TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2023

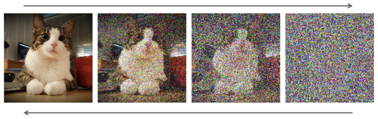
The Mathematics of Diffusion Models

McAllester, arXiv January 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.

SDE Notation

In these slides ϵ will always be a random variable drawn from $\mathcal{N}(0, I)$.

This correspods to "dB" in the more standard way of writing SDEs.

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

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

While the expression in terms of Δt is more verbose, it seems clearer to me.

The Diffusion SDE

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

For the diffusion process (Brownian motion) we have

$$\begin{aligned} z(0) &= y, \quad y \sim \mathrm{Pop}(y) \\ z(t + \Delta t) &= z(t) + \epsilon\sqrt{\Delta t} \qquad (1) \end{aligned}$$

For diffusion we get that (1) holds for all t and Δt.

Probability Notation

In these slides unsubscripted probability notation, such as

$$P(z(t+\Delta t)|z(t)),$$

or a conditiional expectation such as

$$E[f(y)|z(t)] = E_{y\sim P(y|z_t)}[f(y)],$$

refer the joint probability distribution on y and the (continuous) function $z(t)$ defined by the diffusion process.

Reverse-Time Probabilities

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

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

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

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

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

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

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

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 $\sqrt{1+t}$.

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

We then have

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

Computing Bits per Channel (or Perplexity)

For any Markovian VAE we have

$$-\ln \mathrm{Pop}(y) \;=\; -\ln \frac{P(z_N)P(z_{N-1}|z_N)\cdots P(z_1|z_2)P(y|z_1)}{P(z_N|y)P(z_{N-1}|z_N,y)\cdots P(z_1|z_2,y)}$$

$$H(y) \;=\; \left\{\begin{array}{l} E[KL(P(z_N|y),\;P(z_N))] \\ \\ +\sum_{i=2}^N\; E[KL(P(z_{i-1}|z_i,y),\;P(z_{i-1}|z_i))] \\ \\ +\;E[\ln - P(y|z_1)] \end{array}\right. \tag{2}$$

$$\leq \;\left\{\begin{array}{l} E[KL(P(z_N|y),\;P_{\mathrm{gen}}(z_N))] \\ \\ +\sum_{i=2}^N\; E[KL(P(z_{i-1}|z_i,y),\;P_{\mathrm{gen}}(z_{i-1}|z_i))] \\ \\ E[-\ln P_{\mathrm{gen}}(y|z_1)] \end{array}\right. \tag{3}$$

KL-Divergence

$$H(y) \;=\; \begin{cases} E[KL(P(z_N|y),\;P(z_N))] \\ \\ +\sum_{i=2}^{N}\;E[KL(P(z_{i-1}|z_i,y),\;P(z_{i-1}|z_i))] \\ \\ +\;E[\ln -P(y|z_1)] \end{cases}$$

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

$$H(y) \;=\; \begin{cases} E[KL(P(z_N|y),\; P(z_N))] \\ \\ + \sum_{i=2}^{N} \; E[KL(P(z_{i-1}|z_i, y),\; P(z_{i-1}|z_i))] \\ \\ +\; E[\ln - P(y|z_1)] \end{cases}$$

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

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

KL-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

$$H(y) \;=\; \left\{ \begin{array}{l} E[KL(P(z_N|y),\; P(z_N))] \\ \\ +\sum_{i=2}^{N}\; E[KL(P(z_{i-1}|z_i, y),\; P(z_{i-1}|z_i))] \\ \\ +\; E[\ln - P(y|z_1)] \end{array} \right.$$

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

$$t = i\Delta t$$

Passing to the Integral

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

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

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].

Computing Bits per Channel

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

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

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

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

$$\begin{aligned} z(t + \Delta t) &= z(t) + \epsilon\sqrt{\Delta t} \\ dz &= dB \end{aligned}$$

The Fokker-Planck Equation

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

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

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