数理解説：式 (20.4.1) の証明

ここでは、式 (20.4.1) の計算を紹介します。途中を省略せずに記すと、以下のとおりになります。単純化のため、ここではデータが $\boldsymbol{x}^{(0)}$ の 1 つのみの場合で記します。また、$q(\boldsymbol{x}^{(1:T)}|\boldsymbol{x}^{(0)})d\boldsymbol{x}^{(1:T)}$ を $dq(\boldsymbol{x}^{(1:T)})$ で略記します。

なお、この手の複雑な計算では、いきなり意味を理解することは難しいです。理解を試みる方は、まず全ての式変形が正しいことを自分の頭で確認しましょう。その次に、最初と最後の式を見比べて、両者が等しい理由がなにか考えてみると良いでしょう。

$$
\begin{aligned}
&F = F(\boldsymbol{\theta}, X) = \int_{-\infty}^{\infty} \log\left(\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(0:T)})}{q(\boldsymbol{x}^{(1:T)} \mid \boldsymbol{x}^{(0)})}\right) dq(\boldsymbol{x}^{(1:T)}) \\
&= \int_{-\infty}^{\infty} \log\left(p(\boldsymbol{x}^{(T)}) \times \prod_{1\le t\le T} \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(t-1)} \mid \boldsymbol{x}^{(t)}, t)}{q(\boldsymbol{x}^{(t)} \mid \boldsymbol{x}^{(t-1)})}\right) dq(\boldsymbol{x}^{(1:T)}) \\
&= \int_{-\infty}^{\infty} \log\left(p(\boldsymbol{x}^{(T)}) \times \prod_{1\le t\le T} \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(t-1)} \mid \boldsymbol{x}^{(t)}, t)}{q(\boldsymbol{x}^{(t)} \mid \boldsymbol{x}^{(t-1)}, \boldsymbol{x}^{(0)})}\right) dq(\boldsymbol{x}^{(1:T)}) \\
&= \int_{-\infty}^{\infty} \log\left(p(\boldsymbol{x}^{(T)}) \times \prod_{1\le t\le T} \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(t-1)} \mid \boldsymbol{x}^{(t)}, t) q(\boldsymbol{x}^{(t-1)} \mid \boldsymbol{x}^{(0)})}{q(\boldsymbol{x}^{(t-1)} \mid \boldsymbol{x}^{(t)}, \boldsymbol{x}^{(0)}) q(\boldsymbol{x}^{(t)} \mid \boldsymbol{x}^{(0)})}\right) dq(\boldsymbol{x}^{(1:T)}) \\
&= \int_{-\infty}^{\infty} \log\left(p(\boldsymbol{x}^{(T)}) \times \prod_{2\le t\le T} \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(t-1)} \mid \boldsymbol{x}^{(t)}, t) q(\boldsymbol{x}^{(t-1)} \mid \boldsymbol{x}^{(0)})}{q(\boldsymbol{x}^{(t-1)} \mid \boldsymbol{x}^{(t)}, \boldsymbol{x}^{(0)}) q(\boldsymbol{x}^{(t)} \mid \boldsymbol{x}^{(0)})} \times \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(0)} \mid \boldsymbol{x}^{(1)}, 1)}{q(\boldsymbol{x}^{(1)} \mid \boldsymbol{x}^{(0)})}\right) dq(\boldsymbol{x}^{(1:T)}) \\
&= \int_{-\infty}^{\infty} \log\left(p(\boldsymbol{x}^{(T)}) \times \prod_{2\le t\le T} \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(t-1)} \mid \boldsymbol{x}^{(t)}, t)}{q(\boldsymbol{x}^{(t-1)} \mid \boldsymbol{x}^{(t)}, \boldsymbol{x}^{(0)})} \times \frac{q(\boldsymbol{x}^{(1)} \mid \boldsymbol{x}^{(0)})}{q(\boldsymbol{x}^{(T)} \mid \boldsymbol{x}^{(0)})} \times \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(0)} \mid \boldsymbol{x}^{(1)}, 1)}{q(\boldsymbol{x}^{(1)} \mid \boldsymbol{x}^{(0)})}\right) dq(\boldsymbol{x}^{(1:T)}) \\
&= \int_{-\infty}^{\infty} \left(\log \frac{p(\boldsymbol{x}^{(T)})}{q(\boldsymbol{x}^{(T)} \mid \boldsymbol{x}^{(0)})} + \sum_{2\le t\le T} \log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(t-1)} \mid \boldsymbol{x}^{(t)}, t)}{q(\boldsymbol{x}^{(t-1)} \mid \boldsymbol{x}^{(t)}, \boldsymbol{x}^{(0)})} + \log p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(0)} \mid \boldsymbol{x}^{(1)}, 1)\right) dq(\boldsymbol{x}^{(1:T)}) \\
&= -KL\left[q(\boldsymbol{x}^{(T)} \mid \boldsymbol{x}^{(0)}) \,\|\, p(\boldsymbol{x}^{(T)})\right] - \sum_{2\le t\le T} E_{q(\boldsymbol{x}^{(t)} \mid \boldsymbol{x}^{(0)})}\left[KL\left[q(\boldsymbol{x}^{(t-1)} \mid \boldsymbol{x}^{(t)}, \boldsymbol{x}^{(0)}) \,\|\, p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(t-1)} \mid \boldsymbol{x}^{(t)}, t)\right]\right] \\
&\qquad + \int_{-\infty}^{\infty} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(0)} \mid \boldsymbol{x}^{(1)}, 1) dq(\boldsymbol{x}^{(1:T)})
\end{aligned}
$$

ここで、2 行目から 3 行目への式変形では、$q(\boldsymbol{x}^{(t)}|\boldsymbol{x}^{(t-1)})$ が $q(\boldsymbol{x}^{(t)}|\boldsymbol{x}^{(t-1)}, \boldsymbol{x}^{(0)})$ へ変わっています。実は、以下に記す理由で、この 2 者の値は等しいです。拡散過程では、$\boldsymbol{x}^{(t-1)}$ が与えられれば、もともとのデータ $\boldsymbol{x}^{(0)}$ が何であっても、$\boldsymbol{x}^{(t)}$ は確率分布 $q(\boldsymbol{x}^{(t)}|\boldsymbol{x}^{(t-1)})$ に従って決まります。ですので、$\boldsymbol{x}^{(0)}$ の情報が増えても確率分布は $q(\boldsymbol{x}^{(t)}|\boldsymbol{x}^{(t-1)})$ のままであり、$q(\boldsymbol{x}^{(t)}|\boldsymbol{x}^{(t-1)}, \boldsymbol{x}^{(0)}) = q(\boldsymbol{x}^{(t)}|\boldsymbol{x}^{(t-1)})$ が成立しま

す。これを、拡散過程の**マルコフ性(Markov property)**と言います。

4行目への式変形は、条件付きベイズの定理（P405の式19.3.7）を用いた $q(\boldsymbol{x}^{(t)}|\boldsymbol{x}^{(t-1)},\boldsymbol{x}^{(0)})=\dfrac{q(\boldsymbol{x}^{(t-1)}|\boldsymbol{x}^{(t)},\boldsymbol{x}^{(0)})q(\boldsymbol{x}^{(t)}|\boldsymbol{x}^{(0)})}{q(\boldsymbol{x}^{(t-1)}|\boldsymbol{x}^{(0)})}$ を利用しました。5行目への式変形では、確率の積の範囲を $1\leq t\leq T$ から $2\leq t\leq T$ へ変更し、$t=1$ の場合を右に添えました。6行目への式変形では、積の中で $q(\boldsymbol{x}^{(t)}|\boldsymbol{x}^{(0)})$ が打ち消し合うので、$q(\boldsymbol{x}^{(1)}|\boldsymbol{x}^{(0)})$ と $q(\boldsymbol{x}^{(T)}|\boldsymbol{x}^{(0)})$ だけを残して積の外へ出しました。

最後の式変形でKL divergenceが登場する計算は、次のとおりです。まず $\boldsymbol{x}^{(1)}$ から $\boldsymbol{x}^{(t-2)}$ まで、次に $\boldsymbol{x}^{(T)}$ から $\boldsymbol{x}^{(t+1)}$ まで、最後に $\boldsymbol{x}^{(t-1)}, \boldsymbol{x}^{(t)}$ の順で積分すると、

$$
\begin{aligned}
&\int_{-\infty}^{\infty}\log\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(t-1)}|\boldsymbol{x}^{(t)},t)}{q(\boldsymbol{x}^{(t-1)}|\boldsymbol{x}^{(t)},\boldsymbol{x}^{(0)})}dq(\boldsymbol{x}^{(1:T)})=\int_{-\infty}^{\infty}\log\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(t-1)}|\boldsymbol{x}^{(t)},t)}{q(\boldsymbol{x}^{(t-1)}|\boldsymbol{x}^{(t)},\boldsymbol{x}^{(0)})}q(\boldsymbol{x}^{(1:T)}|\boldsymbol{x}^{(0)})d\boldsymbol{x}^{(1:T)}\\
&=\int_{-\infty}^{\infty}\log\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(t-1)}|\boldsymbol{x}^{(t)},t)}{q(\boldsymbol{x}^{(t-1)}|\boldsymbol{x}^{(t)},\boldsymbol{x}^{(0)})}\prod_{1\leq\tau\leq T}q(\boldsymbol{x}^{(\tau)}|\boldsymbol{x}^{(\tau-1)})d\boldsymbol{x}^{(1:T)}\\
&=\int_{-\infty}^{\infty}\log\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(t-1)}|\boldsymbol{x}^{(t)},t)}{q(\boldsymbol{x}^{(t-1)}|\boldsymbol{x}^{(t)},\boldsymbol{x}^{(0)})}\prod_{3\leq\tau\leq T}q(\boldsymbol{x}^{(\tau)}|\boldsymbol{x}^{(\tau-1)})\times q(\boldsymbol{x}^{(2)}|\boldsymbol{x}^{(1)})q(\boldsymbol{x}^{(1)}|\boldsymbol{x}^{(0)})d\boldsymbol{x}^{(1)}d\boldsymbol{x}^{(2:T)}\\
&=\int_{-\infty}^{\infty}\log\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(t-1)}|\boldsymbol{x}^{(t)},t)}{q(\boldsymbol{x}^{(t-1)}|\boldsymbol{x}^{(t)},\boldsymbol{x}^{(0)})}\prod_{3\leq\tau\leq T}q(\boldsymbol{x}^{(\tau)}|\boldsymbol{x}^{(\tau-1)})\times q(\boldsymbol{x}^{(2)}|\boldsymbol{x}^{(0)})d\boldsymbol{x}^{(2:T)}\\
&=\cdots\\
&=\int_{-\infty}^{\infty}\log\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(t-1)}|\boldsymbol{x}^{(t)},t)}{q(\boldsymbol{x}^{(t-1)}|\boldsymbol{x}^{(t)},\boldsymbol{x}^{(0)})}\prod_{t\leq\tau\leq T}q(\boldsymbol{x}^{(\tau)}|\boldsymbol{x}^{(\tau-1)})\times q(\boldsymbol{x}^{(t-1)}|\boldsymbol{x}^{(0)})d\boldsymbol{x}^{(t-1:T)}\\
&=\int_{-\infty}^{\infty}\log\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(t-1)}|\boldsymbol{x}^{(t)},t)}{q(\boldsymbol{x}^{(t-1)}|\boldsymbol{x}^{(t)},\boldsymbol{x}^{(0)})}\prod_{t+1\leq\tau\leq T}q(\boldsymbol{x}^{(\tau)}|\boldsymbol{x}^{(\tau-1)})\times q(\boldsymbol{x}^{(t)}|\boldsymbol{x}^{(t-1)})q(\boldsymbol{x}^{(t-1)}|\boldsymbol{x}^{(0)})d\boldsymbol{x}^{(t-1:T)}\\
&=\int_{-\infty}^{\infty}\log\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(t-1)}|\boldsymbol{x}^{(t)},t)}{q(\boldsymbol{x}^{(t-1)}|\boldsymbol{x}^{(t)},\boldsymbol{x}^{(0)})}\prod_{t+1\leq\tau\leq T}q(\boldsymbol{x}^{(\tau)}|\boldsymbol{x}^{(\tau-1)})\times q(\boldsymbol{x}^{(t)},\boldsymbol{x}^{(t-1)}|\boldsymbol{x}^{(0)})d\boldsymbol{x}^{(t-1:T)}\\
&=\int_{-\infty}^{\infty}\log\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(t-1)}|\boldsymbol{x}^{(t)},t)}{q(\boldsymbol{x}^{(t-1)}|\boldsymbol{x}^{(t)},\boldsymbol{x}^{(0)})}\prod_{t+1\leq\tau\leq T}q(\boldsymbol{x}^{(\tau)}|\boldsymbol{x}^{(\tau-1)})\times q(\boldsymbol{x}^{(t-1)}|\boldsymbol{x}^{(t)},\boldsymbol{x}^{(0)})q(\boldsymbol{x}^{(t)}|\boldsymbol{x}^{(0)})d\boldsymbol{x}^{(t-1:T)}\\
&=\int_{-\infty}^{\infty}\log\frac{p_{\theta}(\boldsymbol{x}^{(t-1)}|\boldsymbol{x}^{(t)},t)}{q(\boldsymbol{x}^{(t-1)}|\boldsymbol{x}^{(t)},\boldsymbol{x}^{(0)})}q(\boldsymbol{x}^{(t-1)}|\boldsymbol{x}^{(t)},\boldsymbol{x}^{(0)})q(\boldsymbol{x}^{(t)}|\boldsymbol{x}^{(0)})\,d\boldsymbol{x}^{(t-1:t)}\\
&=-\int_{-\infty}^{\infty}KL\left[q(\boldsymbol{x}^{(t-1)}|\boldsymbol{x}^{(t)},\boldsymbol{x}^{(0)})\,\|\,p_{\theta}(\boldsymbol{x}^{(t-1)}|\boldsymbol{x}^{(t)},t)\right]q(\boldsymbol{x}^{(t)}|\boldsymbol{x}^{(0)})\,d\boldsymbol{x}^{(t)}\\
&=-E_{q(\boldsymbol{x}^{(t)}|\boldsymbol{x}^{(0)})}\left[KL\left[q(\boldsymbol{x}^{(t-1)}|\boldsymbol{x}^{(t)},\boldsymbol{x}^{(0)})\,\|\,p_{\theta}(\boldsymbol{x}^{(t-1)}|\boldsymbol{x}^{(t)},t)\right]\right]
\end{aligned}
$$

と計算できます。2行目から4行目への計算で $\boldsymbol{x}^{(1)}$ での積分を実行し、結果

として $q(\boldsymbol{x}^{(2)}|\boldsymbol{x}^{(1)}) \times q(\boldsymbol{x}^{(1)}|\boldsymbol{x}^{(0)})$ が $q(\boldsymbol{x}^{(2)}|\boldsymbol{x}^{(0)})$ へ変化します。これを繰り返した後、7 行目から 9 行目では条件付き確率の公式が用いられています。10 行目への式変形では、$\boldsymbol{x}^{(T)}$ から順に $\boldsymbol{x}^{(t+1)}$ まで積分を行いました。11 行目への式変形は $\boldsymbol{x}^{(t-1)}$ での積分で、これは KL divergence の定義そのものです。12 行目への式変形では、積分を期待値に書き換えました。

数理解説：$q(\boldsymbol{x}^{(t-1)}|\boldsymbol{x}^{(t)},\boldsymbol{x}^{(0)})$ の計算

最後に、$q(\boldsymbol{x}^{(t-1)}|\boldsymbol{x}^{(t)},\boldsymbol{x}^{(0)})$ を計算します。元々、$\boldsymbol{x}^{(0)}$ にノイズを加えて $\boldsymbol{x}^{(t-1)}$ を作り、$\boldsymbol{x}^{(t-1)}$ に更にノイズを加えて $\boldsymbol{x}^{(t)}$ を作っていました。そのため、$\boldsymbol{x}^{(t-1)} - \sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}^{(0)}$ は平均 $\mathbf{0}$、分散共分散行列 $(1-\bar{\alpha}_{t-1}) \times 1$ の多変量正規分布に従い、$\boldsymbol{x}^{(t-1)} - \frac{1}{\sqrt{\alpha_t}}\boldsymbol{x}^{(t)}$ 平均 $\mathbf{0}$、分散共分散行列 $\frac{1-\alpha_t}{\alpha_t} \times 1$ の多変量正規分布に従います。19.3 節の正規分布の引き合いの定理で、$\boldsymbol{X} = \sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}^{(0)}$, $A_X = \frac{1}{1-\bar{\alpha}_{t-1}} \times 1$, $\boldsymbol{Y} = \boldsymbol{x}^{(t-1)}$, $\boldsymbol{Z} = \frac{1}{\sqrt{\alpha_t}}\boldsymbol{x}^{(t)}$, $A_Z = \frac{\alpha_t}{1-\alpha_t} \times 1$ と設定すれば、$q(\boldsymbol{x}^{(t-1)}|\boldsymbol{x}^{(t)},\boldsymbol{x}^{(0)})$ の平均 $A_t\boldsymbol{x}^{(0)} + B_t\boldsymbol{x}^{(t)}$ と分散共分散行列 $C_t \times 1$ は、次の式で得られます。

$$
\text{平均}: \frac{\frac{1}{1-\bar{\alpha}_{t-1}}\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}^{(0)} + \frac{\alpha_t}{1-\alpha_t}\frac{1}{\sqrt{\alpha_t}}\boldsymbol{x}^{(t)}}{\frac{1}{1-\bar{\alpha}_{t-1}} + \frac{\alpha_t}{1-\alpha_t}} = A_t\boldsymbol{x}^{(0)} + B_t\boldsymbol{x}^{(t)}
$$

$$
\text{分散}: \left(\frac{1}{\frac{1}{1-\bar{\alpha}_{t-1}} + \frac{\alpha_t}{1-\alpha_t}}\right) \times 1 = C_t \times 1 \tag{20.4.5}
$$

見た目こそ複雑ですが、それは問題ではありません。学習不要でこれらの数値を計算できることが重要なのです。