Math derivation: DDPM

Duc Anh Nguyen

May 2023

This note presents the mathematical derivations related to denoising diffusion probabilistic models (DDPMs) [HJA20]., an emerging class of generative models that have garnered substantial popularity since 2020.

1 Notations

- In general, we shall denote a data item with the symbol $\mathbf{x}$, which is generally a vector in $\mathbb{R}^d$.
- When several data items are sampled from the same distribution, we differentiate between them by subscripts; for examples, $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, and so on.
- We will also deal with sequence of data items. In particular, we will deal with Markov chains in which the next element in the sequence is generated from the previous one. Different elements of the sequence are differentiated by superscripts which are always inside a pair of parentheses; for example $(\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(100)})$.
- Writing $(\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(100)})$ is cumbersome. We can abbreviate it with $\mathbf{x}^{(0:100)}$.

2 Math derivation: DDPM

- Ho et al. uses the following Markov chain:

  $$\begin{aligned} \mathbf{x}^{(0)} &\sim q(\mathbf{x}^{(0)}), \\ \mathbf{x}^{(t)} &\sim \mathcal{N}(\sqrt{1-\beta_t}\mathbf{x}^{(t-1)}, \beta_t I) \end{aligned}$$

  for all $t = 1, 2, \dots, T$. Here, β_1, β_2, $\dots$, β_T are small positive constants collectively called the **variance schedule**.

- They fix $T = 1000$.
- The picked a linear progression where $\beta_1 = 10^{-4}$ and $\beta_T = 0.02$ as the variance schedule.
- Note that we can write $\mathbf{x}^{(t)}$ in terms for $\mathbf{x}^{(t-1)}$ as follows:

  $$\mathbf{x}^{(t)} = \sqrt{1-\beta_t}\mathbf{x}^{(t-1)} + \sqrt{\beta_t}\boldsymbol{\xi}$$

  where $\boldsymbol{\xi}$ is a random variable distributed according to $\mathcal{N}(\mathbf{0}, I)$. The conditional probability of $\mathbf{x}^{(t)}$ given $\mathbf{x}^{(t-1)}$ (i.e., the transition kernel) is given by:

  $$q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}) = \mathcal{N}(\mathbf{x}^{(t)}; \sqrt{1-\beta_t}\mathbf{x}^{(t-1)}, \beta_t I) = \frac{1}{(2\pi\beta_t)^{d/2}} \exp\left(\frac{-\|\mathbf{x}^{(t)} - \sqrt{1-\beta_t}\mathbf{x}^{(t-1)}\|_2^2}{2\beta_t} \right).$$

2.1 The joint probability of forward process given the initial state

Prove that the join probability of sampling $\mathbf{x}^{(1:T)} = (\mathbf{x}^{(1)}, \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(T)})$ given $\mathbf{x}^{(0)}$ is

$$q(\mathbf{x}^{(1:T)}|\mathbf{x}^{(0)}) = \prod_{t=1}^{T} q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}).$$

Proof.

$$\begin{aligned} q(\mathbf{x}^{(1:T)}|\mathbf{x}^{(0)}) &= \prod_{t=1}^{T} q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}, ..., \mathbf{x}^{(0)}) && \text{(chain rule)} \\ &= \prod_{t=1}^{T} q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}) && \text{(Markov property)} \end{aligned}$$

2.2 The closed form of $\mathbf{x}^{(t)}$ at any arbitrary time step t

Let $\alpha_t = 1 - \beta_t$. Let $\overline{\alpha}_t = \prod_{i=1}^{t} \alpha_i$. For any $1 \le t \le T$, prove that:

$$\mathbf{x}^{(t)} = \sqrt{\overline{\alpha}_t}\mathbf{x}^{(0)} + \sqrt{1 - \overline{\alpha}_t}\boldsymbol{\xi}$$

Proof.

$$\begin{aligned} \mathbf{x}^{(t)} &= \sqrt{\alpha_t}\mathbf{x}^{(t-1)} + \sqrt{1-\alpha_t}\boldsymbol{\xi}^{(t-1)} \\ &= \sqrt{\alpha_t}\left(\sqrt{\alpha_{t-1}}\mathbf{x}^{(t-2)} + \sqrt{1-\alpha_{t-1}}\boldsymbol{\xi}^{(t-2)}\right) + \sqrt{1-\alpha_t}\boldsymbol{\xi}^{(t-1)} \\ &= \sqrt{\alpha_t\alpha_{t-1}}\mathbf{x}^{(t-2)} + \sqrt{\alpha_t(1-\alpha_{t-1})}\boldsymbol{\xi}^{(t-2)} + \sqrt{1-\alpha_t}\boldsymbol{\xi}^{(t-1)} \\ &= \sqrt{\alpha_t\alpha_{t-1}}\mathbf{x}^{(t-2)} + \sqrt{\alpha_t(1-\alpha_{t-1}) + (1-\alpha_t)}\boldsymbol{\xi} && \text{(because } \boldsymbol{\xi}^{(t-2)} \perp \boldsymbol{\xi}^{(t-1)}\text{)} \\ &= \sqrt{\alpha_t\alpha_{t-1}}\mathbf{x}^{(t-2)} + \sqrt{1-\alpha_t\alpha_{t-1}}\boldsymbol{\xi}^{(t-1)} \\ &= \dots \\ &= \sqrt{\overline{\alpha}_t}\mathbf{x}^{(0)} + \sqrt{1-\overline{\alpha}_t}\boldsymbol{\xi} \end{aligned}$$

2.3 The joint probability of backward process

Prove that the probability of sampling $\mathbf{x}^{(0:T)}$ is

$$p_{\boldsymbol{\theta}}(\mathbf{x}^{(0:T)}) = p_{\boldsymbol{\theta}}(\mathbf{x}^{(T)}) \prod_{t=1}^{T} p_{\boldsymbol{\theta}}(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}).$$

Proof.

$$\begin{aligned} p_{\boldsymbol{\theta}}(\mathbf{x}^{(0:T)}) &= p_{\boldsymbol{\theta}}(\mathbf{x}^{(T)}) \prod_{t=1}^{T} p_{\boldsymbol{\theta}}(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}, ..., \mathbf{x}^{(T)}) && \text{(chain rule)} \\ &= p_{\boldsymbol{\theta}}(\mathbf{x}^{(T)}) \prod_{t=1}^{T} p_{\boldsymbol{\theta}}(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}) && \text{(Markov property)} \end{aligned}$$

2.4 Derivation of $q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}, \mathbf{x}^{(0)})$

$$\begin{aligned}
q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}, \mathbf{x}^{(0)}) &= \frac{q(\mathbf{x}^{(t-1)}, \mathbf{x}^{(t)}, \mathbf{x}^{(0)})}{q(\mathbf{x}^{(t-1)}, \mathbf{x}^{(0)})} && \text{(Bayes' rule)} \\
&= \frac{q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}, \mathbf{x}^{(0)})q(\mathbf{x}^{(t)}|\mathbf{x}^{(0)})q(\mathbf{x}^{(0)})}{q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(0)})q(\mathbf{x}^{(0)})} && \text{(chain rule)} \\
&= \frac{q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})q(\mathbf{x}^{(t)}|\mathbf{x}^{(0)})}{q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(0)})} && \text{(Cancel out } q(\mathbf{x}^{(0)})\text{)}
\end{aligned}$$

2.5 The closed form of KL divergence between two multivariate Gaussian distributions

Let $\mathbf{q} \sim \mathcal{N}(\mu_q, \mathbf{\Sigma}_q)$ and $\mathbf{p} \sim \mathcal{N}(\mu_p, \mathbf{\Sigma}_p)$. Prove that

$$\mathrm{KL}(\mathbf{q}||\mathbf{p}) = \frac{1}{2}\left(\mathrm{tr}(\mathbf{\Sigma}_p^{-1}\mathbf{\Sigma}_q) + (\mu_p - \mu_q)^T\mathbf{\Sigma}_p^{-1}(\mu_p - \mu_q) - k + \log\frac{\det(\mathbf{\Sigma}_p)}{\det(\mathbf{\Sigma}_q)}\right).$$

Proof.

$$\begin{aligned}
D_{KL}(p||q) &= \mathbb{E}_p\left[\log(p) - \log(q)\right] \\
&= \mathbb{E}_p\left[\frac{1}{2}\log\frac{|\Sigma_q|}{|\Sigma_p|} - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu_p})^T\Sigma_p^{-1}(\mathbf{x} - \boldsymbol{\mu_p}) + \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu_q})^T\Sigma_q^{-1}(\mathbf{x} - \boldsymbol{\mu_q})\right] \\
&= \frac{1}{2}\mathbb{E}_p\left[\log\frac{|\Sigma_q|}{|\Sigma_p|}\right] - \frac{1}{2}\mathbb{E}_p\left[(\mathbf{x} - \boldsymbol{\mu_p})^T\Sigma_p^{-1}(\mathbf{x} - \boldsymbol{\mu_p})\right] + \frac{1}{2}\mathbb{E}_p\left[(\mathbf{x} - \boldsymbol{\mu_q})^T\Sigma_q^{-1}(\mathbf{x} - \boldsymbol{\mu_q})\right] \\
&= \frac{1}{2}\log\frac{|\Sigma_q|}{|\Sigma_p|} - \frac{1}{2}\mathbb{E}_p\left[(\mathbf{x} - \boldsymbol{\mu_p})^T\Sigma_p^{-1}(\mathbf{x} - \boldsymbol{\mu_p})\right] + \frac{1}{2}\mathbb{E}_p\left[(\mathbf{x} - \boldsymbol{\mu_q})^T\Sigma_q^{-1}(\mathbf{x} - \boldsymbol{\mu_q})\right]
\end{aligned}$$

Now, since $(\mathbf{x} - \boldsymbol{\mu_p})^T\Sigma_p^{-1}(\mathbf{x} - \boldsymbol{\mu_p})$ in the second term $\in \mathbb{R}$, we can write it as $tr\left\{(\mathbf{x} - \boldsymbol{\mu_p})^T\Sigma_p^{-1}(\mathbf{x} - \boldsymbol{\mu_p})\right\}$, where $tr\{\}$ is the trace operator. And using the trace trick $\mathrm{tr}\left\{\mathbf{AB}\right\} = \mathrm{tr}\left\{\mathbf{BA}\right\}$, we can write it as $tr\left\{(\mathbf{x} - \boldsymbol{\mu_p})(\mathbf{x} - \boldsymbol{\mu_p})^T\Sigma_p^{-1}\right\}$.

The second term now is,

$$= \frac{1}{2}\mathbb{E}_p\left[tr\left\{(\mathbf{x} - \boldsymbol{\mu_p})(\mathbf{x} - \boldsymbol{\mu_p})^T\Sigma_p^{-1}\right\}\right]$$

The expectation and trace can be interchanged to get,

$$\begin{aligned}
&= \frac{1}{2}\,\mathrm{tr}\left\{\mathbb{E}_p\left[(\mathbf{x} - \boldsymbol{\mu_p})(\mathbf{x} - \boldsymbol{\mu_p})^T\Sigma_p^{-1}\right]\right\} \\
&= \frac{1}{2}\,\mathrm{tr}\left\{\mathbb{E}_p\left[(\mathbf{x} - \boldsymbol{\mu_p})(\mathbf{x} - \boldsymbol{\mu_p})^T\right]\Sigma_p^{-1}\right\}
\end{aligned}$$

We know $\mathbb{E}_p\left[(\mathbf{x} - \boldsymbol{\mu_p})(\mathbf{x} - \boldsymbol{\mu_p})^T\right] = \Sigma_p$. Simplifying it to

$$\begin{aligned}
&= \frac{1}{2}\,\mathrm{tr}\left\{\Sigma_p\Sigma_p^{-1}\right\} \\
&= \frac{1}{2}\,\mathrm{tr}\left\{I_k\right\} \\
&= \frac{k}{2}
\end{aligned}$$

Theorem 1. *Expectation of square forms*
Let $\mathbf{X}$ *be a* k*-dimensional random vector and* $\mathbf{A}$ *be a constant* $k \times k$ *symmetric matrix. If* $\mathbb{E}(\mathbf{X}) = \boldsymbol{\mu}$ *and* $\mathrm{Var}(\mathbf{X}) = \Sigma$*, then*

$$\mathbb{E}(\mathbf{X}^T\mathbf{A}\mathbf{X}) = \mathrm{tr}(\mathbf{A}\Sigma) + \boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu}$$

Proof.

$$\begin{aligned}
\mathbb{E}(\mathbf{X}^T\mathbf{A}\mathbf{X}) &= \mathrm{tr}(\mathbf{A}\Sigma) + \boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu} \\
&= \mathrm{tr}(\mathbb{E}(\mathbf{X}^T\mathbf{A}\mathbf{X})) \\
&= \mathbb{E}(\mathrm{tr}(\mathbf{X}^T\mathbf{A}\mathbf{X})) \\
&= \mathbb{E}(\mathrm{tr}(\mathbf{A}\mathbf{X}\mathbf{X}^T)) \\
&= \mathrm{tr}(\mathbf{A}\mathbb{E}(\mathbf{X}\mathbf{X}^T)) \\
&= \mathrm{tr}(\mathbf{A}(\mathrm{Cov}(\mathbf{X}) + \boldsymbol{\mu}\boldsymbol{\mu}^T)) \\
&= \mathrm{tr}(\mathbf{A}\Sigma + \mathbf{A}\boldsymbol{\mu}\boldsymbol{\mu}^T) \\
&= \mathrm{tr}(\mathbf{A}\Sigma) + \mathrm{tr}(\mathbf{A}\boldsymbol{\mu}\boldsymbol{\mu}^T) \\
&= \mathrm{tr}(\mathbf{A}\Sigma) + \boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu}
\end{aligned}$$

□

We can simplify the third term using the above theorem, we get,

$$\mathbb{E}_p\left[(\mathbf{x} - \boldsymbol{\mu_q})^T\Sigma_q^{-1}(\mathbf{x} - \boldsymbol{\mu_q})\right] = (\boldsymbol{\mu_p} - \boldsymbol{\mu_q})^T\Sigma_q^{-1}(\boldsymbol{\mu_p} - \boldsymbol{\mu_q}) + \mathrm{tr}\left\{\Sigma_q^{-1}\Sigma_p\right\}$$

Combining all this we get,

$$D_{KL}(p||q) = \frac{1}{2}\left[\log\frac{|\Sigma_q|}{|\Sigma_p|} - k + (\boldsymbol{\mu_p} - \boldsymbol{\mu_q})^T\Sigma_q^{-1}(\boldsymbol{\mu_p} - \boldsymbol{\mu_q}) + \mathrm{tr}\left\{\Sigma_q^{-1}\Sigma_p\right\}\right]$$

2.5.1 Simplify $L_{1:(T-1)}$

- To optimize these terms, we have to specify what $p_{\boldsymbol{\theta}}(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})$ is.
- The Ho et al. paper chooses it to be

  $$p(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}) := \mathcal{N}\big(\mathbf{x}^{(t-1)}; \boldsymbol{\mu_\theta}(\mathbf{x}^{(t)}, t), \beta_t I\big)$$

  where $\boldsymbol{\mu_\theta}(\cdot,\cdot)$ is a neural network. The variance of the Gaussian was chosen empirically.
- Now, consider

  $$D_{KL}\big(q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(0)}, \mathbf{x}^{(t)})\big\|p_{\boldsymbol{\theta}}(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})\big).$$

  We know that $q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(0)}, \mathbf{x}^{(t)})$ is a Gaussian distribution of $\mathbf{x}^{(t-1)}$. Moreover, $p_{\boldsymbol{\theta}}(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})$ is a Gaussian distribution of $\mathbf{x}^{(t-1)}$ by definition. So, the KL-divergence can be computed using Corollary **??**. Applying it, we have that

  $$D_{KL}\big(q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(0)}, \mathbf{x}^{(t)}) = \frac{\|\tilde{\boldsymbol{\mu}}_t(\mathbf{x}^{(0)}, \mathbf{x}^{(t)}) - \boldsymbol{\mu_\theta}(\mathbf{x}^{(t)}, t)\|^2}{2\beta_t} + C \tag{1}$$

  where C is a constant that does not depend on $\boldsymbol{\theta}$ or $\mathbf{x}^{(0)}$ or $\mathbf{x}^{(t-1)}$. (More specifically, C can be written in terms of d and the β_t's in the noise schedule.)
- Now,

  $$L_{t-1} = E_{\mathbf{x}^{(0,t)}\sim q}\left[\frac{\|\tilde{\boldsymbol{\mu}}_t(\mathbf{x}^{(0)}, \mathbf{x}^{(t)}) - \boldsymbol{\mu_\theta}(\mathbf{x}^{(t)}, t)\|^2}{2\beta_t}\right] + C$$

2.5.2 The closed form of forward process posteriors, which are tractable when conditioned on x0

It is noteworthy that the reverse conditional probability is tractable when conditioned on $\mathbf{x}_0$:

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}; \tilde{\boldsymbol{\mu}}(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t \mathbf{I})$$

Using Bayes' rule, we have:

$$\begin{aligned} q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) &= q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_0)\frac{q(\mathbf{x}_{t-1}|\mathbf{x}_0)}{q(\mathbf{x}_t|\mathbf{x}_0)} \\ &\propto \exp\Big(-\frac{1}{2}(\frac{(\mathbf{x}_t - \sqrt{\alpha_t}\mathbf{x}_{t-1})^2}{\beta_t} + \frac{(\mathbf{x}_{t-1} - \sqrt{\bar{\alpha}_{t-1}}\mathbf{x}_0)^2}{1-\bar{\alpha}_{t-1}} - \frac{(\mathbf{x}_t - \sqrt{\bar{\alpha}_t}\mathbf{x}_0)^2}{1-\bar{\alpha}_t})\Big) \\ &= \exp\Big(-\frac{1}{2}(\frac{\mathbf{x}_t^2 - 2\sqrt{\alpha_t}\mathbf{x}_t\mathbf{x}_{t-1} + \alpha_t\mathbf{x}_{t-1}^2}{\beta_t} + \frac{\mathbf{x}_{t-1}^2 - 2\sqrt{\bar{\alpha}_{t-1}}\mathbf{x}_0\mathbf{x}_{t-1} + \bar{\alpha}_{t-1}\mathbf{x}_0^2}{1-\bar{\alpha}_{t-1}} - \frac{(\mathbf{x}_t - \sqrt{\bar{\alpha}_t}\mathbf{x}_0)^2}{1-\bar{\alpha}_t})\Big) \\ &= \exp\Big(-\frac{1}{2}((\frac{\alpha_t}{\beta_t} + \frac{1}{1-\bar{\alpha}_{t-1}})\mathbf{x}_{t-1}^2 - (\frac{2\sqrt{\alpha_t}}{\beta_t}\mathbf{x}_t + \frac{2\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_{t-1}}\mathbf{x}_0)\mathbf{x}_{t-1} + C(\mathbf{x}_t, \mathbf{x}_0))\Big) \end{aligned}$$

where $C(\mathbf{x}_t, \mathbf{x}_0)$ is some function not involving $\mathbf{x}_{t-1}$ and details are omitted.

Following the standard Gaussian density function, the mean and variance can be parameterized as follows (recall that $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{T} \alpha_i$):

$$\begin{aligned} \tilde{\beta}_t &= 1/(\frac{\alpha_t}{\beta_t} + \frac{1}{1-\bar{\alpha}_{t-1}}) = 1/(\frac{\alpha_t - \bar{\alpha}_t + \beta_t}{\beta_t(1-\bar{\alpha}_{t-1})}) = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \cdot \beta_t \\ \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) &= (\frac{\sqrt{\alpha_t}}{\beta_t}\mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_{t-1}}\mathbf{x}_0)/(\frac{\alpha_t}{\beta_t} + \frac{1}{1-\bar{\alpha}_{t-1}}) \\ &= (\frac{\sqrt{\alpha_t}}{\beta_t}\mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_{t-1}}\mathbf{x}_0)\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \cdot \beta_t \\ &= \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_0 \end{aligned}$$

Thanks to the nice property, we can represent $\mathbf{x}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t)$ and plug it into the above equation and obtain:

$$\tilde{\boldsymbol{\mu}}_t = \frac{1}{\sqrt{\alpha_t}}\Big(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_t\Big)$$

3 Math derivation: Denoising Score Matching

[SSDK$^+$20]

3.1 Prove the joint density of points in a Brownian motion/Wiener process is a multivariate Gaussian distribution

Quick review of Wiener process: A Wiener process $(W(t)_t \geq 0)$ is a real-valued stochastic process with the following properties:

1. $W(0) = 0$.
2. Independent increments: the random variables $W(t_1), W(t_2) - W(t_1), ..., W(t_n) - W(t_{n-1})$ are independent for any $0 \leq t_1 < t_2 < ... < t_n$.
3. Normal increments: $W(t) - W(s) \sim \mathcal{N}(0, t-s)$ for any $0 \leq s < t$.

4. Continuous sample paths: with probability 1, the function $t \to W(t)$ is continuous.

Proof.

Let $\mathbf{X}_1 = \mathbf{W}(t_1), \mathbf{X}_2 = \mathbf{W}(t_2) - \mathbf{W}(t_1), ..., \mathbf{X}_n = \mathbf{W}(t_n) - \mathbf{W}(t_{n-1})$. We have

$$\begin{aligned} p(\mathbf{X}_1, ..., \mathbf{X}_n) &= p(\mathbf{X}_1)p(\mathbf{X}_2|\mathbf{X}_1)...p(\mathbf{X}_n|\mathbf{X}_1, ..., \mathbf{X}_{n-1}) && \text{(chain rule)} \\ &= p(\mathbf{X}_1)p(\mathbf{X}_2)...p(\mathbf{X}_n) && \text{(independent increments)} \\ &= \mathcal{N}(\mathbf{0}, t_1\mathbf{I})\mathcal{N}(\mathbf{0}, (t_2 - t_1)\mathbf{I})...\mathcal{N}(\mathbf{0}, (t_n - t_{n-1})\mathbf{I}) && \text{(normal increments)} \\ &= \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}) \end{aligned}$$

where $\mathbf{\Sigma} = \text{diag}(t_1, t_2 - t_1, ..., t_n - t_{n-1})$.

Theorem 2. *Suppose that* $\mathbf{X}$ *is a random variable taking values in* $\mathbf{S} \subseteq \mathbb{R}^n$ *and that* $\mathbf{X}$ *has a continuous distribution with probability density function* $p(\mathbf{x})$*. Let* $\mathbf{Y} = f(\mathbf{X})$ *be a random variable taking values in* $\mathbf{T} \subseteq \mathbb{R}^m$ *and let* f *be a continuously differentiable function from* $\mathbf{S}$ *to* $\mathbf{T}$*. Then the probability density function of* $\mathbf{Y}$ *is given by*

$$p(\mathbf{y}) = p(\mathbf{x})\left| det\left(\frac{\partial \mathbf{x}}{\partial \mathbf{y}}\right)\right|$$

where $\frac{\partial \mathbf{x}}{\partial \mathbf{y}}$ *is the Jacobian matrix of* f*.*

Consider the function $f : \mathbb{R}^n \to \mathbb{R}^n$ such that $(\mathbf{y}_1), ..., (\mathbf{y}_n) = f(\mathbf{x}_1, ..., \mathbf{x}_n) = (\mathbf{x}_1, \mathbf{x}_2 + \mathbf{x}_1, ..., \mathbf{x}_n + \mathbf{x}_{n-1}) = (\mathbf{W}(t_1), \mathbf{W}(t_2), ..., \mathbf{W}(t_n))$. Then $f^{-1}(y_1, ..., y_n) = (y_1, y_2 - y_1, ..., y_n - y_{n-1})$.

The Jacobian matrix of $f^{-1}(y_1, ..., y_n)$ is 1. By the theorem above, we have

$$\begin{aligned} p(\mathbf{W}(t_1), ..., \mathbf{W}(t_n)) &= p(\mathbf{y}_1, ..., \mathbf{y}_n) \\ &= p(\mathbf{x}_1, ..., \mathbf{x}_n)\left|\det\left(\frac{\partial \mathbf{x}}{\partial \mathbf{y}}\right)\right| \\ &= p(\mathbf{x}_1, ..., \mathbf{x}_n) \\ &= \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}) \end{aligned}$$

where $\mathbf{\Sigma} = \text{diag}(t_1, t_2 - t_1, ..., t_n - t_{n-1})$.

3.2 The tractable/closed form of the score function

Let $\mathbf{x}^{(\mathbf{t})} = \gamma\mathbf{x}^{(0)} + \sigma^{(t)}\boldsymbol{\xi}$, where $\boldsymbol{\xi} \sim \mathcal{N}((\mathbf{0}), \mathbf{I})$. Prove that

$$\nabla_{\mathbf{x}^{(\mathbf{t})}} \log q(\mathbf{x}^{(t)}|\mathbf{x}^{(0)}) = -\frac{\boldsymbol{\xi}}{\sigma^{(t)}}$$

Proof.

$$\begin{aligned} \nabla_{\mathbf{x}^{(\mathbf{t})}} \log q(\mathbf{x}^{(t)}|\mathbf{x}^{(0)}) &= \nabla_{\mathbf{x}^{(\mathbf{t})}} \log \mathcal{N}(\mathbf{x}^{(t)}|\gamma\mathbf{x}^{(0)}, \sigma^{(t)}) \\ &= \nabla_{\mathbf{x}^{(\mathbf{t})}}\left(-\frac{1}{2}(\mathbf{x}^{(t)} - \gamma\mathbf{x}^{(0)})^T\sigma^{(t)}\mathbb{I}(\mathbf{x}^{(t)} - \gamma\mathbf{x}^{(0)})\right) \\ &= -\frac{1}{2}\frac{2(\mathbf{x}^{(t)} - \gamma\mathbf{x}^{(0)})}{(\sigma^{(t)})^2} \\ &= -\frac{\boldsymbol{\xi}}{\sigma^{(t)}} \end{aligned}$$

References

[HJA20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.

[SSDK+20] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations, 2020.