Exercises on Denoising diffusion probabilistic models

Gabriel V. Cardoso
STIM
Ecole des Mines Paris(PSL)
Fontaineableau, France
gabriel.victorino_cardoso@minesparis.psl.eu

September 15, 2026

Notation

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\mathbf{X} : (\Omega, \mathcal{F}) \to (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and $\mathbf{Y} : (\Omega, \mathcal{F}) \to (\mathbb{R}^m, \mathcal{B}(\mathbb{R}^m))$. We say that a Markov kernel $\mathsf{N} : \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^m) \to \mathbb{R}_+$ is a conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$ if

$$\mathsf{N}(\mathbf{X}, \mathrm{A}) = \mathbb{E}\left[1_{\mathrm{A}}(\mathbf{Y})|\mathbf{X}\right] \quad \mathbb{P} - \text{almost surely.} \tag{1}$$

We use the notation $\mathsf{N} = \mathcal{L}(\mathbf{Y}|\mathbf{X})$.

Proposition 0.1. *Suppose that for* $\mathrm{B} \in \mathcal{B}(\mathbb{R}^{2d})$, $\mathcal{L}(\mathbf{X}, \mathbf{Y})(\mathrm{B}) = \int 1_{\mathrm{B}}(\mathbf{x}, \mathbf{y})\, p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})\mu(\mathrm{d}\mathbf{x})\mathrm{d}\mathbf{y}$[1] *where*

$$p_{\mathbf{Y}|\mathbf{X}} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+ \quad \textit{such that} \quad \mu\left(\left\{\mathbf{x} \in \mathbb{R}^d \middle| \int p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})\mathrm{d}\mathbf{y} = 1\right\}\right) = 1\,. \tag{2}$$

Then

$$\mathsf{N} : \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d) \ni (\mathbf{x}, \mathrm{A}) \to \int 1_{\mathrm{A}}(\mathbf{y})\, p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})\mathrm{d}\mathbf{y}$$

is a conditional distribution (and the natural one) $\mathcal{L}(\mathbf{Y}|\mathbf{X})$.

Proof. It is straightforward to check that it is a Markov Kernel. We will check for the conditional expectation only.

We denote $\hat{f}(\mathbf{x}) = \int 1_{\mathrm{A}}(\mathbf{y}) p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})\mathrm{d}\mathbf{y}$. Note that it is enough (as we implicitly consider everywhere in the course that we are in L^2) that

$$\mathbb{E}\left[(1_{\mathrm{A}}(\mathbf{Y}) - \hat{f}(\mathbf{X}))\hat{f}(\mathbf{X})\right] = 0\,. \tag{3}$$

But

$$\mathbb{E}\left[(1_{\mathrm{A}}(\mathbf{Y}) - \hat{f}(\mathbf{X}))\hat{f}(\mathbf{X})\right] = \int (1_{\mathrm{A}}(\mathbf{y}) - \hat{f}(\mathbf{x}))\hat{f}(\mathbf{x}) p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})\mu(\mathrm{d}\mathbf{x})\mathrm{d}\mathbf{y} \tag{4}$$

$$= \int \underbrace{\left(\int 1_{\mathrm{A}}(\mathbf{y})\, p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})\mathrm{d}\mathbf{y}\right)}_{\hat{f}(\mathbf{x})} \hat{f}(\mathbf{x})\mu(\mathrm{d}\mathbf{x}) - \int \hat{f}(\mathbf{x})^2\mu(\mathrm{d}\mathbf{x}) = 0\,. \tag{5}$$

Thus, $\hat{f}(\mathbf{X}) = \mathsf{N}(\mathbf{X}, \mathrm{A}) = \mathbb{E}\left[1_{\mathrm{A}}(\mathbf{Y})|\mathbf{X}\right]$. □

Some observations are important:

- We can show that conditional distributions exist almost always (if the image space is nice enough) [1, Theorem B.3.11].

[1] This simply means that the conditional admits a density with respect to the Lebesgue measure

1 Noising Markov Kernels

We define for $\mathbf{x} \in \mathbb{R}^d$ and $\mathrm{A} \in \mathcal{B}(\mathbb{R}^d)$, $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}_+$,

$$\mathsf{Q}(\mathbf{x}, \mathrm{A}) := \int 1_{\mathrm{A}}(\mathbf{y}) \mathcal{N}\left(\mathbf{y}; \alpha\mathbf{x}, \beta^2 \mathrm{I}\right) \mathrm{d}\mathbf{y} \,. \tag{6}$$

1. Give and upper bound on $\mathrm{KL}\left(\mu\mathsf{Q} \middle\| \mathcal{N}\left(\mathbf{m}, \sigma^2 \mathrm{I}\right)\right)$. Interpret what it means, at least for well chosen values of $\alpha, \beta, \mathbf{m}$ and σ.
2. Let $\eta = \mathcal{N}(0, \mathrm{I})$. Give a condition on α and β so that η is an invariant distribution for the Q, which means that
$$\eta\mathsf{Q} = \eta \,. \tag{7}$$
Give an intuitive interpretation.
3. What is the relation between the random variables
$$\mathbf{X}_0 \sim \mu \,, \qquad \mathbf{X}_1 = \alpha\mathbf{X}_0 + \beta\mathbf{Z} \,, \qquad \mathbf{Z} \sim \mathcal{N}(0, \mathrm{I}) \,, \qquad \mathbf{X}_0 \perp\!\!\!\perp \mathbf{Z} \,, \tag{8}$$
and Q and μ ?
4. For $n \in \mathbb{N}$, give and expression for the distribution $\mathsf{Q}^n(\mathbf{x})$.
5. Identify the cases where for every $\mathbf{x}$,
$$\lim_{n\to\infty} \mathrm{KL}\left(\mathsf{Q}^n(\mathbf{x}) \| \mathcal{N}(0, \mathrm{I})\right) = 0 \,. \tag{9}$$
Do the speed of convergence depends on $\mathbf{x}$?

2 Bridge distribution on Gaussian Markov Chains

We consider now the following Markov Chain

$$\mathbf{X}_0 \sim \mu \,, \quad \mathbf{X}_1 = \mathbf{X}_0 + \sigma_1\mathbf{Z}_1 \,, \quad \mathbf{X}_2 = \mathbf{X}_1 + \sigma_2\mathbf{Z}_2 \,, \quad \mathbf{Z}_1 \perp\!\!\!\perp \mathbf{Z}_2 \perp\!\!\!\perp \mathbf{X}_0 \,. \tag{10}$$

1. What is $\mathcal{L}(\mathbf{X}_i|\mathbf{X}_0)$ for $i = 1, 2$?

Solution: Let f be a continuous bounded function.

$$\begin{aligned}\mathbb{E}\left[f(\mathbf{X}_0, \mathbf{X}_1, \mathbf{X}_2)\right] &= \int f(\mathbf{x}_0, \mathbf{x}_0 + \sigma_1\mathbf{z}_0, \mathbf{x}_0 + \sigma_2\mathbf{z}_0 + \sigma\mathbf{z}_1)\mathcal{N}(\mathbf{z}_1; 0, \mathrm{I})\,\mathcal{N}(\mathbf{z}_0; 0, \mathrm{I})\,\mu(\mathrm{d}x_0)\mathrm{d}\mathbf{z}_0\mathrm{d}\mathbf{z}_1 \\ &= \int f(\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2)\mathcal{N}\left(\mathbf{x}_1; \mathbf{x}_0, \sigma_1^2\mathrm{I}\right)\mathcal{N}\left(\mathbf{x}_2; \mathbf{x}_1, \sigma_2^2\mathrm{I}\right)\mu(\mathrm{d}\mathbf{x}_0)\mathrm{d}\mathbf{x}_1\mathrm{d}\mathbf{x}_2 \,.\end{aligned}$$

The joint vector $(\mathbf{X}_0, \mathbf{X}_1, \mathbf{X}_2)$ admits a density $(\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2) \to \mathcal{N}\left(\mathbf{x}_1; \mathbf{x}_0, \sigma_1^2\mathrm{I}\right)\mathcal{N}\left(\mathbf{x}_2; \mathbf{x}_1, \sigma_2^2\mathrm{I}\right)$ with respect to the product measure $\mu \otimes \mathrm{Leb} \otimes \mathrm{Leb}$, where Leb is the Lebesgue measure in $\mathbb{R}^d$.

Particularly, if we consider that μ admits a density f_μ with respect to the Lebesgue measure, then $(\mathbf{X}_0, \mathbf{X}_1, \mathbf{X}_2)$ admits a density $(\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2) \to f_\mu(\mathbf{x}_0)\mathcal{N}\left(\mathbf{x}_1; \mathbf{x}_0, \sigma_1^2\mathrm{I}\right)\mathcal{N}\left(\mathbf{x}_2; \mathbf{x}_1, \sigma_2^2\mathrm{I}\right)$ with respect to the product measure $\mathrm{Leb} \otimes \mathrm{Leb} \otimes \mathrm{Leb}$ which is simply the Lebesgue measure in $\mathbb{R}^{3d}$.

We can now directly calculate the conditional distributions using proposition 0.1.

$$\mathcal{L}(\mathbf{X}_1|\mathbf{X}_0)(\mathbf{x}_0, \mathrm{A}) = \int 1_{\mathrm{A}}(\mathbf{x}_1)\mathcal{N}\left(\mathbf{x}_1; \mathbf{x}_0, \sigma_1^2\mathrm{I}\right)\mathrm{d}\mathbf{x}_1 \,, \tag{11}$$

which is equivalent to say that $\mathcal{L}(\mathbf{X}_1|\mathbf{X}_0)(\mathbf{x}_0, \cdot) = \mathcal{N}\left(\mathbf{x}_0, \sigma_1^2\mathrm{I}\right)$, or even that $\mathcal{L}(\mathbf{X}_1|\mathbf{X}_0) = \mathcal{N}\left(\mathbf{X}_0, \sigma_1^2\mathrm{I}\right)$.
For $\mathbf{X}_2$, simply note that $\int \mathcal{N}\left(\mathbf{x}_1; \mathbf{x}_0, \sigma_1^2\mathrm{I}\right)\mathcal{N}\left(\mathbf{x}_2; \mathbf{x}_1, \sigma_2^2\mathrm{I}\right)\mathrm{d}\mathbf{x}_1 = \mathcal{N}\left(\mathbf{x}_2; \mathbf{x}_0, (\sigma_2^2 + \sigma_1^2)\mathrm{I}\right)$, thus

$$\begin{aligned}\mathbb{E}\left[f(\mathbf{X}_0, \mathbf{X}_2)\right] &= \int f(\mathbf{x}_0, \mathbf{x}_2)\mathcal{N}\left(\mathbf{x}_1; \mathbf{x}_0, \sigma_1^2\mathrm{I}\right)\mathcal{N}\left(\mathbf{x}_2; \mathbf{x}_1, \sigma_2^2\mathrm{I}\right)\mu(\mathrm{d}\mathbf{x}_0)\mathrm{d}\mathbf{x}_0\mathrm{d}\mathbf{x}_1 \\ &= \int f(\mathbf{x}_0, \mathbf{x}_2)\mathcal{N}\left(\mathbf{x}_2; \mathbf{x}_0, (\sigma_2^2 + \sigma_1^2)\mathrm{I}\right)\mu(\mathrm{d}\mathbf{x}_0)\mathrm{d}\mathbf{x}_2 \,.\end{aligned} \tag{12}$$

Thus, by proposition 0.1, $\mathcal{L}(\mathbf{X}_2|\mathbf{X}_0) = \mathcal{N}\left(\mathbf{X}_0, (\sigma_1^2 + \sigma_2^2)\mathrm{I}\right)$.

2. What is $\mathcal{L}\left(\mathbf{X}_1|\mathbf{X}_0,\mathbf{X}_2\right)$?

Solution: We start by noting that

$$\begin{aligned}
&\int f(\mathbf{x}_0,\mathbf{x}_1,\mathbf{x}_2)\mathcal{N}\left(\mathbf{x}_1;\mathbf{x}_0,\sigma_1^2\,\mathrm{I}\right)\mathcal{N}\left(\mathbf{x}_2;\mathbf{x}_1,\sigma_2^2\,\mathrm{I}\right)\mu(\mathrm{d}\mathbf{x}_0)\mathrm{d}\mathbf{x}_1\mathrm{d}\mathbf{x}_2\\
&=\int f(\mathbf{x}_0,\mathbf{x}_1,\mathbf{x}_2)\frac{\mathcal{N}\left(\mathbf{x}_1;\mathbf{x}_0,\sigma_1^2\,\mathrm{I}\right)\mathcal{N}\left(\mathbf{x}_2;\mathbf{x}_1,\sigma_2^2\,\mathrm{I}\right)}{\mathcal{N}\left(\mathbf{x}_2;\mathbf{x}_0,(\sigma_1^2+\sigma_2^2)\,\mathrm{I}\right)}\mathcal{N}\left(\mathbf{x}_2;\mathbf{x}_0,(\sigma_2^2+\sigma_1^2)\,\mathrm{I}\right)\mu(\mathrm{d}\mathbf{x}_0)\mathrm{d}\mathbf{x}_1\mathrm{d}\mathbf{x}_2\\
&=\int f(\mathbf{x}_0,\mathbf{x}_1,\mathbf{x}_2)\mathcal{N}\left(\mathbf{x}_1;\mathbf{x}_0\left(1-\frac{\sigma_1^2}{\sigma_2^2+\sigma_1^2}\right)+\mathbf{x}_2\frac{\sigma_1^2}{\sigma_2^2+\sigma_1^2},\frac{\sigma_1^2\sigma_2^2}{\sigma_2^2+\sigma_1^2}\,\mathrm{I}\right)\mathcal{N}\left(\mathbf{x}_2;\mathbf{x}_0,(\sigma_2^2+\sigma_1^2)\,\mathrm{I}\right)\mu(\mathrm{d}\mathbf{x}_0)\mathrm{d}\mathbf{x}_1\mathrm{d}\mathbf{x}_2\,.
\end{aligned}$$

We apply proposition 0.1 with $\mathbf{X}=\mathbf{X}_1$ and $\mathbf{Y}=(\mathbf{X}_0,\mathbf{X}_2)$. Thus,

$$\mathcal{L}\left(\mathbf{X}_1|\mathbf{X}_0,\mathbf{X}_2\right)(\mathbf{x}_0,\mathbf{x}_1,\mathrm{A})=\int 1_{\mathrm{A}}\left(\mathbf{x}_1\right)\mathcal{N}\left(\mathbf{x}_1;\mathbf{x}_0\left(1-\frac{\sigma_1^2}{\sigma_2^2+\sigma_1^2}\right)+\mathbf{x}_2\frac{\sigma_1^2}{\sigma_2^2+\sigma_1^2},\frac{\sigma_1^2\sigma_2^2}{\sigma_2^2+\sigma_1^2}\,\mathrm{I}\right)\mathrm{d}\mathbf{x}_1\,. \quad (13)$$

3. Give an expression for $(\mathbf{m},\Sigma)\in\arg\min \mathrm{KL}\left(\mathcal{L}\left(\mathbf{X}_1|\mathbf{X}_2\right)(\mathbf{x}_2,\cdot)\|\mathcal{N}\left(\mathbf{m},\Sigma\right)\right)$ that depends only on the moments of $\mathcal{L}\left(\mathbf{X}_0|\mathbf{X}_2\right)$.

Solution: We know that $\mathbf{m}=\mathbb{E}\left[\mathbf{X}_1|\mathbf{X}_2\right]$ and that $\Sigma=\mathbb{V}\left[\mathbf{X}_1|\mathbf{X}_2\right]$. By the total expectation formula we have

$$\mathbf{m}=\mathbb{E}\left[\mathbf{X}_1|\mathbf{X}_2=\mathbf{x}_2\right]=\mathbb{E}\left[\mathbb{E}\left[\mathbf{X}_1|\mathbf{X}_2,\mathbf{X}_0\right]|\mathbf{X}_2=\mathbf{x}_2\right]=\mathbb{E}\left[\mathbf{X}_0|\mathbf{X}_2=\mathbf{x}_2\right]\left(1-\frac{\sigma_1^2}{\sigma_2^2+\sigma_1^2}\right)+\mathbf{x}_2\frac{\sigma_1^2}{\sigma_2^2+\sigma_1^2}\,. \quad (14)$$

By the total variation formula, we have that

$$\begin{aligned}
\Sigma&=\mathbb{E}\left[\mathbb{V}\left[\mathbf{X}_1|\mathbf{X}_2,\mathbf{X}_0\right]|\mathbf{X}_2=\mathbf{x}_2\right]+\mathbb{V}\left[\mathbb{E}\left[\mathbf{X}_1|\mathbf{X}_2,\mathbf{X}_0\right]|\mathbf{X}_2=\mathbf{x}_2\right] && (15)\\
&=\frac{\sigma_1^2\sigma_2^2}{\sigma_2^2+\sigma_1^2}\,\mathrm{I}+\left(1-\frac{\sigma_1^2}{\sigma_2^2+\sigma_1^2}\right)^2\mathbb{V}\left[\mathbf{X}_0|\mathbf{X}_2=\mathbf{x}_2\right] && (16)\\
&=\frac{\sigma_1^2\sigma_2^2}{\sigma_2^2+\sigma_1^2}\,\mathrm{I}+\left(\frac{\sigma_1^2\sigma_2^2}{\sigma_2^2+\sigma_1^2}\right)^2\frac{\mathbb{V}\left[\mathbf{X}_0|\mathbf{X}_2=\mathbf{x}_2\right]}{\sigma_1^4}\,. && (17)
\end{aligned}$$

4. In which case can we neglect the terms that depend on $\mathbb{V}\left[\mathbf{X}_0|\mathbf{X}_2\right]$?

3 Generalizing and DDPM [2].

We now consider a generalized version of the previous exercise, where

$$\mathbf{X}_0\sim\mu\,,\quad \mathbf{X}_k=\mathbf{X}_{k-1}+\sigma\mathbf{Z}_k\,,\quad \mathbf{Z}_k\perp\!\!\!\perp\mathbf{Z}_\ell\perp\!\!\!\perp\mathbf{X}_0\quad\text{for}\quad k\neq\ell\,, \quad (18)$$

and $k\in\mathbb{N}$. We suppose for simplicity that $\mathbb{E}\left[\mathbf{X}_0\right]=0$ and $\mathbb{V}\left[\mathbf{X}_0\right]=\mathrm{I}$.

1. Calculate an upper bound for $\mathrm{KL}\left(\mathcal{L}\left(\mathbf{X}_k\right)\middle\|\mathcal{N}\left(0,(k\sigma^2+1)\,\mathrm{I}\right)\right)$.

Solution: We consider the joint laws

$$\mu=\mathcal{L}\left(\mathbf{X}_0,\mathbf{X}_k\right)\qquad\text{and}\qquad\nu=\mathrm{p}_{\mathrm{data}}\otimes\mathcal{N}\left(0,(k\sigma^2+1)\,\mathrm{I}\right)\,. \quad (19)$$

Then, by the data processing inequality,

$$\begin{aligned}
\mathrm{KL}\left(\mathcal{L}\left(\mathbf{X}_k\right)\middle\|\mathcal{N}\left(0,(k\sigma^2+1)\,\mathrm{I}\right)\right)&\leq\mathrm{KL}\left(\mu\|\nu\right)=\mathbb{E}\left[\log\left(\frac{\mathcal{N}\left(\mathbf{X}_k;\mathbf{X}_0,k\sigma^2\,\mathrm{I}\right)}{\mathcal{N}\left(\mathbf{X}_k;0,(k\sigma^2+1)\,\mathrm{I}\right)}\right)\right]\\
&=\frac{1}{2}\left[d\left(\frac{k\sigma^2}{k\sigma^2+1}+\log\left(\frac{k\sigma^2}{k\sigma^2+1}\right)-1\right)+\frac{\mathbb{E}\left[\left\|\mathbf{X}_0\right\|^2\right]}{k\sigma^2+1}\right]\,.
\end{aligned}$$

2. Suppose that you known for all k the functions $\mathbb{E}\left[\mathbf{X}_0|\mathbf{X}_k\right]$ and $\mathbb{V}\left[\mathbf{X}_0|\mathbf{X}_k\right]$. Can you propose a Markov chain $(\overleftarrow{\mathbf{X}}_k)_{k=K}^0$ such that one might expect that

$$\mathrm{KL}\left(\mathcal{L}\left(\mathbf{X}_k\right)\middle\|\mathcal{L}\left(\overleftarrow{\mathbf{X}}_k\right)\right) \approx 0 \tag{20}$$

for all k?

Solution: Again, we use the data processing inequality to show that

$$\mathrm{KL}\left(\mathcal{L}\left(\mathbf{X}_k\right)\middle\|\mathcal{L}\left(\overleftarrow{\mathbf{X}}_k\right)\right) \leq \mathrm{KL}\left(\mathcal{L}\left(\mathbf{X}_k|\mathbf{X}_{k+1}\right)\middle\|\mathcal{L}\left(\overleftarrow{\mathbf{X}}_k|\overleftarrow{\mathbf{X}}_{k+1}\right)\right).$$

Furthermore, we know by the previous exercise that

$$\begin{aligned}
\mathbf{m}_{k+1}(\mathbf{x}_{k+1}) &= \left(1-\frac{k}{k+1}\right)\mathbb{E}\left[\mathbf{X}_0|\mathbf{X}_{k+1}=\mathbf{x}_{k+1}\right] + \frac{k}{k+1}\mathbf{x}_{k+1} \\
\mathbf{\Sigma}_{k+1}(\mathbf{x}_{k+1}) &= \left(1-\frac{k}{k+1}\right)^2\mathbb{V}\left[\mathbf{X}_0|\mathbf{X}_{k+1}=\mathbf{x}_{k+1}\right] + \frac{k}{k+1}\sigma^2\,\mathrm{I}\,.
\end{aligned}$$

is the minimizer of

$$\arg\min_{\mathbf{m},\Sigma} \mathrm{KL}\left(\mathcal{L}\left(\mathbf{X}_k|\mathbf{X}_{k+1}=\mathbf{x}_{k+1}\right)\|\mathcal{N}\left(\mathbf{m}_{k+1},\Sigma_{k+1}\right)\right) \approx 0$$

for every $\mathbf{x}_{k+1}$. Thus, if we take as the kernel for the Markov Chain

$$\mathsf{Q}(\mathbf{x}_{k+1},\cdot) = \mathcal{N}\left(\cdot;\mathbf{m}_{k+1}(\mathbf{x}_{k+1}),\Sigma_{k+1}(\mathbf{x}_{K+1})\right)$$

and if $\mathrm{KL}\left(\mathcal{L}\left(\mathbf{X}_{k+1}\right)\middle\|\mathcal{L}\left(\overleftarrow{\mathbf{X}}_{k+1}\right)\right) \approx 0$, we will keep having small errors. This is of course a bit "hand wavy".

3. Provide an upper bound on $\mathrm{KL}\left(\mathcal{L}\left(\mathbf{X}_0\right)\middle\|\mathcal{L}\left(\overleftarrow{\mathbf{X}}_0\right)\right)$.

Solution: We start by the data processing inequality.

$$\begin{aligned}
\mathrm{KL}\left(\mathcal{L}\left(\mathbf{X}_0\right)\middle\|\mathcal{L}\left(\overleftarrow{\mathbf{X}_0}\right)\right) &\leq \mathrm{KL}\left(\mathcal{L}\left(\mathbf{X}_0,\cdots,\mathbf{X}_K\right)\middle\|\mathcal{L}\left(\overleftarrow{\mathbf{X}_0},\cdots,\overleftarrow{\mathbf{X}_K}\right)\right) \\
&= \mathbb{E}\left[\log\left(\prod_{k=0}^{K-1}\frac{p_{k|k+1}(\mathbf{X}_k|\mathbf{X}_{k+1})}{\mathcal{N}\left(\mathbf{X}_k;\mathbf{m}_{k+1}(\mathbf{X}_{k+1}),\Sigma_{k+1}(\mathbf{X}_{k+1})\right)}\right)\frac{p_K(\mathbf{X}_k)}{\mathcal{N}\left(0,(k\sigma^2+1)\,\mathrm{I}\right)}\right] \\
&\leq \sum_{k=0}^{K-1}\mathrm{KL}\left(\mathcal{L}\left(\mathbf{X}_k|\mathbf{X}_{k+1}\right)\|\mathcal{N}\left(\mathbf{m}_{k+1}(\mathbf{X}_{k+1}),\Sigma_{k+1}(\mathbf{X}_{k+1})\right)\right) + \mathrm{KL}\left(\mathcal{L}\left(\mathbf{X}_K\right)\middle\|\mathcal{L}\left(\overleftarrow{X}_K\right)\right)
\end{aligned}$$

It is now enough to upper bound $\mathrm{KL}\left(\mathcal{L}\left(\mathbf{X}_k|\mathbf{X}_{k+1}\right)\|\mathcal{N}\left(\mathbf{m}_{k+1}(\mathbf{X}_{k+1}),\Sigma_{k+1}(\mathbf{X}_{k+1})\right)\right)$ as the other term is already upper bounded by question 3.1.

To do so, we will use again the data processing inequality. Consider $\mu_{\mathbf{x}_{k+1}} = \mathcal{L}\left(\mathbf{X}_k,\mathbf{X}_0|\mathbf{X}_{k+1}=\mathbf{x}_{k+1}\right)$ and $\nu_{\mathbf{x}_{k+1}} = \mathcal{N}\left(\mathbf{m}_{k+1}(\mathbf{x}_{k+1}),\Sigma_{k+1}(\mathbf{x}_{k+1})\right)\otimes\mathcal{L}\left(\mathbf{X}_0|\mathbf{X}_{k+1}=\mathbf{x}_{k+1}\right)$.

Then,

$$\begin{aligned}
\mathrm{KL}\left(\mu_{\mathbf{x}_{k+1}}\middle\|\nu_{\mathbf{x}_{k+1}}\right) = \int\log\left(\frac{p_{k|k+1,0}(\mathbf{x}_k|\mathbf{x}_k+1,\mathbf{x}_0)p_{0|k+1}(\mathbf{x}_0|\mathbf{x}_{k+1})}{\mathcal{N}\left(\mathbf{m}_{k+1}(\mathbf{x}_{k+1}),\Sigma_{k+1}(\mathbf{x}_{k+1})\right)p_{0|k+1}(\mathbf{x}_0|\mathbf{x}_{k+1})}\right) \\
p_{k|k+1,0}(\mathbf{x}_k|\mathbf{x}_k+1,\mathbf{x}_0)p_{0|k+1}(\mathbf{x}_0|\mathbf{x}_{k+1})\mathrm{d}\mathbf{x}_0\mathrm{d}\mathbf{x}_k
\end{aligned}$$

where $p_{k|k+1,0}$ is the density of $\mathcal{L}\left(\mathbf{X}_k|\mathbf{X}_{k+1},\mathbf{X}_0\right)$ and $p_{k|0}$ the density of $\mathcal{L}\left(\mathbf{X}_0|\mathbf{X}_k\right)$. Thus we have that

$$\begin{aligned}
&\mathrm{KL}\left(\mu_{\mathbf{x}_{k+1}}\middle\|\nu_{\mathbf{x}_{k+1}}\right) \\
&= \mathbb{E}\left[\mathrm{KL}\left(\mathcal{N}\left(\left(1-\frac{k}{k+1}\right)\mathbf{X}_0+\frac{k}{k+1}\mathbf{x}_{k+1},\frac{k}{k+1}\sigma^2\,\mathrm{I}\right)\middle\|\mathcal{N}\left(\mathbf{m}_{k+1}(\mathbf{x}_{k+1}),\Sigma_{k+1}(\mathbf{x}_{k+1})\right)\right)\middle|\mathbf{X}_{k+1}=\mathbf{x}_{k+1}\right].
\end{aligned}$$

The rest follows from the KL formula for two Gaussians.

4. What if you know only $\mathbb{E}\left[\mathbf{X}_0|\mathbf{X}_k\right]$? Provide an upper bound for $\mathrm{KL}\left(\mathcal{L}\left(\mathbf{X}_0\right)\middle\|\mathcal{L}\left(\overleftarrow{\mathbf{X}}_0\right)\right)$ in this case.

4 References

References

[1] R. Douc et al. *Markov Chains*. Springer Series in Operations Research and Financial Engineering. Springer, Cham, 2018. xviii+757. ISBN: 978-3-319-97703-4 978-3-319-97704-1. DOI: 10.1007/978-3-319-97704-1. URL: https://doi.org/10.1007/978-3-319-97704-1.

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising Diffusion Probabilistic Models". In: *Advances in Neural Information Processing Systems*. Vol. 33. 2020, pp. 6840–6851.