1}|z_i, y),\ P_{\mathrm{enc}}(z_{i-1}|z_i)) \\ -E_{\mathrm{enc}}\ [\ln P_{\mathrm{enc}}(y|z_1)] \end{cases}
$$

$$
= \sum_{i=2}^{N} \left(\frac{||y - E[y|t, z(t)]||^2}{2t^2} \right) \Delta t + E_{\mathrm{enc}}\ [-\ln P_{\mathrm{enc}}(y|z_1)]
$$

$$
t = i\Delta t
$$

Passing to the Integral

$$-\ln \mathrm{pop}(y) = \begin{cases} \int_{t_0}^{\infty} dt \;\; E_{z(t)|y} \;\; \left[\frac{||y - E[y|t, z(t)]||^2}{2t^2} \right] \\ \\ + \; E_{z(t_0)|y}[-\ln p(y|z(t_0))] \end{cases}$$

$$H(y) = \begin{cases} \int_{t_0}^{\infty} dt \;\; E_{y,z(t_0)} \;\; \left[\frac{||y - E[y|t, z(t)]||^2}{2t^2} \right] \\ \\ + \; H(y|z(t_0)) \end{cases}$$

Mutual Information

$$H(y) = \begin{cases} \int_{t_0}^{\infty} dt \;\; E_{y,z(t_0)} \left[\frac{||y - E[y|t,z(t)]||^2}{2t^2} \right] \\ \\ + H(y|z(t_0)) \end{cases}$$

$$H(y) - H(y|z(t_0)) = \int_{t_0}^{\infty} dt \;\; E_{y,z(t_0)} \left[\frac{||y - E[y|t,z(t)]||^2}{2t^2} \right]$$

$$I(y,z(t_0)) = \int_{t_0}^{\infty} dt \;\; E_{y,z(t_0)} \left[\frac{||y - E[y|t,z(t)]||^2}{2t^2} \right]$$

This is the information minimum mean squared error relation (I-MMSE) relation [Guo et al. 2005].

Estimating $E[y|t, z(t)]$

$$z(t-\Delta t) = z(t) + \left(\frac{E[y|t, z(t)] - z(t)}{t}\right)\Delta t + \epsilon\sqrt{\Delta t}, \quad \epsilon \sim \mathcal{N}(0, I)$$

We can train a denoising network $\hat{y}(t, z)$ to estimate $E[y|t, z(t)]$ using

$$\hat{y}^* = \underset{\hat{y}}{\text{argmin}}\, E_t\; E\; (\hat{y}(t, z) - y)^2$$

Assuming universality

$$\hat{y}^*(t, z) = E[y|t, z(t)]$$

Estimating $E[y|t, z(t)]$

In practice it is better to train the denoising network on values of the same scale.

If the population values are scaled so as to have scale 1, then the scale of $z(t)$ is $\sqrt{1+t}$.

$$\hat{y}^* = \underset{\hat{y}}{\operatorname{argmin}}\, E_{t,z(t)}\, (\hat{y}(t, z/\sqrt{1+t}) - y)^2$$

We then have

$$E[y|t, z(t)] = \hat{y}^*(t, z/\sqrt{1+t}))$$

The Fokker-Plack Anaylysis (The Score Function)

For $\epsilon \sim \mathcal{N}(0, I)$ a general SDE can be written as

$$z(t + \Delta t) = z(t) + \mu(z(t), t)\Delta t + \sigma(z(t), t)\epsilon\sqrt{\Delta t}$$

$$dz = \mu(z(t), t)dt + \sigma(z(t), t)dB$$

The diffusion process is the special case of Brownian motion

$$z(t + \Delta t) = z(t) + \epsilon\sqrt{\Delta t}$$
$$dz = dB$$

The Fokker-Planck Equation

Let $P_t(z)$ be the probability that $z(t) = z$.

$$\frac{\partial P_t(z)}{\partial t} = -\nabla \cdot \begin{pmatrix} \mu(z(t), t) P_t(z) \\ \\ -\frac{1}{2}\sigma^2(z(t), t) \nabla_z P_t(z) \end{pmatrix}$$

For the special case of diffusion we have

$$\frac{\partial P_t(z)}{\partial t} = -\nabla \cdot \left(-\frac{1}{2} \nabla_z P_t(z) \right)$$

The Score Function

$$\frac{\partial P_t(z)}{\partial t} = -\nabla \cdot \begin{pmatrix} \mu(z(t),t) P_t(z) \\ \\ -\frac{1}{2}\sigma^2(z(t),t) \nabla_z P_t(z) \end{pmatrix}$$

$$\frac{\partial P_t(z)}{\partial t} = -\nabla \cdot \left(-\frac{1}{2} \nabla_z P_t(z) \right)$$

$$\frac{\partial P_t(z)}{\partial t} = -\nabla \cdot \left[\left(-\frac{1}{2} \nabla_z \ln P_t(z) \right) P_t(z) \right]$$

The Score Function

$$\frac{\partial P_t(z)}{\partial t} = -\nabla \cdot \left[\left(-\frac{1}{2} \nabla_z \ln P_t(z) \right) P_t(z) \right]$$

$\ln P_t(z)$ is the score function.

The time evolution of $P_t(z)$ can be written as the result of **deterministic** flow given by

$$\frac{dz}{dt} = -\frac{1}{2} \nabla_z \ln p_t(z)$$

Deterministic Denoising

Following the deterministic flow backward in time samples from the population!

$$z(t - \Delta t) = z(t) + \frac{1}{2}\nabla_z \ln p_t(z)\Delta t$$

No noise!

Solving for the Score Function

$$\begin{aligned} P_t(z) &= E_y\, P_t(z|y) \\ &= E_y\, \frac{1}{Z(t)} e^{-\frac{||z-y||^2}{2t}} \\ \nabla_z P_t(z) &= E_y\, P_t(z|y)\, (y-z)/t \\ &\vdots \\ &= P_t(z) \frac{E[y|t,z]-z}{t} \end{aligned}$$

$$\nabla_z \ln P_t(z) = \frac{E[y|t,z]-z}{t}$$

This is Tweedie's formula, Robbins 1956.

Stochastic vs. Deterministic Denoising

$$z(t-\Delta t) = z(t) + \left(\frac{E[y|t,z(t)] - z(t)}{t}\right)\Delta t + \epsilon\sqrt{\Delta t}$$

$$z(t-\Delta t) = z(t) + \frac{1}{2}\left(\frac{E[y|t,z(t)] - z(t)}{t}\right)\Delta t$$

Interpolating Stochastic and Deterministic

One can show that for $\lambda \in [0, 1]$ the following also samples from the population.

$$z(t - \Delta t) = z(t) + \frac{1 + \lambda}{2} \left(\frac{E[y|t, z(t)] - z(t)}{t} \right) \Delta t + \lambda \epsilon \sqrt{\Delta t}$$

END