拡散モデル（Diffusion Model）を理解するために

20B01392 松本侑真

2023 年 9 月 22 日

概要

近年、機械学習を応用した生成系 AI がブームである。生成系 AI を使うことで、プロンプトに入力したテキストデータから画像データなどを出力することができる。拡散モデルを用いた有名な生成系 AI の Stable Diffusion では、無料でプロンプトに入力した文字（例：かわいい犬）から、かわいい犬の画像を出力することができる。このような生成系 AI で用いられる拡散モデルについて理解するための基礎を説明する。

目次

1　生成モデルでは何を行っているのか

生成モデル（拡散モデル）は、事前に与えられた教師データを元にモデルを学習することで、未知の入力に対して最適化された出力をアウトプットすることができる。例えば、いろいろな人間の顔データを学習させた生成モデルを用いると、「40 歳のおじさん」や「20 代のアイドル」といった入力を元にして、学習されたモデルで生成した顔画像を出力する。

生成モデルの学習というのは、「モデルの入力から欲しいデータを生み出す確率分布」をいかにして見つけるかということである。すなわち、任意のデータ $\vec{x}$ は、とある 1 つの確率分布 $p(\vec{x})$ に従って生み出されるという仮定のもと、自分のモデルをその $p(\vec{x})$ に限りなく近づけることが生成モデルにおける学習である。すなわち、あらゆるデータが従う確率分布 $p(\vec{x})$ そのものを手に入れることができれば、$p(\vec{x})$ の値を出力するような $\vec{x}$ を見つけることが可能であるだろう。これによって、$p(\vec{x})$ =「20 代のアイドル」を満たす画像 $\vec{x}$ を出力することができる。

学習を行う前に得られているもの

モデルの学習を行うためには、そのための学習データが必要である。n 個の学習データ $\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_n$ が手元にあるとする。このとき、

$$p_0(\vec{x}) = \frac{1}{n}\sum_{i=1}^{n} \delta(\vec{x} - \vec{x}_i) \tag{1.1}$$

を経験分布と呼ぶ。この経験分布は、n が十分に大きいとき、あらゆるデータを生み出す確率分布 $p(x)$ に非常に近いであろう：

$$p_0(\vec{x}) \approx p(\vec{x})\ 。 \tag{1.2}$$

なお、データを連続変数として扱う場合は $\delta(\vec{x} - \vec{x}_i)$ をデルタ関数、離散変数として扱う場合はクロネッカーのデルタとして扱う。

学習のために必要なもの

生成モデルにおいて学習させるための確率分布は自ら用意する必要がある。すなわち、パラメータ θ を用いて、

$$p(\vec{x}) \approx q_\theta(\vec{x}) \tag{1.3}$$

となるようなモデル $q_\theta(\vec{x})$ を見つけたい。モデル $q_\theta(\vec{x})$ の関数形は計算しやすい形を用いて、パラメータ θ を最適化することで $p(\vec{x})$ に近づけるというアプローチを取る。

2　ボルツマンマシン

ボルツマンマシンとは、モデル $q_\theta(\vec{x})$ が

$$q_\theta(\vec{x}) = \frac{1}{Z_\theta}\exp(-E_\theta(\vec{x}))\ 。 \tag{2.1}$$

と表されるものである。このような関数形を Gibbs-Boltzmann 分布（指数関数分布族）と呼ぶ。ここで、規格化定数 Z_θ は

$$Z_\theta = \sum_{\vec{x}} \exp(-E_\theta(\vec{x})) \tag{2.2}$$

である。このように定義された $q_\theta(\vec{x})$ を $p(\vec{x})$ に近づけることが目標である。しかし、$p(\vec{x})$ の形は誰も知らない。今手元にあるのは経験分布 $p_0(\vec{x})$ だけである。そのため、学習データの数 n が十分に多いとして、$q_\theta(\vec{x})$ を経験分布 $p_0(\vec{x})$ に近づけることを行う。

2.1 Kullback-Leibler 情報量

学習を行うためには、2 つの確率分布 $p(\vec{x})$, $q(\vec{x})$ の近さを定義する必要がある。この「近さ」は、データ $\vec{x}$ を入力した際の 2 つの確率分布間の「距離 $D(p \parallel q)$」を測ることで定義することができるだろう。このような関数 $D(p \parallel q)$ はどのように設定しても良いのだが、今回は、Kullback-Leibler divergence（KL 情報量）$D_{\mathrm{KL}}(p \parallel q)$ を導入する：

$$D_{\mathrm{KL}}(p(\vec{x}) \parallel q(\vec{x})) = \sum_{\vec{x}} p(\vec{x}) \ln \frac{p(\vec{x})}{q(\vec{x})} \text{。} \tag{2.3}$$

KL 情報量には

A. $D(p \parallel q) \geq 0$（等号成立は $p = q$ のとき）
B. $D(p \parallel q) \neq D(q \parallel p)$（非対称性）

といった性質がある。特に、KL 情報量は 0 以上であるといった性質を良く用いる。この性質を 2 つの方法で証明してみる。

Gibbs の不等式を用いる方法

Gibbs の不等式とは、$x > 0$ において

$$x - 1 \geq \ln x \Longleftrightarrow \ln \frac{1}{x} \geq 1 - x \tag{2.4}$$

が成立する不等式である。

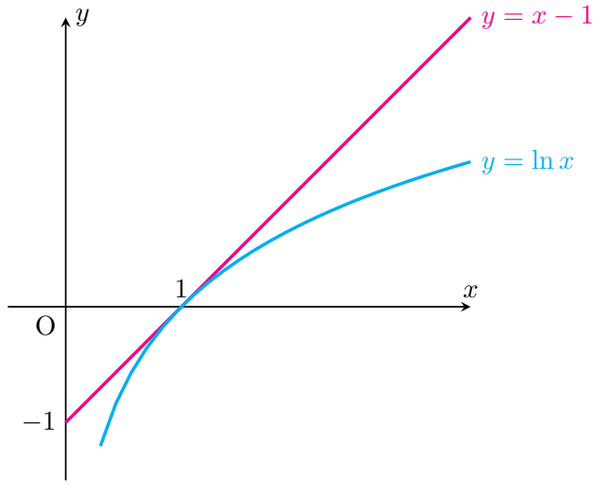

図 1: Gibbs の不等式の直感的な理解

Gibbs の不等式を用いると、

$$D_{\mathrm{KL}}(p \parallel q) = \sum_{\vec{x}} p(\vec{x}) \ln \frac{p(\vec{x})}{q(\vec{x})} \geq \sum_{\vec{x}} p(\vec{x}) \left(1 - \frac{q(\vec{x})}{p(\vec{x})}\right) = \sum_{\vec{x}} p(\vec{x}) - \sum_{\vec{x}} q(\vec{x}) = 0 \tag{2.5}$$

と示される。

Jensen の不等式を用いる方法

Jensen の不等式とは、上に凸な関数 $f(x)$ に対して、

$$f\left(\frac{a+b}{2}\right) \geq \frac{f(a)+f(b)}{2} \tag{2.6}$$

が成立する不等式である。

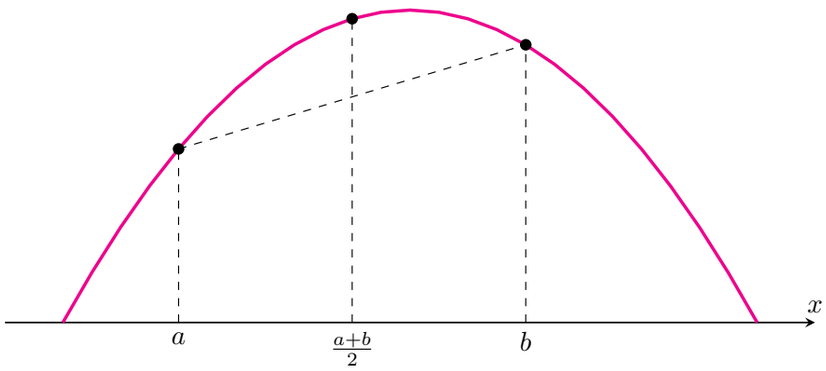

図 2: Jensen の不等式の直感的な理解

Jensen の不等式を用いると、

$$D_{\mathrm{KL}}(p \parallel q) = -\sum_{\vec{x}} p(\vec{x}) \ln \frac{q(\vec{x})}{p(\vec{x})} \geq -\ln\left(\sum_{\vec{x}} p(\vec{x}) \frac{q(\vec{x})}{p(\vec{x})}\right) = 0 \tag{2.7}$$

と示される。

2.2 例：ガウス分布の KL 情報量と学習

モデル $q_\theta(\vec{x})$ をガウス分布とおいて最適化を行う。人間が計算できるのはせいぜいガウス分布程度であるため、これ以上難しいモデルは考えない。まずは $\vec{x}$ が 1 次元のスカラーである場合を考える：

$$q_\theta(x) \propto \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)。 \tag{2.8}$$

規格化定数は

$$Z_\theta = \int_{-\infty}^{\infty} dx\, \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) = \sqrt{2\pi\sigma^2} \tag{2.9}$$

となる。したがって、

$$q_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \tag{2.10}$$

となる。このモデルの未知数は μ と σ^2 であるため、最適化すべきパラメータは $\theta = \{\mu, \sigma^2\}$ となる。次に、KL 情報量を考える。$q_\theta(x)$ を $p(x)$ に近づけることができれば良いのだが、あいにく $p(x)$ の具体形はわからない。そのため、経験分布 $p_0(x) = \sum_{i=1}^n \delta(x - x_i)$ にモデルを近づけることを考える。KL 情報量を計算すると、

$$\begin{aligned} D_{\mathrm{KL}}(p_0 \parallel q_\theta) &= \int dx\, p_0(x)(\ln p_0(x) - \ln q_\theta(x)) \underbrace{=}_{\theta\text{のみ}} - \int dx\, p_0(x)\left(\frac{1}{2\sigma^2}(x-\mu)^2 + \frac{1}{2}\ln\sigma^2\right) \\ &= \frac{1}{2\sigma^2}\frac{1}{n}\sum_{i=1}^n (x_i-\mu)^2 + \frac{1}{2}\ln\sigma^2 \end{aligned} \tag{2.11}$$

となる。KL 情報量を最小化（学習）すれば、経験分布に近いモデルが得られるため、μ と σ^2 で偏微分を行う。

$$\frac{\partial D_{\mathrm{KL}}}{\partial \mu} = -\frac{1}{\sigma^2}\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu) = 0 \qquad \Longleftrightarrow \mu = \frac{1}{n}\sum_{i=1}^{n} x_i \text{（データの平均）} \tag{2.12}$$

$$\frac{\partial D_{\mathrm{KL}}}{\partial \sigma^2} = -\frac{1}{2(\sigma^2)^2}\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 + \frac{1}{2\sigma^2} = 0 \qquad \Longleftrightarrow \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \text{（データの分散）} \tag{2.13}$$

この結果からわかることは、ガウス分布のモデルの平均 μ はデータの平均値に、モデルの分散 σ^2 はデータの分散にすれば、$p(x)$ に近いモデル $q_\theta(x)$ を生成できるということである。

2.3 最尤法（=KL 最小化）

最尤法とは、KL 情報量の最小化と同値なものであるが、良く使われるため紹介する。モデルの学習は、

$$\min_\theta \{D_{\mathrm{KL}}(p_0(\vec{x}) \parallel q_\theta(\vec{x}))\} = \max_\theta \left\{ \sum_{\vec{x}} p(\vec{x}) \ln q_\theta(\vec{x}) \right\} = \max_\theta \left\{ \frac{1}{n}\sum_{i=1}^{n} \ln q_\theta(\vec{x}_i) \right\} \tag{2.14}$$

と変形することができる。$\ln q_\theta(\vec{x}_i)$ を対数尤度関数と呼び、モデルの良さの指標として使われる。対数尤度を最大化することは、KL 情報量の最小化と同値なものである。

2.4 高次元化

先ほどはガウス分布のモデルの 1 次元バージョンの最適化を行った。一般の場合にどのように拡張されるかを見る。すなわち、

$$q_\theta(\vec{x}) \propto \exp\left(-\frac{1}{2}(\vec{x} - \vec{\mu})^\top \Sigma^{-1} (\vec{x} - \vec{\mu}) \right) \tag{2.15}$$

とする。規格化定数は

$$Z_\theta = \int d\vec{x}\, \exp\left(-\frac{1}{2}(\vec{x} - \vec{\mu})^\top \Sigma^{-1} (\vec{x} - \vec{\mu}) \right) \tag{2.16}$$

となる。このままでは規格化定数の計算ができないが、Σ^{-1} の対角化を行うことで計算を進めることができる。ある正方行列 P を用いて、

$$\begin{cases} P^{-1}\Sigma^{-1}P & = \Lambda = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_N \end{pmatrix} \\ P^{-1}(\vec{x} - \vec{\mu}) & = \vec{y} \end{cases} \tag{2.17}$$

と対角化できたとする。Σ は分散共分散行列と呼ばれる。このとき、

$$Z_\theta = \int d\vec{y}\, \exp\left(-\frac{1}{2}\vec{y}^\top \Lambda \vec{y} \right) = \prod_{k=1}^{N} \int d\vec{y}_k\, \exp\left(-\frac{1}{2}\lambda_k y_k^2 \right) = \prod_{k=1}^{N} \sqrt{\frac{2\pi}{\lambda_k}} \tag{2.18}$$

と計算できる。行列の determinant の性質

$$\prod_{k=1}^{N} \lambda_k = \det(\Lambda) = \det\left(\Sigma^{-1}\right) = \frac{1}{\det(\Sigma)} \tag{2.19}$$

を用いると、

$$Z_\theta = \sqrt{(2\pi)^N \det(\Sigma)} \tag{2.20}$$

のように、分散共分散行列 Σ を用いて計算できる。したがって、

$$q_\theta(\vec{x}) = \sqrt{\frac{1}{(2\pi)^N \det(\Sigma)}} \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu})^\top \Sigma^{-1}(\vec{x}-\vec{\mu})\right) \tag{2.21}$$

となる。KL 情報量は、

$$D_{\mathrm{KL}}(p_0 \parallel q_\theta) = \int d\vec{x}\, p_0(\vec{x})(\ln p_0(\vec{x}) - q_\theta(\vec{x})) \underbrace{=}_{\theta\text{のみ}} \frac{1}{2n}\sum_{i=1}^{n}(\vec{x}_i-\vec{\mu})^\top\Sigma^{-1}(\vec{x}_i-\vec{\mu}) + \frac{1}{2}\ln\det(\Sigma) \tag{2.22}$$

と計算できる。KL 情報量の最小化

$$\frac{\partial D_{\mathrm{KL}}}{\partial \vec{\mu}} = -\frac{1}{n}\sum_{i=1}^{n}\Sigma^{-1}(\vec{x}_i-\vec{\mu}) = 0 \qquad \Longleftrightarrow \vec{\mu} = \frac{1}{n}\sum_{i=1}^{n}\vec{x}_i \tag{2.23}$$

$$\frac{\partial D_{\mathrm{KL}}}{\partial \Sigma^{-1}} = \frac{1}{2n}\sum_{i=1}^{n}(\vec{x}_i-\vec{\mu})(\vec{x}_i-\vec{\mu})^\top - \frac{1}{2}\Sigma = 0 \qquad \Longleftrightarrow \Sigma = \frac{1}{n}\sum_{i=1}^{n}(\vec{x}_i-\vec{\mu})(\vec{x}_i-\vec{\mu})^\top \tag{2.24}$$

によって、パラメータの最適化を行うことができる。

2.5 よくあるボルツマンマシン

ボルツマンマシンのエネルギー関数はどのような形でも良いが、一般的にはエネルギー関数として Ising モデルを用いた

$$q_\theta(\vec{x}) = \frac{1}{Z_\theta}\exp\left(\sum_{k\neq l}J_{kl}x_kx_l + \sum_{k=1}^{N}h_kx_k\right), \quad \vec{x} = \{-1,1\}^N \tag{2.25}$$

$$Z_\theta = \sum_{\vec{x}}\exp\left(\sum_{k\neq l}J_{kl}x_kx_l + \sum_{k=1}^{N}h_kx_k\right) \tag{2.26}$$

が使われる。$\sum_{k\neq l}$ は異なる添字のペアについて 1 回ずつ足すことを意味しており、J_{kl} がランダムなモデルをスピングラスモデルと呼ぶ。このときの KL 情報量は

$$\begin{aligned} D_{\mathrm{KL}}(p_0 \parallel q_\theta) &= \sum_{\vec{x}} p_0(\vec{x})(\ln p_0(\vec{x}) - \ln q_\theta(\vec{x})) \\ &\underbrace{=}_{\theta\text{のみ}} -\sum_{\vec{x}} p_0(\vec{x})\left(\sum_{k\neq l}J_{kl}x_kx_l + \sum_{k=1}^{N}h_kx_k\right) + \ln Z_\theta \\ &= -\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{k\neq l}J_{kl}x_k^{(i)}x_l^{(i)} + \sum_{k=1}^{N}h_kx_k^{(i)}\right) + \ln Z_\theta \end{aligned} \tag{2.27}$$

である。パラメータの学習は、

$$\frac{\partial D_{\mathrm{KL}}}{\partial h_k} = -\frac{1}{n}\sum_{i=1}^{n}x_k^{(i)} + \sum_{\vec{x}}x_k\frac{\exp\left(\sum_{k\neq l}J_{kl}x_kx_l + \sum_{k=1}^{N}h_kx_k\right)}{Z_\theta} = 0 \qquad \Longleftrightarrow \langle x_k\rangle_\theta = \frac{1}{n}\sum_{i=1}^{n}x_k^{(i)} \tag{2.28}$$

$$\frac{\partial D_{\mathrm{KL}}}{\partial J_{kl}} = -\frac{1}{n}\sum_{i=1}^{n}x_k^{(i)}x_l^{(i)} + \sum_{\vec{x}}x_kx_l\frac{\exp\left(\sum_{k\neq l}J_{kl}x_kx_l + \sum_{k=1}^{N}h_kx_k\right)}{Z_\theta} = 0 \quad \Longleftrightarrow \langle x_kx_l\rangle_\theta = \frac{1}{n}\sum_{i=1}^{n}x_k^{(i)}x_l^{(i)} \tag{2.29}$$

となる。したがって、モデルの平均値とデータの平均値を同じにするようにパラメータを決定すれば良いことがわかるが、それを実現する J_{kl}, h_k の値は不明なままである。

2.6　勾配法と代理関数

Ising モデルにおけるパラメータの最適化は勾配法によって行うことができる。すなわち、

$$\begin{cases} h'_k = h_k - \eta_k \left(\dfrac{1}{n} \displaystyle\sum_{i=1}^{n} x_k^{(i)} - \langle x_k \rangle_{\theta=\{h,J\}} \right) \\ J'_{kl} = J_{kl} - \eta_{kl} \left(\dfrac{1}{n} \displaystyle\sum_{i=1}^{n} x_k^{(i)} x_l^{(i)} - \langle x_k x_l \rangle_{\theta=\{h,J\}} \right) \end{cases} \tag{2.30}$$

のようにする。$\theta = \{h, J\} \to \theta' = \{h', J'\}$ へ更新する作業を繰り返すことで、パラメータの最適化を行うことができる。なお、η_k, η_{kl} は勾配法の学習率と呼ばれ、小さな値が設定される。パラメータがある値に収束するように学習率を適切に設定する。実用的には学習率の値を適当に決めることが多いが、ここでは勾配法の理論を簡単に紹介する。

2.6.1　勾配法の仕組み

パラメータ θ で特徴づけられた微分可能な関数 $f(\theta)$ を最小化することを考える。勾配法とは、θ を徐々に動かして関数の最小値（極小値）を探索するアルゴリズムである。しかし、θ を動かした際に関数値が小さくならなければ最小値の探索を行うことができない。すなわち、現在のパラメータを θ_0 から、新しいパラメータ θ へ更新する際に、

$$f(\theta) \leq f(\theta_0) \tag{2.31}$$

が成立するようにしなければならない。直感的には、関数の傾きが下がっていく方向に少しだけ θ を変化させれば上記の条件を満たすと考えられる：

$$\theta = \theta_0 - \eta f'(\theta_0), \quad \eta \ll 1 \text{ 。} \tag{2.32}$$

この η の値をある程度見積もるために、$\theta = \theta_0$ 近傍で $f(\theta)$ を上から抑えることのできる簡単な形の関数を考える。例えば、$\theta = \theta_0$ で $f(\theta_0)$ と同じ値を持つ下に凸な二次関数で $f(\theta)$ を上から抑えることができれば、二次関数の頂点を与える θ は確実に $f(\theta) \leq f(\theta_0)$ を満たす。そのような 2 次関数は、$\theta = \theta_0$ で $f(\theta)$ と共通接線を持つ必要がある：

$$f(\theta) \leq f(\theta_0) + f'(\theta_0)(\theta - \theta_0) + \frac{1}{2} L (\theta - \theta_0)^2 \text{ 。} \tag{2.33}$$

ここで、L は十分大きな正の値である。これは $f(\theta)$ の θ_0 における Taylor 展開を 2 次で打ち切り、2 次の係数を調整したと考えることもできる。このとき、

$$f(\theta) \leq f(\theta_0) + \frac{1}{2} L \left(\theta - \theta_0 + \frac{1}{L} f'(\theta_0) \right)^2 - \frac{(f'(\theta_0))^2}{2L} \tag{2.34}$$

と平方完成できるため、頂点を与える $\theta = \theta_0 - f'(\theta_0)/L$ に対して、

$$f(\theta) \leq f(\theta_0) - \frac{(f'(\theta_0))^2}{2L} \tag{2.35}$$

となり、確実に関数の値が減少する。最小化したい関数に対して、簡単な形の上界の最小値を代わりに求めることが勾配法の原理である。そのため、学習率 η は

$$\eta = \frac{1}{L} \tag{2.36}$$

と求まる。

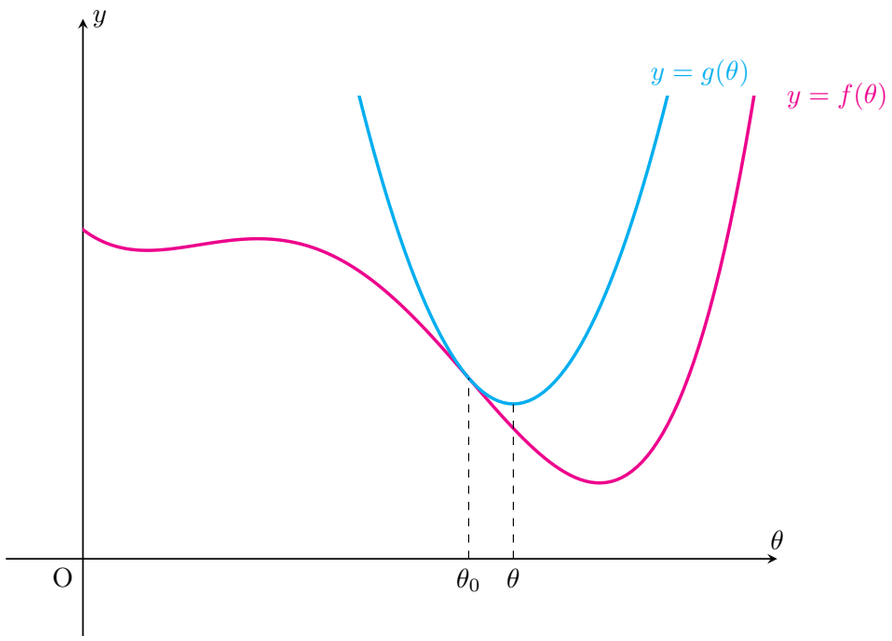

図 3: 勾配法の図形的な意味（最小化したい $y = f(x)$ について、$y = g(x)$ は局所的な上界となっている。）

2.6.2 代理関数

勾配法の原理は、上界最小化として理解されることを見た。先ほどの $g(\theta)$ のように、最適化問題において元の関数の代わりに最適化される関数を代理関数（surrogate function）と呼ぶ。最適化したいスカラー値関数 $f(\vec{\theta})$ に対して、

$$f_\eta(\vec{\theta}, \vec{\theta}_0) = f(\vec{\theta}_0) + \nabla f(\vec{\theta}_0) \cdot (\vec{\theta} - \vec{\theta}_0) + \frac{1}{2}(\vec{\theta} - \vec{\theta}_0)^\top \frac{1}{\eta}(\vec{\theta} - \vec{\theta}_0) \tag{2.37}$$

とする。η はスカラーである。このとき、η が満たすべき条件式は、パラメータ θ に対して

$$d(\vec{\theta}) = f(\vec{\theta}) - f_\eta(\vec{\theta}, \vec{\theta}_0) \leq 0 \tag{2.38}$$

が常に成立することである。特に、$d(\vec{\theta}_0) = 0$ であるため、$f_\eta(\vec{\theta}, \vec{\theta}_0) < f_\eta(\vec{\theta}_0, \vec{\theta}_0)$ を満たす θ に対して $f(\vec{\theta}) < f(\vec{\theta}_0)$ となることがわかる。

η についての条件を具体的に考える。$d(\vec{\theta}_0) = 0$ であるため、全ての成分について $\nabla d(\vec{\theta}) \leq 0$ が常に成立すれば良い。

$$\nabla d(\vec{\theta}) = \nabla f(\vec{\theta}) - \nabla f(\vec{\theta}_0) - \frac{1}{\eta}(\vec{\theta} - \vec{\theta}_0) \tag{2.39}$$

である。したがって、$\nabla d(\vec{\theta}_0) = 0$ であり、

$$(\nabla\nabla^\top) d(\vec{\theta}) = (\nabla\nabla^\top) f(\vec{\theta}) - \frac{1}{\eta} I = H(f) - \frac{1}{\eta} I \tag{2.40}$$

となる。なお、$H(f)$ は Hesse 行列である。したがって、任意のベクトル $\vec{a}$ について

$$\vec{a}^\top \left(H(f) - \frac{1}{\eta} I \right) \vec{a} \leq 0 \tag{2.41}$$

であれば、全ての成分について $\nabla d(\vec{\theta}) \leq 0$ となる。ここで、$f(\vec{\theta})$ は C^2 級とすると、$H(f)_{ij} = H(f)_{ji}$ から、

適当な直交行列 P を用いて $H(f)$ を対角化することができる。したがって、

$$\begin{cases} P^{-1}H(f)P & = \Lambda = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_N \end{pmatrix} \\ P^{-1}(\vec{a}) & = \vec{b} \end{cases} \tag{2.42}$$

とおくと、

$$\vec{b}^{\top}\left(\Lambda - \frac{1}{\eta}I\right)\vec{b} = \sum_{i=k}^{n}(\lambda_k - \frac{1}{\eta})b_k^2 \leq 0 \tag{2.43}$$

が成立すれば良い。したがって、学習率 η の十分条件は、

$$\eta \leq \frac{1}{\max_k \lambda_k} \tag{2.44}$$

である。

3　隠れ変数を導入してリッチなモデルへ

3.1　EM アルゴリズム

ボルツマンマシンにおいて、簡単に計算できるエネルギー関数はせいぜいガウス型までであった。しかし、このままでは単純なモデルしか生成することができない。そのため、異なる平均と分散を持つガウス関数の線形結合を取ったものをモデルとしてみる：

$$q_\theta(\vec{x}) = \sum_{k\in I}\frac{C_k}{\sqrt{(2\pi)^N \det(\Sigma_k)}}\exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu}_k)^{\top}\Sigma_k^{-1}(\vec{x}-\vec{\mu}_k)\right)。 \tag{3.1}$$

ここで、$k \in I$ はガウス分布 $\mathcal{N}(\vec{\mu}_k, \Sigma_k)$ のラベルとその集合を表している。また、規格化条件 $\int dx\, q_\theta(\vec{x}) = 1$ から、

$$\sum_{k\in I} C_k = 1 \tag{3.2}$$

が課されている。最尤法によりパラメータの最適化を行う。すなわち、

$$\max_\theta\left\{\frac{1}{n}\sum_{i=1}^{n}\ln q_\theta(\vec{x}_i)\right\} = \max_\theta\left\{\frac{1}{n}\sum_{i=1}^{n}\ln\left[\sum_{k\in I}\frac{C_k}{\sqrt{(2\pi)^N\det(\Sigma_k)}}\exp\left(-\frac{1}{2}(\vec{x}_i-\vec{\mu}_k)^{\top}\Sigma_k^{-1}(\vec{x}_i-\vec{\mu}_k)\right)\right]\right\} \tag{3.3}$$

を考える。まずは、$\vec{\mu}_k$ についての偏微分を考える。

$$\begin{aligned}\frac{\partial}{\partial\vec{\mu}_k} : \frac{1}{n}\sum_{i=1}^{n}&\underbrace{\frac{C_k\exp(\cdots)/\sqrt{(2\pi)^N\det(\Sigma_k)}}{\sum_{k\in I}C_k\exp(\cdots)/\sqrt{(2\pi)^N\det(\Sigma_k)}}}_{=\gamma_{ik}:\text{ 負担率}}\times\Sigma_k^{-1}(\vec{x}_i-\vec{\mu}_k) \\ &= \frac{1}{n}\sum_{i=1}^{n}\gamma_{ik}\Sigma_k^{-1}(\vec{x}_i-\vec{\mu}_k) = 0 \Longleftrightarrow \vec{\mu}_k = \frac{\sum_{i=1}^{n}\gamma_{ik}\vec{x}_i}{\sum_{i=1}^{n}\gamma_{ik}}\end{aligned} \tag{3.4}$$

i 番目のデータに関して、ラベル k のガウス分布の重み γ_{ik} を負担率と呼ぶ。負担率を用いて、$\vec{\mu}_k$ はデータの重み付き平均を取れば最適化できることがわかる。なお、負担率は $\theta = \{\vec{\mu}_k, \Sigma_k \mid k \in I\}$, $C = \{C_k \mid k \in I\}$ の関数となっている。次に、Σ_k についての偏微分を考える：

$$\frac{\partial}{\partial\Sigma_k^{-1}} : -\frac{1}{2n}\sum_{i=1}^{n}\gamma_{ik}(\vec{x}_i-\vec{\mu}_k)(\vec{x}_i-\vec{\mu}_k)^{\top} + \frac{1}{2n}\sum_{i=1}^{n}\gamma_{ik}\Sigma_k = 0。 \tag{3.5}$$

したがって、ラベル k のガウス分布の分散共分散行列の最適化は

$$\Sigma_k = \frac{\sum_{i=1}^{n} \gamma_{ik}(\vec{x}_i - \vec{\mu}_k)(\vec{x}_i - \vec{\mu}_k)^\top}{\sum_{i=1}^{n} \gamma_{ik}} \tag{3.6}$$

と求まる。しかし、これで終わりではない。線形結合の重みづけについての最適化も考えないといけない。重みづけに制約条件が課されているため、考えないといけない最適化問題は Lagrange の未定乗数 λ を含めた

$$\max_{\theta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ln q_\theta(\vec{x}_i) + \lambda \left(\sum_{k \in I} C_k - 1 \right) \right\} \tag{3.7}$$

である。しかし、先ほどの議論は未定乗数を含めても修正する必要はなく、C_k についての偏微分を新たに考えれば良い。すなわち、

$$\frac{\partial}{\partial C_k} : \frac{1}{n} \sum_{i=1}^{n} \frac{\gamma_{ik}}{C_k} + \lambda = 0 \Longleftrightarrow \lambda C_k = -\frac{1}{n} \sum_{i=1}^{n} \gamma_{ik} \tag{3.8}$$

となる。k についての和を取ると、

$$\lambda = -\frac{1}{n} \sum_{i=1}^{n} \sum_{k \in I} \gamma_{ik} \tag{3.9}$$

と求まるため、C_k について解きなおすと

$$C_k = \frac{\sum_{i=1}^{n} \gamma_{ik}}{\sum_{i=1}^{n} \sum_{k \in I} \gamma_{ik}} \tag{3.10}$$

と求まる。以上の結果から、繰り返しパラメータの更新を行うことで最適解を求めることができる。アルゴリズムとしては以下のようになる：

A. γ_{ik} を適当な値に初期化する
B. γ_{ik} を固定したまま、パラメータ θ, C を最適化する
C. B で最適化したパラメータ θ, C を用いて γ_{ik} を更新する
D. γ_{ik} が収束するまで B と C を繰り返す

これを EM(Expectation-Maximization) アルゴリズムと呼ぶ。

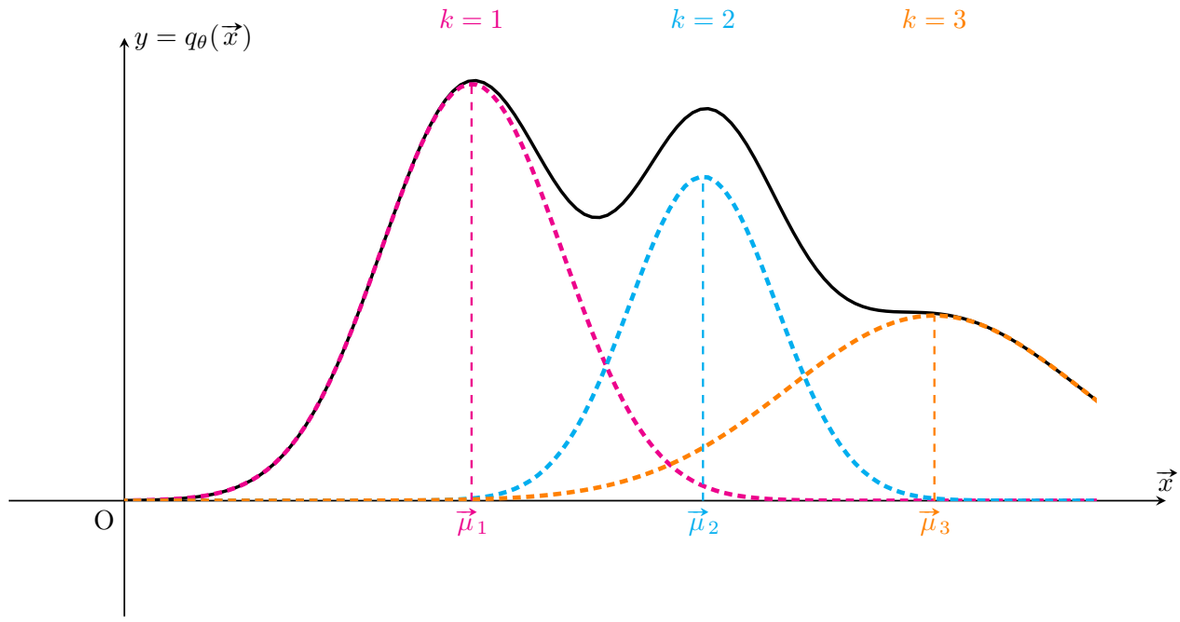

図 4: 3 つの混合ガウス分布によるモデル q_θ（実線）の概念図

3.2 隠れた変数

混合ガウス分布を考えることで、解析的な計算が可能なままモデルを複雑化することができた。このような考え方を一般化するために隠れ変数を導入する。混合ガウス分布は、複数のガウス分布の和によって構成されるモデルであるが、逆に何らかの $q_\theta(\vec{x})$ が与えられた際に、それを構成するそれぞれのガウス分布自身を知ることは難しい。すなわち、モデルを構成するガウス分布は隠れた変数であり、その値を知ることができない。そのため、データ $\vec{x}$ が従うモデル $q_\theta(\vec{x})$ は、隠れた変数 $\vec{z}$ によって生成されていると考える：

$$q_\theta(\vec{x}) = \sum_{\vec{z}} q_\theta(\vec{x}, \vec{z}) = \sum_{\vec{z}} q_\theta(\vec{x}|\vec{z}) q_\theta(\vec{z}) \ 。 \tag{3.11}$$

ここで、$q_\theta(\vec{x}, \vec{z})$ は $\vec{x}$ と $\vec{z}$ の結合確率であり、$q_\theta(\vec{x}|\vec{z})$ は $\vec{x}$ の $\vec{z}$ による条件付き確率である。先ほどの混合ガウス分布では、$\{\vec{\mu}_k, \Sigma_k\}$ を隠れた変数として生成されるモデルと考えることができる。これを踏まえて、混合ガウス分布を隠れた変数の考え方を用いて再考する。すなわち、$\vec{\mu}_k$, Σ_k はデータベクトル $\vec{x}$ の成分に依存しないとすると、

$$q_\theta(z) = C_z, \quad q_\theta(\vec{x}|z) = \frac{1}{\sqrt{(2\pi)^N \det(\Sigma)}} \exp\left(-\frac{1}{2} (\vec{x} - \vec{\mu}_z)^\top \Sigma_z^{-1} (\vec{x} - \vec{\mu}_z) \right) \tag{3.12}$$

と表すことができ、混合確率（モデル）は

$$q_\theta(\vec{x}, z) = \sum_z q_\theta(\vec{x}|z) q(z) = \sum_z \frac{C_z}{\sqrt{(2\pi)^N \det(\Sigma)}} \exp\left(-\frac{1}{2} (\vec{x} - \vec{\mu}_z)^\top \Sigma_z^{-1} (\vec{x} - \vec{\mu}_z) \right) \tag{3.13}$$

となる。EM アルゴリズムにおいて、パラメータの偏微分を行って最適化を行った。偏微分による方法を代理関数の下界最大化という立場から考察する。すなわち、対数尤度関数の代理関数を見つける必要がある。しかし、混合ガウス分布のモデルのように隠れた変数についての和でモデルが表現されている場合、対数の中に和が出てきて計算することができない。そのため、

$$\ln q_\theta(\vec{x}) = \ln \frac{q_\theta(\vec{x}, z)}{q_\theta(z|\vec{x})} \tag{3.14}$$

と変形する。なお、$q_\theta(\vec{x}, z) = q_\theta(z, \vec{x})$ を用いた。右辺にある z の条件付き確率は未知であるため計算できない。しかし、適当な分布 $r(z)$ に従う乱数を用いて期待値を計算してみる：

$$\sum_z r(z) \ln \frac{q_\theta(\vec{x}, z)}{q_\theta(z|\vec{x})} = \sum_z r(z) \ln \frac{q_\theta(\vec{x}, z)}{r(z)} - \sum_z r(z) \ln \frac{q_\theta(z|\vec{x})}{r(z)} \ 。 \tag{3.15}$$

すると、右辺第二項が変数 z に関する KL 情報量の形になっていることがわかる。したがって、

$$\ln q_\theta(\vec{x}) = \sum_z r(z) \ln \frac{q_\theta(\vec{x}, z)}{r(z)} + D_{\mathrm{KL}}(r(z) \parallel q_\theta(z|\vec{x})) \geq \sum_z r(z) \ln \frac{q_\theta(\vec{x}, z)}{r(z)} \tag{3.16}$$

が成立する。すなわち、代理関数として

$$\sum_z r(z) \ln \frac{q_\theta(\vec{x}, z)}{r(z)} \tag{3.17}$$

を最大化すれば良いことがわかる。代理関数は、期待値計算における確率分布 $r(z)$ とパラメータ θ に関して最大化することができる。そのため、まずは

$$r(z) = q_\theta(z|\vec{x}) \tag{3.18}$$

として、最大化したい関数と代理関数のギャップを埋める：

$$\ln q_\theta(\vec{x}) = \sum_z q_\theta(z|\vec{x}) \ln \frac{q_\theta(\vec{x}, z)}{q_\theta(z|\vec{x})} \ 。 \tag{3.19}$$

右辺は対数の中に和が出てこない形をしているため、パラメータの偏微分が簡単に計算できる。すなわち、対数尤度関数の最大化は、$r(z) = q_\theta(z|\vec{x})$ のパラメータ θ を固定した状態で

$$\max_\theta \left(\frac{1}{n} \sum_{i=1}^{n} \ln q_\theta(\vec{x}_i) \right) = \max_\theta \left(\frac{1}{n} \sum_{i=1}^{n} \sum_z q_\theta(z|\vec{x}_i) \ln \frac{q_\theta(\vec{x}_i, z)}{q_\theta(z|\vec{x}_i)} \right) \tag{3.20}$$

を考えることになる。これは Q 関数

$$Q(\theta, \theta_0) \coloneqq \frac{1}{n} \sum_{i=1}^{n} \sum_z q_{\theta_0}(z|\vec{x}_i) \ln q_\theta(\vec{x}_i, z) \tag{3.21}$$

を θ について最大化することと等しい。なお、

$$q_\theta(z|\vec{x}_i) = \frac{q_\theta(\vec{x}_i, z)}{q_\theta(\vec{x}_i)} = \frac{q_\theta(\vec{x}_i, z)}{\sum_z q_\theta(\vec{x}_i, z)} = \frac{C_z \exp(\cdots)}{\sum_z C_z \exp(\cdots)} = \gamma_{iz} \tag{3.22}$$

となることがわかる。すなわち、$q_\theta(z|\vec{x})$ を固定したままパラメータ θ を最適化することは、EM アルゴリズムにおいて負担率 γ_{ik} を固定したままパラメータ θ を最適化したことに対応している。そのため、Q 関数の最大化が EM アルゴリズムのパラメータ更新則と同じものを導くことが期待される。実際に、

$$Q(\theta, \theta_0) = \frac{1}{n} \sum_{i=1}^{n} \sum_z \gamma_{iz} \left[\left(-\frac{1}{2} (\vec{x}_i - \vec{\mu}_z)^\top \Sigma_z^{-1} (\vec{x}_i - \vec{\mu}_z) + \frac{1}{2} \det\left(\Sigma^{-1}\right) \right) + \ln C_z \right] \tag{3.23}$$

であることを用いると、

$$\frac{\partial Q(\theta, \theta_0)}{\partial \vec{\mu}_z} = \frac{1}{n} \sum_{i=1}^{n} \gamma_{iz} \left(\Sigma_z^{-1} (\vec{x}_i - \vec{\mu}_z) \right) = 0 \qquad \Longleftrightarrow \vec{\mu}_z = \frac{\sum_{i=1}^{n} \gamma_{iz} \vec{x}_i}{\sum_{i=1}^{z} \gamma_{iz}} \tag{3.24}$$

$$\frac{\partial Q(\theta, \theta_0)}{\partial \Sigma_z^{-1}} = \frac{1}{n} \sum_{i=1}^{n} \gamma_{iz} \left(-\frac{1}{2} (\vec{x}_i - \vec{\mu}_z)(\vec{x}_i - \vec{\mu}_z)^\top + \frac{1}{2} \Sigma_z \right) = 0 \quad \Longleftrightarrow \Sigma_z = \frac{\sum_{i=1}^{n} \gamma_{iz} (\vec{x}_i - \vec{\mu}_z)(\vec{x}_i - \vec{\mu}_z)^\top}{\sum_{i=1}^{z} \gamma_{iz}} \tag{3.25}$$

となり、EM アルゴリズムの結果と一致する。また、隠れた変数の最適化も行うことができる。これは、混合ガウス分布における重みづけの係数 $q_\theta(z) = C_z$ の最適化のことである。一般に、隠れた変数の束縛条件

$$\sum_{\vec{z}} q_\theta(\vec{z}) = 1 \tag{3.26}$$

が成立しているため、拘束付き最適化問題

$$\max_{C_z} \left(Q(\theta, \theta_0) + \lambda \left(\sum_z C_z - 1 \right) \right) =: \max_{C_z} (Q(\theta,\, \theta_0; C_z)) \tag{3.27}$$

を解けば良い。

$$\frac{\partial Q(\theta, \theta_0; C_z)}{\partial C_z} = \frac{1}{n} \sum_{i=1}^{n} \frac{\gamma_{iz}}{C_z} + \lambda = 0 \Longleftrightarrow \lambda C_z = -\frac{1}{n} \sum_{i=1}^{n} \gamma_{iz} \tag{3.28}$$

となる。z についての和を取ると、

$$\lambda = -\frac{1}{n} \sum_{i=1}^{n} \sum_z \gamma_{iz} \tag{3.29}$$

と求まるため、C_z について解きなおすと

$$C_z = \frac{\sum_{i=1}^{n} \gamma_{iz}}{\sum_{i=1}^{n} \sum_z \gamma_{iz}} \tag{3.30}$$

と求まる。このように、隠れた変数の期待値 (Expectation) を取る乱数の最適化とパラメータ θ による対数尤度関数の最大化 (Maximization) を繰り返すことでモデルの最適化が行える。これが EM アルゴリズムと呼ばれる理由である。

4　制限ボルツマンマシン

計算が可能な範囲でモデルを複雑にするためには、隠れた変数を導入すれば良いということが混合ガウス分布モデルの考察でわかった。隠れた変数をボルツマンマシンで扱えるようにしたものを制限ボルツマンマシンと呼ぶ。データ変数を $\vec{x}$、隠れた変数を $\vec{z}$ とする。このとき、ボルツマンマシンにおいてエネルギー関数が

$$-E_\theta(\vec{x},\vec{z}) = \vec{x}^\top W\vec{z} + \vec{b}^\top\vec{x} + \vec{c}^\top\vec{z} \tag{4.1}$$

と表されるものを制限ボルツマンマシンと呼ぶ。これは、データ変数同士の相互作用と隠れた変数同士の相関がなく、データ変数と隠れた変数間の結合だけが存在するようなモデルと考えることができる。例えば、データと隠れた変数が

$$\{\vec{x}_1,\,\vec{x}_2,\,\vec{x}_3,\,\vec{x}_4\}:\text{データ}\qquad\{\vec{z}_1,\,\vec{z}_2,\,\vec{z}_3\}:\text{隠れた変数} \tag{4.2}$$

とすると、結合の強さ W の制限ボルツマンマシンは以下のようなグラフ構造となる：

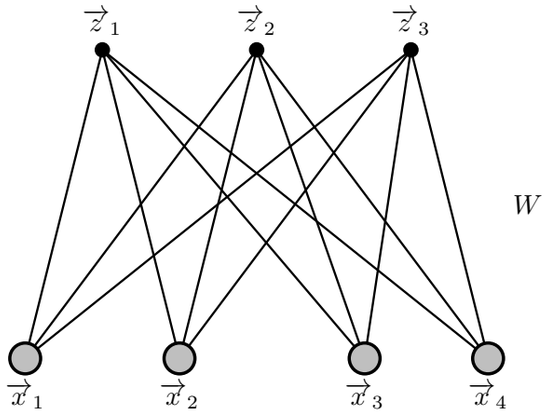

図 5: 制限ボルツマンマシンのグラフ構造

この制限ボルツマンマシンのモデルは

$$q_\theta(\vec{x},\vec{z}) = \frac{1}{Z}\exp\Big(\vec{x}^\top W\vec{z} + \vec{b}^\top\vec{x} + \vec{c}^\top\vec{z}\Big) \tag{4.3}$$

である。データの次元を N、隠れた変数の次元を N_h とする。$k=1,2,\cdots,N$, $l=1,2,\cdots,N_h$ に対して、$\vec{u}_k^\top$ を W の k 行目を取り出した横ベクトル、$\vec{v}_l$ を W の l 行目を取り出した縦ベクトルとすると、

$$q_\theta(\vec{z}) = \sum_{\vec{x}=\{1,-1\}^N} q_\theta(\vec{x},\vec{z}) = \frac{1}{Z}\prod_{k=1}^{N} 2\cosh\big(\vec{u}_k^\top\vec{z} + b_k\big)e^{\vec{c}^\top\vec{z}} \tag{4.4}$$

$$q_\theta(\vec{x}) = \sum_{\vec{z}=\{1,-1\}^{N_h}} q_\theta(\vec{x},\vec{z}) = \frac{1}{Z}\prod_{l=1}^{N_h} 2\cosh\big(\vec{v}_l^\top\vec{x} + c_l\big)e^{\vec{b}^\top\vec{x}} \tag{4.5}$$

が成立する。したがって、$\vec{x}$ の条件付き確率は

$$q_\theta(\vec{x}|\vec{z}) = \frac{q_\theta(\vec{x},\vec{z})}{q_\theta(\vec{z})} = \prod_{k=1}^{N}\frac{\exp\big(x_k\vec{u}_k^\top\vec{z} + b_kx_k\big)}{2\cosh\big(\vec{u}_k^\top\vec{z} + b_k\big)} = \prod_{k=1}^{N} q_\theta(x_k|\vec{z}) \tag{4.6}$$

と計算できる。このように、条件付き確率が各データについての積に分解できるとき、条件付き独立と呼ぶ。同様にして、$\vec{z}$ の条件付き確率は

$$q_\theta(\vec{z}|\vec{x}) = \frac{q_\theta(\vec{x},\vec{z})}{q_\theta(\vec{x})} = \prod_{l=1}^{N_h}\frac{\exp\big(z_l\vec{v}_l^\top\vec{x} + c_lz_l\big)}{2\cosh\big(\vec{v}_l^\top\vec{x} + c_l\big)} = \prod_{l=1}^{N_h} q_\theta(z_l|\vec{x}) \tag{4.7}$$

と計算出来るため、条件付き独立となっている。$q_\theta(\vec{z}|\vec{x})$ とは、データ $\vec{x}$ が与えられたときに隠れた変数 $\vec{z}$ が従う確率分布であった。条件付き独立が成立していることから、それぞれの隠れた変数 $\vec{z}_{l\in N_h}$ を確率分布 $q_\theta(z_l|\vec{x})$ に従ってサンプリングを行うことができる。

5　マルコフ連鎖モンテカルロ法（MCMC）

6　交換モンテカルロ法

7　MCMC を用いないボルツマンマシン

8　変分オートエンコーダー

9　階層変分オートエンコーダ

10 Appendix

10.1 微分に必要な線形代数の公式その 1

$\vec{x}$ を n 次元ベクトル、A を n 次正方行列とする。このとき、$\vec{x}^\top A\vec{x}$ はスカラーであるため、

$$\frac{\mathrm{d}}{\mathrm{d}\vec{x}}\left(\vec{x}^\top A\vec{x}\right) = \frac{\mathrm{d}}{\mathrm{d}\vec{x}}\,\mathrm{tr}(\vec{x}^\top A\vec{x}) = \frac{\mathrm{d}}{\mathrm{d}\vec{x}}\,\mathrm{tr}\left(A\vec{x}\,\vec{x}^\top\right) = \frac{\mathrm{d}}{\mathrm{d}\vec{x}}\sum_{i,j=1}^{n} A_{ij}(\vec{x}\,\vec{x}^\top)_{ji}$$

$$= \sum_{i=1}^{n}\boldsymbol{e}_i\frac{\mathrm{d}}{\mathrm{d}x_i}\left(\sum_{k,l=1}^{n} A_{kl}x_l x_k\right) = \sum_{i=1}^{n}\boldsymbol{e}_i\left(\sum_{k=1}^{n} A_{ki}x_k + \sum_{l=1}^{n} A_{il}x_l\right) = (A + A^\top)\vec{x} \tag{10.1}$$

が成立する。なお、$\boldsymbol{e}_i$ を i 番目の基底ベクトルとして、$\mathrm{d}/\mathrm{d}\vec{x} = \sum_{i=1}^{n}\boldsymbol{e}_i\,\mathrm{d}/\mathrm{d}x_i$ であることを用いた。分散共分散行列は対称行列であるため、

$$\frac{\mathrm{d}}{\mathrm{d}\vec{x}}\left(\vec{x}^\top\Sigma^{-1}\vec{x}\right) = 2\Sigma^{-1}\vec{x} \tag{10.2}$$

となる。

10.2 微分に必要な線形代数の公式その 2

$$\frac{\mathrm{d}}{\mathrm{d}A_{ij}}\left(\vec{x}^\top A\vec{x}\right) = \frac{\mathrm{d}}{\mathrm{d}A_{ij}}\,\mathrm{tr}\left(A\vec{x}\,\vec{x}^\top\right) = \frac{\mathrm{d}}{\mathrm{d}A_{ij}}\sum_{i,j=1}^{n} A_{ij}(\vec{x}\,\vec{x}^\top)_{ji} = x_j x_i = (\vec{x}\,\vec{x}^\top)^\top_{ij} \tag{10.3}$$

が成立するため、

$$\frac{\mathrm{d}}{\mathrm{d}A}\left(\vec{x}^\top A\vec{x}\right) = \vec{x}\,\vec{x}^\top \tag{10.4}$$

となる。

10.3 微分に必要な線形代数の公式その 3

determinant の余因子展開

$$\det(A) = \sum_{j=1}^{n}(-1)^{i+j}A_{ij}m_{ij} \quad (m_{ij}\text{は } A \text{ の } (i,j) \text{ 小行列式}) \tag{10.5}$$

を用いると、

$$\left(\frac{\mathrm{d}}{\mathrm{d}A}\det(A)\right)_{ij} = \frac{\mathrm{d}}{\mathrm{d}A_{ij}}\sum_{j=1}^{n}(-1)^{i+j}A_{ij}m_{ij} = (-1)^{i+j}m_{ij} \tag{10.6}$$

となる。さらに、余因子行列 $\tilde{A}$ を用いて $\det(A) = A\tilde{A}$ と表されるため、

$$\frac{\mathrm{d}}{\mathrm{d}A}\ln\det(A) = \frac{1}{\det(A)}\frac{\mathrm{d}}{\mathrm{d}A}\det(A) = \frac{1}{\det(A)}\tilde{A}^\top = (A^{-1})^\top \tag{10.7}$$

を得る。したがって、

$$\frac{\mathrm{d}}{\mathrm{d}\Sigma^{-1}}\ln\det(\Sigma) = -\Sigma^\top = -\Sigma \tag{10.8}$$

となる。